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

September 23, 2011

ABSTRACT

We extend ensemble Bayesian model averaging (EBMA) for application to binary outcomes and illustrate EBMA's 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 training 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 each forecast is calibrated via its performance in some training 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, stochastic simulations, or subject-matter experts.

The rest of this article will proceed as follows. 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 et al. 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.

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 et al. 2011; Bueno de Mesquita 2011; Brandt et al. 2011). Ward et al. (2010) and Greenhill et al. (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 1990s, scholars of U.S. politics began publishing predictions of presidential elections (Campbell and Wink 1990; Campbell 1992). These efforts were 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 et al. 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-existant. To our knowledge, the only non-IR example is the PollyVote project (c.f. Graefe et al. 2010), which combines multiple predictions using simple averages to forecast U.S. presidential elections.

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). Combining forecasts not only reduces reliance on single data sources and methodologies (which lowers the likelihood of dramatic errors), but also allows for the incorpo-

---

[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 et al. (2004), Enders and Sandler (2005), Leblang and Satyanath (2006), Ward et al. (2007), Brandt et al. (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.

ration of more information than any one theoretical or statistical model is likely to include in isolation.

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 et al. 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 et al. 2007; Smith et al. 2009), and hydrology (Zhang et al. 2009). Indeed, research is underway to extend the method to handle missing data (Fraley et al. 2010; McCandless et al. 2011) as well as calibrate model weights on non-likelihood criteria (e.g., Vrugt et al. 2006).

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 contexts (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.

3.1. *Mathematical foundation*

Assume we have some quantity of interest in the future to forecast, $\mathbf{y}^*$, based on previously collected training data $\mathbf{y}^T$ that is fit to K statistical models, $M_1, M_2, \ldots, M_K$. Each model, M_k, is assumed to come from the prior probability distribution $M_k \sim \pi(M_k)$, and the probability distribution function (PDF) for the training data is $p(\mathbf{y}^T|M_k)$. The outcome of interest is distributed $p(\mathbf{y}^*|M_k)$. Applying Bayes Rule, we get that

$$p(M_k|\mathbf{y}^T) = \frac{p(\mathbf{y}^T|M_k)\pi(M_k)}{\sum\limits_{k=1}^{K} p(\mathbf{y}^T|M_k)\pi(M_k)}. \tag{1}$$

and the marginal predictive PDF for y^* is

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

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 training data. Likewise, we can simply make a deterministic estimate using the weighted predictions of the components, denoted

$$E(\mathbf{y}^*) = \sum_{k=1}^{K} E(\mathbf{y}^*|M_k)p(M_k|\mathbf{y}^T). \tag{3}$$

3.2. *EBMA for dynamic settings*

We now turn to applying this basic BMA technology to prediction in a dynamic setting. In generating predictions of important events (e.g., domestic crises or international disputes), the task is to first build a statistical model for some set of observations S in time periods T, which we refer to as the training period.[3] Using the same statistical model (or general technique in the case of subject-expert predictions), we then generate forecasts, $\mathbf{f}_k$, for observations S in future time periods T^*.

Let us assume, for example, that we have K models forecasting insurgencies in a set of countries S. Each component forecast, $\mathbf{f}_k$, is associated with a component PDF, $g_k(\mathbf{y}|\mathbf{f}_k)$, which may be the original predictive PDF from the forecast model or a bias-corrected forecast. These components are the conditional PDFs of outcome $\mathbf{y}$ given the kth forecast, $\mathbf{f}_k$ assuming that $P(M_k|\mathbf{y}) \equiv w_k = 1$, or that the posterior odds of model k is unity.

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

$$p(\mathbf{y}|\mathbf{f}_1, \ldots, \mathbf{f}_K) = \sum_{k=1}^{K} w_k g_k(\mathbf{y}|\mathbf{f}_k), \tag{4}$$

---

[3]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 training data. Thus, predictions are made sequentially with the entire EBMA procedure being re-calculated for each future event as observations are moved from the out-of-sample period T^* into the training set T. Another alternative is to simply divide *all* the data into discrete training and test periods for the entire procedure. We use both approaches in our examples below.

where the weight, w_k, is based on forecast k's relative predictive performance in the training period T. The w_k's $\in [0, 1]$ are probabilities and $\sum_{k=1}^{K} w_k = 1$. The specific PDF of for an out-of-sample event, y_{st*}, is therefore

$$p(y_{st*}|f_{1st*}, \ldots, f_{Kst*}) = \sum_{k=1}^{K} w_k g_k(y_{st*}|f_{kst*}). \tag{5}$$

3.3. *EBMA for normally distributed outcomes*

When forecasting outcomes that are distributed according to the normal distribution, Raftery et al. (2005) propose approximating the conditional PDF as a normal distribution centered at a linear transformation of the individual forecast, $g_k(\mathbf{y}|\mathbf{f}_k) = N(a_{k0} + a_{k1}\mathbf{f}_k, \sigma^2)$. Using (4) above, the EBMA PDF is then

$$p(\mathbf{y}|\mathbf{f}_1, \ldots, \mathbf{f}_K) = \sum_{k=1}^{K} w_k N(a_{k0} + a_{k1}\mathbf{f}_k, \sigma^2). \tag{6}$$

3.4. *The dichotomous outcome model*

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 et al. (2010) to deal appropriately with binary outcomes.[4]

    We follow Sloughter et al. (2007) and Hamill et al. (2004) in using logistic regression after a power transformation of the forecast to reduce prediction bias. For notational ease, we assume that $\mathbf{f}_k$ is the forecast after the adjustment for bias reduction. Therefore, let $\mathbf{f}'_k \in [0, 1]$ be the forecast on the predicted probability scale and

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

where $I[.]$ is the general indicator function. Hamill et al. (2004) recommend setting $b = 4$, while Sloughter et al. (2007) use $b = 3$. We found that $b = 4$ works best in the examples below, but other analysts may try alternative specifications. This transformation dampens the effect of extreme observations and reduces over-fitting.

    The logistic model for the outcome variables is

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

---

[4]The method for dealing with binary outcomes is implicit in Sloughter et al. (2007) and Sloughter et al. (2010), which assume a discrete-continuous distribution for outcomes that include 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 et al. 2010; Muhlbaier and Polikar 2007) has recently been extended for direct application to binary outcomes (e.g., McCormick et al. 2011; Tomas 2011).

The conditional PDF of some within-sample event, given the forecast f_{kst} and the assumption that k is the true model, can be written

$$g_k(y_{st}|f_{kst}) = P(y_{st} = 1|f_{kst})I[y_{st} = 1] + P(y_{st} = 0|f_{kst})I[y_{st} = 0]. \tag{9}$$

Applying this to (4), the PDF of the final EBMA model for y_{st} is

$$p(y_{st}|f_{1st}, f_{2st}, \ldots, f_{Kst}) = \sum_{k=1}^{K} w_k[P(y_{st} = 1|f_{kst})I[y_{st} = 1] \\ + P(y_{st} = 0|f_{kst})I[y_{st} = 0]]. \tag{10}$$

3.5. *Parameter estimation by maximum likelihood and EM algorithm*

Parameter estimation is conducted using only the data from the training 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 et al. 2008).

With standard independence assumptions, the log-likelihood for the model weights is

$$\ell(w_1, \ldots, w_K|a_{01}, \ldots, a_{0K}; a_{11}, \ldots, a_{1K}) = \sum_{s,t} \log p(y_{st}|f_{1st}, \ldots, f_{Kst}). \tag{11}$$

where the summation is over values of s and t that index all observations in the training time period, and $p(y_{st}|f_{1st}, \ldots, f_{Kst})$ is given by (10). 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_{kst}, which represent the posterior probability for model k for observation y_{st}. The E step involves calculating estimates for these unobserved quantities using the formula

$$\hat{z}_{kst}^{(j+1)} = \frac{\hat{w}_k^{(j)} p^{(j)}(y_{st}|f_{kst})}{\sum_{k=1}^{K} \hat{w}_k^{(j)} p^{(j)}(y_{st}|f_{kst})}, \tag{12}$$

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

It follows that $w_k^{(j)}$ is the estimate of w_k in the jth iteration and $p^{(j)}(.)$ is shown in (10). Assuming these estimates of z_{kst} are correct, it is then straightforward to derive the maximizing value for the model weights. Thus, the M step estimates these as $\hat{w}_k^{(j+1)} = \frac{1}{n}\sum_{s,t}\hat{z}_{kst}^{(j+1)}$, where n represents the number of observations in the training dataset.[5] The E and M steps are iterated until the improvement in the log-likelihood is no larger than some pre-defined tolerance.[6]

---

[5]In the case of normally distributed data, $\hat{\sigma}^{2(j+1)} = \frac{1}{n}\sum_{s,t}\sum_{k=1}^{K}\hat{z}_{kst}^{(j+1)}(y_{st} - f_{kst})^2$.

[6]Although the log-likelihood will increase after each iteration of the algorithm, convergence is only guaranteed

3.6. *Ensemble prediction*

With these parameter estimates, it is now possible to generate ensemble forecasts. If our forecasts, $\mathbf{f}_k$, are generated from a statistical model, we now generate a new prediction, f_{kst^*}, from the previously fitted models. For convenience, let $\hat{\mathbf{a}}_k \equiv (\hat{a}_{k0}, \hat{a}_{k1})$. For some dichotomous observation in country $s \in S$ in the out-of-sample period $t^* \in T^*$, we can see that

$$P(y_{st^*} = 1 | f_{1st^*}, \ldots, f_{Kst^*}; \hat{\mathbf{a}}_1, \ldots, \hat{\mathbf{a}}_K; \hat{w}_1, \ldots, \hat{w}_K) = \sum_{k=1}^{K} \hat{w}_k \text{logit}^{-1}\left(\hat{a}_{k0} + \hat{a}_{k1} f_{kst^*}\right). \quad (13)$$

4. EMPIRICAL APPLICATIONS

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 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 over the subsequent three months.[7] 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: In the remainder of this subsection, we apply EBMA to make predictions for the occurrence of insurgency in these 29 countries. We estimate three exemplar statistical models using data for the in-sample period ranging from January 1999 to December 2008 and fit an EBMA model. We then make out-of-sample forecasts for the period from January 2009 to December 2010 for 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

---

to a local maximum of the likelihood function. Convergence to the global maximum is not assured, and the model may be sensitive to initial conditions. In future research, we will explore these convergence issues more fully with special attention paid to comparison with fully Bayesian implementations. In the examples below, we begin with the assumption that all models are equally likely, $w_k = \frac{1}{K} \; \forall \; k \in [1, \ldots, K]$.

[7] 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.

Analysis Enterprises. It is specified as a simple generalized linear logistic model including 27 different independent variables.[8] 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* and *GDP growth* (both lagged three months).

- **LMER**: This is a generalized linear mixed effects model using a logistic link function and including random country-level intercepts. In addition to the variables from the GLM model, the list of covariates includes: the *executive constraint* and *competitiveness of participation* variables from the Polity IV dataset (Marshall et al. 2009), *proximity to election*,[9] and a *spatial lag* that reflects recent occurrences of insurgencies in the countries' geographic neighbors.[10]

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 in-sample time period. 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 SAE model carries the greatest weight followed by the LMER model. 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: In-sample results. 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 Pacific Rim. EBMA equals or outperforms any single model on all measures.

|	Weight	Constant	Predictor	AUC	PRE	Brier	% Correct
SAE	0.57	0.04	7.46	0.96	0.48	0.04	94.11
LMER	0.43	6.08	28.25	0.96	0.01	0.07	88.79
GLM	0.00	0.57	8.16	0.65	0.00	0.10	88.65
EBMA				0.97	0.55	0.04	94.94
n=3,480							

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 an 89% correct prediction rate. The Brier score is the average squared deviation of the predicted probability

---

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

[9]This is calculated as the number of days to the next or from the last election, whichever is closer.

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

Figure 1: Separation plots for in-sample predictions of the ICEWS data (n=3,480). 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.
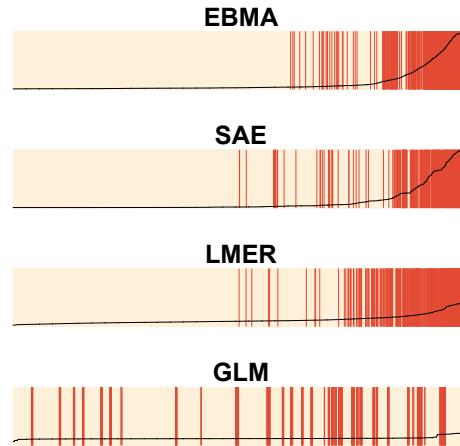
from the true event (0 or 1). Thus, a lower score corresponds to higher forecast accuracy (Brier 1950). Finally, we calculate the percentage of observations that each model would predict correctly using a 0.5 threshold on the predicted probability scale.

There are two aspects of Table 1 that are important to note. First, the EBMA model does at least as well (and usually better) than all of the component models on each our model fit statistics. The EBMA model has the highest AUC, PRE, and % correct. In addition, it is tied for the lowest Brier score with the SAE model. Second, in this example the EBMA procedure assigns probability weights to each model according to their in-sample performance. The largest model weight (0.57) is assigned to the SAE model, which appears to be the best (or tied for the best) component as measured by all of our fit statistics. Meanwhile, the smallest weight (0.00) is assigned to the rudimentary GLM model.

Figure 1 shows separation plots for the EBMA model and the individual components (Greenhill et al. 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 of the SAE model is the best component. More importantly, the overall best performance is associated with the EBMA forecast. The separation plots show that the EBMA model produces few false positives and even fewer false negatives than any of the component models.

The more interesting evaluation of the EBMA method is its out-of-sample predictive power. Table 2 shows fit statistics for the individual components as well as the EBMA forecasts for observations in the 24 months following the training period. While the EBMA model has a marginally smaller area under the ROC curve than the LMER models, it outperforms all component models on the other metrics. In particular, the EBMA model has the highest PRE at 0.18. Since it is possible to predict 89.22% of these observations correctly by forecasting no insurgency, an 18% reduction

Table 2: Out-of-sample results. The table shows fit statistics for the EBMA deterministic forecast and all component model forecasts of insurgency in 29 countries of the Pacific Rim. EBMA equals or outperforms any single model on most measures.

	AUC	PRE	Brier	% Correct
SAE	0.96	0.04	0.06	89.80
LMER	0.97	0.00	0.07	89.37
GLM	0.84	0.00	0.09	89.37
EBMA	0.96	0.18	0.05	91.24
n=696				

Figure 2: Separation plots for out-of-sample predictions of the ICEWS data (n=696). 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.
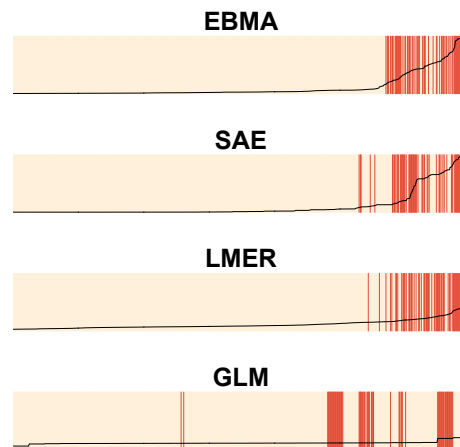
of error relative to the baseline model is quite substantial.

  Figure 2 shows the separation plots for the components as well as the EBMA forecasts for the out-of-sample data. The EBMA model again performs better than any of the individual components with very high predicted probabilities for the majority of actual events. Taking both the fit statistics and the visual evidence together, we can conclude that the EBMA model leads to a substantial improvement in out-of-sample forecasts relative to its components. This is true even in datasets with rare events and even when the individual components are already performing well.

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 forecast 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. We include six of the most widely cited presidential forecasting models.

- **Campbell**: Campbell's "Trial-Heat and Economy Model" (Campbell 2008)

- **Lewis-Beck/Tien**: Lewis-Beck and Tien's "Jobs Model Forecast" (Lewis-Beck and Tien 2008)

- **Erikson/Wlezien**: Erikson and Wlezien's "Leading Economic Indicators and Poll" forecast[11]

- **Fair**: Fair's presidential vote-share model[12]

- **Hibbs**: Hibbs' "Bread and Peace Model" (Hibbs 2000)

- **Abramowitz**: The "Time-for-Change Model" created by Abramowitz (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.[13]

Results: Rather than selecting a single training period (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.[14] We then fit the EBMA model using the components' in-sample 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 in-sample performance of each component for the 1952-1984 period to estimate model weights.[15]

    Table 3 provides exemplar results for the 2004 and 2008 elections. Table 3 shows the weights assigned to each model as well as the in-sample root mean squared error (RMSE) and mean absolute error (MAE) for the components and the EBMA forecasts. The table also shows the out-of-sample prediction errors, calculated as $y_{predicted} - y_{observed}$, for each component model and the EBMA forecast.

---

[11]We replicated Column 2 in Table 2 from Erikson and Wlezien (2008).

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

[13]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 and Douglas Hibbs .

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

[15]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.

Table 3: Prediction errors, model weights, and in-sample fit statistics for component and EBMA forecasts of the 2004 and 2008 elections. Models are sequentially fit using all prior elections. The EBMA model does better than all components on in-sample fit statistics. Although it does not necessarily make the most accurate prediction for any given year, it is less likely to make dramatic forecasting errors.

	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
Lewis-Beck/Tien	0.00	1.67	1.42	−0.41	0.17	1.61	1.33	−2.65
Erikson/Wlezien	0.00	2.67	2.06	4.76	0.17	2.81	2.18	−0.14
Fair	0.48	2.07	1.47	4.82	0.00	2.22	1.80	−2.02
Hibbs	0.12	1.95	1.38	1.54	0.25	1.92	1.38	−1.39
Abramowitz	0.00	1.50	1.18	2.20	0.06	1.53	1.26	−2.37
EBMA		1.29	1.01	2.08		1.30	1.01	−0.53

The example results in Table 3 illustrate three important points. First, the EBMA model again does better than any individual component on in-sample measures of model fit (i.e., RMSE and MAE). Second, these results demonstrate that EBMA is not guaranteed to generate the most accurate prediction for any single observation. Thus, in each year some component models come closer to predicting the actual outcome. However, the EBMA forecasts will very rarely provide egregiously wrong predictions (e.g., the Campbell model in 2008 and the Fair model in 2004) since it borrows predictions 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 it is clear that there is not as clean a relationship between in-sample model performance and model weights as in the insurgency example. For instance, the weight for the Abramowitz model in 2008 is 0.06 even though it has the lowest RMSE and MAE of any component. The diminished relationship between in-sample performance and weight is a result of high in-sample correlations between forecasts.[16] 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.

With the 2004 and 2008 examples in mind, we now turn to the relative out-of-sample performance of the EBMA and component forecasts across the entire 1976-2008 period. Table 4 shows the out-of-sample RMSE and MAE statistics as well as the percentage of observations that fall

---

[16]The correlation matrix between fitted-values of the model for the 1952-2008 period is:

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

within the 67% and 95% predictive intervals for each. 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).

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

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

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% an 67% predictive intervals are nearly perfectly calibrated.

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 of their respective 67% predictive intervals. Poorly calibrated models will tend to be 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 outcomes. 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.

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. Specifically, the models at the bottom of each plot generate relatively wide predictive intervals.
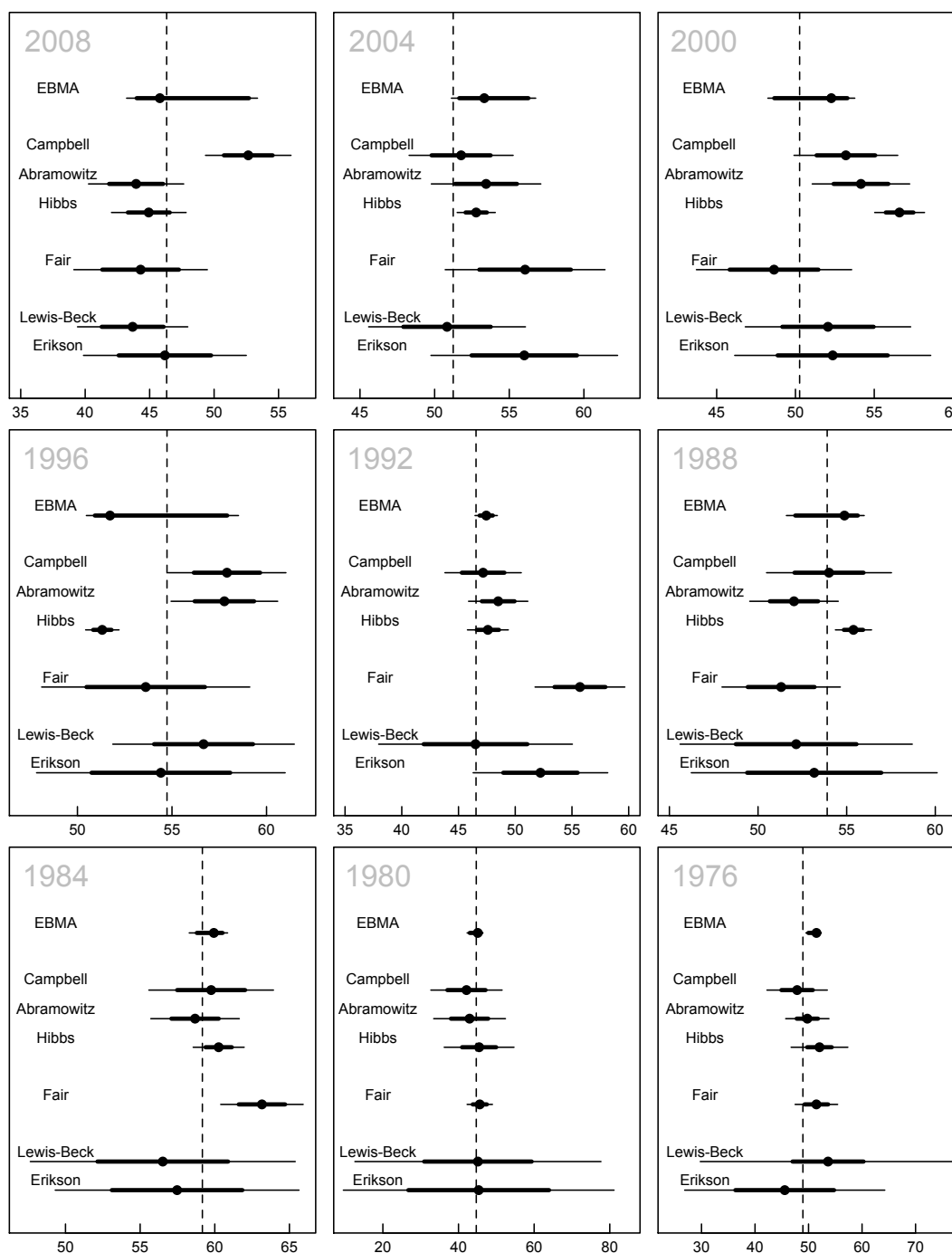
<div style="text-align:center">4.3. *Application to the Supreme Court Forecasting Project*</div>

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).[17] This example highlights the ability of EBMA to handle forecasts generated by classification trees, subject experts, and other sources.

Throughout 2002-2003, a research team consisting of Andrew Martin, Kevin Quinn, Theodore Ruger, and Pauline Kim (henceforward MQRK) generated two sets of forecasts for every pending case. First, using data about case characteristics and justices' past voting patterns, MQRK developed classification trees to generate a binary forecasts for the expected vote of each justice on each case (voting to affirm the lower court opinion is coded as a 1). Second, MQRK recruited a team of 83 legal experts to make forecasts on particular cases in their specialty area. The list 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 statistical model makes predictions for all 67 cases included in the MQRK analysis. Thus, we include the 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 judge, 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)[18] and made EBMA forecasts for the remaining 296 cases with 2002 docket numbers.

Table 5 shows the component weights for the two forecasts and the out-of-sample fit statistics for the MQRK classification trees, subject experts, and EBMA forecasts. 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.[19]

Table 5: Out-of-sample 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					

There is a long-standing debate in may circles of the relative strengths and weaknesses of

---

[17]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`.

[18]The in-sample results are available upon request.

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

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. However, 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 et al. 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.

A related issue is that, as currently constructed, EBMA makes no explicit adjustment for model complexity. That is, model weights are based solely on the components goodness-of-fit with no effort to adjust for their generalizability. This can lead to excessive weighting of complex and over-fit models. Since component forecasts may be agent-based models, stochastic simulations, multi-level models, and the like, it is necessary to go beyond merely penalizing for the number of parameters (e.g., AIC). Complexity measures must take into account functional form and other concerns. As part of continued research, we plan to incorporate several proposed methods for penalizing complexity into the EBMA method (c.f., Pitt et al. 2002; Pitt and Myung 2002; Spiegelhalter et al. 2002).

However, the EBMA method as is it 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 the 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.

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 accu-

racy 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 shown how the method can be adjusted to work for dichotomous dependent variables. The EBMA model developed here, based on previous work Sloughter et al. (2007) and Sloughter et al. (2010), is thus applicable to a large fraction of research in political science, and is particularly suited for the field of international relations.

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.[20]

References

Abramowitz, A. I. (2008). Forecasting the 2008 presidential election with the time-for-change model. *PS: Political Science & Politics 41*(4), 691–695.

Andriole, S. J. and R. A. Young (1977). Toward the development of an integrated crisis warning system. *International Studies Quarterly 21*(1), 107–150.

Armstrong, J. S. (2001). Combining forecasts. In J. S. Armstrong (Ed.), *Principles of Forecasting: A Handbook for Researchers and Practitioners*. Norwell, MA: Kluwer Academic Publishers.

Ascher, W. (1979). *Forecasting, an Appraisal for Policy-Makers and Planners*. Baltimore, MD: Johns Hopkins University Press.

Bartels, L. M. (1997). Specification uncertainty and model averaging. *American Journal of Political Science 41*(2), 641–674.

Bartels, L. M. and J. Zaller (2001). Presidential vote models: A recount. *PS: Political Science and Politics 34*(1), 9–20.

Bates, J. and C. Granger (1969). The combination of forecasts. *Operations Research 20*(4), 451–468.

Bennett, D. S. and A. C. Stam (2009). Revisiting predictions of war duration. *Conflict Management and Peace Science 26*(3), 256–267.

Berrocal, V. J., A. E. Raftery, T. Gneiting, and R. C. Steed (2010). Probabilistic weather forecasting for winter road maintenance. *Journal of the American Statistical Association 105*(490), 522–537.

---

[20]All software and data used to generate the results in this paper will be made available to the public in the authors' dataverse upon publication at `thedata.org`.

Billio, M., R. Casarin, F. Ravazzolo, and H. 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, M., R. Casarin, F. Ravazzolo, and H. 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, P. T., M. Colaresi, and J. R. Freeman (2008). The dynamics of reciprocity, accountability, and credibility. *The Journal of Conflict Resolution 52*(3), 343–374.

Brandt, P. T., J. R. Freeman, and P. A. Schrodt (2011). Real time, time series forecasting of inter- and intra-state political conflict. *Conflict Management and Peace Science 28*(1), 41–64.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review 78*(1), 1–3.

Brock, W. A., S. N. Durlauf, and K. D. West (2007). Model uncertainty and policy evaluation: Some theory and empirics. *Journal of Econometrics 136*(2), 629–664.

Brown, L. B. and H. W. Chappell (1999). Forecasting presidential elections using history and polls. *International Journal of Forecasting 15*(2), 127–135.

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

Bueno de Mesquita, B. (2011). A new model for predicting policy choices: Preliminary tests. *Conflict Management and Peace Science 28*(1), 65–85.

Campbell, J. E. (1992). Forecasting the presidential vote in the states. *American Journal of Political Science 36*(2), 386–407.

Campbell, J. 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, J. E. and K. A. Wink (1990). Trial-heat forecasts of the presidential vote. *American Politics Research 18*(3), 251.

Chmielecki, R. M. and A. E. Raftery (2010). Probabilistic visibility forecasting using Bayesian model averaging. *Monthly Weather Review 139*, 1626–1636.

Choucri, N. and T. W. Robinson (Eds.) (1978). *Forecasting in International Relations: Theory, Methods, Problems, Prospects.* San Francisco: W.H. Freeman.

Clyde, M. (2003). Model averaging. In S. J. Press (Ed.), *Subjective and Objective Bayesian Statistics: Principles, Models and Applications*, pp. 320–335. Hoboken, NJ: Wiley-Interscience.

Clyde, M. and E. I. George (2004). Model uncertainty. *Statistical Science 19*(1), 81–94.

Cuzàn, A. G. and C. M. Bundrick (2008). Forecasting the 2008 presidential election: A challenge for the fiscal model. *PS: Political Science & Politics 41*(4), 717–722.

Davies, J. L. and T. R. Gurr (1998). *Preventive Measures: Building Risk Assessment and Crisis Early Warning Systems*. Lanham, Md: Rowman & Littlefield Publishers.

de Marchi, S., C. Gelpi, and J. D. Grynaviski (2004). Untangling neural nets. *American Political Science Review 98*(2), 371–378.

Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society. Series B (Methodological) 57*(1), 45–97.

Enders, W. and T. M. Sandler (2005). After 9/11: Is it all different now? *The Journal of Conflict Resolution 49*(2), 259–277.

Erikson, R. S. and C. Wlezien (2008). Leading economic indicators, the polls, and the presidential vote. *PS: Political Science & Politics 41*(4), 703–707.

Fair, R. C. (1978). The effect of economic events on votes for president. *The Review of Economics and Statistics 60*(2), 159–173.

Fair, R. 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, R. 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, J. D. and D. D. Laitin (2003). Ethnicity, insurgency and civil war. *American Political Science Review 97*(1), 75–90.

Feder, S. A. (2002). Forecasting for policy making in the post-cold war period. *Annual Review of Political Science 5*, 111–125.

Feldkircher, M. (2011). Forecast combination and Bayesian model averaging: A prior sensitivity analysis. *Journal of Forecasting*, in press.

Fraley, C., A. E. Raftery, and T. Gneiting (2010). Calibrating multimodel forecast ensembles with exchangeable and missing members using Bayesian model averaging. *Monthly Weather Review 138*(1), 190–202.

Fraley, C., A. E. Raftery, T. Gneiting, J. M. Sloughter, and V. J. Berrocal (2011). Probabilistic weather forecasting in R. *R Journal*, in press.

Fraley, C., A. E. Raftery, J. M. Sloughter, T. Gneiting, U. of Washington with contributions from Bobby Yuen, and M. Polokowski (2010). *ensembleBMA: Probabilistic Forecasting using Ensembles and Bayesian Model Averaging*. R package version 4.5.

Freeman, J. R. and B. L. Job (1979). Scientific forecasts in international relations: Problems of definition and epistemology. *International Studies Quarterly 23*(1), 113–143.

Geer, J. and R. R. Lau (2006). Filling in the blanks: A new method for estimating campaign effects. *British Journal of Political Science 36*(2), 269–290.

Gill, J. (2004). Introduction to the special issue. *Political Analysis 12*(4), 647–674.

Gleditsch, K. S. and M. D. Ward (2010). Contentious issues and forecasting interstate disputes. Presented to the 2010 Annual Meeting of the International Studies Association, New Orleans, LA.

Gneiting, T. and T. L. Thorarinsdottir (2010). Predicting inflation: Professional experts versus no-change forecasts. Working Paper. `http://arxiv.org/abs/1010.2318v1` (accessed June 15, 2011).

Graefe, A., A. G. Cuzan, R. J. Jones, and J. S. 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, B., M. D. Ward, and A. Sacks (2011). The separation plot: A new visual method for evaluating the fit of binary data. *American Journal of Political Science*, in press.

Gurr, T. R. and M. I. Lichbach (1986). Forecasting internal conflict: A competitive evaluation of empirical theories. *Comparative Political Studies 19*(3), 3–38.

Hamill, T. S., J. S. Whitaker, and X. Wei (2004). Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts. *Monthly Weather Review 132*(6), 1434 – 1447.

Hibbs, D. (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, D. A. (2000). Bread and peace voting in U.S. presidential elections. *Public Choice 104*(1), 149–180.

Hildebrand, D. K., J. D. Laing, and H. Rosenthal (1976). Prediction analysis in political research. *The American Political Science Review 70*(2), 509–535.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky (1999). Bayesian model averaging: A tutorial. *Statistical Science 14*(4), 382–417.

Holbrook, T. M. (2008). Incumbency, national conditions, and the 2008 presidential election. *PS: Political Science & Politics 41*(4), 709–712.

Huisman, J., L. Breuer, H. Bormann, A. Bronstert, B. Croke, H.-G. Frede, T. Gräff, L. Hubrechts, A. Jakeman, G. Kite, J. Lanini, G. Leavesley, D. Lettenmaier, G. Lindström, J. Seibert, M. Siva-palan, N. 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, K. and G. King (2004). Did illegal overseas absentee ballots decide the 2000 US presidential election? *Perspectives on Politics 2*(3), 537–549.

Jerome, B., V. Jerome, and M. S. Lewis-Beck (1999). Polls fail in France: Forecasts of the 1997 legislative election. *International Journal of Forecasting 15*(2), 163–174.

King, G. and L. Zeng (2001). Improving forecasts of state failure. *World Politics 53*(4), 623–658.

Koop, G. and D. 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, G. A. (1997). Voters, information heterogeneity, and the dynamics of aggregate economic expectations. *American Journal of Political Science 41*(4), 1170–1200.

Leblang, D. and S. Satyanath (2006). Institutions, expectations, and currency crises. *International Organization 60*(1), 245–262.

Lewis-Beck, M. S. (2005). Election forecasting: Principles and practice. *The British Journal of Politics & International Relations 7*(2), 145–164.

Lewis-Beck, M. S. and C. Tien (2008). The job of president and the jobs model forecast: Obama for '08? *PS: Political Science & Politics 41*(4), 687–690.

Lockerbie, B. (2008). Election forecasting: The future of the presidency and the house. *PS: Political Science & Politics 41*(4), 713–716.

Madigan, D. and A. E. Raftery (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association 89*(428), 1535–1546.

Marshall, M. G., K. Jaggers, and T. R. Gurr (2009). Polity IV project: Political regime characteristics and transition 1800-2007. CIDCM: University of Maryland, MD.

Martin, A. D., K. M. Quinn, T. W. Ruger, and P. T. Kim (2004). Competing approaches to predicting supreme court decision making. *Perspectives on Politics 2*(2), 761–767.

McCandless, T. C., S. E. Haupt, and G. S. Young (2011). The effects of imputing missing data on ensemble temperature forecasts. *Journal of Computers 6*(2), 162–171.

McCormick, T. H., A. E. Raftery, D. Madigan, and R. 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, S.-K. and A. 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, S.-K., D. Simonis, and A. 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, J. M. and B. Nyhan (2010). Bayesian model averaging: Theoretical developments and practical applications. *Political Analysis 18*(2), 245–270.

Muhlbaier, M. D. and R. Polikar (2007). An ensemble approach for incremental learning in non-stationary environments. *Multiple Classifier Systems 4472*, 490–500.

Norpoth, H. (2008). On the razor's edge: The forecast of the primary model. *PS: Political Science & Politics 41*(4), 683–686.

O'Brien, S. 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, S. P. (2010). Crisis early warning and decision support: Contemporary approaches and thoughts on future research. *International Studies Review 12*(1), 87–104.

Pevehouse, J. C. and J. S. Goldstein (1999). Serbian compliance or defiance in Kosovo? Statistical analysis and real-time predictions. *The Journal of Conflict Resolution 43*(4), 538–546.

Pitt, M. A. and I. J. Myung (2002). When a good fit can be bad. *TRENDS in Cognitive Sciences 6*(10), 421–425.

Pitt, M. A., I. J. Myung, and S. Zhang (2002). Toward a method of selecting among computational models of cognition. *Psychological Review 109*(3), 472–491.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology 25*(1), 111–163.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski (2005). Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review 133*(5), 1155–1174.

Raftery, A. E., M. Kárnỳ, and P. Ettler (2010). Online prediction under model uncertainty via dynamic model averaging: Application to a cold rolling mill. *Technometrics 52*(1), 52–66.

Raftery, A. E. and Y. Zheng (2003). Long-run performance of Bayesian model averaging. *Journal of the American Statistical Association 98*(464), 931–938.

Ruger, T. W., P. T. Kim, A. D. Martin, and K. 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, G., N. P. Gleditsch, and S. Carey (2011). Forecasting in international relations: One quest, three approaches. *Conflict Management and Peace Science 28*(1), 5–14.

Schrodt, P. A. and D. 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.

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

Sloughter, J. M., T. Gneiting, and A. E. Raftery (2010). Probabilistic wind speed forecasting using ensembles and Bayesian model averaging. *Journal of the American Statistical Association 105*(489), 25–35.

Sloughter, J. M., A. E. Raftery, T. Gneiting, and C. Fraley (2007). Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Monthly Weather Review 135*(9), 3209–3220.

Smith, R. L., C. Tebaldi, D. Nychka, and L. O. Mearns (2009). Bayesian modeling of uncertainty in ensembles of climate models. *Journal of the American Statistical Association 104*(485), 97–116.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. Van der Linde (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society. Series B (Statistical Methodology) 64*(4), 583–639.

Tomas, A. (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, J. E. (1980). Scientific prediction versus crystal ball gazing: Can the unknown be known? *International Studies Quarterly 24*(3), 450–454.

Vrugt, J. A., M. P. Clark, C. G. Diks, Q. Duan, and B. A. Robinson (2006). Multi-objective calibration of forecast ensembles using Bayesian model averaging. *Geophysical Research Letters 33*, L19817.

Vrugt, J. A., C. G. Diks, and M. P. Clark (2008). Ensemble Bayesian model averaging using Markov chain Monte Carlo sampling. *Environmental Fluid Mechanics 8*(5), 579–595.

Ward, M. D., B. D. Greenhill, and K. M. Bakke (2010). The perils of policy by p-value: Predicting civil conflict. *Journal of Peace Research 47*(4), 363–375.

Ward, M. D., R. M. Siverson, and X. Cao (2007). Disputes, democracies, and dependencies: A re-examination of the Kantian peace. *American Journal of Political Science 51*(3), 583–601.

Whiteley, P. F. (2005). Forecasting seats from votes in British general elections. *The British Journal of Politics & International Relations 7*(2), 165–173.

Wright, J. H. (2008). Bayesian model averaging and exchange rate forecasts. *Journal of Econometrics 146*(2), 329–341.

Wright, J. H. (2009). Forecasting US inflation by Bayesian model averaging. *Journal of Forecasting 28*(2), 131–144.

Zhang, X., R. Srinivasan, and D. Bosch (2009). Calibration and uncertainty analysis of the SWAT model using genetic algorithms and Bayesian model averaging. *Journal of Hydrology 374*(3-4), 307–317.