Model averaging in ecology: a review of Bayesian, information-theoretic and tactical approaches for predictive inference

Carsten F. Dormann*[1], Justin M. Calabrese[2], Gurutzeta Guillera-Arroita[3], Eleni Matechou[4], Volker Bahn[5], Kamil Bartoń[6], Colin M. Beale[7], Simone Ciuti[1,20], Jane Elith[3], Katharina Gerstner[8, 9], Jérôme Guelat[10], Petr Keil[9], José J. Lahoz-Monfort[3], Laura J. Pollock[12], Björn Reineking[13, 14], David R. Roberts[1,22], Boris Schröder[15,16], Wilfried Thuiller[12], David I. Warton[17], Brendan A. Wintle[3], Simon N. Wood[18], Rafael O. Wüest[12,21] & Florian Hartig[1,19]

[1]Biometry and Environmental System Analysis, University of Freiburg, Germany
[2]Smithsonian Conservation Biology Institute, Front Royal, USA
[3]School of BioSciences, University of Melbourne, Australia
[4]School of Mathematics, Statistics and Actuarial Science, University of Kent, UK
[5]Department of Biological Sciences, Wright State University, USA
[6]Institute of Nature Conservation, Polish Academy of Sciences, Kraków, Poland
[7]Department of Biology, University of York, UK
[8]Computational Landscape Ecology, Helmholtz Centre for Environmental Research-UFZ, Leipzig, Germany
[9]German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Germany
[10]Swiss Ornithological Institute, Sempach, Switzerland
[12]Univ. Grenoble Alpes, Laboratoire d'Écologie Alpine (LECA), CNRS, Grenoble, France
[13]Univ. Grenoble Alpes, Irstea, UR LESSEM, F-38402 St-Martin-d'Hères, France
[14]Biogeographical Modelling, Bayreuth Center of Ecology and Environmental Research BayCEER, University of Bayreuth, Germany.
[15]Landscape Ecology and Environmental Systems Analysis, Institute of Geoecology, Technische Universität Braunschweig, Germany
[16]Berlin-Brandenburg Institute of Advanced Biodiversity Research (BBIB), Berlin, Germany
[17]School of Mathematics and Statistics, Evolution & Ecology Research Centre, University of New South Wales, Australia
[18]School of Mathematics, Bristol University, UK
[19]Group for Theoretical Ecology, University of Regensburg, Germany
[20]Laboratory of Wildlife Ecology and Behaviour, School of Biology and Environmental Science, University College Dublin, Ireland
[21]Swiss Federal Institute for Forest, Snow and Landscape Research WSL, Birmensdorf, Switzerland
[22]Department of Geography, University of Calgary, Canada

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/ecm.1309

Running head: Model averaging in ecology

Keywords: AIC-weights, ensemble, model combination, model averaging, nominal coverage, prediction averaging, uncertainty

Abstract

In ecology, the true causal structure for a given problem is often not known, and several plausible models and thus model predictions exist. It has been claimed that using weighted averages of these models can reduce prediction error, as well as better reflect model selection uncertainty. These claims, however, are often demonstrated by isolated examples. Analysts must better understand under which conditions model averaging can improve predictions and their uncertainty estimates. Moreover, a large range of different model averaging methods exists, raising the question of how they differ regarding in their behaviour and performance.

Here, we review the mathematical foundations of model averaging along with the diversity of approaches available. We explain that the error in model-averaged predictions depends on each model's predictive bias and variance, as well as the covariance in predictions between models and uncertainty about model weights.

We show that model averaging is particularly useful if the predictive error of contributing model predictions is dominated by variance, and if the covariance between models is low. For noisy data, which predominate in ecology, these conditions will often be met.

Many different methods to derive averaging weights exist, from from Bayesian over information-theoretical to cross-validation optimised and resampling approaches. A general recommendation is difficult, because the performance of methods is often context-dependent. Importantly, estimating weights creates some additional

*corresponding author; Tennenbacher Str. 4, 79106 Freiburg, Email: carsten.dormann@biom.uni-freiburg.de

uncertainty. As a result, estimated model weights may not always outperform arbitrary fixed weights, such as equal weights for all models. When averaging a set of models with many inadequate models, however, estimating model weights will typically be superior to equal weights.

We also investigate the quality of the confidence intervals calculated for model-averaged predictions, showing that they differ greatly in behaviour and seldom manage to achieve nominal coverage. Our overall recommendations stress the importance of non-parametric methods such as cross-validation for a reliable uncertainty quantification of model-averaged predictions.

1 Introduction

Models are an integral part of ecological research, representing alternative, possibly overlapping, hypotheses (Chamberlin, 1890). They are also the standard approach to making predictions about ecological systems (Mouquet et al., 2015). In many cases, it is not possible to clearly identify a single most-appropriate model. For instance, process-based models may differ in the specific ways they represent ecological mechanisms, without a clear empirical or theoretical reason to prefer one option over the other. Statistical analyses rarely offer a single solution, both because the limited amount of data allows for several plausible combinations of predictors, and because different modelling approaches are available for statistical analysis (e.g. Hastie et al., 2009; Kuhn and Johnson, 2013).

Model averaging seemingly solves this dilemma. Proponents of this approach have claimed that calculating a weighted average of the predictions of all candidate models will reduce prediction error through reduced variance and bias (the latter based on arguments described in Madigan and Raftery, 1994), as well as better represent uncertainty about model parametrisation and structure (Wintle et al., 2003, see also section 2.3). For some ecological examples of model averaging see Thuiller (2004); Richards (2005); Brook and Bradshaw (2006); Dormann et al. (2008); Diniz-Filho et al. (2009); Le Lay et al. (2010); Garcia et al. (2012); Cariveau et al. (2013); Meller et al. (2014), and Lauzeral et al. (2015).

Evaluating the utility of this approach is complicated by the large number of different method for model averaging and the subsequent uncertainty quantification of averaged predictions. Several previous reviews on model averaging in ecology and evolution, focussed exclusively on 'information-theoretical model averaging' (Johnson and Omland, 2004; Hobbs and Hilborn, 2006; Burnham et al., 2011; Freckleton, 2011; Grueber et al., 2011; Nakagawa and Freckleton, 2011; Richards et al., 2011; Symonds and Moussalli, 2011), probably under the influence of the AIC-weighted averaging popularised by Burnham & Anderson (2002; Posada and Buckley 2004). *Bayesian* model averaging has been used less frequently in ecology (for an example see Corani and Mignatti, 2015), but for an excellent recent review of this topic in the context of Bayesian model selection see Hooten and Hobbs (2015, see also Hoeting et al. 1999; Ellison 2004; Link and Barker 2006). However, none of the above covers all available model averaging approaches, together with a general discussion of advantages and disadvantages.

Our aim is to provide such a comprehensive review in the light of developments over the last 20 years, summarising the mathematical reasoning behind model averaging, and offering an intuitive but technically sound entry to the field, illustrated by case studies. We primarily address prediction averaging of correlative models, although most of the points will similarly apply to mechanistic/process-based models (see, e.g., Knutti et al., 2010; Diks and Vrugt, 2010, for a review in the context of climate and hydrological models, respectively). We do not consider averaging model parameters, because we agree with the criticism summarised in Banner and Higgs (2017): parameters (such as partial regression coefficients) are estimated conditional on the model structure; as the model structure changes, parameters may become incommensurable (see Posada and Buckley, 2004; Cade, 2015; Banner and Higgs, 2017, and Appendix S1.1 for short review of the parameter-averaging literature). Instead, our focus is on prediction, and predictive inference (sensu Geisser, 1993), as exemplified by model-averaged predictions of species potential occurence for reserve-site selection (Meller et al., 2014) or the effect of roads on occupancy of ponds by frogs (Dai and Wang, 2011). Also, we only focus on averaging sets of models that differ in structure, as opposed to mere differences in initial conditions or parameter values (Gibbs, 1902; Johnson and Bowler, 2009). The latter case is called "ensemble" in the statistical and physical sciences, while in ecology that term is used more loosely.

This review is divided into five parts: first, we present the mathematical logic behind model averaging, and why this alone puts severe constraints on *how* we do model averaging. Then, in the second part, we review the different ways through which model-averaging weights can be derived, comparing Bayesian, information-theoretic, and tactical perspectives (by tactical we mean heuristic approaches to model averaging that are not explicitly based on statistical theory). This is followed by a brief exploration of how to quantify the uncertainty of model-averaged predictions. Finally, we briefly illustrate model averaging with two case studies, before closing with unresolved challenges, and recommendations.

2 The mathematics behind model averaging

In accordance with virtually all discussions of model averaging we encountered, we first focus on how model averaging reduces prediction error, here quantified as mean squared error (MSE) of a prediction $\widehat{Y}_m$ of model m. As for any estimator, we can decompose this error into contributions of bias and variance:

$$\mathrm{MSE}(\widehat{Y}_m) = \left\{\mathrm{bias}(\widehat{Y}_m)\right\}^2 + \mathrm{var}(\widehat{Y}_m). \tag{1}$$

Bias refers to a systematic model error that would not change if a new dataset for the same system became available, while variance refers to the expected spread of model predictions when fit with hypothetical new datasets for the same system.

We can use eqn 1 to examine the error of a weighted average $\widetilde{Y}$ of the predictions of several (M) contributing models, $\widehat{Y}_1, \widehat{Y}_2, \ldots, \widehat{Y}_M$:

$$\widetilde{Y} = \sum_{m=1}^{M} w_m \widehat{Y}_m \ , \quad \text{with} \sum_{m=1}^{M} w_m = 1. \tag{2}$$

The motivation for the weights w_m is to adjust the average such that is has improved properties over a simple average (with equal weights) or a single candidate models (all weight on one model).

We can see from eqn 1 that bias, i.e. the difference between the expectation of the averaged predictions and the truth ($\widetilde{Y} - y^*$), will depend directly on the bias of the contributing models, as well as their weights (eqn 2). The statistical model-averaging literature often assumes that individual models have no bias, and therefore tends to be less interested in its contribution (Bates and Granger, 1969; Buckland et al., 1997; Burnham and Anderson, 2002). In contrast, for *process* models, reducing bias is often names as one of the main motivations (e.g. Solomon et al., 2007; Gibbons et al., 2008; Dietze, 2017). Implicitly, the assumption here is that model biases will tend to fall on both sides of the truth, in which case they may cancel out in an average.

Prediction variance (arising from n hypothetical repeated samplings) is composed of two terms, the variance of each contributing model's prediction,

$$\mathrm{var}(\widehat{Y}_m) = \frac{1}{n-1} \sum_{i=1}^{n} (\overline{\widehat{Y}}_m - \widehat{Y}_m^i)^2,$$

and the covariances between predictions of model m and m':

$$\mathrm{cov}(\widehat{Y}_m, \widehat{Y}_{m'}) = \frac{1}{n-1}\sum_{i=1}^{n}(\overline{\widehat{Y}}_m - \widehat{Y}_m^i)(\overline{\widehat{Y}}_{m'} - \widehat{Y}_{m'}^i).$$

For the average of two predictions, $\widehat{Y}_1$ and $\widehat{Y}_2$, this yields:

$$\mathrm{var}(\widetilde{Y}) = w_1^2\mathrm{var}(\widehat{Y}_1) + w_2^2\mathrm{var}(\widehat{Y}_2) + 2w_1w_2\mathrm{cov}(\widehat{Y}_1, \widehat{Y}_2). \tag{3}$$

When averaging several models, we expand eqn (3) to:

$$\begin{aligned}
\mathrm{var}(\widetilde{Y}) &= \mathrm{var}\left(\sum_{m=1}^{M} w_m\widehat{Y}_m\right) = \sum_{m=1}^{M} w_m^2\mathrm{var}(\widehat{Y}_m) + \sum_{m=1}^{M}\sum_{m'\neq m} w_m w_{m'}\mathrm{cov}(\widehat{Y}_m, \widehat{Y}_{m'}) \\
&= \sum_{m=1}^{M}\sum_{m'=1}^{M} w_m w_{m'}\mathrm{cov}(\widehat{Y}_m, \widehat{Y}_{m'}) \\
&= \sum_{m=1}^{M}\sum_{m'=1}^{M} w_m w_{m'}\rho_{mm'}\mathrm{var}(\widehat{Y}_m)\mathrm{var}(\widehat{Y}_{m'}), \qquad (4)
\end{aligned}$$

where $\rho_{mm'}$ is the correlation between $\widehat{Y}_m$ and $\widehat{Y}_{m'}$.

Combining eqns 1 and 3 we can see that the error of a model-averaged prediction decomposes into

$$\mathrm{MSE}(\widetilde{Y}) = \left(\sum_{m=1}^{M} w_m\left(E(\widehat{Y}_m) - y^*\right)\right)^2 + \sum_{m=1}^{M}\sum_{n=1}^{M} w_m w_{m'}\rho_{mm'}\mathrm{var}(\widehat{Y}_m)\mathrm{var}(\widehat{Y}_{m'}), \tag{5}$$

where $E(\widehat{Y}_m) - y^* = \mathrm{bias}(\widehat{Y}_m)$ represents prediction bias.

2.1 Understanding what influences the error of model-averaged prediction

Equation 5 allows us to make a number of statements about the potential benefits of model averaging. We shall first illustrate the fundamental effects of bias, variance and covariance using simply toy examples. In the next sections, we shall then move from this idealised examples to more realistic situations.

Firstly, when each model produces a distinct prediction, with variances substantially lower than systematic differences between models, bias dominates (Fig. 1

top). How useful model averaging is in this situation depends on the biases of the individual models (see also Fig. 2 top row). As model variance increases (or bias decreases), the error term is increasingly dominated by variance, and assuming covariances are low, the variance of the average (and therefore the mean error) will be smaller than the variance of the single model (Fig. 1 bottom). If the covariance of model predictions is low, increasing the number of models in the average will generally decrease the variance and therefore the prediction error, while the bias of the average has no general connection to the number of averaged models (Fig. 2, right column).

[Fig. 1 approximately here.]

We thus conclude that as bias becomes large relative to prediction variance, model averaging is less and less likely to be useful for reducing variance – but it may still be useful for reducing bias (under the condition of bidirectional bias: Fig. 2, lower row).

[Fig. 2 approximately here.]

Downweighting of variances is the mathematical reason how model averaging reduces the variance over single model predictions, as we briefly explain now.

To understand these effects in more detail, consider the unlikely, but didactically important case that model predictions are independent, meaning that their covariance is 0 and the correlation matrix ρ_{mn} of eqn 5 becomes the identity matrix (or, equivalently, the covariance term of eqn 4 vanishes). If we also assume both predictions have equal variances, $\text{var}(\widehat{Y}_1) = \text{var}(\widehat{Y}_2) = \text{var}(\widehat{Y})$, since $w_2 = 1 - w_1$, the above equation simplifies to $\text{var}(\widetilde{Y}) = (2w_1^2 - 2w_1 + 1)\text{var}(\widehat{Y})$. If one model gets all the weight, we have $\text{var}(\widetilde{Y}) = \text{var}(\widehat{Y})$. If the two models receive equal weight, we have $\text{var}(\widetilde{Y}) = (2 \cdot 0.5^2 - 2 \cdot 0.5 + 1)\text{var}(\widehat{Y}) = 0.5\text{var}(\widehat{Y})$, a considerable improvement in prediction variance (and the minimum of this equation). Other weights fall in-between these values. In other words, model averaging can reduce prediction error because weights enter as quadratic terms in eqn 3, rather than linearly.

Indeed, Bates and Granger (1969) showed that for unbiased models with uncorrelated predictions, the variance in the average is never greater than the smaller of the individual predictions (making the important assumption that the weights are known, which will be discussed below).

The next thing to note is that the correlation between model predictions, i.e. the matrix $(\rho_{ij}) \in \mathbb{R}^{M \times M}$, substantially affects the benefit of model averaging (see also Fig. 3 and interactive tool in Data S1). In the best case, correlations between model predictions are negative or at least absent, and the second term of eqn (5) is negative or vanishes. Under these conditions, averaging can substantially increase the variance of the predictions. As correlations between predictions increase, the covariance term contributes more and more to the overall prediction error. In the extreme case of perfectly correlated predictions of the single models, model averaging has no benefit for reducing prediction variance.

[Fig. 3 approximately here.]

The effect of correlations on the potential reduction of prediction error has an analogy in biodiversity studies, where it is called the 'portfolio effect' (e.g. Thibaut and Connolly, 2013). It states that the fluctuation in biomass of a community is less than the fluctuations of biomass of its members, because the species respond to the environment differently. This asynchrony in response is analogous to negative covariance in community members' biomass, buffering the *sum* of their biomasses.

This point also provides some important insights about why machine learning methods, which often average a large number of bad models, can work so well. When averaging *poor* models, e.g. trees in a Random Forest, covariance is negligible, but the variance of each model prediction is high. Because w_m becomes very small with hundreds of models (approximately $1/M$), the variance of many averaged poor models (with similar variance) tends to be low: $\text{var}(\widetilde{Y}) =$

$\sum_{m=1}^{M} \frac{1}{M^2}\text{var}(\widehat{Y}_m) + \frac{1}{M^2}\sum_{m=1}^{M}\sum_{m\neq n}\text{cov}(\widehat{Y}_m, \widehat{Y}_n) \approx M\frac{1}{M^2}\text{var}(\widehat{Y}) = \frac{1}{M}\text{var}(\widehat{Y})$,

where the second term disappears due to lack of correlations among predictions. We may speculate that poor models typically also exhibit substantial but bidirectional bias, which again would be reduced by averaging.

Putting bias, variance and correlation together (Fig. 2), we note that model averaging will deliver smaller prediction error when bias is bidirectional (i.e. model predictions over- *and* underestimate the true value: bottom row of Fig. 2) and predictions are negatively correlated (Fig. 2 bottom right). Uni-directional bias will remain problematic (top row of Fig. 2), irrespective of covariances among predictions.

Thus, for a given set of weights, the prediction error of model-averaged predictions depends on three things: the bias of the model average, as emerging from the bias of the individual models, the prediction variances of the individual models, and the covariance of those predictions.

2.2 Estimating weights can thwart the benefit of model averaging

So far, we have assumed that weights have fixed values, or that weights are not random variates, and thus there is no uncertainty about them. Yet, the aim of optimising predictive performance suggests that weights need to be estimated from the data. But estimation brings associated uncertainty with it, and this has implications for the actual benefits of model averaging: estimated "optimal" weights will be suboptimal (Nguefack-Tsague, 2014). With such an error, even for only mildly correlated predictions, the averaged prediction will more likely be biased than if the (unknown) truly optimal weights were used (Claeskens et al., 2016). It may in fact be often no better than one obtained using arbitrary weights, e.g. equal weights (Clemen, 1989; Smith et al., 2009; Graefe et al., 2014, 2015). The "simple theoretical explanation" provided by Claeskens et al. (2016) demonstrates that estimating weights introduces additional variance into the prediction. As a consequence, the predictions averaged with estimated weights may be worse than that of a single model (in contrast to the assertion of Bates and Granger 1969; see Claeskens et al. 2016 for an example).

Apart from the error of the estimate, a further open problem is to obtain a good estimator for the optimal weight in the first place. Currently no closed solution is available, not even for linear models (Liang et al., 2011). Neither Bayesian nor information-theoretical model weights are designed to minimise prediction error, and their weights will in general not be optimal for that purpose. Some tactical approaches estimate model weights explicitly to minimise prediction error on hold-out data (in particular jackknife model averaging and stacking; see section 3.3). Only these approaches are at least trying to estimate optimal weights for minimizing predictive error. The interactive tool we provide (Fig. 3) allows readers to explore this issue in a simple 2-model case. It shows that, in this simple case, estimating weights substantially reduces the parameter space where model averaging is superior to the best single model. Thus, the bias-variance trade-off applies also to model averaging, in the sense that weight estimation introduces additional parameters and therefore higher model complexity to the analysis. It is therefore important to think carefully about when to use model averaging, as it can add unnecessary complexity.

Uncertainty about the optimal weights does not imply that estimated weights are of no use, or that the use of arbitrary weights (e.g. equal weights) is generally superior. While uncertainty in estimated weights increases prediction error, the ability to statistically downweight or wholly remove unsuitable models from the prediction set is a substantial benefit. In Claeskens et al. (2016) and similar simulations, all models considered are “alright” (bias-free and with similar prediction variance), which obviously need not be the case in practical applications. Thus, the question is not if estimated model weights are useful in general, but how useful they are beyond their function of filtering out inferior models from the average. We believe there is a benefit beyond this filter function, but we recognise that there is a need for further research to better demonstrate this benefit, and understand when it occurs.

2.3 Model averaging (typically) reduces prediction errors

To complement these theoretical considerations, we examined 180 studies (a random draws from the results of a systematic literature search: see Appendix S1.7) regarding reported benefits from model averaging.

The majority of studies we encountred used an empirical approach to assess predictive performance, i.e. forecasting, hindcasting or cross-validation to observed data (e.g. Namata et al., 2008; Marmion et al., 2009*a*,*b*; Grenouillet et al., 2010; Montgomery et al., 2012; Smith et al., 2013; Engler et al., 2013; Edeling et al., 2014; Trolle et al., 2014). Most Model averaging generally yielded lower prediction errors than the individual contributing models. Most of these studies used test datasets to estimate predictive success, and rely critically on the assumption of independence between test and training datasets (Roberts et al., 2017). Few studies used simulated data to examine the performance of model averaging under specific conditions (e.g. small sample size, model structure uncertainty, missing data: Ghosh and Yuan, 2009; Schomaker, 2012), and even fewer employ analytical mathematics (Shen and Huang, 2006; Potempski and Galmarini, 2009; Chen et al., 2012; Zhang et al., 2013).

2.4 Quantifying uncertainty of model-averaged predictions

So far, we have shown that model averaging can produce predictions with a smaller error than any of the contributing models by averaging away their variance and bias. Those gains, however, generally decrease with i) increasing covariance of the individual model predictions, and ii) increasing mean bias of the contributing models. Moreover, iii) weighted averaging allows reducing the weight of models poorly supported by data, but at the expense of introducing additional variance in the average, induced by the weight estimation.

Besides having an estimate with low error, the second goal of most statistical methods is to provide a measure of (un)certainty of that estimate. The nature of this measure differs between tactical, Bayesian, and frequentist approaches. Tactical aproaches, such as machine learning, are usually satisfied with providing an estimate of predictive error on new data, typically obtained through cross-validation. This procedure can be directly extended to model-averaged predictions.

For Bayesian and frequentist methods, the issue of extending the conventional methods for estimating uncertainty to model-averaging is somewhat more complicated. Bayesian methods quantify uncertainty via the posterior distribution, which can be summarized by a Bayesian credible interval. One would interpret a 95% credible interval as displaying a 95% certainty for the true value to be contained in the interval. Frequentist methods traditionally provide a confidence interval. Under repeated sampling of new data sets under identical conditions, a correctly defined 95% confidence interval should contain the true value in 95% of the cases.

To construct a frequentist confidence interval for a model-averaged prediction, we have to ask ourselves how this model-averaged prediction will spread around the true value under repeated sampling. Fortunately, we have already derived this result in eqs. 1-5. For simple cases, we can directly convert this into a confidence interval. For example, for an unbiased average, with uncorrelated models of equal weight and variance, the standard deviation of the average, and thus its confidence interval, should decrease with one over the square root of the number of contributing models, times the confidence interval of the single models. In general, however, the calculation of the confidence interval of the average will have to take the confidence intervals of all contributing models, as well as their weights, covariance and bias into account.

Buckland et al. (1997) proposed a simplification of eqn (5), which considers bias and variance of the averaged models (for derivation see Burnham and Anderson, 2002, p. 159-162):

$$\mathrm{var}(\widetilde{Y}) = \left(\sum_{m=1}^{M} w_m \sqrt{\mathrm{var}(\widehat{Y}_m) + \gamma_m^2} \right)^2 . \qquad (6)$$

Misspecification bias of model m is computed as $\gamma_m = \widehat{Y}_m - \widetilde{Y}$, thus assuming (explicitly on page 604 of Buckland et al. 1997) that the averaged point estimate $\widetilde{Y}$ is unbiased and can hence be used to compute the bias of the individual predictions. This assumption can be visualised in Fig. 2 as the situation where the empty triangles always sit right on top of ‘truth’. This assumption is problematic, as it cannot be met by unidirectionally biased model predictions, nor when weights w_m fail to get the weighting *exactly* right and thus $\widetilde{Y}$ remains biased. Less problematically, Buckland et al. (1997) also assumed that predictions from different models are *perfectly* correlated, making the covariance term as large as possible, and variance estimation conservative. The distribution theory behind this approach has been criticised as “not (even approximately) correct” (Claeskens and Hjort, 2008, p. 207), but shown to work well in simulations (Lukacs et al., 2010; Fletcher and Dillingham, 2011).

Improving on eqn (6) requires knowledge of the covariance of model predictions $\rho_{mm'}$ (eqn 5). The key problem is that there is no analytical way to compute $\rho_{mm'}$.

Bootstrapping, although computationally costly, offers a good solution to this problem.

While the obstacles to calculate confidence intervals for model-averaged predictions may seem somewhat discouraging, it should be noted that alternatives to model averaging do not necessarily fare better. Predictions from a selected single-best model *always* underestimate the true prediction error (e.g. Namata et al., 2008; Fletcher and Turek, 2012; Turek and Fletcher, 2012). The reason is that the uncertainty about which model is correct is not included in this final prediction: we predict as if we had not carried out model selection but had known from the beginning which model would be the best (as if the model had been "prescribed": Harrell, 2001). Thus, even if we were able to choose, from our model set M, the model closest to truth, we would still need to adjust the confidence distribution for model selection; and a perfect adjustment was analytically shown not to exist (Kabaila et al., 2015).

Accordingly, simulations studies that have suggested that model averaging may improve coverage (Namata et al., 2008; Wintle et al., 2003; Zhao et al., 2013), presumably because the process of averaging allows us to take into account model uncertainty (Liang et al., 2011). Yet, given the diversity of approaches to computing model weights encountered in section 3, these studies cannot be seen as conclusive, only as suggestive, for the improvement of nominal coverage using model averaging. For example Fletcher and Turek (2012) and Turek and Fletcher (2012) explore how model averaging can improve the tail areas of the confidence distribution. These two studies, however, as well as those cited before, assumed that the full model, referring to the model that includes all sub-models prior to any model selection (see Appendix S1.3), is not in the set. The approach by Fletcher and Turek (2012) and Turek and Fletcher (2012) was re-analysed by Kabaila et al. (2015). The key finding of this latter study is that the full model coverage was still superior to all other model averaging approaches, suggesting that the full model should currently be kept in mind, both for inference, minimal bias and correct prediction intervals (see also Harrell, 2001, p. 59). Such findings sit uncomfortably with the bias-variance trade-off (Hastie et al., 2009), which states that overly complex models have poor predictive performance; and indeed the full model has high prediction variance.

Regrettably, such reasoning cannot be extended in an obvious way to non-nested models, process models, or machine learning models. Here, model averaging seems without alternative for propagating model selection uncertainty into prediction uncertainty more fairly.

Our final option to quantify uncertainty, the Bayesian credible interval, can be interpreted as a **mixture distribution**. In a two-step process, the model weights first determine the probability of any model to be correct, and the uncertainty of each model is then mixed additively into a averaged uncertainty. If the predictions of all individual models are identical, the final distribution will remain the same. From the perspective of 5, this is identical to assuming that the average models are maximally correlated, although the logical motivation for the mixing is different. If predictions differ widely, e.g. due to bias, the mixed confidence distribution will be much wider and possibly multi-modal.

To illustrate the various Bayesian and frequentist options, we calculated predictive uncertainties and coverage for four different options for a set of simple linear regressions in Fig. 5:

1. Make the assumption that model-averaged predictions are unbiased. Use bootstrapping to estimate covariances of predictions for each model. From these estimates, compute prediction variance according to eqn (5). This solution is computer-intensive, but it takes into account covariance of model predictions. On the other hand, it cannot account for bias, and should thus not be used when bias of the estimator is suspected, for example from cross-validation.

2. Make the assumption that model-averaged predictions are unbiased. Use Buckland et al. (1997)'s approach (eqn 6). This will yield wider estimates than option 1, because assumptions about bias and correlation are more conservative.

3. Use a mixture distribution to compute the confidence distribution of the average, assuming effectively that predictions from different models are perfectly correlated, but possibly biased.

4. Fit the full model (if available) and use its confidence distribution, which can rarely be improved on (Kabaila et al., 2015).

[Figure 5 approximately here.]

When averaging models with largely independent (i.e. uncorrelated) predictions, only the bootstrap-estimated covariance matrix (option 1 above) will also compute lower variances (according to eqn 4). In our example (Fig. 5, see Data S1 for details), "propagation" produced the tightest confidence interval (and hence lowest coverage), followed by "Buckland" and "mixing". However, neither of these confidence intervals seemed large enough, as all had too low coverage. Only the full model produces accurate confidence intervals and coverage. Further simulations along these lines will have to show how these approaches perform for more complex models and situations.

3 Approaches to estimating model-averaging weights

So far, we have discussed the properties of a weighted model average, but we have not discussed how to estimate the model-averaging weights. Estimating weights aims at abating poorly fitting, and elevating well-predicting models, and the actual method for estimating weights has obvious fundamental importance for the quality of an averaged prediction. Different perspectives on model-averaging weights have emerged (Table 1), which can be broadly classified into four categories of decreasing probabilistic interpretability:

1. In the Bayesian perspective, model weights are probabilities that model M_i is the 'true' model (e.g. Link and Barker, 2006; Congdon, 2007).
2. In the information-theoretic framework, model weights are measures of how closely the proposed models approximate the true model as measured by the Kullback-Leibler divergence, relative to other models.
3. In a 'tactical' perspective, model weights are parameters to be chosen in such a way as to achieve best predictive performance of the average. No specific interpretation of the model is attached to the weights; they only have to work.
4. Assigning fixed, equal weights to all predictions can be seen as a reference naïve approach, representing the situation without adjusting for differences in models' predictive abilities.

We shall address these four perspectives in turn, also hinting at relationships among them.

[Table 1 approximately here.]

3.1 Bayesian model weights

Theory Bayes' formula can be applied to choosing among models in much the same way as to parameter values (Wasserman, 2000). To perform inference with multiple models and their parameters at the same time, one can write down the joint posterior probability $P(M_i, \Theta_i|D)$ of model M_i with parameter vector Θ_i, given the observed data D, as

$$P(M_i, \Theta_i|D) \propto L(D|M_i, \Theta_i) \cdot p(\Theta_i) \cdot p(M_i), \tag{7}$$

where $L(D|M_i, \Theta_i)$ is the likelihood of model M_i, $p(\Theta_i)$ is the prior distribution of the parameters of the respective model M_i, and $p(M_i)$ is the prior weight on model M_i.

In practice, one is often interested in some simplified statistics from this distribution, such as the model with the highest posterior model probability, or the distribution of a parameter or prediction including model selection uncertainty. To obtain this information, we can marginalise (i.e. integrate) over parameter space, or marginalise over model space.

If we marginalise over parameter space, we obtain posterior model weights that represent the relative probability of each model (whilst marginalising over model space yields averaged parameters, which we shall not address here). We can alternatively calculate these weights by calculating the marginal likelihood of each model, defined as the average of eqn (7) across all k parameters for any given model:

$$P(D|M_i) \propto \int_{\Theta_1} \cdots \int_{\Theta_k} L(D|M_i, \Theta_i) p(\Theta_i) d\Theta_1 \cdots d\Theta_k. \tag{8}$$

From the marginal likelihood, we can compare models via the **Bayes factor**, defined as the ratio of their marginal likelihoods (e.g. Kass and Raftery, 1995):

$$\mathrm{BF}_{i,j} = \frac{P(D|M_i)}{P(D|M_j)} = \frac{\int L(D|M_i, \Theta_i) p(\Theta_i) d\Theta_i}{\int L(D|M_j, \Theta_j) p(\Theta_j) d\Theta_j}, \tag{9}$$

with the multiple integral now pulled together for notational convenience. For more than two models, however, it is more useful to standardise this quantity across all models in question, calculating a Bayesian posterior model weight $p(M_i|D)$ (including model priors $p(M_i)$: Kass and Raftery, 1995,) as

$$\text{posterior model weight}_i = \; p(M_i|D) = \frac{P(D|M_i)\, p(M_i)}{\sum_j P(D|M_j) p(M_j)}. \tag{10}$$

Estimation in practice While the definition of Bayesian model weights and averaged parameters is straightforward, the estimation of these quantities can be challenging. In practice, there are two options to numerically estimate the quantities defined above, both with caveats.

The first option is to sample directly from the joint posterior (eqn 7) of the models and the parameters. Basic algorithms such as rejection sampling can do that without any modification (e.g. Toni et al., 2009), but they are inefficient for higher-dimensional parameter spaces. More sophisticated algorithms such as MCMC and SMC (see Hartig et al., 2011, for a basic review) require modifications to deal with the issue of different number of parameters when changing between models. Such modifications (mostly the reversible-jump MCMCs, **rjMCMC**: Green, 1995, see Appendix S1.5.1) are often difficult to program, tune and generalise, which is the reason why they are typically only applied in specialised, well-defined settings. The posterior model probabilities of the rjMCMC are estimated as the proportion of time the algorithm spent with each model, measured as the number of iterations the algorithm drew a particular model divided by the total number of iterations.

The second option is to approximate the marginal likelihood in eqn (8) of each model independently, renormalise that into weights, and then average predictions based on these weights. The challenge here is to get a stable approximation of the marginal likelihood, which can be problematic (Weinberg, 2012, see Appendix S1.5.1). Still, because of the relatively simple implementation, this approach is a more common choice than rjMCMC (e.g. Brandon and Wade, 2006).

Influence of priors A problem for the computation of model weights when performing Bayesian inference across multiple models is the influence of the choice of *parameter priors*, especially "uninformative" ones (see section 5 in Hoeting et al., 1999;

Chickering and Heckerman, 1997).

The challenge arises because in eqns (8) and (9) the prior density $p(\theta_i)$ enters the marginal likelihood and hence the Bayes factor multiplicatively. This has the somewhat unintuitive consequence that increasing the width of an uninformative parameter prior will linearly decrease the model's marginal likelihood (e.g. Link and Barker, 2006). That Bayesian model weights are strongly dependent on the width of the prior choice has sparked discussion of the appropriateness of this approach in situations with uninformative priors. For example, in situations where multiple nested models are compared, the width of the uninformative prior may completely determine the complexity of models that are being selected. One suggestion that has been made is to *not* perform multi-model inference *at all* with uninformative priors, or that at least additional corrections are necessary to apply Bayes factors weights (O'Hagan, 1995; Berger and Pericchi, 1996). One such correction is to calibrate the model on a part of the data first, use the result as new priors and then perform the analysis described above (intrinsic Bayes factor: Berger and Pericchi 1996, fractional Bayes factor: O'Hagan 1995). If enough data are available so that the likelihood is sufficiently peaked by the calibration step, this approach should eliminate any complication resulting from the prior choice (for an ecological example see van Oijen et al., 2013).

Bayesian-flavoured approaches Apart from the natural Bayesian average (see also Yao et al., 2017), there are a number of other approaches that are connected to or inspired by Bayesian thinking.

In a set of influential publications, Raftery et al. (1997), Hoeting et al. (1999) and Raftery et al. (2005) introduced *post-hoc* Bayesian model averaging, i.e. for vectors of predictions from already fitted models. The key idea is to iteratively estimate the proportion of times a model would yield the highest likelihood within the set of models (through expectation maximisation, see Appendix S1.5.2 for details), and use this proportion as model weight. In the spirit of the inventors, we refer to this approach as **Bayesian model averaging using Expectation-Maximisation** (BMA-EM), but place it closer to a frequentist than a Bayesian approach, as the models were not necessarily (and in none of their examples) fitted within the Bayesian framework. It has been used regularly, often for process models (e.g. Gneiting et al., 2005; Zhang et al., 2009), where a rjMCMC-procedure would require substantial programming work at little perceived benefit, but also in data-poor situations in the political sciences (Montgomery et al., 2012).

Chickering and Heckerman (1997) investigate approximations of the marginal likelihood in eqn (9), such as the **Bayesian Information Criterion** (BIC, as defined in the next section; see also Appendix S1.5.3) and find them to work well for model selection, but *not* for model averaging. In contrast, Kass and Raftery (1995) state (on p. 778) that $\mathbf{e}^{\mathbf{BIC}}$ is an acceptable approximation of the Bayes factor, and hence suitable for model averaging, despite being biased even for large sample sizes. These approximations may be improved when using more complex versions of BIC (SPBIC and IBIC: Bollen et al., 2012).

The "widely applicable information criterion" **WAIC** (Watanabe 2010 and an equivalent **WBIC**: Watanabe 2013) are motivated and actually analytically derived in a Bayesian framework (Gelman et al., 2014). With an uninformative prior, it can be seen as a variation of AIC (see next section). The WAIC is computed, for each model, from two terms (Gelman et al., 2014): (1) the log pointwise predicted density (lppd) across the posterior simulations for each of the n predicted values, defined as $\text{lppd} = \log \prod_{i=1}^{n} p_{\text{posterior}}(y_i)$; and (2) a bias-correction term $p_{\text{WAIC}} = \sum_{i=1}^{n} \text{var}(\log(p(y_i|\theta_s)))$, where var is the *sample* variance over all S samples of the posterior distributions of parameters θ. The WAIC is then defined as

$\text{WAIC} = -2\,\text{lppd} + 2\,p_{\text{WAIC}}$. In other words, the WAIC is the likelihood of observing the data under the posterior parameter distributions, corrected by a penalty of model complexity proportional to the variance of these likelihoods across the MCMC samples. Model weights are computed from WAIC analogously to equation 11 below.

3.2 Information-theoretic model weights

In the *information-theoretic* perspective, models closer to the data, as measured by the Kullback-Leibler divergence, should receive more weight than those further away.

There are several approximations of the KL-divergence, most famously Akaike's Information Criterion (AIC: Akaike, 1973; Burnham and Anderson, 2002). AIC and related indices can be computed only for likelihood-based models with known number of parameters (p_m), restricting the information-theoretic approach to GLM-like models (incl. GAM):

$$\text{AIC}_m = -2\ell_m + 2p_m \quad \text{and} \quad w_m = \frac{e^{-0.5(\text{AIC}_m - \text{AIC}_{\min})}}{\sum_{i \in \mathcal{M}} e^{-0.5(\text{AIC}_i - \text{AIC}_{\min})}}, \tag{11}$$

where ℓ_m is the log-likelihood of model m.

In the ecological literature, AIC (and its sample-size corrected version AICc, and its adaptations to quasi-likelihood models such as QIC: Pan 2001; Claeskens and Hjort 2008) is by far the most common approach to determine model weights (for recent examples see, e.g., Dwyer et al., 2014; Rovai et al., 2015), despite the fact that the reasoning behin this choice is not entirely clear. **AIC-weights** (eqn 11) have been interpreted as Bayesian model probabilities (Burnham and Anderson 2002, p. 75; Link and Barker 2006), assuming a specific, model complexity and sample size-dependent, "savvy prior" (Burnham and Anderson 2002, p. 302; see also Hooten and Hobbs 2015, p. 16, for reformulation as regularisation prior). An alternative interpretation is the proportion of times a model would be chosen as the best model under repeated sampling (Hobbs and Hilborn, 2006), but such an interpretation is contentious (Richards, 2005; Bolker, 2008; Claeskens and Hjort, 2008). In an anecdotal comparison, Burnham and Anderson (2002, p. 178) showed that AIC-weights are substantially different from **bootstrapped model weights**. The latter were proposed by Buckland et al. (1997) and represent the proportion of bootstraps a model is performing best in terms of AIC: see case study 1 below. In simulations, AIC-weights did not reliably identify the model with the known lowest KL-divergence or prediction error (Richards, 2005; Richards et al., 2011). Instead, **Mallows' model averaging** (MMA) has been shown to yield the lowest mean squared error for *linear* models (Hansen, 2007; Schomaker et al., 2010). Mallows' C_p penalises model complexity equivalent to $-2\ell_m - n + 2p_m$ (for n data points; rather than AIC's $-2\ell_m + 2p_m$, eqn 11).

Schwartz' Bayesian Information Criterion was derived to find the most probable model given the data (Schwartz, 1978; Shmueli, 2010), equivalent to having the largest Bayes factor (see previous section). **BIC** uses $\log(n)$ rather than AIC's "2" as penalisation factor for model complexity (Appendix S1.5.3). A particularly noteworthy modification of the AIC exist, where the model fit is assessed with respect to a focal predictor value, e.g. a specific age or temperature range, yielding the Focussed Information Criterion (FIC: Claeskens and Hjort 2008). We are not aware of a systematic simulation study comparing the performance of these model averaging weights, but AIC's dominance should not indicate its superiority (see also case study 1 below).

The weighting procedure can additionally be wrapped into a cross-validation and model pre-selection, which leads to the ARMS-procedure (**Adaptive Regression by Mixing with model Screening**: Yang, 2001; Yuan and Yang, 2005; Yuan and Ghosh, 2008). We shall not present details on ARMS here (for cross-validation see next section), because we regard model pre-selection as an unresolved issue (see section 5.3).

3.3 Tactical approaches to computing model weights

Methods covered in this section share the "tactical" goal of choosing weights to optimise prediction (e.g. reduce prediction error), without a specific reference to a statistical theory such as Bayesian inference or information theory.

The most straightforward approach in this area is to make the averaging weight dependent on an estimate of the predictive error of each model, usually obtained by cross-validation. **Cross-validation** approximates a model's predictive performance on new data by predicting to a hold-out part of the data (typically between 5 and 20 folds, down to **leave-one-out cross-validation**, which omits each single data point in turn). The fit to the hold-out can be quantified in different ways. If the data can be reasonably well described by a specific distribution with log-likelihood function ℓ (even if the model algorithm itself is non-parametric), the log-likelihood of the data in the k folds can be computed and summed (van der Laan et al., 2004; Wood, 2015, p. 36):

$$\ell^m_{CV} = \sum_{i=1}^{k} \ell(y_{[i]}|\hat{\theta}^m_{y_{[-i]}}), \tag{12}$$

where the index $[-i]$ indicates that the data $y_{[i]}$ in fold i were not used for fitting model m and estimating model parameters $\hat{\theta}^m_{y_{[-i]}}$. It can be shown that leave-one-out *cross-validation log-likelihood* is asymptotically equivalent to AIC and thus KL-distance (Stone, 1977), albeit at a higher computational cost. Hence, computing model weights w^m_{CV} (Hauenstein et al., 2017):

$$w^m_{\mathrm{CV}} = \frac{e^{\ell^m_{\mathrm{CV}}}}{\sum_{i \in \mathcal{M}} e^{\ell^i_{\mathrm{CV}}}} \tag{13}$$

creates a weighting scheme very similar to AIC-weights, which implicitly penalises overfitting.

Other measures of model fit to the hold-out folds have been used, largely as *ad hoc* proxies for a likelihood function (e.g. in likelihood-free models): pseudo-R^2 (e.g Nagelkerke, 1991; Nakagawa and Schielzeth, 2013), area under the ROC-curve (AUC:

Marmion et al., 2009*a*; Ordonez and Williams, 2013; Hannemann et al., 2015), or True Skill Statistic (Diniz-Filho et al., 2009; Garcia et al., 2012; Engler et al., 2013; Meller et al., 2014). In these cases, weights were computed by substituting ℓ_{CV} in eqn (13) by the respective measure, or given a value of $1/S$ for a somewhat arbitrarily defined subset of S (out of M) models, e.g. those above an arbitrary threshold considered minimal satisfactory performance (Crossman and Bass, 2008; Crimmins et al., 2013; Ordonez and Williams, 2013).

Largely ignored by the ecological literature are two other non-parametric approaches to compute model weights: *stacking* and *jackknife model averaging* (see Appendix S1.4 for discussion of averaging *within* machine-learning algorithms). Both are cross-validation based, but unlike simple cross-validation weights, which are based on the performace of each contributing model on hold-out data, stacking and jacknife model averaging explicitly optimise weights to reduce the *error of the average* on hold-out data.

Stacking (Wolpert, 1992; Smyth and Wolpert, 1998; Ting and Witten, 1999) finds the optimised model weights to reduce prediction error (or maximise likelihood) on a test hold-out of size H. This is, for RMSE and likelihood, respectively:

$$\underset{w_m}{\arg\min}\left\{\sqrt{\frac{1}{H}\sum_{i=1}^{H}\left(y_{[i]}-\sum_{m=1}^{M}w_m\hat{f}\left(X_i\left|\hat{\theta}^m_{[-i]}\right.\right)\right)^2}\right\}$$

(Hastie et al., 2009) and

$$\underset{w_m}{\arg\max}\left\{\ell\left(y_{[i]}\left|\sum_{m=1}^{M}w_m\hat{f}\left(X_i\left|\hat{\theta}^m_{[-i]}\right.\right)\right.\right)\right\},$$

where $\hat{f}(X_i|\hat{\theta}^m_{[-i]})$ is the prediction of model m, fitted without using data i, to data i. This procedure is repeated many times, each time yielding a vector of optimised model weights, w_m, which are then averaged across repetitions and rescaled to sum to 1. Yao et al. (2017) extend this approach also to Bayesian models to provide a clear prediction-error minimising goal. Smyth and Wolpert (1998) and Clarke (2003) report stacking to generally outperform the cross-validation approach from two paragraphs earlier, and Bayesian model averaging, respectively (see also the case studies in section 4 and Appendix S5).

In **Jackknife Model Averaging** (JMA: Hansen and Racine, 2012), each data point is omitted in turn from fitting and then predicted to (thus actually a leave-one-out cross-validation rather than a "jackknife"). Then, weights are optimised so as to minimise RMSE (or maximise likelihood) between the observed and the fitted value across all N "jackknife" samples. The optimisation function is the same as for stacking, except that $H = N$. Thus, in stacking, weights are optimised once for each run, while for the jackknife only one optimisation over all N leave-one-out-cross-validations is required (further details and examples with R-code are given in Appendix S1.5.6).

The forecasting (i.e. time-predictions) literature (reviewed in Armstrong, 2001; Stock and Watson, 2001; Timmermann, 2006) offers two further approaches. Bates and Granger (1969)'s **minimal variance** approach attributes more weight to models with low-variance predictions. More precisely, it uses the inverse of the variance-covariance matrix of predictions, $\mathbf{\Sigma}^{-1}$, to compute model weights. In the multi-model generalisation (Newbold and Granger, 1974) the weights-vector w is calculated as:

$$w_{\text{minimal variance}} = (\mathbf{1}'\mathbf{\Sigma}^{-1}\mathbf{1})^{-1}\mathbf{1}\mathbf{\Sigma}^{-1}, \tag{14}$$

where $\mathbf{1}$ is an M-length vector of ones. This is the analytical solution of eqn 5, assuming no bias and ignoring the problem that weights are random variates, under the weights-sum-to-one constraint. Equation 14 does not ensure all-positive weights, nor is it obvious how to estimate $\mathbf{\Sigma}$. One option (used in our case studies) is to base $\mathbf{\Sigma}$ on the deviation from a prediction to test data in lieu of measure of past performance (following recommendation of Bates and Granger, 1969).

Finally, Garthwaite and Mubwandarikwa (2010) devised a rarely used method, called the “**cos-squared weighting** scheme”, designed to adjust for correlation in predictions by different models. It was motivated by (i) giving lower weight to models highly correlated with others (thereby reducing the prediction variance contributed through covariances in eqn 5), (ii) division of weights when a new, near-identical model prediction is added to the set, and (iii) reducing all weights when more models are added to the set. Weights are computed as proportional to the amount of rotation the predictions would require to make them orthogonal in prediction space, hence the trigonometric name of their approach.

Modelling model weights

So far, weights were always constant. However, one might also consider making weights dependent on other variables. This approach, which we term “model-based model combinations” (**MBMC**, also called “superensemble modelling”) was first proposed by Granger and Ramanathan (1984). Here a statistical model f is used to combine the predictions from different models, as if they were predictors in a regression: $\widetilde{Y} \sim f(\widehat{Y}_1, \widehat{Y}_2, \ldots, \widehat{Y}_m)$ (see Fig. 4 left). The regression-type model f can be of any type, such as a linear model or a neural network. We call this regression the “supra-model” in order to distinguish between different modelling levels.

A very simple supra-model would compute the **median of predictions** for each point $\mathbf{X}_i$(e.g. Marmion et al., 2009*a*). Different models are used in the “average” without requiring any additional parameter estimation. Median predictions imply varying weights, as the one or two models considered for computing the median may change between different $\mathbf{X}_i$.

An ideal model combination could switch, or gently transition, between models (such as manually constructed by Crisci et al., 2017). Since the predictions are combined more or less freely in model-based model combinations to yield the best possible fit to the observed data, MBMC should be superior to any constant-weight-per-model approach (see Fig. 4 right), as was indeed found by Diks and Vrugt (2010). This advantage comes with a severe drawback: a high proclivity to overfitting, as we fit the same data twice (once to each model, then again to their prediction regression).

[Fig. 4 approximately here.]

This does not seem to be recognised as a problem (despite being a key message of Hastie et al., 2009), as all studies we found incorrectly cross-validate the supra-model only, not the *entire* workflow (if at all; e.g. Krishnamurti et al., 1999; Thomson et al., 2006; Diks and Vrugt, 2010; Breiner et al., 2015; Romero et al., 2016). To correctly cross-validate MBMCs, one has to produce hold-outs *before* fitting the contributing models, and evaluate the MBMC prediction on this hold-out (Fig. 4, Appendix S5.9 and case studies).

Note that supra-models may differ substantially in their ability to harness the contributing models. As it is a yet fairly unexplored field in model averaging, analysts are advised to try different supra-model types (Fig. 4).

3.4 Equal weights

Last, we discuss the most trivial weighting scheme: in many fields of science (climate modelling, economics, political sciences), model averaging proceeds with giving the structurally different models equal weight, i.e. $1/M$ (e.g. Johnson and Bowler, 2009; Knutti et al., 2010; Graefe et al., 2014; Rougier, 2016). In ecology, studies analysing species distributions reported equal weights to be a very good choice when assessed using cross-validation (Crossman and Bass, 2008; Marmion et al., 2009*a*; Rapacciuolo et al., 2012), but no better than the single models on validation with independent data (Crimmins et al., 2013). Equal weights may serve as a reference approach to see whether estimating weights reduces prediction error for this specific set of models. In that sense, we may argue, all the above weight estimation approaches only serve to separate the wheat from the chaff; once a set of reasonable models has been identified, equal weights are apparently a good approach.

4 Case studies

All methods discussed above can be applied to simple regression models, while some explicitly rely on a model's likelihood and can thus not be used for non-parametric approaches. We therefore devised two case studies, the first being a rather simple example to illustrate the use of all methods in Table 1, and the second a more complicated species distribution case study based on a reduced set of methods. Note that we do not include adaptive regression by mixing with model screening (ARMS: Yang, 2001) because its more sophisticated variations (Yuan and Yang, 2005) are not readily implemented in R, and the basic ARMS is barely different from AIC-model averaging for a preselected set of models.

4.1 Case study 1: Simulation with Gaussian response, many models and few data points

In this first, simulation-based case study, we explore the variability of model-averaging approaches in the common case where several partially nested models are fit (see Data S1 for details and code). The simulation was set up so that several of the fitted models have similar support as explanations for the data. This was achieved by generating the response differently in each of two groups (using similar, but not identical predictors). We simulated 70 data points with 4 predictors yielding $2^4 = 16$ candidate models, and another 70 data points for validation. We computed model weights in 19 different ways (Table 1) and compared the prediction error of weighted averages as well as the individual models to the validation data points. Simulation and analyses were repeated 100 times.

Two results emerged from this simulation that are worth reporting. First, prediction error (quantified as RMSE) was similar across the 19 weight-computing approaches, with a few noticeably poor exceptions (the two MBMC approaches, minimal variance and the cos-squared scheme: Fig. 6), and most were no better than those of the best nine single model predictions. Second, most averaging approaches gave some weight ($w > 0.01$) to ten or more models (Table 2), despite models being overlapping and partially nested, so that we have actually only five (more or less) independent models (those containing only one predictor: m2, m3, m5, m9 and intercept-only m1). In real data sets, such spreading of weight is the result of data sparseness or extreme noise, making important effects stand out less; indeed, half of our candidate models are not hugely different, i.e. within ΔAIC < 4.

[Figure 6 approximately here.]

[Table 2 approximately here.]

4.2 Case study 2: Real species presence-absence data, many data points and a moderate number of predictors

In the second case study, we use data on the real distribution of short-finned eel (*Anguilla australis*) in New Zealand (from Elith et al., 2008). The data are provided in the R-package dismo, already split into a 1000-rows training and a 500-rows test data set, and featuring 10 predictors. We ran four different model types (GAM, Random Forest - rF, artificial neural network - ANN, support vector machine - SVM) using all 10 predictors, along with two variations of the GLM (best models selected by AIC and BIC from the full model containing the 10 predictors, relevant quadratic terms and all first-order interactions). For details see Data S1.

The number of averaging approaches that can be used to compute model weights is smaller than in the previous case study, as three of the six models do not report a likelihood or the number of parameters, precluding the use of rjMCMC, Bayes factor, (W)AIC, BIC, and Mallows' Cp. Because we do not know the underlying data-generating model, we evaluate the models on the randomly pre-selected test data provided.

[Table 3 approximately here.]

One interesting result is that model averaging was effectively a model selection tool in several cases (Table 3). Stacking, bootstrapping, JMA, and to a lesser degree minimal variance, BMA-EM and the model-based model combinations yielded non-zero weights for only 1 (or 2) models. Apparently, these approaches yielded sub-optimal model weights, as these "model selection"-outcomes of model averaging fared worse than those that kept all models in the set (equal weight, leave-one-out and cos-squared).

Secondly, the best two model averaging algorithms in this case study, apart from the median where varying weights are used, identified an approximately equal weighting as optimal strategy. That is somewhat surprising, given that SVM performed relatively poorly (and was excluded by BMA-EM, but favoured by cos-squared as a more independent contribution). The likely reason of high weights for the poor SVM is that averaging-in less correlated predictions reduces covariances in eqn (5).

The good performance of the median in both case studies suggests that using the central value of *each prediction*, rather than give constant weights to the model itself, may be even more effective in reducing variance and thus prediction error. Further research is needed to clarify if this principle is indeed valid across many applications.

5 Recommendations

In this review, we have firstly explained the mechanisms by which model averaging can improve model predictions, and secondly, we have discussed the large diversity of methods that are available to compute averaging weights. While our general results and outlook on this field are positive, in the sense that model averaging is often useful, the complexity of the topic prevents us from providing final answers about the best approach for ecologists. Surprisingly many issues seem to be statistically unresolved, or addressed by quick-fixes and even fundamental questions remain open, which we will discuss next. It is unsatisfactory to see the large variance in weights and performance of the different averaging approaches in our case studies, but also the literature provides too few comparisons of model weights to provide robust advice. In general, our recommendations are thus guided by reducing harm, rather than suggesting an optimal solution.

5.1 Averaged prediction should be accompanied by uncertainty estimates

Just like any other statistical approach, model averaging can be used wrongly. Focussing entirely on the predictions, rather than their uncertainty, can be misleading, as Knutti et al. (2010) showed for combining precipitation predictions: spatial heterogeneity cancelled out across models, giving the erroneous impression of little change when in fact all models predict large changes (albeit in different regions). Similarly, King et al. (2008) found that averaging parameters from two competing models led to no effect of two hypothesised impacts, although in both models a (different) driver was very influential. We thus strongly encourage including at least model-averaged confidence intervals alongside any prediction, possibly in addition to the individual model predictions, to prevent erroneous interpretation of averaged predictions. Also, more attention should be paid to the full model. It has many desirable properties (unbiased parameter estimates, very good coverage), but suffers from violation of the parsimony principle ("Occam's razor") and requires more consideration in which form covariates should be fit. Its larger prediction error, compared to the over-optimistic single-best partial model, is the reason for correct confidence intervals.

5.2 Dependencies among model predictions should be addressed

Statistical models, which aim to describe the data to which they are fitted, will often have correlated parameters and fits; process models may overlap in modelled processes. Having highly similar models in the model set will inflate the cumulative weight given to them (as illustrated in Appendix S1.6) . One way to handle inflation of weights by highly-related models is to assign prior model probabilities in a Bayesian framework. Another approach would be to pre-select models of different types (see next point). Alternatively, the cos-square scheme of Garthwaite and Mubwandarikwa (2010) uses the correlation matrix of model projections to appropriately change weights of correlated models. Of the weighting schemes considered here, it is the only approach doing so, but it should be noted that the performance of this approach in our case study was rather poor (Fig. 6, Tables 2 and 3).

5.3 Validation-based weighting or validation-based pre-selection of models

Madigan and Raftery (1994), Draper (1995), Burnham and Anderson (2002) and more recently Yuan and Yang (2005) and Ghosh and Yuan (2009), have argued that only “good” models should be averaged. Different ways of combining model averaging with a model screening step have been proposed (Augustin et al., 2005; Yuan and Yang, 2005; Ghosh and Yuan, 2009), in which model selection precedes averaging (pre-selection). This will happen implicitly, and in a single step, if any of the model weight algorithms discussed above attributes a weight of effectively zero to a model, as happened in case study 2. How prevalent this effect is in real world studies is unclear, as weights are rarely reported.

In contrast, some studies select models *after* the predictions are made (e.g. Thuiller, 2004; Forester et al., 2013). These studies have averaged models which predict in the same direction (along the “consensus axis”: Grenouillet et al. 2010), which are the best 50% in the set (Marmion et al., 2009*a*), or however many one should combine to minimise prediction error. Such approaches necessitate addressing the challenge of using data twice (Lauzeral et al., 2015). Post-selection reduces the ability of “dissenting voices” (i.e. less correlated predictions) to reduce prediction error and instead reinforce the trend of the model type most represented in the set. As a consequence, their uncertainty estimation will be overly optimistic. We do not advocate their use.

We suggest to employ **validation-based methods of model averaging** rather than relying on model-based estimates of error. That is, we recommend (leave-one out) cross-validation and stacking rather than AIC (in line with recommendations of Hooten and Hobbs, 2015). Using (semi-)independent test data gives us some capacity to estimate predictive bias. In such a setting, it may be less relevant whether models are pre-selected by validation-based estimates of error and then averaged with equal weights or weighted by validation-based estimates of error without pre-selection. For this to work, however, it is crucial that (cross)-validation strategies are designed to ensure independence of the validation data, which is a non-trivial problem in many practical ecological applications (Roberts et al., 2017).

5.4 Process models are no different

In fishery science, averaging process models is relatively common (Brodziak and Piner, 2010), as it is in weather and climate science (Krishnamurti et al., 1999; Knutti et al., 2010; Bauer et al., 2015). There are at least two connected challenges such enterprises face: validation and weighting. Often process models are tuned/calibrated on all sets of data available, in the sensible attempt to describe all relevant processes in the best possible way. That means, however, that no independent validation data are available, so that we cannot use the prediction accuracy of different models to compute model weights. Consequently, all models receive the same weight (e.g. in IPCC reports, or for economic models), or some reasonable but statistically ad-hoc construction of weights is employed (e.g. Giorgi and Mearns, 2002). In recent years, hind-casting has gained in popularity, i.e. evaluating models by predicting to past data. This will only be a useful approach if historic data were not already used to derive or tune model parameters, and if hindcasting success is related to prediction success (which it need not be, if processes or drivers change).

Cross-validation is often infeasible for large models, as run-times are prohibitively long. However, the greatest obstacle to averaging process models is the absence of truly equivalent alternative models, which predict the same state variable. Fishery science is one of the few areas of ecology in which commensurable models exist and are being averaged in a variety of ways (e.g. Stanley and Burnham, 1998; Brodziak and Legault, 2005; Brandon and Wade, 2006; Katsanevakis, 2006; Hill et al., 2007; Katsanevakis and Maravelias, 2008; Jiao et al., 2009; Hollowed et al., 2009; Brodziak and Piner, 2010). Carbon and biomass assessments are also moving in that direction (Hanson et al., 2004; Butler et al., 2009; Wang et al., 2009; Picard et al., 2012). These fields could profit from exploring averaging methods such as minimal variance and cos-squared, which do not require cross-validation and may perform better than either equal weights or BMA-EM, and probably better than MBMC's potentially overfitted supra-models.

Finally, irrespective of the approach chosen, model averaging studies should report model weights, and predictions should be accompanied by estimates of prediction uncertainty.

5.5 Overall conclusion and recommendations

In conclusion, we find that:

1. Model averaging may, but need not necessarily reduce prediction errors. Model averaging benefits generally increase with i) decreasing covariance of the individual model predictions, and ii) decreasing mean bias of the contributing models. Moreover, iii) while estimating model weights allows reducing the weight of poor models, this comes at the expense of introducing additional variance in the average, reducing the benefits of model averaging.

2. There are currently no generally reliable analytical methods to calculate frequentist confidence intervals (or p-values) on model-averaged predictions. Non-parametric methods, however, such as cross-validation remain reliable for estimating predictive errors, and should therefore be preferred for quantifying predictive uncertainties of model averages. Bayesian credible intervals are in principle valid as well, if the typical assumption for Bayesian model selection, that the true model is among the candidates, is met.

3. From general considerations, we believe that non-parametric methods that directly target predictive error (e.g. cross-validation or stacking) are a robust and straightforward choice for choosing weights. Parametric methods such as AIC, BIC are faster, but may not always perform equally well. Cross-validation can be used to test if fixed or estimated weights perform better than the full or the best

model.

Acknowledgements

We are grateful to five meticulous reviewers and the editor for providing substantial and constructive feedback, which greatly improved the previous versions of this manuscript. We like to thank the German Science Foundation (DFG) for funding the workshop “Model averaging in Ecology”, held in Freiburg 2-6 March 2015 (DO 786/9-1). Part of this work was carried out during a research stay of CFD at the University of Melbourne, co-funded by the DFG (DO 786/10-1). BS is supported by the DFG (SCHR1000/6-2 and SCHR1000/8-2). DIW is supported by an Australian Research Council (ARC) Future Fellowship (grant number FT120100501). DRR is supported by the Alexander von Humboldt Foundation through the German Federal Ministry of Education and Research. GGA is supported by an ARC Discovery Early Career Research Award (project DE160100904). JE is supported by ARC’s FT0991640 and DP160101003. JJLM is supported by Australia’s National Environmental Research Program (NERP) Environmental Decisions Hub and ARC DP160101003. KB is supported by a grant from the National Science Center (DEC-2015/16/S/NZ8/00158). KG is supported by the German Federal Ministry of Education and Research (BMBF 01LL0901A). WT received funding from the European Research Council under the European Community’s FP7/2007-2013 Grant Agreement no. 281422 (TEEMBIO).

References

Akaike, H. 1973. Information theory as an extension of the maximum likelihood principle. In B. Petrov and F. Csaki, editors, 2nd International Symposium on Information Theory, page 267–281. Akademiai Kiado, Budapest.

Armstrong, J. S. 2001. Combining forecasts. In J. S. Armstrong, editor, Principles of Forecasting: A Handbook for Researchers and Practitioners, pages 417–439. Springer, New York.

Augustin, N., W. Sauerbrei, and M. Schumacher. 2005. The practical utility of incorporating model selection uncertainty into prognostic models for survival data. Statistical Modelling, **5**:95–118.

Banner, K. M. and M. D. Higgs. 2017. Considerations for assessing model averaging of regression coefficients. Ecological Applications, **28**:78–93.

Bates, J. M. and C. W. J. Granger. 1969. The combination of forecasts. Journal of the Operational Research Society, **20**:451–468.

Bauer, P., A. Thorpe, and G. Brunet. 2015. The quiet revolution of numerical weather prediction. Nature, **525**:47–55.

Berger, J. O. and L. R. Pericchi. 1996. The intrinisic Bayes factor for model selection and prediction. Journal of the American Statistical Association, **91**:109–122.

Bolker, B. M. 2008. Ecological Models and Data in R. Princeton University Press, Princeton, NJ.

Bollen, K. A., S. Ray, J. Zavisca, and J. J. Harden. 2012. A comparison of Bayes factor approximation methods including two new methods. Sociological Methods & Research, **41**:294–324.

Brandon, J. R. and P. R. Wade. 2006. Assessment of the Bering-Chukchi-Beaufort seas stock of bowhead whales using Bayesian model averaging. Journal of Cetacean Research Management, **8**:225–239.

Breiner, F. T., A. Guisan, A. Bergamini, and M. P. Nobis. 2015. Overcoming limitations of modelling rare species by using ensembles of small models. Methods in Ecology and Evolution, **6**:1210–1218.

Brodziak, J. and C. M. Legault. 2005. Model averaging to estimate rebuilding targets for overfished stocks. Canadian Journal of Fisheries and Aquatic Sciences, **62**:544–562.

Brodziak, J. and K. Piner. 2010. Model averaging and probable status of North Pacific striped marlin, *Tetrapturus audax*. Canadian Journal of Fisheries and Aquatic Sciences, **67**:793–805.

Brook, B. W. and C. J. A. Bradshaw. 2006. Strength of evidence for density dependence in abundance time series of 1198 species. Ecology, **87**:1445–1451.

Buckland, S. T., K. P. Burnham, and N. H. Augustin. 1997. Model selection: an integral part of inference. Biometrics, **53**:603–618.

Burnham, K. P. and D. R. Anderson. 2002. Model Selection and Multi-Model Inference: a Practical Information-Theoretical Approach. Springer, Berlin, 2nd edition.

Burnham, K. P., D. R. Anderson, and K. P. Huyvaert. 2011. AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. Behavioral Ecology and Sociobiology, **65**:23–35.

Butler, A., R. M. Doherty, and G. Marion. 2009. Model averaging to combine simulations of future global vegetation carbon stocks. Environmetrics, **20**:791–811.

Cade, B. 2015. Model averaging and muddled multimodal inferences. Ecology, **96**:2370–2382.

Cariveau, D. P., N. M. Williams, F. E. Benjamin, and R. Winfree. 2013. Response diversity to land use occurs but does not consistently stabilise ecosystem services provided by native pollinators. Ecology Letters, **16**:903–911.

Chamberlin, T. C. 1890. The method of multiple working hypotheses. Science, **15**:92–96.

Chen, X., G. Zou, and X. Zhang. 2012. Frequentist model averaging for linear mixed-effects models. Frontiers of Mathematics in China, **8**:497–515.

Chickering, D. M. and D. Heckerman. 1997. Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables. Machine Learning,

29:181–212.

Claeskens, G. and N. L. Hjort. 2008. Model Selection and Model Averaging. Cambridge University Press, Cambridge, UK.

Claeskens, G., J. R. Magnus, A. L. Vasnev, and W. Wang. 2016. The forecast combination puzzle: A simple theoretical explanation. International Journal of Forecasting, **32**:754–762.

Clarke, B. 2003. Comparing Bayes model averaging and stacking when model approximation error cannot be ignored. The Journal of Machine Learning Research, **4**:683–712.

Clemen, R. 1989. Combining forecasts: a review and annotated bibliography. International Journal of Forecasting, **5**:559–581.

Congdon, P. 2007. Model weights for model choice and averaging. Statistical Methodology, **4**:143–157.

Corani, G. and A. Mignatti. 2015. Robust Bayesian model averaging for the analysis of presence–absence data. Environmental and Ecological Statistics, **22**:513–534.

Crimmins, S. M., S. Z. Dobrowski, and A. R. Mynsberge. 2013. Evaluating ensemble forecasts of plant species distributions under climate change. Ecological Modelling, **266**:126–130.

Crisci, C., R. Terra, J. Pablo, B. Ghattas, M. Bidegain, G. Goyenola, J. José, G. Méndez, and N. Mazzeo. 2017. Multi-model approach to predict phytoplankton biomass and composition dynamics in a eutrophic shallow lake governed by extreme meteorological events. Ecological Modelling, **360**:80–93.

Crossman, N. D. and D. A. Bass. 2008. Application of common predictive habitat techniques for post-border weed risk management. Diversity and Distributions, **14**:213–224.

Dai, Q. and Y. Wang. 2011. Effect of road on the distribution of amphibians in wetland area – test with model-averaged prediction. Polish Journal of Ecology, **59**:813–821.

Dietze, M. C. 2017. Ecological Forecasting. Princeton University Press, Princeton, N.J.

Diks, C. G. H. and J. A. Vrugt. 2010. Comparison of point forecast accuracy of model averaging methods in hydrologic applications. Stochastic Environmental Research and Risk Assessment, **24**:809–820.

Diniz-Filho, J. A. F., L. Mauricio Bini, T. Fernando Rangel, R. D. Loyola, C. Hof, D. Nogués-Bravo, M. B. Araújo, L. M. Bini, and T. F. L. V. B. Rangel. 2009. Partitioning and mapping uncertainties in ensembles of forecasts of species turnover under climate change. Ecography, **32**:897–906.

Dormann, C. F., O. Schweiger, P. Arens, I. Augenstein, S. Aviron, D. Bailey, C. F. Dormann, J. Baudry, R. Billeter, R. Bugter, R. Bukácek, F. Burel, M. Cerny, R. D. Cock, G. D. Blust, R. DeFilippi, T. Diekötter, J. Dirksen, W. Durka, P. J. Edwards, M. Frenzel, R. Hamersky, F. Hendrickx, F. Herzog, S. Klotz, B. Koolstra, A. Lausch, D. L. Coeur, J. Liira, J.-P. P. Maelfait, P. Opdam, M. Roubalova, A. Schermann-Legionnet, N. Schermann, T. Schmidt, M. J. M. Smulders, M. Speelmans, P. Simova, J. Verboom, W. K. R. E. van Wingerden, and M. Zobel. 2008. Prediction uncertainty of environmental change effects on temperate european biodiversity. Ecology Letters, **11**:235–244.

Draper, D. 1995. Assessment and propagation of model uncertainty. Journal of the Royal Statistical Society B, **57**:45–97.

Dwyer, J. F., R. E. Harness, and K. Donohue. 2014. Predictive model of avian electrocution risk on overhead power lines. Conservation Biology, **28**:159–68.

Edeling, W. N., P. Cinnella, and R. P. Dwight. 2014. Predictive RANS simulations via Bayesian Model-Scenario Averaging. Journal of Computational Physics, **275**:65–91.

Elith, J., J. R. Leathwick, and T. Hastie. 2008. A working guide to boosted regression trees. Journal of Animal Ecology, **77**:802–13.

Ellison, A. M. 2004. Bayesian inference in ecology. Ecology Letters, **7**:509–520.

Engler, R., L. T. Waser, N. E. Zimmermann, M. Schaub, S. Berdos, C. Ginzler, and A. Psomas. 2013. Combining ensemble modeling and remote sensing for mapping individual tree species at high spatial resolution. Forest Ecology and Management, **310**:64–73.

Fletcher, D. and P. W. Dillingham. 2011. Model-averaged confidence intervals for factorial experiments. Computational Statistics and Data Analysis, **55**:3041—-3048.

Fletcher, D. and D. Turek. 2012. Model-averaged profile likelihood intervals. Journal of Agricultural Biological and Environmental Statistics, **17**:38–51.

Forester, B. R., E. G. DeChaine, and A. G. Bunn. 2013. Integrating ensemble species distribution modelling and statistical phylogeography to inform projections of climate change impacts on species distributions. Diversity and Distributions, **19**:1480–1495.

Freckleton, R. P. 2011. Dealing with collinearity in behavioural and ecological data: model averaging and the problems of measurement error. Behavioral Ecology and Sociobiology, **65**:91–101.

Garcia, R. A., N. D. Burgess, M. Cabeza, C. Rahbek, and M. B. Araújo. 2012. Exploring consensus in 21st century projections of climatically suitable areas for African vertebrates. Global Change Biology, **18**:1253–1269.

Garthwaite, P. H. and E. Mubwandarikwa. 2010. Selection of weights for weighted model averaging. Australian & New Zealand Journal of Statistics, **52**:363–382.

Geisser, S. 1993. Predictive Inference. CRC Press, Boca Raton, FL.

Gelman, A., J. Hwang, and A. Vehtari. 2014. Understanding predictive information criteria for Bayesian models. Statistics and Computing, **24**:997–1016.

Ghosh, D. and Z. Yuan. 2009. An improved model averaging scheme for logistic regression. Journal of Multivariate Analysis, **100**:1670–1681.

Gibbons, J. M., G. M. Cox, A. T. A. Wood, J. Craigon, S. J. Ramsden, D. Tarsitano, and N. M. J. Crout. 2008. Applying Bayesian Model Averaging to mechanistic models: An example and comparison of methods. Environmental Modelling & Software, **23**:973–985.

Gibbs, J. W. 1902. Elementary Principles in Statistical Mechanics. Charles Scribner's Sons, New York.

Giorgi, F. and L. O. Mearns. 2002. Calculation of average, uncertainty range, and reliability of regional climate changes from AOGCM simulations via the "reliability ensemble averaging" (REA) method. Journal of Climate, **15**:1141–1158.

Gneiting, T., A. E. Raftery, A. H. Westveld, and T. Goldman. 2005. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. Monthly Weather Review, **133**:1098–1118.

Graefe, A., J. S. Armstrong, R. J. Jones, and A. G. Cuzan. 2014. Combining forecasts: An application to elections. International Journal of Forecasting, **30**:43–54.

Graefe, A., H. Küchenhoff, V. Stierle, and B. Riedl. 2015. Limitations of Ensemble Bayesian Model Averaging for forecasting social science problems weights. International Journal of Forecasting, **31**:943–951.

Granger, C. W. J. and R. Ramanathan. 1984. Improved methods of combining forecasts. Journal of Forecasting, **3**:194–204.

Green, P. J. P. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika, **82**:711–732.

Grenouillet, G., L. Buisson, N. Casajus, and S. Lek. 2010. Ensemble modelling of species distribution: the effects of geographical and environmental ranges. Ecography, **34**:9–17.

Grueber, C. E., S. Nakagawa, R. J. Laws, and I. G. Jamieson. 2011. Multimodel inference in ecology and evolution: challenges and solutions. Journal of Evolutionary Biology,

24:699–711.

Hannemann, H., K. J. Willis, and M. Macias-Fauria. 2015. The devil is in the detail: unstable response functions in species distribution models challenge bulk ensemble. Global Ecology and Biogeography, **25**:26–35.

Hansen, B. E. 2007. Least squares model averaging. Econometrica, **75**:1175–1189.

Hansen, B. E. and J. S. Racine. 2012. Jackknife model averaging. Journal of Econometrics, **167**:38–46.

Hanson, P. J., J. S. Amthor, S. D. Wullschleger, K. B. Wilson, R. F. Grant, A. Hartley, D. Hui, J. E. R. Hunt, D. W. Johnson, J. S. Kimball, A. W. King, Y. Luo, S. G. McNulty, G. Sun, P. E. Thornton, S. Wang, M. Williams, D. D. Baldocchi, and R. M. Cushman. 2004. Oak forest carbon and water simulations: Model intercomparisons and evaluations against independent data. Ecological Monographs, **74**:443–489.

Harrell, F. E. 2001. Regression Modeling Strategies - with Applications to Linear Models, Logistic Regression, and Survival Analysis. Springer, New York.

Hartig, F., J. M. Calabrese, B. Reineking, T. Wiegand, and A. Huth. 2011. Statistical inference for stochastic simulation models - theory and application. Ecology Letters, **14**:816–827.

Hastie, T., R. J. Tibshirani, and J. H. Friedman. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, Berlin, 2nd edition.

Hauenstein, S., C. F. Dormann, and S. N. Wood. 2017. Computing AIC for black-box models using Generalised Degrees of Freedom: a comparison with cross-validation. Communications in Statistics-Simulation and Computation, **in press**:1–17.

Hill, S. L., G. M. Watters, A. E. Punt, M. K. McAllister, C. L. Quéré, and J. Turner. 2007. Model uncertainty in the ecosystem approach to fisheries. Fish and Fisheries, **8**:315–336.

Hobbs, N. T. and R. Hilborn. 2006. Alternatives to statistical hypothesis testing in ecology: a guide to self teaching. Ecological Applications, **16**:5–19.

Hoeting, J., D. Madigan, A. Raftery, and C. Volinsky. 1999. Bayesian model averaging: a tutorial. Statistical Science, **14**:382–401.

Hollowed, A. B., N. A. Bond, T. K. Wilderbuer, W. T. Stockhausen, Z. T. A'mar, R. J. Beamish, J. E. Overland, and M. J. Schirripa. 2009. A framework for modelling fish and shellfish responses to future climate change. ICES Journal of Marine Science, **66**:1584–1594.

Hooten, M. B. and N. T. Hobbs. 2015. A guide to Bayesian model selection for ecologists. Ecological Monographs, **85**:3–28.

Jiao, Y., K. Reid, and E. Smith. 2009. Model selection uncertainty and Bayesian model averaging in fisheries recruitment modeling. In R. J. Beamish and B. J. Rothschild, editors, The Future of Fisheries Science in North America, pages 505–524. Springer, New York.

Johnson, C. and N. Bowler. 2009. On the reliability and calibration of ensemble forecasts. Monthly Weather Review, **137**:1717–1720.

Johnson, J. B. and K. S. Omland. 2004. Model selection in ecology and evolution. Trends in Ecology and Evolution, **19**:101–108.

Kabaila, P., A. H. Welsh, and W. Abeysekera. 2015. Model-averaged confidence intervals. Scandinavian Journal of Statistics, **43**:35–48.

Kass, R. E. and A. E. Raftery. 1995. Bayes factors. Journal of the American Statistical Association, **90**:773–795.

Katsanevakis, S. 2006. Modelling fish growth: Model selection, multi-model inference and model selection uncertainty. Fisheries Research, **81**:229–235.

Katsanevakis, S. and C. D. Maravelias. 2008. Modelling fish growth: multi-model inference as a better alternative to a priori using von Bertalanffy equation. Fish and Fisheries, **9**:178–187.

King, R., S. P. Brooks, C. Mazzetta, S. N. Freeman, and B. J. T. Morgan. 2008. Identifying and diagnosing population declines: A Bayesian assessment of lapwings in the UK. Journal of the Royal Statistical Society C, **57**:609–632.

Knutti, R., R. Furrer, C. Tebaldi, J. Cermak, and G. A. Meehl. 2010. Challenges in combining projections from multiple climate models. Journal of Climate, **23**:2739–2758.

Krishnamurti, T. N., C. M. Kishtawal, T. E. LaRow, D. R. Bachiochi, Z. Zhang, C. E. Williford, S. Gadgil, and S. Surendran. 1999. Improved weather and seasonal climate forecasts from multimodel superensemble. Science, **285**:1548–1550.

Kuhn, M. and K. Johnson. 2013. Applied Predictive Modeling. Springer, Berlin.

Lauzeral, C., G. Grenouillet, and S. Brosse. 2015. The iterative ensemble modelling approach increases the accuracy of fish distribution models. Ecography, **38**:213–220.

Le Lay, G., R. Engler, E. Franc, and A. Guisan. 2010. Prospective sampling based on model ensembles improves the detection of rare species. Ecography, **33**:1015–1027.

Liang, H., G. Zou, A. T. K. Wan, and X. Zhang. 2011. Optimal weight choice for frequentist model average estimators. Journal of the American Statistical Association, **106**:1053–1066.

Link, W. A. and R. J. Barker. 2006. Model weights and the foundations of multimodel inference. Ecology, **87**:2626–2635.

Lukacs, P. M., K. P. Burnham, and D. R. Anderson. 2010. Model selection bias and Freedman's paradox. Annals of the Institute of Statistical Mathematics, **62**:117–125.

Madigan, D. and A. E. Raftery. 1994. Model selection and accounting for model uncertainty in graphical models using Occam's window. Journal of the American Statistical Association, **89**:1535–1546.

Marmion, M., J. Hjort, W. Thuiller, and M. Luoto. 2009*a*. Statistical consensus methods for improving predictive geomorphology maps. Computers & Geosciences,

35:615–625.

Marmion, M., M. Parviainen, M. Luoto, R. K. Heikkinen, and W. Thuiller. 2009*b*. Evaluation of consensus methods in predictive species distribution modelling. Diversity and Distributions, **15**:59–69.

Meller, L., M. Cabeza, S. Pironon, M. Barbet-Massin, L. Maiorano, D. Georges, and W. Thuiller. 2014. Ensemble distribution models in conservation prioritization: from consensus predictions to consensus reserve networks. Diversity and Distributions, **20**:309–321.

Montgomery, J. M., F. M. Hollenbach, and M. D. Ward. 2012. Improving predictions using Ensemble Bayesian Model Averaging. Political Analysis, **20**:271–291.

Mouquet, N., Y. Lagadeuc, V. Devictor, L. Doyen, A. Duputié, D. Eveillard, D. Faure, E. Garnier, O. Gimenez, P. Huneman, F. Jabot, P. Jarne, D. Joly, R. Julliard, S. Kéfi, G. J. Kergoat, S. Lavorel, L. L. Gall, L. Meslin, S. Morand, X. Morin, H. Morlon, G. Pinay, R. Pradel, F. M. Schurr, W. Thuiller, and M. Loreau. 2015. Predictive ecology in a changing world. Journal of Applied Ecology, **52**:1293–1310.

Nagelkerke, N. J. D. 1991. A note on a general definition of the coefficient of determination. Biometrika, **78**:691–692.

Nakagawa, S. and R. P. Freckleton. 2011. Model averaging, missing data and multiple imputation: a case study for behavioural ecology. Behavioral Ecology and Sociobiology, **65**:91–101.

Nakagawa, S. and H. Schielzeth. 2013. A general and simple method for obtaining R^2 from generalized linear mixed-effects models. Methods in Ecology and Evolution, **4**:133–142.

Namata, H., M. Aerts, C. Faes, and P. Teunis. 2008. Model averaging in microbial risk assessment using fractional polynomials. Risk Analysis, **28**:891–905.

Newbold, P. and C. W. J. Granger. 1974. Experience with forecasting univariate time series and the combination of forecasts. Journal of the Royal Statistical Society A, **131**:131–165.

Nguefack-Tsague, G. 2014. On optimal weighting scheme in model averaging. American Journal of Applied Mathematics and Statistics, **2**:150–156.

Ordonez, A. and J. W. Williams. 2013. Climatic and biotic velocities for woody taxa distributions over the last 16 000 years in eastern North America. Ecology Letters, **16**:773–781.

O'Hagan, A. 1995. Fractional Bayes factors for model comparison. Journal of the Royal Statistical Society B, **57**:99–138.

Pan, W. 2001. Akaike's Information Criterion in Generalized Estimating Equations. Biometrics, **57**:120–125.

Picard, N., M. Henry, F. Mortier, C. Trotta, and L. Saint-André. 2012. Using Bayesian model averaging to predict tree aboveground biomass in tropical moist forests. Forest Science, **58**:15–23.

Posada, D. and T. R. Buckley. 2004. Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. Systematic Biology, **53**:793–808.

Potempski, S. and S. Galmarini. 2009. Est modus in rebus: analytical properties of multi-model ensembles. Atmospheric Chemistry and Physics, **9**:9471–9489.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski. 2005. Using Bayesian Model Averaging to calibrate forecast ensembles. Monthly Weather Review, **133**:1155–1174.

Raftery, A. E., D. Madigan, and J. A. Hoeting. 1997. Bayesian model averaging for linear regression models. Journal of The American Statistical Association, **92**:179–191.

Rapacciuolo, G., D. B. Roy, S. Gillings, R. Fox, K. Walker, and A. Purvis. 2012. Climatic associations of British species distributions show good transferability in time but low predictive accuracy for range change. PLoS ONE, **7**:e40212.

Richards, S. A. 2005. Testing ecological theory using the information-theoretic approach: examples and cautionary results. Ecology, **86**:2805–2814.

Richards, S. A., M. J. Whittingham, and P. A. Stephens. 2011. Model selection and model averaging in behavioural ecology: the utility of the IT-AIC framework. Behavioral Ecology and Sociobiology, **65**:77–89.

Roberts, D. R., V. Bahn, S. Ciuti, M. S. Boyce, J. Elith, G. Guillera-Arroita, S. Hauenstein, J. J. Lahoz-Monfort, B. Schröder, W. Thuiller, D. I. Warton, B. A. Wintle, F. Hartig, and C. F. Dormann. 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography, **40**:913–929.

Romero, D., J. Olivero, J. C. Brito, and R. Real. 2016. Comparison of approaches to combine species distribution models based on different sets of predictors. Ecography, **39**:561–571.

Rougier, J. 2016. Ensemble averaging and mean squared error. Journal of Climate, **29**:8865–8870.

Rovai, A. S., P. Riul, R. R. Twilley, E. Castañeda-Moya, V. H. Rivera-Monroy, A. A. Williams, M. Simard, M. Cifuentes-Jara, R. R. Lewis, S. Crooks, and et al. 2015. Scaling mangrove aboveground biomass from site-level to continental-scale. Global Ecology and Biogeography, **25**:286–298.

Schomaker, M. 2012. Shrinkage averaging estimation. Statistical Papers, **53**:1015–1034.

Schomaker, M., A. T. K. Wan, and C. Heumann. 2010. Frequentist model averaging with missing observations. Computational Statistics and Data Analysis, **54**:3336–3347.

Schwartz, G. E. 1978. Estimating the dimension of a model. Annals of Statistics, **6**:461–464.

Shen, X. and H.-C. Huang. 2006. Optimal model assessment, selection, and combination. Journal of American Statistical Association, **101**:554–568.

Shmueli, G. 2010. To explain or to predict? Statistical Science, **25**:289–310.

Smith, A. B., M. J. Santos, M. S. Koo, K. M. C. Rowe, K. C. Rowe, J. L. Patton, J. D. Perrine, S. R. Beissinger, and C. Moritz. 2013. Evaluation of species distribution models by resampling of sites surveyed a century ago by Joseph Grinnell. Ecography, **36**:1017–1031.

Smith, R. L., C. Tebaldi, D. W. Nychka, and L. O. Mearns. 2009. Bayesian modeling of uncertainty in ensembles of climate models. Journal of the American Statistical Association, **104**:97–116.

Smyth, P. and D. Wolpert. 1998. An Evaluation of Linearly Combining Density Estimators via Stacking. Technical Report No. 98-25. Information and Computer Science Department, University of California, Irvine, CA.

Solomon, S., D. Qin, M. Manning, M. Marquis, K. Averyt, M. Tignor, H. L. M. Jr., and Z. Chen, editors. 2007. IPCC: Climate Change 2007. The Physical Science Basis. Cambridge University Press.

Stanley, T. R. and K. P. Burnham. 1998. Information-theoretic model selection and model averaging for closed-population capture-recapture studies. Biometrical Journal, **40**:475–494.

Stock, J. and M. Watson. 2001. A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series. In R. Engle and H. White, editors, Cointegration, Causality, and Forecasting: A Festschrift in Honour of Clive Granger, page 1–44. Oxford University Press, Oxford, UK.

Stone, M. 1977. An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. Journal of the Royal Statistical Society B, page 44–47.

Symonds, M. R. E. and A. Moussalli. 2011. A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike's information criterion. Behavioral Ecology and Sociobiology, **65**:13–21.

Thibaut, L. M. and S. R. Connolly. 2013. Understanding diversity-stability relationships: towards a unified model of portfolio effects. Ecology Letters, **16**:140–150.

Thomson, M. C., F. J. Doblas-Reyes, S. J. Mason, R. Hagedorn, S. J. Connor, T. Phindela, A. P. Morse, and T. N. Palmer. 2006. Malaria early warnings based on seasonal climate forecasts from multi-model ensembles. Nature, **439**:576–579.

Thuiller, W. 2004. Patterns and uncertainties of species' range shifts under climate change. Global Change Biology, **10**:2020–2027.

Timmermann, A. G. 2006. Forecast combinations. In G. Elliott, C. Granger, and A. Timmermann, editors, Handbook of Economic Forecasting, pages 135–196. Elsevier, Dordrecht.

Ting, K. M. and I. H. Witten. 1999. Issues in stacked generalization. Journal of Artificial Intelligence Research, **10**:271–289.

Toni, T., D. Welch, N. Strelkowa, A. Ipsen, and M. P. H. Stumpf. 2009. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. Journal of the Royal Society Interface, **6**:187–202.

Trolle, D., J. A. Elliott, W. M. Mooij, J. H. Janse, K. Bolding, D. P. Hamilton, and E. Jeppesen. 2014. Advancing projections of phytoplankton responses to climate change through ensemble modelling. Environmental Modelling & Software, **61**:371–379.

Turek, D. and D. Fletcher. 2012. Model-averaged Wald confidence intervals. Computational Statistics & Data Analysis, **56**:2809–2815.

van der Laan, M. J., S. Dudoit, and S. Keles. 2004. Asymptotic optimality of likelihood-based cross-validation. Statistical Applications in Genetics and Molecular Biology, **3**:Article 4.

van Oijen, M., C. Reyer, F. J. Bohn, D. R. Cameron, G. Deckmyn, M. Flechsig, S. Harkonen, F. Hartig, A. Huth, A. Kiviste, P. Lasch, A. Makela, T. Mette,

F. Minunno, and W. Rammer. 2013. Bayesian calibration, comparison and averaging of six forest models, using data from Scots pine stands across Europe. Forest Ecology and Management, **289**:255–268.

Wang, Y.-P., C. M. Trudinger, and I. G. Enting. 2009. A review of applications of model–data fusion to studies of terrestrial carbon fluxes at different scales. Agricultural and Forest Meteorology, **149**:1829–1842.

Wasserman, L. 2000. Bayesian model selection and model averaging. Journal of Mathematical Psychology, **44**:92–107.

Watanabe, S. 2010. Asymptotic equivalence of Bayes cross validation and Widely Applicable Information Criterion in singular learning theory. Journal of Machine Learning Research, **11**:3571–3594.

Watanabe, S. 2013. A Widely applicable Bayesian Information Criterion. Journal of Machine Learning Research, **14**:867–897.

Weinberg, M. D. 2012. Computing the Bayes factor from a Markov chain Monte Carlo simulation of the posterior distribution. Bayesian Analysis, **7**:737–770.

Wintle, B. A., M. A. McCarthy, C. T. Volinsky, and R. P. Kavanagh. 2003. The use of Bayesian model averaging to better represent uncertainty in ecological models. Conservation Biology, **17**:1579–1590.

Wolpert, D. H. 1992. Stacked generalization. Neural Networks, **5**:241–259.

Wood, S. N. 2015. Core Statistics. Cambridge University Press, Cambridge, UK.

Yang, Y. 2001. Adaptive regression by mixing. Journal of the American Statistical Association, **96**:574–588.

Yao, Y., A. Vehtari, D. Simpson, and A. Gelman. 2017. Using stacking to average Bayesian predictive distributions. arXiv, **(stat.ME)**:1704.02030v3.

Ye, J. 1998. On measuring and correcting the effects of data mining and model selection. Journal of the American Statistical Association, **93**:120–131.

Yuan, Z. and D. Ghosh. 2008. Combining multiple biomarker models in logistic regression. Biometrics, **64**:431–439.

Yuan, Z. and Y. H. Yang. 2005. Combining linear regression models: When and how? Journal of the American Statistical Association, **100**:1202–1214.

Zhang, X., R. Srinivasan, and D. Bosch. 2009. Calibration and uncertainty analysis of the SWAT model using Genetic Algorithms and Bayesian Model Averaging. Journal of Hydrology, **374**:307–317.

Zhang, X., A. T. K. Wan, and G. Zou. 2013. Model averaging by jackknife criterion in models with dependent data. Journal of Econometrics, **174**:82–94.

Zhao, K., D. Valle, S. Popescu, X. Zhang, and B. Mallick. 2013. Hyperspectral remote sensing of plant biochemistry using Bayesian model averaging with variable and band selection. Remote Sensing of Environment, **132**:102–119.

Table 1: Approaches to model averaging, in particular to deriving model weights, their computational speed, likelihood/number of parameter requirement, as well as references to implementation in R.

Model averaging approach	speed	likelihood value $\vert p_m$ required?[1]	comments (R-package)[2]
Reversible jump MCMC	slow	yes|no	Requires individual coding of each model. (rjmcmc)
Bayes factor	slow	yes|no	Requires specification of priors. (BayesianTools, BayesVarSel)
Bayesian model averaging using expectation maximisation (BMA-EM)	moderate	yes|no	Requires validation step. (BMA, EBMAforecast)
Fit-based weights	rapid-slow	yes|yes[3]	AIC, BIC and Cp can be easily computed from fitted models (stats, MuMIn). (LOO-CV as option in MuMIn,[4] also in loo, cvTools, caret, crossval). DIC & WAIC should be implemented in a Bayesian approach for full benefit. (BayesianTools)
Adaptive regression by mixing with model screening (ARMS)	moderate	yes|yes	No up-to-date implementation. (ARMS[5])
Bootstrapped model weights	slow	no|no	(MuMIn,[4] boot, resample)
Stacking	slow	no|no	Requires validation step. (MuMIn[4])
Jackknife model averaging (JMA)	slow	no|no	Computation time increases linearly with n. (MuMIn,[4] boot, resample)
Minimal variance	rapid	no|no	Based only on predictions. (MuMIn[4])
Cos-squared	rapid	no|no	Based only on predictions. (MuMIn[4])
Model-based model combinations	moderate	no|no	Requires setting up regression-type analysis with model predictions, plus validation step.[2]
equal weight ($1/M$)	rapid	no|no	M is number of models considered.

[1] Does this method require a maximum-likelihood fit and/or number of parameters (p_m) of the model? Typically these two are linked, since maximum-likelihood approaches typically employ the GLM, which provides both information.

[2] See also Appendix for details and case studies in Data S1 for examples of implementation in R.

[3] While non-parametric models have no readily extractable number of parameters, a Generalised Degrees of Freedom-approach could be used to compute them (Ye, 1998). Similarly, but more efficiently, cross-validation can be used to estimate the effective number of parameters (Hauenstein et al., 2017).

[4] Implemented in MuMIn as part of this publication.

[5] `http://users.stat.umn.edu/~sandy/courses/8053/handouts/Aaron/ARMS/`

Table 2: Model weights (averaged across 100 repetitions) given to the 16 linear regression models of case study 1 by different weighting methods (see Table 1 for abbreviations), arranged by increasing prediction error (last column, median across replications). Only the best (m10) and the full model are shown from the 16 candidate models. LOO-CV: leave-one-out cross-validation using R^2 or RMSE as measure of model performance. For code see case study 1 in Data S1.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	RMSE
rjMCMC median	0.00	0.01	0.00	0.11	0.00	0.00	0.08	0.11	0.00	0.14	0.00	0.09	0.14	0.13	0.10	0.09	1.069
BIC	0.00	0.01	0.00	0.18	0.00	0.03	0.17	0.04	0.00	0.19	0.00	0.04	0.24	0.05	0.05	0.01	1.074
median[1]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	1.075
m10[2]	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1.076
rjMCMC weights	0.00	0.01	0.00	0.11	0.00	0.00	0.08	0.11	0.00	0.14	0.00	0.09	0.14	0.13	0.10	0.09	1.076
boot	0.00	0.01	0.00	0.15	0.00	0.04	0.17	0.03	0.00	0.16	0.00	0.08	0.22	0.04	0.07	0.03	1.076
AIC	0.00	0.00	0.00	0.13	0.00	0.02	0.13	0.08	0.00	0.14	0.00	0.08	0.18	0.09	0.09	0.05	1.077
WAIC	0.00	0.00	0.00	0.13	0.00	0.02	0.11	0.09	0.00	0.14	0.00	0.08	0.16	0.10	0.11	0.06	1.078
MMA	0.00	0.00	0.00	0.13	0.00	0.02	0.12	0.08	0.00	0.14	0.00	0.09	0.18	0.10	0.10	0.06	1.078
stacking	0.00	0.07	0.02	0.08	0.04	0.06	0.13	0.07	0.04	0.06	0.06	0.07	0.11	0.07	0.08	0.04	1.079
JMA	0.00	0.01	0.00	0.16	0.00	0.05	0.22	0.01	0.00	0.19	0.03	0.01	0.29	0.02	0.02	0.00	1.079
full[2]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1.086
BMA-EM	0.00	0.08	0.01	0.08	0.02	0.07	0.14	0.06	0.03	0.08	0.10	0.04	0.15	0.06	0.06	0.03	1.104
BayesFactor	0.07	0.06	0.06	0.07	0.06	0.06	0.06	0.06	0.06	0.06	0.07	0.06	0.06	0.06	0.06	0.06	1.109
equal weight	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	1.110
LOO-CV (R^2)	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	1.110
LOO-CV (RMSE)	0.09	0.06	0.08	0.06	0.07	0.06	0.06	0.06	0.07	0.06	0.06	0.06	0.06	0.06	0.06	0.06	1.123
MBMC (LM)[3]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	1.135
MBMC (rF)[3]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	1.181
minimal variance	−1.15	0.42	0.19	0.00	0.64	0.00	0.00	0.00	0.91	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.208
cos-squared	0.00	0.00	0.30	0.00	0.21	0.21	0.02	0.01	0.00	0.00	0.24	0.00	0.00	0.00	0.01	0.00	1.209

[1] Weights not available, as different models contribute to the median at each replication.

[2] Prediction from individual model.

[3] Weights are variable. LM and rF refer to a linear model and a Random Forest as supra-model, respectively.

Table 3: Model weights given to the six model types of case study 2 (GLM, GAM, Random Forest, artificial neural networks and support vector machine) by different weighting methods (see Table 1 for abbreviations), arranged by decreasing fit of the averaged predictions to test data, assessed as log-likelihood (ℓ) (last column). LOO-CV: leave-one-out cross-validation using R^2 or RMSE as measure of model performance. For code see case study 2 in Data S1.

Method	GLM_{AIC}	GLM_{BIC}	GAM	rF	ANN	SVM	ℓ
median [1]	(0.176)	(0.216)	(0.212)	(0.162)	(0.146)	(0.088)	−182.84
LOO-CV	0.168	0.168	0.166	0.169	0.165	0.164	−184.82
equal weight	0.167	0.167	0.167	0.167	0.167	0.167	−184.86
cos-squared	0.122	0.104	0.178	0.188	0.186	0.221	−185.02
BMA-EM	0.388	0.192	0.000	0.420	0.000	0.000	−185.24
stacking	0.000	0.000	0.000	1.000	0.000	0.000	−186.82
bootstrap	0.000	0.000	0.000	1.000	0.000	0.000	−186.83
minimal variance	0.155	0.469	−0.036	0.58	−0.026	−0.141	−188.45
MBMC (GAM) [3]	–	–	*	*	–	–	−198.23
MBMC (rF) [3]	–	–	–	–	–	–	−200.20
JMA	0.000	0.000	0.000	0.000	0.000	1.000	−214.68
MBMC (GLM) [3]	–	–	*	*	–	–	−268.52
rF [2]	0	0	0	1	0	0	−186.83
GAM [2]	0	0	1	0	0	0	−193.40
ANN [2]	0	0	0	0	1	0	−194.28
GLM_{AIC} [2]	1	0	0	0	0	0	−197.48
GLM_{BIC} [2]	0	1	0	0	0	0	−197.73
SVM [2]	0	0	0	0	0	1	−214.68

[1] Weights are proportion of times this model was actually used to compute the median value divided by two.

[2] Prediction from individual model.

[3] Weights are variable. Asterisk indicates that a model's prediction was a significant term in the supra-model. GAM, rF and GLM refer to three different types of supra-model: a generalised additive model, a Random Forest, and a generalised linear model.

Figure 1: Conceptual depiction of the contributions of error to model averaging. A) Contributing models have larger bias than variance. Then, the error of the average depends on how the bias is averaged out. It can increase or decrease compared to the best model. Adding a lot more models will not change the error, unless this reduces bias. B) Contributing models have similar bias and variance. In this case, averaging an increasing number of models can reduce the variance of the error, while the bias remains. C) Contributing models are unbiased, but have large variance. In this case (assuming covariances between models are low), an increasing number of models can, in principle, make the error arbitrarily small.

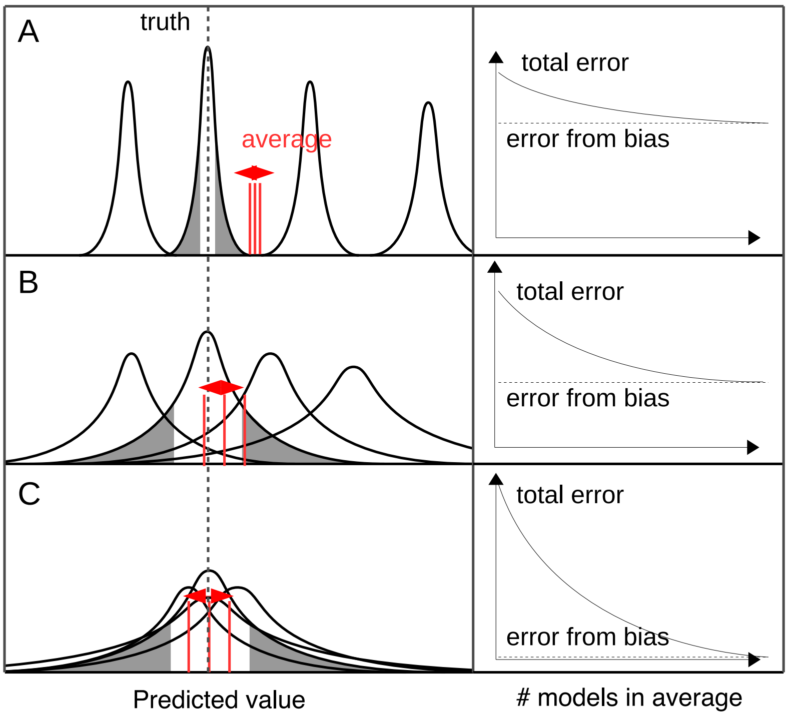
Figure 2: Conceptualised outcomes of model averaging. Sampling distributions of model predictions are depicted as stylised empty triangle on the see-saw (wider means less certain). Filled triangles represent the model predictions with unidirectionally bias (top row) or straddling truth (bottom row), and positive, no, or negative covariances among model predictions in columns. In the top row, grey shaded quadrants indicate model combinations with bias in the same direction, leading to a biased average (tilted see-saw). In the bottom row, grey shaded quadrants indicate opposite biases, which *may* lead to less biased averaged prediction, assuming optimal model weights were found. Changes in prediction covariance (columns) affect the uncertainty of the average, with negatively correlated predictions (right) yielding lowest uncertainty.

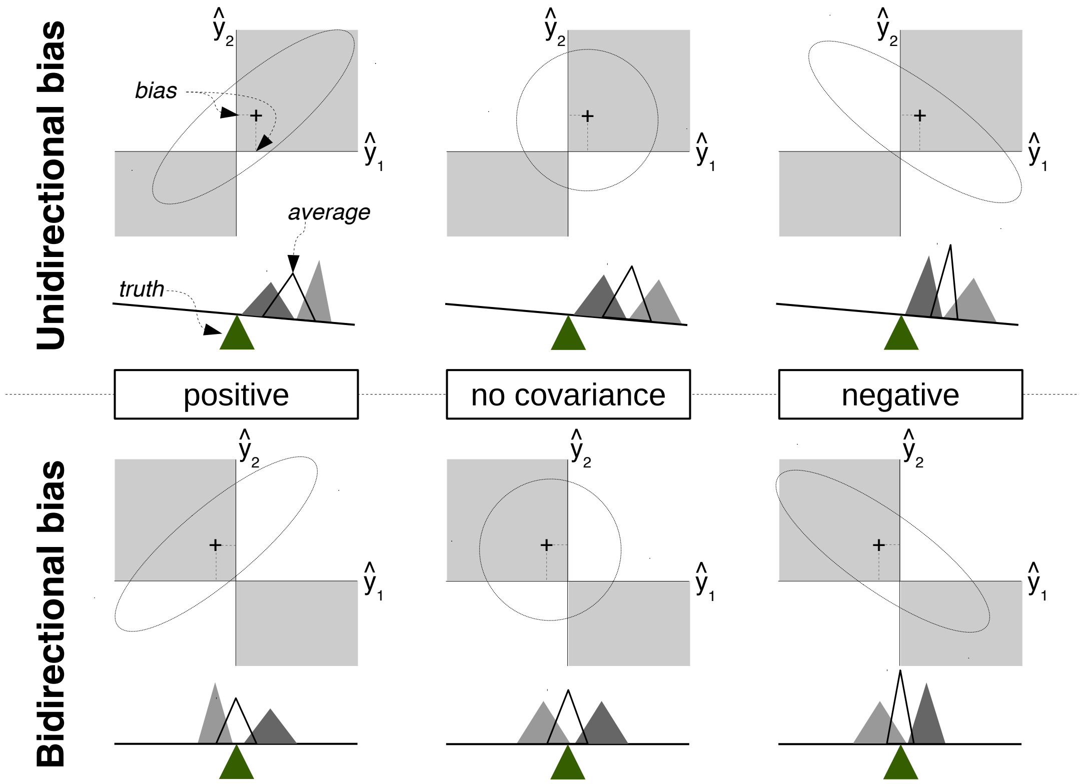
Figure 3: When averaging is optimal, in the simplest case of two models that make correlated Gaussian predictions. The models are here described by their biases (b_1, b_2, not shown), their standard deviations (σ_1, σ_2), and by the correlation (ρ) between them. Each panel shows the regions in the (σ_1, ρ) plane where model 1 is best (blue shading and contour line), model 2 is best (orange shading and contour line), and where the optimal average is best (colour gradient between blue and orange). Top row represents the case where weights are known (i.e. without error: $\sigma_w = 0$), while the second row represents exactly the same settings, but with estimated weights (with uncertainty $\sigma_w = 0.2$). Notice that when w is estimated with uncertainty, the contours marking the transition between each single model and the average move into the washed-out colours, i.e. deviate from the fixed w situation in the upper panels. These curves now represent a level set at the values $\bar{w}_1^* = 1 - \sigma_w$ (blue curve) and $\bar{w}_2^* = \sigma_w$ (orange curve). As a consequence, the area where model averaging with estimated weights is superior to the better single model decreases substantially relative to the fixed w case, and disappears completely for $\sigma_w \geq 0.5$. Formal derivations for the contours and the critical weights is given in Appendix S1.2, the interactive tool itself in Data S1. Biases are set to $b_1 = 3$ and $b_2 = 2$.

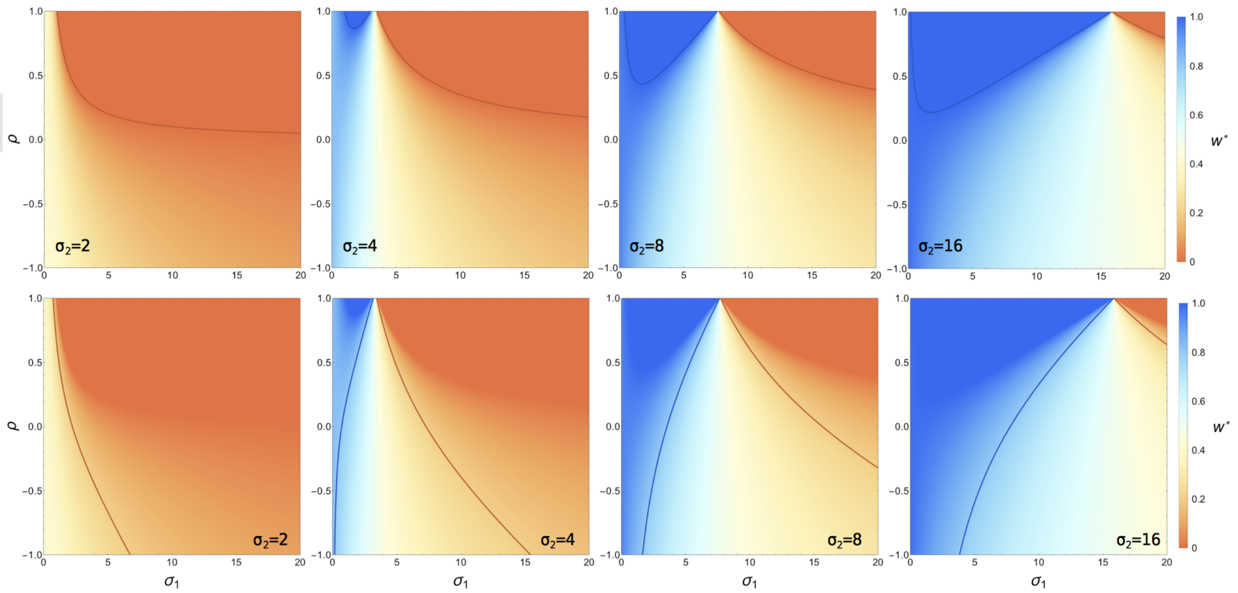
Figure 4: A simple model-based model combination example. *Left*: Three models (solid grey lines: constant, linear and quadratic) fitted separately to a data set (points, following the thin black line). Using a linear model (with quadratic terms: red) to combine the three models' fits may improve fit, even more so than the full model (green), and with narrower confidence intervals. Dotted lines indicate the weight that each model receives at each point in the linear model. Such MBMC did not necessarily improve fit, as Random Forest-based model combinations showed (blue). *Right*: Using 5-fold cross-validation around the entire workflow shows that the linear supra-model (Supra-LM) indeed improved prediction (decreased root mean squared prediction error), while the Random Forest-supra-model (Supra-rF) did not. The full model (as reference) comprised all terms present in Supra-LM, but was fitted directly.

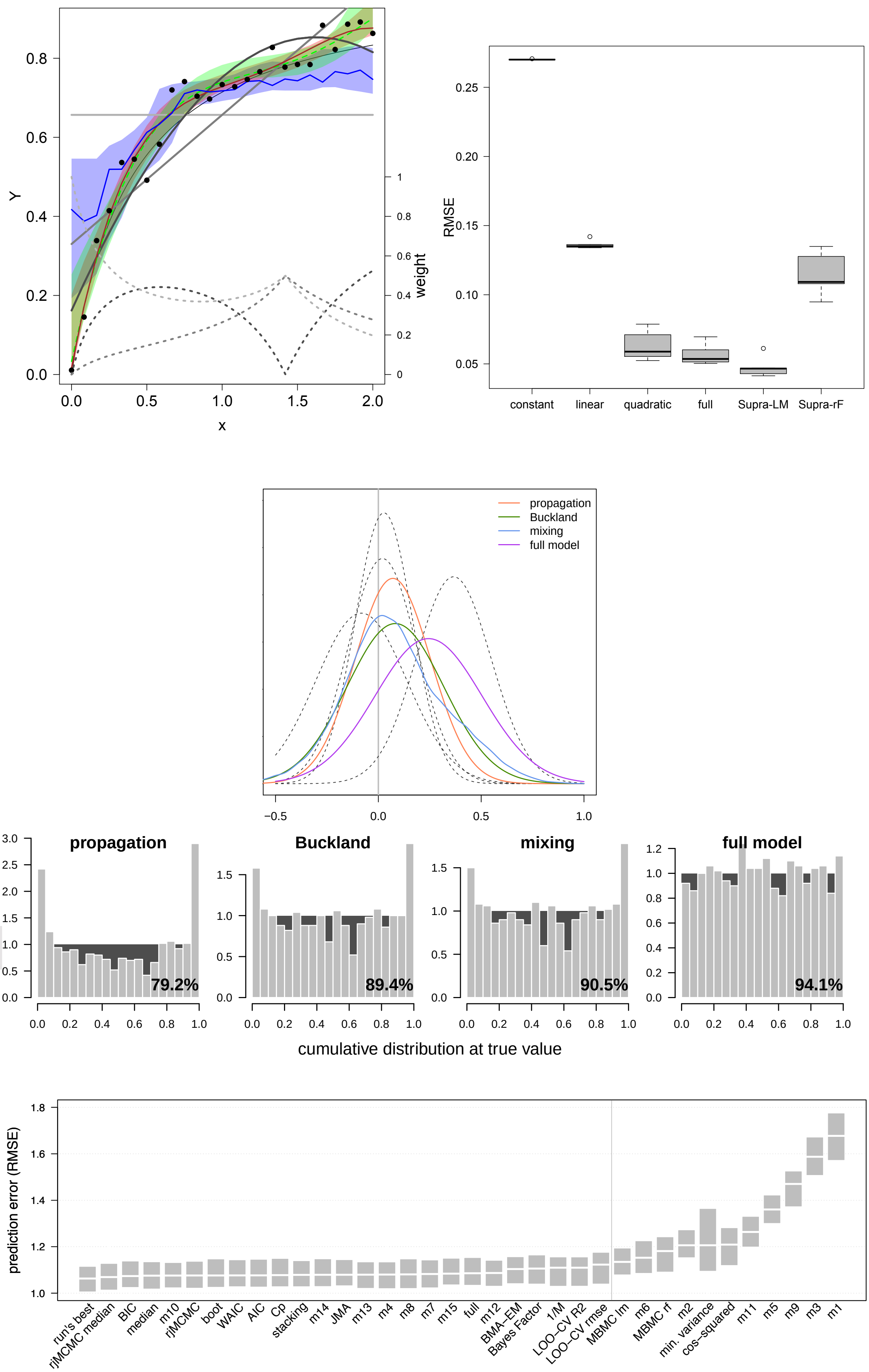
Figure 5: A comparison of different approaches to quantifying uncertainty when combining predictions from four linear models (dashed curves) with equal weights. *Top*: Estimates of predictive uncertainty in a single example run. Truth is indicated by the vertical line. Error propagation based on bootstrapped estimates for eqn (5), Buckland et al.'s correction and model mixing yield (substantially) smaller uncertainties than the full model. *Bottom*: Histograms of the cumulative density of the estimated uncertainties at the true values. The numbers display the coverage for the 95% confidence interval.

Figure 6: Prediction error of different model averaging approaches (100 repetitions) for case study 1. Box represents quartiles, white line the median. Approaches to the left of the vertical line are very similar, and no better than nine of the candidate models. See Table 1 for list of approaches, and case study 1 in Data S1 for list and fits of the individual models.


A
truth
average
total error
error from bias
B
total error
error from bias
C
total error
error from bias
Predicted value
models in average


Unidirectional bias
bias
average
truth
positive
no covariance
negative
Bidirectional bias
$\hat{y}_1$
$\hat{y}_2$


σ_2=2
σ_2=4
σ_2=8
σ_2=16
ρ
σ_1
w^*


Y
x
weight
RMSE
constant
linear
quadratic
full
Supra-LM
Supra-rF
propagation
Buckland
mixing
full model
density
79.2%
89.4%
90.5%
94.1%
cumulative distribution at true value
prediction error (RMSE)
run's best
rjMCMC median
BIC
median
m10
rjMCMC
boot
WAIC
AIC
Cp
stacking
m14
JMA
m13
m4
m8
m7
m15
full
m12
BMA-EM
Bayes Factor
1/M
LOO-CV R2
LOO-CV rmse
MBMC lm
m6
MBMC rf
m2
min. variance
cos-squared
m11
m5
m9
m3
m1