Ecological Applications, 27(1), 2017, pp. 78–93

Considerations for assessing model averaging of regression coefficients

Katharine M. Banner[1,3] and Megan D. Higgs[1, 2]

[1]*Department of Mathematical Sciences, Montana State University, Wilson Hall 2-214, P.O. Box 172400, Bozeman, Montana 59717 USA*

Abstract. Model choice is usually an inevitable source of uncertainty in model-based statistical analyses. While the focus of model choice was traditionally on methods for choosing a single model, methods to formally account for multiple models within a single analysis are now accessible to many researchers. The specific technique of model averaging was developed to improve predictive ability by combining predictions from a set of models. However, it is now often used to average regression coefficients across multiple models with the ultimate goal of capturing a variable's overall effect. This use of model averaging implicitly assumes the same parameter exists across models so that averaging is sensible. While this assumption may initially seem tenable, regression coefficients associated with particular explanatory variables may not hold equivalent interpretations across all of the models in which they appear, making explanatory inference about covariates challenging. Accessibility to easily implementable software, concerns about being criticized for ignoring model uncertainty, and the chance to avoid having to justify choice of a final model have all led to the increasing popularity of model averaging in practice. We see a gap between the theoretical development of model averaging and its current use in practice, potentially leaving well-intentioned researchers with unclear inferences or difficulties justifying reasons for using (or not using) model averaging. We attempt to narrow this gap by revisiting some relevant foundations of regression modeling, suggesting more explicit notation and graphical tools, and discussing how individual model results are combined to obtain a model averaged result. Our goal is to help researchers make informed decisions about model averaging and to encourage question-focused modeling over method-focused modeling.

Key words: Bayesian model averaging; explanatory inference; linear regression; model averaging; model selection; multimodel inference; predictive inference.

Introduction

In practice, there typically exists more than one reasonable model as a basis for statistical inference and for over 50 years the potential implications of ignoring uncertainty in the process of choosing an inferential model have been discussed. Leamer (1978) cautioned that inferences conditional on one model may result in inflated precision for estimates and predictions. Similarly, Hodges (1987) explicitly laid out three types of uncertainty he argued should be addressed in any analysis: uncertainty in structure (model), uncertainty in parameter estimates conditional on model, and uncertainty in measurement (inherent in data collection).

Algorithms for implementing Bayesian model selection and variable selection were developed in the late 1990s and the formal assessment of model uncertainty became possible (e.g., George and McCulloch 1993, Green 1995, Geweke 1996, George and McCulloch 1997, Raftery et al. 1997, Kuo and Mallick 1998, Hoeting et al. 1999). A substantial amount of work went into improving algorithms' computational efficiency (e.g., Clyde et al. 1996, Clyde 1999, Clyde et al. 2011) and exploring their sensitivities to different prior specifications (e.g., Chipman 1996, Chipman et al. 2001, Link and Barker 2006, Feldkircher and Zeugner 2009). Relatively little work has gone into assessing the practical implications of incorporating model uncertainty into analyses, particularly for explanatory (rather than predictive) goals. Concerns raised previously have received little formal attention or are still unsettled within the statistical community, as evidenced by recent publications (Cade 2015, Fieberg and Johnson 2015, Hooten and Hobbs 2015, Ver Hoef and Boveng 2015).

Bayesian model averaging was originally developed as a method for improving out-of-sample predictions by combining predictions from multiple models with weights based on their posterior model probabilities. Algorithms for obtaining posterior model probabilities were developed with this predictive goal in mind. Software to implement model averaging has developed to the point where researchers can almost automatically obtain results regardless of their statistical backgrounds and the potential benefits of model averaging are being advertised in a very broad sense. In general, discussions about

Manuscript received 21 April 2016; revised 21 April 2016; accepted 31 May 2016. Corresponding Editor: K. Ogle.

[2]Present address: Neptune & Company, Inc., Bozeman, Montana 59715 USA

[3]E-mail: katharine.banner@montana.edu

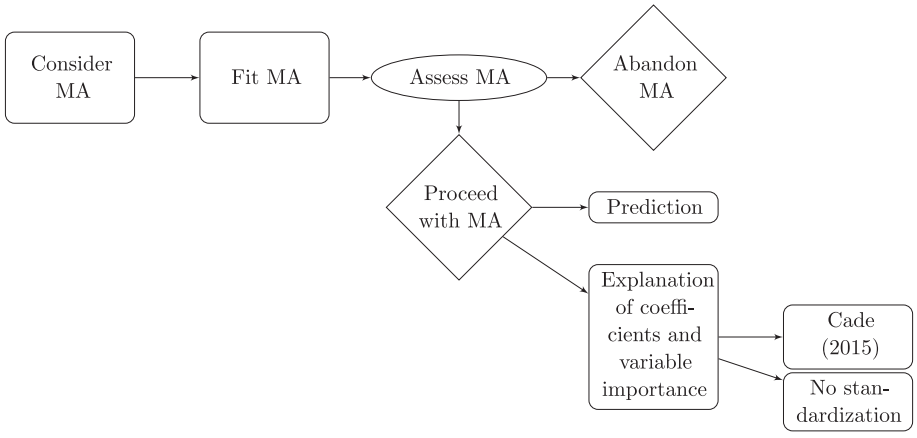

FIG. 1. This flow chart depicts steps a researcher could take when considering model averaging. Much of the model-averaging literature falls under the rectangular nodes. Articles related to algorithms and computational efficiency fall under Fit MA (e.g., Clyde , 1999, 2012, Link and Barker 2006, Feldkircher and Zeugner 2009, Clyde et al. 2011, Barker and Link 2013), articles related to predictive performance fall under Prediction (e.g., Raftery et al. 1997, Hoeting et al. 1999), and literature related to using model averaging for explanatory inferences falls under Explanation of coefficients and variable importance (e.g., Burnham and Anderson 2002, Cade 2015). Some of the literature raises concerns with particular aspects of model averaging and multimodel inference (e.g., Cade 2015, Fieberg and Johnson 2015, Ver Hoef and Boveng 2015). This paper falls under the Assess node.

the benefits of model averaging have blurred the distinction between predictive and explanatory inference, leading researchers to believe model averaging is always an advantage. This is concerning because the interpretation of partial regression coefficients can depend on other variables that have been included in the model, so averaging regression coefficients across models may not be practically meaningful.

In their text geared towards biologists, Burnham and Anderson (2002) advocate for accounting for model uncertainty using AIC-based (Akaike's information criterion; non-Bayesian) model averaging in problems where the estimates of coefficients are of primary interest (not just for prediction). Similarly, Montgomery and Nyhan (2010) advocate for Bayesian model averaging for political scientists, arguing it can "help applied researchers to ensure that their estimates of effects of key independent variables are robust to a wide range of possible model specifications." In their widely cited paper, *Bayesian model averaging: a tutorial*, Hoeting et al. (1999) suggest that model averaging for prediction can help researchers avoid having to defend a particular choice of model, with the benefit of a simplified presentation. While these are attractive qualities for predictive inference, we do not believe these statements were intended to extend to model averaging of regression coefficients.

The advertised allures of model averaging are strong, and we see researchers being pulled toward the method regardless of their research objectives. Our observations and concerns are, in general, consistent with those recently raised by Cade (2015), who argues that the use of model averaging of regression coefficients may result in misleading inferences while leaving researchers "strangely, feeling satisfied that model uncertainty has been addressed." Cade (2015) suggests using partial standard deviations to adjust the regression coefficients for their changing scales among models (due to multicollinearity).

There are many things a researcher must consider before deciding if model averaging is useful and appropriate for a particular problem (Fig. 1). After the decision to use model averaging has been made, there are additional considerations and decisions to be made about how model averaging will be implemented, such as whether to employ the standardization methods described in Cade (2015). Much of the model averaging literature falls into categories represented with rectangles in Fig. 1 (e.g., Consider MA, Fit MA, etc.), with little falling into the Assess MA category. In this paper, we focus on the assessment step in the model averaging process.

Assessing the appropriateness of model averaging must be done on a case-by-case basis, as it is difficult to understand the model-averaged result without understanding how the individual model results combine to create it. We look closer at the difference between model-averaging predictions and regression coefficients, review foundations of linear regression, suggest helpful notation, and introduce graphical tools to help understand how results from individual models are combined to form model averaged results. We also highlight the importance of making modeling decisions in the context of the research questions, which we term question-focused modeling. These considerations help researchers make informed decisions about model averaging on a case-by-case basis and provide a foundation for arguing against its use in cases where it is unnecessary or inappropriate.

We use two examples (see *Example 1: Haul-out Behavior of Weddell Seals* and *Example 2: When Prediction Leads to Explanation*) with different analysis

objectives to show how these considerations and graphical tools can be used in practice. In our first example, we demonstrate the practical importance of considering whether the added complexity of model averaging is worth the potential gains, particularly if inferences change very little compared to conditioning on one reasonable model. We highlight question-focused modeling and illustrate advantages it can have over model averaging. In our second example, we point out the often hidden secondary goals of analyses even when the stated goal is prediction. We use both examples to illustrate how we can assess the implications of the model-averaging process relative to the problem at hand.

The information in this paper should be relevant to a broad spectrum of researchers, ranging from those with little statistical background to quantitative ecologists and applied statisticians. The information presented is meant to be foundational for some, a refresher for others, and for all, an aid for assessing when and why model averaging might be used as an effective research tool.

Model Averaging Within Multimodel Inference

Multimodel inference is an umbrella term for incorporating multiple models into a single analysis, including both model selection (selection of a model for inference from a clearly defined set of models) and model combination (e.g., model averaging; Hooten and Hobbs 2015). To conduct multimodel inference, a model set $\mathcal{M} \equiv \{M_1, M_2, ..., M_J\}$ must be defined, where J is the total number of models considered. Ideally, $\mathcal{M}$ is specified using sound science and expert knowledge prior to observing the data (e.g., Burnham and Anderson 2002). However, for processes that are not well understood and do not have a pre-defined set of potential models, it is common for researchers to define $\mathcal{M}$ as the set of all possible regression models made up from combinations of a set of potential input variables. The consideration of all first-order combinations of quantitative input variables without any interactions is commonly referred to as *all subsets* regression and is the default in multiple software packages. Other strategies for choosing model sets are discussed elsewhere (e.g., Doherty et al. 2012).

Model averaging was originally developed in a Bayesian framework, where all unknowns are modeled with probability distributions and the hierarchy stemming from models and parameters within models is naturally incorporated. Knowledge about parameters before data are collected (prior) is combined with information from the data (through the likelihood) to form posterior distributions for all unknowns of interest, as opposed to focusing only on a likelihood function for inference. Including the model set $\mathcal{M}$ as an unknown allows model uncertainty to be directly incorporated into an analysis. Prior distributions must be placed on $\mathcal{M}$ and all of the parameters in each model. A prior for the discrete random variable $\mathcal{M}$ is defined by assigning prior probabilities to each model in $\mathcal{M}$, such that $p(\mathcal{M}) = \{\Pr(M_1), \Pr(M_2), ..., \Pr(M_J)\}$, and $\sum_{j=1}^{J} \Pr(M_j) = 1$. Similarly, we denote the joint prior distribution on the parameters in the jth model as $p(\boldsymbol{\theta}_j|M_j)$, where $\boldsymbol{\theta}_j$ is the parameter vector for the jth model (including regression coefficients and any other parameters in model M_j, such as, but not limited to, (co)variance terms). The collection of posterior model probabilities $\{\Pr(M_j|\mathbf{y})$ for $j = 1,2,...,J\}$ defines the posterior distribution for $\mathcal{M}$, conditional on the observed data ($\mathbf{y}$). The posterior model probability for an individual model is connected to the prior model probability and the distribution $p(\mathbf{y}|M_j)$ through Bayes' theorem:

$$\Pr(M_j|\mathbf{y}) = \frac{p(\mathbf{y}|M_j)\Pr(M_j)}{\sum_{j=1}^{J} p(\mathbf{y}|M_j)\Pr(M_j)} \quad (1)$$

where

$$p(\mathbf{y}|M_j) = \int_{\boldsymbol{\theta}_j} p(\mathbf{y}|\boldsymbol{\theta}_j, M_j) p(\boldsymbol{\theta}_j|M_j) d\boldsymbol{\theta}_j. \quad (2)$$

Specifically, the likelihood function associated with model M_j is based on $p(\mathbf{y}|\boldsymbol{\theta}_j, M_j)$, and $p(\mathbf{y}|M_j)$ is obtained by integrating over the posterior distribution of the parameters in M_j. The integration is typically intractable and therefore done computationally. Note in Eq. 2 that the prior distribution for the parameters is contributing to $p(\mathbf{y}|M_j)$, which in turn contributes to the posterior model probabilities.

To provide clear meaning to the posterior model probabilities, it is natural to assume truth is in the model set so that the posterior model probabilities represent the probability that a model is true, given the data and the priors. While this assumption is surely false in practice, we appeal to the rationale presented in Link and Barker (2006) and Barker and Link (2015) as to why it is reasonable to proceed. Collectively, the posterior model probabilities are also used to combine posteriors of quantities of interest from multiple models with the goal of incorporating model uncertainty into inferences and are also often used to discuss the degree of uncertainty in the model set.

Let ϕ be a quantity of interest, such as a prediction of a new observation or a regression coefficient. To generally define a posterior distribution for a model-averaged ϕ, the marginal posterior distributions of ϕ from individual models in $\mathcal{M}$ are combined using posterior model probabilities to form the distribution commonly written as

$$p(\phi_{\mathrm{MA}}|\mathbf{y}) = \sum_{j=1}^{J} Pr(M_j|\mathbf{y}) p(\phi|M_j, \mathbf{y}). \quad (3)$$

Eq. 3 is useful conceptually, but sometimes, as we will see in Eq. 9, the posterior distribution for a model-averaged ϕ can be more complicated than it appears. For more details about the formal setup of Bayesian model averaging that is accessible to ecologists and

environmental scientists, we refer the reader to Link and Barker (2010).

We present the fully Bayesian implementation of multimodel inference, but the concepts and discussion can be extended to the non-Bayesian approaches through a focus on combining point estimates and their standard errors, rather than on parameters and their posterior distributions. Formulas for the non-Bayesian context can be found elsewhere (e.g., Burnham and Anderson 2002, Link and Barker 2010). The broad use of Akaike's information criterion (AIC) among ecologists has naturally led to its popularity for use in model averaging, and while AIC and Bayesian information criterion (BIC) approximations initially seem a simpler option, practitioners should understand the implications of their choice for results (see Link and Barker 2006, Hooten and Hobbs 2015).

Model averaging in regression

In the context of linear regression, including generalized linear regression, model uncertainty typically enters through the variable selection process. For *all subsets* regression with p first-order input variables, $\mathcal{M}$ contains $J = 2^p$ elements, and a common motivating question is: Which of p potential input variables should define the model(s) ultimately used for inference? We follow convention and focus on the case where only first-order terms of p quantitative input variables are considered. The issues we discuss in the context of this common convention are even more relevant and complicated when higher-order terms such as interactions and polynomials are considered. Thus, the challenges in implementing Bayesian model averaging with higher order terms often naturally leads to their exclusion, and we discuss concerns with this in *Example 1: Haul-out Behavior of Weddell Seals*.

We use typical regression notation, defining $\mathbf{y}$ as a ($n\times1$) column vector of observations of a response variable and $\mathbf{X}$ as a ($n \times (p + 1)$) matrix with a column vector of 1s for the intercept and additional column vectors for the p potential input variables, $\mathbf{X} = [\mathbf{1},\mathbf{x_1},\mathbf{x_2},...,\mathbf{x_p}]$. Let $\boldsymbol{\beta}_j$ be the (p + 1) column vector of regression coefficients associated with M_j. Assuming normal, independent errors with constant variance ($\mathbf{y} \sim N_n(\boldsymbol{\mu}\{\mathbf{y}|\mathbf{X}\},\sigma^2 I_{n\times n})$), we have our familiar multiple linear regression model. The mean of $\mathbf{y}$ is modeled as a function of the first order terms of the p input variables and the intercept. More generally, it can be assumed that $\mathbf{y}$ follows another distribution (such as binomial or Poisson), and the mean of $\mathbf{y}$ can be connected to the input variables through a link function $g()$,

$$g(\boldsymbol{\mu}\{\mathbf{y}|\mathbf{X}\}) = \beta_{0,j} + \beta_{1,j}\mathbf{x}_1 + \beta_{2,j}\mathbf{x}_2 + \ldots + \beta_{p,j}\mathbf{x}_p = \mathbf{X}\boldsymbol{\beta}_j, \quad (4)$$

where $g()$ is simply the identity function for multiple linear regression.

We include an index for model in our notation for regression coefficients to make it explicit that the meaning of the regression coefficient associated with input variable X_v in model M_j ($\beta_{v,j}$) is model dependent. All models in $\mathcal{M}$ can be defined by excluding certain input variables from the fullest model, which is equivalent to setting the $\beta_{v,j}$ associated with those variables equal to 0.

A considerable amount of work has been put into investigating the implications and limitations of different prior specifications for $\mathcal{M}$ and $\boldsymbol{\theta}_j$ in a regression context (e.g., George, and McCulloch 1993, 1997, Carlin and Chib 1995, Geweke 1996, Raftery et al. 1997, Kuo and Mallick 1998, Feldkircher and Zeugner 2009) and also the implementation of Markov chain Monte Carlo (MCMC) and other sampling schemes, such as Bayesian adaptive sampling and reversible jump MCMC (e.g., Green 1995, Link, and Barker 2006, 2010, Clyde et al. 2011, Barker and Link 2013). In depth discussions or critiques of the different methods available are beyond the scope of this paper. We focus on understanding how model averaging may be used when the default (or easier to use) priors are chosen for implementation.

One of the most common priors for $\mathcal{M}$ is the discrete uniform prior, which places equal prior weight on each of the models considered, $\Pr(M_j) = 1/J$ for j = 1,2,...,J. Assuming normal likelihoods, forms of the conjugate normal prior for regression parameters are commonly used because of their computational advantages and availability in R (R Core Team 2016). One such prior is Zellner's g-prior (Zellner 1984, Feldkircher and Zeugner 2009), which specifies a vague multivariate normal prior on the partial regression coefficients in each model and an improper uniform prior on the standard deviation parameter. Specification of one hyper-parameter, g, allows the researcher to control how diffuse the prior will be, and the resulting posterior distributions for the regression coefficients from an individual model are multivariate t distributions (see Appendix S1 for details).

Knowing the form of the posterior distributions under the g-prior greatly simplifies the implementation of multimodel inference from a computational perspective because it only requires approximating the first two moments, rather than the whole posterior distribution. However, the validity of the multivariate t form of the posterior distributions is contingent upon no severe violations of regression assumptions for all models in $\mathcal{M}$. In the *all subsets* setting, the size of $\mathcal{M}$ grows exponentially with the number of input variables considered, which can make model checking prohibitive even for moderately sized p. This poses a major limitation for multimodel inference, as recently pointed out by Ver Hoef and Boveng (2015).

Model averaging of predictions

The original motivation for the development of model averaging was to improve prediction, with the quantity of interest being a new observation at specified values of the input variables, denoted $\phi_j = \tilde{y}_j$ for model M_j (consistent with Gelman et al. 2013). Although the values of

predictions will change among models, $\tilde{y}_j$ holds the same meaning for all models and can be directly compared among models or combined in a weighted average to provide an average prediction over models. It has been demonstrated that model averaging can be advantageous under many out-of-sample prediction criteria (e.g., Raftery et al. 1997, Hoeting et al. 1999).

Within Bayesian inference, we obtain a posterior distribution of predictions from M_j for a particular set of input variables in $\mathbf{X}^{\text{new}}$. This posterior predictive distribution $p(\tilde{y}|\mathbf{y})$ is defined by

$$p(\tilde{y}_j|\mathbf{y},M_j)=\int_{\boldsymbol{\theta}_j} p(\tilde{y}_j|\boldsymbol{\theta}_j,M_j)p(\boldsymbol{\theta}_j|\mathbf{y},M_j)d\boldsymbol{\theta}_j \tag{5}$$

and it describes the current knowledge about the prediction by combining uncertainty in the parameters of the model with variability coming from the distribution defining the likelihood (i.e., the two components on the right of Eq. 5).

We can think of the weighted average $\tilde{y}_{\text{MA}}=\sum_{j=1}^{J}\Pr(M_j|y)\tilde{y}_j$ as a model-averaged prediction in both the Bayesian and non-Bayesian context. In the Bayesian context, $\tilde{y}_{\text{MA}}$ is an unknown of interest with a distribution. Specifically, its distribution is the mixture distribution of the posterior predictive distributions from the individual models, with weights equal to the posterior model probabilities (as in Eq. 3). In the non-Bayesian context, the point prediction of $\tilde{y}_{\text{MA}}$, is constructed using the point predictions from the individual models, and formulas are available to obtain the approximate standard error of a model-averaged point prediction (see Burnham and Anderson 2002, Link and Barker 2010).

This approach of directly averaging predictions works for linear and nonlinear models, and even for combining predictions from models with different forms. However, for normal linear models, a shortcut using the regression coefficients has become the most common way to obtain model averaged predictions because it bypasses the need to actually obtain predictions (or posterior predictive distributions) for the J models. This shortcut is particularly useful when there are many more models than there are input variables. For normal linear models, the posterior prediction from M_j can be written $\tilde{y}_j=\mathbf{X}^{\text{new}}\boldsymbol{\beta}_j$, leading to

$$\begin{aligned}\tilde{y}_{\text{MA}}&=\sum_{j=1}^{J}\Pr(M_j|y)\tilde{y}_j\\&=\sum_{j=1}^{J}\Pr(M_j|y)\left[\mathbf{X}^{\text{new}}\boldsymbol{\beta}_j\right], \qquad (6)\\&=\mathbf{X}^{\text{new}}\left[\sum_{j=1}^{J}\Pr(M_j|y)\boldsymbol{\beta}_j\right]=\mathbf{X}^{\text{new}}\boldsymbol{\beta}_{\text{MA}}. \qquad (7)\end{aligned}$$

Again, in a Bayesian setting $\boldsymbol{\beta}_{\text{MA}}$ describes unknowns with distributions. In a non-Bayesian setting, carrying through point predictions and point estimates for regression parameters provides the same endpoint and illustrates the shortcut.

To better understand the equality in Eq. 7, we consider an example with two possible predictors, for which the *all subsets* model set is defined by four models: (intercept only, X_1, X_2, or X_1 and X_2) where $j = 1, 2, 3, 4$, respectively. The vectors of partial regression coefficients for each model can be written: $\boldsymbol{\beta}_1=(\beta_{0,1},0,0)^T$, $\boldsymbol{\beta}_2=(\beta_{0,2},\beta_{1,2},0)^T$, $\boldsymbol{\beta}_3=(\beta_{0,3},0,\beta_{2,3})^T$, and $\boldsymbol{\beta}_4=(\beta_{0,4},\beta_{1,4},\beta_{2,4})^T$. We rewrite Eqs. 6 and 7 to explicitly show the connection between model averaging of predictions and this shortcut through model averaged regression coefficients (for this equation we denote $\Pr(M_j|\mathbf{y})=w_j$),

$$\begin{aligned}\tilde{y}_{\text{MA}}&=\mathbf{X}^{\text{new}}\begin{pmatrix}w_1\beta_{0,1}+w_2\beta_{0,2}+w_3\beta_{0,3}+w_4\beta_{0,4}\\w_1\times 0+w_2\beta_{1,2}+w_3\times 0+w_4\beta_{1,4}\\w_1\times 0+w_2\times 0+w_3\beta_{2,3}+w_4\beta_{2,4}\end{pmatrix}\\&=\mathbf{X}^{\text{new}}\begin{pmatrix}\beta_{0,\text{MA}}\\\beta_{1,\text{MA}}\\\beta_{2,\text{MA}}\end{pmatrix}. \qquad (8)\end{aligned}$$

We suspect the current emphasis on interpreting model-averaged partial regression coefficients (as opposed to predictions) evolved from this practice. Model-averaged regression coefficients are mixtures of regression coefficients from different models (e.g., $\beta_{1,\text{MA}}=w_2\beta_{1,2}+w_4\beta_{1,4}$ in Eq. 8), and while the natural desire is for an "overall effect", we need to carefully think about what is actually represented by the weighted average. To build a foundation for this discussion, we revisit the properties of partial regression coefficients in the context of linear regression.

Model averaging of partial regression coefficients

A concept critical to understanding the potential challenges with interpreting model-averaged regression coefficients is that of partial regression coefficients. An appreciation of the meaning of the term partial is commonly overlooked, forgotten, or not given adequate attention in introductions to regression; it reflects the potential change in meaning of a regression coefficient associated with a particular variable when the other variables in the model change. To help illustrate this and discuss the challenges in the context of model averaging, we use a subset of data from Sacher and Staffeldt (1974) also used as an example in *The Statistical Sleuth* (Ramsey and Schafer 2013). These data are average values for brain weight (g), gestation length (days), and body size (kg) for 96 species of mammals. Natural log transformations on average brain weight (*lbrain*), body size (*lbody*), and gestation length (*lgest*) were performed as the relationships are approximately linear on the log–log scale.

In this example, primary interest is in the relationship between gestation length and mean brain weight. Given the available variables and the common model averaging choice of not including higher order terms, we consider

two models representing two distinctly different research goals: (1) we can investigate the overall relationship between gestation length and mean brain size using $\mu\{lbrain|M_g\} = \beta_{0,g} + \beta_{1,g}lgest$, and (2) we can investigate the relationship between gestation length and mean brain weight conditional on body size using $\mu\{lbrain|M_{gb}\} = \beta_{0,gb} + \beta_{1,gb}lgest + \beta_{2,gb}lbody$.

As one would suspect, body size and gestation length are collinear on the log–log scale ($r = 0.85$). For those taught to view multicollinearity as having only negative consequences, this degree of collinearity may seem alarming. However, the more interesting of the two questions (from a biological point of view) is the second. In an observational study like this one, it would have been impossible for the researchers to experimentally control body size while manipulating gestation length of mammals. However, using regression, we can assess evidence of a relationship between gestation length and mean brain weight, after accounting for the strong (but not very interesting) relationship between body size and brain weight. We fully accept the larger posterior variances (or standard errors) that are a consequence of having less information in the data to estimate the relationship of interest conditional on body size; it is a small price to pay to be able to address the specific question of interest.

Returning to the two models, we do not expect the relationship between gestation length and mean brain weight over all the mammal species to be the same as the relationship between gestation length and mean brain weight for mammal species with the same (or similar) body sizes. In other words, we do not expect the parameters $\beta_{1,g}$ and $\beta_{1,gb}$ to be equal. Note that with no interaction in the model, we are assuming a common $\beta_{1,gb}$ applies to all body sizes (i.e., the relationship between gestation length and mean brain weight is the same across all body sizes). To gain intuition for the potential difference in meaning between these two parameters, we use the data to create arbitrary categories of body size and estimate the slope and intercept within each category (Fig. 2). This visual aid helps us compare the estimated relationships between gestation length and brain weight for groups of mammals with similar body sizes to the estimated relationship observed over all mammals (ignoring body size). As expected, the estimated relationship changes when we group by similar body sizes, though the separate slopes look very similar, which is consistent with the exclusion of the interaction term in the second model.

Using exploratory tools like Fig. 2 can be extremely informative when there are two collinear variables and we want to control for one of them while making inference about the other. Similar tools are partial residual (or added variable) plots and partial regression plots (presented in Cade 2015), which can aid in understanding partial relationships captured in estimated coefficients for more than two explanatory variables. Although the scatterplot is limited to two explanatory variables, it provides a more intuitive way to visualize and illustrate the different information used to estimate the two different regression coefficients.

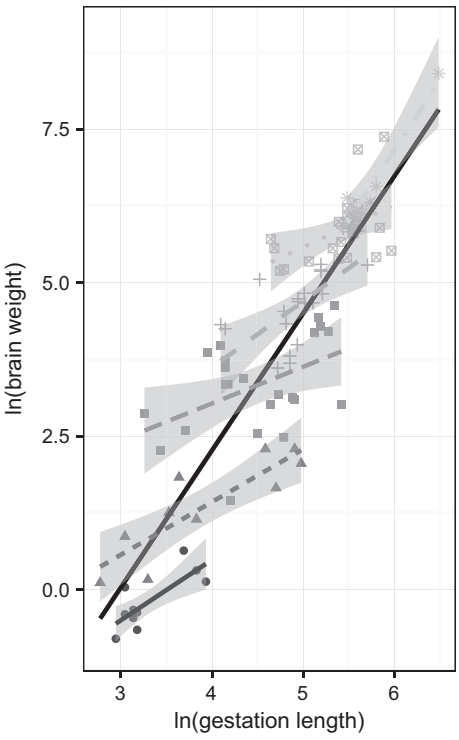

FIG. 2. Interest lies in the relationship between *lgest* and *lbrain*, so we categorize the continuous variable we wish to account for, *lbody*, into subgroups of similar body size and create a scatterplot with points coded by these groups. We call this type of plot a cut-and-coded scatterplot. The corresponding lines represent the estimated relationships between *lgest* and *lbrain* conditional on the subgroups (we can think about the average slope across the body size subgroups as an approximate estimate of $\beta_{1,gb}$). This figure was created using the ggplot function in the ggplot2 R package (Wickham 2009, R Core Team 2016). Code is available in Appendix S2. A script file, can also be found in Appendix Data S1 (Rcode-appendix-MAPP.R).

Conventional notation for regression coefficients does a poor job distinguishing between $\beta_{1,g}$ and $\beta_{1,gb}$ by labeling both parameters β_1, implying $\beta_{1,g} = \beta_{1,gb}$, which is rarely true. In general, the only case for which partial regression coefficients associated with a particular explanatory variable hold the same interpretation across models is when the explanatory variables are orthogonal. This case is only realized in balanced experimental designs and other settings where variables are constructed to be orthogonal (e.g., principle components, orthogonal polynomials); it is very rare in observational studies. Therefore, we recommend explicit notation connecting a partial regression coefficient to the model it is defined in, as also suggested by others (e.g., Hocking 2003, Jewell 2004, Cade 2015).

Vague notation has also contributed to the common mistake of misinterpreting partial regression coefficients as effects of the explanatory variable they precede, regardless

of which variables are in the model. We believe it also contributes to the attractiveness of capturing an "overall effect" through model averaging. Often, users fail to consider that (1) the meaning of the relationship associated with a partial regression coefficient is not exchangeable across models, and (2) the word "effect" implies a causal relationship between the input variable and the response, which is often inappropriate for observational studies without paying careful attention to principles and methods of causal modeling. We prefer interpreting partial regression coefficients as the linear relationship between the input variable and the mean response, after accounting for the other variables in the model.

Two components of the model-averaged posterior distribution.— The total probability of the posterior distribution of a model-averaged regression coefficient is made up of two components, one coming from the models where the coefficient is set to zero, and the other coming from the models where the coefficient is estimated. We explicitly include both components by differentiating between the subset of models including the vth input variable ($\mathcal{M}_v = \{M_j|\beta_{v,j} \neq 0\}$), and its complement, the subset of models excluding the vth input variable ($\overline{\mathcal{M}}_v = \{M_j|\beta_{v,j} = 0\}$). For the remainder of the paper, we will refer to the posterior distribution conditional on $\mathcal{M}_v$ as the continuous component, and the part conditional on $\overline{\mathcal{M}}_v$ as the zero component (describing a point mass at zero). The notation may seem cumbersome, but it is important to formally represent both pieces because inferences depend on which components are used. This property of the model-averaged posterior distribution is not intuitively obvious, and is rarely discussed in detail in the Bayesian model averaging literature, with the exception of Hoeting et al. (1999), who explicitly show it graphically.

The posterior distribution of the model-averaged partial regression coefficient associated with the input variable $\mathbf{x}_v$ is a mixed distribution, which can be written as

$$p(\beta_{v,\mathrm{MA}}|\mathbf{y}) = \begin{cases} \sum_{j\in\overline{\mathcal{M}}_v} \Pr(M_j|\mathbf{y}) & \text{for } \beta_{v,\mathrm{MA}} = 0 \\ \sum_{j\in\mathcal{M}_v} \Pr(M_j|\mathbf{y}) p(\beta_{v,j}|\mathbf{y}, M_j) & \text{for } \beta_{v,\mathrm{MA}} \neq 0. \end{cases} \quad (9)$$

It is most typical to see reference to the two parts separately, and for different reasons; the continuous component is used for summarizing "overall effects", and the zero component is used for measuring variable importance. The complement of the zero component is the posterior probability that the variable is included, termed the posterior inclusion probability (PIP). The PIP for input variable $\mathbf{x}_v$ is $\mathrm{PIP}_v = 1 - \sum_{j\in\overline{\mathcal{M}}_v} \Pr(M_j|\mathbf{y})$, and a variable is often referred to as important if its PIP is greater than or equal to 0.5 (threshold suggested in the AIC-based context, see Burnham and Anderson 2002, Burnham 2015).

The zero component can be ignored without any practical consequences when using model averaging for prediction, and so it is often forgotten or omitted when the model-averaging is extended to regression coefficients. When model-averaging regression coefficients, it is important to be aware of the point mass at zero, which is part of the posterior distribution and therefore is included when performing Bayesian model averaging. The inclusion of the zeros leads to shrinkage of regression coefficients toward zero. However, a common version of the AIC-based method for model-averaging regression coefficients conditions on $\mathcal{M}_v$ and therefore ignores the zero component in constructing the average (see Burnham and Anderson 2002, Cade 2015).

To illustrate the difference, suppose we are interested in three models with the following forms: M_1:$\beta_{0,1}$, M_2:$\beta_{0,2} + \beta_{1,2}\mathbf{x}_1$, and M_3:$\beta_{0,3} + \beta_{1,3}\mathbf{x}_1 + \beta_{2,3}\mathbf{x}_2$. Let the posterior distribution for $\mathcal{M}$ be $p(\mathcal{M}|\mathbf{y}) = \{0.34,0.33,0.33\}$. To model average the partial regression coefficients associated with $\mathbf{x}_1$ over all models ($\mathcal{M}$), we use both components to obtain $\beta_{1,\mathrm{MA}} = (0.34 \times 0) + (0.33 \times \beta_{1,2}) + (0.33 \times \beta_{1,3})$. When users choose to average over only the continuous component, the weights are normalized over the models in $\mathcal{M}_v$ to obtain $\beta_{1,\mathrm{MA}_{\mathcal{M}_v}} = (0.5 \times \beta_{1,2}) + (0.5 \times \beta_{1,3})$. Again, this latter type of weighting is atypical in Bayesian model-averaging, but common in AIC-based averaging, and therefore researchers should be aware of the differences for inference.

Graphical tools for assessing model averaging

In practice, regression analysis should include both exploratory data analysis and diagnostic plots to assess appropriateness of the proposed models (e.g., scatterplots, various plots of residuals, posterior predictive checks). Similar practices are also important, and perhaps even more so, for model averaging because we must assess the appropriateness of many individual regression models and also consider whether it is appropriate to combine them. However, what we have observed in practice suggests less exploratory data analysis and model checking is actually done in the context of model averaging, which in part can be attributed to prohibitively large or at least inconveniently large $\mathcal{M}$ (Ver Hoef and Boveng 2015). Hoeting et al. (1999) recommend checking the usual diagnostics for the fullest model in the model set before proceeding with model averaging for predictions, which is similar to common advice when considering a set of nested models for multiple linear regression. The method for checking assumptions for many models simultaneously is inherently context and situation dependent, and therefore we do not provide broad suggestions here. However, in our next two points, we do help with assessing whether it is appropriate to combine results from individual models, assuming assumptions are not severely violated.

First, we have observed that choosing model averaging as an inferential tool may lead to a lack of consideration of higher order terms such as interactions or

polynomials. This seems to be done out of convention or ease of implementation and is often not justified by plots of the data or subject matter expertise. Many of the usual regression plotting tools could help researchers identify interactions and higher order terms to include in an analysis. Including such terms complicates the model averaging process and the researcher will need to restrict the *all subsets* model set by forcing interactions and higher order terms into models together (Chipman 1996). We provide an example of forcing variables into a model in *Example 1: Haul-out Behavior of Weddell Seals*, but we do not discuss the implications for the prior model probabilities when the model set is manipulated in this way because it is beyond the scope of this paper. We refer the reader to Chipman (1996) and Clyde and George (2004).

Second, we see the need for an accessible graphical tool to help researchers efficiently compare results from individual models and understand how they are combined into model-averaged posterior distributions. Such a tool allows researchers to assess the implicit assumption that the parameters (or estimates) being averaged have a common (enough) meaning across models and to understand the weights given to each. Software for model averaging makes it relatively easy to obtain the posterior for the model averaged partial regression coefficient, but the ability to visualize how individual posteriors are combined to create it is absent. Another common and slightly more informative visualization tool displays how the signs of the posterior means of partial regression coefficients associated with each input variable change across the top models (commonly called an image plot). However, the individual posterior distributions could have means with the same sign, and still be noticeably different. A comparison of the entire posterior distribution is much more desirable.

We developed the model averaged posteriors plot (MAP plot hereafter) to provide a visual summary of all components going into, and resulting from, the averaging of partial regression coefficients across models. The plot is designed to be used in the process of deciding whether model averaging is an appropriate tool for inference, not as a way to display the results of model averaging at the end of an analysis. An effective way to assess whether model averaging is potentially appropriate for a particular case is to actually implement it and then use the MAP plot to carefully digest and critically evaluate the potential usefulness of model averaging for the problem at hand. This plot is not meant to replace the usual regression diagnostics and modeling decisions mentioned previously, but it can help identify models that should be checked. For example, if model averaging makes sense for a particular example, and the model-averaged result is essentially a combination of five models, the researcher could easily check diagnostics for those five models. In the following sections, we use the MAP plot for assessing model averaging in two examples with different analysis objectives.

Example 1: Haul-out Behavior of Weddell Seals

In 2010, researchers studying Weddell Seal (*Leptonychotes weddellii*) haul-out behavior in Erebus Bay, Antarctica, set up cameras to take photos of a portion of Big Razorback haul-out site (Fig. 3a). Photos were taken every 45 minutes for the purpose of aligning the time of day for data collection to the time of day most seals were expected on the ice. One of the peak haul-out times is between the hours of 8:00 and 20:00, so we use only the data collected during these convenient times. Other explanatory variables thought to be related to the number of seals on the ice at a given time were also collected at each time point. The specific goal of this analysis is to make inference about the time of day associated with maximum mean seal counts, after accounting for : wind speed (m/s), temperature (° C), and tide height (m). Note that interest in the maximum necessitates a quadratic term for time of day.

For this data set and question of interest, we have already identified a useful model (up to consideration of interactions). However, suppose we are asked by reviewers to use model averaging to incorporate model uncertainty into the analysis (not an unrealistic scenario given our personal experiences). In an attempt to satisfy the reviewers, but still keep the question of interest in sight, we force the linear and quadratic components of time of day into all models in $\mathcal{M}$. It is possible that interactions should be investigated to allow the relationship between time of day and maximum expected seal count to depend on the values of the other explanatory variables, but we follow the common implementation of model averaging and do not include interactions in our model set. For this analysis, models are indexed by the explanatory variables they include. We define $\mathcal{M} \equiv \{M_T, M_{Tw}, M_{Tt}, M_{Td}, M_{Td}, M_{Twt}, M_{Ttd}, M_{Twd}, M_{Ttwd}\}$, where T is defined to include both the linear and quadratic components of time of day (time and time^2), and others are defined as w = wind speed, t = temp, and d = tide.

Given the count data, we could implement a Poisson generalized linear model using reversible jump MCMC (RJMCMC) similar to Link and Barker (2006). However, in this case all seal counts are greater than zero, averaging 15.79 seals, and natural log-transformed counts reveal that linearity, normality, and constant variance assumptions for normal linear regression are reasonable. Therefore, we expect Poisson log-linear regression and normal linear regression on the log-transformed counts to provide very similar results for this example, and they do. We choose to use the simpler and computationally faster alternative because the choice of method does not change the conclusions or the points we wish to make with this example. Thus, the models in $\mathcal{M}$ are of the form $\mu\{\log(\mathbf{y})\} = \mathbf{X}\boldsymbol{\beta}_j$, where errors are assumed independent and normally distributed.

FIG. 3. A digital image used for enumerating Weddell seals on the fast ice in the Big Razorback haul-out site, Erebus Bay, Antarctica. This image was taken at 17:30 on 18 October 2010. Weddell seal imagery obtained under authority of permit NMFS Permit No. 1032-1917-02. The data for this example are in Appendix Data S1 (seal-count-data.csv). Code is available in Appendix S2. A script file, can also be found in Appendix Data S1 (Rcode-appendix-MAPP.R).

There is an obvious violation of independence through temporal autocorrelation among residuals within a day. However, the violation exists for every model in the model set and ignoring it for this exercise does not affect the comparisons we use to illustrate our points. Preliminary analyses revealed an AR(1) correlation structure was sufficient to account for the correlation within days. Incorporating an AR(1) correlation structure into the analysis would result in us recommending that the researchers start sampling about 40 minutes earlier and end sampling about 40 minutes later than if the correlation was left unaccounted for. A sampling effort will span multiple hours, so the practical implications of the difference in results are minimal given the objective of the analysis.

The quantity of interest is the time of day that maximizes the expected number of seals, which can be expressed as a function of the regression coefficients associated with time and time^2 using the formula for the vertex of a parabola, $\{\text{Time at max}\}_j = -(\beta_{\text{time},j}/2\beta_{\text{time}^2,j})$ for M_j. Using Bayesian inference, it is straightforward to computationally obtain a posterior distribution for this quantity, and in a likelihood setting, the delta method is a common choice for obtaining an associated standard error for the point estimate.

We implement Bayesian model averaging with the bms function in the BMS R package (Feldkircher and Zeugner 2009, R Core Team 2016). We use the default unit information formulation of the g-prior ($g = n = 769$) on the parameters and a discrete uniform prior on $\mathcal{M}$. The standard output returned by the bms function tabulates posterior means, standard deviations for the model averaged partial regression coefficients, posterior inclusion probabilities for each explanatory variable, and the frequency with which the posterior means are positive across models (which gets at the partial regression coefficient idea, but requires a sign switch, see *Graphical tools for assessing model averaging*). The user can also easily extract posterior model probabilities and other posterior summaries from the bms output for all models in $\mathcal{M}$. Based on the prior we use, we can appeal to analytical results providing the form of the posterior distributions as t-distributions using moments obtained directly from bms (details available in Appendix S1 and S2). We display all relevant information about the individual models and the resulting model-averaged posterior distributions in a MAP plot (Fig. 4).

For this example, the MAP plot is composed of five panels, one for each of the explanatory variables considered. A column (panel) displays the distributions of partial regression coefficients associated with the same explanatory variable across models. A row (going across panels) provides the posterior distributions for the partial regression coefficients from a particular model, with the model-averaged posterior distribution in the bottom row. The probability associated with the point mass at zero of the model-averaged distribution is provided as text. All explanatory variables were standardized to have zero mean and unit standard deviation prior to analysis due to large differences in their ranges (Fig. 4). An alternative

form of standardization proposed in Cade (2015) uses partial standard deviations to adjust for changing multicollinearity among explanatory variables. However, as explained in *Graphical tools for assessing model averaging*, we are focusing on the early stages of assessing the appropriateness of model averaging, rather than deciding how to present results of model averaging. Creating the MAP plot at this stage promotes investigation of the differences in results across models before making the decision to possibly implement the standardization in Cade (2015).

There are several observations gleaned from the information provided by the MAP plot (Fig. 4). First, we observe that only two of the eight models received non-negligible posterior mass (M_{Ttd} with posterior model probability 0.884 and M_{Ttwd} with posterior model probability 0.116). The remaining models had such small posterior model probabilities that, in this case, model averaging is the result of a weighted average of only two models. The PMPs as visualized through the MAP plot is a useful tool for identifying when results of model averaging are a combination from a few models.

We also observe that the continuous component of the posterior distributions for the model-averaged coefficients associated with tide and temp are essentially the same as the model-averaged posterior including both the zero and continuous components. For example, the panel for tide shows that all models excluding tide have very small posterior model probabilities (given as ≈ 0, with values $\ll 0.0001$). Therefore, the probability associated with the point mass at zero for the model-averaged coefficients associated with tide is essentially zero, and the total posterior probability is almost entirely contained in the continuous component. Similar observations hold for temperature. The panels for time and time^2 have posteriors for all models because both explanatory variables were forced into all models in $\mathcal{M}$, and therefore, the model averaged posterior distributions for time and time^2 do not have zero components.

In contrast to the previous observations, we highlight the panel summarizing the partial regression coefficient for wind. One of the models excluding wind (M_{Ttd}) has large posterior model probability (0.884), and the other models excluding wind (M_{Tt}, M_T, and M_{Td}) have small posterior model probabilities (given as ≈ 0, with values $\ll 0.0001$). The zero component of the model-averaged distribution associated with wind is the sum of posterior model probabilities from these models, totaling approximately 0.884, and the continuous component comes from just one model, M_{Ttwd}, with posterior model probability 0.116 ($1 - 0.884 = 0.116$). Therefore, the model-averaged distribution associated with wind is essentially made up of one model contributing to the zero component and one model contributing to the continuous component.

It is also clear from the plot that while the posterior model probability for the largest model M_{Ttwd} is small, the posterior distributions for partial regression coefficients from M_{Ttwd} are similar to their model averaged counterparts. This is largely due to the fact that the addition of wind to the model with the highest posterior model probability, M_{Ttd}, did not appreciably change the posterior distributions of the partial regression coefficients for temp, tide, time, or time^2. Observations such as these are meant to encourage readers to use the MAP plot to further investigate how information from individual models combines to create the model averaged posterior, particularly in the presence of strong multicollinearity, and also to evaluate whether model averaging makes sense given the question of interest and desired inferences.

Now, we use the MAP plot to help assess averaging of the quantity of interest, time of day associated with maximum mean seal counts. We use posterior draws of $\beta_{\text{time},j}|\mathbf{y},M_j$ and $\beta_{\text{time}^2,j}|\mathbf{y},M_j$, to create posterior distributions for $\{\text{Time at max}\}_j$ for all models, and then create the model averaged distribution $\{\text{Time at max}\}_{\text{MA}}$ (Fig. 5). Recall that inference for the quantity of interest could have been directly addressed using the model including all of the available covariates. Therefore, it is interesting to compare inferences drawn from model averaging to those drawn from the largest model, which is straightforward because the posterior distribution of $\{\text{Time at max}\}_{\text{MA}}$ does not have a zero component. We compare a posterior credible interval for $\{\text{Time at max}\}_{\text{MA}}$ to that for $\{\text{Time at max}\}_{Ttwd}$. After back-transforming, we have 95% posterior credible intervals of 16:13–17:13 for the model averaged result and 16:14–17:14 for the largest model. There is a direct interpretation for the posterior interval coming from the largest model: There is a 95% chance the time of day at which the maximum number of seals are on the ice is between 16:14 and 17:14, after accounting for tide, wind, and temp. However, the interpretation of the interval from the model-averaged result is less clear because we average over two models that do not account for the same covariates. In this example, we used model averaging to incorporate model uncertainty but ended up with a posterior interval no wider than what we would have obtained had we chosen a reasonable model based solely on the question of interest.

By definition, the variance of the model-averaged quantity is based on the entire posterior distribution, including both the zero and continuous components. It is not always reasonable to compare posterior variances from individual models to the posterior variance coming from only the continuous component. The posterior variance of the model-averaged quantity of interest can be larger than the variance based only on the continuous component when the zero component is large and the continuous component is far from zero. Similarly, it can be smaller when the zero component is large and the continuous component is near zero. The only coefficient in this example with this potential is that associated with wind, where the variance from the continuous component of the model-averaged distribution, Var_{cts}, is larger than the variance from the model averaged distribution considering both the zero and continuous components, Var_{MA}, ($\text{Var}_{\text{MA}} = 3.51 \times 10^{-5}$, $\text{Var}_{cts} = 9.22 \times 10^{-5}$). For the other variables, it is reasonable to compare the individual model posterior variances to those of the

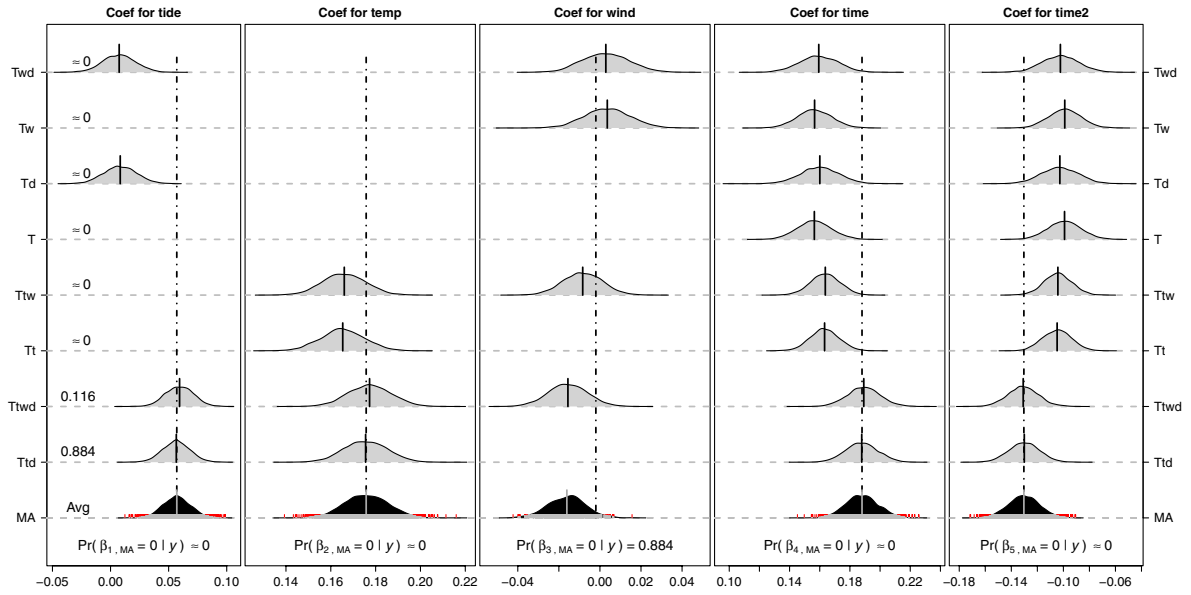

FIG. 4. The model-averaged posteriors (MAP) plot is split into five panels, one for each of the explanatory variables considered. Rows across panels display output from one model and are ordered by increasing magnitude of posterior model probability (printed along the leftmost y-axis. Probabilities given as ≈ 0 have a value < 0.0001.). The posterior distribution of the model-averaged partial regression coefficient associated with each explanatory variable is located at the bottom of the plot (continuous component is shown as a density plot and the zero component is shown in text). Note that the point mass associated with the zero components associates with the coefficients time and time^2 are actually exactly zero because they appear in every model (although the label suggests otherwise). The vertical line shows the posterior mean of the model averaged distribution. We created the MApp R package (Kampstra 2008, Feldkircher and Zeugner 2009, Albert 2014, Wickham and Francois 2015, R Core Team 2016) that includes a MAP plot function to automatically construct the plot from different sources of model averaging output, along with a vignette. This package is available at https://github.com/kbanner14/MApp-Rpackage, Code to recreate the figure is available in Appendix S2, a script file, can also be found in Appendix Data S1 (Rcode-appendix-MAPP.R).

continuous component of the model-averaged posterior because the probability associated with the zero component is essentially 0, and therefore the point mass at zero is not contributing to the variance. These comparisons can be easily made using Figs. 4 and 5 and the posterior standard deviations in Table 1. In this example there is little difference. This may seem unexpected, as discussions of multimodel inference lead us to expect that accounting for model uncertainty should always result in increased posterior variance (or decreased precision).

From this example, we see that the incorporation of model uncertainty into an analysis does not always result in what we expect, and there is a need for more scrutiny on a case-by-case basis. This is important information for weighing whether the complexity of model averaging, along with its often added simplifying assumptions, is worth the supposed gains. When a research question is well defined, it can be addressed by building a question-driven model justified by the objective. This can be achieved by clearly defining the relationship(s) of interest, considering what variables should be accounted for (or controlled for) so that the relationship(s) of interest can be addressed, and possibly iterating on that model, as described by Ver Hoef and Boveng (2015). This process of building such a model can seem difficult because of the effort needed in justifying decisions and assumptions. Therefore, it is often tempting to elicit a more automatic approach to find the best model or combine results over multiple models. While model averaging may hold the allure of a decision-free model building technique (just use all of them), we must consider whether the added complexity is necessary. In this example, we believe model averaging was unnecessary for the following reasons: (1) there was a well defined research question that could have been addressed directly by a single model (the largest model), and (2) inferences from the largest model were nearly identical to the model averaged result.

EXAMPLE 2: WHEN PREDICTION LEADS TO EXPLANATION

In this section, we revisit an example from Hoeting et al. (1999) using data from Penrose et al. (1985) and Johnson (1996). The objective is to predict percentage of body fat with age, weight, height, and 10 body circumference measures. Training data are available for 251 adult males between the ages of 22 and 81 years. We refer to the variables as predictors rather than explanatory variables, because the analysis objective is prediction, rather than explanation using regression coefficients. We choose this example because these data appear multiple times in the model averaging literature (e.g., Hoeting et al. 1999, Burnham and Anderson 2002, Zeugner 2011, Burnham 2015).

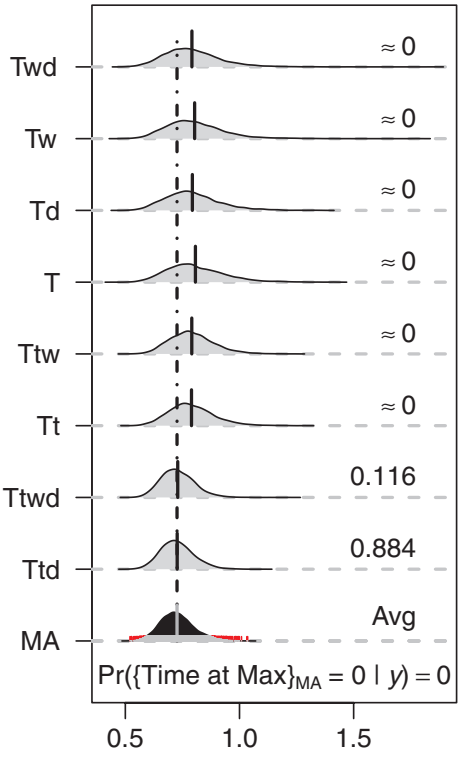

FIG. 5. Posteriors for the time at max parameter were created by finding the time of day corresponding to the maximum log total seals on the ice, which is a function of the partial regression coefficients associated with time and time^2 from Fig. 4. The MAP plot for time at max shows the posterior distributions for the quantity of interest across the models in $\mathcal{M}$ along with the model averaged counterpart. Code to recreate the figure is available in Appendix S2, a script file, can also be found in Appendix Data S1 (Rcode-appendix-MAPP.R).

Even when the stated goal of model averaging is to improve predictive ability, it is often attractive to use the results to quantify the relative importance of predictors and interpret their overall effects. Both of these secondary goals typically lead to predictors being considered in isolation, outside the context of models created with collections of predictors. As previously discussed (*Model averaging of predictions*), the appropriateness of model averaging for prediction does not imply it is appropriate for explaining overall effects. Careful examination of the individual model results using the MAP plot can help researchers decide if reporting conclusions focused on a variable in isolation may be appropriate.

We keep the objective consistent with Hoeting et al. (1999) and suppose the researchers have no a priori information about which variables are useful for predicting mean percent body fat and no specific question of interest regarding any of the variables beyond using first order combinations of the variables for prediction. We use the same *all subsets* model set with $J = 2^{13} = 8{,}192$ models, the same discrete uniform prior on $\mathcal{M}$, and a formulation of the *g*-prior for the regression coefficients that approximates the prior used in Hoeting et al. (1999; see Zeugner 2011) and the bms function (Feldkircher and Zeugner 2009).

The bms implementation of this example computes posterior model probabilities for all models in $\mathcal{M}$ and by default stores only the top 500. The results for this analysis indicate a substantial amount of model uncertainty; the 10 models with the highest posterior model probabilities account for about 74% of the posterior probability, and the top 500 models capture about 99.8%. We create model-averaged distributions using only these 500 models. Using the MAP plot, we display results for three predictors (weight, abdomen circumference, and wrist circumference) using the 10 models with highest probabilities (Fig. 6). The reader is encouraged to use the plot to investigate how information related to these variables in different models is combined to form the model-averaged results.

In *Two components of the model-averaged posterior distribution*, we introduced the PIP for the *v*th variable (PIP_v) as the sum of the posterior model probabilities over the models including that variable ($\mathcal{M}_v$). PIPs are often interpreted as relative measures of overall importance for each explanatory variable over all models considered. See Cade (2015) for an alternative measure of relative importance and Burnham (2015) for a recent discussion about using AIC weights for this purpose. Rather than repeat arguments here, we tie our discussion of overall importance to the previously discussed implications of ignoring the partial in partial regression coefficients (*Model averaging of partial regression coefficients*).

TABLE 1. Posterior standard deviations for the partial regression coefficients associated with each explanatory variable (each row) are shown for each model in the model set.

	MA	Ttd	Ttwd	Tt	Ttw	T	Td	Tw	Twd
Tide	0.0119	0.0119	0.0124	…	…	…	0.0138	…	0.014
Temp	0.0097	0.0098	0.0098	0.0097	0.0096	…	…	…	…
Wind	0.0095	…	0.0096	…	0.0096	…	…	0.0112	0.0113
Time	0.0107	0.0107	0.0109	0.0096	0.0095	0.0111	0.0126	0.0113	0.0128
Time^2	0.0119	0.0118	0.0121	0.0106	0.0107	0.0126	0.0139	0.0125	0.0141
Time at max	0.0716	0.0717	0.0711	0.0971	0.0987	0.1236	0.1215	0.1243	0.1185
PMP	…	0.8837	0.1158	4e-04	0	0	0	0	0

Notes: The columns define models in the model set (columns Ttd–Twd). The posterior standard deviation coming from the continuous component of the model averaged distribution is shown in the MA column. The row for time at max was appended to the results. This table is standard output from our plotting function, MApp_bms which was programmed in R (R Core Team 2016). Code is provided in Appendix S1 and S2.

The posterior model probabilities provide information about a collection of predictors working together to model the mean response, and each predictor occurs in isolation in only one model. Making statements about the importance of a single predictor with a posterior inclusion probability may imply (particularly to those with little background in regression or model averaging) that the predictor alone will be an effective predictor, regardless of what else accompanies it in the model. We show why such statements are potentially misleading by considering the PIP for wrist ($\mathbf{PIP}_{\text{wrist}} = 0.55$) which is large enough to be considered "important." Now, consider how wrist performs by itself as compared to when it is coupled with weight and/or abdomen. The model with wrist alone does not even rank in the top 500 with respect to posterior model probability, but when wrist is combined with weight and abdomen, the resulting model ranks second (i.e., after accounting for abdomen and weight, wrist adds meaningful information for the prediction). However, one would not want to collect only information about wrist circumference to predict percent body fat. We bring up this point to promote discussion of the issues that arise when the importance of a variable is assessed in isolation and to encourage careful thought about which quantities are chosen as a basis for conclusions.

Discussion

In the last 20 years, it has become relatively easy to implement model averaging with automatic priors and closed form posteriors, and also through non-Bayesian approximations (e.g., AIC or BIC weights). This ease of implementation, coupled with the allures of addressing model uncertainty and avoiding having to justify choosing one model, have made model averaging an increasingly popular method with many researchers in the ecological and environmental sciences. We suggest a critical look at this method, especially when it is used with partial regression coefficients, before it becomes more commonplace.

In general, potential benefits of new methods are readily communicated, popularized, and propagated through the peer review process, while challenges in using them are often left in the shadows. We believe statisticians have a responsibility to help researchers navigate the hard decisions about whether methods are appropriate or necessary for a particular problem. We are not the first to raise red flags suggesting a more careful look at multimodel inference and model averaging, and many of our concerns are consistent with those expressed elsewhere (e.g., Thomas et al. 2007, Cade 2015, Fieberg and Johnson 2015, Ver Hoef and Boveng 2015). In this paper, we offered researchers accessible information to help provide a basis for critically evaluating model averaging of partial regression coefficients on a case-by-case basis.

In *Model Averaging Within Multimodel Inference*, we provided notation for partial regression coefficients and the posterior distribution for model averaged regression coefficients. While subtle, these suggested changes for notation have the potential to remind the user of the implicit assumption underlying model averaging regression coefficients and may prevent errant use of the method.

In *Model averaging of partial regression coefficients*, we briefly discussed question-focused modeling, which we believe is directly tied to discussions about the challenges of model-averaging regression coefficients. In general, we find linear regression is a useful tool for three purposes: (1) estimating relationships between an explanatory variable and the mean response, often after accounting for other control variables, (2) estimating effects from experimental designs, and (3) building predictive models. It is typically easier in designed experiments to think about what variables should be controlled for or accounted for when making the treatment comparisons. However, this same thought exercise can, and should, extend to observational studies. Often, we can identify the relationship(s) of interest, think about what we would have controlled for in the design if random assignment had been possible, and then use regression as a way to adjust for the identified variables. This can be thought of as working toward ceteris paribus (holding all else constant except the treatment assignment), even in observational studies. This describes question-focused modeling because the research question drives the choice of what variables to control for rather than allowing the analysis method to automatically choose them (method-focused modeling). Fieberg and Johnson (2015) and Ver Hoef and Boveng (2015) agree we have much to gain by carefully considering what variables are directly of interest and what variables we would like to account for. Automatic approaches like model averaging certainly hold an appeal, but model-based automatic approaches can be dangerous in the context of explanatory inference.

A common motivation for model averaging is to incorporate model uncertainty into the analysis, inferences, and ultimately management decisions. This implies an increase in posterior variance of predictions on average (with a decrease in bias), and it has been stated that the same should hold for the posterior variance of model averaged partial regression coefficients (Leamer 1978, Raftery et al. 1997). Beyond the scope of this paper, we are investigating the extent to which this may be the case in practice and how this relates to the two components (the zero component and the continuous component) of the model averaged posterior distribution. If the results are essentially the same for model averaging as compared to using one model, model averaging may not be a mistake, but it may be adding unnecessary complexity to an analysis for a perceived goal that is left unrealized.

When model averaging is used for prediction, it is natural to inquire which predictors are most useful for predicting the mean response. Posterior inclusion probabilities (PIPs) have been used as such a summary

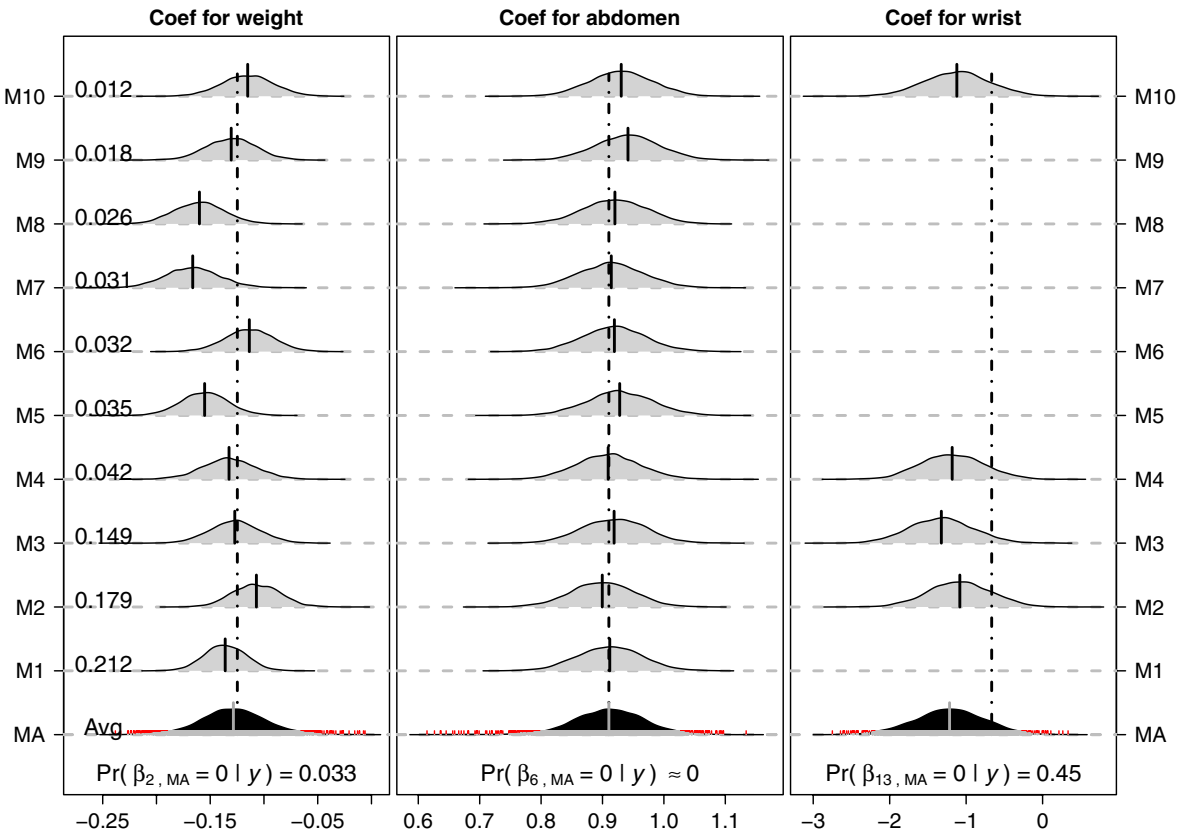

FIG. 6. This MAP plot includes variables weight, abdomen, and wrist from the body fat example. The models with the two highest posterior model probabilities are those including weight and abdomen, and weight, abdomen, and wrist, represented in rows M_1 and M_2, respectively. For the other rows, not all variables included in the model are shown in the plot. This plot shows the posterior distributions for the partial regression coefficients associated with these three predictors for the ten models with the highest posterior model probabilities (note that the model-averaged distribution is an approximation, computed using the top 500 models which amount to 99.8% of the posterior model mass). mass). See Fig. 4 for definition of abbreviations. Code to recreate the figure is available in Appendix S2, a script file, can also be found in Appendix Data S1 (Rcode-appendix-MAPP.R).

for individual predictors. However, a good model (as a collection of predictors) should be distinguished from a good predictor in isolation. We do not criticize the use of PIPs in general but suggest careful thought about when they are appropriate and how they are interpreted when they are used for stating broad conclusions. Although PIPs are a standard output from software packages designed for multimodel inference, we do not think they should be used simply because they are readily available. We deliberately use the posterior exclusion probabilities in the model averaged posteriors (MAP) plot because of their critical role in fully defining the posterior distribution of model averaged regression coefficients.

While the goal of accounting for as many sources of uncertainty as possible is certainly worthy, we believe the complexity of the method used to achieve this goal, and the potential loss in the interpretability of results in the context of the question of interest, should be weighed against sheer sophistication and popularity of the method. The utility of a modeling approach should be measured by how well it serves the intended purpose of the research, by whether it is accessible for others to critique, and by the degree to which it facilitates practical interpretations and meaningful conclusions. In *Example 1: Haul-out Behavior of Weddell Seals* and *Example 2: When Prediction Leads to Explanation*, we demonstrated the utility of the MAP plot for all three of the aforementioned criteria. The MAP plot can help researchers understand the differences between the results from model averaging and those from individual models, assess the appropriateness and usefulness of model averaging for a particular problem, and find a starting place for justifying their choice of method to peers. We presented the MAP plot and discussion within a Bayesian framework, but the information contained in the plot and the ideas motivating it can directly be translated to model averaging performed using AIC. The main difference is displaying confidence intervals instead of posterior distributions, and such a plot can be used to elicit the same type of discussion and thought about model averaging (there is a function in our R package to create the MAP plot for AIC results MApp IC(); see Appendix S1 for details).

Conclusion

Negative attention directed toward not accounting for model uncertainty has led to researchers feeling as though they must account for it in their analyses. In such cases, model averaging may be the low hanging fruit (e.g., Burnham and Anderson 2002, Montgomery and Nyhan 2010) or may be suggested by peer reviewers. We have experienced this ourselves as statistical consultants through reviewer requests to conduct model selection or model averaging among competing models when it was, at best, unnecessary, and at worst, inappropriate. Reviewer comments have reflected both a lack of understanding of interpretation of partial regression coefficients and a loyalty to more automatic model selection techniques as a way to justify a model. This experience is echoed by other statisticians and scientists with solid statistical foundations (Thomas et al. 2007, Brewer 2015, Cade 2015).

Maintaining a focus on question-focused modeling in cases where it is appropriate can be difficult to justify to peers who may be expecting some form of multimodel inference in an analysis. With our tools for assessing model averaging, we shed light on the complexity of the posterior distribution of model-averaged partial regression coefficients and address a common misconception that model averaging is an easier, nearly automatic, alternative to question-focused modeling or other model selection methods. Crainiceanu et al. (2008) and Wilson and Reich (2014) describe methods that may fall closer to middle ground between multimodel inference and question-focused modeling by proposing algorithms to help identify a subset of models addressing a particular question of interest and where the meanings of the partial regression coefficients of interest are similar enough across models to be reasonably averaged.

This paper is meant to provoke careful thought and continued discussions about model averaging, as well as provide common foundations to facilitate assessment and discussion of model averaging on a case-by-case basis. We agree with Ver Hoef and Boveng (2015) that iterating on one model can often be better than considering many models, and we stress that adding complexity to analyses to account for model uncertainty should be seriously weighed against the cost of interpretation and the simplifying assumptions made in the process.

Acknowledgments

We would like to thank the subject matter editor, Dr. Kiona Ogle, as well as Dr. Brian S. Cade, Dr. Mevin Hooten, and two anonymous reviewers, all of whom provided invaluable feedback to strengthen and focus the manuscript. Additionally, we would like to thank Dr. Brian S. Cade for sharing his work throughout the process of developing this manuscript. The Weddell seal field research was supported by the National Science Foundation OPP-0635739 grant to R. A. Garrott, J. J. Rotella, and D. B. Siniff. All of the camera images were obtained under authority of permit NMFS Permit No.1032-1917-02, and Dr. Robert Garrott put in a large effort obtaining the photos. Jay Rotella, Jesse DeVoe, and Michelle LaRue were largely involved in providing data sets and a wealth of information about their involvement in the Weddell Seal Project, and we thank them for their help and contributions.

Literature Cited

Albert, J. 2014. LearnBayes: functions for learning Bayesian inference. R package version 2.15. https://CRAN.R-project.org/package=LearnBayes

Barker, R. J., and W. A. Link. 2013. Bayesian multimodel inference by rjmcmc: a Gibbs sampling approach. American Statistician 67:150–156.

Barker, R. J., and W. A. Link. 2015. Truth, models, model sets, aic, and multimodel inference: a Bayesian perspective. Journal of Wildlife Management 79:730–738.

Brewer, M. 2015. Ten top tips for reviewing statistics: a guide for ecologists. https://methodsblog.wordpress.com/2015/06/03/reviewing_statistics/

Burnham, K. P. 2015. Multimodel inference: understanding AIC relative variable importance values. http://warnercnr.colostate.edu/kenb/pdfs/KenB/AICRelativeVariableImportanceWeights-Burnham.pdf

Burnham, K. P., and D. R. Anderson. 2002. Model selection and multimodel inference a practical information-theoretic approach. Springer, New York, NY.

Cade, B. S. 2015. Model averaging and muddled multimodel inference. Ecology 96:2370–2382.

Carlin, B. P., and S. Chib. 1995. Bayesian model choice via Markov chain Monte Carlo methods. Journal of the Royal Statistical Society. Series B (Methodological) 57:473–484.

Chipman, H. 1996. Bayesian variable selection with related predictors. Canadian Journal of Statistics 24:17–36.

Chipman, H., E. I. George, and R. E. McCulloch. 2001. The practical implementation of Bayesian model selection. IMS Lecture Notes Monograph Series 38:67–134.

Clyde, M. 2016. BAS: Bayesian adaptive sampling for Bayesian model averaging. R package version 1.4.1.

Clyde, M., H. Desimone, and G. Parmigiani. 1996. Prediction via orthoganilized model mixing. Journal of the American Statistical Association 91:1197–1208.

Clyde, M., and E. I. George. 2004. Model uncertainty. Statistical Science 19:81–94.

Clyde, M. A. 1999. Bayesian model averaging and model search strategies. Bayesian Statistics 6:157–185.

Clyde, M. A., J. Ghosh, and M. L. Littman. 2011. Bayesian adaptive sampling for variable selection and model averaging. Journal of Computational and Graphical Statistics 20:80–101.

Crainiceanu, C. M., F. Dominici, and G. Parmigiani. 2008. Adjustment uncertainty in effect estimation. Biometrika 95:635–651.

Doherty, P. F., G. C. White, and K. P. Burnham. 2012. Comparison of model building and selection strategies. Journal of Ornithology 125(Suppl 2):S317–S323.

Feldkircher, M., and S. Zeugner. 2009. Benchmark priors revisited: on adaptive shrinkage and the supermodel effect in Bayesian model averaging. International Monetary Fund (IMF) Working Paper, Pages 1–39. Finance Department.

Fieberg, J., and D. H. Johnson. 2015. Mmi: multimodel inference or models with management implications. Journal of Wildlife Management 79:708–718.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. 2013. Bayesian data analysis. Third edition. Chapman and Hall CRC Press, Boca Raton, FL.

George, E. I., and R. E. McCulloch. 1993. Variable selection via Gibbs sampling. Journal of the American Statistical Association 88:881–889.

George, E. I., and R. E. McCulloch. 1997. Approaches for Bayesian variable selection. Statatistica Sinica 7:339–373.
Geweke, J. 1996. Variable selection and model comparison in regression. Bayesian Statistics 5:609–620.
Green, P. J. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82:711–732.
Hocking, R. R. 2003. Methods and applications of linear models. John Wiley & Sons, Hoboken, NJ.
Hodges, J. S. 1987. Uncertainty, policy analysis and statistics. Statistical Science 2:259–291.
Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky. 1999. Bayesian model averaging: a tutorial. Statistical Science 14:382–401.
Hooten, M. B., and N. T. Hobbs. 2015. A guide to bayesian model selection for ecologists. Ecological Monographs 85:3–28.
Jewell, N. P. 2004. Texts in Statistical Science: Statistics for Epidemiology. Chapman & Hall CRC, Boca Raton, FL.
Johnson, R. W. 1996. Fitting percentage of body fat to simple body measurements. Journal of Statistics Education 4:1.
Kampstra, P. 2008. Beanplot: a boxplot alternative for visual comparison of distributions. Journal of Statistical Software, Code Snippets 28:1–9.
Kuo, L., and B. Mallick. 1998. Variable selection for regression models. Sankhyā: The Indian Journal of Statistics, Series B 60:65–81.
Leamer, E. E. 1978. Specification searches. Wiley, New York, New York, USA.
Link, W. A., and R. J. Barker. 2006. Model weights and the foundations of multimodel inference. Ecology, 87:2626–35.
Link, W. A., and R. J. Barker. 2010. Bayesian inference with ecological applications. Elsevier, London, NW1 7BY, UK.
Montgomery, J. M., and B. Nyhan. 2010. Bayesian model averaging: theoretical developments and practical applications. Political Analysis 18:245–270.
Penrose, K., A. Nelson, and A. Fisher. 1985. Generalized body composition prediction equation for men using simple measurement techniques (abstract). Medicine and Science in Sports Exercise 17:189.
R Core Team. 2016. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.r-project.org
Raftery, A. E., D. Madigan, and J. A. Hoeting. 1997. Bayesian model averaging for linear regression models. Journal of the American Statistical Association 92:179.
Ramsey, F. L., and D. W. Schafer. 2013. The statistical sleuth: a course in methods of data analysis. Third edition. Brooks/Cole, Cengage Learning, Boston, MA.
Sacher, G. A., and E. Staffeldt. 1974. Relation of gestation time to brain weight for placental mammals: implications for the theory of vertebrate growth. American Naturalist 108:593–613.
Thomas, D. C., M. Jerrett, N. Kuenzli, T. A. Louis, F. Dominici, S. Zeger, J. Schwarz, R. T. Burnett, D. Krewski, and D. Bates. 2007. Bayesian model averaging in time-series studies of air pollution and mortality. Journal of Toxicology and Environmental Health 70:311–315.
Ver Hoef, J. M., and P. L. Boveng. 2015. Iterating on a single model is a viable alternative to multimodel inference. Journal of Wildlife Management 79:719–729.
Wickham, H. 2009. ggplot2: elegant graphics for data analysis. Springer, New York, New York, USA.
Wickham, H., and R. Francois. 2015. dplyr: a grammar of data manipulation. R package version 0.4.3. https://CRAN.R-project.org/package=dplyr
Wilson, A., and B. J. Reich. 2014. Confounder selection via penalized credible regions. Biometrics 70:852–861.
Zellner, A. 1984. Posterior odds ratios for regression hypotheses: general considerations and some specific results. University of Chicago Press, Chicago, IL.
Zeugner, S. 2011. Replicating the body fat example from "Bayesian model averaging: a tutorial" (1999) with bms in R. https://modelaveraging.wordpress.com/2011/02/07/replicating-the-body-fat-example-from-bayesian-model-averaging-a-tutorial-1999-with-bms-in-r/

Supporting Information

Additional Supporting Information may be found online at: http://onlinelibrary.wiley.com/doi/10.1002/eap.1419/full