ECOGRAPHY

Research

Testing whether ensemble modelling is advantageous for maximising predictive performance of species distribution models

Tianxiao Hao, Jane Elith, José J. Lahoz-Monfort and Gurutzeta Guillera-Arroita

T. Hao (http://orcid.org/0000-0003-4363-1956) ✉ (tianxiaoh@student.unimelb.edu.au), J. Elith (http://orcid.org/0000-0002-8706-0326), J. J. Lahoz-Monfort (http://orcid.org/0000-0002-0845-7035) and G. Guillera-Arroita (http://orcid.org/0000-0002-8387-5739), School of BioSciences, The Univ. of Melbourne, Parkville, VIC 3010, Australia.

Ecography
43: 549–558, 2020
doi: 10.1111/ecog.04890

Subject Editor: Dan Warren
Editor-in-Chief: Hanna Tuomisto
Accepted 29 November 2019

Predictive performance is important to many applications of species distribution models (SDMs). The SDM 'ensemble' approach, which combines predictions across different modelling methods, is believed to improve predictive performance, and is used in many recent SDM studies. Here, we aim to compare the predictive performance of ensemble species distribution models to that of individual models, using a large presence–absence dataset of eucalypt tree species. To test model performance, we divided our dataset into calibration and evaluation folds using two spatial blocking strategies (checkerboard-pattern and latitudinal slicing). We calibrated and cross-validated all models within the calibration folds, using both repeated random division of data (a common approach) and spatial blocking. Ensembles were built using the software package 'biomod2', with standard ('untuned') settings. Boosted regression tree (BRT) models were also fitted to the same data, tuned according to published procedures. We then used evaluation folds to compare ensembles against both their component untuned individual models, and against the BRTs. We used area under the receiver-operating characteristic curve (AUC) and log-likelihood for assessing model performance. In all our tests, ensemble models performed well, but not consistently better than their component untuned individual models or tuned BRTs across all tests. Moreover, choosing untuned individual models with best cross-validation performance also yielded good external performance, with blocked cross-validation proving better suited for this choice, in this study, than repeated random cross-validation. The latitudinal slice test was only possible for four species; this showed some individual models, and particularly the tuned one, performing better than ensembles. This study shows no particular benefit to using ensembles over individual tuned models. It also suggests that further robust testing of performance is required for situations where models are used to predict to distant places or environments.

Keywords: BIOMOD, block cross-validation, consensus forecast, model performance, model tuning, spatial autocorrelation, spatial blocking

NORDIC SOCIETY OIKOS

www.ecography.org

Introduction

Species distribution models (SDMs), also known as ecological niche models or habitat suitability models, are models that fit species–environment relationships to explain and predict distributions of species. SDMs are important tools in ecology, widely used in applications including exploring ecological and evolutionary hypotheses, invasive species management, reserve planning and predicting the impact of past and future climate change on species and communities (Guisan and Thuiller 2005, Guillera-Arroita et al. 2015, Guisan et al. 2017). A range of modelling algorithms are now available for building SDMs (e.g. generalised linear models, regression trees Maxent; Elith et al. 2006). It is not necessarily straightforward for users of SDMs to decide which algorithm is optimal for their situation (Elith and Graham 2009). In the absence of targeted application-specific metrics of model performance, SDM users often seek the algorithm that yields most accurate predictions. However, the relative predictive performance of algorithms is often situation-dependent, and past algorithm comparisons have not been able to identify any single class of algorithms as consistently superior at prediction (Segurado and Araújo 2004, Elith et al. 2006, Pearson et al. 2006). Rather than choosing a single modelling method ('individual model' hereafter), it has been suggested that one could instead build multiple models using different modelling methods, and combine predictions from these models to produce together an 'ensemble' prediction to achieve better prediction (Araújo and New 2007). We note here that this is a particular use of the word ensemble: it considers ensembles across modelling methods. This excludes machine learning methods such as random forests and boosted regression trees (Hastie et al. 2009) which are in one sense an ensemble, but are different conceptually and based on just one model type (decision trees). Hence in this paper we consider random forests and boosted regression trees as 'individual' models.

Ensemble modelling is widely applied by SDM users (Araújo and New 2007, Hao et al. 2019a) and many modellers believe ensemble models are superior for prediction tasks compared to individual models (Hao et al. 2019a). However, there is limited empirical investigation about how well ensemble models predict compared to individual SDMs. Two studies (Crimmins et al. 2013, Zhu and Peterson 2017) specifically investigated the relative performance of ensemble and individual models when used to predict into novel environments (also referred to as 'model transfer'), as such use is popular among ensemble modellers. Using species data independent from original data for validating model predictions, both studies found ensemble models to perform no better than individual models. In contrast, Marmion et al. (2009) found ensemble models to perform better than individual models when validated using a subset of the full dataset, providing evidence that ensembles can predict well to withheld data within the same space and time as the training data. Given that the structure and source of validation data can strongly impact the assessment of model performance (Roberts et al. 2017), and the limited testing of ensemble models to date, it can be argued that our understanding about ensemble performance in different settings is still fairly limited.

Here, we aim to contribute to this knowledge gap by testing the predictive performance of widely-used ensemble and individual models using a large set of tree occurrence data in southeast Australia. Rather than using the common strategy of random partitioning of data into calibration and validation subsets (Marmion et al. 2009), in this study we apply spatial blocking. Spatial blocking involves partitioning the data into mutually-exclusive spatial blocks, which are used to either calibrate or validate models. The advantage of spatial blocking over random partitioning is that test data are spatially more distant from training data, and therefore likely more independent (Roberts et al. 2017, Valavi et al. 2019). Spatial dependence (spatial autocorrelation, SAC) can occur in both species and predictor data, because neighbouring sites are more likely to experience similar environmental conditions and there may be spatial dependence in biological processes (Roberts et al. 2017). Models might overfit these dependencies, and random splits may not reveal such overfitting (Roberts et al. 2017). Hence spatial blocks are useful, enabling a thorough comparison of ensembles (expected to be complex), other complex models and simpler ones (sensu Merow et al. 2014). Through this study, we aim to complement the existing knowledge on performance of ensemble models versus individual models and investigate how spatial blocking affects our understanding of model performance.

Methods

Species and predictor data

Species data are from a large (32 256 sites) presence–absence dataset of 36 eucalypt tree species in New South Wales, Australia. These were obtained from vegetation survey data stored in the Flora Survey Module of the Atlas of NSW Wildlife, Office of Environment and Heritage. Creating spatial blocks for species with few presence records is difficult. To retain a reasonable amount of data to calibrate and validate models, we omitted all species with fewer than 500 presences in the full dataset, resulting in a selection of 14 tree species (13 in the genus *Eucalyptus* and one in the genus *Corymbia*). These 14 species were represented by 534 to 2003 presence records (see Table 1 for number of presence records and Supplementary material Appendix 1 for maps of records). For covariate data, we used 11 predictor variables at nine arc-second (~250 × 250 m) raster resolution (details in Supplementary material Appendix 1). Both species and predictor data were compiled and used previously in Fithian et al. (2015), and made openly available with that paper.

Spatial blocking

We devised spatial blocks in two different ways described below.

Table 1. Number of presence records for all species; all species are modelled under 'checkerboard' blocking design, and names of those species also modelled under 'latitudinal' blocking design are in bold.

Species	Abbreviation	No. of presences
Eucalyptus dives	eucadive	905
Eucalyptus pauciflora	eucapauc	1094
Eucalyptus agglomerata	eucaaggl	1025
Eucalyptus cypellocarpa	eucacype	1290
Eucalyptus fastigata	eucafast	753
Eucalyptus obliqua	eucaobli	953
Eucalyptus pilularis	eucapilu	1773
Eucalyptus piperita	eucapipe	1762
Eucalyptus robusta	eucarobu	534
Eucalyptus sieberi	eucasieb	2003
Eucalyptus moluccana	eucamolu	804
Eucalyptus punctata	eucapunc	2102
Eucalyptus rossii	eucaross	613
Corymbia maculata	corymacu	1387

Checkerboard blocking

In our first strategy for blocking, we divided our study region into a 'checkerboard' pattern with 53 equal-sized square blocks, each with a size of ~83 × 83 km, using the 'blockCV' package (Valavi et al. 2019) within the R ver. 3.4.1 statistical language environment (<www.r-project.org>). The block size was selected using the 'spatialAutoRange' function in 'blockCV', which suggests block sizes based on the median range of spatial autocorrelation in all predictor variables. The 53 blocks were then allocated into five roughly equal-sized folds, with similar number of presence and absence records in each fold (using 'selection = random' argument in the function 'spatialBlock' in 'blockCV'; for an example of fold arrangement see Supplementary material Appendix 1 Fig. A2). For each modelling run, we reserved one fold as the external fold. These data were strictly kept away from the model calibrating process, and only used to validate external model predictions. We used the remaining four folds (internal folds from here on) to calibrate and internally cross-validate models. In total five modelling runs were completed for each species and method, by cycling through the data, with each fold acting as the external fold exactly once.

Latitudinal blocking

The checkerboard design tests prediction to different parts of the landscape to some extent, but alternative designs can enforce testing in more geographically distant areas. This is useful, because ensemble modelling is a popular approach for predicting to distant areas (e.g. invasion or range expansion), justified by a widely stated view that ensembles are superior at such prediction tasks (Hao et al. 2019a). Because our study area experiences warmer temperatures in the north and vice versa, here we tested how well models predict to latitudinal extremities within the study area. We evenly divided the landscape into five latitudinal slices between latitudes −38 and −28, each slice with a width of 2° latitude (~222 km). We then used either the northernmost or the southernmost slice as the external fold and the remaining four folds to calibrate and internally cross-validate models. We applied this approach to only four species – those whose distribution spanned all five slices: *Corymbia maculata*, *Eucalyptus obliqua*, *Eucalyptus pauciflora* and *Eucalyptus pilularis* (ranging from 24 to 809 presence records per species per slice). We tested whether these slices require models to extrapolate using multivariate environmental similarity surface (MESS, Supplementary material Appendix 2).

Individual models

We used eight different approaches to build individual models, including popular methods from both statistics and machine learning. The models, and their acronyms, are detailed in Table 2. We built these individual models using the 'biomod2' package ver.3.3-7 (Thuiller et al. 2009) for the R statistical language. 'biomod2' provides a range of methods and functionalities relevant to the problem of modelling distributions, including a streamlined framework for building ensemble SDMs. Ensemble SDM workflows can be carried out using other toolkits such as BioEnsembles (Diniz-Filho et al. 2009) or without specialised tool sets (Hardy et al. 2011, Crimmins et al. 2013). However, we choose 'biomod2' because it is most popular among ensemble SDM users and is therefore representative of the typical ensemble SDM workflow (Hao et al. 2019a).

The configurations of our selected modelling methods are governed by tuning choices, including settings for how each model is fitted and what terms are allowed in the model. For instance, for a GLM, choices include the complexity of basis expansions allowed in the model (e.g. linear, quadratic or cubic terms; Hastie et al. 2009), and the approach used for selecting the final model (e.g. stepwise selection, or use of a full model). In practice, the evidence points towards common use of default tunings with biomod2 (Hao et al. 2019a). To be consistent with that, here we also used 'biomod2' in-built functions to build individual models as well as 'biomod2' default tuning choices (see our archived data for our R code, including tuning parameters used).

The tuning of model algorithms is known to affect predictive performance (Elith et al. 2008, Hastie et al. 2009, Merow et al. 2013). Here, we wish to briefly explore whether default tuning choices in 'biomod2' are optimal for our dataset or if they can be improved. We wish to also test how tuned models compare against ensembles of models with 'biomod2' default tunings. To this aim, for each modelling run, we tuned a BRT model (tuned BRT hereafter), using the package 'dismo' (Hijmans et al. 2017) in R and advice from Elith et al. (2008), and compared its predictive performance with 'biomod2' models. We chose to tune only the BRT model because we are experienced with the algorithm, and tuning all individual models is beyond the scope of this study. In a BRT model, the main tuning parameters include learning rate, depth of tree, bag fraction and the total number of trees in the ensemble (for an explanation see Elith et al. 2008). Our tuned BRT model is set with a slow learning rate (0.002), relatively

Table 2. Individual models used to build ensemble models.

Model	Abbreviation	Overview of model fitting in biomod2*	R packages called
Generalised linear model	GLM	A regression model that fits quadratic response curves with no interactions between covariates, with stepwise backward selection using Akaike's information criterion.	glm
Generalised additive model	GAM	A regression model that fits smoothed additive response curves through the mgcv package, allowing no interactions between covariates.	gam, mgcv
Multivariate adaptive regression splines	MARS	Similar to a GAM, MARS can fit complex response curves by joining together linear segments. The defaults allow no interaction terms, and use default penalties from the earth package.	earth
Artificial neural networks	ANN	A single-hidden-layer neural network that uses a five-fold internal cross-validation to choose the best number of units in the hidden layer and weight decay. These two parameters control model complexity.	nnet
Classification tree analysis	CTA	A decision tree model fitted with default settings in the underlying rpart package. Under biomod2 defaults it fits complex trees with many nodes. A five-fold internal cross-validation is used to choose the best model.	rpart
Flexible discriminant analysis	FDA	This method first fits a MARS model (fitted through mda package) then performs dimensionality reduction before attempting classification.	mda
Random forest	RF	A machine-learning method that ensembles predictions from 500 classification trees, fitted on randomly selected subsets of all training data. Individual trees are controlled to have at least five data points in their terminal nodes, but are otherwise allowed to grow as many nodes as possible.	randomForest
Boosted regression trees	BRT	A machine-learning method that ensembles regression trees through gradient boosting. A maximum of 2500 relatively deep trees are fitted, and best iteration of trees is selected using an internal three-fold cross-validation.	gbm

* Using default tuning in 'biomod2' package ver.3.3-7; no models use spatial terms. Further information on default settings can be viewed using the function `Print_Default_ModelingOptions()` in 'biomod2'.

deep trees (5), bag fraction of 0.75 and a maximum of 20 000 trees. The optimal number of trees is estimated in the 'gbm.step' function in 'dismo', a method that gradually adds more trees to the ensemble until optimal prediction on cross-validated data is achieved. We chose these settings because our presence–absence dataset is large, thus it likely can support relatively complex relationships (i.e. deep trees), and because these settings allowed us to fit at least 1000 trees, which is desirable to reduce variance between different runs of these stochastic models (Elith et al. 2008). This contrasts with the 'biomod2' ver. 3.3-7 defaults, which use the 'gbm' package, has a slow learning rate (0.001) and deep tree depth (7), uses a bag fraction of 0.5 and only allows up to 2500 trees. In 'biomod2' the optimal number of trees are then chosen by a three-fold cross-validation across the set fitted.

Internal cross-validation

Cross-validation (CV) is often used by modellers to understand the performance of their models and to build a special class of 'weighted' ensemble models popular among 'biomod2' users. CV divides data multiple times into different subsets used for calibrating (i.e. fitting or training) and validating (i.e. testing) models. Although CV is thought to be less ideal than independent validation as a test of true model performance (Marmion et al. 2009, Crimmins et al. 2013), we believe it is nevertheless important to investigate what information on model performance a typical ensemble modeller would gain from using CV. Therefore, we performed CV in all modelling runs in both spatial blocking designs, using only the internal folds. To distinguish our CV from external validations using the external fold, we will refer to CVs as 'internal' to reflect the use of internal folds only, and to cross-validated models as 'internal models' (see Fig. 1 for an illustration of how we divided data for internal and external models). We designed internal CVs in two different ways. Firstly, we randomly selected 75% of the samples in the internal folds for calibrating the internal models, and used the remaining 25% to validate them; we repeat this process four times in each modelling run. This repeated random CV strategy is provided by 'biomod2' as a default option, and is widely used in 'biomod2' studies (Hao et al. 2019a). Therefore, we included it to represent the model performance information that typical 'biomod2' users obtain, and to assess whether it yields different results than those obtained from spatially blocked validation methods. For our second internal CV, we used the existing spatially-blocked folds within internal folds by using three internal folds to calibrate models and the remaining one for validation. We cycled through the four internal folds until each fold had been used for validation once, effectively conducting a four-fold CV (Hastie et al. 2009). In both CV designs, we averaged model performance statistics across the four repeated runs/folds, using area under the curve (AUC) of the receiver-operating characteristic plot (Fielding and Bell 1997) as a metric of predictive performance. AUC is a confusion matrix-derived measure of discrimination, reflecting how successful models are at correctly discriminating presences from absences. It is extensively

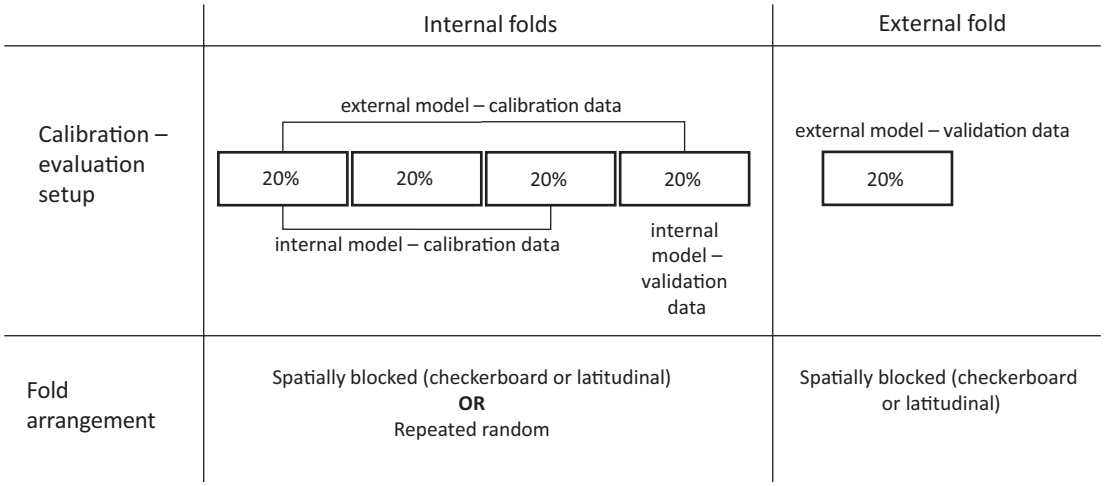

Figure 1. Schematic illustrating the amount (in % of full data) and arrangement of data used for calibration and validation of internal and external models.

employed by SDM users, including most ensemble modelling studies (Hao et al. 2019a). AUC values range between 0 and 1, with 0.5 representing random discrimination and one representing perfect discrimination. We did not subject tuned BRT models to internal CVs as we aimed to compare their performance against other models in external validations (rather than to understand how their performance would be perceived by typical 'biomod2' users using CV).

Ensemble models

We built ensembles using the two most popular methods amongst 'biomod2' users (Hao et al. 2019a) – Mean, and weighted average (WA). Mean produces the ensemble prediction by averaging predictions across individual models. WA also averages predictions, but weights them based on CV performance of individual models (so the ensemble is more strongly influenced by models performing well on CV). For our WA ensembles, we weighted the model predictions based on their internal cross-validation AUC, with weights calculated as in eq. 1 (Hartley et al. 2006, Marmion et al. 2009):

$$\text{WAprediction}_i = \frac{\sum_j \left(\text{AUC}_j \times \text{prediction}_{ij}\right)}{\sum_j \text{AUC}_j} \quad (1)$$

that is, for a given site i, the WA ensemble prediction, WAprediction_i, is calculated as the sum of predictions for site i across j individual models weighted by their respective AUC, AUC_j and normalized by the sum of all AUCs.

For each modelling run, we produced one Mean and two WA ensembles, giving a total of three ensembles. One WA ensemble used AUCs from repeated random internal cross-validation (random WA hereafter), and the other used AUCs from spatially-blocked internal CV (block WA).

External validation

When investigating the performance of SDMs, it is often difficult to find a single best way to validate models. In practice, 'biomod2' ensembles are often evaluated using cross-validation on the same data used to calculate weights to build WA ensembles. As recognised by its authors (<https://rstudio-pubs-static.s3.amazonaws.com/38564_747d4bbf87704f0394734977bd4905c4.html>), this test may provide biased results across the various ensemble types and individual models, because the weighting is in favour of models that performed better on validation data. Therefore, performance of WA ensembles is possibly optimistic as judged by internal cross-validation. Since our study aims to investigate performance properties, to avoid this potential artificial advantage to WA ensembles, we validated models externally on held-out data from the external folds (Fig. 1), and averaged their external performance across all external folds (five for checkerboard and two for latitudinal). Our external validation also serves a second purpose of allowing us to assess if models that performed well on internal cross-validations also perform well on the external data. This is important to the scenarios in which modellers choose the individual model with best internal performance as their final model, instead of combining models in an ensemble (equivalent to the 'Best' ensemble approach in Marmion et al. 2009). We wish to understand how effective this strategy of choosing the 'best internal model' is and how it compares to model-combining ensembles and conventional single-model approaches. We calibrated the models for external evaluation on the entirety of internal folds (~80% of all data), then validated them using all data from the external fold. We will refer to these models as 'external models' hereafter, to reflect the use of external evaluations. For both external random WA ensembles and external block WA ensembles, we used internal cross-validation AUCs, averaged across four repeats, to calculate the weights. We chose AUC and log-likelihood as validation statistics in external validations. We used log-likelihood to complement AUC because while AUC tests model discrimination, log-likelihood tests a

different aspect of model performance – model calibration, which is how closely fitted values of models match observations (Pearce and Ferrier 2000).

Methods summary

In summary, we used eight algorithms for individual models and three ensemble methods in our analysis; in addition, we fine-tuned separate BRT models for comparison, thus resulting in 12 models per species per modelling run in total. We modelled all 14 species using a checkerboard blocking design and, for four selected species, we also built models using a latitudinally sliced blocking design. In both designs, we divided data into five folds, and for each modelling run, we designated four as internal folds and the remaining one as the external fold. Under checkerboard blocking, we cycled through all folds to act as the external fold for a total of five modelling runs per species. Under latitudinal blocking, we only used the northernmost and southernmost folds as the external fold, thus completing only two modelling runs per species. In all modelling runs, we built and cross-validated internal models using data from only internal folds, and models were cross-validated using a repeated random strategy and a spatially blocked strategy. We then built external models using all data from internal folds and validated them on the held-out external fold (checkerboard or latitudinal). The external validations allowed fair comparison between ensemble and individual models, and the internal cross-validations helped to understand whether internally best-performing models are still superior on external validations. Note that we necessarily calibrated and tested the internal and external models on different subsets of data (Fig. 1), so while they both can offer narratives about ranking of model performance, the internal and external validation results are not directly comparable. We used AUC to measure model performance in internal validations, and used AUC and log-likelihood in external validations.

In addition to the methods described above, we also repeated all analyses with 90% less calibration data to simulate a data-poor situation. Because they yielded similar patterns to those in the main results presented below, we report data-thinned analyses and log-likelihood results in Appendix 2.

Results

Checkerboard blocking

Most models performed well when assessed with AUC based on checkerboard blocking (Fig. 2). Across all species, blocked internal CVs tended to yield similar AUC values to those produced by external validations, with a slightly lower performance estimate internally, understandable given that the models in internal tests were fitted to less data. In contrast, repeated random CVs consistently estimated higher AUC values. This is consistent with the concept that spatial blocks provide test data more independent from the training data. We note that even relatively smooth models like GLMs show this pattern, whereas complex models, like RF, show even greater overoptimism under random CVs (Fig. 2).

Focusing on the comparison between performance of ensembles and individual models on external data, we found that, across all species, all three ensembles (mean, random WA, block WA) outperformed all untuned individual models (paired two-tailed Wilcoxon signed-rank test, $p < 0.05$ for all models except RF, for which $p = 0.05$, Supplementary material Appendix 2). The improvements in AUC, whilst significant, tended to be very small (≤ 0.03 AUC unit difference between ensembles and all individual models except CTA and FDA, which performed significantly worse than every other model, $p < 0.01$). The three ensembles performed indistinguishably from one another ($p > 0.05$), so weighting did not significantly alter performance in our case. Among untuned individual models, RF performed best, with its mean AUC score only ~0.01 unit less than that of the ensembles. However, RF only outperformed other strong-performing models (i.e. GLM, BRT, GAM and MARS) by 0.01–0.02 AUC units. Tuned BRTs outperformed all untuned individual models ($p < 0.05$), including untuned BRTs, but again most differences were small. The differences, albeit small, imply that default BRT tuning choices in 'biomod2' are not quite optimal for our dataset. Furthermore, tuned BRTs appeared to match the performance of all ensemble models when judged by mean AUC across species ($p > 0.5$).

Choosing the model that performed best per species on internal cross-validation also yielded good performance on external validation, comparable to that of ensembles (see 'block best' and 'random best' rows in Supplementary material Appendix 2 Table A2). Blocked internal CVs were more successful at predicting best external models than random internal CVs (11 correct predictions out of 14 species versus 4; Table 3). This appears to be driven by RFs performing particularly strongly under random internal CVs, but not always under the other two validation designs. However, inability to identify the model with the best external performance would not be problematic in this dataset, because best internal models tended to perform well on external validation even if they are not the top models – the AUC difference between best internal and external models are always within 0.02 AUC units ('AUC loss' columns in Table 3).

Latitudinal blocking

Latitudinal extremities were more remote though mostly still within the range of environments in the training data (Supplementary material Appendix 2). AUC of models on latitudinally blocked internal cross-validations were lower than that observed when checkerboard blocking was used (mean of 0.78 with a range from 0.54 to 0.92, compared to mean of 0.87 with a range from 0.61 to 0.96). AUC values from latitudinal external validation were also lower than those from checkerboard external validation (mean = 0.76, min = 0.46, max = 0.92 compared to mean = 0.87, min = 0.51, max = 0.99; Fig. 3). In addition to lower overall AUC, ranking

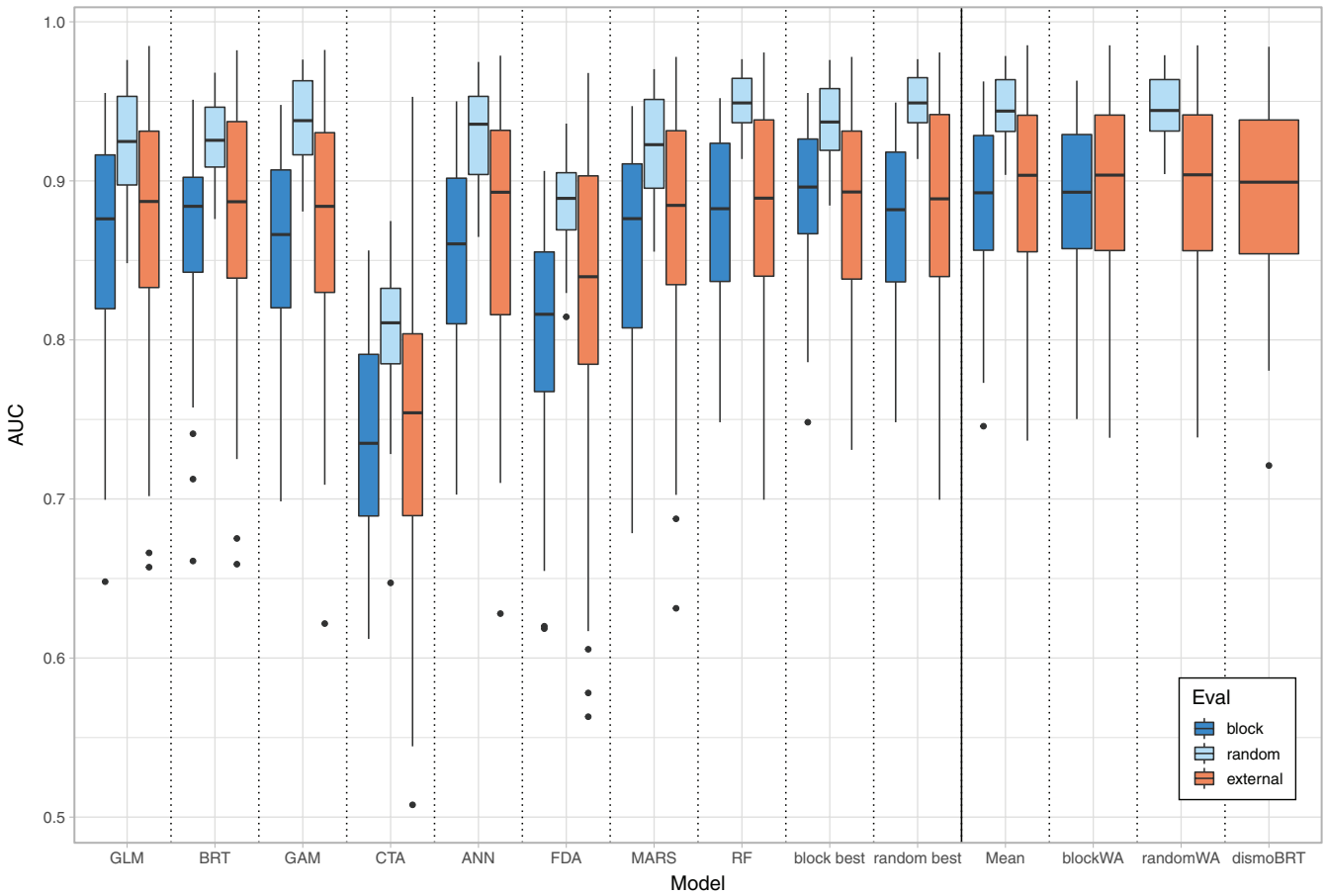

Figure 2. Model performance as measured by AUC across 14 species and five folds using a checkerboard-blocked validation design, for different modelling approaches (x-axis). AUC scores are colour-coded according to the data used for evaluation (Eval: block = spatially blocked internal cross-validation, random = repeated random internal cross-validation, external = external validation). 'Block best' and 'random best' correspond to the best-performing individual model for each species under the respective internal cross-validation method. Boxes span 1st and 3rd quartile values, with the horizontal line indicating the median.

Table 3. Best individual models as identified by either repeated random (Random) or spatially blocked (Block) internal cross-validations, compared to best individual models on external validations. 'Blocking' represents whether external validation results were based on checkerboard (C) or latitudinal blocking designs (L). 'AUC loss' indicates how much worse (in AUC units) the best cross-validation models performed on external validation compared to the best external models. Models that performed best on both cross-validation and external validation are in bold.

Blocking	Species	Random	AUC loss	Block	AUC loss	External best	Best external AUC
C	corymacu	RF	0.01	**GAM**	0.00	**GAM**	0.82
C	eucaaggl	**RF**	0.00	**RF**	0.00	**RF**	0.83
C	eucacype	**RF**	0.00	**RF**	0.00	**RF**	0.92
C	eucadive	RF	0.00	MARS	0.01	ANN	0.95
C	eucafast	GAM	0.01	**RF**	0.00	**RF**	0.93
C	eucamolu	RF	0.02	**GLM**	0.00	**GLM**	0.88
C	eucaobli	**RF**	0.00	**RF**	0.00	**RF**	0.92
C	eucapauc	RF	0.00	**GLM**	0.00	**GLM**	0.93
C	eucapilu	RF	0.00	RF	0.00	GAM	0.83
C	eucapipe	RF	0.01	**GAM**	0.00	**GAM**	0.87
C	eucapunc	RF	0.01	**MARS**	0.00	**MARS**	0.90
C	eucarobu	RF	0.01	BRT	0.00	MARS	0.90
C	eucaross	GAM	0.00	**GLM**	0.00	**GLM**	0.94
C	eucasieb	**RF**	0.00	**RF**	0.00	**RF**	0.88
L	corymacu	RF	0.11	MARS	0.00	BRT	0.74
L	eucaobli	RF	0.04	RF	0.04	MARS	0.84
L	eucapauc	RF	0.00	RF	0.00	MARS	0.87
L	eucapilu	RF	0.02	BRT	0.01	FDA	0.80

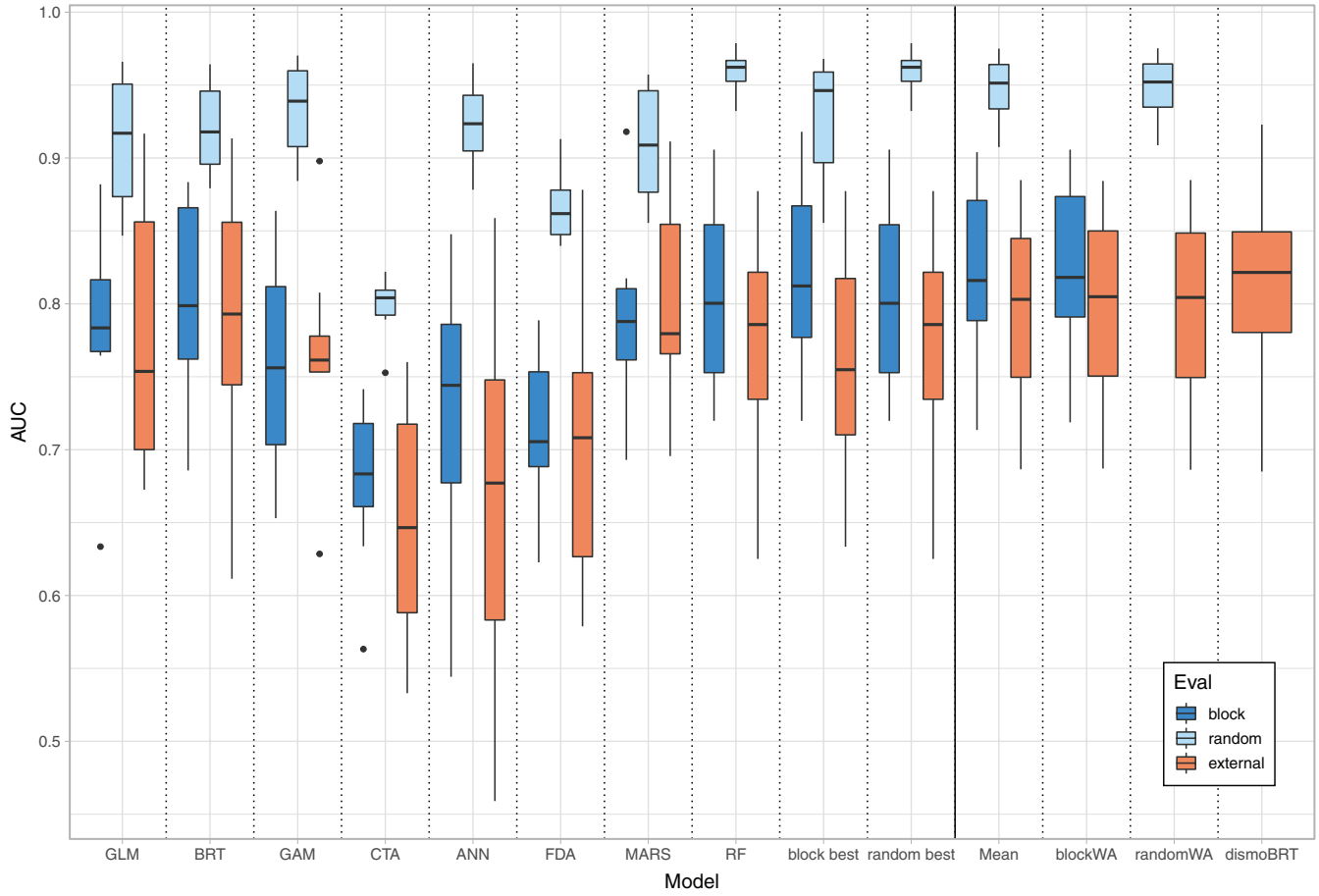

Figure 3. Model performance as measured by AUC across four species and two folds using a latitudinally-blocked validation design, for different modelling approaches (x-axis). AUC scores are colour-coded according to the data used for evaluation (Eval: block = spatially blocked internal cross-validation, random = repeated random internal cross-validation, external = external validation). 'Block best' and 'random best' correspond to the best-performing individual models for each species under the respective internal cross-validation method. Boxes span 1st and 3rd quartile values, with the horizontal line indicating the median.

of model performance under latitudinal blocking was also different from the checkerboard blocking scenario. The three ensembles performed equally well, and all outperformed most untuned individual models, but were only better than GLMs and untuned BRTs by 0.01–0.02 AUC units (see Supplementary material Appendix 2 Table A3 for details). However, MARS outperformed all ensembles for three out of four species, and its mean AUC across species was higher than all ensembles by 0.01 AUC units (this pattern is consistent when data are thinned, Supplementary material Appendix 2). Moreover, tuned BRTs obtained higher AUC than all other models (ensembles included), and appeared to outperform other models most often (Supplementary material Appendix 2 Table A3, note that we did not perform significance tests due to small sample size, n = 4).

Models that performed best on internal cross-validations were consistently not the best performing on latitudinal external validation (Table 3). Most internal best models were not noticeably worse than external best models, with the exception of RF for *Corymbia maculata*, which performed best on internal repeated random CV but on external validation did worse than the best model (BRT) by over 0.1 AUC units. In addition, repeated random CV consistently yielded higher AUC values than blocked CV and external validation for every model (Fig. 3).

For both checkerboard and latitudinal blocking designs, detailed model-to-model AUC comparisons are available in Supplementary material Appendix 2 Table A2, A3, which also contains additional results from log-likelihood and thinned-data analyses.

Discussion

By validating models on spatially blocked data, we found that: 1) ensemble models performed well compared to untuned individual models, but their performance gain was small; 2) ensembles can be outperformed by untuned individual models when predicting to distant areas; 3) the approach where one chooses the individual model with best internal validation performance also yielded good performance, only marginally worse than that of ensembles; 4) ensembles of untuned

individual models could not consistently outperform a tuned individual model. Hence in this dataset we observe strong performance across a range of approaches, without clear superiority of ensembles.

In our AUC tests using latitudinal design, ensembles were outperformed by tuned BRTs and untuned MARS on both full and thinned training data, and by untuned GLMs under thinned training data (Supplementary material Appendix 2). This suggests that ensembles may not perform as well as individual models when predicting to more distant areas. This is consistent with previous investigations into ensemble performance showing that they perform well at interpolating tasks (Marmion et al. 2009), but are not always the best at transfer tasks (Crimmins et al. 2013, Zhu and Peterson 2017). This is worth further investigation, because ensembles are popular with such transferring tasks (Hao et al. 2019a). In our latitudinal tests, the observed decrease in ensemble performance relative to others may be explained by a combination of: 1) spatial dependencies between train and test sets are decreased in latitudinal blocking, and 2) complex models can overfit spatial patterning, erroneously attributing geographic patterns of SAC to environmental covariates (Merow et al. 2014, Roberts et al. 2017). It is likely that the ensembles are the most complex models in our study, because there are many component models, all fitted with default settings. On the other hand, models outperforming ensembles on latitudinal validations were all tuned to have relatively simple response curves (see Table 2 for default tunings of GLMs and MARS) or their tuning was optimised (tuned BRT), thus they were less likely overfit. Since our explorations of extrapolation revealed relatively few environments outside those in the training data, it is unlikely that this decreased performance in latitudinal test folds is driven by extrapolation problems (sensu Sequeira et al. 2018).

Interestingly, the performance of ensemble models did not appear to be affected by inclusion of poor models (e.g. CTAs). Initial investigations revealed that the CTA models in this study tended to produce nearly flat response curves. This means that, when combined with other models in an ensemble, predictions from these poor CTA models did not influence the discrimination capacity of the final prediction greatly (i.e. they did not affect the ranking of the final model predictions). One would imagine that with different tuning the range of predictions from a CTA would increase, and may either improve the performance of the CTA or, if not, have more (negative) impact on the ensemble performance. Since the CTA models were the most consistently poorly performing ones in our study, we have limited scope to test whether ensembles would be similarly unaffected by other model algorithms that perform poorly in other datasets. This requires further research, to identify whether specific types of prediction failure more severely impact ensemble performance (Dormann et al. 2018).

The tuning of modelling methods is known to affect predictive performance (Elith et al. 2008, Hastie et al. 2009), but the approach to model tuning is rarely reported in studies employing 'biomod2' ensembles (Hao et al. 2019a). It appears that 'biomod2' default tuning choices are used in many ensemble modelling studies, therefore we also used 'biomod2' defaults to emulate a typical 'biomod2' modelling scenario. However, with only simple tuning of our BRTs, using slightly different code and settings, our 'tuned' BRTs consistently achieved slightly yet consistently better and less variable (Fig. 2, 3) discrimination performance than those fitted with 'biomod2' defaults. Other methods are also likely to respond to individual tuning. For example, biomod2 default tuning for CTA and RF allows for complex trees (Table 2) and does not use spatially-blocked validations to control for overfitting – this could encourage models to be overfitted to spatial structures in data, and thus underperform on spatially-blocked validations. We believe it is worthwhile investigating how individual models respond to tuning, and how they perform versus ensembles of well-tuned individual models across a range of datasets. Nevertheless, we acknowledge that tuning multiple modelling methods with different underlying techniques is a non-trivial undertaking. Practically, modellers often only have the skills to fine-tune a single modelling technique. In such cases it may be important to ask whether the tuned single model is better or worse than an ensemble of untuned models. In our case study, either approach is suitable for maximising prediction accuracy. However, there may be an advantage to using tuned single models, as they are more interpretable than ensembles (response curves and variable importance are readily available) and time can be spent focussing on optimal tuning for the task at hand.

In both of our spatial blocking designs, repeated random cross-validations yielded higher performance estimates than spatially blocked cross-validations and external validations. Depending on the application in which predictions are to be used, blocked cross-validation may be a more realistic estimate of model performance (Roberts et al. 2017). The consistency between the internal and external block cross-validation estimates is in some ways not surprising: they are both testing similar things – i.e. capacity to predict to spatially distinct sites. The consistency is also encouraging: estimates on internal data were consistent with those seen on external data. Because 'biomod2' and ensemble modelling are often used for predicting to new environments (Hao et al. 2019a) and blocked CV may better test prediction to new environments and will reveal overfitting to spatial dependencies in the data, our findings suggest that using blocked CV instead of the 'biomod2' default repeated random CV may be beneficial to many 'biomod2' users. We remark that blocked CV can be implemented in 'biomod2' through its in-built 'BIOMOD_cv' function, which can create spatially sliced blocks (Wenger and Olden 2012). Alternatively, one can use the external R package 'blockCV' (Valavi et al. 2019), which provides a wider range of options for spatial and environmental blocking for SDMs, and can create blocked data specifically formatted for use with 'biomod2'.

In summary, we tested the predictive performance of ensemble models compared to individual models using a presence–absence dataset, and found ensemble models to perform slightly better than untuned individual models in

most situations, but not consistently better than tuned individual models on external validation. With future research on ensemble performance testing the breadth of applications and data types commonly used in ensemble modelling, knowledge of ensemble performance will be improved and used to inform best practice in ensemble modelling.

Data availability statement

Data are available from the Dryad Digital Repository: <http://dx.doi.org/10.5061/dryad.tqjq2bvv2> (Hao et al. 2019b).

Acknowledgements – *Funding* – This work was supported by a Discovery Project grant to José J. Lahoz-Monfort and Jane Elith (DP160101003), and a Discovery Early Career Research Award to Gurutzeta Guillera-Arroita (DE160100904), both from the Australian Research Council.

References

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

Crimmins, S. M. et al. 2013. Evaluating ensemble forecasts of plant species distributions under climate change. – Ecol. Model. 266: 126–130.

Diniz-Filho, J. A. F. et al. 2009. Partitioning and mapping uncertainties in ensembles of forecasts of species turnover under climate change. – Ecography 32: 897–906.

Dormann, C. F. et al. 2018. Model averaging in ecology: a review of Bayesian, information–theoretic and tactical approaches for predictive inference. – Ecol. Monogr. 88: 485–504.

Elith, J. and Graham, C. H. 2009. Do they? How do they? Why do they differ? On finding reasons for differing performances of species distribution models. – Ecography 32: 66–77.

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

Elith, J. et al. 2008. A working guide to boosted regression trees. – J. Anim. Ecol. 77: 802–813.

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

Fithian, W. et al. 2015. Bias correction in species distribution models: pooling survey and collection data for multiple species. – Methods Ecol. Evol. 6: 424–438.

Guillera-Arroita, G. et al. 2015. Is my species distribution model fit for purpose? Matching data and models to applications. – Global Ecol. Biogeogr. 24: 276–292.

Guisan, A. and Thuiller, W. 2005. Predicting species distribution: offering more than simple habitat models. – Ecol. Lett. 8: 993–1009.

Guisan, A. et al. 2017. Habitat suitability and distribution models: with applications in R. – Cambridge Univ. Press.

Hao, T. et al. 2019a. A review of evidence about use and performance of species distribution modelling ensembles like BIOMOD. – Divers. Distrib. 25: 839–852.

Hao, T. et al. 2019b. Data from: Testing whether ensemble modelling is advantageous for maximising predictive performance of species distribution models. – Dryad Digital Repository, <http://dx.doi.org/10.5061/dryad.tqjq2bvv2>.

Hardy, S. M. et al. 2011. Predicting the distribution and ecological niche of unexploited snow crab (*Chionoecetes opilio*) populations in Alaskan waters: a first open-access ensemble model. – Integr. Comp. Biol. 51: 608–622.

Hartley, S. et al. 2006. Quantifying uncertainty in the potential distribution of an invasive species: climate and the Argentine ant. – Ecol. Lett. 9: 1068–1079.

Hastie, T. et al. 2009. The elements of statistical learning: data mining, inference and prediction. – Springer.

Hijmans, R. J. et al. 2017. dismo: species distribution modeling. – <https://cran.r-project.org/web/packages/dismo>.

Marmion, M. et al. 2009. Evaluation of consensus methods in predictive species distribution modelling. – Divers. Distrib. 15: 59–69.

Merow, C. et al. 2013. A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. – Ecography 36: 1058–1069.

Merow, C. et al. 2014. What do we gain from simplicity versus complexity in species distribution models? – Ecography 37: 1267–1281.

Pearce, J. and Ferrier, S. 2000. Evaluating the predictive performance of habitat models developed using logistic regression. – Ecol. Model. 133: 225–245.

Pearson, R. G. et al. 2006. Model-based uncertainty in species range prediction. – J. Biogeogr. 33: 1704–1711.

Roberts, D. R. et al. 2017. Cross-validation strategies for data with temporal, spatial, hierarchical or phylogenetic structure. – Ecography 40: 913–929.

Segurado, P. and Araújo, M. B. 2004. An evaluation of methods for modelling species distributions. – J. Biogeogr. 31: 1555–1568.

Sequeira, A. M. M. et al. 2018. Transferring biodiversity models for conservation: opportunities and challenges. – Methods Ecol. Evol. 9: 1250–1264.

Thuiller, W. et al. 2009. BIOMOD – a platform for ensemble forecasting of species distributions. – Ecography 32: 369–373.

Valavi, R. et al. 2019. blockCV: an r package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. – Methods Ecol. Evol. 10: 225–232.

Wenger, S. J. and Olden, J. D. 2012. Assessing transferability of ecological models: an underappreciated aspect of statistical validation. – Methods Ecol. Evol. 3: 260–267.

Zhu, G.-P. and Peterson, A. T. 2017. Do consensus models outperform individual models? Transferability evaluations of diverse modeling approaches for an invasive moth. – Biol. Invas. 19: 2519–2532.

Supplementary material (available online as Appendix ecog-04890 at <www.ecography.org/appendix/ecog-04890>). Appendix 1–2.