Received: 26 June 2018 | Revised: 25 November 2018 | Accepted: 3 December 2018

DOI: 10.1111/ddi.12892

BIODIVERSITY REVIEW

WILEY Diversity and Distributions

A review of evidence about use and performance of species distribution modelling ensembles like BIOMOD

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

School of BioSciences, The University of Melbourne, Parkville, Victoria, Australia

Correspondence
Tianxiao Hao, School of BioSciences, The University of Melbourne, Parkville, Victoria, Australia.
Email: tianxiaoh@student.unimelb.edu.au

Funding information
Australian Research Council, Grant/Award Number: DE160100904 and DP160101003

Editor: Josep Serra-Diaz

Abstract

Aim: The idea of combining predictions from different models into an ensemble has gained considerable popularity in species distribution modelling, partly due to free and comprehensive software such as the R package BIOMOD. However, despite proliferation of ensemble models, we lack oversight of how and where they are used for modelling distributions, and how well they perform. Here, we present such an overview.

Location: Global.

Methods: Since BIOMOD is freely available and widely used by ensemble species distribution modellers, we focused on articles that apply BIOMOD, filtering the initial 852 papers identified in our structured literature search to a relevant final subset of 224 eligible peer-reviewed journal articles.

Results: BIOMOD-based ensembles are used across many taxa and locations, with terrestrial plants being the most represented group of species (n = 72) and Europe being the most represented continent (n = 106). These studies often focus on forecasting distributions in the future (n = 109), and commonly use presence-only species data (n = 139) and climatic environmental predictors (n = 219). An average of six models are used in ensembles, and approximately half of ensembles weight contributions of models by their cross-validation performance. However, discussion about choices made in the modelling process and unambiguous information on the performance of ensemble models versus individual models are limited. The use of independent data to validate model performance is particularly uncommon.

Main conclusions: We document the breadth of ensemble applications, but could not draw strong quantitative conclusions about the predictive performance of ensemble models, due to lack of unambiguous information reported. Understanding how and where ensembles are best used when modelling species distributions is important for enabling best choices for different applications. To enable this objective to be achieved, we provide recommendations for thorough reporting practices in a BIOMOD-based ensemble workflow.

KEYWORDS
BIOMOD, consensus forecast, ecological niche models, ensemble, habitat suitability models, species distribution model

1 | INTRODUCTION

Species distribution models (SDMs) are an increasingly important tool in ecology, biogeography and conservation sciences. Correlative SDMs are models that use species–environment relationships to explain and predict distributions of species. SDMs have broad 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 (Guillera-Arroita et al., 2015; Guisan & Thuiller, 2005; Hijmans & Graham, 2006). There are now many methods used for distribution modelling, varying in how they approach model selection, how they define fitted functions and interactions, whether they can handle imperfect detection and sampling biases and so on (Franklin, 2010; Guisan, Thuiller, & Zimmermann, 2017; Peterson et al., 2011). For example, statistical regression models such as generalized linear models (GLMs, McCullagh, 1984) compute species occurrence as parametric functions of environmental variables whereas random forests (RF, Breiman, 2001) are based on machine-learning decision tree approaches. Predictive outcomes across SDM methods are known to be variable, and the choice of modelling method can significantly affect model predictive performance. However, no one method is consistently superior in performance across species, regions and applications (Elith et al., 2006; Pearson et al., 2006; Segurado & Araujo, 2004). This makes it difficult to choose which method (*individual model* hereafter) to use, prompting the idea of combining predictive outputs from different models in a so-called *ensemble* (Araújo & New, 2007).

The underlying philosophy of ensemble modelling is that each individual model carries both some true "signal" about the relationships the model is aiming to capture, and some "noise" created by errors and uncertainties in the data and the model structure. Ensembles combine models with the intention of obtaining better separation of the signal from noise (Araújo & New, 2007; Dormann et al., 2018). Ensemble modelling has a history in other fields dealing with complex and dynamic systems, such as economics (Gregory, Smith, & Yetman, 2001) and meteorology (Sanders, 1963), and is thought to be traceable to the earliest days of statistical sciences (de Laplace, 1818). The concept of ensemble is widely used in machine learning, often with complex classifiers built by combining many simple modelling units. These types of ensembles have been shown in many cases to have superior predictive performance to individual models (Friedman & Popescu, 2008; Seni & Elder, 2010). They use a variety of methods (including bagging and boosting, Dietterich, 2000) to form the ensemble and are integral to a number of methods used in species distribution modelling (e.g., Cutler et al., 2007; Elith, Leathwick, & Hastie, 2008; Hardy, Lindgren, Konakanchi, & Huettmann, 2011).

A popular variant of SDM ensembles emerged over a decade ago from within the species distribution modelling research community, and here, we are interested to study that variant. Early ideas emphasised use of SDM ensembles for forecasting species distribution changes under future climates (the term "consensus forecast" is often used in this context, Araújo & New, 2007; Thuiller, Lafourcade, Engler, & Araújo, 2009) and focused on ensembles across modelling methods. This contrasts with the idea of ensembles or averages across multiple instances of the same method, such as in model averaging of regression models with different combination of predictors (Burnham & Anderson, 2003), or ensembles of trees such as boosted trees or RF (Hastie, Tibshirani, & Friedman, 2009). From here on we focus on the ensembles-across-methods ideas that have emerged in the SDM literature, in the interest of targeting a widely used approach in species distribution modelling and understanding its use and performance. We note that, within this literature, some users focusing on climate change view the consensus forecast as consensus across model predictions *and* future climate change scenarios (e.g., combinations of emission scenarios and global circulation models). However, to enable review across the range of applications, we focus here on ensembles-across-methods (i.e., algorithms), regardless of whether they are also across future climate scenarios.

Various strategies exist to combine predictions from individual models into an ensemble, the most intuitive of which is simply taking the mathematical mean or median across predictions, irrespective of whether such predictions are binary or continuous. More complex approaches involve "weighting," scaling predictions of different models by weights based on some measure of predictive performance (Araújo & New, 2007). These weights are often derived by validating predictions from individual models on some test data. Weighting is thought to improve how well the ensemble predicts (Araújo & New, 2007; Marmion, Parviainen, Luoto, Heikkinen, & Thuiller, 2009), although weighted ensembles also require more effort to produce as the individual models need to be validated before they can be combined (although users of unweighted ensembles may also choose to do so).

We choose BIOMOD (Thuiller, 2003; Thuiller et al., 2009) as our focus because BIOMOD is the most well-known and well-established ensemble software created within the SDM community, it is freely available, and we expect its use to continue. Other approaches are also available but often tailor-made by specific groups of researchers, and not necessarily widely used by the broader community. BioEnsembles is a notable example (Diniz-Filho et al., 2009), but there are also several ensemble modelling examples without specialised tool sets (e.g., Crimmins, Dobrowski, & Mynsberge, 2013; Hardy et al., 2011). Mention of BIOMOD-based ensembles is increasingly popular in the SDM literature. A search of citations of the two BIOMOD introductory papers (Thuiller, 2003; Thuiller et al., 2009) on ISI Web of Science database (searched on 07/06/2017) yields 829 unique results, with 254 of them published in the 3 years prior to June 2017. Despite this proliferation, there is not yet an overview of published studies employing these types of ensemble methods. Such an overview is an important part of understanding modelling approaches and how they perform. Here, we aim to fill this knowledge gap by providing a review of SDM studies that used BIOMOD-like ensemble modelling. We summarize these in terms of their studied

species, geography, data and choices in modelling methods, as a prelude to analysing evidence on model performance of both the ensembles and the individual models, as reported in these studies. As we started this study, our ultimate aim was to enhance understanding of how these types of ensemble models perform in their applications, how sensitive they are to choices made in the modelling process, and whether they are particularly suited to some situations over others.

2 | BACKGROUND: USE OF ENSEMBLES LIKE BIOMOD

Many users cite the superior predictive performance of ensembles over individual models (as seen in Crossman & Bass, 2008; Marmion et al., 2009) as justification for choosing them. However, individual models have also been shown to outperform ensemble models (Crimmins et al., 2013). It is reasonable to believe another important contributor to the popularity of ensembles is the existence of comprehensive free ensemble modelling tools, such as BIOMOD.

BIOMOD provides a suite of methods and tools relevant to the problem of modelling distributions, such as the ability to quickly build individual models and to combine them in different ways. BIOMOD was first developed in the S-Plus language environment in 2003 (Thuiller, 2003) and was later ported to the R statistical language environment (R Core Team, 2016) as a package under the name "biomod2." The latest version of "biomod2" (ver. 3.3-7; Thuiller, Georges, Engler, & Breiner, 2016) is capable of computing SDMs with up to 10 different modelling methods (for a full list, including abbreviations of model names used from here on, see Table 1). "biomod2" can combine these individual models using various approaches to form ensembles, including: *Weighted Mean*, *Mean*, *Median*, and *Committee Averaging*. The first three approaches correspond to the ensemble strategies previously introduced, while the last one, Committee Averaging, transforms probabilistic predictions from the single models into binary predictions using a threshold, and then averages them. Users of "biomod2" can choose which individual models they want to include in the ensemble. For example, they could use all individual models available, or use only models that performed better than a specified threshold on a selected metric.

Another major module in "biomod2" is model evaluation. The evaluation statistics, useful in their own right, are also used as a basis for weighting models in Weighted Mean ensembles, and for determining which individual models to exclude from the ensemble due to poor performance. "biomod2" allows evaluation of models based on several metrics derived from a confusion matrix, including sensitivity (proportion of true positives correctly identified), specificity (proportion of true negatives correctly identified), area under curve (AUC) of receiver operating characteristics (Hanley & McNeil, 1982), and true skill statistics (TSS; Allouche, Tsoar, & Kadmon, 2006). These latter two evaluation metrics are indicators of discrimination capacity, which quantifies how well the

TABLE 1 Classes of SDM modelling methods available in BIOMOD

Name	Abbreviation[a]	Required data[b]	References
Generalized linear models	GLM	Presence/absence	McCullagh (1984)
Generalised additive models	GAM	Presence/absence	Hastie and Tibshirani (2004)
Multivariate adaptive regression splines	MARS	Presence/absence	Friedman (1991)
Classification tree analysis	CTA	Presence/absence	Breiman (1984)
Mixture discriminant analysis	MDA	Presence/absence	Hastie, Tibshirani, and Buja (1994)
Artificial neural networks	ANN	Presence/absence	Ripley (2007)
Random forests	RF	Presence/absence	Breiman (2001)
Boosted regression tree	BRT	Presence/absence	Elith et al. (2008)
Maximum Entropy	MaxEnt	Presence-background	Phillips, Anderson, and Schapire (2006)
Bioclimatic envelope	BIOCLIM	Presence-only	Busby (1991)

[a]Note that the "biomod2" package uses different names for some of the classes (cf. Thuiller et al., 2009), that is generalized boosted model (GBM) instead of boosted regression tree (BRT), and surface-range envelope (SRE) instead of BIOCLIM. [b]Presence-only data can be supplied to some presence/absence methods by selecting background points, however one should interpret model outputs in these situations as relative but not absolute probability of species occurrence, consistent with other presence–background methods.

model can distinguish presences from absences (or presences from background samples, when absences are unavailable). As held-out subsets of the original dataset are often used for model evaluation (known as cross-validation, CV), strategies for partitioning data into subsets are provided in "biomod2." The package also allows users to define their own subsets or use a completely independent dataset for evaluation.

3 | METHODS FOR LITERATURE REVIEW

To scan for BIOMOD-relevant studies, we considered those articles that have cited any of the two introductory BIOMOD papers (Thuiller, 2003; Thuiller et al., 2009). We searched the literature using both ISI Web of Science and Google Scholar, with our own search protocol modified from the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) scheme for systematic reviews and meta-analyses (Moher, Liberati, Tetzlaff, & Altman, 2009). We directly searched for all citing articles (up to October 2016 when this review was initiated) of Thuiller (2003) and Thuiller et al. (2009) in the Web of Science database, yielding 443 and 475 results respectively (787 in total after removing duplicates). We complemented this search result with Google Scholar database, where we searched for the keyword "ensemble" within the citing articles of Thuiller (2003) and Thuiller et al. (2009), yielding 260 and 577 results each (648 in total after removing duplicates). These were a pruned subset of all returned in Google Scholar: We manually excluded grey literature and non-eligible results (non-English publications and inaccessible articles). Alternative keywords "model averaging" and "consensus" were also tested in Google Scholar but the SDM-relevant results tended to be a subset of the "ensemble" search results and thus these terms were not pursued further.

After combining the pools of results and removing duplicates, we further removed any publication that was not a peer-reviewed journal article (theses, book chapters etc.). We then screened the abstracts of remaining articles, removing any that did not include a specific study using correlative SDMs. These removed papers were generally either not SDM-oriented or were SDM discussion/review papers that did not include actual analysis. After screening, 694 papers were retained and then read in full (see Supporting Information Appendix S1 for the complete list of assessed papers). From these, we only retained those papers that unambiguously used an SDM ensemble of the type we defined earlier. On this basis 473 were rejected; for a visualisation of the selection workflow and further details of reasons for rejection, see Figure 1.

Our literature search resulted in a final sample of 224 eligible papers which were reviewed in detail (including reviewing Appendices and Supplementary Materials where applicable, Appendix S2). From the 224 papers retained for synthesis, we documented the following items of information:

3.1 | Target species

We categorised modelled species into eight broad categories: plants, birds, herpetofauna (amphibians and reptiles), mammals, invertebrates, fish, mixed taxa and viruses or bacteria; we also made the distinction between marine and terrestrial/freshwater species.

3.2 | Data details

The quality and quantity of data is especially important for performance of models (Araújo & Guisan, 2006) and to their suitability for particular end-uses (Guillera-Arroita et al., 2015). Therefore, we documented: (a) the number of species records; (b) whether species absence records were used in modelling; and (c) spatial extent and resolution of data.

3.3 | Predictor variables

We documented the number of predictors considered by studies, and broadly categorised predictors into four classes based on the nature of predictors: (a) climatic: climate and vegetation; (b) biological: presence of another species (excluding those relating to vegetation); (c) geological: topography (including bathymetry and hydrology for non-terrestrial studies), landcover types, soil and geology; and (d) anthropogenic: presence of human activities (including human-modified landcover types).

3.4 | Model transfer

Some applications of SDMs, such predicting species responses to climate change or potential biological invasions, require prediction to new times and places (referred to as model transfer hereafter; Randin et al., 2006; Sequeira, Bouchet, Yates, Mengersen, & Caley, 2018). Model transfer often involves extrapolation of species–environment relationships outside the environments sampled by the species data. The performance of SDMs for such applications is known to be affected by several factors, and predictions often deemed less reliable (Dormann, Gruber, Winter, & Herrmann, 2010; Elith & Leathwick, 2009; Fitzpatrick & Hargrove, 2009; Guisan & Thuiller, 2005; Thuiller, 2004). As ensemble modelling is proposed as a useful approach for dealing with the uncertainties of extrapolation (Araújo & New, 2007; Marmion et al., 2009; Thuiller et al., 2009), we aimed to assess whether the majority of ensemble modelling papers indeed involved model transfer, and to summarize why transfer was required.

3.5 | Modelling methodology

We documented the various choices made in the process of modelling: (a) Number and type of individual models built; (b) Tuning of individual models, meaning the approach and settings used for fitting individual models; options vary with algorithms, and include aspects such as choice of model selection technique for a GLM

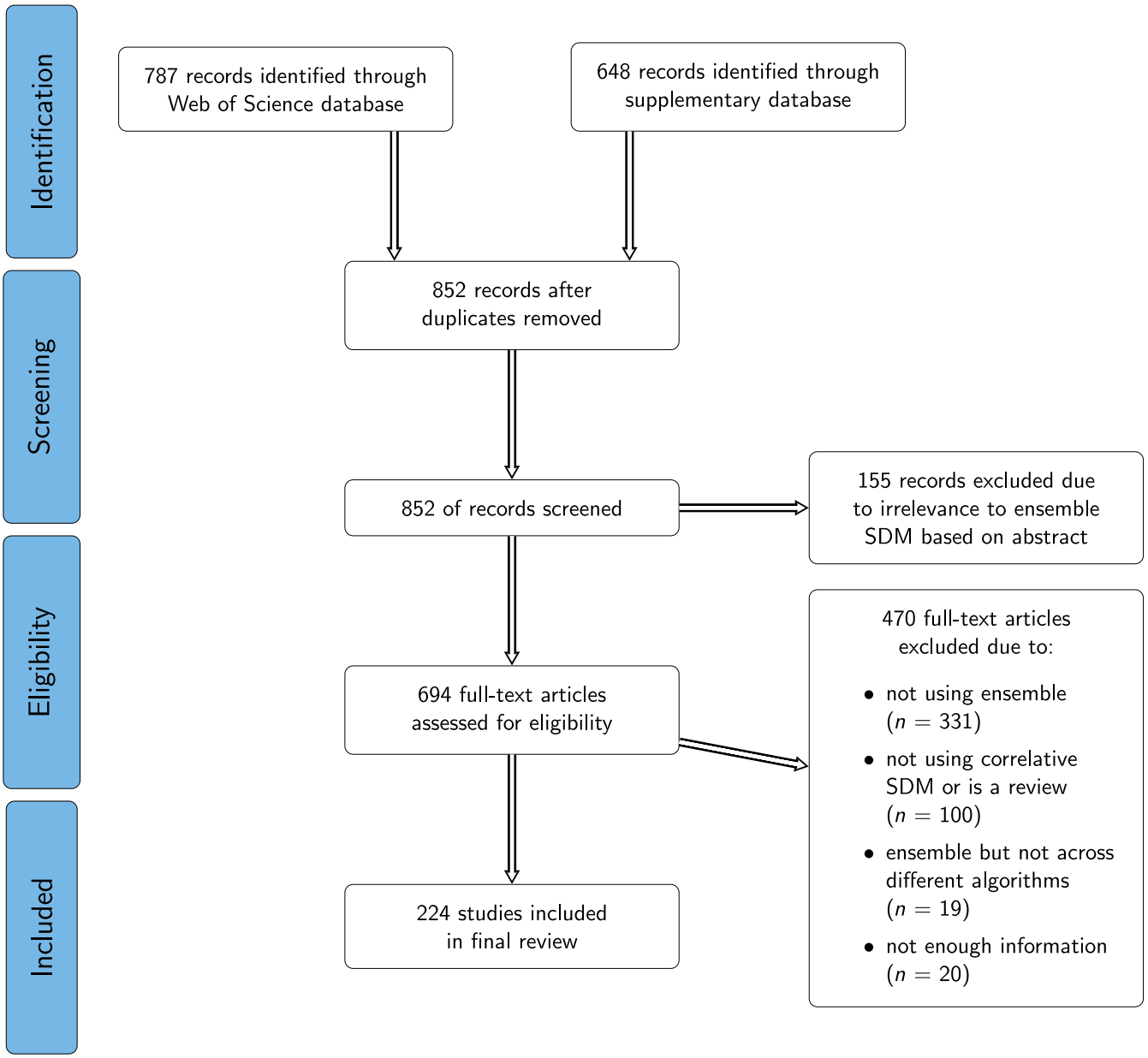

FIGURE 1 Workflow diagram illustrating our literature review protocol, modified from the PRISMA scheme (Moher et al., 2009)

(e.g., forwards stepwise selection or all subsets selection, Hastie et al., 2009) and regularization methods for machine-learning approaches (e.g., *learning rate* for BRT, *beta multiplier* in MaxEnt, Elith et al. 2011); (c) Ensemble methods (e.g., Weighted Average, Mean); (d) Model evaluation procedures (e.g., cross-validation) and metrics (e.g., AUC, TSS).

3.6 | Performance results reported

We documented results on model performance where provided. In particular, we noted if ensemble models were compared to individual models, and whether models were cross-validated or validated on independent data (here we define independent data as data that are independently collected from data used in model fitting; e.g., data collected in a different time period or location, or data collected by a different group of recorders).

4 | RESULTS AND DISCUSSION

We observe that BIOMOD-like ensembles have been used in a diverse range of applications; here, we summarize the main aspects of these ensemble SDM studies.

4.1 | Species and geography

A wide variety of species are modelled. The most represented taxonomical group is terrestrial plants, followed by invertebrates and

birds (Figure 2a), consistent with the broader applications of SDMs (Elith & Leathwick, 2009). Seventy reviewed studies model distribution of a single species and the remaining 154 model more than one species (no. of species in multispecies models ranges from 2 to 8,472 with median = 29 and mean = 434). Among reviewed studies, 26 model species across taxonomical groups, or model distribution of response variables other than biological species (e.g., landform processes in Aalto & Luoto, 2014). The geographical focus of reviewed studies is heavily biased towards Europe (106 studies, Figure 2b), which in some studies also include surrounding regions (the Mediterranean areas of Asia and North Africa). Terrestrial or freshwater species are modelled more often than marine species (208 and 16 studies, respectively). Data used in these studies vary greatly in raster resolution, with grid cell size ranging from 5 by 5 m in small scale regional studies, for example Engler et al., (2013) to ~110 by 110 km (or 1-degree longitude/latitude) commonly used in global models. The median value for cell size is ~5 by 5 km and the mean value is ~16 by 16 km. The above summary of cell size excludes 30 studies that use stream segments as spatial units to model aquatic species or that do not make spatial predictions.

4.2 | Data and predictors

Among reviewed studies, 85 use species absence data in their models (among which two also use data on species abundance, or density in grid cells). The remaining 139 studies build their models with only species presence information. The quantity of species records used in modelling varies, and in multispecies modelling studies it is often difficult to discern the exact number of records used per species, therefore here we summarize only for single species studies (n = 70): presence–absence (PA) records used range from 19 to 39,645 (median = 1,002 and mean = 4,347) and presence-only (PO) records range from 14 to 128,653 (median = 360 and mean = 3,851). Nearly all studies use climatic predictors (n = 219), unsurprisingly as

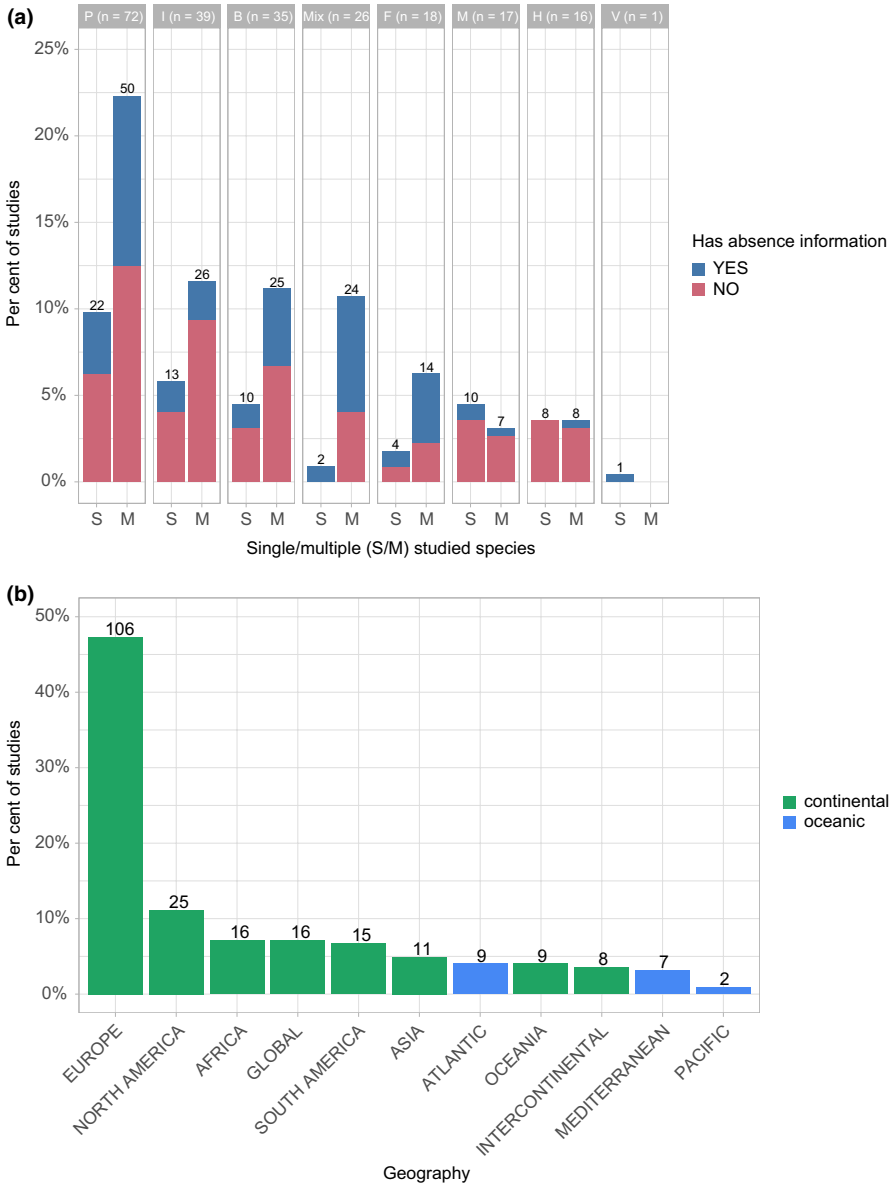

FIGURE 2 (a) Summary of studied species and data type among a sample of 224 ensemble species distribution model studies identified through reviewing literature citing BIOMOD. Floating numbers represent total count of single/multispecies studies within each taxonomical group. For abbreviation of taxonomical groups: B = Birds, F = Fish, M = Mammals, H = Herpetofauna (reptiles and amphibians), I = Invertebrates, P = Plants, V = Viruses, Mix = organisms from more than one of the aforementioned categories, or non-biological response variables. (b) Frequency of continental or oceanic bodies being targeted by sampled studies. Floating numbers represent count of papers

climate likely drives distribution patterns of most species at large scales. Geological and anthropogenic predictors are used less frequently (*n* = 113 and 42, respectively), and biological predictors are only used in three studies. The median and mean number of predictors fitted to models are 8 and 8.8, respectively (excluding studies that do not specify the predictor used or use varying numbers of predictors to fit multiple sets of models, *n* = 26).

4.3 | Model transfer

We observe that the scope of ensemble SDM applications has expanded beyond the initial dialogue around forecasting distribution changes into the future, although predicting species responses to future climate change is still the most common motivation for users of BIOMOD-like ensembles (109 studies). Sixteen model-transfer studies project distribution to past conditions (e.g., past climatic conditions simulated by global circulation models or past land-use conditions prior to human modification), and seven studies predict distribution in both past and future conditions. Other studies focus on transferring predictions across spatial horizons, with eight studies predicting invasion or expansion outside native habitats, and two predicting invasions under climate change scenarios (spatial and temporal transfer). Finally, three studies predict to new environments in which species data have been collected (different time period or invaded range) to specifically evaluate the transferability of their models (e.g., Crimmins et al., 2013). The remaining 79 studies do not transfer their models.

4.4 | Methodology

In contrast to the diverse applications of models, modelling methodologies employed by these studies are generally similar. This is largely a consequence of how we construct our review set, and the consistency in approach across ensemble applications. Among reviewed studies, 12 explicitly report not using BIOMOD but using their own codes or other toolboxes such as BioEnsembles. However, the methodologies employed by these 12 studies are closely related to those of BIOMOD and thus relevant to our review focus. The remaining studies all use BIOMOD, likely in the form of the package "biomod2" implemented in R (specific version number of BIOMOD reported in 57 studies).

On average six individual models are used in ensembles (median = 6 and mean = 6.2). Regarding specific modelling approaches (see details and acronyms, Table 1), GLMs are most frequently used, with BRTs, RFs and GAMs closely following behind (Figure 3a); BIOCLIM is least frequently used (excluding the non-BIOMOD studies). The MaxEnt algorithm, popular among SDM users using PO data (Morales, Fernández, & Baca-González, 2017), was not implemented in BIOMOD until 2012 (as inferred from documentation of the "biomod2" package, Thuiller et al., 2016). It is not possible to unequivocally analyse choice of methods among those available because users often do not always specify what methods were available or what version of BIOMOD they used. Using the 2012 inclusion date as a cut-off (Figure 3a) MaxEnt can be seen to be a popular choice among the set of methods used for PO data. Other methods (e.g., GLM, GAM, BRT, RF) are more often used, perhaps because users prefer them or perhaps because—despite the post-2012 date—their analyses use earlier versions of BIOMOD without MaxEnt. While all reviewed studies report the algorithms used for their individual models, we are largely unable to identify procedures used to tune these algorithms, or the tuned parameters used in final individual models. This hinders our ability to understand approaches taken by modellers (Golding et al., 2018; Naimi & Araújo, 2016). Only 47 out of 224 reviewed papers report tuning information, either by listing the specific tuning parameters used or by providing codes used for analyses. Furthermore, 40 studies report the use of default tuning options in BIOMOD. However, unless the default parameters are specified, our ability to faithfully reproduce these default models is limited, because BIOMOD defaults may have and may continue to change over time with ongoing development of the software, and default settings for older versions are not necessarily easy to determine even when version numbers are specified.

In our reviewed studies, individual models are often subject to an initial round of validation before they are combined, as a mean to provide weighting score for Weighted Mean ensembles and to inform if some models should be excluded from ensembles due to poor performance. These procedures often assess predictive accuracy using one or more measures of discrimination derived from a confusion matrix (e.g., 168 studies using AUC and 116 studies using TSS). It is noteworthy that when "biomod2" is used to validate models, by default sensitivity and specificity are always reported, although not all users choose to discuss these results. In rare cases (*n* = 7), some measures of model calibration (how well predicted values fit observed values) are also used to assess individual model performance (e.g., Root Mean Square Error in Folmer et al., 2016). None of these seven studies use BIOMOD, which does not provide functionality for estimating calibration performance. Cross-validation is used in most of these individual model validations (*n* = 204). In cross-validation, different strategies can be used to divide data into calibration and validation subsets (Roberts et al., 2017), but the vast majority in our sample (*n* = 198) use BIOMOD's default strategy of repeatedly and randomly resampling data into a calibration set and a validation set (e.g., dividing all data by 70/30% into each set, repeated 10 times). The remaining six use either jack-knife (holding a single data record out for validation each time; *n* = 2), or *k*-fold CV (dividing data into *k* equal sized folds and cycling through folds with one being used for validation each time; *n* = 4). Although CV is often used in the initial validation of individual models, final individual models are often trained with all data available before they are combined in an ensemble. Aside from these 204 cross-validating studies, three studies use independent data to validate their individual models, and 10 studies use both independent validation and CV. The remaining seven studies do not explicitly report validating their individual models. Furthermore, we note that individual model performance is used to guide selection of models

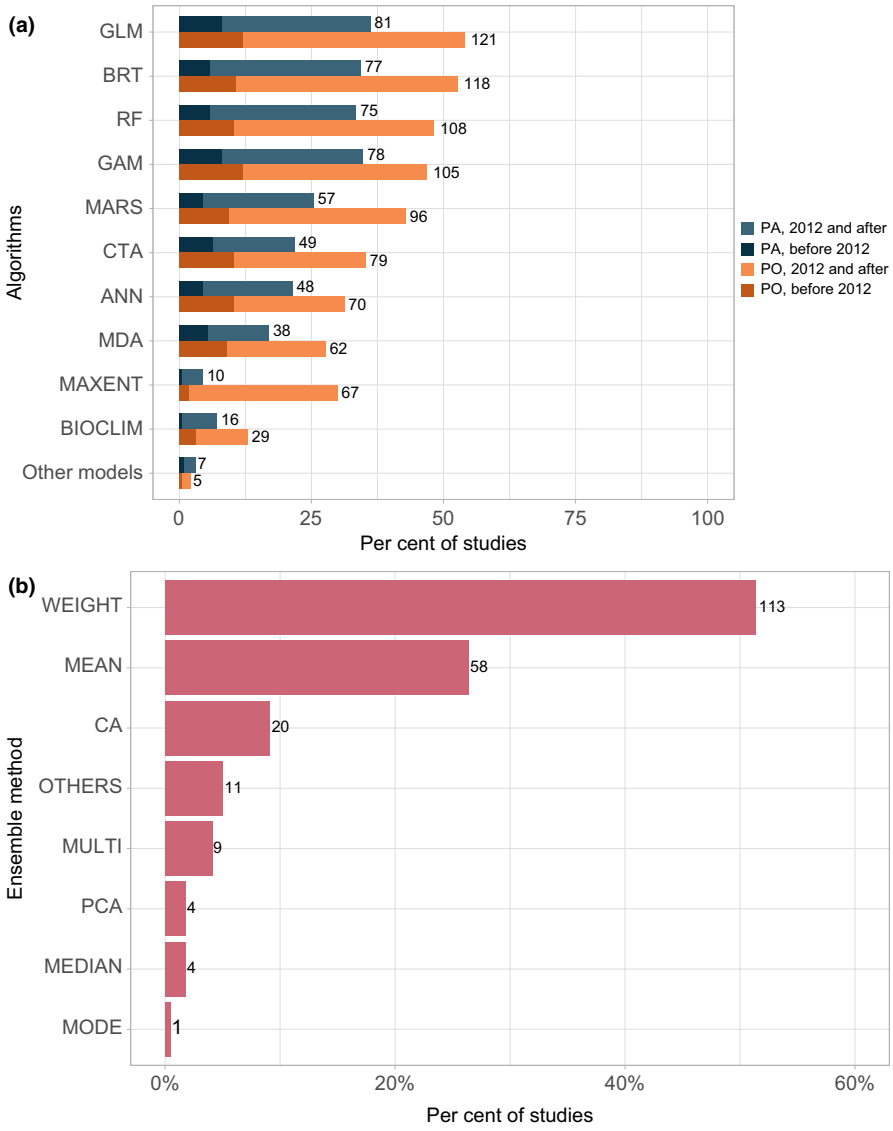

FIGURE 3 (a) Frequency of ensemble species distribution model studies employing particular individual modelling algorithms in their ensembles. For abbreviation of algorithm names see Table 1. We distinguish those using presence-only data (PO) from those using presence–absence data (PA) because MaxEnt and BIOCLIM are specifically designed for PO data and thus we expect higher usage of these methods among PO studies. We further distinguish those published before 2012 from those in 2012 or later, because MaxEnt only became available in 2012. Floating numbers represent total count of papers employing each method for either PA or PO data. (b) Frequency of ensemble methods among ensemble SDM studies (WEIGHT = weighted mean, CA = committee averaging, MULTI = using multiple ensemble methods, PCA = a special case of median where the ensemble uses median of models selected by a principle component analysis on all models, see Marmion et al., 2009 for detailed explanation). Floating numbers represent count of papers

in ensembles by some studies (n = 85). In these cases, modellers either exclude individual models from ensembles based on a pre-selected arbitrary cut-off threshold of performance (e.g., AUC > 0.8), or choose to ensemble only the best performing individual models (e.g., the top four algorithms out of eight).

In terms of ensemble methods, Weighted Mean, used by 113 studies, is by far the most popular (Figure 3b), possibly because it has been suggested to perform best (Marmion et al., 2009). The next most popular approaches were unweighted Mean and Committee Averaging (58 and 20 studies, respectively).

4.5 | Performance

Perhaps the most interesting result scientifically is how these ensemble models perform. Here the review results are disappointing. Despite the large number of papers (694 in our literature search) referring to BIOMOD, our review reveals that there is limited unambiguous information on the performance of ensemble models relative to individual modelling approaches. Only 46 out of 224 reviewed studies report the performance of ensemble models to individual models, predominantly by comparing ensemble models to the individual models used within the ensemble (rather than to independently tuned individual models). This is common practice in the BIOMOD applications, since most applications are not primarily exploring ensemble performance in comparison with that of other models or other model selection approaches. Furthermore, as many of our reviewed studies involve model transfer into novel environments, the performance of those models could not be directly evaluated in the unobserved prediction space.

Results from the 46 studies that do compare performance (Figure 4) generally favour ensembles over individual models: Ensembles are the best performing models in 21 studies, and in 17 studies while they do not always outperform the best performing individual models (i.e., are not always the best across species or across different metrics of model validation) they are still the best in some instances. Only eight studies provide evidence that ensembles are consistently, across species and metrics, worse than

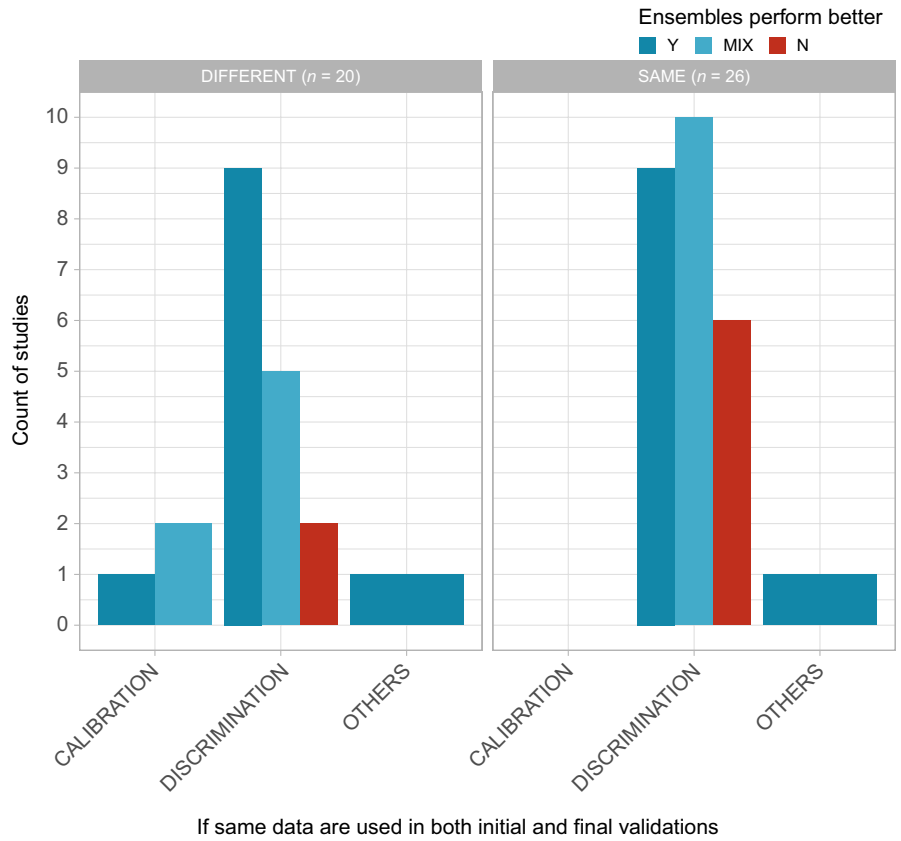

FIGURE 4 Evidence of ensemble model performance relative to individual model performance, based on 46 studies reporting such information among a sample of 224 papers identified through reviewing literature citing BIOMOD. We distinguish whether ensembles models were validated on the same set of data as those used in initial validations for individual models, or whether an additional round of final validation is performed for all models, using a different set of data. Performance comparison results are categorised as: Y = ensembles are the best performing models; MIX = ensembles do not always outperform the best performing individual models (i.e., are not always the best across species or across different metrics of model validation) but are still the best in some instances; N = ensembles are consistently, across species and metrics, worse than at least one individual model

at least one individual model. Among the 46 studies, 20 report performing both initial validations for individual models and additional final validations for both individual and ensemble models. The final validations use different data from those used in initial validations (eight use a held-out proportion of original data different from those used in initial validation, and 12 use independent data). The remaining 26 studies use the same data for validating both initial individual models and final ensemble models. In theory, we can expect that model performance assessed in this way could be biased towards "performance-informed" ensembles (i.e., Weighted Mean or ensembles selecting models performing best on evaluation data). This is because these ensembles could favour individual models that performed better on the held-out evaluation data, and therefore are primed to behave well on that held-out dataset in the final validation. However, we do not observe evidence of such a bias in the studies reviewed (there is no evidence that the performance of ensembles relative to individual models is different between those 26 studies and the other 20 studies; Pearson's chi-squared test, p-value ≈0.40; also see Figure 4).

In our selected set of papers, two studies specifically aim to compare performances between ensemble and individual models (Crimmins et al., 2013; Marmion et al., 2009). Using a held-out proportion of original data for evaluation, Marmion et al. (2009) find ensemble models to outperform their individual counterparts. In contrast, using temporally independent data, Crimmins et al. (2013) find ensembles to be outperformed by GLMs. The remaining papers in our selected set report performance of both ensemble and individual models either because they investigate other factors affecting model performance (e.g., Breiner, Guisan, Bergamini, & Nobis, 2015 investigate performance of generic ensemble models vs. ensembles of single-predictor models), or perhaps because BIOMOD functions provide performance results by default (when using the same initial validation procedures for individual models and ensembles). In summary, while we were highly interested in understanding model performance and were initially planning a quantitative analysis of results to understand whether ensembles perform better in some circumstances or with some data compared with others, the evidence is insufficient for such analysis. The main problems

are lack of evaluation on data not used to tune the ensembles and, more generally, lack of reporting of model performance.

4.6 | Use of ensembles

In the Introduction, we noted the continual growth in popularity of the ensemble approach in SDMs, and this review provides relevant data. Figure 5 presents the number of reviewed papers published each year since the initial release of BIOMOD in 2003. The figure also shows some key events in the development of ensemble SDMs, which may have contributed to the proliferation of ensembles.

The results above summarize the main trends in ensemble SDM use. Further specifics that may be of interest to particular readers are available in Supporting Information Appendix S2.

5 | LEARNING MORE

Our review provides a broad overview of the patterns and habits of BIOMOD ensemble users. However, we were unable to insightfully characterize these studies in terms of modelling choices and predictive performance of models, because documentation on these aspects is often lacking or ambiguous. For instance, while most studies reported what algorithms are included in the ensemble, relatively few detailed the version of BIOMOD being used and whether they used default settings or how they fine-tuned the individual models. We know that these methodological details are not central to the main aim of many authors, but documenting this information would create a deposit of knowledge on ensemble modelling that would enable meaningful meta-analyses regarding performance. Particularly, it could help to build-up a strong, data-driven picture of how ensembles perform under different conditions, and how the choices made in the modelling process affect model performance. This opportunity would only be possible if data and modelling methods were thoroughly documented by authors, and code and data made available. In this section, we discuss the methodological choices less explored and documented by authors in our sampled studies, and give suggestions to BIOMOD users on how to more usefully report their modelling processes, so that their publications can be used to provide substantial evidence about performance and to support best practice modelling.

Among modelling choices, data partitioning for cross-validation and model tuning are least discussed among the sampled papers. A suite of methods exists for partitioning data into training and evaluation sets, such as *k*-fold cross-validation or bootstrapping (Hastie et al., 2009). However, we observe most of our sampled studies (198 of the 217 studies providing details on validation) use repeated random splitting of data, which is the default strategy in "biomod2." There is growing recognition that different cross-validation strategies test different aspects of model performance, and that the strategy should match the application of the models—for

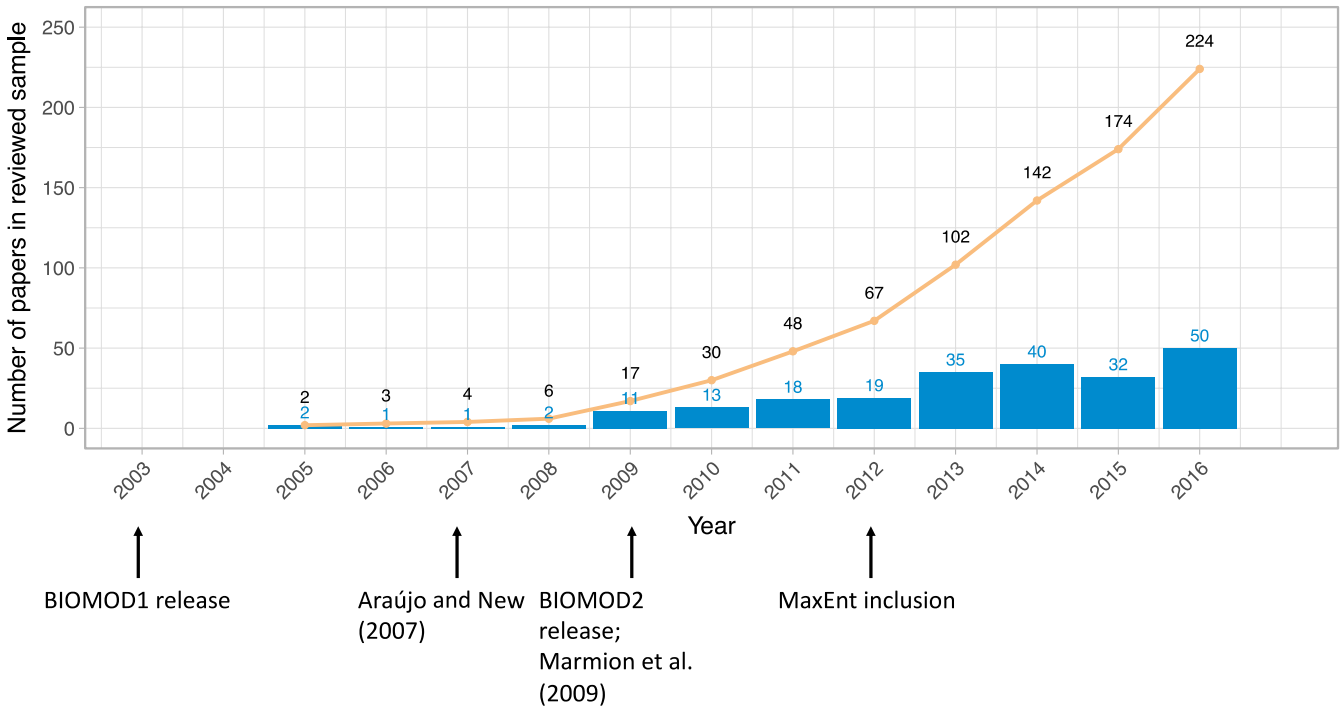

FIGURE 5 The number of papers published each year from 2003 to 2016 in a sample of 224 BIOMOD-related papers. Some key events in BIOMOD's timeline are marked: the initial release for BIOMOD 1 (Thuiller, 2003), the publication of Araújo and New, (2007) laying a popular theoretical framework for ensemble modelling in SDMs, Marmion et al., (2009) demonstrating predictive superiority of ensemble models (particularly weighted ensembles), the release of "biomod2" (Thuiller et al., 2009), and implementation of the popular MaxEnt algorithm in "biomod2" (Thuiller et al., 2016)

instance, at times it may be appropriate to spatially or environmentally separate the folds using block cross-validation (Roberts et al., 2017; Valavi, Elith, Lahoz-Monfort, & Guillera-Arroita, 2018). Different strategies will impact how models are weighted in a weighted ensemble, and thus, potentially affect how the ensemble models predict. We suggest this is well worth exploring further for users of BIOMOD ensembles, to build understanding of optimal strategies for the range of likely applications. In addition, for users who aim to fairly compare performance of "performance-informed" ensembles to other models but who do not have independent validation data, we suggest a "two-step" internal validation approach (Figure 6), which we observe in Marmion et al. (2009) and Meller et al. (2014). This approach involves first dividing all data into "outer" training and testing sets, and then further dividing outer training sets into "inner" training and testing sets. The inner sets are used to train and test individual models to provide weights to Weighted Mean ensembles. In the next step, all training data are used to train the final individual and ensemble models, which are all validated on outer testing sets. This approach avoids informing ensembles about individual model performance on final testing data, and therefore ensures fair comparison.

We gathered limited evidence from our sampled studies on if and how model tuning affects performance of these types of ensembles. The lack of information on model tuning may reflect a belief among authors that when an ensemble approach is used, tuning of individual models is no longer relevant, or it may the practical difficulty of implementing both tuning and ensemble procedures, as both can be complex. However, one would expect model tuning to be important because: (a) there is strong evidence that model tuning affects performance of individual models (e.g., Anderson & Gonzalez, 2011); and (b) it has been repeatedly suggested that ensemble model performance is dependent on the quality of individual models that compose the ensemble (Araújo & New, 2007; Araújo, Whittaker, Ladle, & Erhard, 2005; Marmion et al., 2009). Presumably many authors used BIOMOD defaults, either due to inexperience with the modelling methods, or because they assumed these settings to be the optimal tuning. Evidence is lacking for whether these defaults are optimal. By extension, we also cannot discern based on the available literature how ensembles with default tuning of individual models predict compared to well-tuned individual models or ensembles of well-tuned models.

To increase our knowledge regarding how performance of these types of ensembles is affected by choices in the modelling process, it would be beneficial for future ensemble SDM literature to have thorough coverage about these methodological choices. We emphasise again that, while model performance is not the main focus of many SDM studies, reporting methodological details is both beneficial to the accumulation of knowledge in the area and to the reproducibility of SDM studies in general. We provide an outline of the types of information that could be documented in Figure 6. This figure is by no means an exhaustive checklist of all possible choices in an ensemble modelling workflow, especially if the modelling workflow is very different from that of BIOMOD. However, with either explicit reporting of software version and selected methodology, or ideally with published code and data, one could fully reproduce ensemble SDM studies and build a comprehensive picture of the performance of ensembles across datasets and applications. We encourage authors publishing new work to provide such details to support an enhanced

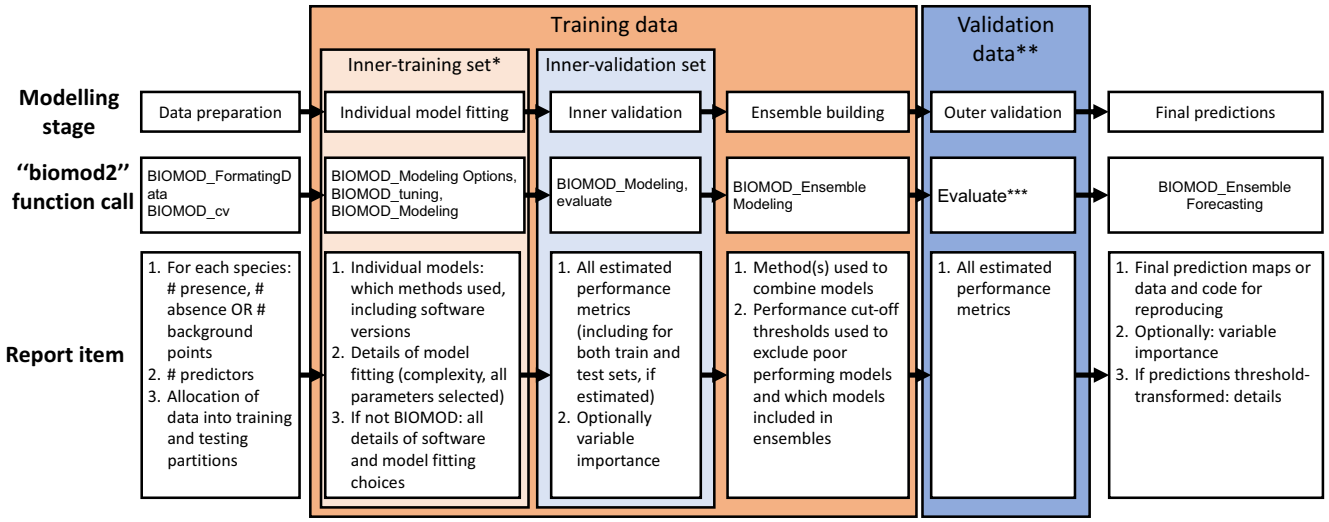

FIGURE 6 Schematic of a typical ensemble modelling workflow, including BIOMOD function calls and partitioning of data used in each stage of modelling, and items to report for transparent and reproducible documentation of the modelling process. *Partitioning data for inner-training and validation ensures unbiased comparison between performance of "performance-informed" ensembles and that of other models. When only "performance-naïve" ensembles are used (e.g., Mean or Median), such step is not necessary. **Validation set can be a held-out proportion of original data or independently collected data. ***Currently "biomod2" does not support two-step validations within its natural workflow, so users must manually reserve the outer validation set. However, users can use the "biomod2" function "evaluate" to validate models on any data, including those held-out in the validation set

understanding of how these models perform across the breadth of applications.

ACKNOWLEDGEMENTS

The authors were 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.

DATA ACCESSIBILITY

The full list of articles read and included in this review, and the documentation of reviewed articles are available in Supporting Information Appendices S1 and S2.

ORCID

Tianxiao Hao https://orcid.org/0000-0003-4363-1956
Jane Elith https://orcid.org/0000-0002-8706-0326
Gurutzeta Guillera-Arroita https://orcid.org/0000-0002-8387-5739
José J. Lahoz-Monfort https://orcid.org/0000-0002-0845-7035

REFERENCES

Aalto, J., & Luoto, M. (2014). Integrating climate and local factors for geomorphological distribution models. *Earth Surface Processes and Landforms*, *39*, 1729–1740. https://doi.org/10.1002/esp.3554

Allouche, O., Tsoar, A., & Kadmon, R. (2006). Assessing the accuracy of species distribution models: Prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology*, *43*, 1223–1232. https://doi.org/10.1111/j.1365-2664.2006.01214.x

Anderson, R. P., & Gonzalez, I. (2011). Species-specific tuning increases robustness to sampling bias in models of species distributions: An implementation with Maxent. *Ecological Modelling*, *222*, 2796–2811. https://doi.org/10.1016/j.ecolmodel.2011.04.011

Araújo, M. B., & Guisan, A. (2006). Five (or so) challenges for species distribution modelling. *Journal of Biogeography*, *33*, 1677–1688. https://doi.org/10.1111/j.1365-2699.2006.01584.x

Araújo, M. B., & New, M. (2007). Ensemble forecasting of species distributions. *Trends in Ecology and Evolution*, *22*, 42–47. https://doi.org/10.1016/j.tree.2006.09.010

Araújo, M., Whittaker, R., Ladle, R., & Erhard, M. (2005). Reducing uncertainty in projections of extinction risk from climate change. *Global Ecology and Biogeography*, *14*, 529–538. https://doi.org/10.1111/j.1466-822X.2005.00182.x

Breiman, L. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth International Group.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Breiner, F. T., Guisan, A., Bergamini, A., & Nobis, M. P. (2015). Overcoming limitations of modelling rare species by using ensembles of small models. *Methods in Ecology and Evolution*, *6*, 1210–1218. https://doi.org/10.1111/2041-210X.12403

Burnham, K. P., & Anderson, D. R. (2003). *Model selection and multimodel inference: A practical information-theoretic approach*. Berlin, Germany: Springer Science & Business Media.

Busby, J. R. (1991).BIOCLIM – a bioclimate analysis and prediction system. In C. R. Margules & M. P. Austin (Eds.), *Nature conservation: cost effective biological surveys and data analysis* (pp. 64–68). Melbourne, Australia: CSIRO

Crimmins, S. M., Dobrowski, S. Z., & Mynsberge, A. R. (2013). Evaluating ensemble forecasts of plant species distributions under climate change. *Ecological Modelling*, *266*, 126–130. https://doi.org/10.1016/j.ecolmodel.2013.07.006

Crossman, N. D., & Bass, D. A. (2008). Application of common predictive habitat techniques for post-border weed risk management. *Diversity and Distributions*, *14*, 213–224. https://doi.org/10.1111/j.1472-4642.2007.00436.x

Cutler, D. R., Edwards, T. C. Jr, Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology*, *88*, 2783–2792. https://doi.org/10.1890/07-0539.1

de Laplace, P. S. (1818). *Théorie analytique des probabilités: Supplément a la théorie analytique des probabilités*. Paris, France: Courcier.

Dietterich, T. G. (2000). *Ensemble methods in machine learning. International Workshop on Multiple Classifier Systems*, 1–15.

Diniz-Filho, J. A. F., Bini, L. M., Rangel, T. F., Loyola, R. D., Hof, C., Nogués-Bravo, D., & Araújo, M. B. (2009). Partitioning and mapping uncertainties in ensembles of forecasts of species turnover under climate change. *Ecography*, *32*, 897–906. https://doi.org/10.1111/j.1600-0587.2009.06196.x

Dormann, C. F., Calabrese, J. M., Guillera-Arroita, G., Matechou, E., Bahn, V., Bartoń, F., … Hartig, F. (2018). Model averaging in ecology: A review of Bayesian, information-theoretic and tactical approaches. *Ecological Monographs*, *88*, 485–504. https://doi.org/10.1002/ecm.1309

Dormann, C. F., Gruber, B., Winter, M., & Herrmann, D. (2010). Evolution of climate niches in European mammals? *Biology Letters*, *6*, 229–232.

Elith, J., Graham, C. H., Anderson, R. P., Dudík, M., Ferrier, S., Guisan, A., … Zimmermann, N. E. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, *29*, 129–151. https://doi.org/10.1111/j.2006.0906-7590.04596.x

Elith, J., & Leathwick, J. R. (2009). Species distribution models: Ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, *40*, 677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, *77*, 802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x

Elith, J., Phillips, S. J., Hastie, T., Dudik, M., Chee, Y. E., & Yates, C. J. (2011). A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, *17*, 43–57.

Engler, R., Waser, L. T., Zimmermann, N. E., Schaub, M., Berdos, S., Ginzler, C., & Psomas, A. (2013). Combining ensemble modeling and remote sensing for mapping individual tree species at high spatial resolution. *Forest Ecology and Management*, *310*, 64–73. https://doi.org/10.1016/j.foreco.2013.07.059

Fitzpatrick, M. C., & Hargrove, W. W. (2009). The projection of species distribution models and the problem of non-analog climate. *Biodiversity and Conservation*, *18*, 2255–2261. https://doi.org/10.1007/s10531-009-9584-8

Folmer, E. O., van Beusekom, J. E. E., Dolch, T., Gräwe, U., van Katwijk, M. M., Kolbe, K., & Philippart, C. J. M. (2016). Consensus forecasting of intertidal seagrass habitat in the Wadden Sea. *Journal of Applied Ecology*, *53*, 1800–1813. https://doi.org/10.1111/1365-2664.12681

Franklin, J. (2010). *Mapping species distributions: spatial inference and prediction*. Cambridge, UK: Cambridge University Press.

Friedman, J. H. (1991). Multivariate Adaptive Regression Splines. *The Annals of Statistics*, *19*, 1–67. https://doi.org/10.1214/aos/1176347963

Friedman, J. H., & Popescu, B. E. (2008). Predictive learning via rule ensembles. *The Annals of Applied Statistics*, *2*, 916–954. https://doi.org/10.1214/07-AOAS148

Golding, N., August, T. A., Lucas, T. C. D., Gavaghan, D. J., van Loon, E. E., & McInerny, G. (2018). The ZOON R package for reproducible and shareable species distribution modelling. *Methods in Ecology and Evolution*, *9*, 260–268. https://doi.org/10.1111/2041-210x.12858

Gregory, A. W., Smith, G. W., & Yetman, J. (2001). Testing for Forecast Consensus. *Journal of Business and Economic Statistics*, *19*, 34–43. https://doi.org/10.1198/07350010152472599

Guillera-Arroita, G., Lahoz-Monfort, J. J., Elith, J., Gordon, A., Kujala, H., Lentini, P. E., … Wintle, B. A. (2015). Is my species distribution model fit for purpose? Matching data and models to applications. *Global Ecology and Biogeography*, *24*, 276–292. https://doi.org/10.1111/geb.12268

Guisan, A., & Thuiller, W. (2005). Predicting species distribution: Offering more than simple habitat models. *Ecology Letters*, *8*, 993–1009. https://doi.org/10.1111/j.1461-0248.2005.00792.x

Guisan, A., Thuiller, W., & Zimmermann, N. E. (2017). *Habitat suitability and distribution models: With applications in R*. Cambridge, UK: Cambridge University Press.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*, 29–36. https://doi.org/10.1148/radiology.143.1.7063747

Hardy, S. M., Lindgren, M., Konakanchi, H., & Huettmann, F. (2011). Predicting the distribution and ecological niche of unexploited snow crab (*Chionoecetes opilio*) populations in Alaskan waters: A First open-access ensemble model. *Integrative and Comparative Biology*, *51*, 608–622. https://doi.org/10.1093/icb/icr102

Hastie, T., & Tibshirani, R. (2004). *Generalized additive models. Encyclopedia of statistical sciences*. Chichester, UK: John Wiley & Sons Inc.

Hastie, T., Tibshirani, R., & Buja, A. (1994). Flexible discriminant analysis by optimal Scoring. *Journal of the American Statistical Association*, *89*, 1255–1270. https://doi.org/10.1080/01621459.1994.10476866

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. New York, NY: Springer.

Hijmans, R. J., & Graham, C. H. (2006). The ability of climate envelope models to predict the effect of climate change on species distributions. *Global Change Biology*, *12*, 2272–2281. https://doi.org/10.1111/j.1365-2486.2006.01256.x

Marmion, M., Parviainen, M., Luoto, M., Heikkinen, R. K., & Thuiller, W. (2009). Evaluation of consensus methods in predictive species distribution modelling. *Diversity and Distributions*, *15*, 59–69. https://doi.org/10.1111/j.1472-4642.2008.00491.x

McCullagh, P. (1984). Generalized linear models. *European Journal of Operational Research*, *16*, 285–292. https://doi.org/10.1016/0377-2217(84)90282-0

Meller, L., Cabeza, M., Pironon, S., Barbet-Massin, M., Maiorano, L., Georges, D., & Thuiller, W. (2014). Ensemble distribution models in conservation prioritization: From consensus predictions to consensus reserve networks. *Diversity and Distributions*, *20*, 309–321. https://doi.org/10.1111/ddi.12162

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G., Group TP (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine*, *6*, 1–6. https://doi.org/10.1371/journal.pmed.1000097

Morales, N. S., Fernández, I. C., & Baca-González, V. (2017). MaxEnt's parameter configuration and small samples: Are we paying attention to recommendations? A Systematic Review. *Peerj*, *5*, e3093. https://doi.org/10.7717/peerj.3093

Naimi, B., & Araújo, M. B. (2016). sdm: A reproducible and extensible R platform for species distribution modelling. *Ecography*, *39*, 368–375. https://doi.org/10.1111/ecog.01881

Pearson, R. G., Thuiller, W., Araujo, M. B., Martinez-Meyer, E., Brotons, L., McClean, C., … Lees, D. C. (2006). Model-based uncertainty in species range prediction. *Journal of Biogeography*, *33*, 1704–1711. https://doi.org/10.1111/j.1365-2699.2006.01460.x

Peterson, A. T., Soberón, J., Pearson, R. G., Anderson, R. P., Martínez-Meyer, E., Nakamura, M., & Araújo, M. B. (2011). *Ecological niches and geographic distributions*. Princeton, NJ: Princeton University Press.

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, *190*, 231–259. https://doi.org/10.1016/j.ecolmodel.2005.03.026

R Core Team (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Randin, C. F., Dirnböck, T., Dullinger, S., Zimmermann, N. E., Zappa, M., & Guisan, A. (2006). Are niche-based species distribution models transferable in space? *Journal of Biogeography*, *33*, 1689–1703. https://doi.org/10.1111/j.1365-2699.2006.01466.x

Ripley, B. D. (2007). *Pattern recognition and neural networks*. Cambridge, UK: Cambridge University Press.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., … Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, *40*, 913–929. https://doi.org/10.1111/ecog.02881

Sanders, F. (1963). On subjective probability forecasting. *Journal of Applied Meteorology*, *2*, 191–201. https://doi.org/10.1175/1520-0450(1963)002<0191:OSPF>2.0.CO;2

Segurado, P., & Araujo, M. (2004). An evaluation of methods for modelling species distributions. *Journal of Biogeography*, *31*, 1555–1568. https://doi.org/10.1111/j.1365-2699.2004.01076.x

Seni, G., & Elder, J. F. (2010). Ensemble methods in data mining: Improving accuracy through combining predictions. *Synthesis Lectures on Data Mining and Knowledge Discovery*, *2*, 1–126. https://doi.org/10.2200/S00240ED1V01Y200912DMK002

Sequeira, A. M. M., Bouchet, P. J., Yates, K. L., Mengersen, K., & Caley, M. J. (2018). Transferring biodiversity models for conservation: Opportunities and challenges. *Methods in Ecology and Evolution*, *9*, 1250–1264. https://doi.org/10.1111/2041-210X.12998

Thuiller, W. (2003). BIOMOD – optimizing predictions of species distributions and projecting potential future shifts under global change. *Global Change Biology*, *9*, 1353–1362. https://doi.org/10.1046/j.1365-2486.2003.00666.x

Thuiller, W. (2004). Patterns and uncertainties of species' range shifts under climate change. *Global Change Biology*, *10*, 2020–2027. https://doi.org/10.1111/j.1365-2486.2004.00859.x

Thuiller, W., Georges, D., Engler, R., & Breiner, F. (2016). *'biomod2': Ensemble platform for species distribution modeling*.

Thuiller, W., Lafourcade, B., Engler, R., & Araújo, M. B. (2009). BIOMOD–a platform for ensemble forecasting of species distributions. *Ecography*, *32*, 369–373. https://doi.org/10.1111/j.1600-0587.2008.05742.x

Valavi, R., Elith, J., Lahoz-Monfort, J. J., & Guillera-Arroita, G. (2018). blockCV: An R package for generating spatially or environmentally separated folds for *k*-fold cross-validation of species distribution models. *Methods in Ecology and Evolution*, https://doi.org/10.1111/2041-210X.13107

BIOSKETCH

The first author of this manuscript, Tianxiao Hao, is a research student in quantitative ecology. His main research interest is in the methodology and application of species distribution models (SDMs), including questions of predictive performance, model selection and data-based uncertainties. Tianxiao's current research projects investigate: (a) the predictive performance of "ensemble"-type SDMs, and (b) using citizen science dataset to model distribution of Australian macrofungi and to describe Australian macrofungal biogeography.

Author contributions: The ideas for this paper were jointly conceived by all authors; T.H. conducted the literature review and analysed the data; T.H. led the writing, with contributions from all.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Hao T, Elith J, Guillera-Arroita G, Lahoz-Monfort J.J. A review of evidence about use and performance of species distribution modelling ensembles like BIOMOD. *Divers Distrib*. 2019;00:1–14. https://doi.org/10.1111/ddi.12892