Supporting information to "Model averaging in ecology: a review of Bayesian, information-theoretical and tactical approaches"

Carsten F. Dormann Justin M. Calabrese Gurutzeta Guillera-Arroita
Eleni Matechou Volker Bahn Kamil Bartoń Colin M. Beale Simone Ciuti
Jane Elith Katharina Gerstner Jérôme Guelat Petr Keil
José J. Lahoz-Monfort Laura J. Pollock Björn Reineking David R. Roberts
Boris Schröder Wilfried Thuiller David I. Warton Brendan A. Wintle
Simon N. Wood Rafael O. Wüest Florian Hartig

April 19, 2018

Contents

Appendix S1

S1.1 Model averaging for parameter estimation

The value of a parameter common to several models could also be the target of model averaging. Since model structure affects the estimation of each parameter, one could think of averaging estimates across models to accommodate model structural uncertainty. Fundamentally, this is a problematic enterprise, as the actual meaning of a parameter changes when the model structure changes, as they are conditional on the form of the model and the other covariates included in it (see Cade, 2015; Banner and Higgs, 2017, for a thorough coverage of this argument).

We attribute the origin of this idea to one of two possible misunderstandings.

1. In meteorological process models and alike, uncertainty about initial and boundary conditions is often accommodated by running *the same model* with different setting, and deriving the average parameter by averaging the fits from each set of conditions. Since it is always the same model, no *structural* uncertainty is included. This approach is perfectly fine, but does not translate readily to models of different structure.

2. For *linear models*, predicting from a parameter-averaged model is mathematically identical to averaging predictions, but this is not the case for non-linear models. For linear models we can use the mathematical shortcut of averaging model parameters and predicting with these (Burnham and Anderson, 2002; Lukacs et al., 2010; Freckleton, 2011). Indeed, many mathematical papers and proofs are concerned entirely with linear models, despite their lack of transferability to most real-world situations (e.g. Hansen, 2007; Liang et al., 2011; Nguefack-Tsague, 2014).

   $$\tilde{y} = \frac{1}{m}\sum_{i=1}^{m} Xb_i = X\frac{\sum_{i=1}^{m} b_i}{m} \tag{S1}$$

   For non-linear models, such as GLMs with log or logit link functions $g(x)$, such coefficient averaging is *not* equivalent to prediction averaging, as

   $$\frac{1}{m}\sum_{i=1}^{m} g(Xb_i) \neq g\left(X\frac{\sum_{i=1}^{m} b_i}{m}\right) \tag{S2}$$

   for any non-linear function g (a consequence of Jensen's inequality, which also holds for concave functions). Thus in general, parameter averaging is not equivalent to prediction averaging. In practice, however, both log and logit are often sufficiently linear, making coefficient averaging an acceptable approximation.

Another problematic issue with parameter averaging is which value to assign parameters that are not in a given model. Depending on the amount of collinearity among predictors, the conditional or unconditional approach of Burnham and Anderson (2002, p. 152) may work better, but clearly not both can be right. The more common solution is to set omitted parameter estimates to 0 (the "unconditional" approach).

Cade (2015) suggests an interpretation of MA regression coefficients after employing a two-level standardisation. Banner and Higgs (2017) recommend against any interpretations and provide a graphical tool for assessment. These publications suggest that at least interpretation is challenging, in contrast to common practice especially in conservation and behavioural ecology, where averaged parameters are often seen as best scientific knowledge. Interpretation of model-averaged parameters after best-subset regressions is also flawed statistically, since the averaged parameters of the resulting model (which is often the full model) differ from the maximum likelihood estimates of the structurally identical model, which is part of the best-subset regression. So we know that the averaged parameters are statistically non-optimal: why should they then be "better" suited for interpretation than the MLE estimates for the same model structure?

Model selection inevitably inflates effect sizes of estimated parameters (i.e. distorts the ratio of estimate and its standard error), which was the motivation to develop shrinkage estimators. Parameter averaging also *de facto* acts as a shrinkage, as some models will not have a given predictor and the parameter is hence set to 0 (Lukacs et al., 2010). When averaging, this pulls estimates back towards 0, acting similar to the regularisation term in ridge or lasso regression. So, while the parameter averaging as such is conceptually problematic, it may actually work to reduce parameter estimation bias, although for a different reason than originally anticipated. Using lasso or ridge regression seems a much more direct and statistically sound way to reach this goal.

Model averaging may reduce parameter estimation bias: evidence from literature review

According to the (non-ecological) literature, parameter estimation is consistently improved by model averaging (e.g. Turkheimer et al., 2003). This is largely due to the fact that parameter averaging is dominated by *process models* averaging over different initial conditions. In this case, there is no difference in model structure, and parameters retain their meaning.

For linear models, several studies show that model averaging acts similar to shrinkage, i.e. downwards bias of model coefficients. In fact, Ghosh and Yuan (2009) make this link mathematically explicitly and show that model weights are effectively shrinkage parameters. Still, model-averaged parameters were found to regularly be closer to the 'truth' than single-model estimates, thus confirming the expectation that model averaging improves parameter estimation.

Thus, parameter averaging may work, but for the wrong reason. Also, it would be false to *assume* that model averaging is a useful procedure to reduce bias in parameter estimation. Overall, we agree with Cade (2015) that parameter averaging for regression-type models leads to inconsistent interpretations of these parameters and should be discouraged. The interpretation of *variable importance* is distorted when simply summing model weights, for the same reason: when a variable is omitted, its effect will be partially transferred onto other model parameters and hence wrongly attributed to them (Galipaud et al., 2014).

S1.2 Analytical treatment of averaging predictions from two correlated, normally distributed models (J.M.C., B.R.)

In this appendix, we consider the simple case of averaging two models whose predictions, under repeated sampling, are jointly normally distributed, and (possibly) correlated. We focus on obtaining analytical results on the conditions under which the averaged prediction is better than either of the individual models. Our set-up is a generalization of that studied by Bates and Granger (1969).

Let $\widehat{Y}_1 \sim \mathcal{N}(b_1, \sigma_1^2)$ be the prediction of model 1, $\widehat{Y}_2 \sim \mathcal{N}(b_2, \sigma_2^2)$ the prediction of model 2, and assuming the two models are jointly normally distributed, let $\operatorname{cor}(\widehat{Y}_1, \widehat{Y}_2) = \rho$. We assume that the true value of the quantity being predicted is 0, and thus the means b_1 and b_2 of each prediction measure the bias in the prediction. The mean squared prediction error of each model is thus

$$\mathrm{MSE}_1 = b_1^2 + \sigma_1^2 \tag{S3}$$

$$\mathrm{MSE}_2 = b_2^2 + \sigma_2^2. \tag{S4}$$

We define the prediction with the lowest MSE to be the best. Next, we consider a weighted average of the predictions of models 1 and 2

$$\widehat{Y}_A = w\widehat{Y}_1 + (1-w)\widehat{Y}_2, \tag{S5}$$

where $0 \leq w \leq 1$ is the weight assigned to model 1 and $1-w$ is the weight assigned to model 2. As multiplying a random variable by a constant changes its parameters but not its distribution, the weighted random variables $w\widehat{Y}_1$ and $(1-w)\widehat{Y}_2$ are distributed as $\mathcal{N}(wb_1, w^2\sigma_1^2)$ and $\mathcal{N}((1-w)b_2, (1-w)^2\sigma_2^2)$, respectively. Furthermore, the correlation between two random variables is invariant to separate scale changes of those variables, thus $\mathrm{Corr}(w\widehat{Y}_1, (1-w)\widehat{Y}_2) = \rho$. Finally, we denote the covariance in predictions between the two models as $\sigma_{12} = \rho\sigma_1\sigma_2$. The sum of two jointly distributed correlated normal random variables is also normally distributed, thus the averaged prediction is normally distributed with mean and variance

$$b_A = w(b_1 - b_2) + b_2 \tag{S6}$$

$$\sigma_A^2 = w^2\sigma_1^2 + 2(1-w)w\sigma_{12} + (1-w)^2\sigma_2^2, \tag{S7}$$

respectively, which follow from standard formulae for the moments of sums of correlated variables. The mean squared prediction error of the average is then

$$\mathrm{MSE}_A = w^2\mathrm{MSE}_1 + (1-w)^2\mathrm{MSE}_2 + 2(1-w)w(b_1b_2 + \sigma_{12}). \tag{S8}$$

We are interested in identifying the conditions under which the average is best. We therefore focus on identifying the boundaries between the region where the average has minimum MSE and the regions where either model 1 or model 2 have minimum MSE. In general, the value of w that defines the transition between regions must satisfy two conditions. Specifically, for the i^{th} single model, we seek w_i^* such that

$$w_i^* = \underset{w\in[0,1]}{\arg\min}\, \mathrm{MSE}_A(w) \tag{S9}$$

$$\mathrm{MSE}_i = \mathrm{MSE}_A(w_i^*). \tag{S10}$$

Clearly, the transition between regions must occur where the MSE of the average equals that of the focal single model. However, while this condition is necessary, it is not sufficient. To see this, consider that at any point in the region where the average is favoured that is not exactly on the boundary, it will always be possible to find a value of MSE_A that equals that of the focal single model. However, in this region it will also always be possible to find values of MSE_A that are lower than that of the single model (e.g. $\min(\mathrm{MSE}_A)$), and so this cannot be a transition point between regions. Only on the boundary between regions will $\min(\mathrm{MSE}_A)$ be the only value of the MSE of the average that equals that of the focal single model.

To find w_i^*, we therefore must first solve for the weight, w^*, that minimizes the MSE of the average. Differentiating eqn S8 with respect to w, we obtain

$$\frac{\mathrm{dMSE}_A}{\mathrm{d}w} = 2(w\mathrm{MSE}_1 - (1-w)\mathrm{MSE}_2 + (1-2w)(b_1 b_2 + \sigma_{12})). \tag{S11}$$

We then set

$$\frac{\mathrm{dMSE}_A}{\mathrm{d}w} = 0 \tag{S12}$$

and solve for w, which gives

$$w^* = \frac{1}{C}(\mathrm{MSE}_2 - b_1 b_2 - \sigma_{12}), \tag{S13}$$

where we have defined $C \equiv \mathrm{MSE}_1 + \mathrm{MSE}_2 - 2b_1 b_2 - 2\sigma_{12}$. Setting $w = w^*$ in eqn (S8), the minimum possible value of the MSE of the average for a given set of model parameters is then

$$\min(\mathrm{MSE}_A) = \frac{1}{C}(\mathrm{MSE}_1 \sigma_2^2 + b_2^2 \sigma_1^2 - (2b_1 b_2 + \sigma_{12})\sigma_{12}). \tag{S14}$$

The second condition that defines w_i^* is then satisfied when $\mathrm{MSE}_i = \min(\mathrm{MSE}_A)$. To understand the conditions under which averaging is favoured, it is useful to visualize these relationships. We thus consider the two-dimensional parameter space, (z, ρ), where z is one of the model parameters (b_1, b_2, σ_1, or σ_2), and the remaining parameters are held constant. Based on the MSE values of the individual models and of the average, the (z, ρ) plane can be partitioned into regions where model 1 is best, where model 2 is best, and where the average is best. Specifically, the contour separating the region where model 1 is best from that where the average is best, $c_1(z)$, occurs where $\mathrm{MSE}_1 = \min(\mathrm{MSE}_A)$ in the (z, ρ) plane. Thus solving the implicit equation $\mathrm{MSE}_1 = \min(\mathrm{MSE}_A)$ for ρ yields

$$c_1(z) = \frac{\mathrm{MSE}_1 - b_1 b_2}{\sigma_1 \sigma_2} \tag{S15}$$

Similarly, the contour separating the region where model 2 is best from the region where the average is best occurs when $\mathrm{MSE}_2 = \min(\mathrm{MSE}_A)$, and solving this equation for ρ gives

$$c_2(z) = \frac{\mathrm{MSE}_2 - b_1 b_2}{\sigma_1 \sigma_2}. \tag{S16}$$

To use these expressions, one chooses a particular parameter of interest for z. For example, setting $z = b_1$, one would plot $c_1(b_1)$ and $c_2(b_1)$ in the (b_1, ρ) plane, and hold the values of the remaining parameters b_2, σ_1, and σ_2 constant. Finally, upon substituting eqn (S15) and eqn (S16) separately into eqn (S13), we obtain the critical weights

$$w_1^* = 1 \tag{S17}$$

and

$$w_2^* = 0, \tag{S18}$$

respectively. Therefore the average is favoured for $0 < w < 1$ when w is fixed and known. While the critical weights are intuitively obvious in the fixed w case, the approach we have used to derive them is general and transfers to the less intuitive variable weights case that we consider next.

So far we have assumed the optimal weights are known, but in practice the weights must be estimated from the data. We now consider the case when w is estimated with error, specifically when $w \sim \mathcal{N}(\bar{w}, \sigma_w^2)$, where the true value being estimated is $\bar{w}$. In other words, we assume that the estimate of w is unbiased under repeated sampling.

We obtain the MSE of the average in this case by integrating eqn (S8) term-by-term over the distribution of w

$$t_1 = \mathrm{MSE}_1 \int_{-\infty}^{\infty} w^2 f(w; \bar{w}, \sigma_w)\,\mathrm{d}w \tag{S19}$$

$$= \mathrm{MSE}_1(\bar{w}^2 + \sigma_w^2) \tag{S20}$$

$$t_2 = \mathrm{MSE}_2 \int_{-\infty}^{\infty} (1-w)^2 f(w; \bar{w}, \sigma_w)\,\mathrm{d}w \tag{S21}$$

$$= \mathrm{MSE}_2((1-\bar{w})^2 + \sigma_w^2) \tag{S22}$$

$$t_3 = 2(b_1 b_2 + \sigma_{12}) \int_{-\infty}^{\infty} (1-w) w f(w; \bar{w}, \sigma_w)\,\mathrm{d}w \tag{S23}$$

$$= 2(b_1 b_2 + \sigma_{12})((1-\bar{w})\bar{w} - \sigma_w^2), \tag{S24}$$

where

$$f(w; \bar{w}, \sigma_w) = \frac{e^{-\frac{(w-\bar{w})^2}{2\sigma_w^2}}}{\sqrt{2\pi}\sigma_w} \tag{S25}$$

is the PDF of the normal distribution. Summing terms and simplifying, the MSE for the average when w is normally distributed is then

$$\mathrm{MSE}_A^{'} = \bar{w}^2 \mathrm{MSE}_1 + (1-\bar{w})^2 \mathrm{MSE}_2 + 2(1-\bar{w})\bar{w}(b_1 b_2 + \sigma_{12}) + C\sigma_w^2, \tag{S26}$$

where the $'$ notation is used to distinguish the variable-w MSE from the fixed-w MSE. It is apparent that eqn (S26) has the same form as eqn (S8), but with an extra term that is scaled by the sampling variance of w, σ_w^2. We note that C takes its minimum value when the covariance between models is maximal, which occurs when $\rho = 1$ given that $\sigma_{xy} = \rho\sigma_x\sigma_y$. Setting $\rho = 1$ and simplifying, we obtain

$$\min(C) = (b_1 - b_2)^2 + (\sigma_1 - \sigma_2)^2, \tag{S27}$$

which is always positive. Therefore $\mathrm{MSE}_A^{'} > \mathrm{MSE}_A$ whenever $\sigma_w > 0$, implying that estimating w with uncertainty will always increase the MSE of the average, all else being equal.

Intuitively, we expect that the increase in the MSE of the average resulting from uncertain weight estimation will narrow the range of conditions that favour averaging. To see this, consider that while the MSE of the average increases when weights must be estimated, MSE_1 and MSE_2 remain the same as before because they do not involve weight estimation. To quantify the magnitude of this effect, we proceed as before to derive the contour lines in the (z, ρ) plane that separate regions where the single models are favoured from the region where the average has the lowest MSE when w is uncertain. Differentiating eqn (S26) with respect to $\bar{w}$, setting the result equal to 0, and solving for $\bar{w}$ yields $\bar{w}^* = w^*$. This is reassuring as we have assumed unbiased estimation of w, and so the value of w that minimizes the MSE of the average in the fixed weights case will also be the value of the mean, $\bar{w}$, that minimizes the MSE in the uncertain weights case.

Next, we set $\bar{w} = \bar{w}^*$ in eqn (S26), yielding

$$\min(\mathrm{MSE}_A^{'}) = \frac{1}{C}(b_2^2\sigma_1^2 - 2b_1 b_2\sigma_{12} + \mathrm{MSE}_1\sigma_2^2 - \sigma_{12}^2 + C^2\sigma_w^2). \tag{S28}$$

We then set $\mathrm{MSE}_1 = \min(\mathrm{MSE}_A^{'})$ and $\mathrm{MSE}_2 = \min(\mathrm{MSE}_A^{'})$ and solve for ρ in each case, which gives the contour lines

$$c_1^{'}(z) = \frac{\mathrm{MSE}_1 - b_1 b_2}{\sigma_1\sigma_2} - \frac{\sigma_w(\mathrm{MSE}_1 - \mathrm{MSE}_2)}{\sigma_1\sigma_2(2\sigma_w - 1)} \tag{S29}$$

and

$$c_2^{'}(z) = \frac{\mathrm{MSE}_2 - b_1 b_2}{\sigma_1 \sigma_2} - \frac{\sigma_w(\mathrm{MSE}_2 - \mathrm{MSE}_1)}{\sigma_1 \sigma_2 (2\sigma_w - 1)}, \tag{S30}$$

respectively. Equations (S29) and (S30) have the same form as eqns (S15) and (S16), but with an extra term that is a function of σ_w and acts to pull the contours into the region where the average is favoured in the fixed weights case. This has the effect of expanding the size of the regions where the single models are favoured, while contracting the region where the average has minimum MSE.

As before, we then substitute eqns (S29) and (S30) separately into eqn (S13), yielding

$$\begin{aligned} \bar{w_1^*} &= w_1^* - \sigma_w & \text{(S31)} \\ &= 1 - \sigma_w & \text{(S32)} \\ \bar{w_2^*} &= w_2^* + \sigma_w & \text{(S33)} \\ &= \sigma_w. & \text{(S34)} \end{aligned}$$

Therefore, when the weights must be estimated, the region where the average is favoured shrinks from $0 < w < 1$ to $\sigma_w < w < 1 - \sigma_w$. This implies that, in this particular setup, the region where the average has minimum MSE disappears completely when $\sigma_w = 0.5$.

The interactive tool provided in appendix II shows these contours in the (b_1, ρ) plane (left panel) and the (σ_1, ρ) plane (right panel) on top of a density plot that is colour-coded according to the value of w^* at each point. The tool provides interactive parameter sliders so that users can explore the parameter space of the two-model set-up to understand the conditions under which averaging is advantageous. The sliders that control model parameters are placed above each panel. In addition to sliders for the parameters of each of the two models, there is also a slider that controls the standard deviation, σ_w, of the sampling distribution of w. The default value is 0, which yields the fixed w case described above, while $\sigma_w > 0$ allows one to explore the effects of uncertain estimation of w on the relative advantages of averaging. Finally, a slider controlling the resolution, R, of the underlying density plot is located below each panel. This resolution slider allows the user to control the trade-off between the quality (i.e., smoothness) of the density plot and the amount of time it takes to render the plot after changing the values of model parameters. Selecting a lower resolution will facilitate quickly exploring the parameter space. However, if the colour gradient appears “wavy”, rough, or displays other artifacts, increasing the resolution will typically produce a better result. The interactive tool requires the freely available Wolfram CDF Player, which can be downloaded at: `https://www.wolfram.com/cdf-player/`.

Several key insights emerge from this analysis. First, there is no guarantee that averaging will produce the best solution. Indeed, it is easy to find large regions of parameter space where either model 1 or model 2 alone yield the minimum MSE. It is also possible to find areas where the optimal solution transitions rapidly from one model to the other with only a narrow region where averaging is favoured in between. All of the parameters involved in the problem, including the biases and variances of each model, as well as the correlation between models, affect the conditions under which averaging is favoured. In particular, positive correlations between predictions often lead to one of the individual models being preferred over the average. In contrast, negative correlations between models broaden the conditions under which averaging is advantageous. The conditions that favour averaging models that are biased in the same direction are much more restrictive than those that favour averaging models that are either unbiased or that have opposite biases. In real-world averaging problems, the predictions from different models will unfortunately tend to be highly correlated, which will limit the utility of averaging. Overall, the surprising richness of the results produced by this “toy problem” analysis suggests that more caution may be warranted when approaching complex, real-world averaging problems.

A further note of caution emerges when we allow the weights to be estimated with uncertainty. Specifically, the regions that favour averaging contract, often dramatically, as a function of σ_w. Thus even in scenarios where averaging would beat either of the single models if the optimal weights were known exactly, estimating w with uncertainty can reverse the situation such that the average no longer yields the minimum MSE. This happens for two reasons. First, uncertainty in w is guaranteed to increase the MSE of the average, all else being equal, while the MSE values of the single models are unaffected because they do not involve weight estimation. Second, over much of the parameter space where averaging is favoured in the fixed-w case, the reduction in MSE achieved by the average is often very small, and is thus easily washed away by the increase in MSE that results from estimating w with uncertainty. Finally, we have assumed that w is estimated without bias, which will typically not be the case in reality. Allowing both bias and uncertainty in the estimation of w would very likely further erode the conditions under which averaging is advantageous.

The value of this simplified set-up is that it is possible to clearly establish the conditions that favour averaging. The two main messages that emerge from this exercise are that the conditions that favour averaging become much narrower when model predictions are positively correlated, and when the weights are estimated with uncertainty. Unfortunately, both of these conditions will frequently occur together in real-world averaging problems. However, it is important to appreciate the limitations of this exercise as well. Some types of models may make predictions that are not normally, or even uni-modally, distributed. Insights gleaned from a case where models are jointly normally distributed might be of limited utility is such situations. Furthermore, real model averaging problems will tend to feature many more than two models to (potentially) be averaged. On the one hand, this may lead to models having a range of biases such that there are at least some models that fall on each side of the truth. On the other hand, averaging a large number of highly correlated models may often not be better than the best single model, and the need to estimate many weights (instead of just one) with uncertainty will further restrict the potential advantages of averaging. As researchers frequently ignore both the correlations among models (but see Marmion et al., 2009), and the effects of estimating weights when averaging, these issues may be considerably more problematic than is currently realized.

S1.3 Fitting the full model

The "full model" refers to a model which contains all predictors of interest, possibly in different forms (e.g. as linear terms, interactions, polynomials, or non-linear functions). The full model does not need to have all true covariates in it, but it has all covariates that are used in *any* of the other models. Thus, it must contain all polynomials and interactions that are present in all other models combined. The full model potentially contains more parameters than there are data points (i.e. it is over-parameterised), and the predictors may be highly correlated. The latter is a problem for the stability of estimates (leading to inflated parameter estimates), the first prevents the use of algebra-based approaches.

Markov chain Monte Carlo-based approaches can be used to fit an over-parameterised *Bayesian model* (see Reichert and Omlin, 1997, for a discussion of why this may be a good idea). In addition to being able to accommodate so many covariates, it seems indicated to regularize such a model to reduce inflation of parameter estimates. For the Bayesian full model, a Bayesian Lasso would be the simplest way to achieve this. Priors strongly centered on zero are used to down-biase estimates (Laplace or double-exponential priors). Alternatively, one could *fit a range of univariate models and (unconditionally) average them* afterwards (see Breiner et al., 2015, for an example). Note that unconditional model averaging of the univariate regressions in fact acts like a regularization (Ghosh and Yuan, 2009). The machine-learning version is *bagging, specifically Random Forest*, where only a subset of predictors is trialled at each node.

In appendix VI we present a small and highly artificial data set (20 data points, 50 predictors with parameter value 1 each) as an illustration of how to fit the full model with the three approaches mentioned. We achieved the best fit with Random Forest (RMSE=2.20), followed by the Bayesian approach (Lasso or not give the same point estimates; RMSE=3.99). The averaged bivariate regressions are substantially worse (RMSE=8.23).

S1.4 Model averaging in machine-learning algorithms

In machine learning, several algorithms use the combination of many models to stabilise model predictions. Most famous are bagging (Breiman, 1996) and boosting (Schapire, 1990), nicely put into context by Hastie et al. (2009). Here we briefly describe the idea and approach behind these two approaches, before discussing similarities and differences with model averaging as described in the main text.

Bagging, or **b**ootstrap **agg**regation, samples from the original data with replacement and runs the model of interest on each of these bootstrap samples, creating b bootstrapped models. For prediction, each of these models makes a prediction and all predictions are then averaged (in the case of a continuous response). In the original idea no weighting of the bootstrapped models is involved, but one could also use the performance of a bootstrap on the unused data for that bootstrap (the so-called out-of-bag estimate) as a way to generate weights. Bagging can be used for any kind of statistical model. Its most common implementation is in Random Forest (Breiman, 2001), where classification and regression trees are "bagged" (requiring also that a random subset of the predictors is used for each node of each tree; this double-randomisation stabilises the Random Forest substantially more than the mere bagging would do).

Boosting (Schapire, 1990), with later adaptive boosting (AdaBoost: Freund and Schapire, 1996), has seen some development since its original conception, with stochastic gradient boosting now being the algorithm typically referred to (Friedman, 2002; Hastie et al., 2009). The key idea is that (1) a first model F_1 (such as a multiple regression or a tree) is fitted; (2) its residuals are computed; (3) a new model F_2 is fitted to these residuals; (4) a new overall model after m iterations $F_{\text{overall}} = \sum_{i=1}^{m} F_i$ is used to compute new residuals; (5) the overall model so far is evaluated on a test data set. The steps 3-5 are repeated until improvement converges. Thus, we can view boosting as a series of (diminishing) corrections of the previous set of models.

Both approaches assume that many poorly fitting models ('weak learners') can be combined to a well-fitting model ('strong learner') (Kearns and Valiant, 1989). No single model is expected to be particularly good. However, as pointed out in Hastie et al. (2009, p. 285), the issue is slightly more complicated for 0/1 data (and hence 0-1-loss), where good models improve through bagging, while poor models may well deteriorate further. Both bagging and boosting work on typically hundreds to thousands of models, although the eventual averaging actually leads to these algorithms having few actual degrees of freedom (Elder, 2003).

Clearly, bagging and boosting *employ* model averaging as part of their algorithm. They are thus on a par with model averaging of, say, linear models using AIC-weights. In fact, there is nothing special about the model averaging within bagging or boosting approaches, despite the awe in which they are held.

Another important point is that the great performance of bagging/boosting is no evidence that model averaging is a good idea *per se*. They differ substantially in their set-up to the model averaging approaches discussed in the main text:

1. Each model is a poor model *on purpose*, since weak base learners can be boosted to lower predictive errors than strong leaners (Hastie et al., 2009, p. 362).

2. They combine many (> 500) different models.

3. The focus is purely on prediction, not on parameter estimation or identification of model structure.

In summary, machine-learning methods may also internally use model averaging, in exactly the same way discussed in the text. Therefore, model aggregation as part of a machine-learning algorithm works in the same way as discussed in the main text. In fact, we hope that the main text sheds light on *why* such model aggregation works in machine learning algorithms.

S1.5 Method details

S1.5.1 Model averaging via Bayesian posterior model weights (F.H., E.M.)

As explained in the main text, Bayesian inference provides a natural framework for mixing models and parameters, where the joint posterior distribution across all possible models with all possible parameter is expressed as

$$P(M_i, \Theta_i|D) \propto L(D|M_i, \Theta_i) \cdot p(\Theta_i) \cdot p(M_i), \tag{S35}$$

The result joint distribution $P(M_i, \Theta_i|D)$ can then be used to make predictions for any desired quantity.

There are two ways to calculate $P(M_i, \Theta_i|D)$ in practice:

1. Calculating the joint probability $P(M_i, \Theta_i|D)$ directly, which is technically nearly always done via reversible jump MCMC sampling

2. Calculating marginal posterior model weights, which can then be used to weight predictions of the independent models. In practice, this requires approximating the marginal likelihood for each model, a quantity that will be further explained below.

Reversible jump MCMC as a method is arguably more elegant, and often numerically more efficient, but typically incurs a considerable time investment for implementing and tuning the reversible jump algorithm. Marginal likelihoods calculations are easier to generalize, but computational effort for stable estimates may be considerably higher than for reversible jump MCMCs.

In the rest of this section, we provide an example for either option.

Setting up a reversible jump MCMC analysis

Reversible jump MCMC (Green, 1995) is a trans-dimensional iterative algorithm developed in the Bayesian framework which allows us to obtain the posterior model probabilities and samples from the posterior distributions of model parameters. The latter posterior distributions can be summarised either conditionally on a model, for example the model with the highest posterior probability, or they can be averaged across models.

The algorithm involves two types of updates: a between-models update for moving between models with different numbers of parameters and a within-model update for updating the parameters in the current model, that is the model where the algorithm has moved to in the current iteration.

The posterior model probabilities are estimated as the proportion of time the algorithm spent in each model, which is the number of iterations the algorithm moved to a particular model divided by the total number of iterations. We note that a move to the same model is possible if the proposal to move to a different model is rejected.

Suppose that we are proposing to move from model M_i with parameter vector Θ_i to model M_j with parameter vector Θ_j. We need to define a one-to-one function, h, which satisfies $(\Theta_j, \mathbf{u}') = h(\Theta_i, \mathbf{u})$, where $\mathbf{u}$ and $\mathbf{u}'$ are sets of random variables (King et al., 2009).

The move is accepted with probability

$$A = \min\left(1, \frac{\mathrm{P(M_j, \Theta_j|D)}\mathcal{P}(\mathrm{M_i|M_j})\mathrm{q}'(\mathbf{u}')}{\mathrm{P(M_i, \Theta_i|D)}\mathcal{P}(\mathrm{M_j|M_i})\mathrm{q}(\mathbf{u})}\,|\mathrm{J}|\right), \tag{S36}$$

where $P(M_j, \Theta_j|D)$ is the posterior distribution of model M_j, as defined in eq. (S35), $\mathcal{P}(M_i|M_j)$ is the probability of proposing to move to model M_i from model M_j, given that the algorithm is currently in model M_j, $q(\mathbf{u})$ and $q'(\mathbf{u}')$ are the proposal density functions of $\mathbf{u}$ and $\mathbf{u}'$, and $|J|$ is the Jacobian matrix, defined as

$$|J| = \left|\frac{\partial(\Theta_j, \mathbf{u}')}{\partial(\Theta_i, \mathbf{u})}\right| \tag{S37}$$

For the example presented in case study 1, the four models are not nested and they are of different dimension: 20, 15, 10 and 5 predictors. We define,

$$\begin{aligned}
\Theta_1 &= (\beta_0, \beta_1, \ldots, \beta_{20}, \sigma_1^2)\\
\Theta_2 &= (\beta_0', \beta_{21}, \ldots, \beta_{35}, \sigma_2^2)\\
\Theta_3 &= (\beta_0^*, \beta_{36}, \ldots, \beta_{45}, \sigma_3^2)\\
\Theta_4 &= (\beta_0'', \beta_{46}, \ldots, \beta_{50}, \sigma_4^2)
\end{aligned}$$

We choose independent $N(0,1)$ priors for all coefficients in the model and equal prior model probabilities, i.e. $p(M_i) = 1/4$, $i = 1, \ldots, 4$. Similarly, we choose independent $N(0,1)$ proposal distributions for all coefficients.

Suppose we propose to move to model M_3 from model M_4. We need to define function h such that

$$\begin{aligned}
&\beta_0'' = \beta_0^*\\
&\beta_{36}, \ldots, \beta_{45} = \mathbf{u}\\
&\mathbf{u}' = \beta_{46}, \ldots, \beta_{50}
\end{aligned}$$

Hence, h is just the identify function which results in $|J| = 1$. In this case, the acceptance probability simplifies to

$$A = \min\left(1, \frac{\mathrm{L(D|M_3, \Theta_3)}}{\mathrm{L(D|M_4, \Theta_4)}}\right), \tag{S38}$$

since all other terms cancel out because we have used the same prior probabilities for all models, all moves between models are proposed with the same probabilities and we have set the prior and proposal distributions of model parameters to be the same. Note that the latter is not generally recommended as it can lead to low acceptance rates and poor mixing but in this case we are able to run the algorithm for long enough to compensate because of the simplicity of the model.

Now suppose we propose to move to model M_4 from model M_3. We need to define function h such that

$$\begin{aligned}
&\beta_0^* = \beta_0''\\
&\beta_{46}, \ldots, \beta_{50} = \mathbf{u}\\
&\mathbf{u}' = \beta_{36}, \ldots, \beta_{45}
\end{aligned}$$

which again gives

$$A = \min\left(1, \frac{\mathrm{L(D|M_4, \Theta_4)}}{\mathrm{L(D|M_3, \Theta_3)}}\right), \tag{S39}$$

So generally, we have set up the RJMCMC algorithm so that the acceptance probabilities for between models moves are simply the ratios of the likelihood values evaluated for each model, proposed and current.

Computing marginal posterior model weights via marginal likelihood estimation

Looking at the joint posterior eq. S35, note that we can marginalize out the parametric uncertainty for each individual model by integrating over the prior space. This quantity

$$P(D|M_i) \propto \int L(D|M_i, \Theta_i) p(\Theta_i) d\Theta_i \tag{S40}$$

is called the marginal likelihood of model i. Using this in eq. S35, we can write the marginal posterior probability for each model (e.g. the posterior support for each individual model that can be used as weight for the averaging) as

$$\mathrm{BMW}_i = \frac{P(D|M_i) \cdot p(M_i)}{\sum_j P(D|M_j) \cdot p(M_i)}. \tag{S41}$$

The technical problem that remains is estimating the marginal likelihood eq. S40. A brute-force approach would be to sample parameters from the prior, calculate the likelihood for these parameters, and take the mean of all calculated values. However, the approach is not particularly efficient when the parameter space is large.

Various alternatives and approximations have been suggested to provide better approximations. One route to improve ML estimates is concentrating sampling effort in the space of high posterior values, because only those parts of the parameter space contribute to the integral for which likelihood times prior is large. A natural idea is therefore to use MCMC samplers to make such a selection. Simple approximations using MCMC samples such as the harmonic mean harmonic mean estimator of Newton and Raftery (1994) often perform poorly in high-dimensional parameter space with wide priors. Currently, a number of alternative algorithms have been proposed, but it is not clear to us whether one is clearly preferable over the other.

Model averaging using mixtures of g-priors

We learned of this approach only after the first revision of the manuscript (many thanks to the reviewer pointing it out to us!) and hence did not implement or discuss it within the main body of this paper. The approach is another variation on Bayesian model averaging and apparently is computationally less costly. It is available, in R, in the **BMS** package (Zeugner and Feldkircher, 2015, altough models are here subsets of the specified maximal model and cannot be freely specified; hence **BMS** could not be used as part of our case study 1).

The g-prior was invented by Zellner (1986) as a multivariate prior for regression parameters β (with the co-variances being proportional to the inverse of Fisher's Information matrix $\psi(\theta)$, and g being the constant[1]):

$$\beta|\psi \sim \mathrm{MVN}\left(\beta_0, g\psi^{-1}(\theta)\right). \tag{S42}$$

One advantage of the g-prior over any other is that we only fit one parameter, g, as the covariance structure is fixed by the (data- and model-derived) Fisher matrix.

The g-prior approach was then extended to multiple models by Ley and Steel (2012), who assign a hyper-prior on g *across* regression models. Knowing g, Ley and Steel (2012) derive a formula for the marginal likelihood of each model, and thus their Bayes-factor. Hence, the g-prior-approach is an alternative way to yield the Bayesian model weights of eqn S41.

Technical details

In the case study 1 (appendix IV) we provide an example of the implementation of both rjMCMC and Bayes factor model weights. The rjMCMC code is build-for-purpose and contained in that appendix (`RJMCMCfunctions.R`).

For the Bayes factor example, we use the method by Chib and Jeliazkov (2001). This method uses existing draws from an MCMC sampler, but calculates additional points around the existing values to approximate the Marginal Likelihood. For the MCMC sampling of each single model, we used a Metropolis-Hastings MCMC with prior optimization and 30000 steps from the **BayesianTools** package, followed by an estimation of the marginal likelihood with 1000 draws from the posterior after burn-in, using an implementation of the method by Chib and Jeliazkov (2001) in the **BayesianTools** package.[2]

[1] See also `https://en.wikipedia.org/wiki/G-prior`

[2] This can be installed in R by running the following commands:

```
library(devtools)
install_url("https://dl.dropboxusercontent.com/s/hy9l6mokresqyel/BayesianTools_0.0.0.9000.tar.gz")
```

S1.5.2 BMA using Expectation Maximisation ('BMA-EM') (C.F.D.)

Bayesian Model Averaging (BMA) in general refers to any kind of model averaging within the framework of Bayesian statistics. In the literature, BMA is often associated with a particular approach, championed by Adrian Raftery and co-authors in several publications (particularly Raftery et al., 2005), using a specific approach (expectation maximisation). This BMA-EM-approach is described here.

Fundamentally, BMA-EM is *post-hoc* model averaging: it takes the fits (to training data) of several models and then essentially iteratively computes model weights to optimally fit the ensemble to the training data. These weights are then used to combine predictions from the training-data models.

More specifically, the steps comprise (Raftery et al., 2005; Zhang et al., 2009):

1. Get predictions from each of the M models ($f_m = \hat{y}_m$ in our notation; we use that of Raftery et al. to avoid too many hats in the following equations).

2. Start with equal weights ($w_1 = \ldots = w_M = 1/M$) and compute the BMA-prediction probability density as $P(y|f_1, \ldots, f_K) = \sum_{m=1}^{M} w_m g_m(y|f_m)$, where $g(.)$ is a normal PDF for each model's prediction, with mean $a_m + b_m f_m$ and standard deviation σ (which is implicitly assumed to be identical across models by Raftery et al. 2005).
   The BMA prediction is simply the expectation of this expression: $E(y|f_1, \ldots, f_M) = \sum_{m=1}^{M} w_m g_m(y|f_m)$. w_m carries the interpretation of model probability, or more precisely the probability of prediction f_m being correct, given the observed data D.
   The role of a_m and b_m is to remove bias in the predictions. They can easily be derived from fitting a linear regression of the predictions of model f_m against the (training) data y and remain constant thereafter.

3. Compute the log-likelihood of the model average,
   $$\ell_{\text{BMA}} = \sum_{i=1}^{N} \log \left(\sum_{m=1}^{M} w_m g_m(y|f_m) \right),$$
   where i refers to data points. (Note that this assumes that model predictions are independent, which is unlikely to be true. Raftery et al. (2005, p. 1159) argue that estimates of w_m are unlikely to be affected by this non-independence.)

4. Let z_{im} be a latent indicator variable, with value 1 when model m is the best prediction to data point i. We do not know the values of z_{im}, but we can estimate them as $\hat{z}_{im} = \frac{w_m g(y_i|f_{im},\sigma)}{\sum_m w_m g(y_i|f_{im},\sigma)}$. (We here omit an index for the iterations, but note that the values of $\hat{z}_{im}, w_m$ and σ are being updated with every following iteration.) Raftery et al. (2005) call this the 'expectation' or E-step.

5. From $\hat{z}_{im}$ we can now compute new values for w_m and σ (the 'maximisation' or M-step): $w_m = \frac{1}{n} \sum_{i=1}^{N} \hat{z}_{im}$, and $\sigma^2 = \frac{1}{n} \sum_{i=1}^{N} \sum_{m=1}^{M} \hat{z}_{im}(y_i - f_{im})^2$.

6. Steps 3-5 are repeated until ℓ_{BMA} converged, yielding estimates for model probabilities w_m.

Note that the models themselves are not updated or cross-validated. Vrugt et al. (2008) introduce an MCMC-based computing scheme with better convergence properties, particularly for small data sets.

Several studies apply a **power-transformation** to the prediction $f_m = \hat{y}_m^i$ for data i from model m, so that $f'_m = (\hat{y}_m^i)^{1/b}$, with b taking the value of 3 (Sloughter et al., 2007) or 4 (Hamill et al., 2004), in order to shrink predictions towards zero and thereby hope to reduce prediction bias due to extreme observations. Montgomery et al. (2012) extend this approach to binary responses, turning the power-transformation into a more complicated transformation:

$$f'_m = \left((1 + \text{logit}(f_m))^{1/b} - 1 \right) \left(\text{I}(f_m > 0.5) - \text{I}(f_m < 0.5) \right), \quad \text{(S43)}$$

with 'I' being an indicator function. (Note that in Montgomery et al. (2012)'s equation A1 absolute values appear, which is unnecessary for values of $f \in [0, 1]$.)

For this version of Bayesian model averaging there is no mechanism that controls for overfitted models. If a model near-perfectly fits the training data, it will inevitably yield a high weight, since there is no way that the calibration of the model average can tell that these fits are due to an over-parameterised model. This is fine for ensembles, where members represent the same model with different initial conditions or are known to be structurally of similar complexity. For correlational models in general, however, another **split of the training data into a calibration and a validation set** is required (Montgomery et al., 2012): the models are fitted to the calibration data, then predict to the validation hold-out, which is then used for computing the model averaging weights, which are then used on the test data for evaluation. The purpose of the validation data is to get an estimate of the true predictive performance of the models, rather than their fit. To date, no study has analysed how to split the training data, but a 50/50% or a 66/34% split seem typical (Hastie et al., 2009, p. 222). If we were to repeatedly compute BMA-weights this way, it would conceptually approach the stacking method described in section S1.5.5, although with a different optimisation function. Note that this procedure requires the number of data points to be at least twice the number of models!

Technical details

For computation (case study 2) we used the **EBMAforecast** package of R (version 0.5: Montgomery et al., 2013), setting b to a value of 3, and splitting the training data randomly into two equal-sized calibration and validation sets. (Note that we added a value of 0.0001 to all predictions with value 0 as the `calibrateEnsemble` function requires p^i_m to be in $(0, 1)$.) Although not strictly a requirement, weights are estimated more stably when $N_{\text{train}} > M$. As we split the data into half to estimate predictive performance before computing weights, this means there should be at least $2M$ data points.

Acknowledgements

Many thanks to Florian M. Hollenbach and Michael Ward for their help and access to a pre-release version of **EBMAforecast**, which is by now succeeded by the CRAN version.

S1.5.3 Information-theory-based model averaging weights

Here we briefly discuss approaches that use the fit of the model and some measure of model complexity. Most prominent is the AIC, and its small-sample version AICc. Furthermore we use the BIC, which penalises heavier and leads to smaller models being favoured. e^{BIC} can be seen as an approximation of the Bayes factor (as described in the main text). We also include here the widely applicable information criterion (WAIC, computed using R-package **blmeco**), as generalisation of the AIC, as well as Mallows'Cp (which is equivalent to AIC for normally distributed data, computed using R-package **MuMIn**).

Note that all these indices can be derived as approximations of the Kullback-Leibler distance, and as such as approximations of the predictive performance of a model (Gelman et al., 2014). The WAIC requires, as detailed in the main text, two ingredients, each themselves estimated from the model. The **blmeco** implementation does not do justice to this estimation and is largely proportional to the AIC, due to the way this package computes the WAIC. WAIC in package **BayesianTools**, in contrast, uses Gelman et al. (2014)'s definition, but works only for models fitted with that specific package.

The key challenge for computing the Bayes factor, and hence the posterior model probability, is the integral over the parameters, i.e. the marginal likelihood of a model. As mentioned in the main text, the BIC is an approximation of the Bayes factor (more specifically, e^{BIC} is), but other approximations exist. **BayesianTools**' uses MCMC-integration (i.e. summing over MCMC samples and standardising), which Chickering and Heckerman (1997) consider as 'gold standard'. That publication also mentions Laplace approximation (which is used in INLA Rue et al., 2009), the candidate method and the Cheeseman-Stutz-approximation. Having not seen them implemented for model averaging in any of the publications we encountered during this review, we refer the reader to Chickering and Heckerman (1997) for details and references for these approaches.

All these information criteria were derived under large-sample asymptotic assumptions. To allow for small-sample effects, we also include leave-one-out-cross-validation (LOO) to directly assess model fit on the data at hand. The most common criterion under LOO seems to be RMSE. We additionally investigate the log-likelihood of the omitted data point, and, for binary data, AUC and e^{AUC} (see case study 2).

S1.5.4 Naïve bootstrap model weights

Starting with a machine-learning approach to model averaging, more specifically bagging, we see that bagged models are averages across bootstraps (see appendix S1.4). In bagging, the same model type is bootstrapped and aggregated, while in model averaging (in a wider sense) we may also want to average predictions from several different types of models. However, we can still borrow the bagging idea and use bootstrapping to compute model weights, rather than performing the actual averaging of predictions.

A different way of looking at this proposal is to realise that the fit of a model to data is a function of the data (as well as the model): $\ell = f(\text{data}|\text{model})$. Thus, changing the data may lead to a model m being better or worse than its competitors. Through bootstrapping, we can quantify how often a model is the best model in the model set $\mathcal{M}$, and use this as weight when averaging model predictions. This *selection frequency* approach to model weights was proposed by Buckland et al. (1997), but later shown to be biased (Wagenmakers et al., 2004).

The procedure is computer-intensive, but simple: Bootstrap the data, fit all models, compute a measure of model fit (e.g. RMSE, log-likelihood) on the unused data, score models and tally which is the best across all bootstraps.[3]

It is generally recommended to bootstrap at least 10000 times,[4] particularly when interested in the tails of the bootstrap distribution. In model averaging low weights have a very low effect on the averaged value and we thus deemed 1000 bootstraps to be adequate for our analyses. If computing resources allow, an order of magnitude more bootstraps (> 10000) would be advisable.

Technical details

In the case studies (appendix IV), we provide an example implementing this algorithm. This is also now available in **MuMIn** as function `bootWeights`, but is there (currently) restricted to GLMs.

S1.5.5 Stacking models (C.F.D.)

Stacking describes the averaging of models based on weights computed during a cross-validation-like procedure. The process works like this:

1. Fit each model to a subset of the data (the training data), typically 1/2 of the full data set.
2. Predict to the unused subset (the test or hold-out data).

[3] There may actually be smarter ways than tallying, e.g. using the rank, or weighting ranks by absolute differences in RMSE. That is not a trivial task and we found no literature on this. Note, for example, that it would be worth to consider scaling the measure between the extremes possible, e.g. for RMSE between the RMSE for the saturated model (not 0) and an intercept-only model.

[4] `http://stats.stackexchange.com/questions/86040/rule-of-thumb-for-number-of-bootstrap-samples`

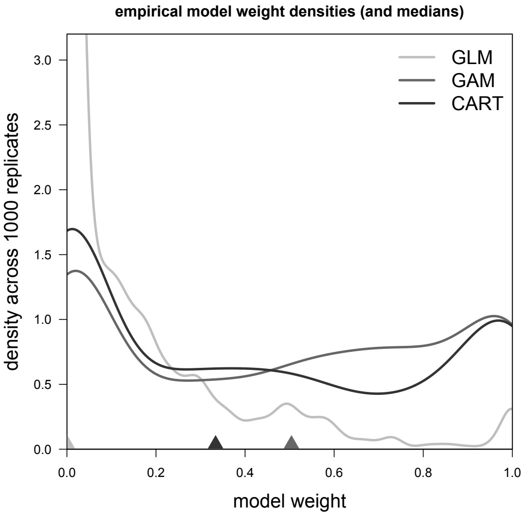

Figure S1: Densities of weights as computed across 1000 random training/test splits (80/20% each). Triangles represent median weights for each of the three model types.

3. Fit these hold-out predictions to the hold-out observations, by optimising the weights by which the models are combined. (Weights are constrained to sum to 1, and to be in [0, 1].)
4. Store optimised weights.
5. Repeat steps 1-4 many (1000) times , yielding a distribution of weights for each model (which Smyth and Wolpert (1998) referred to as an empirical Bayesian estimate of posterior model probability).
6. Compute mean or median of model weights for each model, and re-scale to sum to 1.
7. Multiply re-scaled mean (or median) weight for each model with the models' predictions or fits to compute the stacked predictions or fits.

Note that this approach requires a sample size of at least twice the number of models. By increasing the training size relative to the test this can be slightly improved. Small data sets (i.e. where N is not *much* larger than M) will often have optimisation issues. While for case study 2, 500 data points with $M = 6$ models lead to only a handful of failed optimisations, the 100 data points of case study 4 on $M = 49$ models saw a failure rate of 60-70%.

We illustrate the use of this procedure using the species richness data from Washington (see appendix S1.6 for details). In this case, we did not attempt to evaluate whether stacking was better or worse for prediction or alike, which would require another cross-validation around the stacking. Repeating the steps 1–4 1000 times yields a wide range of weights for each method (Fig. S1). The median weights we computed for GLM, GAM and CART were 0.091, 0.630 and 0.019, respectively. To evaluate the approach, we kept a 20% hold-out before the first step. In this case, mean and median stacked fits were similar (RMSE of 2.494 and 2.308, respectively), slightly worse than the fits of the two best models (GLM: 2.379, GAM: 2.381), but better than the worst (CART: 4.540). While such a simple illustration does not allow an evaluation of this approach, it at least shows that model stacking is not necessarily better than each of the stacked models.

Note that changing the size of the training/test fraction for small data sets may have a substantial effect on weights, but luckily not so much on RMSE. Using 2/3 for the training, for example, leads to the following RMSE values: mean: 2.643; median: 2.562; GLM: 2.379; GAM: 2.381; CART: 4.544. The median weights for GLM, GAM and CART are 0.029, 0.689 and 0.165, respectively (note in particular the changes for GLM and CART).

Technical details

In the case studies (appendices IV and V), we provide an example implementing this algorithm (see function `stacking` in appendices IV and V). This function was invoked 1000 times to estimate model weights. We used the mean of stacked model weights in our analyses. There is no systematic investigation into best splitting of data, minimum number of stacking runs, or the use of mean vs median stacked weights. This is also now available in **MuMIn** as function `stackingWeights`, but is there (currently) restricted to GLMs.

S1.5.6 Jackknife model averaging (C.F.D.)

Jackknife model averaging (JMA: Hansen and Racine, 2012) was proposed as an optimal weights strategy even under heteroscedastic errors. JMA has since been shown to also be optimal for mixed models and other dependence structures (Zhang et al., 2013). It optimises the model weights for jackknifed model fits. However, Nguefack-Tsague (2014) showed that such optimal models will not necessarily outperform AIC-, BIC- or unpenalised likelihood weights. Note that "jackknife" is actually not the correct term, both in its original definition (Tukey, 1958) and today's use (`https://en.wikipedia.org/wiki/Jackknife_resampling`), but it should rather be called "leave-one-out model averaging". One refers to "jackknife" if a statistics is computed on the $n-1$ data (i.e. on $y_{[-i]}$), while LOO refers to statistics on the data that were omitted (i.e. on y_i).

Let $\hat{f}\left(X_i \middle| \hat{\theta}^m_{[-i]}\right)$ be the prediction of model m, which was fitted omitting data point i, to this omitted data point. Then JMA chooses model weights w_m so as to minimise

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

under the constraint that $\sum_{m=1}^{M} w_m = 1$. Note that this is the same equation as in the main text under stacking, but replacing the H there by N, the number of data points. Instead of this RMSE-minimisation, one could also choose to maximise the likelihood:

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

Either requires a jackknife run for all models, from which the predictions to the left-out-data point are stored in a $N \times M$ matrix $\mathbf{J}$. Thus, the optimisation simplifies, in the case of RMSE-minimisation, to minimising $\sum(\mathbf{Jw}-\mathbf{Y})^2$.

Note that computational costs are relatively high and proportional to N, as it requires computing matrix $\mathbf{J}$ from omitting each data point once and recomputing *all* models.

Technical details

In the case studies (appendices IV and V), we provide examples implementing this algorithm. It is also now available in **MuMIn** as function `jackknifeWeights`, but is there (currently) restricted to GLMs.

S1.5.7 Bates-Granger model weights

This approach uses prediction covariance to compute model weights. To get an estimate of prediction covariance, we split the data into two, fit models to one half and predict to the other half. These predictions are then used to compute the $M \times M$ variance-covariance between models, $\mathbf{\Sigma}$.

From $\mathbf{\Sigma}$, model weights w are algebraicly computed as:

$$w_{\text{Bates-Granger}} = (\mathbf{1}'\mathbf{\Sigma}^{-1}\mathbf{1})^{-1}\mathbf{1}\mathbf{\Sigma}^{-1}$$

which in R is:
`ginv(t(ones)%*%ginv(Sigma)%*%ones)%*%ones%*%ginv(Sigma)`, where `ones` is a vector of 1s of length M. Note that we mostly use `MASS::ginv` rather than `solve` for matrix inversion as `ginv` seems to be more stable near singularities.

A peculiar feature of Bates-Granger weights is that although they sum to 1, they need not be from $[0, 1]$. Indeed, for small data sets we repeatedly get very high and very negative weights. As the estimated model averages display very poor fits in these cases, we assume that the Bates-Granger method works best with a large number of data points. The crucial problem seems to be the stable estimation of $\mathbf{\Sigma}$.

Technical details

In the case studies (appendices IV and V), we provide examples implementing this algorithm (see also function `BatesGranger` in appendices IV and V). It is also now available in **MuMIn** as function `BGWeights`, but is there (currently) restricted to GLMs.

S1.5.8 Cos-squared model weights

We implemented this approach following the algorithm outlined in the appendix of Garthwaite and Mubwandarikwa (2010), which we reproduce here, using our notation. The R-function is included in the case studies.

The following algorithm determines the weights $w_1, \dots, w_M$, corresponding to a $M \times M$ correlation matrix $\mathbf{R}$. It is being computed as the correlation of the predictions of each model with each other, i.e. the (i, j)th element of $\mathbf{R}$ is cor$(\hat{y}_i, \hat{y}_j)$.

1. Set $\mathbf{D}_1$ equal to the $M \times M$ identity matrix and put $i = 1$.

2. At the ith iteration, perform a spectral decomposition of $\mathbf{D}_i\mathbf{R}\mathbf{D}_i$,giving
$$\mathbf{D}_i\mathbf{R}\mathbf{D}_i = \mathbf{Q}_i\mathbf{\Lambda}a_i\mathbf{Q}_i^\top,$$
where $\mathbf{\Lambda}_i$ is a diagonal matrix whose non-zero elements are the eigenvalues of$\mathbf{D}_i\mathbf{R}\mathbf{D}_i$ and the columns of $\mathbf{Q}_i$ are the corresponding orthonormalized eigenvectors. [For $j = 1, \dots, k$,the jth column of $\mathbf{Q}_i$ is the eigenvector corresponding to the eigenvalue in the (j, j)th element of $\mathbf{\Lambda}_i$.]

3. Put $\mathbf{E}_i = \mathbf{D}_i^{-1}\mathbf{Q}_i\mathbf{\Lambda}_i^{1/2}\mathbf{Q}_i^\top$, where $\mathbf{\Lambda}_i^{1/2}$ is a diagonal matrix and $\mathbf{\Lambda}_i^{1/2}\mathbf{\Lambda}_i^{1/2} = \mathbf{\Lambda}_i$.

4. Set $\mathbf{D}_{i+1}$ equal to a diagonal matrix, with diagonal equal to the diagonal of $\mathbf{E}_i$.

5. Return to step (2) until convergence, when $\mathbf{D}_{i+1} \approx \mathbf{D}_i$.

6. Let $d_1, \dots, d_M$ denote the diagonal elements of $\mathbf{D}^*$, where $\mathbf{D}^*$ is the value of $\mathbf{D}_i$ at convergence. For $i = 1, \dots, M$ put $w_i = d_i^2 / \sum_{j=1}^M d_j^2$.

Note that convergence is not guaranteed, especially when $\mathbf{R}$ is based on a small number of predicted points.

Technical details

In the case studies (appendices IV and V), we provide examples implementing this algorithm. The function `csweights` (in appendices IV and V) is also now available in **MuMIn** as function `cos2Weights`, but is there (currently) restricted to GLMs.

S1.5.9 Model-based model combinations (“ensemble supra-model”) (C.F.D.)

We can use any regression model approach to combine predictions from the different models (i.e., GLMs, GAMs, Random Forest, neural networks). However, there are two points worth keeping in mind:

1. Model fits are used to train the supra-model. Hence any model that overfits will look good to the MBMC algorithm, while it actually is too optimistic. This requires a train/test-splitting of the data, training on half of them and using the predictions to the other half as a fairer assessment of a models fit. The MBMC is then trained on these predictions to the test data.

2. The idea of MBMC is to have weights that depend on the predicted value. For low values of $\hat{y}$ one (set of) model predictions is used, for high values another one. To allow model weights to vary smoothly along the prediction value gradient, GLM and GAM seem more appropriate than CARTs or Random Forest. This is a conceptual decision, however, not a statistical one.

We used a GLM, a GAM and a Random Forest to combine model predictions.

The steps of the MBMC-approach are as follows:

1. Split the data randomly into two sets: train and test.

2. Fit all modelling approaches to the train data (or, if they were all fitted before the full data set, update these models with the train data only).

3. Predict with each model to the test data set.

4. Use a regression-like approach of your choice to fit the predictions (as predictors) to the response data in the test data. This is the MBMC-model.

5. Fit the models to the full data set. Predict with the full models to new data.

6. Use the predictions to new data as new data for predicting from fitted the MBMC-model.

Clearly we are using the data twice: once for training the MBMC, and once for fitting each model before predicting to new data. This is not ideal. When we have many data points (relative to the number of models), we should, in step 5, predict with the models of step 2 to new data, rather than re-fitting the full model.

More important is however the training of the MBMC on the test data, rather than on the training data, which would lead to a weights in favour of the most overfitted model in the set.

Non-averaging approaches for combining models

Several publications in the remote sensing literature use Bayesian updating to explore the value of adding new layers of information (going back to Strahler, 1980). Starting with one model (e.g. a digital elevation model), new layers are added, using the fit of the first layer as prior (e.g. Pereira and Itami, 1991; Osborne et al., 2001; Romero et al., 2016). While this may not be a bad idea, models are not averaged but sequentially combined. As a consequence, predictions from this approach depend on the arbitrary order in which models were combined.

The problem of simultaneously optimising a supra-model along with its components

Instead of the approach just outlined, we could think of fitting the models **and** the supra-model simultaneously. This may lead to very differing optimal values for the single models’ parameters, as it is part of a larger model construction that fits the data. The MBMC-approach above uses the supra-model only to squeeze out the most from the models’ predictions, while the simultaneous fitting can be seen as a single model in itself, which just happens to have the single models as sub-components. Here we address the problems arising when trying to follow the simultaneous optimisation of the models and their combination in a supra-model.

Simultaneous fitting of models and supra-model faces the problem of parameter identifiability. If several models share an additive term then either model can contribute through it to the final prediction, and no single optimal solution exists. We illustrate this point by two obvious and one less obvious examples, each combining three models.

Example 1 In a first, near-trivial example we show that three models can be combined to yield the same result as a single, "full model". Mathematically that is little surprising, but it is supposed to drive home the point that one can conceptually view this three-model mixture as the same thing as a "full model". We may not always be able to formulate this "full model", but we can use mixture models to represent it. In this simple-most case, we consider a constant model, a linear model without intercept, and a quadratic model without intercept or linear term. A moments reflection shows that this is equivalent to fitting a 2nd-order polynomial function:

$$w_1 b_1 + w_2 b_2 x + w_3 b_3 x^2 = r + sx + tx^2,$$

where $r = w_1 b_1$, $s = w_2 b_2$ and $t = w_3 b_3$. The additional constraint that $\sum w_i = 1$ is not sufficient to lead to a unique solution. Imagine the true value of r, s and t is 1. We can now find different combinations of w_i and b_i to yield these values: $w = (0.5, 0.3, 0.2)$ and $b = (2, 3.33, 5)$, or $w = (0.1, 0.1, 0.8)$ and $b = (10, 10, 1.25)$, and so forth.

This means, while we can find solutions, by mixture models, that are identical to the 'full model' solution, the mixture parameters w_i and model parameters b_i are by themselves meaningless and can only be interpreted in combination.

Example 2 Instead of three terms, we now fit three simple, nested models to a data set with a threshold. The models are i) a horizontal line, ii) a linear function, and iii) a quadratic function. As before, these three models are combined as weighted sum, with weights summing to 1 and being positive. Since the horizontal line is equivalent to the intercept in models ii and iii, and the slope of model ii also appears in model iii, we can imagine an infinite number of possibilities of how to choose weights and intercepts/slopes to yield the exact-same line. For this set of nested models no unique solution exists, even if we would set mixture weights to a fixed value.

As in the previous example, depending on which starting values we provide we get a different set of optimal parameters, all perfectly equivalent in terms of the resulting function.

$$w_1 a + w_2(b + cx) + w_3(d + ex + fx^2) = (w_1 a + w_2 b + w_3 d) + (w_2 c + w_3 e)x + fx^2 = r + sx + tx^2$$

Obviously, the model parameters are not uniquely identifiable, requiring use to express it differently (as quadratic polynomial).

Example 3 We shall now turn to a less obvious example, three non-linear growth functions, neither of which is the special case of another (the function being the logistic, Gompertz and Richards growth equations).

We fit three growth curve models to a simulated data set, which is actually the 0.3/0.5/0.2-mixture of the three models (see below for R code). Although the models are mathematically not nested, we still cannot correctly identify model and mixing parameters. Depending on the starting conditions, we get different 'optimal' parameters and mixture weights (or the algorithm doesn't converge at all due to this ill-posed problem).

The reason is the same as in example 2: we can choose arbitrary mixture weights and adjust model parameters (within limits defined by the non-linearity of the function) accordingly. Hence, no unique solution exists, but the infinitely many ‘optimal’ solutions identified are identical in terms of the resulting mixture function. Again, model parameters and mixture weights have no interpretable meaning by themselves.

To our knowledge, model averaging through mixtures of models has been discussed by statisticians, but has not been implemented yet.[5] Our examples may illustrate why.

```
library(grofit)
data(grofit.time)
X ← t(grofit.time[1,])
# invent a data set as a composite of three growth equations:
Y ← 0.3*gompertz(X, 0.23, 0.02, 3) + 0.5*logistic(X, 0.24, 0.03, 3.5)
        + 0.2*richards(X, 0.24, 0.04, 6.5, 25) + rnorm(65, 0, 0.005)

# now fit each of the three models to it:
(g1 ← optim(par=c(.2,.02,4), fn=function(par) sum((Y — gompertz(X,
        par[1], par[2], par[3]))^2), method=”BFGS”))
(g2 ← optim(par=c(.2,1,1), fn=function(par) sum((Y — logistic(X,
        par[1], par[2], par[3]))^2), method=”BFGS” ))
(g3 ← optim(par=c(.2,.05, 6, 23), fn=function(par) sum((Y —
        richards(X, par[1], par[2], par[3], par[4]))^2), method=”BFGS”))

#### plot the result:
plot(Y ~ X, las=1)
curve(gompertz(x, g1$par[1], g1$par[2], g1$par[3]), add=T, col=30)
curve(logistic(x, g2$par[1], g2$par[2], g2$par[3]), add=T, col=60)
curve(richards(x, g3$par[1], g3$par[2], g3$par[3], g3$par[4]),
        add=T, col=80)

mixgrofun ← function(par){
        ww ← par[1:2]
        ww ← c(1, exp(ww))
        w ← ww/sum(ww)
        RMSE ← w[1] * sum((Y—gompertz(X, par[3], par[4], par[5]))^2) +
                  w[2] * sum((Y—logistic(X, par[6], par[7], par[8]))^2) +
                  w[3] * sum((Y—richards(X, par[9], par[10], par[11],
                  par[12]))^2)
        return(—RMSE)
}
mixgrofun(par=c(0,1, 0.2,.05,5, 0.2,.05,5, 0.2,.05,5,20))
(ops ← optim(par=c(0,0, 0.2,.05,5, 0.2,.05,4, 0.2,.05,5,20), fn=mixgrofun, method=”Nelder—Mead”, hessian=F,
     control=list(maxit=100000)))
# this yields a convergence—warning ‘10’, which means
# ‘‘degeneracy of the Nelder—Mead algorithm’’!
```

[5] Rings et al. (2012) use mixture models, but they still use standard BMA to combine these into an ensemble forecast.

S1.6 Non-independence of models affects model weights

Here we give a simple example illustrating the effect of having variations of the same model multiple times in the ensemble. We replaced the best-performing model by three variations of the same model, varying only in a complexity parameter and re-computed model weights.

Specifically, we calculated model weights based on 10-fold cross-validation performance, measured as log-likelihood on the hold out (see R-code for details). The data set of Aho and Weaver (2010) (provided as `wash.rich` in package **asbio**) comprises 40 data points (vascular species richness) and 5 predictors (soil N, slope, aspect, rock cover, soil pH). We used a GLM with quadratic polynomial predictors, a GAM with cubic shrinkage and a maximum of 3 knots per predictor and a Classification and Regression Tree (CART) with two variables trialled per split. To produce the three variants, we changed the type of spline used in the GAM (the original cubic shrinkage plus thin-plate with shrinkage for GAM 1 and Duchon spline for GAM 3).

Fig. S2 shows that having similar versions of a model approach in the model set leads to a down-weighting of all models. As the replaced version is now present thrice, it absorbs most of the model weights. For all practical purposes, we now have predictions dominated by one model approach, and we lost some of the model structural uncertainty by dilution with replicates of one model variant.

In a real-world setting, such non-independence of model approaches is far less obvious, but potentially has a similar effect.

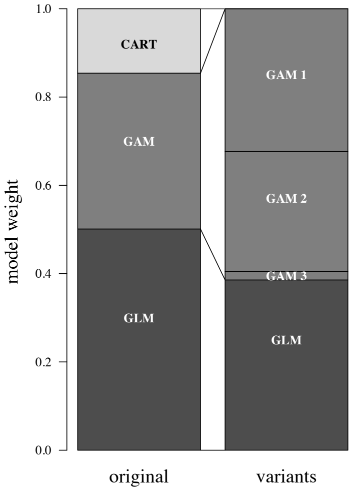

Figure S2: Effect of including variations of Generalised Additive Models thrice in the model set on model weights. GLM and CART refer to Generalised Linear Model and Classification and Regression Tree, respectively. In the three-GAM-variant, CART's weight is effectively reduced to 0.

Bibliography

Aho, K. and T. Weaver. 2010. Ecology of alpine nodes: environments, communities, and ecosystem evolution (Mount Washburn; Yellowstone Natl. Park). Arctic, Antarctic, and Alpine Research, **40**:139–151.

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.

Breiman, L. 1996. Bagging predictors. Machine Learning, **24**:123–140.

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

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.

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.

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

Chib, S. and I. Jeliazkov. 2001. Marginal likelihood from the Metropolis-Hastings output. Journal of the American Statistical Association, **96**:270–281.

Chickering, D. M. and D. Heckerman. 1997. Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables. Machine Learning, **29**:181–212.

Elder, J. F. 2003. The generalization paradox of ensembles. Journal of Computational and Graphical Statistics, **12**:853–864.

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.

Freund, Y. and R. E. Schapire. 1996. Experiments with a new boosting algorithm. In L. Saitta, editor, Machine Learning: Proceedings of the Thirteenth International Conference (ICML '96), pages 148–156. Morgan Kaufmann.

Friedman, J. H. 2002. Stochastic gradient boosting. Computational Statistics and Data Analysis, **38**:367–378.

Galipaud, M., M. A. F. Gillingham, M. David, and F.-X. Dechaume-Moncharmont. 2014. Ecologists overestimate the importance of predictor variables in model averaging: a plea for cautious interpretations. Methods in Ecology and Evolution, **5**:983–991.

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

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.

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

Hamill, T. M., J. S. Whitaker, and X. Wei. 2004. Ensemble re-forecasting: Improving medium-range forecast skill using retrospective forecasts. Bulletin of the American Meteorological Society, **132**:3825–3830.

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.

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

Kearns, M. and L. Valiant. 1989. Crytographic limitations on learning Boolean formulae and finite automata. Symposium on Theory of Computing (ACM), **21**:433–444.

King, R., O. Gimenez, B. Morgan, and S. Brooks. 2009. Bayesian Analysis for Population Ecology. Chapman & Hall/Crc Interdisciplinary Statistics Series, Chapman & Hall/CRC.

Ley, E. and M. F. Steel. 2012. Mixtures of g-priors for Bayesian model averaging with economic applications. Journal of Econometrics, **171**:251–266.

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.

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.

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

Montgomery, J. M., F. Hollenbach, and M. D. Ward. 2013. EBMAforecast: Ensemble BMA Forecasting. R package version 0.42.

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

Newton, M. A. and A. E. Raftery. 1994. Approximate Bayesian inference with the weighted likelihood bootstrap. Journal of the Royal Statistical Society B, **56**:3–48.

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

Osborne, P. E., J. C. Alonso, and R. G. Bryant. 2001. Modelling landscape-scale habitat use using gis and remote sensing: a case study with great bustards. Journal of Applied Ecology, **38**:458–471.

Pereira, J. M. C. and R. M. Itami. 1991. Gis-based habitat modeling using logistic multiple regression: a study of the mt. graham red squirrel. Photogrammetric Engineering & Remote Sensing, **57**:1475–1486.

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.

Reichert, P. and M. Omlin. 1997. On the usefulness of overparameterized ecological models. Ecological Modelling, **95**:289–299.

Rings, J., J. A. Vrugt, G. Schoups, J. A. Huisman, and H. Vereecken. 2012. Bayesian model averaging using particle filtering and Gaussian mixture modeling: Theory, concepts, and simulation experiments. Water Resources Research, **48**:W05520.

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.

Rue, H., S. Martino, and N. Chopin. 2009. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. Journal of the Royal Statistical Society. Series B: Statistical Methodology, **71**:319–392.

Schapire, R. E. 1990. The strength of weak learnability. Machine Learning, **5**:197–227.

Sloughter, J. M. L., A. E. Raftery, T. Gneiting, and C. Fraley. 2007. Probabilistic quantitative precipitation forecasting using Bayesian model averaging. Monthly Weather Review, **135**:3209–3220.

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.

Strahler, A. H. 1980. The use of prior probabilities in maximum likelihood classification of remotely sensed data. Remote Sensing of Environment, **10**:135–163.

Tukey, J. W. 1958. Bias and confidence in not quite large samples. The Annals of Mathematical Statistics, **29**:614–623.

Turkheimer, F. E., R. Hinz, and V. J. Cunningham. 2003. On the undecidability among kinetic models: From model selection to model averaging. Journal of Cerebral Blood Flow & Metabolism, **23**:490–498.

Vrugt, J. A., C. G. H. Diks, and M. P. Clark. 2008. Ensemble Bayesian model averaging using Markov Chain Monte Carlo sampling. Environmental Fluid Mechanics, **8**:579–595.

Wagenmakers, E.-J., S. Farrell, and R. Ratcliff. 2004. Naive nonparametric bootstrap model weights are biased. Biometrics, **60**:281–283.

Zellner, A. 1986. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In P. Goel and A. Zellner, editors, Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti, page 233–243. Studies in Bayesian Econometrics, Elsevier.

Zeugner, S. and M. Feldkircher. 2015. Bayesian model averaging employing fixed and flexible priors: The BMS package for R. Journal of Statistical Software, **68**.

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.

S1.7 Model-averaging literature scanned

To get an overview of the model averaging literature, we conducted a literature search using a topic search in Web of Science and the following keywords:

```
((("predict*" OR "parameter") AND ("variance" OR "uncertainty" OR "error")) AND ("model␣averaging" OR "model␣
    weighting" OR "consensus␣model" OR "combining␣forecasts" OR "model␣mixing" OR "combining␣models" OR "
    model␣aggregation" OR "rjMCMC" OR "multi−model" OR "model␣ensemble"))
```

This revealed 827 studies (16 March 2015) from which we randomly chose 181 studies for detailed examination. From the 181 studies, only 87 studies made claims why model averaging may lead to model improvements such as better coverage, less biased parameter estimates, lower variance or lower prediction errors, or compared model averaging with single model approaches. We extracted from these 87 studies the field of application, the type of data used (simulated or empirical data or both), and whether they found evidence that model averaging leads to improved models. Their findings are summarised, citing only the most general and relevant studies, in section 2.3 of the main text.

Below we list all publications we considered in our assessment of claims about model averaging. With such a large number of potential publications it is inevitable to overlook some potentially important ones. This list demonstrates that we opted for a wide, cross-disciplinary search, across many years and journals. Additional studies cited in the main paper, which we found through cross-references or other means than the literature review, are not included here.

Ajami, N.K., Duan, Q. & Sorooshian, S. (2007) An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction. Water Resources Research, 43, W01403.

Araujo, M.B. & New, M. (2007) Ensemble forecasting of species distributions. Trends in Ecology & Evolution, 22, 42–47.

Arnold, T.W. (2010) Uninformative parameters and model selection using Akaike's Information Criterion. Journal of Wildlife Management, 74, 1175–1178.

Ball, R. D (2001) Bayesian methods for quantitative trait loci mapping based on model selection: approximate analysis using the Bayesian Information Criterion. Genetics, 159: 1351–1364.

Bates, J.M. and Granger, C.W.J. (1969) The combination of forecasts. Operational Research Quarterly 20(4): 451–468.

Bohn, T.J., Sonessa, M.Y. & Lettenmaier, D.P. (2010) Seasonal hydrologic forecasting: do multimodel ensemble averages always yield improvements in forecast skill? Journal of Hydrometeorology, 11, 1358–1372.

Bradley, B.A. (2009) Regional analysis of the impacts of climate change on cheatgrass invasion shows potential risk and opportunity. Global Change Biology, 15, 196–208.

Buisson, L., W. Thuiller, et al. (2010) Uncertainty in ensemble forecasting of species distribution. Global Change Biology 16(4): 1145–1157.

Bunnin, F.O., Guo, Y. & Ren, Y. (2002) Option pricing under model and parameter uncertainty using predictive densities. Statistics and Computing, 12, 37–44.

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

Butler, A., Doherty, R. M. & Marion, G. (2009) Model averaging to combine simulations of future global vegetation carbon stocks. Environmetrics, 20, 791–911.

Calcagno, V. & De Mazancourt, C. (2010) glmulti: An R package for easy automated model selection with (Generalized) Linear Models. Journal of Statistical Software, 34.

Chen, X., Zou, G. & Zhang, X. (2012) Frequentist model averaging for linear mixed–effects models. Frontiers of Mathematics in China, 8, 497–515.

Chickering, D.M. & Heckerman, D. (1997) Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables. Machine Learning, 29, 181–212.

Clark, A.J., Gallus, W.A., Xue, M. & Kong, F. (2009) A Comparison of Precipitation Forecast Skill between Small Convection–Allowing and Large Convection–Parameterizing Ensembles. Weather and Forecasting, 24, 1121–1140.

Crimmins, S. M., S. Z. Dobrowski, et al. (2013) Evaluating ensemble forecasts of plant species distributions under climate change. Ecological Modelling 266: 126–130.

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

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

Dochtermann, N.A. & Jenkins, S.H. (2011) Developing multiple hypotheses in behavioral ecology. Behavioral Ecology and Sociobiology, 65, 37–45.

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

Droguett, E.L. & Mosleh, A. (2013) Integrated treatment of model and parameter uncertainties through a Bayesian approach. Proceedings of the Institution of Mechanical Engineers Part O-Journal of Risk and Reliability, 227, 41–54.

Duan, Q., Ajami, N.K., Gao, X. & Sorooshian, S. (2007) Multi-model ensemble hydrologic prediction using Bayesian model averaging. Advances in Water Resources, 30, 1371–1386.

Ebert, E.E. (2001) Ability of a poor man's ensemble to predict the probability and distribution of precipitation. Monthly Weather Review, 129, 2461–2480.

Edeling, W.N., Cinnella, P. & Dwight, R.P. (2014) Predictive RANS simulations via Bayesian Model–Scenario Averaging. Journal of Computational Physics, 275, 65–91.

Efficace, F., Therasse, P., Piccart, M.J., Coens, C., Van Steen, K., Welnicka–Jaskiewicz, M., Cufer, T., Dyczka, J., Lichinitser, M., Shepherd, L., de Haes, H., Sprangers, M.A. & Bottomley, A. (2004) Health-related quality of life parameters as prognostic factors in a nonmetastatic breast cancer population: An international multicenter study. Journal of Clinical Oncology, 22, 3381–3388.

Elder, J.F. (2003) The generalization paradox of ensembles. Journal of Computational and Graphical Statistics, 12, 853–864.

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

Engler, R., L. T. Waser, et al. (2013) Combining ensemble modeling and remote sensing for mapping individual tree species at high spatial resolution. Forest Ecology and Management 310: 64–73.

Feldkircher, M. (2012) Forecast Combination and Bayesian Model Averaging: A Prior Sensitivity Analysis. Journal of Forecasting, 31, 361–376.

Fernandez, C., Ley, E. & Steel, M.F.J. (2001b) Model uncertainty in cross–country growth regressions. Journal of Applied Econometrics, 16, 563–576.

Fischer, E.M., Lawrence, D.M. & Sanderson, B.M. (2011) Quantifying uncertainties in projections of extremes–a perturbed land surface parameter experiment. Climate Dynamics, 37, 1381–1398.

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

Forstmeier, W. & Schielzeth, H. (2011) Cryptic multiple hypotheses testing in linear models: overestimated effect sizes and the winner's curse. Behavioral Ecology and Sociobiology, 65, 47–55.

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.

Galipaud, M., Gillingham, M. A. F., David, M. & Dechaume–Moncharmont, F.–X. (2014) Ecologists overestimate the importance of predictor variables in model averaging: a plea for cautious interpretations. Methods in Ecology and Evolution, in press.

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

Garratt, A., Koop, G. & Vahey, S.P. (2008) Forecasting substantial data revisions in the presence of model uncertainty. Economic Journal, 118, 1128–1144.

Garroway, C.J., Bowman, J. & Wilson, P.J. (2011) Using a genetic network to parameterize a landscape resistance surface for fishers, *Martes pennanti*. Molecular Ecology, 20, 3978–3988.

Ghosh, D., and Yuan, Z. (2009), An Improved Model Averaging Scheme for Logistic Regression. Journal of Multivariate Analysis,100, 1670–1681.

Giannini, T.C., Chapman, D.S., Saraiva, A.M., Alves–dos–Santos, I. & Biesmeijer, J.C. (2013) Improving species distribution models using biotic interactions: a case study of parasites, pollinators and plants. Ecography, 36, 649–656.

Giudice, J.H., Fieberg, J.R. & Lenarz, M.S. (2012) Spending degrees of freedom in a poor economy: A case study of building a sightability model for moose in northeastern Minnesota. Journal of Wildlife Management, 76, 75–87.

Grenouillet, G., Buisson, L., Casajus, N. & Lek, S. (2011) Ensemble modelling of species distribution: the effects of geographical and environmental ranges. Ecography, 34, 9–17.

Grueber, C.E., Nakagawa, S., Laws, R.J. & Jamieson, I.G. (2011) Multimodel inference in ecology and evolution: challenges and solutions. Journal of Evolutionary Biology, 24, 699–711.

Hamilton, G., McVinish, R. & Mengersen, K. (2009) Bayesian model averaging for harmful algal bloom prediction. Ecological Applications, 19, 1805–1814.

Hartley, S., Harris, R. and Lester, P. J. (2006), Quantifying uncertainty in the potential distribution of an invasive species: climate and the Argentine ant. Ecology Letters, 9: 1068–1079.

Hickerson, M.J., Stone, G.N., Lohse, K., Demos, T.C., Xie, X., Landerer, C. & Takebayashi, N. (2014) Recommendations for Using Msbayes to Incorporate Uncertainty in Selecting an Abc Model Prior: A Response to Oaks Et Al. Evolution, 68, 284–294.

Hoeting, J.A., Madigan, D., Raftery, A.E. & Volinsky, C.T. (1999) Bayesian model averaging: a tutorial. Statistical Science, 14, 382–401.

Holan, S., McElroy, T. & Chakraborty, S. (2009) A Bayesian Approach to Estimating the Long Memory Parameter. Bayesian Analysis, 4, 159–190.

Jackson, C.H., Bojke, L., Thompson, S.G., Claxton, K. & Sharples, L.D. (2011) A Framework for Addressing Structural Uncertainty in Decision Models. Medical Decision Making, 31, 662–674.

Jackson, C.H., Sharples, L.D. & Thompson, S.G. (2010) Structural and parameter uncertainty in Bayesian cost–effectiveness models. Journal of the Royal Statistical Society Series C–Applied Statistics, 59, 233–253.

Jackson, C.H., Thompson, S.G. & Sharples, L.D. (2009) Accounting for uncertainty in health economic decision models by using model averaging. Journal of the Royal Statistical Society Series A–Statistics in Society, 172, 383–404.

Jackson, C.R., Meister, R. & Prudhomme, C. (2011) Modelling the effects of climate change and its uncertainty on UK Chalk groundwater resources from an ensemble of global climate model projections. Journal of Hydrology, 399, 12–28.

Johnson C, Bowler N (2009) On the reliability and calibration of ensemble forecasts. Mon. Wea. Rev. 137:1717–1720

Kar, S.C., Hovsepyan, A. & Park, C.K. (2006) Economic values of the APCN multi–model ensemble categorical seasonal predictions. Meteorological Applications, 13, 267–277.

Koop, G. & Korobilis, D. (2014) A new index of financial conditions. European Economic Review, 71, 101–116.

Koyama, H. & Watanabe, M. (2010) Reducing Forecast Errors Due to Model Imperfections Using Ensemble Kalman Filtering. Monthly Weather Review, 138, 3316–3332.

Kuehnlein, C., Keil, C., Craig, G.C. & Gebhardt, C. (2014) The impact of downscaled initial condition perturbations on convective–scale ensemble forecasts of precipitation. Quarterly Journal of the Royal Meteorological Society, 140, 1552–1562.

Kulkarni, M.A., Acharya, N., Kar, S.C., Mohanty, U.C., Tippett, M.K., Robertson, A.W., Luo, J.-J. & Yamagata, T. (2012) Probabilistic prediction of Indian summer monsoon rainfall using global climate models. Theoretical and Applied Climatology, 107, 441–450.

Landrum, M.B. & Becker, M.P. (2001) A multiple imputation strategy for incomplete longitudinal data. Statistics in Medicine, 20, 2741–2760.

Lawler, J. J., D. White, R. P. Neilson, and A. R. Blaustein. 2006. Predicting climate–induced range shifts: model differences and model reliability. Global Change Biology 12:1568–1584.

LeBlanc, M. & Tibshirani, R. (1996) Combining Estiamates in Regression and Classification. Journal of the American Statistical Association, 91, 1641–1650.

Lee, S.-S., Lee, J.-Y., Ha, K.-J., Wang, B. & Schemm, J.K.E. (2011) Deficiencies and possibilities for long–lead coupled climate prediction of the Western North Pacific–East Asian summer monsoon. Climate Dynamics, 36, 1173–1188.

Lee, T.-H. & Yang, Y. (2006) Bagging binary and quantile predictors for time series. Journal of Econometrics, 135, 465–497.

Leon, A. & Gonzalez, R.L. (2009) Predicting soil organic carbon percentage from loss–on-ignition using Bayesian Model Averaging. Australian Journal of Soil Research, 47, 763–769.

Li, J. & Chen, W. (2014) Forecasting macroeconomic time series: LASSO–based approaches and their forecast combinations with dynamic factor models. International Journal of Forecasting, 30, 996–1015.

Li, Y., Kinzelbach, W., Zhou, J., Cheng, G.D. & Li, X. (2012b) Modelling irrigated maize with a combination of coupled–model simulation and uncertainty analysis, in the northwest of China. Hydrology and Earth System Sciences, 16, 1465–1480.

Lin, Z. & Beck, M.B. (2012) Accounting for structural error and uncertainty in a model: An approach based on model parameters as stochastic processes. Environmental Modelling & Software, 27–28, 97–111.

Logutov, O.G. & Robinson, A.R. (2005) Multi–model fusion and error parameter estimation. Quarterly Journal of the Royal Meteorological Society, 131, 3397–3408.

Marmion, M., Hjort, J., Thuiller, W., Luoto, M. (2009) Statistical consensus methods for improving predictive geomorphology maps. Computer and Geosciences, 35, 615–625

Marmion, M., M. Parviainen, et al. (2009) Evaluation of consensus methods in predictive species distribution modelling. Diversity and Distributions 15(1): 59–69.

Mbogga, M.S., Wang, X. & Hamann, A. (2010) Bioclimate envelope model predictions for natural resource management: dealing with uncertainty. Journal of Applied Ecology, 47, 731–740.

McAlpine, C.A., Rhodes, J.R., Bowen, M.E., Lunney, D., Callaghan, J.G., Mitchell, D.L. & Possingham, H.P. (2008) Can multiscale models of species' distribution be generalized from region to region? A case study of the koala. Journal of Applied Ecology, 45, 558–567.

Montgomery, J.M. & Nyhan, B. (2010) Bayesian model averaging: Theoretical developments and practical applications. Political Analysis, 18, 245–270.

Montgomery, J.M., Hollenbach, F.M. & Ward, M.D. (2012) Improving Predictions Using Ensemble Bayesian Model Averaging. Political Analysis, 20, 271–291.

Nagib, G., Gharieb, W. & Binder, Z. (1992) Qualitative Multimodel Control Using a Learning Approach. International Journal of Systems Science, 23, 855–869.

Najafi, M.R., Moradkhani, H. & Jung, I.W. (2011) Assessing the uncertainties of hydrologic model selection in climate change impact studies. Hydrological Processes, 25, 2814–2826.

Nakaegawa, T., Kusunoki, S., Sugi, M., Kitoh, A., Kobayashi, C. & Takano, K. (2007) A study of dynamical seasonal prediction of potential water resources based on an atmospheric GCM experiment with prescribed sea–surface temperature. Hydrological Sciences Journal–Journal Des Sciences Hydrologiques, 52, 152–165.

Nakagawa, S. & Freckleton, R.P. (2010) Model averaging, missing data and multiple imputation: a case study for behavioural ecology. Behavioral Ecology and Sociobiology, 65, 103–116.

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

Namhata, A., Small, M.J. & Karamalidis, A.K. (2014) Multi–model weighted predictions for CH4 and H2S solubilities in freshwater and saline formation waters relevant to unconventional oil and gas extraction. International Journal of Coal Geology, 131, 177–185.

Nandi, M., Dewanji, A., Roy, B. & Sarkar, S. (2014) Model Selection Approach for Distributed Fault Detection in Wireless Sensor Networks. International Journal of Distributed Sensor Networks, 148234.

Narayan, A., Marzouk, Y. & Xiu, D. (2012) Sequential data assimilation with multiple models. Journal of Computational Physics, 231, 6401–6418.

Nasseri, M., Zahraie, B., Ajami, N.K. & Solomatine, D.P. (2014) Monthly water balance modeling: Probabilistic, possibilistic and hybrid methods for model combination and ensemble simulation. Journal of Hydrology, 511, 675–691.

Naujokaitis–Lewis, I.R., Curtis, J.M.R., Tischendorf, L., Badzinski, D., Lindsay, K. & Fortin, M.–J. (2013) Uncertainties in coupled species distribution–metapopulation dynamics models for risk assessments under climate change. Diversity and Distributions, 19, 541–554.

Negrin, M.A. & Vazquez–Polo, F.–J. (2008) Incorporating model uncertainty in cost–effectiveness analysis: A Bayesian model averaging approach. Journal of Health Economics, 27, 1250–1259.

Negrin, M.A., Vazquez–Polo, F.J., Martel, M., Moreno, E. & Giron, F.J. (2010) Bayesian Variable Selection in Cost–Effectiveness Analysis. International Journal of Environmental Research and Public Health, 7, 1577–1596.

Neto, E.A., Rodrigues, M.A. & Odloak, D. (2000) Robust predictive control of a gasoline debutanizer column. Brazilian Journal of Chemical Engineering, 17, 967–977.

Neuman, S.P. (2003) Maximum likelihood Bayesian averaging of uncertain model predictions. Stochastic Environmental Research and Risk Assessment, 17, 291–305.

Neuman, S.P., Xue, L., Ye, M. & Lu, D. (2012) Bayesian analysis of data–worth considering model and parameter uncertainties. Advances in Water Resources, 36, 75–85.

Nguefack–Tsague, Georges. On optimal weighting scheme in model averaging. American Journal of Applied Mathematics and Statistics 2.3 (2014): 150–156.

Oaks, J.R., Linkem, C.W. & Sukumaran, J. (2014) Implications of uniformly distributed, empirically informed priors for phylogeographical model selection: A reply to Hickerson et al. Evolution, 68, 3607–3617.

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

Paeth, H. & Hense, A. (2002) Sensitivity of climate change signals deduced from multi–model Monte Carlo experiments. Climate Research, 22, 189–204.

Palmer, T.N., Shutts, G.J., Hagedorn, R., Doblas–Reyes, F.J., Jung, T. & Leutbecher, M. (2005) Representing model uncertainty in weather and climate prediction. Annual Review of Earth and Planetary Sciences, 33, 163–193.

Pan, W., Xiao, G.H. & Huang, X.H. (2006) Using input dependent weights for model combination and model selection with multiple sources of data. Statistica Sinica, 16, 523–540.

Pan, Z., Andrade, D., Segal, M., Wimberley, J., McKinney, N. & Takle, E. (2010) Uncertainty in future soil carbon trends at a central US site under an ensemble of GCM scenario climates. Ecological Modelling, 221, 876–881.

Park, I. & Grandhi, R.V. (2014) A Bayesian statistical method for quantifying model form uncertainty and two model combination methods. Reliability Engineering & System Safety, 129, 46–56.

Parrish, M.A., Moradkhani, H. & DeChant, C.M. (2012) Toward reduction of model uncertainty: Integration of Bayesian model averaging and data assimilation. Water Resources Research, 48, W03519.

Parsa, S., Ccanto, R., Olivera, E., Scurrah, M., Alcazar, J. & Rosenheim, J.A. (2012) Explaining Andean Potato Weevils in Relation to Local and Landscape Features: A Facilitated Ecoinformatics Approach. Plos One, 7, e36533.

Pena, D. & Redondas, D. (2006) Bayesian curve estimation by model averaging. Computational Statistics & Data Analysis, 50, 688–709.

Pesaran, M.H., Schleicher, C. & Zaffaroni, P. (2009) Model averaging in risk management with an application to futures markets. Journal of Empirical Finance, 16, 280–305.

Peters, G.W., Shevchenko, P.V. & Wuethrich, M.V. (2009) Model Uncertainty in Claims Reserving Within Tweedie's Compound Poisson Models. Astin Bulletin, 39, 1–33.

Picard, N., Henry, M., Mortier, F., Trotta, C. & Saint–Andre, L. (2012) Using Bayesian Model Averaging to Predict Tree Aboveground Biomass in Tropical Moist Forests. Forest Science, 58, 15–23.

Planque, B., Loots, C., Petitgas, P., Lindstrom, U. & Vaz, S. (2011) Understanding what controls the spatial distribution of fish populations using a multi–model approach. Fisheries Oceanography, 20, 1–17.

Platts, P.J., McClean, C.J., Lovett, J.C. & Marchant, R. (2008) Predicting tree distributions in an East African biodiversity hotspot: model selection, data bias and envelope uncertainty. Ecological Modelling, 218, 121–134.

Posada, D. & Buckley, T.R. (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. & Galmarini, S. (2009) Est modus in rebus: analytical properties of multi–model ensembles. Atmospheric Chemistry and Physics, 9, 9471–9489.

Powell, J.R. (2012) Accounting for uncertainty in species delineation during the analysis of environmental DNA sequence data. Methods in Ecology and Evolution, 3, 1–11.

Prein, A.F., Gobiet, A. & Truhetz, H. (2011) Analysis of uncertainty in large scale climate change projections over Europe. Meteorologische Zeitschrift, 20, 383–395.

Prost, L., Makowski, D. & Jeuffroy, M.–H. (2008) Comparison of stepwise selection and Bayesian model averaging for yield gap analysis. Ecological Modelling, 219, 66–76.

Pulkkinen, H. & Mantyniemi, S. (2013) Maximum survival of eggs as the key parameter of stock–recruit meta–analysis: accounting for parameter and structural uncertainty. Canadian Journal of Fisheries and Aquatic Sciences, 70, 527–533.

Raftery, A.E. (1996) Approximate Bayes factors and accounting for model uncertainty in generalised linear models. Biometrika, 83, 251–266.

Raftery, A.E., Gneiting, T., Balabdaoui, F. & Polakowski, M. (2005) Using Bayesian model averaging to calibrate forecast ensembles. Monthly Weather Review, 133, 1155–1174.

Raftery, A.E., Karny, M. & Ettler, P. (2010) Online Prediction Under Model Uncertainty via Dynamic Model Averaging: Application to a Cold Rolling Mill. Technometrics, 52, 52–66.

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

Raisanen, J. & Ylhaisi, J.S. (2012) Can model weighting improve probabilistic projections of climate change? Climate Dynamics, 39, 1981–1998.

Raisanen, J., Ruokolainen, L. & Ylhaisi, J. (2010) Weighting of model results for improving best estimates of climate change. Climate Dynamics, 35, 407–422.

Rapacciuolo, G., D. B. Roy, et al. (2012) Climatic Associations of British Species Distributions Show Good Transferability in Time but Low Predictive Accuracy for Range Change. Plos One 7(7)

Rapach, D.E. & Strauss, J.K. (2012) Forecasting US state–level employment growth: An amalgamation approach. International Journal of Forecasting, 28, 315–327.

Rasouli, S. & Timmermans, H.J.P. (2014) Using ensembles of decision trees to predict transport mode choice decisions: Effects on predictive success and uncertainty estimates. European Journal of Transport and Infrastructure Research, 14, 412–424.

Regehr, E.V., Hunter, C.M., Caswell, H., Amstrup, S.C. & Stirling, I. (2010) Survival and breeding of polar bears in the southern Beaufort Sea in relation to sea ice. Journal of Animal Ecology, 79, 117–127.

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

Ridley, J., Lowe, J., Brierley, C. & Harris, G. (2007) Uncertainty in the sensitivity of Arctic sea ice to global warming in a perturbed parameter climate model ensemble. Geophysical Research Letters, 34, L19704.

Rixen, M., Book, J.W., Carta, A., Grandi, V., Gualdesi, L., Stoner, R., Ranelli, P., Cavanna, A., Zanasca, P., Baldasserini, G., Trangeled, A., Lewis, C., Trees, C., Grasso, R., Giannechini, S., Fabiani, A., Merani, D., Berni, A., Leonard, M., Martin, P., Rowley, C., Hulbert, M., Quaid, A., Goode, W., Preller, R., Pinardi, N., Oddo, P., Guarnieri, A., Chiggiato, J., Carniel, S., Russo, A., Tudor, M., Lenartz, F. & Vandenbulcke, L. (2009) Improved ocean prediction skill and reduced uncertainty in the coastal region from multi–model super–ensembles. Journal of Marine Systems, 78, S282–S289.

Roberts, D. R. and A. Hamann (2012) Method selection for species distribution modelling: are temporally or spatially independent evaluations necessary? Ecography 35(9): 792–802.

Roberts, D. R., S. E. Nielsen, and G. B. Stenhouse (2014) Idiosyncratic responses of grizzly bear habitat to climate change based on projected food resource changes Ecological Applications 24:5, 1144–1154

Rodriguez-Rey, M., A. Jimenez–Valverde, et al. (2013) Species distribution models predict range expansion better than chance but not better than a simple dispersal model. Ecological Modelling 256: 1–5.

Rodriguez-Soto, C., O. Monroy-Vilchis, et al. (2011) Predicting potential distribution of the jaguar (Panthera onca) in Mexico: identification of priority areas for conservation. Diversity and Distributions 17(2): 350–361.

Rohner, B., Bugmann, H. & Bigler, C. (2013) Estimating the age-diameter relationship of oak species in Switzerland using nonlinear mixed-effects models. European Journal of Forest Research, 132, 751–764.

Rondina, F. (2012) The role of model uncertainty and learning in the US postwar policy response to oil prices. Journal of Economic Dynamics & Control, 36, 1009–1041.

Rotter, R.P., Palosuo, T., Kersebaum, K.C., Angulo, C., Bindi, M., Ewert, F., Ferrise, R., Hlavinka, P., Moriondo, M., Nendel, C., Olesen, J.E., Patil, R.H., Ruget, F., Takac, J. & Trnka, M. (2012) Simulation of spring barley yield in different climatic zones of Northern and Central Europe: A comparison of nine crop models. Field Crops Research, 133, 23–36.

Roura–Pascual, N., L. Brotons, et al. (2009) Consensual predictions of potential distributional areas for invasive species: a case study of Argentine ants in the Iberian Peninsula. Biological Invasions 11(4): 1017–1031.

Sahlin, U., Smith, H.G., Edsman, L. & Bengtsson, G. (2010) Time to establishment success for introduced signal crayfish in Sweden – a statistical evaluation when success is partially known. Journal of Applied Ecology, 47, 1044–1052.

Samore, A.B., Canavesi, F., Rossoni, A. & Bagnato, A. (2012) Genetics of casein content in Brown Swiss and Italian Holstein dairy cattle breeds. Italian Journal of Animal Science, 11, 196–202.

Sanderson, B.M. (2013) On the estimation of systematic error in regression–based predictions of climate sensitivity. Climatic Change, 118, 757–770.

Schmidt-Lebuhn, A.N., Knerr, N.J. & González-Orozco, C.E. (2012) Distorted perception of the spatial distribution of plant diversity through uneven collecting efforts: the example of Asteraceae in Australia. Journal of Biogeography, 39, 2072–2080.

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

Semenov, M.A. & Stratonovitch, P. (2010) Use of multi–model ensembles from global climate models for assessment of climate change impacts. Climate Research, 41, 1–14.

Shang, H.L. (2012) Point and interval forecasts of age–specific life expectancies: A model averaging approach. Demographic Research, 27, 593–643.

Sillanpaa, M. L and Corander, J. K (2002) Model choice in gene mapping: what and why. TRENDS in Genetics, 18, 301–307.

Singh, A., Mishra, S. & Ruskauff, G. (2010) Model Averaging Techniques for Quantifying Conceptual Model Uncertainty. Ground Water, 48, 701–715.

Sitters, H., Christie, F.J., Di Stefano, J., Swan, M., Penman, T., Collins, P.C. & York, A. (2014) Avian responses to the diversity and configuration of fire age classes and vegetation types across a rainfall gradient. Forest Ecology and Management, 318, 13–20.

Sloughter, J.M., Gneiting, T. & Raftery, A.E. (2013) Probabilistic Wind Vector Forecasting Using Ensembles and Bayesian Model Averaging. Monthly Weather Review, 141, 2107–2119.

Smith, A. B., M. J. Santos, et al. (2013) Evaluation of species distribution models by resampling of sites surveyed a century ago by Joseph Grinnell. Ecography 36(9): 1017–1031.

Smith, A.C., Koper, N., Francis, C.M. & Fahrig, L. (2009) Confronting collinearity: comparing methods for disentangling the effects of habitat loss and fragmentation. Landscape Ecology, 24, 1271–1285.

Smith, S.K. and Shahidullah, M. (1995) An evaluation of population projection errors for census tracts. Journal of the American Statistical Association 90(429): 64–71.

Song, S.O.K., Hogg, J., Peng, Z.–Y., Parker, R., Kellum, J.A. & Clermont, G. (2012) Ensemble Models of Neutrophil Trafficking in Severe Sepsis. Plos Computational Biology, 8, e1002422.

Stephenson, D.B., Collins, M., Rougier, J.C. & Chandler, R.E. (2012) Statistical problems in the probabilistic prediction of climate change. Environmetrics, 23, 364–372.

Stohlgren, T. J., P. Ma, et al. (2010) Ensemble Habitat Mapping of Invasive Plant Species. Risk Analysis 30(2): 224–235.

Strauch, M., Bernhofer, C., Koide, S., Volk, M., Lorz, C. & Makeschin, F. (2012) Using precipitation data ensemble for uncertainty analysis in SWAT streamflow simulation. Journal of Hydrology, 414, 413–424.

Sun, B. & Wang, H. (2013) Larger variability, better predictability? International Journal of Climatology, 33, 2341–2351.

Symonds, M.R.E. & Moussalli, A. (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.

Tang, G., Mayes, M.A., Parker, J.C., Yin, X.L., Watson, D.B. & Jardine, P.M. (2009) Improving parameter estimation for column experiments by multi–model evaluation and comparison. Journal of Hydrology, 376, 567–578.

Thomson, J. R., E. Fleishman, R. Mac Nally, and D. S. Dobkin. 2007. Comparison of predictor sets for species richness and the number of rare species of butterflies and birds. Journal of Biogeography 34:90–101.

Thomson, J.R., Mac Nally, R., Fleishman, E. & Horrocks, G. (2007) Predicting bird species distributions in reconstructed landscapes. Conservation Biology, 21, 752–766.

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

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

Tsai, F.T.–C. & Li, X. (2008a) Multiple Parameterization for Hydraulic Conductivity Identification. Ground Water, 46, 851–864.

Tsaparis, D., Katsanevakis, S., Ntolka, E. & Legakis, A. (2009) Estimating dung decay rates of roe deer (Capreolus capreolus) in different habitat types of a Mediterranean ecosystem: an information theory approach. European Journal of Wildlife Research, 55, 167–172.

Van Oijen, M., Reyer, C., Bohn, F.J., Cameron, D.R., Deckmyn, G., Flechsig, M., Harkonen, S., Hartig, F., Huth, A., Kiviste, A., Lasch, P., Makela, A., Mette, T., Minunno, F. & Rammer, W. (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.

Warren, D.L., Wright, A.N., Seifert, S.N. & Shaffer, H.B. (2014) Incorporating model complexity and spatial sampling bias into ecological niche models of climate change risks faced by 90 California vertebrate species of concern (ed J Franklin) Diversity and Distributions, 20, 334–343.

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

Weidemann, F., Dehnert, M., Koch, J., Wichmann, O. & Hoehle, M. (2014) Bayesian parameter inference for dynamic infectious disease modelling: rotavirus in Germany. Statistics in Medicine, 33, 1580–1599.

Weigel, A.P., Liniger, M.A. & Appenzeller, C. (2008) Can multi–model combination really enhance the prediction skill of probabilistic ensemble forecasts? Quarterly Journal of the Royal Meteorological Society, 134, 241–260. Wenger, S. J., N. A. Som, et al. (2013) Probabilistic accounting of uncertainty in forecasts of species distributions under climate change. Global Change Biology 19(11): 3343–3354.

Wilby, R.L. (2005) Uncertainty in water resource model parameters used for climate change impact assessment. Hydrological Processes, 19, 3201–3219.

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

Yen, H., Wang, X., Fontane, D.G., Harmel, R.D. & Arabi, M. (2014) A framework for propagation of uncertainty contributed by parameterization, input data, model structure, and calibration/validation data in watershed modeling. Environmental Modelling & Software, 54, 211–221.

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

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

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

Zhang, Z. & Townsend, J.P. (2009) Maximum–Likelihood Model Averaging To Profile Clustering of Site Types across Discrete Linear Sequences. Plos Computational Biology, 5, e1000421.

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

Zhou, B. & Du, J. (2010) Fog Prediction from a Multimodel Mesoscale Ensemble Prediction System. Weather and Forecasting, 25, 303–322.

Zhou, H. & Wu, Y. (2014) A Generic Path Algorithm for Regularized Statistical Estimation. Journal of the American Statistical Association, 109, 686–699.

Zhou, Y., Aston, J.A.D. & Johansen, A.M. (2013) Bayesian model comparison for compartmental models with applications in positron emission tomography. Journal of Applied Statistics, 40, 993–1016.

Zhu, J., Forsee, W., Schumer, R. & Gautam, M. (2013a) Future projections and uncertainty assessment of extreme rainfall intensity in the United States from an ensemble of climate models. Climatic Change, 118, 469–485.

Zhu, J., Huang, B., Balmaseda, M.A., Kinter, J.L., Peng, P., Hu, Z.–Z. & Marx, L. (2013b) Improved reliability of ENSO hindcasts with multi–ocean analyses ensemble initialization. Climate Dynamics, 41, 2785–2795.

Zhu, Y.J. (2005) Ensemble forecast: A new approach to uncertainty and predictability. Advances in Atmospheric Sciences, 22, 781–788.

Ziehmann, C. (2000) Comparison of a single–model EPS with a multi–model ensemble consisting of a few operational models. Tellus Series a–Dynamic Meteorology and Oceanography, 52, 280–299.

Zou, H.F., Xia, G.P., Yang, F.T. & Wang, H.Y. (2007) An investigation and comparison of artificial neural network and time series models for Chinese food grain price forecasting. Neurocomputing, 70, 2913–2923.