Improving Predictions Using Ensemble Bayesian Model Averaging *

Jacob M. Montgomery
Department of Political Science
Washington University in St. Louis
Campus Box 1063, One Brookings Drive
St. Louis, MO, USA, 63130-4899

Florian Hollenbach
Department of Political Science
Duke University
Perkins Hall 326 Box 90204
Durham, NC, USA, 27707-4330

Michael D. Ward
Department of Political Science
Duke University
Perkins Hall 326 Box 90204
Durham, NC, USA, 27707-4330
corresponding author: michael.d.ward@duke.edu

January 18, 2012

*For generously sharing their data and models with us, we thank Alan Abramowitz, James Campbell, Robert Erikson, Ray Fair, Douglas Hibbs, Michael Lewis-Beck, Andrew D. Martin, Kevin Quinn, Stephen Shellman, Charles Tien, & Christopher Wlezien. We especially want to thank Adrian Raftery and Brendan Nyhan for their encouragement and feedback as this project evolved. The editor and the reviewers of *Political Analysis* provided especially salient and important suggestions that substantially improved our research. This work was supported by the

Information Processing Technology Office of the Defense Advanced Research Projects Agency through a holding grant is to the Lockheed Martin Corporation [FA8650-07-C-7749].

ABSTRACT

We present ensemble Bayesian model averaging (EBMA) and illustrate its ability to aid scholars in the social sciences to make more accurate forecasts of future events. In essence, EBMA improves prediction by pooling information from multiple forecast models to generate ensemble predictions similar to a weighted average of component forecasts. The weight assigned to each forecast is calibrated via its performance in some validation period. The aim is not to choose some "best" model, but rather to incorporate the insights and knowledge implicit in various forecasting efforts via statistical postprocessing. After presenting the method, we show that EBMA increases the accuracy of out-of-sample forecasts relative to component models in three applied examples: predicting the occurrence of insurgencies around the Pacific Rim, forecasting vote shares in U.S. presidential elections, and predicting the votes of U.S. Supreme Court Justices.

1. INTRODUCTION

Testing systematic predictions about future events against observed outcomes is generally seen as the most stringent validity check of statistical and theoretical models. Yet, political scientists rarely make predictions about the future. Empirical models are seldom applied to out-of-sample data and are even more rarely used to make predictions about future outcomes. Instead, researchers typically focus on developing and validating theories that explain past events.

In part, this lack of emphasis on forecasting results from the fact that it is so difficult to make accurate predictions about complex social phenomena. However, research in political science could gain immensely in its policy relevance if predictions were more common and more accurate. Improved forecasting of important political events would make research more germane to policymakers and the general public who may be less interested in explaining the past than anticipating and altering the future. From a scientific standpoint, greater attention to forecasting would facilitate stringent validation of theoretical and statistical models since truly causal models should perform better in out-of-sample forecasting.

In this article, we extend a promising statistical method – ensemble Bayesian model averaging (EBMA) – and introduce software that will aid researchers across disciplines to make more accurate forecasts. In essence, EBMA makes more accurate predictions possible by pooling information from multiple forecast models to generate ensemble predictions similar to a weighted average of component forecasts. The weight assigned to each forecast is calibrated via its performance in

some prior period. These component models can be diverse. They need not share covariates, functional forms, or error structures. Indeed, the components may not even be statistical models, but may be predictions generated by agent-based models or subject-matter experts.

In the rest of this article we briefly review existing political science research aimed at forecasting and then present the mathematical details of the EBMA method. We then illustrate the benefits of EBMA by applying it to predict insurgency events on the Pacific Rim, U.S. presidential elections, and voting on the U.S. Supreme Court.

2. DYNAMIC FORECASTING IN POLITICAL SCIENCE

Although forecasting is a rare exercise in political science, there are an increasing number of exceptions. In most cases, "forecasts" are conceptualized as an exercise in which the predicted values of a dependent variable are calculated based on a specific statistical model and then compared with observed values (e.g., Hildebrand, Laing and Rosenthal 1976). In many instances, this reduces to an analysis of residuals. In others, the focus is on randomly selecting subsets of the data to be excluded during model development for cross-validation. However, there is also a more limited tradition of making true forecasts about events that have not yet occurred. (See Brandt, Freeman and Schrodt (2011*a*) for a recent and thorough survey of forecasts in political science and economics with a focus on strategies to perform more meticulous comparisons of their accuracy.)

An early proponent of using statistical models to make predictions in the realm of international relations (IR) was Stephen Andriole (Andriole and Young 1977). In 1978, a volume edited by Nazli Choucri and Thomas Robinson provided an overview of the then current work in forecasting in IR. Much of this work was done in the context of policy-oriented research for the U.S. government during the Vietnam War. Subsequently, there were a variety of efforts to create or evaluate forecasts of international conflict, including Freeman and Job (1979), Singer and Wallace (1979), and Vincent (1980). In addition, a few efforts began to generate forecasts of domestic conflict (e.g., Gurr and Lichbach 1986). Recent years, however, have witnessed increasing interest in prediction across a wide array of contexts in IR.[1] The 2011 special issue of *Conflict Management and Peace Science* on prediction in the field of IR exemplifies this growing emphasis on forecasting (c.f., Schneider, Gleditsch and Carey 2011; Bueno de Mesquita 2011; Brandt, Freeman and Schrodt 2011*b*). Ward, Greenhill and Bakke (2010) and Greenhill, Ward and Sacks (2011) provide additional discussion of forecasting in IR.

Outside of IR, forecasting in political science has largely taken place in the context of election research. In the 1960s, de Sola Pool, Abelson and Popkin (1964) published a volume describing their work on the 1960 and 1964 presidential elections. They reported their efforts to use a computer simulation to predict election outcomes, which was initially undertaken in the context of providing campaign management advice for the 1960 campaign of John F. Kennedy. Rosenstone (1983) published perhaps the most influential early work on elections forecasting, which surveyed the then state-of-the-art and included examples going

back to 1932.

In the 1990s, political scientists renewed their interest in predicting presidential elections (Campbell and Wink 1990; Campbell 1992). This work was anticipated by the efforts of several economists, most notably the forecasts established by Ray C. Fair (1978). As we discuss below, predicting U.S. presidential and congressional elections has since developed into a regular exercise. Moreover, researchers have begun to forecast election outcomes in France (e.g., Jerome, Jerome and Lewis-Beck 1999) and the United Kingdom (e.g., Whiteley 2005).[2]

While efforts to predict future outcomes remain uncommon, research that combines multiple forecasts are nearly non-existent. To our knowledge, the only non-IR examples are the PollyVote project (c.f. Graefe et al. 2010), which combines multiple predictions using simple averages of forecasts to predict U.S. presidential elections, and Lock and Gelman (2010), who use Bayesian methodology to combine information from historical state-level election returns, current polling data, and forecasting models to generate election forecasts.

Yet, combining forecasts, and ensemble methods in particular, have been shown to substantially reduce prediction error in two important ways. First, across subject domains, ensemble predictions are usually more accurate than any individual component model. Second, they are significantly less likely to make dramatically incorrect predictions (Bates and Granger 1969; Armstrong 2001; Raftery et al. 2005).[3] Combining forecasts not only reduces reliance on single data sources and methodologies (which lowers the likelihood of dramatic errors), but also allows for the incorporation of more information than any one model is likely to include

in isolation.

The idea of ensemble learning itself has a long history in the machine learning community. The most thorough treatment is found in Hastie, Tibshirani, and Friedman (2009). A wide range of statistical approaches including bagging, random forests, as well as boosting and penalized methods may be properly considered ensemble approaches. They are different from EBMA however, which comes from another branch on the ensemble family tree – Bayesian statistics. Bayesian methods themselves can generally be viewed as ensemble methods, since they produce a large number of candidate "models" that are averaged to create a posterior distribution of parameters (Hastie, Tibshirani and Friedman 2009, p. 605).

These advances in the statistical literature parallel additional research in formal theory, which shows that groups of agents using diverse decision-rules or composed of agents with different viewpoints on a problem can produce superior outcomes in difficult decision environments (Page, Sander and Schneider-Mizell 2007; Page 2008, 2011). That is, social systems, organizations, and institutions that are better able to combine insights and knowledge from diverse actors are more functional, successful, and adaptive in complex environments.

This last strain of thought is related to research that suggests the use of prediction markets as a method of aggregating a large number of individual predictions about particular events. For example, Berg, Nelson and Rietz (2008) discuss prediction markets and demonstrate that they can be more accurate than polls when forecasting elections. One important prediction market in political science is the Iowa Electronic Market, in which individuals buy futures on politicians which are

paid after election results are revealed.

3. ENSEMBLE BAYESIAN MODEL AVERAGING

Predictive models remain underutilized, yet an increasing number of scholars have developed forecasting models for specific research domains. As the number of forecasting efforts proliferate, however, there is a growing benefit from developing methods to pool across models and methodologies to generate more accurate forecasts. Very often, specific predictive models prove to be correct only for certain subsets of observations. Moreover, specific models tend to be more sensitive to unusual events or particular data issues than ensemble methods.

To aid the newfound emphasis on prediction in political science, we are advancing recent statistical research aimed at integrating multiple predictions into a single improved forecast. In particular, we are adapting an ensemble method first developed for application to the most mature prediction models in existence – weather forecasting models. To generate predictive distributions of outcomes (e.g., temperature), weather researchers apply ensemble methods to forecasts generated from multiple models (Raftery et al. 2005). Thus, state-of-the-art ensemble forecasts aggregate multiple runs of (often multiple) weather prediction models into a single unified forecast.

The particular ensemble method we are extending for application to political outcomes is ensemble Bayesian model averaging (EBMA). First proposed by Raftery et al. (2005), EBMA pools across various forecasts while meaningfully

incorporating *a priori* uncertainty about the "best" model. It assumes that no particular model or forecasting method can fully encapsulate the true data-generating process. Rather, various research teams or statistical techniques will reflect different facets of reality. EBMA collects *all* of the insights from multiple forecasting efforts in a coherent manner. The aim is not to choose some "best" model, but rather to incorporate the insights and knowledge implicit in various forecasting efforts via statistical post-processing. In recent years, variants of the EBMA method have been applied to subjects as diverse as inflation (Wright 2009; Koop and Korobilis 2009; Gneiting and Thorarinsdottir 2010), stock prices (Billio et al. 2011), economic growth and policymaking (Brock, Durlauf and West 2007; Billio et al. 2010), exchange rates (Wright 2008), industrial production (Feldkircher 2011), ice formation (Berrocal et al. 2010), visibility (Chmielecki and Raftery 2010), water catchment streamflow (Huisman et al. 2009), climatology (Min and Hense 2006; Min, Simonis and Hense 2007; Smith et al. 2009), and hydrology (Zhang, Srinivasan and Bosch 2009). Indeed, research is underway to extend the method to handle missing data (Fraley, Raftery and Gneiting 2010; McCandless, Haupt and Young 2011) as well as calibrate model weights on non-likelihood criteria (e.g., Vrugt et al. 2006).

3.1. *Overview of method*

EBMA is designed for application in the context of a subject-domain with ongoing forecasting efforts. That is, it assumes the existence of multiple teams or individ-

uals making regular predictions about a common set of outcomes. For example, there may be multiple analysts or teams making predictions about the likelihood of violent conflict in specific regions of the world, quarterly economic growth for the United States, or the votes of members on bills before Congress. As we show in our examples below, these predictions might originate from the insights and intuitions of individual subject-experts, traditional statistical models, non-linear classification trees, neural networks, agent based models, or anything in between.

EBMA is a method for taking the predictions made by multiple teams and combining them – based on their past performance and uniqueness – to create a new ensemble forecasting model. This ensemble model can then make predictions about unobserved outcomes in the future and usually outperforms its components. Roughly speaking, it creates forecasts by creating weighted averages of component predictions, or component predictive probability distribution functions (PDFs). The weight assigned to each component forecast, denoted w_k below, reflects two aspects of the components' past forecasts. First, *ceteris paribus*, the EBMA model will give greater weight to forecasts that were more accurate in the past. Second, *ceteris paribus*, it will assign a greater weight to models that made unique (but correct) predictions. That is, component forecasts that are too highly correlated may jointly have a large weight, but will individually be penalized.

There are two important aspects of EBMA that distinguish it from the alternative model-selection or averaging methods referenced above. First, EBMA is more flexible in not requiring any information about the actual covariates that go into the component models. A second, and related point is that EBMA does

not require researchers to develop metrics to penalize component forecasts for the number of parameters included, the number of covariates, or their complexity more generally. In the case of subject-expert opinions, there may not even be any covariates or statistical models involved, a point we return to in the Supreme Court example below. Another non-statistical component model that researchers might include is prices on prediction markets. In other instances, predictions may come from models whose "complexity" is not easily defined or enumerated (e.g., agent-based models). Of course, overly-complex "garbage can" models will generally perform poorly when making predictions over any outcome and therefore receive a lower weight in the ensemble forecast. Yet, this lower weight is a function of the model's predictive performance rather than a pre-specified preference for parsimony. The upshot is that EBMA forecasts will implicitly penalize over-fitting of component models since the weight assigned to component models is based on predictive performance. However, it can do so without explicitly penalizing components for complexity.[4]

3.2. *Mathematical foundations*

EBMA itself is an extension of the Bayesian model averaging (BMA) methodology (c.f., Madigan and Raftery 1994; Draper 1995; Raftery 1995; Hoeting et al. 1999; Clyde 2003; Raftery and Zheng 2003; Clyde and George 2004) that has received considerable attention in the field of statistics. BMA was first introduced to political science by Bartels (1997) and has been applied in a number of con-

texts (e.g., Bartels and Zaller 2001; Gill 2004; Imai and King 2004; Geer and Lau 2006). Montgomery and Nyhan (2010) provide a more in-depth discussion of BMA and its applications in political science.

Assume we have some quantity of interest to forecast, $\mathbf{y}^{t^*}$, in some future period $t^* \in T^*$. Further assume that we have extant forecasts for events $\mathbf{y}^t$ for some past period $t \in T$ that were generated from K forecasting models or teams, $M_1, M_2, \ldots, M_K$. Each model, M_k, is assumed to come from the prior probability distribution $M_k \sim \pi(M_k)$, and the PDF for $\mathbf{y}^t$ is $p(\mathbf{y}^t|M_k)$. The outcome of interest is distributed $p(\mathbf{y}^{t^*}|M_k)$. Applying Bayes Rule, we get that

$$p(M_k|\mathbf{y}^t) = \frac{p(\mathbf{y}^t|M_k)\pi(M_k)}{\sum_{k=1}^{K} p(\mathbf{y}^t|M_k)\pi(M_k)} \tag{1}$$

and the marginal predictive PDF is

$$p(\mathbf{y}^{t^*}) = \sum_{k=1}^{K} p(\mathbf{y}^{t^*}|M_k)p(M_k|\mathbf{y}^t). \tag{2}$$

The BMA PDF (2) can be viewed as the weighted average of the component PDFs where the weights are determined by each model's performance within the already-observed period T. Likewise, we can simply make a deterministic estimate using the weighted predictions of the components, denoted $E(\mathbf{y}^{t^*}) = \sum_{k=1}^{K} E(\mathbf{y}^{t^*}|M_k)p(M_k|\mathbf{y}^t)$.

EBMA for dynamic settings: In generating predictions of future events, the task is to first build a set of statistical models for some set of observations S in the past time periods T', which we refer to as the training period. Using these models, we then generate predictions $\mathbf{f}_k^{s|t}$ for some period T, which has already occurred but

which was not included in the training sample. We refer to this as the validation period, and we will use this data to calibrate the EBMA model.[5] Finally, using the same K models, we assume that there are true forecasts ($\mathbf{f}_k^{s|t^*}$) for observations $s \in S$ in future time periods $t^* \in T^*$.[6] We either (a) treat these raw predictions as a component model in the steps below or (b) statistically post-process the predictions for out-of-sample bias reduction and treat these re-calibrated predictions as a component model.

As a running example, let us assume that we have K forecasting efforts for modeling insurgencies in a set of countries S ongoing throughout the training (T') validation (T) and test (T^*) periods. We will associate each component forecast with a component PDF, $g_k(\mathbf{y}|\mathbf{f}_k^{s|t,t^*})$, which may be the original prediction from the forecast model or the bias-corrected forecast.

The EBMA PDF is then a finite mixture of the K component PDFs, denoted

$$p(\mathbf{y}|\mathbf{f}_1^{s|t}, \ldots, \mathbf{f}_K^{s|t}) = \sum_{k=1}^{K} w_k g_k(\mathbf{y}|\mathbf{f}_k^{s|t}), \tag{3}$$

The w_k's $\in [0, 1]$ are model probabilities and $\sum_{k=1}^{K} w_k = 1$. Roughly speaking, they are associated with each component model's predictive performance in the validation period controlling for the degree to which they offer unique insight (i.e., a model's predictions are distinct from those of other component models). We provide additional discussion about the model weights in the election forecasting example below. Details for parameter estimation are provided in Appendix A. The ensemble PDF for an insurgency in the test period t^* in country s is then:

$$p(y|f_1^{s|t^*}, \ldots, f_K^{s|t^*}) = \sum_{k=1}^{K} w_k g_k(y|f_k^{s|t^*}). \tag{4}$$

EBMA for normally distributed outcomes: To gain a fuller understanding of the EBMA method, it is easiest to imagine an effort to predict some normally distributed outcome. When forecasting outcomes that are distributed normally, Raftery et al. (2005) propose approximating the conditional PDF as a normal distribution centered at a linear transformation of the individual forecast, $g_k(\mathbf{y}|\mathbf{f}_k^{s|t}) = N(a_{k0} + a_{k1}\mathbf{f}_k^{s|t}, \sigma^2)$. Prior applications have found that this adjustment of the component models' forecasts reduces over-fitting and improves the performance of the final ensemble forecasting model (Raftery et al. 2005).[7] Using (3) and (4) above, the EBMA PDF is then

$$p(\mathbf{y}|\mathbf{f}_1^{s|t}, \ldots, \mathbf{f}_K^{s|t}) = \sum_{k=1}^{K} w_k N(a_{k0} + a_{k1}\mathbf{f}_k^{s|t}, \sigma^2), \tag{5}$$

and the predictive distribution for some observation y is

$$p(y|f_1^{s|t^*}, \ldots, f_K^{s|t^*}) = \sum_{k=1}^{K} w_k N(a_{k0} + a_{k1}f_k^{s|t^*}, \sigma^2) \tag{6}$$

Thus, the predictive PDF is a mixture of K normal distributions each of whose mean is determined by the component prediction ($f_k^{s|t^*}$) and whose "height" (i.e, the total area under the curve for component k) is determined by the model weight w_k.

4. EMPIRICAL APPLICATIONS

In this section, we provide empirical applications of EBMA to predict insurgency in the Pacific Rim, presidential election outcomes in the United States, and votes

of Justices of the United States Supreme Court. These three examples demonstrate the usefulness of the method for diverse domains of political science research, different types of outcomes of interest (i.e., dichotomous and continuous), and different forms of component models (i.e., statistical models versus expert predictions).

4.1. *Application to insurgency forecasting*

Our first example applies the EBMA method to data collected for the Integrated Crisis Early Warning Systems (ICEWS) project sponsored by the Defense Advanced Research Projects Agency (DARPA). The task of the ICEWS project is to train models on data (focusing on five outcomes of interest) for 29 countries for every month from 1997 through the present and to then make accurate predictions about expected crisis events.[8] For purposes of demonstration, we focus on only one of these outcomes – violent insurgency.

The bulk of the data for the ICEWS project is gleaned from natural language processing of a continuously updated harvest of news stories (primarily taken from Lexus/Nexus and Factiva archives). These are digested with a version of the TABARI processor for events developed by Philip Schrodt and colleagues in the context of the Event Data Project (see `http://eventdata.psu.edu/` for more details). These data are augmented with a variety of covariates including: country-level attributes (coded on a monthly or yearly basis) from the Polity and World Bank datasets, information about election cycles (if any), events in

neighboring countries, and the length of shared borders with neighboring countries.

Component models: We apply EBMA to make predictions for the occurrence of insurgency in these 29 countries for each month in the year 2010 (the last year in the dataset). As a first step in the process, we must choose some set of observations that we wish to treat as a training period for the component statistical models and a second set of data to treat as a validation set with which to calibrate the EBMA model.

Unfortunately, there is no clear guidance available on how to choose the relative sizes of the training set, the validation set, or the test set. Hastie, Tibshirani, and Friedman (2009, Chpt. 7) discuss this in some detail, wherein they note (p. 195): "It is difficult to give a general rule on how to choose the number of observations in each of the ... parts, as this depends on the signal-to-noise ratio in the data and the training sample size." The basic point is that there is no general rule. Moreover, the appropriate size of training and test set depends on the prevalence of the signal in the training data. In the case of predictive studies like ours, the only general rule is: it depends.

In this case, we estimated three exemplar statistical models using data for the *training period* ranging from January 1999[9] to December 2007, and fit an EBMA model using the component model predictions for the *validation period* ranging from January 2008 to December 2009. We then make forecasts for the *test period* ranging from January 2010 to December 2010 using both the component

and EBMA models. To provide variation in the complexity (as well as accuracy) of the components, we included the following models.

- **SAE**: This is one model developed as part of the ICEWS project and was designed by Strategic Analysis Enterprises. It is specified as a simple generalized linear logistic model including 27 different independent variables.[10] All of the variables are taken from the ICEWS event-stream data.

- **GLM**: For the purposes of demonstrating the properties of the EBMA method, we estimated a crude logistic model that includes only *population size*, *GDP growth* (both lagged three months), the number of *minority groups at risk* in the country, and a measure of *anocracy* supplied in the *Polity IV* dataset (Marshall, Jaggers and Gurr 2009).

- **LMER**: This is a generalized linear mixed effects model using a logistic link function and including a random effects term for lagged *GDP per capita* and the lagged number of *conflictual events involving the United states* in the country of interest.[11] The list of additional covariates includes: the number of *conflictual events involving the military* within the country of interest (lagged three months), the number of *days elapsed since the last election*,[12] the number of *new insurgencies* that began in the previous two years, and a *spatial lag* that reflects recent occurrences of domestic crises in the countries' geographic neighbors.[13]

Results: Table 1 shows the EBMA model parameters as well as fit statistics associated with the individual component models and the EBMA predictions for

the validation time period (2008-2009). The first column shows the weights that the EBMA model assigned to each component. As can be seen, the GLM model is effectively excluded, while the LMER model carries the greatest weight ($w_k = 0.85$) followed by the SAE model ($w_k = 0.14$). The constant term associated with each component corresponds to the term a_{k0} in (8), while the predictor corresponds to a_{k1}. The other columns in Table 1 are fit statistics. AUC is the area under the Receiver-Operating Characteristic (ROC) curve. The advantage of using ROC curves is that it evaluates forecasts in a way that is less dependent on an arbitrary cutoff point. A value of 1 would mean that all observations were predicted correctly at all possible cutoff points (King and Zeng 2001).

Table 1 about here

We compare the models using three additional metrics. The proportional reduction in error (PRE) is the percentage increase of correctly predicted observations relative to some pre-defined base model. In this case, the base model is predicting "no insurgencies" for all observations. Insurgencies are relatively rare events. Thus, predicting a zero for all observations leads to a 91.8% correct prediction rate. The Brier score is the average squared deviation of the predicted probability from the true event (0 or 1). Thus, a lower score corresponds to higher forecast accuracy (Brier 1950). Finally, we calculate the percentage of observations that each model would predict correctly using a 0.5 threshold on the predicted probability scale.

Figure 1 about here

Note that the EBMA model does at least as well (and usually better) than all of the component models on almost all of our model fit statistics. The EBMA model has the highest PRE and % correct and the lowest Brier score. Although

the LMER model has a slightly higher AUC, that model's overall performance suggests that it may be over-fit.

Figure 1 shows separation plots for the EBMA model and the individual components (Greenhill, Ward and Sacks 2011). In each plot, the observations are ordered from left to right by increasing predicted probabilities of insurgency (as predicted by the particular model). The black line corresponds to the predicted probability produced by the relevant model for each observation and actual occurrences of insurgencies are colored red. Figure 1 shows visually that the GLM model performs very poorly, whereas the LMER model performs very well, but tends to assign high probabilities to a large number of observations where we observe no insurgencies (i.e., it over-predicts insurgencies). More importantly, the overall best performance is associated with the EBMA forecast. The separation plots show that the EBMA model produces few false negatives and significantly fewer false positives than any of the component models.

However, the more interesting evaluation of the EBMA method is its test-period predictive power. Table 2 shows fit statistics for the individual components as well as the EBMA forecasts for observations in the 12 months following the validation period. The EBMA model outperforms the component models on all metrics. In particular, the EBMA model has the highest PRE at 0.43. Since it is possible to predict 89.9% of these observations correctly by forecasting "no insurgency," a 43% reduction of error relative to the baseline model is quite substantial. **Table 2 about here**

Importantly, EBMA clearly outperforms all component models in regard to the Brier score. Research regarding scoring rules for forecasts has shown that

the Brier score is one of the best statistically proper scoring rules for evaluating predictions of binary dependent variables (Gneiting and Raftery 2007). Thus, to generally compare and rank the different models one should use the Brier score (Gneiting and Raftery 2007). As can be seen in Tables 1 and 2, the EBMA model has the lowest Brier score in both the validation and test-period forecasts.[14]

Figure 2 shows the separation plots for the components as well as the EBMA forecasts for 2010. The EBMA model again performs better than any of the individual components with very high predicted probabilities for the majority of actual events without over-predicting too many events. Taking both the fit statistics and the visual evidence together, we can conclude that the EBMA model leads to a substantial improvement in test-period forecasts relative to its components. This is true even in datasets with rare events and even when the individual components are already performing well.

Figure 2 about here

4.2. *Application to US presidential election forecasts*

For the past several U.S. election cycles, a number of research teams have developed forecasting models and published their predictions in advance of Election Day. For example, before the 2008 election, a symposium of forecasts was published in *PS: Political Science and Politics* with forecasts of presidential and congressional vote shares developed by Campbell (2008), Norpoth (2008), Lewis-Beck and Tien (2008), Abramowitz (2008), Erikson and Wlezien (2008), Holbrook (2008), Lockerbie (2008) and Cuzàn and Bundrick (2008). Responses to

the forecasts were published in a subsequent issue. Earlier, in 1999, an entire issue of the *International Journal of Forecasting* was dedicated to the task of predicting presidential elections (Brown and Chappell 1999). Predicting presidential elections has also drawn the attention of economists seeking to understand the relationship between economic fundamentals and political outcomes. Two prominent examples include work by Ray Fair (2010) and Douglas Hibbs (2000).

Component models: In the rest of this subsection, we replicate several of these models and demonstrate the usefulness of the EBMA methodology for improving the prediction of single important events.[15] We include forecasting models from six of the most widely cited presidential forecasting teams. Note that we do not, in every case, replicate the particular model which the authors identify as their definitive forecast.

- **Campbell**: Campbell's "Trial-Heat and Economy Model" (Campbell 2008)
- **Abramowitz**: The "Time-for-Change Model" created by Abramowitz (2008)
- **Hibbs**: Hibbs' "Bread and Peace Model" (Hibbs 2000)
- **Fair**: Fair's presidential vote-share model[16]
- **Lewis-Beck/Tien**: Lewis-Beck and Tien's "Jobs Model Forecast" (Lewis-Beck and Tien 2008)
- **EWT2C2**: Column 2 from Table 2 in Erikson and Wlezien (2008)

With the exception of the Hibbs forecast, the models are simple linear regressions. The dependent variable is the share of the two-party vote received by the incumbent-party candidate.[17]

Results: Rather than selecting a single partition of the data into training, validation, and test periods (as in the insurgency analysis) we generate sequential predictions. For each year from 1976 to 2008, we use all available prior data to fit the component models.[18] We then fit the EBMA model using the components' performances for election years beginning with 1952 (the year when all models begin generating predictions). For example, to generate predictions for the 1988 election, we used the performance of each component for the 1952-1984 period to estimate model weights.[19]

Table 3 about here

Table 3 provides exemplar results for the 2004 and 2008 elections. Table 3 shows the weights assigned to each model as well as the validation-period root mean squared error (RMSE) and mean absolute error (MAE) for the components and the EBMA forecasts.[20] Table 3 also shows the prediction errors, calculated as the difference between predicted and actual values in each year for each component model and the EBMA forecast.

The example results in Table 3 illustrate three important points. First, the EBMA model does better than any individual component on validation-sample measures of model fit such as RMSE and MAE (see Footnote 20). Second, these results demonstrate that EBMA is not guaranteed to generate the most accurate prediction for any single observation. In each year, at least one component model

comes closer to predicting the actual outcome. However, the EBMA forecasts will very rarely provide egregiously wrong predictions (e.g., as found in the Campbell model in 2008 and the Fair model in 2004) since it borrows strength from multiple components. Moreover, as we show below, in the aggregate the EBMA model tends to provide the best forecast over time.

Third, Table 3 shows that there is no clean a relationship between validation-sample model performance and model weights. For instance, the weight for the Abramowitz model in 2008 is 0.001 even though it has the lowest RMSE and MAE of any component. The diminished relationship between validation-sample performance and weight is a result of high correlations between forecasts.[21] For instance, fitted values for the Abramowitz model are correlated at 0.94 with the Campbell model and at 0.96 with the Lewis-Beck/Tien model. Thus, conditioned on knowing these forecasts, the Abramowitz component provides limited additional information.

In general, as the number of models included as EBMA components increases, the risk of including highly correlated forecasts will also rise. Researchers should be aware of the fact that adding additional forecasts as components will not necessarily improve the performance of EBMA. EBMA performance will instead be improved by the inclusion of increasing numbers of *diverse* and *accurate* forecasts (see also, Graefe et al. 2010). Including a large number of extremely correlated forecasts may actually reduce the benefits of ensemble forecasting. However, we note that in practice this is unlikely to be a significant concern as there are few domains in political science for which there are large numbers of ongoing fore-

casting efforts.

With the 2004 and 2008 examples in mind, we now turn to the relative test-sample performance of the EBMA and component forecasts across the entire 1976-2008 period. Table 4 shows the test-sample RMSE and MAE statistics as well as the percentage of observations that fall within the 67% and 90% predictive intervals for each model. For our purposes here, the main result in Table 4 is that the EBMA model again outperforms all components. The first two columns show this to be true in terms of predicted error (RMSE and MAE).[22]

Table 4 about here

In addition, the coverage statistics demonstrate better calibration of EBMA forecasts relative to its component models. For instance, the observed outcome falls within the 67% predictive interval for the Abramowitz model only three out of nine times, while it covers the observed values eight out of nine times for the Lewis-Beck/Tien model. Meanwhile, the EBMA 90% and 67% predictive intervals are nearly perfectly calibrated.

Figure 3 about here

In a well-calibrated forecasting model, out-of-sample outcomes should fall within predictive intervals at a rate corresponding to their size. For instance, the goal is for two-thirds of all out-of-sample observations to fall within their respective 67% predictive intervals. Poorly calibrated models will tend to produce predictive intervals that are either too narrow, generating inaccurate predictions, or too large, generating predictions that are accurate but too vague to be useful. The better calibration of the EBMA model can be seen visually in Figure 3. The plot shows the point predictions and the 67% and 90% predictive intervals for each model in each year. The vertical dashed lines show the actual observed out-

comes. Note that two of the most accurate forecasts, the Lewis-Beck/Tien and Erikson/Wlezien models, make very imprecise predictions. Thus, although they have very good coverage, it is at least partly because their estimates are so inexact. The Campbell, Abramowitz, and Hibbs models provide more reasonable predictive intervals, but are less accurate than EBMA. Meanwhile, the Fair model falls somewhere in between these two groupings.

Finally, it is worth noting an example – very noticeable in this data – of the kinds of problems that may arise when relying on a single model for making predictions. From 1952 to 2004, the Campbell model was consistently one of the strongest performers. Indeed, it made the most accurate forecast of the 2004 election. However, one of the crucial variables in this model comes from polling data measured in early September. As a result of the particularly late timing of the Republican Convention in 2008, it was the only model to forecast a victory for John McCain. By relying on a wider array of data sources and methodologies, EBMA reduces the likelihood of such large misses without completely eliminating the general insights captured by individual models that may on occasion be wide of the mark.

4.3. *Application to the Supreme Court Forecasting Project*

Our final application of EBMA is a re-analysis of data from the Supreme Court Forecasting Project (Ruger et al. 2004; Martin et al. 2004).[23] This example is especially interesting and important as it shows a particular strength of EBMA that

was not utilized in the previous two examples. That is, not only is EBMA able to combine the forecasts from multiple statistical models, in addition, it can also combine statistical predictions with forecasts generated by classification trees, subject experts, and other sources. As is shown below, the EBMA model is able to combine the strength of a statistical forecasting model with the particular strength of subject-expert predictions and improves on the accuracy of both. Furthermore the Supreme Court Forecasting Project offers a clean example for our purpose. The weights for the EBMA model are calibrated on the performance of the components on actual predictions of Supreme Court Justice votes. That is, even for the validation period, the predictions were made before the court decisions were issued. Thus, we can use the performance of the component models on actual predictions to calculate the weights for the EBMA model. We then compare the EBMA forecasts with the component model predictions on a separate test-sample.

A large literature in political science is concerned with constitutional courts and the Supreme Court in particular, with one of the most prominent strands of the literature on courts trying to analyze and explain justices' voting behavior. In general, however, the theories and models attempt to explain behavior *ex post* (e.g., Hausegger and Baum 1999; Segal and Cover 1989; Richards and Kritzer 2002; Klein and Hume 2003; Songer, Segal and Cameron 1994).

In contrast to most of this literature, a research team consisting of Andrew Martin, Kevin Quinn, Theodore Ruger, and Pauline Kim (henceforward MQRK) set out to develop methods to predict Supreme Court decisions in the future. Throughout 2002-2003, MQRK generated two sets of forecasts for every pending

case. First, for each case MQRK collected data on six different case characteristics, such as the "circuit of origin of the case" or "ideological direction of the lower courts ruling," which were then used as explanatory variables (Martin et al. 2004, 762). The authors then used classification trees to generate a binary forecast for the expected vote of each justice on each case (voting to affirm the lower court opinion is coded as a 1).

As a second method to forecast Supreme Court decisions, MQRK recruited a team of 83 legal experts. These experts would then predict the decision for each justice and the court as a whole for particular cases in their specialty area. The list of experts included academics, appellate attorneys, former Supreme Court clerks, and law school deans. MQRK attempted to recruit three expert forecasts for each case, although this was not possible for all cases.

The classification trees made predictions for all 67 cases included in the MQRK analysis. We include these binary model predictions as one component forecast. However, the individual legal experts made predictions on only a handful of cases. Owing to the paucity of the data for each expert, we pooled them together and treat all of the expert opinions as part of a single forecasting effort. We coded the expert forecast to be the mean expert prediction. This implies that the expert forecast predicts a vote to affirm if a majority of experts polled for that case predict an affirming vote. We fit an EBMA model using all cases with docket numbers dating from 2001 (n=395) and made EBMA forecasts for the remaining 214 cases with 2002 docket numbers.[24]

Table 5 shows the component weights for the two forecasts and the test-period

fit statistics for the MQRK classification trees, subject experts, and EBMA forecasts. The EBMA model weights the subject experts about twice as much as the statistical model. Once again, the results show that the EBMA procedure outperforms all components (even when there are only two). In terms of AUC, Brier scores, and correct predictions, the EBMA forecast outperforms both the statistical model and the combined subject experts. In addition, EBMA scores substantially better on the PRE metric.[25]

Table 5 about here

There is a long-standing debate in many circles of the relative strengths and weaknesses of statistical models and subject experts for making predictions (e.g., Ascher 1979). Models that use quantifiable measurements and widely available (if sometimes crude) data to make predictions can make egregious errors in particular cases. Some cases may be decided by forces invisible to the statistical model but obvious to experts familiar with the case. Subject experts, on the other hand, can become too focused on minutia and miss larger (if more subtle) trends in the data easily recognized by more advanced methodologies. The EBMA technique offers a theoretically motivated way to combine the strengths of both methods, while smoothing over their relative weaknesses, to make more accurate predictions.

5. DISCUSSION

As currently implemented, EBMA already offers a method for aiding the accurate prediction of future events. However, we envision several paths forward for future research in this area. First, we are planning to extend EBMA into a fully Bayesian

framework. Markov chain Monte Carlo estimation of EBMA models promises to more efficiently handle a wider variety of outcome distributions and will provide additional information regarding our uncertainty about model weights and within-model variances (Vrugt, Diks and Clark 2008).

Second, EBMA estimates model weights based exclusively on the point predictions of component forecasts. Even for continuous data (e.g., the presidential vote forecasts), the current procedure assumes that the within-forecast variance (σ^2) is constant across models. In other words, model weights do not reflect the uncertainty associated with each model's predictions. Applying both Bayesian and bootstrap methods, we intend to incorporate the entire predictive PDFs of component forecasts so that model weights reflect not only components' accuracy, but also their precision. Poorly calibrated models should be penalized and receive less posterior weight.

However, EBMA as it is currently implemented shows considerable promise for aiding systematic social inquiry. For many important and interesting events, it is almost impossible for social scientists to find the "true" data-generating process. Socially determined events are inherently difficult to predict because of nonlinearities and the unpredictability of human behavior. This may be one of the main reasons political scientists so rarely make systematic predictions about the future. Yet, we believe it should be one ultimate goal of the discipline to make sensible and reliable forecasts. Doing so would make the discipline more relevant to policymakers and provide more avenues for rigorous testing of theoretical models and hypothesized empirical regularities.[26]

EBMA uses the accuracy of in-sample predictions of individual models to calibrate a combined weighted-average forecast and to make more accurate predictions. Moreover, it does so in a transparent and theoretically motivated manner that allows us to see which component models are most important in informing the broader EBMA model. Thus, EBMA can enhance the accuracy of forecasts in political science, while also allowing the continued development of multiple theoretical and empirical approaches to the study of important topics. In addition, we have adjusted EBMA to work for dichotomous dependent variables. The EBMA model, therefore, can now be used in large fraction of research in political science. However, the method depends fundamentally on the existence of relatively good individual models, otherwise the ensemble is empty. Thus, EBMA and other ensemble methods should not discourage the development of individual prediction models, but rather leverage their individual contributions with those from other models in order to achieve more accurate predictions.

Finally, we demonstrated the utility of the EBMA method for improving out-of-sample forecasts in three empirical analyses. In each, the EMBA model outperformed its components and was less sensitive to idiosyncratic data issues than the individual models. The EBMA method was applied to improve the prediction of insurgencies around the Pacific Rim, U.S. presidential election results, and the votes of U.S. Supreme Court Justices. However, we believe these applications represent only a portion of the areas to which the EBMA method could be fruitfully applied. Using the software we have developed for this project, it will be possible for researchers to increase the accuracy of forecasts of a wide array of

important events.[27]

APPENDIX A: ADJUSTMENTS FOR DICHOTOMOUS OUTCOMES

Past work on EBMA does not apply directly to the prediction of many political events because the assumed PDFs are normal, Poisson, or gamma. In many settings (e.g., international conflicts), the data are not sufficiently fine-grained to justify these distributional assumptions. Usually, the outcomes of interest are dichotomous indicators for whether an event (e.g., civil war) has occurred in a given time period and country. Thus, none of the distributional assumptions used in past work are appropriate in this context. Fortunately, it is a straightforward extension of Sloughter et al. (2007) and Sloughter, Gneiting and Raftery (2010) to deal appropriately with binary outcomes.[28]

We follow Sloughter et al. (2007) and Hamill, Whitaker and Wei (2004) in using logistic regression after a power transformation of the forecast to reduce prediction bias. That is, point predictions are raised to a power, $\frac{1}{b} \leq 1$. This shrinks predictions downward towards zero. The transformation dampens the effect of extreme observations and helps to reduce over-fitting which might occur because certain models do slightly better in predicting high-leverage observations. Since the predictions for dichotomous outcomes are necessarily between -1 and 1, our adjustment process is slightly more complex. Nonetheless, the results for bias reduction are the same.

For notational ease, we assume that $\mathbf{f}_k$ is the forecast after the adjustment for bias reduction (we will omit the superscripts for the moment). Therefore, let

$\mathbf{f}'_k \in [0,1]$ be the forecast on the predicted probability scale and

$$\mathbf{f}_k = \left[(1+\text{logit}(\mathbf{f}'_k))^{1/b} - 1\right] I\left[\mathbf{f}'_k > \frac{1}{2}\right] - \left[(1+\text{logit}(|\mathbf{f}'_k|))^{1/b} - 1\right] I\left[\mathbf{f}'_k < \frac{1}{2}\right], \quad (7)$$

where $I[.]$ is the general indicator function. Hamill, Whitaker and Wei (2004) recommend setting $b = 4$, while Sloughter et al. (2007) use $b = 3$. We use $b = 3$ in the insurgency example above and $b = 4$ in the courts example. However, we found that this choice makes very little difference for these examples.

The logistic model for the outcome variables is

$$\text{logit } P(\mathbf{y} = 1|\mathbf{f}_k) \equiv \log \frac{P(\mathbf{y} = 1|\mathbf{f}_k)}{P(\mathbf{y} = 0|\mathbf{f}_k)} = a_{k0} + a_{k1}\mathbf{f}_k. \quad (8)$$

The conditional PDF of some within-sample event, given the forecast $f_k^{s|t}$ and the assumption that k is the true model, can be written

$$g_k(y|f_k^{s|t}) = P(y = 1|f_k^{s|t})I[y = 1] + P(y = 0|f_k^{s|t})I[y = 0]. \quad (9)$$

Applying this to (3), the PDF of the final EBMA model for y is

$$\begin{aligned} p(y|f_1^{s|t}, f_2^{s|t}, \ldots, f_K^{s|t}) = & \sum_{k=1}^{K} w_k[P(y = 1|f_k^{s|t})I[y = 1] \\ & + P(y = 0|f_k^{s|t})I[y = 0]]. \end{aligned} \quad (10)$$

Parameter estimation is conducted using only the data from the validation period T. The parameters a_{0k} and a_{1k} are specific to each individual component model. For model k, these parameters can be estimated as traditional linear models where $\mathbf{y}$ is the dependent variable and the covariate list includes only $\mathbf{f}_k$ and a constant term.

The difficulty is in estimating the weighting parameters, $w_k \, \forall \, k \in [1, 2, \ldots, K]$. For the moment, we have followed Raftery et al. (2005) and Sloughter et al. (2007)

in using maximum likelihood methods. In future work we plan to implement a fully Bayesian analysis by placing priors on all parameters and using Markov chain Monte Carlo techniques to estimate model weights (c.f. Vrugt, Diks and Clark 2008).

The log-likelihood function cannot be maximized analytically, but Raftery et al. (2005) and Sloughter et al. (2007) suggest using the expectation-maximization (EM) algorithm. We introduce the unobserved quantities $z_k^{s|t}$, which represent the posterior probability for model k for observation $s|t$. The E step involves calculating estimates for these unobserved quantities using the formula

$$\hat{z}_k^{(j+1)s|t} = \frac{\hat{w}_k^{(j)} p^{(j)}(y|f_k^{s|t})}{\sum_{k=1}^{K} \hat{w}_k^{(j)} p^{(j)}(y|f_k^{s|t})}, \tag{11}$$

where the superscript j refers to the jth iteration of the EM algorithm.

$w_k^{(j)}$ is the estimate of w_k in the jth iteration and $p^{(j)}(.)$ is shown in (10). Assuming these estimates of $z_k^{s|t}$ are correct, it is then straightforward to derive the maximizing value for the model weights. Thus, the M step estimates these as $\hat{w}_k^{(j+1)} = \frac{1}{n}\sum_{s,t} \hat{z}_k^{(j+1)s|t}$, where n represents the number of observations in the validation dataset.[29] The E and M steps are iterated until the improvement in the log-likelihood is no larger than some pre-defined tolerance.[30]

Ensemble prediction for dichotomous outcomes: With these parameter estimates, it is now possible to generate ensemble forecasts. If our forecasts $\mathbf{f}_k^{s|t}$ are each generated from a statistical model, we now generate a new prediction $f_k^{s|t^*}$ from the previously fitted models. For convenience, let $\hat{\mathbf{a}}_k \equiv (\hat{a}_{k0}, \hat{a}_{k1})$. For some

dichotomous observation in the test period $t^* \in T^*$, we can see that

$$P(y = 1 | f_1^{s|t^*}, \ldots, f_K^{s|t^*}; \hat{\mathbf{a}}_1, \ldots, \hat{\mathbf{a}}_K; \hat{w}_1, \ldots, \hat{w}_K) = \sum_{k=1}^{K} \hat{w}_k \text{logit}^{-1} \left(\hat{a}_{k0} + \hat{a}_{k1} f_k^{s|t^*} \right). \tag{12}$$

References

Abramowitz, Alan I. 2008. "Forecasting the 2008 Presidential Election with the Time-for-Change Model." *PS: Political Science & Politics* 41(4):691–695.

Andriole, Stephen J. and Robert A. Young. 1977. "Toward the Development of an Integrated Crisis Warning System." *International Studies Quarterly* 21(1):107–150.

Armstrong, J. Scott. 2001. Combining Forecasts. In *Principles of Forecasting: A Handbook for Researchers and Practitioners*, ed. J. Scott Armstrong. Norwell, MA: Kluwer Academic Publishers.

Bartels, Larry M. 1997. "Specification Uncertainty and Model Averaging." *American Journal of Political Science* 41(2):641–674.

Bartels, Larry M. and John Zaller. 2001. "Presidential Vote Models: A Recount." *PS: Political Science and Politics* 34(1):9–20.

Bates, J.M. and Clive W.J. Granger. 1969. "The Combination of Forecasts." *Operations Research* 20(4):451–468.

Bennett, D. Scott and Allan C. Stam. 2009. "Revisiting Predictions of War Duration." *Conflict Management and Peace Science* 26(3):256–267.

Berg, Joyce E., Forrest D. Nelson and Thomas A. Rietz. 2008. "Prediction Market Accuracy in the Long Run." *International Journal of Forecasting* 24(2):285–300.

Berrocal, Veronica J., Arian E. Raftery, Tilmann Gneiting and Richard C. Steed. 2010. "Probabilistic Weather Forecasting for Winter Road Maintenance." *Journal of the American Statistical Association* 105(490):522–537.

Billio, Monica, Roberto Casarin, Francesco Ravazzolo and Herman K. Van Dijk. 2010. "Combining Predictive Densities Using Bayesian Filtering with Applications to US Economics Data." Norges Bank Working Paper. `http://ssrn.com/abstract=1735421` (accessed June 1, 2011).

Billio, Monica, Roberto Casarin, Francesco Ravazzolo and Herman K. Van Dijk. 2011. "Bayesian Combinations of Stock Price Predictions with an Application to the Amsterdam Exchange Index." Tinbergen Institute Discussion Paper No. 2011-082/4. `http://www.tinbergen.nl/discussionpapers/11082.pdf` (accessed June 1, 2011).

Brandt, Patrick T., John R. Freeman and Philip A. Schrodt. 2011*a*. "Racing Horses: Constructing and Evaluating Forecasts in Political Science." Paper prepared for the 28th Annual Summer Meeting of the Society for Political Methodology. `http://polmeth.wustl.edu/media/Paper/RHMethods20110721small_1.pdf` (accessed August 20, 2011).

Brandt, Patrick T., John R. Freeman and Philip A. Schrodt. 2011*b*. "Real Time, Time Series Forecasting of Inter- and Intra-State Political Conflict." *Conflict Management and Peace Science* 28(1):41–64.

Brandt, Patrick T., Michael Colaresi and John R. Freeman. 2008. "The Dynamics of Reciprocity, Accountability, and Credibility." *The Journal of Conflict Resolution* 52(3):343–374.

Brier, Glenn W. 1950. "Verification of Forecasts Expressed in Terms of Probability." *Monthly Weather Review* 78(1):1–3.

Brock, William A., Steven N. Durlauf and Kenneth D. West. 2007. "Model Uncertainty and Policy Evaluation: Some Theory and Empirics." *Journal of Econometrics* 136(2):629–664.

Brown, Lloyd B. and Henry W. Chappell. 1999. "Forecasting Presidential Elections Using History and Polls." *International Journal of Forecasting* 15(2):127–135.

Bueno de Mesquita, Bruce. 2002. *Predicting Politics*. Columbus, OH: Ohio State University Press.

Bueno de Mesquita, Bruce. 2011. "A New Model for Predicting Policy Choices: Preliminary Tests." *Conflict Management and Peace Science* 28(1):65–85.

Campbell, James E. 1992. "Forecasting the Presidential Vote in the States." *American Journal of Political Science* 36(2):386–407.

Campbell, James E. 2008. "The Trial-heat Forecast of the 2008 Presidential Vote: Performance and Value Considerations in an Open-Seat Election." *PS: Political Science & Politics* 41(4):697–701.

Campbell, James E. and Ken A. Wink. 1990. "Trial-heat Forecasts of the Presidential Vote." *American Politics Research* 18(3):251–269.

Chmielecki, Richard M. and Arian E. Raftery. 2010. "Probabilistic Visibility Forecasting Using Bayesian Model Averaging." *Monthly Weather Review* 139(5):1626–1636.

Choucri, Nazli and Thomas W. Robinson, eds. 1978. *Forecasting in International Relations: Theory, Methods, Problems, Prospects*. San Francisco, CA: W.H. Freeman.

Clyde, Merlise. 2003. Model Averaging. In *Subjective and Objective Bayesian Statistics: Principles, Models and Applications*, ed. S. James Press. Hoboken, NJ: Wiley-Interscience pp. 320–335.

Clyde, Merlise and Edward I. George. 2004. "Model Uncertainty." *Statistical Science* 19(1):81–94.

Cuzàn, Alfred G. and Charles M. Bundrick. 2008. "Forecasting the 2008 Presidential Election: A Challenge for the Fiscal Model." *PS: Political Science & Politics* 41(4):717–722.

Davies, John L. and Ted Robert Gurr. 1998. *Preventive Measures: Building Risk Assessment and Crisis Early Warning Systems*. Lanham, Md: Rowman & Littlefield Publishers.

Dawid, A. Phillip. 1982. "The well-calibrated Bayesian (with Discussion)." *Journal of the American Statistical Association* 77(379):605–613.

Dawid, A. Phillip. 1984. "Present position and potential developments: Some personal views: Statistical theory: The prequential approach (with Discussion)." *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 147(2):278–292.

de Marchi, Scott, Christopher Gelpi and Jeffrey D. Grynaviski. 2004. "Untangling Neural Nets." *American Political Science Review* 98(2):371–378.

de Sola Pool, Ithiel, Robert P. Abelson and Samuel L. Popkin. 1964. *Candidates, issues, and strategies : a computer simulation of the 1960 and 1964 Presidential elections*. Cambridge, MA: MIT Press.

Draper, David. 1995. "Assessment and Propagation of Model Uncertainty." *Journal of the Royal Statistical Society. Series B (Methodological)* 57(1):45–97.

Enders, Walter and Todd Michael Sandler. 2005. "After 9/11: Is It All Different Now?" *The Journal of Conflict Resolution* 49(2):259–277.

Erikson, Robert S. and Christopher Wlezien. 2008. "Leading Economic Indicators, the Polls, and the Presidential Vote." *PS: Political Science & Politics* 41(4):703–707.

Fair, Ray C. 1978. "The Effect of Economic Events on Votes For President." *The Review of Economics and Statistics* 60(2):159–173.

Fair, Ray C. 2010. "Presidential and Congressional Vote-Share Equations: November 2010 Update." Working Paper, Yale University. `http://fairmodel.econ.yale.edu/RAYFAIR/PDF/2010C.pdf` (accessed June 07, 2011).

Fair, Ray C. 2011. "Vote-Share Equations: November 2010 Update." Working Paper, Yale University. `http://fairmodel.econ.yale.edu/vote2012/index2.htm` (accessed March 07, 2011).

Fearon, James D. and David D. Laitin. 2003. "Ethnicity, Insurgency and Civil War." *American Political Science Review* 97(1):75–90.

Feder, Stanley A. 2002. "Forecasting For Policy Making in the Post-Cold War Period." *Annual Review of Political Science* 5:111–125.

Feldkircher, Martin. 2011. "Forecast Combination and Bayesian Model Averaging: A Prior Sensitivity Analysis." *Journal of Forecasting* p. in press.
URL: *http://dx.doi.org/10.1002/for.1228*

Fraley, Chris, Adrian E. Raftery, J. McLean Sloughter and Tilmann Gneiting. 2010. *ensembleBMA: Probabilistic Forecasting using Ensembles and Bayesian*

Model Averaging. R package version 4.5.
URL: *http://CRAN.R-project.org/package=ensembleBMA*

Fraley, Chris, Adrian E. Raftery and Tilmann Gneiting. 2010. "Calibrating Multimodel Forecast Ensembles with Exchangeable and Missing Members Using Bayesian Model Averaging." *Monthly Weather Review* 138(1):190–202.

Fraley, Chris, Adrian E. Raftery, Tilmann Gneiting, J. McLean Sloughter and Veronica J. Berrocal. 2011. "Probabilistic Weather Forecasting in R." *R Journal* 3(1):55–63.

Freeman, John R. and Brian L. Job. 1979. "Scientific Forecasts in International Relations: Problems of Definition and Epistemology." *International Studies Quarterly* 23(1):113–143.

Geer, John and Richard R. Lau. 2006. "Filling in the Blanks: A New Method for Estimating Campaign Effects." *British Journal of Political Science* 36(2):269–290.

Gill, Jeff. 2004. "Introduction to the Special Issue." *Political Analysis* 12(4):647–674.

Gleditsch, Kristian Skrede and Michael D. Ward. 2010. "Contentious Issues and Forecasting Interstate Disputes." Presented to the 2010 Annual Meeting of the International Studies Association, New Orleans, LA.

Gneiting, Tilmann and Adrian E. Raftery. 2007. "Strictly Proper Scoring Rules, Prediction, and Estimation." *Journal of the American Statistical Association* 102(477):359–378.

Gneiting, Tilmann and Thordis L. Thorarinsdottir. 2010. "Predicting Inflation: Professional Experts Versus No-Change Forecasts." Working Paper. `http://arxiv.org/abs/1010.2318v1` (accessed June 15, 2011).

Graefe, Andreas, Aldfred G. Cuzan, Randal J. Jones and J. Scott Armstrong. 2010. "Combining Forecasts for U.S. Presidential Elections: The PollyVote." Working Paper. `http://dl.dropbox.com/u/3662406/Articles/Graefe_et_al_Combining.pdf` (accessed May 15, 2011).

Greenhill, Brian, Michael D. Ward and Audrey Sacks. 2011. "The Separation Plot: A New Visual Method for Evaluating the Fit of Binary Data." *American Journal of Political Science* 55(4):990–1002.

Gurr, Ted Robert and Mark Irving Lichbach. 1986. "Forecasting Internal Conflict: A Competitive Evaluation of Empirical Theories." *Comparative Political Studies* 19(3):3–38.

Hamill, Thomas S., Jeffrey S. Whitaker and Xue Wei. 2004. "Ensemble Reforecasting: Improving Medium-range Forecast Skill Using Retrospective Forecasts." *Monthly Weather Review* 132(6):1434 – 1447.

Hastie, Trevor, Robert Tibshirani and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer.

Hausegger, Lori and Lawrence Baum. 1999. "Inviting Congressional Action: A Study of Supreme Court Motivations in Statutory Interpretation." *American Journal of Political Science* 43(1):162–185.

Hibbs, Douglas. 2011. "The Bread and Peace Model Applied to the 2008 U.S. Presidential Election." `http://douglas-hibbs.com/Election2008/2008Election-MainPage.htm` (accessed March 08, 2011).

Hibbs, Douglas A. 2000. "Bread and Peace Voting in U.S. Presidential Elections." *Public Choice* 104(1):149–180.

Hildebrand, David K., James D. Laing and Howard Rosenthal. 1976. "Prediction Analysis in Political Research." *The American Political Science Review* 70(2):509–535.

Hoeting, Jennifer A., David Madigan, Adrian E. Raftery and Christopher T. Volinsky. 1999. "Bayesian Model Averaging: A Tutorial." *Statistical Science* 14(4):382–417.

Holbrook, Thomas M. 2008. "Incumbency, National Conditions, and the 2008 Presidential Election." *PS: Political Science & Politics* 41(4):709–712.

Huisman, J.A., L. Breuer, H. Bormann, A. Bronstert, B.F.W. Croke, H.-G. Frede, T. Gräff, L. Hubrechts, A.J. Jakeman, G. Kite, J. Lanini, G. Leavesley, D.P. Lettenmaier, G. Lindström, J. Seibert, M. Sivapalan, N.R. Viney and P. Willems. 2009. "Assessing the Impact of Land Use Change on Hydrology by Ensemble Modelling (LUCHEM) II: Ensemble Combinations and Predictions." *Advances in Water Resources* 32(2):147–158.

Imai, Kosuke and Gary King. 2004. "Did Illegal Overseas Absentee Ballots Decide the 2000 US Presidential Election?" *Perspectives on Politics* 2(3):537–549.

Jerome, Bruno, Veronique Jerome and Michael S. Lewis-Beck. 1999. "Polls Fail in France: Forecasts of the 1997 Legislative Election." *International Journal of Forecasting* 15(2):163–174.

King, Gary and Langche Zeng. 2001. "Improving Forecasts of State Failure." *World Politics* 53(4):623–658.

Klein, David E. and Robert J. Hume. 2003. "Fear of Reversal as an Explanation of Lower Court Compliance." *Law & Society Review* 37(3):579–606.

Koop, Gary and Dimitris Korobilis. 2009. "Forecasting Inflation Using Dynamic Model Averaging." Working Paper. `http://personal.strath.ac.uk/gary.koop/koop_korobilis_forecasting_inflation_using_DMA.pdf` (accessed May 25, 2011).

Krause, George A. 1997. "Voters, Information Heterogeneity, and the Dynamics of Aggregate Economic Expectations." *American Journal of Political Science* 41(4):1170–1200.

Leblang, David and Shanker Satyanath. 2006. "Institutions, Expectations, and Currency Crises." *International Organization* 60(1):245–262.

Lewis-Beck, Michael S. 2005. "Election Forecasting: Principles and Practice." *The British Journal of Politics & International Relations* 7(2):145–164.

Lewis-Beck, Michael S. and Charles Tien. 2008. "The Job of President and the Jobs Model Forecast: Obama for '08?" *PS: Political Science & Politics* 41(4):687–690.

Lock, Kari and Andrew Gelman. 2010. "Bayesian Combination of State Polls and Election Forecasts." *Political Analysis* 18(3):337–348.

Lockerbie, Brad. 2008. "Election Forecasting: The Future of the Presidency and the House." *PS: Political Science & Politics* 41(4):713–716.

Madigan, David and Adrian E. Raftery. 1994. "Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window." *Journal of the American Statistical Association* 89(428):1535–1546.

Marshall, Monty G., Keith Jaggers and Ted Robert Gurr. 2009. "Polity IV Project: Political Regime Characteristics and Transition 1800-2007." CIDCM: University of Maryland, MD.

Martin, Andrew D., Kevin M. Quinn, Theodore W. Ruger and Pauline T. Kim. 2004. "Competing Approaches to Predicting Supreme Court Decision Making." *Perspectives on Politics* 2(4):761–767.

McCandless, Tyler C., Sue Ellen Haupt and George S. Young. 2011. "The Effects of Imputing Missing Data on Ensemble Temperature Forecasts." *Journal of Computers* 6(2):162–171.

McCormick, Tyler H., Adrian E. Raftery, David Madigan and Randall S. Burd. 2011. "Dynamic Logistic Regression and Dynamic Model Averaging for Binary Classification." Working Paper. `http://www.stat.columbia.edu/~madigan/PAPERS/ldbma27.pdf` (accessed March 26, 2011).

Min, Seung-Ki and Andreas Hense. 2006. "A Bayesian Approach to Climate Model Evaluation and Multi-Model Averaging with an Application to Global Mean Surface Temperatures from IPCC AR4 Coupled Climate Models." *Geophysical Research Letters* 33(8):L08708.

Min, Seung-Ki, Daniel Simonis and Andreas Hense. 2007. "Probabilistic Climate Change Predictions Applying Bayesian Model Averaging." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 365(1857):2103–2116.

Montgomery, Jacob M. and Brendan Nyhan. 2010. "Bayesian Model Averaging: Theoretical Developments and Practical Applications." *Political Analysis* 18(2):245–270.

Montgomery, Jacob M., Florian Hollenbach and MIchael D. Ward. 2012. ""'Replication data for: Improving Predictions Using Ensemble Bayesian Model Averaging"."'. http://hdl.handle.net/1902.1/17286 IQSS Dataverse Network.

Muhlbaier, Michael D. and Robi Polikar. 2007. "An Ensemble Approach for Incremental Learning in Nonstationary Environments." *Multiple Classifier Systems* 4472:490–500.

Norpoth, Helmut. 2008. "On the Razor's Edge: The Forecast of the Primary Model." *PS: Political Science & Politics* 41(4):683–686.

O'Brien, Sean P. 2002. "Anticipating the Good, the Bad, and the Ugly: An Early Warning Approach to Conflict and Instability Analysis." *Journal of Conflict Resolution* 46(6):791–811.

O'Brien, Sean P. 2010. "Crisis Early Warning and Decision Support: Contemporary Approaches and Thoughts on Future Research." *International Studies Review* 12(1):87–104.

Page, Scott E. 2008. "Uncertainty, Difficulty, and Complexity." *Journal of Theoretical Politics* 20(2):115–149.

Page, Scott E. 2011. *Diversity and Complexity*. Princeton,N.J.: Princeton University Press.

Page, Scott E., Leonard M. Sander and Casey M. Schneider-Mizell. 2007. "Conformity and Dissonance in Generalized Voter Models." *Journal of Statistical Physics* 128(6):1279–1287.

Pevehouse, Jon C. and Joshua S. Goldstein. 1999. "Serbian Compliance or Defiance in Kosovo? Statistical Analysis and Real-Time Predictions." *The Journal of Conflict Resolution* 43(4):538–546.

Raftery, Adrian E. 1995. "Bayesian Model Selection in Social Research." *Sociological Methodology* 25(1):111–163.

Raftery, Adrian E., Miroslav Kárnỳ and Pavel Ettler. 2010. "Online Prediction Under Model Uncertainty via Dynamic Model Averaging: Application to a Cold Rolling Mill." *Technometrics* 52(1):52–66.

Raftery, Adrian E., Tilmann Gneiting, Fadoua Balabdaoui and Michael Polakowski. 2005. "Using Bayesian Model Averaging to Calibrate Forecast Ensembles." *Monthly Weather Review* 133(5):1155–1174.

Raftery, Adrian E. and Yingye Zheng. 2003. "Long-run Performance of Bayesian Model Averaging." *Journal of the American Statistical Association* 98(464):931–938.

Richards, Mark J. and Herbert M. Kritzer. 2002. "Jurisprudential Regimes in Supreme Court Decision Making." *The American Political Science Review* 96(2):305–320.

Rosenstone, Steven J. 1983. *Forecasting Presidential Elections*. New Haven, CT: Yale University Press.

Ruger, Theadore W., Pauline T. Kim, Andrew D. Martin and Kevin M. Quinn. 2004. "The Supreme Court Forecasting Project: Legal and Political Science Approaches to Predicting Supreme Court Decisionmaking." *Columbia Law Review* 104(4):1150–1210.

Schneider, Gerald, Nils Petter Gleditsch and Sabine Carey. 2011. "Forecasting in International Relations: One Quest, Three Approaches." *Conflict Management and Peace Science* 28(1):5–14.

Schrodt, Philip A. and Deborah J. Gerner. 2000. "Using Cluster Analysis to Derive Early Warning Indicators for Political Change in the Middle East, 1979-1996." *American Political Science Review* 94(4):803–818.

Segal, Jeffrey A. and Albert D. Cover. 1989. "Ideological Values and the Votes of U.S. Supreme Court Justices." *The American Political Science Review* 83(2):557–565.

Singer, J. David and Michael D. Wallace. 1979. *To Augur Well: Early Warning Indicators in World Politics*. Beverly Hills, CA: Sage Publications.

Sloughter, J. McLean, Adrian E. Raftery, Tilmann Gneiting and Chris Fraley. 2007. "Probabilistic Quantitative Precipitation Forecasting Using Bayesian Model Averaging." *Monthly Weather Review* 135(9):3209–3220.

Sloughter, J. McLean, Tilmann Gneiting and Adrian E. Raftery. 2010. "Probabilistic Wind Speed Forecasting Using Ensembles and Bayesian Model Averaging." *Journal of the American Statistical Association* 105(489):25–35.

Smith, Richard L., Claudia Tebaldi, Doug Nychka and Linda O. Mearns. 2009. "Bayesian Modeling of Uncertainty in Ensembles of Climate Models." *Journal of the American Statistical Association* 104(485):97–116.

Songer, Donald R., Jeffrey A. Segal and Charles M. Cameron. 1994. "The Hierarchy of Justice: Testing a Principal-Agent Model of Supreme Court-Circuit Court Interactions." *American Journal of Political Science* 38(3):673–696.

Spirtes, Peter, Clark N. Glymour and Richard Scheines. 2000. *Causation, Prediction, and Search*. Vol. 81 Cambridge, MA: MIT Press.

Tomas, Amber. 2011. "A Dynamic Logistic Multiple Classifier System for Online Classification." Working Paper. `http://www.stats.ox.ac.uk/~tomas/html_links/T2011.pdf` (accessed June 1, 2011).

Vincent, Jack E. 1980. "Scientific Prediction versus Crystal Ball Gazing: Can the Unknown be Known?" *International Studies Quarterly* 24(3):450–454.

Vrugt, Jasper A., Cees G.H. Diks and Martyn P. Clark. 2008. "Ensemble Bayesian Model Averaging Using Markov Chain Monte Carlo Sampling." *Environmental Fluid Mechanics* 8(5):579–595.

Vrugt, Jasper A., Martyn P. Clark, Cees G.H. Diks, Qinyun Duan and Bruce A. Robinson. 2006. "Multi-Objective Calibration of Forecast Ensembles Using Bayesian Model Averaging." *Geophysical Research Letters* 33:L19817.

Ward, Michael D., Brian D. Greenhill and Kristin M. Bakke. 2010. "The Perils of Policy by p-value: Predicting Civil Conflict." *Journal of Peace Research* 47(4):363–375.

Ward, Michael D., Randolph M. Siverson and Xun Cao. 2007. "Disputes, Democracies, and Dependencies: A Re-examination of the Kantian Peace." *American Journal of Political Science* 51(3):583–601.

Whiteley, Paul F. 2005. "Forecasting Seats from Votes in British General Elections." *The British Journal of Politics & International Relations* 7(2):165–173.

Wright, Jonathan H. 2008. "Bayesian Model Averaging and Exchange Rate Forecasts." *Journal of Econometrics* 146(2):329–341.

Wright, Jonathan H. 2009. "Forecasting US Inflation by Bayesian Model Averaging." *Journal of Forecasting* 28(2):131–144.

Zhang, Xuesong, Raghavan Srinivasan and David Bosch. 2009. "Calibration and Uncertainty Analysis of the SWAT Model Using Genetic Algorithms and Bayesian Model Averaging." *Journal of Hydrology* 374(3-4):307–317.

Notes

[1]An incomplete list of recent work would include Krause (1997), Davies and Gurr (1998), Pevehouse and Goldstein (1999), Schrodt and Gerner (2000), King and Zeng (2001), O'Brien (2002), Bueno de Mesquita (2002), Fearon and Laitin (2003), de Marchi, Gelpi and Grynaviski (2004), Enders and Sandler (2005), Leblang and Satyanath (2006), Ward, Siverson and Cao (2007), Brandt, Colaresi and Freeman (2008), Bennett and Stam (2009), and Gleditsch and Ward (2010). A summary of classified efforts is reported in Feder (2002). An overview of some of the historical efforts along with a description of current thinking about forecasting and decision-support is given by O'Brien (2010).

[2]Lewis-Beck (2005) provides a more in-depth discussion of election forecasting in a comparative context.

[3]The case for using predictions heuristically can also be found in early work by Dawid (1982, 1984).

[4]To fully capture the ability of EBMA to reduce over-fitting when using statistical models it is necessary to divide the data into three periods (Hastie, Tibshirani and Friedman 2009). The first period, the training period, is used to fit the parameters for each component model. The second period, the validation period, is used to calculate model weights and other parameters for the EBMA model using out-of-sample predictions generated from the component models. We then generate ensemble predictions for the third period, the test period, using the EBMA model parameters calculated in period two. This approach is explicit in the insurgency forecasting example below and implicit in the Supreme Court example below since the subject-experts and classification algorithm were "trained" on data not included in the study. This three-stage method is adjusted in the election forecasting example as the component models are already sparse (somewhat ameliorating concerns about over-fitting) and there are a much smaller number of observations. In this example, component models are trained over the period beginning in 1916 and the EBMA parameters

are calculated only for the period beginning in 1952. However, there is significant overlap in the training and validation samples.

[5]In the case of subject-experts, the training period is implicitly the period over which experts have gained their experience. Forecasts will only be necessary for the validation period.

[6]Sloughter et al. (2007) make predictions for only one future time period, and use only a subset of past time-periods (they recommend 30) in their validation period. Thus, predictions are made sequentially with the entire EBMA procedure being re-calculated for each future event as observations are moved from the test period T^* into the validation period T. Another alternative is to simply divide *all* the data into discrete training, validation and test periods for the entire procedure. We use both approaches in our examples below.

[7]Our adjustments to the basic EBMA method for application to dichotomous outcomes, as well as details of parameter estimation, are shown in Appendix A.

[8]The twenty-nine countries are Australia, Bangladesh, Bhutan, Cambodia, China, Comoros, Fiji, India, Indonesia, Japan, Laos, Madagascar, Malaysia, Mauritius, Mongolia, Myanmar, Nepal, New Zealand, North Korea, Papua New Guinea, Philippines, Russia, Singapore, Solomon Islands, South Korea, Sri Lanka, Taiwan, Thailand, and Vietnam. This set is not a random sample, but rather constitutes the countries of population greater than $500,000$ that are in the area of responsibility of the US Pacific Command.

[9]Because some of the models include lagged data, this is the first year for which all of the component models produce fitted values or predictions.

[10]See `strategicanalysisenterprises.com` for more details. All data and models will be included in a replication dataset at the time of publication.

[11]It is worth noting that the mixed effects model is a kind of ensemble mixture in that in averages the so-called within model with the between model.

[12]This is calculated as the number of days between the middle of the current month and the last federal election regardless of the legitimacy of the election.

[13]Geographical proximity is measured in terms of the length of the shared border between the two countries.

[14]One alternative approach to generating ensemble forecasts would be to use the simple average of each component forecast. However, this causes difficulties because the researcher must use their own judgement to decide which alternative models are sufficiently accurate and diverse for inclusion. EBMA offers a more statistically motivated and straightforward method for achieving the same end. In any case, these simple averages do not perform well against the EBMA forecast. In the current example, a simple unweighted average results in $AUC = 0.885$, PRE=0.123, Brier = 0.052, and % Correct = 92.8 for the test-period. This is not surprising given that simple averaging weights an inaccurate model the same as an accurate one. EBMA on the other hand is able to detect the superiority of components and calibrates weights accordingly. Likewise, simple averages cannot identify pairs or groups of highly correlated forecasts and will tend to give these groupings too much weight.

[15]It is important to note that we attempted to replicate each of the models for the 2008 election as closely as possible given the model descriptions in the articles and the data provided by the authors. We then proceeded to use the same model specifications as used in the 2008 articles to forecast all elections previous to 2008. Thus, prior to 2008 the individual model results are not exact replications of the author's given prediction for that election year and results may vary from what was presented by the authors as the forecast for a given election. This may be due to changes in the model specification over time and data updates. Thus, we neither attempted nor succeeded in replicating the exact forecasts for all election years for all components.

[16]The model here replicates Equation 1 in Fair (2010).

[17]The data to replicate the models by Abramowitz (2008), Campbell (2008), Erikson and Wlezien

(2008), and Lewis-Beck and Tien (2008) were provided in personal correspondence with the respective authors. The remaining data were downloaded from the web sites of Ray C. Fair (`http://fairmodel.econ.yale.edu/vote2012/tbl1.txt`) and Douglas Hibbs (`http://www.douglas-hibbs.com/`).

[18]For example, the Fair model uses data for election results beginning in 1916 while the Abramowitz model begins with data from the 1952 election.

[19]See footnote 4 for additional discussion of the implications of the overlapping training and validation samples. Results in this section were computed using modifications of the 'ensembleBMA' package (Fraley et al. 2010, 2011). Because of the paucity of data, we did not apply any bias correction to these forecasts. Thus, the predictor and constant, denoted a_{0k} and a_{1k} above, are constrained to zero and one respectively.

[20]As we noted above, these models are fit sequentially, so the validation periods change. For example, the validation period for the forecast of the 2004 election is 1952-2000. The validation-period RMSE is therefore calculated for those observations. For the 2008 election, the validation period is 1952-2004.

[21]The correlation matrix between fitted-values of the model for the 1952-2004 period is:

	C	**A**	**H**	**F**	**L**	**E**
Campbell	1.00					
Abramowitz	0.94	1.00				
Hibbs	0.91	0.93	1.00			
Fair	0.87	0.89	0.89	1.00		
Lewis-Beck/Tien	0.93	0.96	0.91	0.88	1.00	
EWT2C2	0.85	0.90	0.87	0.91	0.86	1.00

[22]Brandt, Freeman and Schrodt (2011*a*) survey a variety of metrics in addition to those we employ here. These include measures of average prediction errors, measures using medians and geometric averages, measures that compare the complete difference in probability distributions, and sequential rank-based methods. Although there are many candidate metrics, at least for the alternative metrics we have calculated so far, the substantive conclusions we reach do not change for our examples and they are not presented due to space constraints. However, as suggested by a helpful reviewer, we note that there are reasons to doubt that RMSE or MAE will necessarily provide a ranking of component models based on accuracy. A more complete approach evaluating the accuracy of the component models is to examine the results displayed in Figure 3 below.

[23]Additional details about the project, replication files, as well as a complete listing of cases and expert forecasts are available at: `http://wusct.wustl.edu/index.php`.

[24]As noted above, these cases were heard in the 2002-2003 period. We note that the dates on the docket number do not necessarily reflect the order in which they were argued before the court and the order in which cases were argued did not correspond to when the decisions were handed down. Thus, there is no obvious way to partition the data into validation and test periods. However, in general, the docket numbers roughly correspond to the age of the case. Although partitioning the data in this manner is slightly arbitrary, it serves the limited purpose of demonstrating the method.

[25]The baseline model here is the prediction that all votes will be to reverse the lower court. This baseline model is correct for roughly 70% of the votes in the test period.

[26]In addition, some scholars have advanced the argument that prediction is closely related to the identification of causal processes (e.g., Spirtes, Glymour and Scheines 2000). However, this is far from a universally accepted position and is not the basis for our advocacy of increased forecasting in political science.

[27]All data used to generate the results in this article will be made available to the public

in the journal's dataverse upon publication at `http://hdl.handle.net/1902.1/17286` (Montgomery, Hollenbach and Ward 2012). The package for ensemble Bayesian Model Averaging, `EBMAforecast` is available through the Comprehensive R-Archive Network at `http://cran.r-project.org/`.

[28]The method for dealing with binary outcomes is implicit in Sloughter et al. (2007) and Sloughter, Gneiting and Raftery (2010), which assume a discrete-continuous distribution for outcomes that includes a logistic component. However, they do not explicitly and fully develop the model for dichotomous outcomes. A related strain of research on Dynamic Model Averaging (c.f., Raftery, Kárnỳ and Ettler 2010; Muhlbaier and Polikar 2007) has recently been extended for direct application to binary outcomes (e.g., McCormick et al. 2011; Tomas 2011).

[29]In the case of normally distributed data, $\hat{\sigma}^{2(j+1)} = \frac{1}{n}\sum_{s,t}\sum_{k=1}^{K}\hat{z}_k^{(j+1)s|t}(y - f_k^{s|t})^2$.

[30]In the examples above, we begin with the assumption that all models are equally likely, $w_k = \frac{1}{K}\ \forall\ k \in [1, \ldots, K]$. Critics of MLE methods and the EM algorithm have raised concerns that convergence to a local rather than a global maximum may occur. We have found no differences in our results based on different starting values, but convergence can be slow if starting values are too dissimilar from the final estimates. Although we feel confident in the results reported here, in future research, we plan to expand the model estimation technique to include Bayesian methods. This will facilitate comparisons of estimates resulting from multiple estimation techniques.

List of Figures

1 Separation plots for validation-period predictions of the ICEWS data (n=696). For each model, observations are shown from left to right in order of increasing predicted probability of insurgency (shown as the black line). Observations where insurgency actually occurred are shown in red. EBMA outperforms all component models in assigning high predicted probabilities to *more* observed insurgencies and to *fewer* non-insurgencies. 54

2 Separation plots for the test-period predictions of the ICEWS data (n=348). For each model, observations are shown from left to right in order of increasing predicted probability (shown as the black line). Observations where insurgency actually occurred are shown in red. EBMA outperforms all component models in assigning high predicted probabilities to *more* observed insurgencies and to *fewer* non-insurgencies. 55

3 The predicted and actual percentage of the two-party vote going to the incumbent party in U.S. presidential elections from six component models and the EBMA forecast. For each year, the plots show the point predictions (circles), 67% predictive intervals (thick horizontal lines), and 90% predictive intervals (thin horizontal lines). The vertical dashed line is the observed outcome. The EBMA model is better calibrated than its components. 56

Table 1: Validation-period results (2008-2009). The table shows estimated model weights, parameters, and fit statistics for the EBMA deterministic forecast and all component forecasts of insurgency in 29 countries of the US Pacific Command. EBMA outperforms any single model on most measures.

	Weight	Constant	Predictor	AUC	PRE	Brier	% Correct
LMER	0.85	-1.89	2.58	0.97	-0.58	0.08	87.07
SAE	0.14	-1.25	3.11	0.92	-0.21	0.07	90.09
GLM	0.00	-1.76	1.42	0.66	0.00	0.08	91.81
EBMA				0.96	0.65	0.04	97.13

n=696

Table 2: Test-period results (2010). The table shows fit statistics for the EBMA deterministic forecast and all component model forecasts of insurgency in 29 countries of the Pacific Rim for the test-period. EBMA equals or outperforms any single model on all measures.

	AUC	PRE	Brier	% Correct
LMER	0.97	0.11	0.08	91.09
SAE	0.96	0.20	0.06	91.95
GLM	0.72	0.00	0.09	89.94
EBMA	0.97	0.43	0.04	94.25

n=348

Table 3: Test-period prediction errors, model weights, and validation-period fit statistics for component and EBMA forecasts of the 2004 and 2008 elections. The models are trained using *all* prior data and the EBMA model is validated on the observations beginning in 1952. The EBMA model does better than all components on validation-sample fit statistics. In addition, although it does not necessarily make the most accurate prediction for any given year, it is less likely to make dramatic forecasting errors for the test-period.

	2004 Election				*2008 Election*			
	Weights	RMSE	MAE	Pred. Error	Weights	RMSE	MAE	Pred. Error
Campbell	0.40	1.71	1.33	0.53	0.36	1.65	1.28	6.33
Abramowitz	0.00	1.50	1.18	2.20	0.06	1.53	1.26	-2.37
Hibbs	0.12	1.95	1.38	1.54	0.25	1.92	1.38	-1.39
Fair	0.48	2.07	1.47	4.82	0.00	2.22	1.80	-2.02
Lewis-Beck/Tien	0.00	1.67	1.42	-0.41	0.17	1.61	1.33	-2.64
EWT2C2	0.00	2.67	2.06	4.76	0.17	2.81	2.18	-0.14
EBMA		1.29	1.02	2.08		1.30	1.01	-0.53

Table 4: Fit statistics and observed coverage probabilities for sequentially generated test-sample predictions of presidential elections from 1976-2008. EBMA outperforms its component models on all metrics.

			Coverage	
	RMSE	MAE	67%	90%
Campbell	2.74	1.99	0.67	0.78
Abramowitz	2.27	2.05	0.33	0.78
Hibbs	2.81	2.24	0.22	0.56
Fair	4.01	3.20	0.44	0.78
Lewis-Beck	2.27	1.82	0.89	1.00
EWT2C2	2.88	2.16	0.78	1.00
EBMA	1.72	1.47	0.67	0.89

Table 5: Test-period results for U.S. Supreme Court example. The table shows fit statistics for the EBMA deterministic forecast and component forecasts of U.S. Supreme Court votes on cases in the 2002-2003 session with 2002 docket numbers. EBMA outperforms its component models on all metrics.

	Weight	AUC	PRE	Brier	% Correct
MQRK model	0.32	0.66	-0.02	0.29	70.56
Subject experts	0.68	0.62	0.15	0.23	75.23
EBMA forecast		0.70	0.21	0.18	77.10

n=214

Figure 1: Separation plots for validation-period predictions of the ICEWS data (n=696). For each model, observations are shown from left to right in order of increasing predicted probability of insurgency (shown as the black line). Observations where insurgency actually occurred are shown in red. EBMA outperforms all component models in assigning high predicted probabilities to *more* observed insurgencies and to *fewer* non-insurgencies.

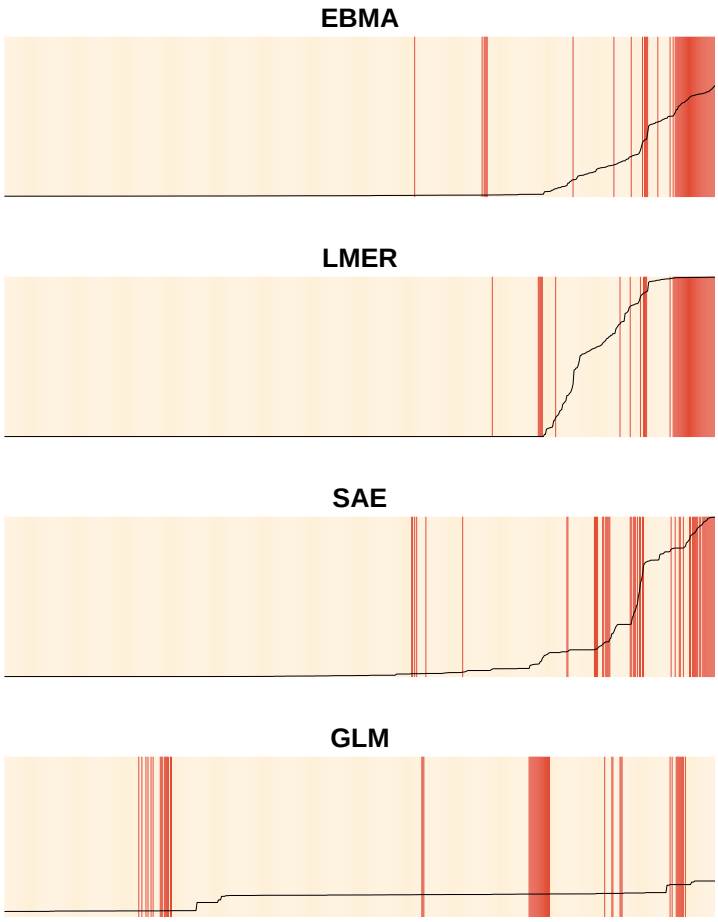

Figure 2: Separation plots for the test-period predictions of the ICEWS data (n=348). For each model, observations are shown from left to right in order of increasing predicted probability (shown as the black line). Observations where insurgency actually occurred are shown in red. EBMA outperforms all component models in assigning high predicted probabilities to *more* observed insurgencies and to *fewer* non-insurgencies.

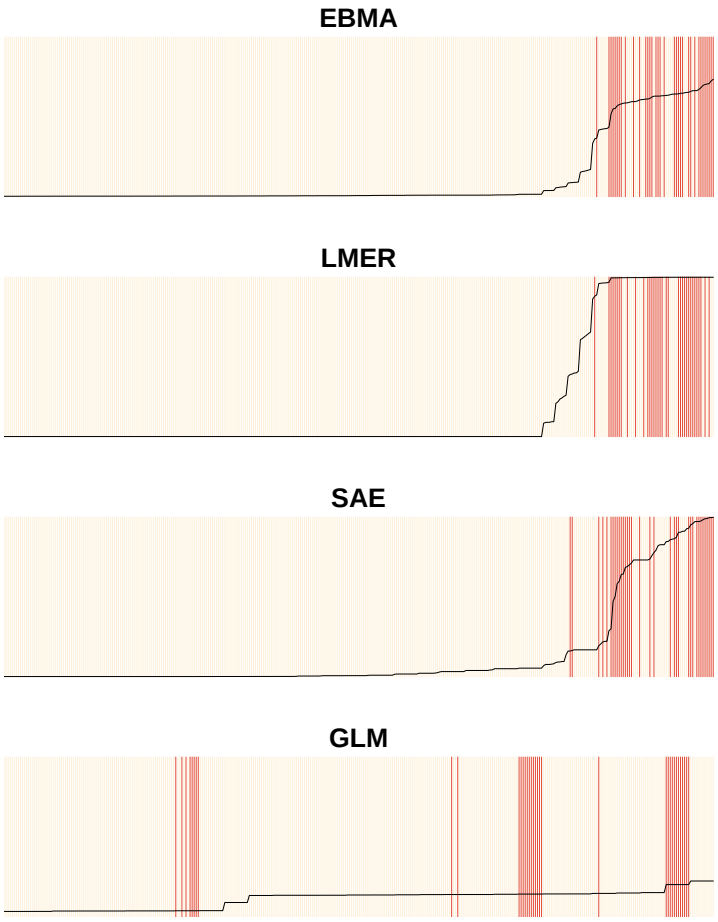

Figure 3: The predicted and actual percentage of the two-party vote going to the incumbent party in U.S. presidential elections from six component models and the EBMA forecast. For each year, the plots show the point predictions (circles), 67% predictive intervals (thick horizontal lines), and 90% predictive intervals (thin horizontal lines). The vertical dashed line is the observed outcome. The EBMA model is better calibrated than its components.