



Deposited via The University of York.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/id/eprint/127516/>

Version: Accepted Version

Article:

Dormann, Carsten F., Calabrese, Justin M., Guillera-Arroita, Gurutzeta et al. (2018) Model averaging in ecology:: a review of Bayesian, information-theoretic and tactical approaches for predictive inference. *Ecological Monographs*. pp. 485-504. ISSN: 0012-9615

<https://doi.org/10.1002/ecm.1309>

Reuse

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

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

Carsten F. Dormann^{*1}, Justin M. Calabrese², Gurutzeta Guillera-Arroita³, Eleni Matechou⁴, Volker Bahn⁵, Kamil Bartoń⁶, Colin M. Beale⁷, Simone Ciuti^{1,20}, Jane Elith³, Katharina Gerstner^{8,9}, Jérôme Guelat¹⁰, Petr Keil⁹, José J. Lahoz-Monfort³, Laura J. Pollock¹², Björn Reineking^{13,14}, David R. Roberts¹, Boris Schröder^{15,16}, Wilfried Thuiller¹², David I. Warton¹⁷, Brendan A. Wintle³, Simon N. Wood¹⁸, Rafael O. Wüest¹² & Florian Hartig^{1,19}

¹Biometry and Environmental System Analysis, University of Freiburg, Germany

²Smithsonian Conservation Biology Institute, Front Royal, USA

³School of BioSciences, University of Melbourne, Australia

⁴School of Mathematics, Statistics and Actuarial Science, University of Kent, UK

⁵Department of Biological Sciences, Wright State University, USA

⁶Department of Wildlife Conservation, Institute of Nature Conservation PAS, Kraków, Poland

⁷Department of Biology, University of York, UK

⁸Computational Landscape Ecology, Helmholtz Centre for Environmental Research-UFZ, Leipzig, Germany

⁹iDIV, Leipzig, Germany

¹⁰Swiss Ornithological Institute, Sempach, Switzerland

¹²Univ. Grenoble Alpes, Laboratoire d'Écologie Alpine (LECA), CNRS, Grenoble, France

¹³Univ. Grenoble Alpes, IRSTEA, UR EMGR, Grenoble, France

¹⁴Biogeographical Modelling, Bayreuth Center of Ecology and Environmental Research BayCEER, University of Bayreuth, Germany.

¹⁵Landscape Ecology and Environmental Systems Analysis, Institute of

Geocology, Technische Universität Braunschweig, Germany

¹⁶Berlin-Brandenburg Institute of Advanced Biodiversity Research (BBIB),
Berlin, Germany

¹⁷School of Mathematics and Statistics, Evolution & Ecology Research
Centre, University of New South Wales, Australia

¹⁸School of Mathematics, Bristol University, UK

¹⁹Group for Theoretical Ecology, University of Regensburg, Germany

²⁰Laboratory of Wildlife Ecology and Behaviour, School of Biology and
Environmental Science, University College Dublin, Ireland

December 6, 2017

Running head: Model averaging in ecology

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

Abstract

In ecology, the true causal structure for a given problem is often not known, and several plausible models exist. It has been claimed that using weighted averages of these models can reduce prediction error, as well as better reflect model selection uncertainty. However, a large range of different model averaging methods exists, raising the question of how they differ regarding these goals. A core question for an analyst is thus to understand under which circumstances model averaging can improve predictions and their uncertainty estimates.

Here we review the mathematical foundations of model averaging along with the diversity of approaches available. The terms contributing to error in model-averaged predictions are each model's bias (i.e. the deviation of each model prediction from the

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

12 unknown truth), variance of, and covariance among, model predictions, and
13 uncertainty of model weights.

14 If bias of contributing model predictions is substantially larger than their variance,
15 the advantage of reduced variance through weighted averages is greatly reduced. For
16 noisy data, which predominate in ecology, variance is probably often larger than bias
17 and model averaging becomes an option to reduce prediction error. Correlation
18 between model predictions also reduces the effect of model averaging, and to
19 counteract this effect, model weights could be adjusted to maximise the variance
20 reduction.

21 Model-averaging weights have to be estimated from the data, and this estimation
22 process carries some uncertainty, so that “optimised” model weights may not be better
23 than the use of arbitrary weights, such as equal weights for all models. In the presence
24 of inadequate models, however, estimating model weights is still likely to be superior
25 to equal weights. Many different methods to derive averaging weights exist, from
26 Bayesian over information-theoretical to optimised and resampling approaches, as
27 reviewed here.

28 We also investigate the coverage of the confidence interval of the prediction for
29 different ways to combine model prediction distributions, showing that they differ
30 greatly, and that the full model has very good coverage properties. Our overall
31 recommendations stress the importance of validation-based approaches and of
32 uncertainty quantification to avoid unreflected use of model averaging.

33 **1 Introduction**

34 Models are an integral part of ecological research, representing alternative, possibly
35 overlapping, hypotheses (Chamberlin, 1890). They are also the key approach to making

36 predictions about ecological systems (Mouquet et al., 2015). In many cases it is not
37 possible to clearly identify a single most-appropriate model. For instance,
38 process-based models may differ in the specific ways they represent ecological
39 mechanisms, but several different process models may accord with our ecological
40 understanding. Statistical models are limited in their complexity by the amount of data
41 available for fitting, making several combinations of predictors plausible, and different
42 modelling approaches are available for statistical analysis (e.g. Hastie et al., 2009; Kuhn
43 and Johnson, 2013).

44 Model averaging, as the weighted sum of predictions from several candidate
45 models, provides a potential avenue to avoid selecting a single model over others
46 similarly plausible. Scientists average model predictions for different reasons, most
47 prominently: (a) reducing prediction error through reduced variance, and partially by
48 (b) reducing prediction bias (based on arguments described in Madigan and Raftery,
49 1994), and (c) accommodating/quantifying uncertainty about model parametrisation
50 and structure (Wintle et al., 2003, see also section 2.3).

51 Here we focus on averaging sets of models that differ in structure, as opposed to
52 mere differences in initial conditions or parameter values (Gibbs, 1902; Johnson and
53 Bowler, 2009). The latter case in the statistical and physical literature is called
54 “ensemble”, while in ecology that term is used more loosely. For some ecological
55 examples of model averaging see Wintle et al. (2003); Thuiller (2004); Richards (2005);
56 Brook and Bradshaw (2006); Dormann et al. (2008); Diniz-Filho et al. (2009); Le Lay
57 et al. (2010); Garcia et al. (2012); Cariveau et al. (2013); Meller et al. (2014), and Lauzeral
58 et al. (2015).

59 Several previous publications have reviewed model averaging in ecology and
60 evolution, focussing exclusively on ‘information-theoretical model averaging’
61 (Johnson and Omland, 2004; Hobbs and Hilborn, 2006; Burnham et al., 2011; Freckleton,

2011; Grueber et al., 2011; Nakagawa and Freckleton, 2011; Richards et al., 2011; Symonds and Moussalli, 2011), probably under the influence of the AIC-weighted averaging popularised by Burnham & Anderson (2002; Posada and Buckley 2004). *Bayesian* model averaging has been treated less frequently in ecology (for an example see Corani and Mignatti, 2015), but for an excellent recent review of this topic in the context of Bayesian model selection see Hooten and Hobbs (2015, see also Hoeting et al. 1999; Ellison 2004; Link and Barker 2006). However, none of the above is a comprehensive review of the state of knowledge across the available model averaging approaches.

Our aim is to provide such a comprehensive review in the light of developments over the last 20 or so years, summarising the actual mathematical reasoning and offering an intuitive as well as technical entry, illustrated by case studies. We primarily address averaging of predictions from correlative models, although most of the points will similarly apply to mechanistic/process-based models (see, e.g., Knutti et al., 2010; Diks and Vrugt, 2010, for a review in the context of climate and hydrological models, respectively). We do not concentrate on averaging model parameters, because we agree with the criticism summarised in Banner and Higgs (2017): parameters are estimated conditional on the model structure; as the model structure changes, parameters may become incommensurable (see Posada and Buckley, 2004; Cade, 2015; Banner and Higgs, 2017, and Appendix S1 for short review of the parameter-averaging literature).

This review is divided into two parts: theoretical and practical. In the first we present the mathematical logic behind model averaging, and why this alone puts severe constraints on *how* we do model averaging. Then, in the second part, we review the different ways model-averaging weights can be derived, comparing Bayesian, information-theoretic and other tactical perspectives (i.e. those not derived from statistical theory but still with a clear objective). This is followed by a brief exploration

88 of how to quantify model-averaged prediction uncertainty. We briefly illustrate model
89 averaging with two case studies, before closing with unresolved challenges and
90 recommendations.

91 **2 The mathematics behind model averaging**

92 Model averaging refers to the computation of a weighted-average prediction \tilde{Y} based
93 on the predictions of several (M) contributing models, $\hat{Y}_1, \hat{Y}_2, \dots, \hat{Y}_M$:

$$\tilde{Y} = \sum_{m=1}^M w_m \hat{Y}_m, \quad \text{with} \quad \sum_{m=1}^M w_m = 1. \quad (1)$$

94 Conceptually, the role of weights w_m is to adjust predictions such that the average
95 prediction has improved properties over selecting a single among a number of
96 candidate models (for example, less bias, lower variance or closer-to-nominal coverage).
97 In accordance with virtually all applications of model averaging we encountered, we
98 first focus on how model averaging reduces prediction error, here quantified as mean
99 squared error (MSE) of a prediction \hat{Y}_m of model m , which is composed of prediction
100 bias and prediction variance:

$$\text{MSE}(\hat{Y}_m) = \left\{ \text{bias}(\hat{Y}_m) \right\}^2 + \text{var}(\hat{Y}_m). \quad (2)$$

101 We shall now decompose this equation to understand what contributes to prediction
102 error in the context of model averaging.

103 Bias, i.e. the difference between the prediction expectation and the truth (y^*), will
104 depend directly on the bias of the contributing models, as well as their weights (eqn 1).
105 As the truth is unknown (except in simulations), the statistical model-averaging
106 literature typically makes the assumption that individual models have no bias (Bates
107 and Granger, 1969; Buckland et al., 1997; Burnham and Anderson, 2002). In contrast,
108 the focus of averaging process models is primarily on removing bias (e.g. Solomon
109 et al., 2007; Gibbons et al., 2008; Dietze, 2017).

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

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

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

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

110 For the average of two predictions, \widehat{Y}_1 and \widehat{Y}_2 , we have:

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

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

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

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

113 Putting eqns 2 and 3 together we get:

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

114 where $E(\widehat{Y}_m) - y^* = \text{bias}(\widehat{Y}_m)$ represents model misspecification bias.

115 2.1 Influences on the error of model-averaged prediction

116 Equation 5 allows us to make a number of statements about the potential benefits of
 117 model averaging. Firstly, bias will typically remain unknown, as truth y^* is unknown,
 118 but it can be estimated through (cross-)validation, and hence also the relative
 119 importance of bias to variability of predictions can be quantified (Fig. 1). When each
 120 model produces a distinct prediction, with variances substantially lower than

121 differences between models, then bias dominates (Fig. 1 top). As variance increases (or
122 bias decreases), the different model predictions overlap more and more, until bias is
123 small relative to variance (Fig. 1 bottom). Predictions from any model will now
124 typically have higher variance than the averaged prediction. Also, averaging can
125 reduce bias, if predictions scatter around the truth, but not unidirectional bias, that is if
126 all (most) model predictions err on the same side (see also Fig. 2 top row). However, if
127 predictions scatter around the truth, bias can be reduced by averaging.

128 [Fig. 1 approximately here.]

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

132 [Fig. 2 approximately here.]

133 Downweighting of variances is the mathematical reason how model averaging
134 reduces the variance over single model predictions. In the unlikely, but didactically
135 important case that predictions are independent, their covariance is 0 and the
136 correlation matrix ρ_{mn} of eqn 5 becomes the identity matrix (or, equivalently, the
137 covariance term of eqn 4 vanishes). If we also assume both predictions have equal
138 variances ($\text{var}(\hat{Y}_1) = \text{var}(\hat{Y}_2) = \text{var}(\hat{Y})$), and since $w_2 = 1 - w_1$, the above equation
139 simplifies to $\text{var}(\tilde{Y}) = (2w_1^2 - 2w_1 + 1)\text{var}(\hat{Y})$. If one model gets all the weight, we
140 have $\text{var}(\tilde{Y}) = \text{var}(\hat{Y})$. If the two models receive equal weight, we have
141 $\text{var}(\tilde{Y}) = (2 \cdot (0.5)^2 - 2 \cdot 0.5 + 1)\text{var}(\hat{Y}) = 0.5\text{var}(\hat{Y})$, a considerable improvement
142 in prediction variance (and the minimum of this equation). Other weights fall
143 in-between these values. More generally, Bates and Granger (1969) showed that for
144 unbiased models with uncorrelated predictions, the variance in the average is never
145 greater than the smaller of the individual predictions (making the important
146 assumption that the weights are known, which will be discussed below). In other

147 words, model averaging can reduce prediction error because weights enter as quadratic
148 terms in eqn 3, rather than linearly.

149 The correlation between model predictions, i.e. the matrix $(\rho_{ij}) \in \mathbb{R}^{M \times M}$,
150 substantially affects the benefit of model averaging (see also Fig. 3 and interactive tool
151 in the Appendix Data S2). In the best case, correlations between model predictions are
152 negative or at least absent, and the second term of eqn (5) is negative or vanishes. Here,
153 the variance in the average is dominated by individual models' prediction variances. As
154 correlation between predictions increases, the covariance-term contributes more and
155 more to the overall prediction error, making the averaging of perfectly correlated
156 predictions exactly outweigh the benefit gained by the quadratic weights-effect for the
157 variances.

158 [Fig. 3 approximately here.]

159 This point provides some important insights about why some machine learning
160 methods that average a large number of bad models work so well. When averaging *poor*
161 models, e.g. trees in a randomForest, covariance is negligible, but the variance of each
162 model prediction is high. Because w_m becomes very small with hundreds of models
163 (around $1/M$), the variance of many averaged poor models (with similar variance)
164 tends to be low: $\text{var}(\tilde{Y}) = \sum_{m=1}^M \frac{1}{M^2} \text{var}(\hat{Y}_m) + \frac{1}{M^2} \sum_{m=1}^M \sum_{m \neq n} \text{cov}(\hat{Y}_m, \hat{Y}_n) \approx$
165 $M \frac{1}{M^2} \text{var}(\hat{Y}) = \frac{1}{M} \text{var}(\hat{Y})$, where the second term disappears due to lack of
166 correlations among predictions. We may speculate that poor models typically also
167 exhibit substantial but undirected bias, which again would be reduced by averaging.

168 The effect of correlations in the potential reduction of prediction error is rather
169 intuitive. If a prediction from a given model is extreme (e.g. on the high end of the
170 distribution), negative correlation will tend to balance out, while positive correlation
171 will accentuate total variance (e.g. Bohn et al., 2010). Ecologists know an analogous
172 effect from biodiversity studies, where it is called the 'portfolio effect'

173 (e.g. Thibaut and Connolly, 2013). It states that the fluctuation in biomass of a
174 community is less than the fluctuations of biomass of its members, because the species
175 respond to the environment differently. This asynchrony in response is analogous to
176 negative covariance in community members' biomass, buffering the *sum* of their
177 biomasses.

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

183 Thus, for a given set of weights, the prediction error of model-averaged predictions
184 depends on three things: the bias of the model average, the individual model prediction
185 variances, and the correlation between individual model predictions.

186 **2.2 Estimating weights can thwart the benefit of model** 187 **averaging**

188 Equation 5 assumes that the values of the weights are set a priori, and thus there is no
189 uncertainty about them. However, that would imply that an arbitrary set of weights is
190 used. Instead, the aim of optimising predictive performance suggests weights need to
191 be estimated from the data. But estimation brings associated uncertainty with it, and
192 this has implications for the actual benefits of model averaging: estimated “optimal”
193 weights will be suboptimal (Nguefack-Tsague, 2014), so the averaged prediction even
194 for only mildly correlated predictions will more likely be biased than if the (unknown)
195 truly optimal weights were used (Claeskens et al., 2016). It may in fact be often no
196 better than one obtained using some arbitrary weights, e.g. equal weights (Clemen,

197 1989; Smith et al., 2009; Graefe et al., 2014, 2015). The “simple theoretical explanation”
198 provided by Claeskens et al. (2016) demonstrates that estimating weights introduces
199 additional variance into the prediction. As a consequence, the predictions averaged
200 with estimated weights may be worse than that of a single model (in contrast to the
201 assertion of Bates and Granger 1969; see Claeskens et al. 2016 for an example).

202 Finding optimal weights now becomes far more involved, and currently no closed
203 solution is available, not even for linear models (Liang et al., 2011). The interactive tool
204 we provide (Fig. 3) allows readers to explore this issue in a simple 2-model case. It
205 shows that, in this simple case, estimating weights substantially reduces the parameter
206 space where model averaging is superior to the best single model.

207 The performance reduction does not however imply that estimated weights are of
208 no use, or that the use of arbitrary weights (e.g. equal weights) is generally superior.
209 While uncertainty in estimated weights increases prediction error, the ability to
210 downweight or wholly remove unsuitable models from the prediction set is a
211 substantial benefit. In Claeskens et al. (2016) and similar simulations, all models
212 considered are “alright” (bias-free and with similar prediction variance), which
213 obviously need not be the case. Model weights are a measure of suitability for
214 prediction, which can be derived most logically from validation on (semi-)independent
215 data (see section 3 for details). If the unknown optimal model weights deviate strongly
216 from $1/m$, their estimation uncertainty is then a price worth paying.

217 **2.3 Model averaging (typically) reduces prediction errors**

218 The majority of studies we encountered (as random draws from the results of a
219 systematic literature search: see Appendix S7) used an empirical approach to assess
220 predictive performance, i.e. forecasting, hindcasting or cross-validation to observed
221 data (e.g. Namata et al., 2008; Marmion et al., 2009*a,b*; Grenouillet et al., 2010;

222 Montgomery et al., 2012; Smith et al., 2013; Engler et al., 2013; Edeling et al., 2014;
223 Trolle et al., 2014). Across the 180 studies we examined, model averaging generally
224 yielded lower prediction errors than the individual contributing models. Most of these
225 studies used test datasets to estimate predictive success, and rely critically on the
226 assumption of independence between test and training datasets (Roberts et al., 2017).
227 Few studies used simulated data to examine the performance of model averaging under
228 specific conditions (e.g. small sample size, model structure uncertainty, missing data:
229 Ghosh and Yuan, 2009; Schomaker, 2012). Very few studies provide mathematical
230 analyses (Shen and Huang, 2006; Potempski and Galmarini, 2009; Chen et al., 2012;
231 Zhang et al., 2013).

232 Summarising section 2 so far, we observe that

- 233 1. model averaging reduces prediction error by reducing prediction variance and
234 bias;
- 235 2. the more positively correlated predictions are, the smaller is the benefit gained
236 from averaging them;
- 237 3. when bias is large relative to the prediction variance of individual models, the
238 least-biased model will be a better choice than the model average; and
- 239 4. estimating weights introduces additional variance, outweighing, in some
240 situations, the benefits of model averaging.

241 **2.4 Quantifying uncertainty of model-averaged**

242 **predictions**

243 In random sampling, in addition to a statistic of interest, say a point prediction, we are
244 typically interested in the uncertainty of this statistic, e.g. as quantified by its variance
245 (goal 2 at the beginning of the paper). A relevant question is whether the associated

246 confidence intervals have nominal coverage, i.e. whether the true value is in the 95%-CI
247 indeed 95% of the time in repeated experiments.

248 If we attempt an analogy between random sampling and model averaging, the first
249 catch is that predictions from different models will be non-independent. In this case the
250 standard deviation does not decrease as square root of n , but more slowly. The second
251 catch is that models are almost certainly not random draws from the population of
252 models (if we just think of all the models which we did *not* include). Non-random
253 draws from a distribution are almost certain to yield biased estimates of that
254 distribution's parameters.

255 The first catch can be taken care of by taking into account the variance-covariance
256 matrix of model predictions (see section 2, eqns 3-5). The second catch (models are
257 non-random draws) is harder and the severity of this problem depends on whether
258 model predictions are biased in the same direction (the "unidirectional bias" in Fig. 2)
259 or in different ways. Model averaging can only successfully unite diverging biased
260 predictions when they are biased in different directions. The approaches to computing
261 prediction variance below rely on the assumption that model predictions in fact do
262 scatter around the truth, and that the (weighted) average of model predictions is
263 unbiased. Since truth is unknown, this assumption cannot be tested. When models
264 share their fundamental structure (e.g. process models relying on the same equations),
265 it is more likely that they are unidirectionally biased.

266 **2.4.1 Simplified error propagation in model-averaged predictions**

267 To approximate the predictive variance of model-averaged predictions, Buckland et al.
268 (1997) proposed a simplification of eqn (5) (for derivation see Burnham and Anderson,

269 2002, p. 159-162):

$$\text{var}(\tilde{Y}) = \left(\sum_{m=1}^M w_m \sqrt{\text{var}(\hat{Y}_m) + \gamma_m^2} \right)^2. \quad (6)$$

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

282 Improving on eqn (6) requires knowledge of the correlation matrix ρ_{mn} of eqn (5).
283 The key problem is that there is no analytical way to compute the correlation of model
284 predictions. While bootstrapping models and their prediction can provide an estimate
285 of ρ_{mn} , it can more directly provide an estimate of $\text{var}(\tilde{Y})$, rendering the indirect route
286 via eqn (6) unnecessary.

287 **2.4.2 Coverage of the model-averaged prediction**

288 Predictions from a selected single-best model *always* underestimate the true prediction
289 error (e.g. Namata et al., 2008; Fletcher and Turek, 2012; Turek and Fletcher, 2012). The
290 reason is that the uncertainty about which model is correct is not included in this final
291 prediction: we predict as if we had not carried out model selection but had known from
292 the beginning which model would be the best (as if the model had been “prescribed”:

293 Harrell, 2001). Thus, even if we were able to choose, from our model set M , the model
294 closest to truth, we would still need to adjust the confidence distribution for model
295 selection; however, a perfect adjustment was analytically shown not to exist (Kabaila
296 et al., 2015).

297 For statistical models, it is less clear whether the full model (i.e. prior to any model
298 selection; see Appendix S3) or model averaging computes the uncertainty intervals
299 correctly. Simulations suggest that model averaging may improve coverage (Namata
300 et al., 2008; Wintle et al., 2003; Zhao et al., 2013, none of who tested the full model),
301 which can be understood to happen because the process of averaging allows us to take
302 into account model uncertainty (Liang et al., 2011). Given that model averages need not
303 be normal (at the link scale), Fletcher and Turek (2012) and Turek and Fletcher (2012)
304 explore how to improve the tail areas of the confidence distribution, albeit under the
305 assumption that the true model is in the model set. Their approach was re-analysed by
306 Kabaila et al. (2015) under model selection. The key finding of this latter study is that
307 the full model coverage was still superior to all other model averaging approaches,
308 suggesting that the full model should currently be kept in mind, both for inference,
309 minimal bias and correct prediction intervals (see also Harrell, 2001, p. 59). Such
310 findings sit uncomfortably with the bias-variance trade-off (Hastie et al., 2009), which
311 states that overly complex models have poor predictive performance; and indeed the
312 full model has high prediction variance. However, our statements are about the
313 confidence intervals, rather than the point predictions, and those will be incorrectly
314 narrow for model selection without selection-correction. Regrettably, such reasoning
315 cannot be extended in an obvious way to models that do not have a “full model”
316 (non-nested models, process models, or machine learning models). Here model
317 averaging provides a way forward in representing prediction coverage more fairly.

318 Given the diversity of approaches to computing model weights encountered in

319 section 3, these studies cannot be seen as conclusive, only as suggestive, for the
320 improvement of nominal coverage using model averaging.

321 In a different approach to characterising the uncertainty in model predictions, model
322 averaging can be interpreted as computing the distribution of a random variable that is
323 derived from a collection of random variables (the model predictions), also known as a
324 **mixture distribution** (Claeskens and Hjort, 2008, p. 217). In a two-step process, the
325 model weights determine the probability of choosing the model, and then the model
326 prediction is drawn from its confidence distribution. If predictions are unbiased, they
327 stack up high around the mean, and yield the same value as the equation for the
328 standard error of the mean. If predictions differ widely, e.g. due to bias, the mixed
329 confidence distribution will be much wider and possibly multi-modal. Mixing
330 distributions assumes their independence, i.e. the random draw of a value from one
331 model prediction is uncorrelated with the next draw of model and prediction. As model
332 predictions are likely to be positively correlated, assuming (conditional) independence
333 will *underestimate* variance (i.e. correlated draws would yield wider confidence
334 distributions).

335 Overall, this leaves us with the following options for computing the confidence
336 intervals of averaged predictions (which we will compute for a set of simple linear
337 regressions in Fig. 5):

- 338 1. Make the assumption that model-averaged predictions are unbiased (i.e. that y^*
339 can be estimated as \tilde{Y}). Use bootstrapping to estimate covariances of predictions
340 for each model. From these estimates, compute prediction variance according to
341 eqn (5). This solution is computer-intensive, but it takes into account covariance
342 of model predictions. (Note that simply averaging predictions from bootstrapped
343 models is not correct, as it does not incorporate model misspecification bias.)

- 344 2. Make again the assumption that model-averaged predictions are unbiased. Use
345 Buckland et al. (1997)'s approach (eqn 6). This will yield wider estimates than
346 option 1, because assuming perfect correlation is conservative.
- 347 3. Make the assumption that predictions from different models are effectively
348 uncorrelated. Use model mixing to compute the confidence distribution of the
349 average.
- 350 4. Fit the full model (if available) and use its confidence distribution, which can
351 rarely be improved on (Kabaila et al., 2015).

352 [Figure 5 approximately here.]

353 When averaging models with largely independent (i.e. uncorrelated) predictions,
354 only the bootstrap-estimated covariance matrix (option 1 above) will also compute
355 lower variances (according to eqn 4). In our illustration (Fig. 5, see Appendix S8), the
356 first three options ("propagation", "Buckland" and "mixing") hardly differ, while the full
357 model has a different location and is wider. The coverage of the 95% confidence
358 interval, computed through 1000 simulations, is best matched by the full model, while
359 the propagation approach is overly conservative. Buckland's equation and mixing have
360 slightly too low coverage.

361 **3 Approaches to estimating model-averaging**

362 **weights**

363 When faced with predictions from very different models, estimating weights aims at
364 abating poorly, and elevating well predicting ones. For the resulting averaged
365 predictions, the actual method for estimating weights has obvious fundamental
366 importance. We now review approaches to estimate model-averaging weights and

367 elucidate on their interconnections (Table 1). Different perspectives on
368 model-averaging weights have emerged, which we present in somewhat arbitrary four
369 categories of decreasing probabilistic interpretability:

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

381 We shall address these four perspectives in turn, also hinting at relationships
382 between them.

383 [Table 1 approximately here.]

384 **3.1 Bayesian model weights**

385 Our outline of Bayesian model weights follows that of Wasserman (2000), paying
386 attention to recent computational advances in the field.

387 **Theory** Bayes’ formula can be applied to models in much the same way as to
388 parameters. Hence, to perform inference with multiple models, one can write down the
389 joint posterior probability $P(M_i, \Theta_i | D)$ of model M_i with parameter vectors Θ_i , given
390 the observed data D , as

$$P(M_i, \Theta_i | D) \propto L(D | M_i, \Theta_i) \cdot p(\Theta_i) \cdot p(M_i), \quad (7)$$

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

393 The joint distribution provides all information necessary for inference. Often, in
 394 practice, we want to extract some simplified statistics from this distribution such as the
 395 model with the highest posterior model probability, or the distribution of a parameter
 396 or prediction including model selection uncertainty. To obtain this information, we can
 397 marginalise (average, integrate) over parameter space, or marginalise over model space.

398 If we marginalise over parameter space, we obtain model weights (whilst
 399 marginalising over model space yields averaged parameters, which we shall not
 400 address here). The first step is to calculate the marginal likelihood, defined as the
 401 average of eqn (7) across all k parameters for any given model:

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

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

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

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

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

408 **Estimation in practice** While the definition of Bayesian model weights and
 409 averaged parameters is straightforward, the estimation of these quantities can be

410 challenging. In practice, there are two options to numerically estimate the quantities
411 defined above, both with caveats.

412 The first option is to sample directly from the joint posterior (eqn (7)) of the models
413 and the parameters. Basic algorithms such as rejection sampling can do that without
414 any modification (e.g. Toni et al., 2009), but they are inefficient for higher-dimensional
415 parameter spaces. More sophisticated algorithms such as MCMC and SMC (see Hartig
416 et al., 2011, for a basic review) require modifications to deal with the issue of different
417 number of parameters when changing between models. Such modifications (mostly the
418 reversible-jump MCMCs, **rjMCMC**: Green, 1995, see Appendix S5.1.1) are often
419 difficult to program, tune and generalise, which is the reason why they are typically
420 only applied in specialised, well-defined settings. The posterior model probabilities of
421 the rjMCMC are estimated as the proportion of time the algorithm spent with each
422 model, measured as the number of iterations the algorithm drew a particular model
423 divided by the total number of iterations.

424 The second option is to approximate the marginal likelihood in eqn (8) of each
425 model independently e.g. compute the maximum a posteriori model probability,
426 renormalise that into weights, and then average predictions based on these weights.
427 The challenge here is to get a stable approximation of the marginal likelihood, which
428 can be very problematic (Weinberg, 2012, see Appendix S5.1.2). Because of the
429 relatively simple implementation, this approach is a more common choice than
430 rjMCMC (e.g. Brandon and Wade, 2006).

431 **Influence of priors** A problem for the computation of model weights when
432 performing Bayesian inference across multiple models, is the influence of the choice of
433 *parameter priors*, especially “uninformative” ones (see section 5 in Hoeting et al., 1999;
434 Chickering and Heckerman, 1997).

435 The challenge arises because in eqns (8) and (9) the prior density $p(\theta_i)$ enters the
436 marginal likelihood and hence the Bayes factor multiplicatively. This has the somewhat
437 unintuitive consequence that increasing the width of an uninformative parameter prior
438 will linearly decrease the model's marginal likelihood (e.g. Link and Barker, 2006).
439 That Bayesian model weights are strongly dependent on the width of the prior choice
440 has sparked discussion of the appropriateness of this approach in situations with
441 uninformative priors. For example, in situations where multiple nested models are
442 compared, the width of the uninformative prior may completely determine the
443 complexity of models that are being selected. One suggestion that has been made is to
444 *not* perform multi-model inference *at all* with uninformative priors, or that at least
445 additional corrections are necessary to apply Bayes factors weights (O'Hagan, 1995;
446 Berger and Pericchi, 1996). One such correction is to calibrate the model on a part of the
447 data first, use the result as new priors and then perform the analysis described above
448 (intrinsic Bayes factor: Berger and Pericchi 1996, fractional Bayes factor: O'Hagan
449 1995). If sufficient data are available so that the likelihood is sufficiently peaked
450 strongly during the calibration step, this approach should eliminate any complication
451 resulting from the prior choice (for an ecological example see van Oijen et al., 2013).

452 **Bayesian variations** In a set of influential publications, Raftery et al. (1997),
453 Hoeting et al. (1999) and Raftery et al. (2005) introduced *post-hoc* Bayesian model
454 averaging, i.e. for vectors of predictions from already fitted models. The key idea is to
455 iteratively estimate the proportion of times a model would yield the highest likelihood
456 within the set of models (through expectation maximisation, see Appendix S5.2 for
457 details), and use this proportion as model weight. In the spirit of the inventors, we refer
458 to this approach as **Bayesian model averaging using Expectation-Maximisation**
459 (BMA-EM), but place it closer to a frequentist than a Bayesian approach, as the models

460 were not necessarily (and in none of their examples) fitted within the Bayesian
 461 framework. It has been used regularly, often for process models (e.g. Gneiting et al.,
 462 2005; Zhang et al., 2009), where a rjMCMC-procedure would require substantial
 463 programming work at little perceived benefit, but also in data-poor situations in the
 464 political sciences (Montgomery et al., 2012).

465 Chickering and Heckerman (1997) investigate approximations of the marginal
 466 likelihood in eqn (9), such as the **Bayesian Information Criterion** (BIC, as defined
 467 in the next section; see also Appendix S5.3) and find them to work well for model
 468 selection, but *not* for model averaging. In contrast, Kass and Raftery (1995) state (on
 469 p. 778) that e^{BIC} is an acceptable approximation of the Bayes factor, and hence suitable
 470 for model averaging, despite being biased even for large sample sizes. These
 471 approximations may be improved when using more complex versions of BIC (SPBIC
 472 and IBIC: Bollen et al., 2012).

473 The “widely applicable information criterion” **WAIC** (Watanabe 2010 and an
 474 equivalent **WBIC**: Watanabe 2013) are motivated and actually analytically derived in a
 475 Bayesian framework (Gelman et al., 2014). Its uninformative prior implementation
 476 should be seen as a variation of AIC (see next section), while the implementation with
 477 model priors is based on posterior distribution of parameter estimates, and computed,
 478 for each model, from two terms (Gelman et al., 2014): (1) the log pointwise predicted
 479 density (lppd) across the posterior simulations for each of the n predicted values,
 480 defined as $\text{lppd} = \log \prod_{i=1}^n p_{\text{posterior}}(y_i)$; and (2) a bias-correction term
 481 $p_{\text{WAIC}} = \sum_{i=1}^n \text{var}(\log(p(y_i|\theta_s)))$, where *var* is the *sample* variance over all S samples
 482 of the posterior distributions of parameters θ . Then the WAIC is defined as
 483 $\text{WAIC} = -2 \text{lppd} + 2 p_{\text{WAIC}}$. In words, the WAIC is the likelihood of observing the data
 484 under the posterior parameter distributions, corrected by a penalty of model
 485 complexity proportional to the variance of these likelihoods across the MCMC samples.

486 Model weights are computed from WAIC analogously to equation 11 below.

487 **3.2 Information-theoretic model weights**

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

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

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

495 where ℓ_m is the log-likelihood of model m .

496 In the ecological literature, AIC (and its sample-size corrected version AICc, and its
497 adaptations to quasi-likelihood models such as QIC: Pan 2001; Claeskens and Hjort
498 2008) is by far the most common approach to determine model weights (for recent
499 examples see, e.g., Dwyer et al., 2014; Rovai et al., 2015). **AIC-weights** (eqn (11)) have
500 been interpreted as Bayesian model probabilities (Burnham and Anderson 2002, p. 75;
501 Link and Barker 2006), although we are not aware of a convincing theoretical
502 justification. An alternative interpretation is the proportion of times a model would be
503 chosen as the best model under repeated sampling (Hobbs and Hilborn, 2006), but such
504 an interpretation is contentious (Richards, 2005; Bolker, 2008; Claeskens and Hjort,
505 2008). In an anecdotal comparison, Burnham and Anderson (2002, p. 178) showed that
506 AIC-weights are substantially different from **bootstrapped model weights**. The
507 latter were proposed by Buckland et al. (1997) and represent the proportion of
508 bootstraps a model is performing best in terms of AIC: see case study 1 below. In

509 simulations, AIC-weights did not reliably identify the model with the known lowest
510 KL-divergence or prediction error (Richards, 2005; Richards et al., 2011). Instead,
511 **Mallows’ model averaging** (MMA) has been shown to yield the lowest mean
512 squared error for *linear* models (Hansen, 2007; Schomaker et al., 2010). Mallows’ C_p
513 penalises model complexity equivalent to $-2\ell_m - n + 2p_m$ (for n data points; rather
514 than AIC’s $-2\ell_m + 2p_m$, eqn 11).

515 Other approximations of the KL-divergence include Schwartz’ Bayesian
516 Information Criterion (see previous section), which was designed to find the most
517 probable model given the data (Schwartz, 1978; Shmueli, 2010), equivalent to having
518 the largest Bayes factor (see previous section). **BIC** uses $\log(n)$ rather than AIC’s “2”
519 as penalisation factor for model complexity (Appendix S5.3). A particularly noteworthy
520 modification of the AIC exist, where the model fit is assessed with respect to a focal
521 predictor value, e.g. a specific age or temperature range, yielding the Focused
522 Information Criterion (FIC: Claeskens and Hjort 2008). We are not aware of a
523 systematic simulation study comparing the performance of these model averaging
524 weights, but AIC’s dominance should not indicate its superiority (see also case study 1
525 below).

526 The weighting procedure can additionally be wrapped into a cross-validation and
527 model pre-selection, which leads to the ARMS-procedure (**Adaptive Regression by**
528 **Mixing with model Screening**; Yang, 2001; Yuan and Yang, 2005; Yuan and Ghosh,
529 2008). We shall not present details on ARMS here (for cross-validation see next section),
530 because we regard model pre-selection as an unresolved issue (see section 5.3).

531 **3.3 Tactical approaches to computing model weights**

532 Methods covered in this section share the “tactical” goal of choosing weights to
533 optimise prediction (e.g. reduce prediction error). These weighting schemes are not

534 explicitly building on Bayes or information theory thus most general in application.

535 **Cross-validation** approximates a model’s predictive performance on new data by
536 predicting to a hold-out part of the data (typically between 5 and 20 folds).

537 **Leave-one-out cross-validation** disturbs the data least, omitting each single data
538 point in turn. The fit to the hold-out can be quantified in different ways. If the data can
539 be reasonably well described by a specific distribution with log-likelihood function ℓ
540 (even if the model algorithm itself is non-parametric), the log-likelihood of the data in
541 the k folds can be computed and summed (van der Laan et al., 2004; Wood, 2015, p. 36):

$$\ell_{CV}^m = \sum_{i=1}^k \ell(y_{[i]} | \hat{\theta}_{y_{[-i]}}^m), \quad (12)$$

542 where the index $[-i]$ indicates that the data $y_{[i]}$ in fold i were not used for fitting model
543 m and estimating model parameters $\hat{\theta}_{y_{[-i]}}^m$. *Cross-validation log-likelihood*, specifically
544 leave-one-out cross-validation, is asymptotically equivalent to AIC and thus
545 KL-distance (Stone, 1977), albeit at a higher computational cost. The use of hold-out
546 data in cross-validation implicitly penalises overfitting, and we can hence compute
547 model weights w_{CV}^m in the same way as AIC-weights (Hauenstein et al., 2017):

$$w_{CV}^m = \frac{e^{\ell_{CV}^m}}{\sum_{i \in \mathcal{M}} e^{\ell_{CV}^i}}. \quad (13)$$

548 Other measures of model fit to the hold-out folds have been used, largely as *ad hoc*
549 proxies for a likelihood function (e.g. in likelihood-free models): pseudo- R^2 (e.g
550 Nagelkerke, 1991; Nakagawa and Schielzeth, 2013), area under the ROC-curve (AUC:
551 Marmion et al., 2009a; Ordonez and Williams, 2013; Hannemann et al., 2015), or True
552 Skill Statistic (Diniz-Filho et al., 2009; Garcia et al., 2012; Engler et al., 2013; Meller
553 et al., 2014). In these cases, weights were computed by substituting ℓ_{CV} in eqn (13) by
554 the respective measure, or given a value of $1/S$ for a somewhat arbitrarily defined
555 subset of S (out of M) models, e.g. those above an arbitrary threshold considered
556 minimal satisfactory performance (Crossman and Bass, 2008; Crimmins et al., 2013;

Largely ignored by the ecological literature are two other non-parametric approaches to compute model weights: *stacking* and *jackknife model averaging* (see Appendix S4 for discussion of averaging *within* machine-learning algorithms). Both are cross-validation based, and both optimise model weights on hold-out data. **Stacking** (Wolpert, 1992; Smyth and Wolpert, 1998; Ting and Witten, 1999) finds the optimised model weights to reduce prediction error (or maximise likelihood) on a test hold-out of size H . This is, for RMSE and likelihood, respectively:

$$\arg \min_{w_m} \left\{ \sqrt{\frac{1}{H} \sum_{i=1}^H \left(y_{[i]} - \sum_{m=1}^M w_m \hat{f} \left(X_i \mid \hat{\theta}_{[-i]}^m \right) \right)^2} \right\}$$

(Hastie et al., 2009) and

$$\arg \max_{w_m} \left\{ \ell \left(y_{[i]} \mid \sum_{m=1}^M w_m \hat{f} \left(X_i \mid \hat{\theta}_{[-i]}^m \right) \right) \right\},$$

558 where $\hat{f}(X_i \mid \hat{\theta}_{[-i]}^m)$ is the prediction of model m , fitted without using data i , to data i .

559 This procedure is repeated many times, each time yielding a vector of optimised model
560 weights, w_m , which are then averaged across repetitions and rescaled to sum to 1.

561 Smyth and Wolpert (1998) and Clarke (2003) reports stacking to generally outperform
562 the cross-validation approach from two paragraphs earlier, and Bayesian model
563 averaging, respectively (see also the case studies in section 4 and Appendix S5).

564 In **Jackknife Model Averaging** (JMA: Hansen and Racine, 2012), each data point
565 is omitted in turn from fitting and then predicted to (thus actually a leave-one-out
566 cross-validation rather than a “jackknife”). Then, weights are optimised so as to
567 minimise RMSE (or maximise likelihood) between the observed and the fitted value
568 across all “jackknife” samples. The optimisation function is the same as for stacking,
569 except that $H = N$. Thus, in stacking, weights are optimised once for each run, while
570 for the jackknife only one optimisation over all N leave-one-out-cross-validations is
571 required (further details and examples with R-code are given in Appendix S5.6).

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

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

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

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

593 **Model-based model combination: varying weights**

594 Combining model predictions using statistical models, an approach we term
595 “model-based model combinations” (**MBMC**, also called “superensemble modelling”)

596 was first proposed by Granger and Ramanathan (1984). Here a statistical model f is
597 used to combine the predictions from different models, as if they were predictors in a
598 regression: $\tilde{Y} \sim f(\hat{Y}_1, \hat{Y}_2, \dots, \hat{Y}_m)$ (see Fig. 4 left). The regression-type model f can be
599 of any type, such as a linear model or a neural network. We call this regression the
600 “supra-model” in order to distinguish between different modelling levels.

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

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

613 [Fig. 4 approximately here.]

614 This does not seem to be recognised as a problem (despite being a key message of
615 Hastie et al., 2009), as all studies we found incorrectly cross-validate the supra-model
616 only, not the *entire* workflow (if at all; e.g. Krishnamurti et al., 1999; Thomson et al.,
617 2006; Diks and Vrugt, 2010; Breiner et al., 2015; Romero et al., 2016). To correctly
618 cross-validate MBMCs, one has to produce hold-outs *before* fitting the contributing
619 models, and evaluate the MBMC prediction on this hold-out (Fig. 4, Appendix S5.9 and
620 case studies).

621 Note that supra-models may differ substantially in their ability to harness the

622 contributing models. As it is a yet fairly unexplored field in model averaging, analysts
623 are advised to try different supra-model types (Fig. 4).

624 **3.4 Equal weights**

625 In many fields of science (climate modelling, economics, political sciences), model
626 averaging proceeds with giving the structurally different models equal weight, i.e.
627 $1/M$ (e.g. Johnson and Bowler, 2009; Knutti et al., 2010; Graefe et al., 2014; Rougier,
628 2016). In ecology, studies analysing species distributions reported equal weights to be a
629 very good choice when assessed using cross-validation (Crossman and Bass, 2008;
630 Marmion et al., 2009a; Rapacciuolo et al., 2012), but no better than the single models on
631 validation with independent data (Crimmins et al., 2013). Equal weights may serve as a
632 reference approach to see whether estimating weights reduces prediction error for this
633 specific set of models. In that sense, we may argue, all the above weight estimation
634 approaches only serve to separate the wheat from the chaff; once a set of reasonable
635 models has been identified, equal weights are apparently a good approach.

636 **4 Case studies**

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

645 preselected set of models.

646 **4.1 Case study 1: Simulation with Gaussian response,** 647 **many models and few data points**

648 In this first, simulation-based case study, we explore the variability of model-averaging
649 approaches in the common case where several partially nested models are fit (see
650 Appendix S9 for details and code). The simulation was set up so that several of the
651 fitted models have similar support as explanations for the data. This was achieved by
652 generating the response differently in each of two groups (using similar, but not
653 identical predictors). We simulated 70 data points with 4 predictors yielding $2^4 = 16$
654 candidate models, and another 70 for validation. We computed model weights in 19
655 different ways (Table 1) and compared the prediction error of weighted averages as
656 well as the individual models to the validation data points. Simulation and analyses
657 were repeated 100 times.

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

[Figure 6 approximately here.]

[Table 2 approximately here.]

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

In the second case study we use data on the real distribution of short-finned eel (*Anguilla australis*) in New Zealand (from Elith et al., 2008). The data are provided in the R-package *dismo*, already split into a 1000-rows training and a 500-rows test data set, and featuring 10 predictors. We ran four different model types (GAM, randomForest-rF, artificial neural network-ANN, support vector machine-SVM), along with two variations of the GLM (best models selected by AIC and BIC). For details see Appendix S10.

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

[Table 3 approximately here.]

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

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

699 The good performance of the median in both case studies suggests that using the
700 central value of *each prediction*, rather than give constant weights to the model itself,
701 may be even more effective in reducing variance and thus prediction error.

702 **5 Recommendations**

703 Despite setting out to review the field of model averaging for ecologists, the complexity
704 of the topic prevents us from providing final answers. The recent mathematical
705 explanation why estimating optimal weights makes the averaged predictions perform
706 poorly (Claeskens et al., 2016) is an example for fundamental limitations of model
707 averaging. Many issues seem to be statistically unresolved, or addressed by quick-fixes
708 and even fundamental questions remain open, which we will discuss in the final
709 section.

710 It is unsatisfactory to see the large variance in weights and performance of the
711 different averaging approaches in our case studies. Also the literature provides too few
712 comparisons of model weights to provide robust advice. In general, our
713 recommendations are thus guided by reducing harm, rather than suggesting an optimal
714 solution.

715 **5.1 Averaged prediction should be accompanied by** 716 **uncertainty estimates**

717 Just like any other statistical approach, model averaging can also be misapplied.
718 Focussing entirely on the predictions rather than their spread can mislead, as Knutti
719 et al. (2010) showed for combining precipitation predictions: spatial heterogeneity
720 cancelled out across models, giving the erroneous impression of little change when in
721 fact all models predict large changes (albeit in different regions). Similarly, King et al.
722 (2008) found that averaging parameters from two competing models led to no effect of
723 two hypothesised impacts, although in both models a (different) driver was very
724 influential. We thus strongly encourage including at least model-averaged confidence
725 intervals alongside any prediction, possibly in addition to the individual model
726 predictions, to prevent erroneous interpretation of averaged predictions. Also, more
727 attention should be paid to the full model. It has many desirable properties (unbiased
728 parameter estimates, very good coverage), but suffers from violation of the parsimony
729 principle (“Occam’s razor”) and requires more consideration in which form covariates
730 should be fit. Its larger prediction error, compared to the over-optimistic single-best
731 partial model, is the reason for correct confidence intervals.

732 **5.2 Dependencies among model predictions should be** 733 **addressed**

734 Statistical models, which aim to describe the data to which they are fitted, will often
735 have correlated parameters and fits; process models may overlap in modelled processes.
736 Having highly similar models in the model set will inflate the cumulative weight given
737 to them (as illustrated in Appendix S6) . One way to handle inflation of weights by
738 highly-related models is to assign prior model probabilities in a Bayesian framework.

739 Another approach would be to pre-select models of different types (see next point).
740 Alternatively, the cos-square scheme of Garthwaite and Mubwandarikwa (2010) uses
741 the correlation matrix of model projections to appropriately change weights of
742 correlated models. It is the only approach currently doing so, and, while the jury is still
743 out on this method, our case study results look only mildly promising (Fig. 6, Tables 2
744 and 3).

745 **5.3 Validation-based weighting or validation-based** 746 **pre-selection of models**

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

755 In contrast, some studies select models *after* the predictions are made (e.g. Thuiller,
756 2004; Forester et al., 2013). These studies have averaged models which predict in the
757 same direction (along the “consensus axis”: Grenouillet et al. 2010), which are the best
758 50% in the set (Marmion et al., 2009a), or however many one should combine to
759 minimise prediction error. Such approaches necessitate addressing the challenge of
760 using data twice (Lauzeral et al., 2015). Post-selection reduces the ability of “dissenting
761 voices” (i.e. less correlated predictions) to reduce prediction error and instead reinforce
762 the trend of the model type most represented in the set. As a consequence, their

763 uncertainty estimation will be overly optimistic. We do not advocate their use.

764 We suggest to employ **validation-based methods of model averaging** rather
765 than relying on model-based estimates of error, i.e. (leave-one out) cross-validation and
766 stacking rather than AIC. On account of us rarely believing our models in ecology, test
767 data give us some capacity to make allowances for predictive bias. It is probably of
768 little practical relevance whether models are pre-selected by validation-based estimates
769 of error and then averaged with equal weights or weighted by validation-based
770 estimates of error without pre-selection.

771 **5.4 Process models are no different**

772 In fishery science, averaging process models is relatively common (Brodziak and Piner,
773 2010), as it is in weather and climate science (Krishnamurti et al., 1999; Knutti et al.,
774 2010; Bauer et al., 2015). There are at least two connected challenges such enterprises
775 face: validation and weighting. Often process models are tuned/calibrated on all sets of
776 data available, in the logical attempt to describe all relevant processes in the best
777 possible way. That means, however, that no independent validation data are available,
778 so that we cannot use the prediction accuracy of different models to compute model
779 weights. Consequently, all models receive the same weight (e.g. in IPCC reports, or for
780 economic models), or some reasonable but statistically ad-hoc construction of weights
781 is employed (e.g. Giorgi and Mearns, 2002). In recent years, hind-casting has gained in
782 popularity, i.e. evaluating models by predicting to past data. This will only be a useful
783 approach if historic data were not used already to derive or tune model parameters,
784 and if hindcasting success is related to prediction success (which it need not be, if
785 processes or drivers change).

786 Cross-validation is often infeasible for large models, as run-times are prohibitively
787 long. However, the greatest obstacle to averaging process models is the absence of truly

788 equivalent alternative models, which predict the same state variable. Fishery science is
789 one of the few areas of ecology in which commensurable models exist and are being
790 averaged in a variety of ways (e.g. Stanley and Burnham, 1998; Brodziak and Legault,
791 2005; Brandon and Wade, 2006; Katsanevakis, 2006; Hill et al., 2007; Katsanevakis and
792 Maravelias, 2008; Jiao et al., 2009; Hollowed et al., 2009; Brodziak and Piner, 2010).
793 Carbon and biomass assessments are also moving in that direction (Hanson et al., 2004;
794 Butler et al., 2009; Wang et al., 2009; Picard et al., 2012). These fields would profit from
795 averaging methods such as minimal variance and cos-squared, which do not require
796 cross-validation and may perform better than either equal weights or BMA-EM, and
797 probably better than MBMC's potentially overfitted supra-models.

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

801 **Acknowledgements**

802 We like to thank the German Science Foundation (DFG) for funding the workshop
803 "Model averaging in Ecology", held in Freiburg 2-6 March 2015 (DO 786/9-1). Part of
804 this work was carried out during a research stay of CFD at the University of Melbourne,
805 co-funded by the DFG (DO 786/10-1). BS is supported by the DFG (SCHR1000/6-2 and
806 SCHR1000/8-2). DIW is supported by an Australian Research Council (ARC) Future
807 Fellowship (grant number FT120100501). DRR is supported by the Alexander von
808 Humboldt Foundation through the German Federal Ministry of Education and
809 Research. GGA is the recipient of a Discovery Early Career Research Award from the
810 Australian Research Council (project DE160100904). JE is supported by ARC's
811 FT0991640 and DP160101003. JJLM is supported by Australia's National Environmental

812 Research Program (NERP) Environmental Decisions Hub and ARC DP160101003. KG is
813 supported by the German Federal Ministry of Education and Research (BMBF
814 01LL0901A). WT received funding from the European Research Council under the
815 European Community's FP7/2007-2013 Grant Agreement no. 281422 (TEEMBIO).

816 **References**

- 817 Akaike, H. 1973. Information theory as an extension of the maximum likelihood
818 principle. In B. Petrov and F. Csaki, editors, 2nd International Symposium on
819 Information Theory, page 267–281. Akademiai Kiado, Budapest.
- 820 Armstrong, J. S. 2001. Combining forecasts. In J. S. Armstrong, editor, Principles of
821 Forecasting: A Handbook for Researchers and Practitioners, pages 417–439. Springer,
822 New York.
- 823 Augustin, N., W. Sauerbrei, and M. Schumacher. 2005. The practical utility of
824 incorporating model selection uncertainty into prognostic models for survival data.
825 *Statistical Modelling*, 5:95–118.
- 826 Banner, K. M. and M. D. Higgs. 2017. Considerations for assessing model averaging of
827 regression coefficients. *Ecological Applications*, 28:78–93.
- 828 Bates, J. M. and C. W. J. Granger. 1969. The combination of forecasts. *Journal of the*
829 *Operational Research Society*, 20:451–468.
- 830 Bauer, P., A. Thorpe, and G. Brunet. 2015. The quiet revolution of numerical weather
831 prediction. *Nature*, 525:47–55.
- 832 Berger, J. O. and L. R. Pericchi. 1996. The intrinsic bayes factor for model selection and
833 prediction. *Journal of the American Statistical Association*, 91:109–122.
- 834 Bohn, T. J., M. Y. Sonessa, and D. P. Lettenmaier. 2010. Seasonal hydrologic forecasting:
835 do multimodel ensemble averages always yield improvements in forecast skill?

836 Journal of Hydrometeorology, **11**:1358–1372.

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

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

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

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

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

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

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

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

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

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

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

864 Cade, B. 2015. Model averaging and muddled multimodal inferences. *Ecology*,
865 **96**:2370–2382.

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

869 Chamberlin, T. C. 1890. The method of multiple working hypotheses. *Science*,
870 **15**:92–96.

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

873 Chickering, D. M. and D. Heckerman. 1997. Efficient approximations for the marginal
874 likelihood of Bayesian networks with hidden variables. *Machine Learning*,
875 **29**:181–212.

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

878 Claeskens, G., J. R. Magnus, A. L. Vasnev, and W. Wang. 2016. The forecast combination
879 puzzle: A simple theoretical explanation. *International Journal of Forecasting*,
880 **32**:754–762.

881 Clarke, B. 2003. Comparing Bayes model averaging and stacking when model
882 approximation error cannot be ignored. *The Journal of Machine Learning Research*,
883 **4**:683–712.

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

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

- 888 Corani, G. and A. Mignatti. 2015. Robust Bayesian model averaging for the analysis of
889 presence–absence data. *Environmental and Ecological Statistics*, **22**:513–534.
- 890 Crimmins, S. M., S. Z. Dobrowski, and A. R. Mynsberge. 2013. Evaluating ensemble
891 forecasts of plant species distributions under climate change. *Ecological Modelling*,
892 **266**:126–130.
- 893 Crisci, C., R. Terra, J. Pablo, B. Ghattas, M. Bidegain, G. Goyenola, J. José, G. Méndez,
894 and N. Mazzeo. 2017. Multi-model approach to predict phytoplankton biomass and
895 composition dynamics in a eutrophic shallow lake governed by extreme
896 meteorological events. *Ecological Modelling*, **360**:80–93.
- 897 Crossman, N. D. and D. A. Bass. 2008. Application of common predictive habitat
898 techniques for post-border weed risk management. *Diversity and Distributions*,
899 **14**:213–224.
- 900 Dietze, M. C. 2017. *Ecological Forecasting*. Princeton University Press, Princeton, N.J.
- 901 Diks, C. G. H. and J. A. Vrugt. 2010. Comparison of point forecast accuracy of model
902 averaging methods in hydrologic applications. *Stochastic Environmental Research
903 and Risk Assessment*, **24**:809–820.
- 904 Diniz-Filho, J. A. F., L. Mauricio Bini, T. Fernando Rangel, R. D. Loyola, C. Hof,
905 D. Nogués-Bravo, M. B. Araújo, L. M. Bini, and T. F. L. V. B. Rangel. 2009.
906 Partitioning and mapping uncertainties in ensembles of forecasts of species turnover
907 under climate change. *Ecography*, **32**:897–906.
- 908 Dormann, C. F., O. Schweiger, P. Arens, I. Augenstein, S. Aviron, D. Bailey, C. F.
909 Dormann, J. Baudry, R. Billeter, R. Bugter, R. Bukáček, F. Burel, M. Cerny, R. D. Cock,
910 G. D. Blust, R. DeFilippi, T. Diekötter, J. Dirksen, W. Durka, P. J. Edwards, M. Frenzel,
911 R. Hamersky, F. Hendrickx, F. Herzog, S. Klotz, B. Koolstra, A. Lausch, D. L. Coeur,
912 J. Liira, J.-P. P. Maelfait, P. Opdam, M. Roubalova, A. Schermann-Legionnet,
913 N. Schermann, T. Schmidt, M. J. M. Smulders, M. Speelmans, P. Simova, J. Verboom,

914 W. K. R. E. van Wingerden, and M. Zobel. 2008. Prediction uncertainty of
915 environmental change effects on temperate european biodiversity. *Ecology Letters*,
916 **11**:235–244.

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

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

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

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

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

926 Engler, R., L. T. Waser, N. E. Zimmermann, M. Schaub, S. Berdos, C. Ginzler, and
927 A. Psomas. 2013. Combining ensemble modeling and remote sensing for mapping
928 individual tree species at high spatial resolution. *Forest Ecology and Management*,
929 **310**:64–73.

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

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

934 Forester, B. R., E. G. DeChaine, and A. G. Bunn. 2013. Integrating ensemble species
935 distribution modelling and statistical phylogeography to inform projections of
936 climate change impacts on species distributions. *Diversity and Distributions*,
937 **19**:1480–1495.

938 Freckleton, R. P. 2011. Dealing with collinearity in behavioural and ecological data:
939 model averaging and the problems of measurement error. *Behavioral Ecology and*

940 Sociobiology, **65**:91–101.

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

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

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

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

950 Gibbons, J. M., G. M. Cox, A. T. A. Wood, J. Craigon, S. J. Ramsden, D. Tarsitano, and
 951 N. M. J. Crout. 2008. Applying Bayesian Model Averaging to mechanistic models: An
 952 example and comparison of methods. *Environmental Modelling & Software*,
 953 **23**:973–985. WOS:000255770300002.

954 Gibbs, J. W. 1902. *Elementary Principles in Statistical Mechanics*. Charles Scribner’s
 955 Sons, New York.

956 Giorgi, F. and L. O. Mearns. 2002. Calculation of average, uncertainty range, and
 957 reliability of regional climate changes from aogcm simulations via the “reliability
 958 ensemble averaging” (rea) method. *Journal of Climate*, **15**:1141–1158.

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

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

964 Graefe, A., H. Küchenhoff, V. Stierle, and B. Riedl. 2015. Limitations of Ensemble
 965 Bayesian Model Averaging for forecasting social science problems weights.

966 International Journal of Forecasting, **31**:943–951.

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

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

971 Grenouillet, G., L. Buisson, N. Casajus, and S. Lek. 2010. Ensemble modelling of species
972 distribution: the effects of geographical and environmental ranges. *Ecography*,
973 **34**:9–17.

974 Grueber, C. E., S. Nakagawa, R. J. Laws, and I. G. Jamieson. 2011. Multimodel inference
975 in ecology and evolution: challenges and solutions. *Journal of Evolutionary Biology*,
976 **24**:699–711.

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

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

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

983 Hanson, P. J., J. S. Amthor, S. D. Wullschleger, K. B. Wilson, R. F. Grant, A. Hartley,
984 D. Hui, J. E. R. Hunt, D. W. Johnson, J. S. Kimball, A. W. King, Y. Luo, S. G. McNulty,
985 G. Sun, P. E. Thornton, S. Wang, M. Williams, D. D. Baldocchi, and R. M. Cushman.
986 2004. Oak forest carbon and water simulations: Model intercomparisons and
987 evaluations against independent data. *Ecological Monographs*, **74**:443–489.

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

990 Hartig, F., J. M. Calabrese, B. Reineking, T. Wiegand, and A. Huth. 2011. Statistical
991 inference for stochastic simulation models - theory and application. *Ecology Letters*,

992 **14:816–827.**

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

995 Hauenstein, S., C. F. Dormann, and S. N. Wood. 2017. Computing aic for black-box
996 models using generalised degrees of freedom: a comparison with cross-validation.
997 Communications in Statistics-Simulation and Computation, **in press**:1–17.

998 Hill, S. L., G. M. Watters, A. E. Punt, M. K. McAllister, C. L. Quéré, and J. Turner. 2007.
999 Model uncertainty in the ecosystem approach to fisheries. Fish and Fisheries,
1000 **8:315–336.**

1001 Hobbs, N. T. and R. Hilborn. 2006. Alternatives to statistical hypothesis testing in
1002 ecology: a guide to self teaching. Ecological Applications, **16:5–19.**

1003 Hoeting, J., D. Madigan, A. Raftery, and C. Volinsky. 1999. Bayesian model averaging: a
1004 tutorial. Statistical Science, **14:382–401.**

1005 Hollowed, A. B., N. A. Bond, T. K. Wilderbuer, W. T. Stockhausen, Z. T. A’mar, R. J.
1006 Beamish, J. E. Overland, and M. J. Schirripa. 2009. A framework for modelling fish
1007 and shellfish responses to future climate change. ICES Journal of Marine Science,
1008 **66:1584–1594.**

1009 Hooten, M. B. and N. T. Hobbs. 2015. A guide to Bayesian model selection for
1010 ecologists. Ecological Monographs, **85:3–28.**

1011 Jiao, Y., K. Reid, and E. Smith. 2009. Model selection uncertainty and bayesian model
1012 averaging in fisheries recruitment modeling. In R. J. Beamish and B. J. Rothschild,
1013 editors, The Future of Fisheries Science in North America, pages 505–524. Springer,
1014 New York.

1015 Johnson, C. and N. Bowler. 2009. On the reliability and calibration of ensemble
1016 forecasts. Monthly Weather Review, **137:1717–1720.**

1017 Johnson, J. B. and K. S. Omland. 2004. Model selection in ecology and evolution.

1018 Trends in Ecology and Evolution, **19**:101–108.

1019 Kabaila, P., A. H. Welsh, and W. Abeysekera. 2015. Model-averaged confidence
1020 intervals. *Scandinavian Journal of Statistics*, **43**:35–48.

1021 Kass, R. E. and A. E. Raftery. 1995. Bayes factors. *Journal of the American Statistical*
1022 *Association*, **90**:773–795.

1023 Katsanevakis, S. 2006. Modelling fish growth: Model selection, multi-model inference
1024 and model selection uncertainty. *Fisheries Research*, **81**:229–235.

1025 Katsanevakis, S. and C. D. Maravelias. 2008. Modelling fish growth: multi-model
1026 inference as a better alternative to a priori using von Bertalanffy equation. *Fish and*
1027 *Fisheries*, **9**:178–187.

1028 King, R., S. P. Brooks, C. Mazzetta, S. N. Freeman, and B. J. T. Morgan. 2008. Identifying
1029 and diagnosing population declines: A Bayesian assessment of lapwings in the UK.
1030 *Journal of the Royal Statistical Society C*, **57**:609–632.

1031 Knutti, R., R. Furrer, C. Tebaldi, J. Cermak, and G. A. Meehl. 2010. Challenges in
1032 combining projections from multiple climate models. *Journal of Climate*,
1033 **23**:2739–2758.

1034 Krishnamurti, T. N., C. M. Kishtawal, T. E. LaRow, D. R. Bachiochi, Z. Zhang, C. E.
1035 Williford, S. Gadgil, and S. Surendran. 1999. Improved weather and seasonal climate
1036 forecasts from multimodel superensemble. *Science*, **285**:1548–1550.

1037 Kuhn, M. and K. Johnson. 2013. *Applied Predictive Modeling*. Springer, Berlin.

1038 Lauzeral, C., G. Grenouillet, and S. Brosse. 2015. The iterative ensemble modelling
1039 approach increases the accuracy of fish distribution models. *Ecography*, **38**:213–220.

1040 Le Lay, G., R. Engler, E. Franc, and A. Guisan. 2010. Prospective sampling based on
1041 model ensembles improves the detection of rare species. *Ecography*, **33**:1015–1027.

1042 Liang, H., G. Zou, A. T. K. Wan, and X. Zhang. 2011. Optimal weight choice for
1043 frequentist model average estimators. *Journal of the American Statistical*

1044 Association, **106**:1053–1066.

1045 Link, W. A. and R. J. Barker. 2006. Model weights and the foundations of multimodel
1046 inference. *Ecology*, **87**:2626–2635.

1047 Lukacs, P. M., K. P. Burnham, and D. R. Anderson. 2010. Model selection bias and
1048 Freedman’s paradox. *Annals of the Institute of Statistical Mathematics*, **62**:117–125.

1049 Madigan, D. and A. E. Raftery. 1994. Model selection and accounting for model
1050 uncertainty in graphical models using Occam’s window. *Journal of the American
1051 Statistical Association*, **89**:1535–1546.

1052 Marmion, M., J. Hjort, W. Thuiller, and M. Luoto. 2009*a*. Statistical consensus methods
1053 for improving predictive geomorphology maps. *Computers & Geosciences*,
1054 **35**:615–625.

1055 Marmion, M., M. Parviainen, M. Luoto, R. K. Heikkinen, and W. Thuiller. 2009*b*.
1056 Evaluation of consensus methods in predictive species distribution modelling.
1057 *Diversity and Distributions*, **15**:59–69.

1058 Meller, L., M. Cabeza, S. Pironon, M. Barbet-Massin, L. Maiorano, D. Georges, and
1059 W. Thuiller. 2014. Ensemble distribution models in conservation prioritization: from
1060 consensus predictions to consensus reserve networks. *Diversity and Distributions*,
1061 **20**:309–321.

1062 Montgomery, J. M., F. M. Hollenbach, and M. D. Ward. 2012. Improving Predictions
1063 Using Ensemble Bayesian Model Averaging. *Political Analysis*, **20**:271–291.

1064 Mouquet, N., Y. Lagadeuc, V. Devictor, L. Doyen, A. Duputié, D. Eveillard, D. Faure,
1065 E. Garnier, O. Gimenez, P. Huneman, F. Jabot, P. Jarne, D. Joly, R. Julliard, S. Kéfi, G. J.
1066 Kergoat, S. Lavorel, L. L. Gall, L. Meslin, S. Morand, X. Morin, H. Morlon, G. Pinay,
1067 R. Pradel, F. M. Schurr, W. Thuiller, and M. Loreau. 2015. Predictive ecology in a
1068 changing world. *Journal of Applied Ecology*, **52**:1293–1310.

1069 Nagelkerke, N. J. D. 1991. A note on a general definition of the coefficient of

determination. *Biometrika*, **78**:691–692.

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

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

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

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

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

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

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

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

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

Posada, D. and T. R. Buckley. 2004. Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches

1096 over likelihood ratio tests. *Systematic Biology*, **53**:793–808.

1097 Potempski, S. and S. Galmarini. 2009. Est modus in rebus: analytical properties of
1098 multi-model ensembles. *Atmospheric Chemistry and Physics*, **9**:9471–9489.

1099 Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski. 2005. Using Bayesian
1100 Model Averaging to calibrate forecast ensembles. *Monthly Weather Review*,
1101 **133**:1155–1174.

1102 Raftery, A. E., D. Madigan, and J. A. Hoeting. 1997. Bayesian model averaging for linear
1103 regression models. *Journal of The American Statistical Association*, **92**:179–191.

1104 Rapacciuolo, G., D. B. Roy, S. Gillings, R. Fox, K. Walker, and A. Purvis. 2012. Climatic
1105 associations of British species distributions show good transferability in time but
1106 low predictive accuracy for range change. *PLoS ONE*, **7**:e40212.

1107 Richards, S. A. 2005. Testing ecological theory using the information-theoretic
1108 approach: examples and cautionary results. *Ecology*, **86**:2805–2814.

1109 Richards, S. A., M. J. Whittingham, and P. A. Stephens. 2011. Model selection and
1110 model averaging in behavioural ecology: the utility of the IT-AIC framework.
1111 *Behavioral Ecology and Sociobiology*, **65**:77–89.

1112 Roberts, D. R., V. Bahn, S. Ciuti, M. S. Boyce, J. Elith, G. Guillera-Aroita, S. Hauenstein,
1113 J. J. Lahoz-Monfort, B. Schröder, W. Thuiller, D. I. Warton, B. A. Wintle, F. Hartig,
1114 and C. F. Dormann. 2017. Cross-validation strategies for data with temporal, spatial,
1115 hierarchical, or phylogenetic structure. *Ecography*, **40**:913–929.

1116 Romero, D., J. Olivero, J. C. Brito, and R. Real. 2016. Comparison of approaches to
1117 combine species distribution models based on different sets of predictors.
1118 *Ecography*, **39**:561–571.

1119 Rougier, J. 2016. Ensemble averaging and mean squared error. *Journal of Climate*,
1120 **29**:8865–8870.

1121 Rovai, A. S., P. Riul, R. R. Twilley, E. Castañeda-Moya, V. H. Rivera-Monroy, A. A.

1122 Williams, M. Simard, M. Cifuentes-Jara, R. R. Lewis, S. Crooks, and et al. 2015.
1123 Scaling mangrove aboveground biomass from site-level to continental-scale. *Global*
1124 *Ecology and Biogeography*, **25**:286–298.

1125 Schomaker, M. 2012. Shrinkage averaging estimation. *Statistical Papers*, **53**:1015–1034.

1126 Schomaker, M., A. T. K. Wan, and C. Heumann. 2010. Frequentist model averaging with
1127 missing observations. *Computational Statistics and Data Analysis*, **54**:3336–3347.

1128 Schwartz, G. E. 1978. Estimating the dimension of a model. *Annals of Statistics*,
1129 **6**:461–464.

1130 Shen, X. and H.-C. Huang. 2006. Optimal model assessment, selection, and
1131 combination. *Journal of American Statistical Association*, **101**:554–568.

1132 Shmueli, G. 2010. To explain or to predict? *Statistical Science*, **25**:289–310.

1133 Smith, A. B., M. J. Santos, M. S. Koo, K. M. C. Rowe, K. C. Rowe, J. L. Patton, J. D.
1134 Perrine, S. R. Beissinger, and C. Moritz. 2013. Evaluation of species distribution
1135 models by resampling of sites surveyed a century ago by Joseph Grinnell.
1136 *Ecography*, **36**:1017–1031.

1137 Smith, R. L., C. Tebaldi, D. W. Nychka, and L. O. Mearns. 2009. Bayesian modeling of
1138 uncertainty in ensembles of climate models. *Journal of the American Statistical*
1139 *Association*, **104**:97–116.

1140 Smyth, P. and D. Wolpert. 1998. An Evaluation of Linearly Combining Density
1141 Estimators via Stacking. Technical Report No. 98-25. Information and Computer
1142 Science Department, University of California, Irvine, CA.

1143 Solomon, S., D. Qin, M. Manning, M. Marquis, K. Averyt, M. Tignor, H. L. M. Jr., and
1144 Z. Chen, editors. 2007. IPCC: Climate Change 2007. The Physical Science Basisf.
1145 Cambridge University Press.

1146 Stanley, T. R. and K. P. Burnham. 1998. Information-theoretic model selection and
1147 model averaging for closed-population capture-recapture studies. *Biometrical*

1148 Journal, **40**:475–494.

1149 Stock, J. and M. Watson. 2001. A comparison of linear and nonlinear univariate models
1150 for forecasting macroeconomic time series. In R. Engle and H. White, editors,
1151 Cointegration, Causality, and Forecasting: A Festschrift in Honour of Clive Granger,
1152 page 1–44. Oxford University Press, Oxford, UK.

1153 Stone, M. 1977. An asymptotic equivalence of choice of model by cross-validation and
1154 Akaike’s criterion. *Journal of the Royal Statistical Society B*, page 44–47.

1155 Symonds, M. R. E. and A. Moussalli. 2011. A brief guide to model selection, multimodel
1156 inference and model averaging in behavioural ecology using Akaike’s information
1157 criterion. *Behavioral Ecology and Sociobiology*, **65**:13–21.

1158 Thibaut, L. M. and S. R. Connolly. 2013. Understanding diversity-stability relationships:
1159 towards a unified model of portfolio effects. *Ecology Letters*, **16**:140–150.

1160 Thomson, M. C., F. J. Doblas-Reyes, S. J. Mason, R. Hagedorn, S. J. Connor, T. Phindela,
1161 A. P. Morse, and T. N. Palmer. 2006. Malaria early warnings based on seasonal
1162 climate forecasts from multi-model ensembles. *Nature*, **439**:576–579.

1163 Thuiller, W. 2004. Patterns and uncertainties of species’ range shifts under climate
1164 change. *Global Change Biology*, **10**:2020–2027.

1165 Timmermann, A. G. 2006. Forecast combinations. In G. Elliott, C. Granger, and
1166 A. Timmermann, editors, *Handbook of Economic Forecasting*, pages 135–196.
1167 Elsevier, Dordrecht.

1168 Ting, K. M. and I. H. Witten. 1999. Issues in stacked generalization. *Journal of Artificial*
1169 *Intelligence Research*, **10**:271–289.

1170 Toni, T., D. Welch, N. Strelkowa, A. Ipsen, and M. P. H. Stumpf. 2009. Approximate
1171 Bayesian computation scheme for parameter inference and model selection in
1172 dynamical systems. *Journal of the Royal Society Interface*, **6**:187–202.

1173 Trolle, D., J. A. Elliott, W. M. Mooij, J. H. Janse, K. Bolding, D. P. Hamilton, and

1174 E. Jeppesen. 2014. Advancing projections of phytoplankton responses to climate
1175 change through ensemble modelling. *Environmental Modelling & Software*,
1176 **61**:371–379.

1177 Turek, D. and D. Fletcher. 2012. Model-averaged Wald confidence intervals.
1178 *Computational Statistics & Data Analysis*, **56**:2809–2815.

1179 van der Laan, M. J., S. Dudoit, and S. Keles. 2004. Asymptotic optimality of
1180 likelihood-based cross-validation. *Statistical Applications in Genetics and Molecular*
1181 *Biology*, **3**:Article4.

1182 van Oijen, M., C. Reyer, F. J. Bohn, D. R. Cameron, G. Deckmyn, M. Flechsig,
1183 S. Harkonen, F. Hartig, A. Huth, A. Kiviste, P. Lasch, A. Makela, T. Mette,
1184 F. Minunno, and W. Rammer. 2013. Bayesian calibration, comparison and averaging
1185 of six forest models, using data from Scots pine stands across Europe. *Forest Ecology*
1186 *and Management*, **289**:255–268.

1187 Wang, Y.-P., C. M. Trudinger, and I. G. Enting. 2009. A review of applications of
1188 model–data fusion to studies of terrestrial carbon fluxes at different scales.
1189 *Agricultural and Forest Meteorology*, **149**:1829–1842.

1190 Wasserman, L. 2000. Bayesian model selection and model averaging. *Journal of*
1191 *Mathematical Psychology*, **44**:92–107.

1192 Watanabe, S. 2010. Asymptotic equivalence of Bayes cross validation and Widely
1193 Applicable Information Criterion in singular learning theory. *Journal of Machine*
1194 *Learning Research*, **11**:3571–3594.

1195 Watanabe, S. 2013. A Widely applicable Bayesian Information Criterion. *Journal of*
1196 *Machine Learning Research*, **14**:867–897.

1197 Weinberg, M. D. 2012. Computing the Bayes factor from a Markov chain Monte Carlo
1198 simulation of the posterior distribution. *Bayesian Analysis*, **7**:737–770.

1199 Wintle, B. A., M. A. McCarthy, C. T. Volinsky, and R. P. Kavanagh. 2003. The use of

1200 Bayesian model averaging to better represent uncertainty in ecological models.
1201 Conservation Biology, **17**:1579–1590.

1202 Wolpert, D. H. 1992. Stacked generalization. Neural Networks, **5**:241–259.

1203 Wood, S. N. 2015. Core Statistics. Cambridge University Press, Cambridge, UK.

1204 Yang, Y. 2001. Adaptive regression by mixing. Journal of the American Statistical
1205 Association, **96**:574–588.

1206 Ye, J. 1998. On measuring and correcting the effects of data mining and model selection.
1207 Journal of the American Statistical Association, **93**:120–131.

1208 Yuan, Z. and D. Ghosh. 2008. Combining multiple biomarker models in logistic
1209 regression. Biometrics, **64**:431–439.

1210 Yuan, Z. and Y. H. Yang. 2005. Combining linear regression models: When and how?
1211 Journal of the American Statistical Association, **100**:1202–1214.

1212 Zhang, X., R. Srinivasan, and D. Bosch. 2009. Calibration and uncertainty analysis of
1213 the SWAT model using Genetic Algorithms and Bayesian Model Averaging. Journal
1214 of Hydrology, **374**:307–317.

1215 Zhang, X., A. T. K. Wan, and G. Zou. 2013. Model averaging by jackknife criterion in
1216 models with dependent data. Journal of Econometrics, **174**:82–94.

1217 Zhao, K., D. Valle, S. Popescu, X. Zhang, and B. Mallick. 2013. Hyperspectral remote
1218 sensing of plant biochemistry using Bayesian model averaging with variable and
1219 band selection. Remote Sensing of Environment, **132**:102–119.

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

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

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

² See also appendix for details and case studies for examples of implementation in R.

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

⁴ Implemented in MuMIn as part of this publication.

⁵ <http://users.stat.umn.edu/~sandy/courses/8053/handouts/Aaron/ARMS/>

Table 2: Model weights (averaged across 100 repetitions) given to the 16 linear regression models of case study 1 by different weighting methods (see Table 1 for abbreviations), arranged by increasing prediction error (last column, median across replications). Only the best (m10) and the full model are shown from the 16 candidate models. LOO-CV: leave-one-out cross-validation using R^2 or RMSE as measure of model performance. For code see case study Appendix S9.

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

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

² Prediction from individual model.

³ Weights are variable. LM and rF refer to a linear model and a randomForest as supra-model, respectively,

Table 3: Model weights given to the six model types of case study 2 (GLM, GAM, randomForest, artificial neural networks and support vector machine) by different weighting methods (see Table 1 for abbreviations), arranged by decreasing fit of the averaged predictions to test data, assessed as log-likelihood (ℓ) (last column). LOO-CV: leave-one-out cross-validation using R^2 or RMSE as measure of model performance. For code see case study Appendix S10.

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

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

² Prediction from individual model.

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

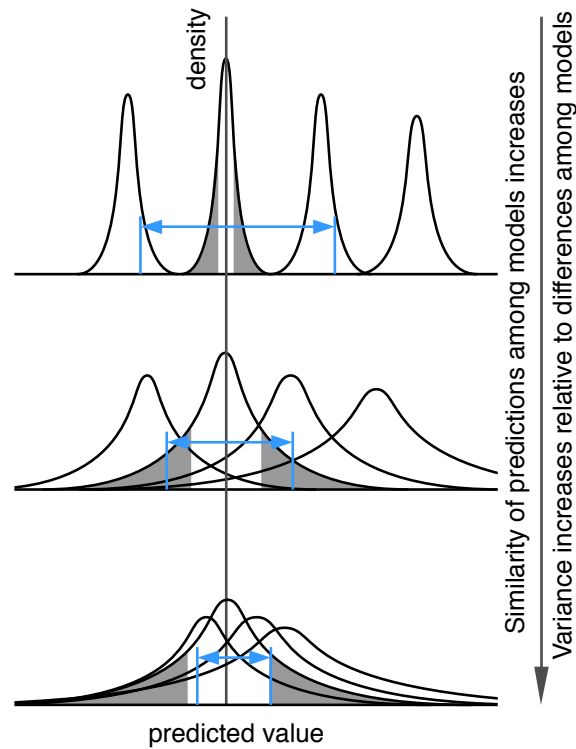


Figure 1: Conceptual depiction of 50% model averaging uncertainty intervals (blue) for different cases of bias and variance in four models (using equal weights). Distributions are the sampling distribution of a prediction from the four models. Truth is indicated by a vertical line. Shaded areas are outer 50% predictions of the best model, illustrating that the best model increasingly predicts to outside the model average's interval as variance becomes large relative to bias. From top to bottom, the source of error morphs from between-model variance to within-model variance. Accordingly, model selection would be more appropriate in the top situation, while model averaging would be superior in the lower situation.

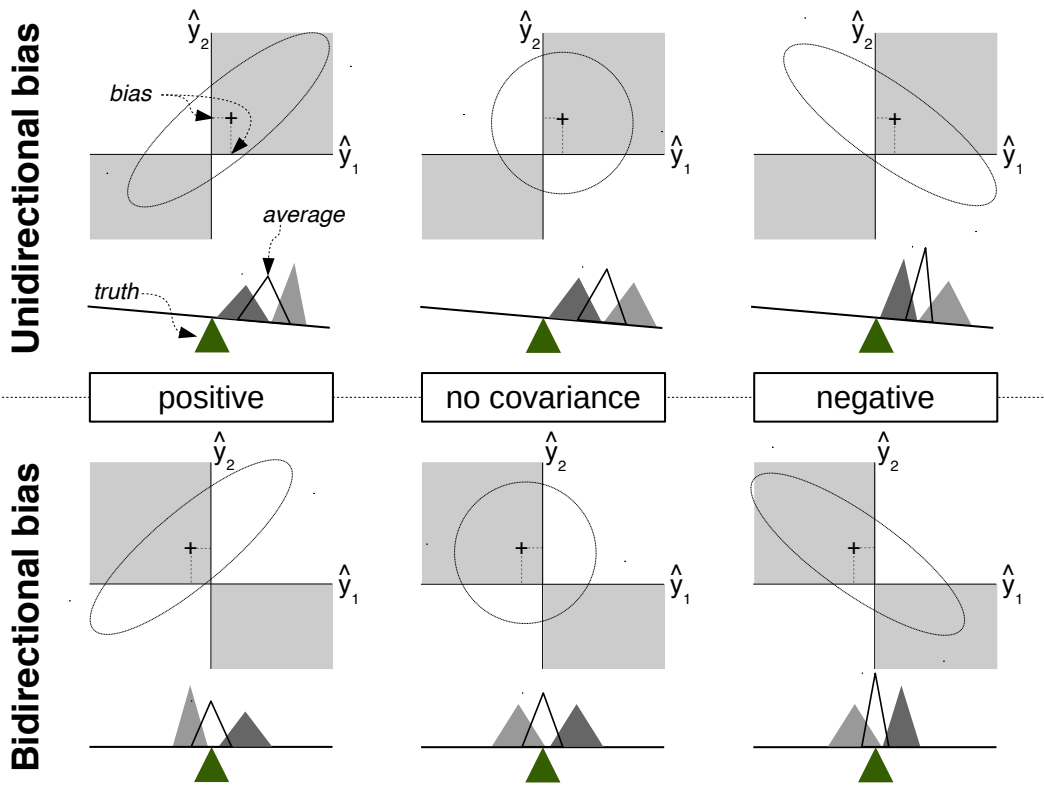


Figure 2: Conceptualised outcomes of model averaging. Sampling distributions of model predictions are depicted as stylised empty triangle on the see-saw (taller means less uncertain). Filled triangles represent the model predictions with unidirectionally bias (top row) or straddling truth (bottom row),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

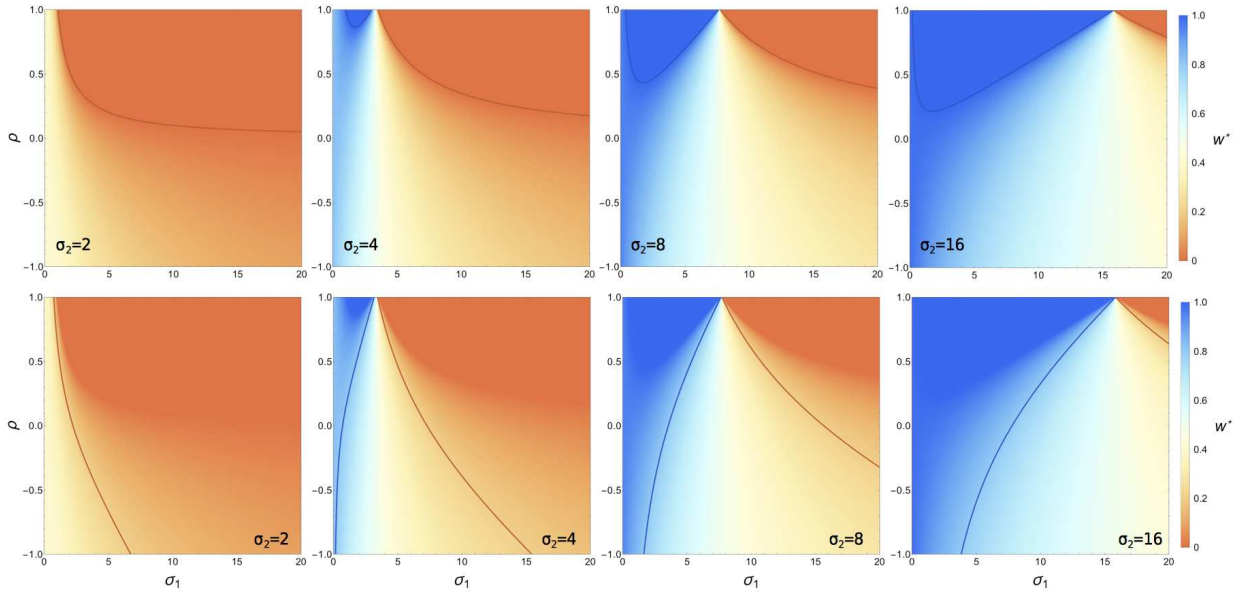


Figure 3: When to average, in the simplest case of two models that make correlated Gaussian predictions. The models are here described by their biases (b_1, b_2 , not shown), their standard deviations (σ_1, σ_2), and by the correlation (ρ) between them. Each panel shows the regions in the (σ_1, ρ) plane where model 1 is best (blue shading and contour line), model 2 is best (orange shading and contour line), and where the optimal average is best (colour gradient between blue and orange). Top row represents the case where weights are known (i.e. without error: $\sigma_w = 0$), while the second row represents exactly the same settings, but with estimated weights (with uncertainty $\sigma_w = 0.2$). Notice that when w is estimated with uncertainty, the contours marking the transition between each single model and the average move into the washed-out colours, i.e. deviate from the fixed w situation in the upper panels. These curves now represent a level set at the values $\bar{w}_1^* = 1 - \sigma_w$ (blue curve) and $\bar{w}_2^* = \sigma_w$ (orange curve). As a consequence, the area where model averaging with estimated weights is superior to the better single model decreases substantially relative to the fixed w case, and disappears completely for $\sigma_w \geq 0.5$. Formal derivations for the contours and the critical weights is given in Appendix S2, the interactive tool itself in Data S1. Biases are set to $b_1 = 3$ and $b_2 = 2$.

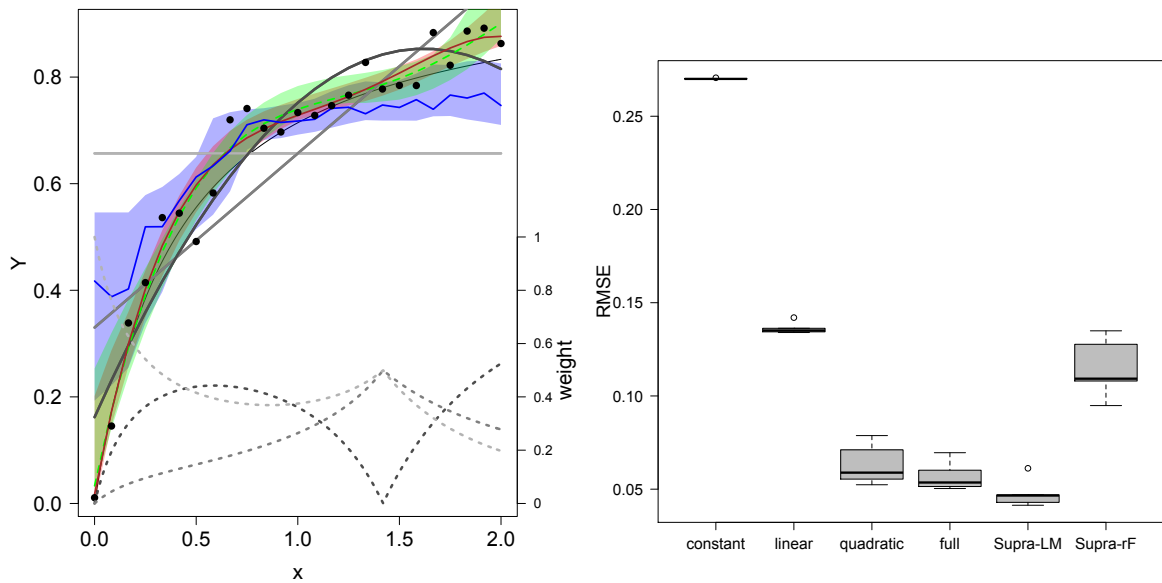


Figure 4: A simple *model-based model combination* example. *Left*: Three models (solid grey lines: constant, linear and quadratic) fitted separately to a data set (points, following the thin black line). Using a linear model (with quadratic terms: red) to combine the three models' fits may improve fit,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

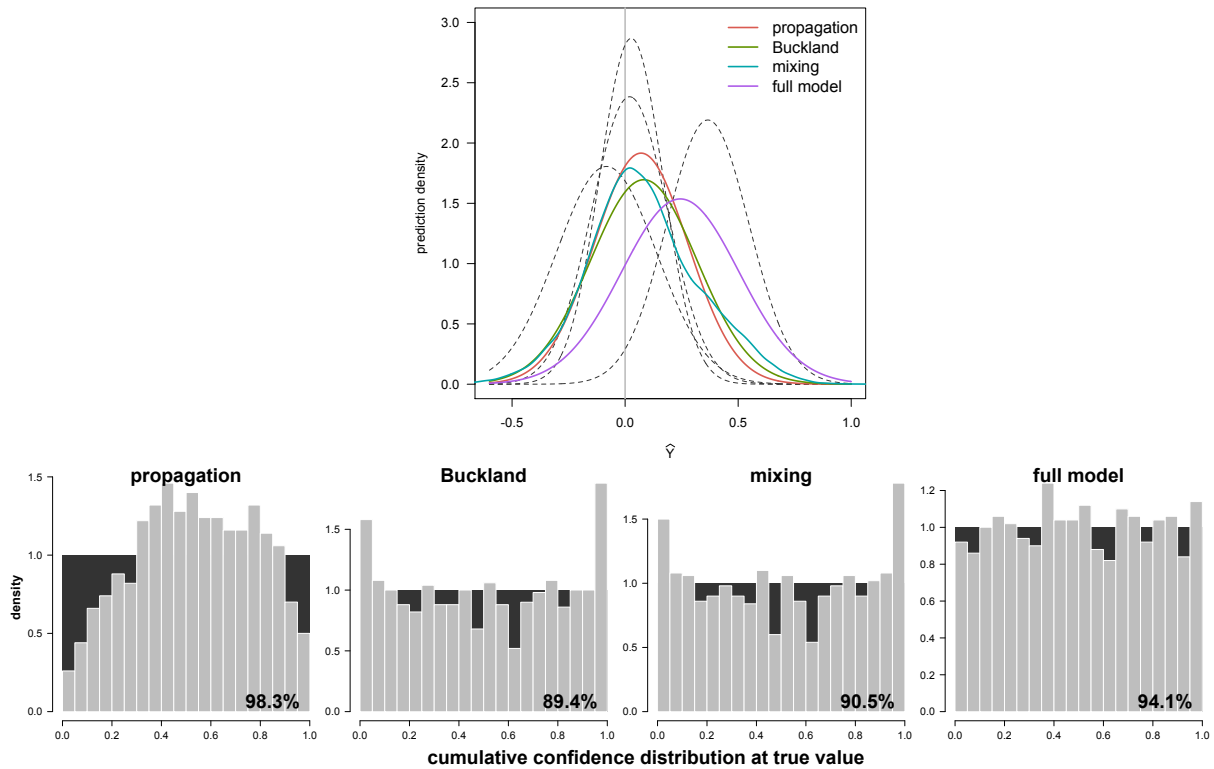


Figure 5: A comparison of different approaches to quantifying uncertainty when combining predictions from four linear models (dashed curves) with equal weights. *Top*: Truth is indicated by the vertical line. Error propagation based on bootstrapped estimates for eqn (5), Buckland et al.'s correction and model mixing yield similar averaged confidence distributions, while the full model is shifted. *Bottom*: Histograms of Bayesian p-values (the quantile of the true value in the posterior distribution across 1000 simulations) for each of the four methods, which should be uniform (black background). Number gives actual coverage for the 95% confidence interval. The error propagation estimate is too conservative with coverage close to 100%. Coverage of the approach of Buckland et al. and mixing are slightly too low in this example.

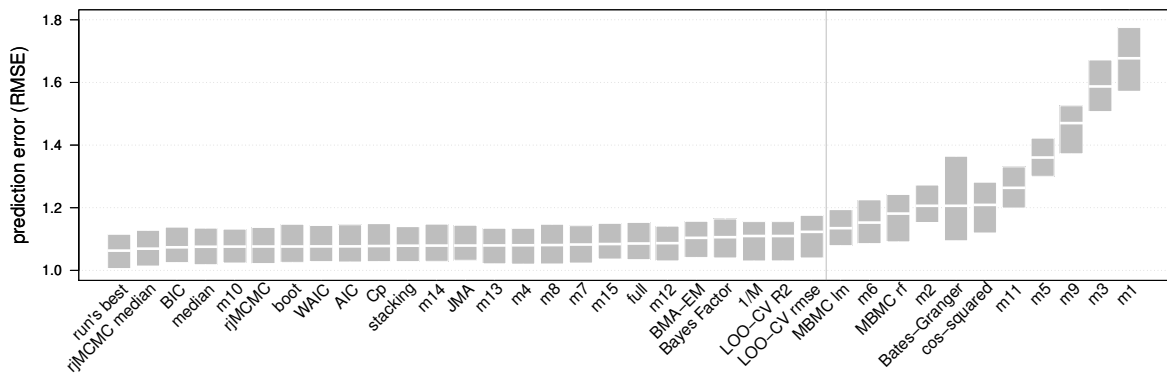


Figure 6: Prediction error of different model averaging approaches (100 repetitions) for case study 1. Box represents quartiles, white line the median. Approaches to the left of the vertical line are very similar, and no better than nine of the candidate models. See Table 1 for list of approaches, and Appendix S9 for list and fits of the individual models.