PROJECT DESCRIPTION

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 results from the fact that it is 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 in 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.

We propose to extend a promising statistical method – ensemble Bayesian model averaging (EBMA) – and to develop software that will aid scholars across disciplines to make more accurate forecasts employing realistic assessments of the uncertainty, calibration, and sharpness of their predictions. This project builds on work in the fields of meteorology and statistics to (1) extend the method for application to a wider array of outcomes (e.g., binary data), (2) provide freely available software that implements both maximum likelihood and Bayesian estimation techniques, and (3) publish papers that provide accessible explanations of the method and its social science applications.

1. DYNAMIC FORECASTING IN POLITICAL SCIENCE

Although forecasting remains a rare exercise in political science, there is 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 other cases, 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 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 several 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*

[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).

on prediction in the field of IR typifies this growing emphasis on forecasting (c.f., Schneider, Gleditsch and Carey 2011; Bueno de Mesquita 2011; Brandt, Freeman and Schrodt 2011*b*).

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 forecast 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). Lewis-Beck (2005) provides a more in-depth discussion of election forecasting in a comparative context. More recently, Brandt, Freeman and Schrodt (2011*a*) have provided a thorough survey of how forecasts have been compared in political science and economics, with a focus on strategies for more meticulous comparison of the accuracy of forecasts.

While efforts to predict future outcomes remain uncommon, research that combines multiple forecasts is nearly non-existent. 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.

EBMA is a statistical approach that promises to improve prediction of outcomes determined by complex social processes and build upon the increased interest in generating political forecasts. 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. 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.

Ensemble methods have been shown to significantly reduce prediction error in two important ways. First, 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).[2] Combining forecasts not only reduces reliance on single data sources and methodologies, but generally allows for the incorporation of more information than any one theoretical or statistical model is likely to include in isolation.

2. 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, standalone models tend to be more sensitive to unusual events or particular data issues than ensemble methods.

The research proposed here will aid the newfound emphasis on prediction by 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

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

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 a model, but rather to incorporate the insights and knowledge implicit in various forecasting efforts via statistical post-processing.

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). 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.

2.1. **Mathematical foundation.** Assume we have some quantity of interest in the future to forecast, $\mathbf{y}^*$, based on 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

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

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

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

2.2. **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^*.

[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 recalculated 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.

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

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

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

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

2.3. **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

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

2.4. **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, Gneiting and Raftery (2010) to deal appropriately with binary outcomes.

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. 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, Whitaker and Wei (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

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

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

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

2.5. **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 with a constant and the covariate $\mathbf{f}_k$.

The difficulty is in estimating the weighting parameters, $w_k \ \forall \ k \in [1, 2, \ldots, K]$. One approach we propose to implement with NSF support is to place priors on all parameters and conduct a fully Bayesian analysis with Markov chain Monte Carlo (MCMC) techniques (c.f. Vrugt, Diks and Clark 2008). For the moment, however, we have followed Raftery et al. (2005) and Sloughter et al. (2007) in using maximum likelihood methods.

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. The E and M steps are iterated until the improvement in the log-likelihood is no larger than some predefined tolerance.[4]

[4] Although the log-likelihood will increase after each iteration of the algorithm, convergence is only guaranteed 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. As part of our proposed research, we will explore these convergence issues more fully and give special attention to comparison with fully Bayesian implementations. In the examples below, we begin with the assumption that all models are equally likely, $w_k = \frac{1}{K} \ \forall \ k \in [1, \ldots, K]$.

2.6. **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

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

3. Empirical applications

3.1. **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 predictions about expected crisis events over the subsequent three months.[5] For purposes of demonstration, we focus only on predicting violent insurgency.

The bulk of the data for the ICEWS project is gleaned from natural language processing of a continuously updated harvest of news. 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 from the Polity and World Bank datasets, information about election cycles (if any), events in neighboring countries, and the length of shared borders.

3.1.1. *Component Models.* 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 Analysis Enterprises. It is specified as a simple logistic model including 27 different independent variables.[6] All of the variables are taken from the ICEWS event-stream data.
- **GLM**: For the purposes of demonstrating the performance of EBMA, 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 includes random country-level intercepts as well as random country-level coefficients for *per capita GDP*. The list of covariates includes: *population size*, the *executive constraint* and *competitiveness of participation* variables from the Polity IV dataset (Marshall, Jaggers and Gurr 2009), *proximity to election*,[7] and a *spatial lag* that reflects recent occurrences of insurgencies in the countries' geographic neighbors.[8]

[5] 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.

[6] See `strategicanalysisenterprises.com` for more details.

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

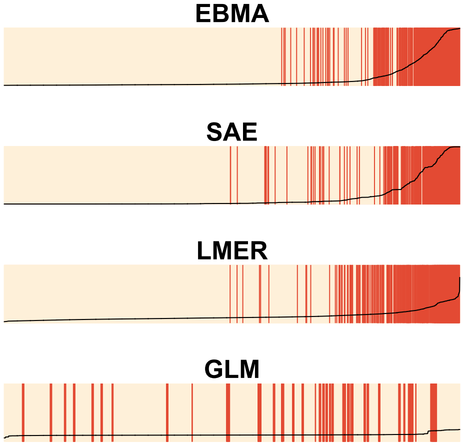
[8] Geographical proximity is measured in terms of the length of the shared border between the two countries.

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

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.



3.1.2. *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 refers 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 they evaluate forecasts in a way that is less dependent on cutoff points. A value of 1 means that all observations are predicted correctly at all possible cutoff points (King and Zeng 2001).

We compare the models using three additional metrics.[9] 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 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.

[9] 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. As we discuss below, we intend to explore the relative merits of the various model comparison techniques as part of this project. In any case, 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

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 its in-sample performance. The largest weight (0.57) is assigned to the SAE model, which appears to be the strongest component as measured by all the fit statistics. Meanwhile, the smallest weight (0.00) is assigned to the rudimentary GLM model.

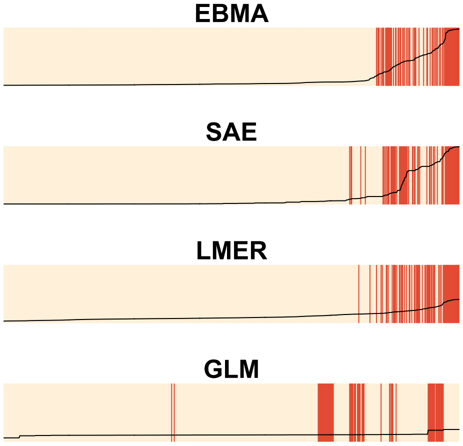
Figure 1 shows separation plots for the EBMA model and the individual components (Greenhill, Ward and Sacks 2011). In each plot, the observations are ordered from left to right by increasing predicted probabilities of insurgency (as predicted by the particular model). The black line corresponds to the 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 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 it produces few false positives and even fewer false negatives than any of the component models.

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.



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 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 performs better than any of the individual components, assigning high predicted probabilities for most observed insurgencies. 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.

3.2. **Application to U.S. 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 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). Earlier, in 1999, an entire issue of the *International Journal of Forecasting* was dedicated to the prediction of presidential elections (Brown and Chappell 1999). This topic has also drawn the attention of economists seeking to understand the relationship between economic fundamentals and political outcomes. Two prominent examples include work by Fair (2010) and Hibbs (2000).

3.2.1. *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's (2008) "Trial-Heat and Economy Model," Lewis-Beck and Tien's (2008) "Jobs Model Forecast," Erikson and Wlezien's (2008) "Leading Economic Indicators and Poll" forecast, Fair's (2010) presidential vote-share model, Hibbs' (2000) "Bread and Peace Model," and 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.[10]

3.2.2. *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.[11] 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.[12]

Results for the individual elections (not shown here) illustrate three 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, EBMA is not guaranteed to generate the most accurate prediction for any single observation, and in each year some component models come closer to predicting the actual outcome. However, the EBMA model does not provide egregiously incorrect predictions since it borrows predictions from multiple components. Moreover, as we show below, in the aggregate the EBMA model tends to provide the best forecast. Third, there is not always a crisp relationship between in-sample model performance and model weights. For instance, the weight for the Abramowitz model in 2008 is a modest 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.

We now turn to the relative out-of-sample performance of the EBMA and component forecasts across the entire 1976-2008 period. Table 3 presents the out-of-sample statistics as well as the

[10] We replicated Column 2 in Table 2 in Erikson and Wlezien (2008) and Equation 1 in Fair (2010). 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.

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

[12] 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.

percentage of observations that fall within the 67% and 95% predictive intervals for each. The main result is that the EBMA model again outperforms all components.

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.

TABLE 3. 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.

			Coverage	
	RMSE	MAE	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 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. For example, 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, Fair, and Hibbs models provide more reasonable predictive intervals, but are less accurate than EBMA.

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 in the general election. 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.

3.3. **Application to 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).[13] 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' voting patterns, MQRK developed classification trees to generate a binary forecasts for the vote of each justice on each case (voting to affirm the lower court is coded as a 1). Second, MQRK recruited a team of 83 prominent legal

[13] 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`.

experts to make forecasts on particular cases in their specialty area. 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. Specifically, we coded the expert forecast to be the mean expert prediction. We fit an EBMA model using all cases with docket numbers dating from 2001 (n=395) and made EBMA forecasts for the remaining 296 cases with 2002 docket numbers.

3.3.1. *Results.* Table 4 shows the component weights for the two forecasts and the out-of-sample fit statistics for the MQRK classification trees, the subject experts, and the EBMA forecast. Once again, the results show that the EBMA procedure outperforms all components (even when there are only two). In terms of PRE,[14] AUC, Brier scores, and correct predictions, the EBMA forecast outperforms both the statistical model and the combined subject experts.

There is a long-standing debate in many circles over the relative strengths and weaknesses of statistical models and subject experts in forecasting (e.g., Ascher 1979). Models that use quantifiable measurements and widely available data to make predictions can make egregious errors in particular cases that are 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 minutiae 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.

TABLE 4. 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 all component models.

	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

4. PROPOSED RESEARCH AND TIMELINE

Thus far, we have extended prior research to make EBMA applicable to binary and continuous outcomes in political science. As the above examples demonstrate, the method already increases the accuracy of predictions. However, we propose to expand this research in several ways.

Fully Bayesian estimation: MCMC estimation of EBMA models can more efficiently handle a wider variety of outcome distributions (Vrugt, Diks and Clark 2008). Standard Bayesian methods, such as data augmentation, will allow us to build a set of statistical results and computer algorithms appropriate for an array of assumed outcome distributions. Specifically, we plan to conduct basic research to develop a class of models and MCMC samplers to handle continuous, censored-continuous, binary, and event count data. Moreover, as the number of parameters

[14] The baseline model predicts 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.

is relatively moderate, Bayesian algorithms also promise to provide fast results while eliminating concerns about false convergence to local maxima that can occur when using EM methods.

This is not just an opportunity to improve computing efficiency. It is also an opportunity to calculate predictive density functions (rather than point estimates) for both component and ensemble forecasts. This permits a broader range of heuristic evaluations, including point, interval, and density comparisons. It will also facilitate our development and implementation of techniques to combine and average models in a manner that reflects the full degree of uncertainty associated with component forecasts.

Forecast comparison metrics: Some considerable thought has gone into how best to combine and compare predictive ensembles, yet there is no widely agreed upon set of metrics (see Brandt, Freeman and Schrodt (2011*a*) for a survey). Indeed, it has been shown that some of the most widely used metrics (e.g., RMSE) may give faulty results for some kinds of models (Geweke and Amisano 2011). While it is relatively easy to contribute to the proliferation of comparison metrics (i.e., use mean, median, mode, and geometric mean versions of standard approaches) it is rather more difficult and more important to heuristically evaluate existing and proposed model-comparison methods for a range of political science models using real-world (rather than simulated) data. One planned avenue for our research, therefore, is to: (1) implement a significant number of existing model comparison methods within our software, and (2) explore their value not in a rarefied simulation context, but in the context of real political science models making predictions about contemporary political phenomena. Our sense is that while making a wide range of metrics computationally available is probably helpful, in the end some smaller set of dominant metrics or comparison heuristics will evolve and proliferate within the discipline. Thus, we aim to provide some evidence-based evidence on this point.

Vintage data: One important practical issue relates to the vintage of the data used to generate forecasts. Most of the presidential forecasting models, for example, use data of approximately the same vintage. The predictors in these models are economic indicators or polling results measured in the months just prior to the election. However, a few models rely more heavily on data measured relatively far in advance of the election. The vintage of the data may affect the accuracy of the forecasts during calibration, and therefore partially determine the weights assigned to each model in the ensemble. Thus, a topic for future research is methods to incorporate the uncertainty introduced by information vintage into the ensemble weights.

Missing data: Thus far, we have assumed that all of the models in the ensemble are working over the same time periods, both in- and out-of-sample. Yet, for EBMA to be practicable in many political contexts, this constraint must be relaxed. Doing so will involve developing and implementing methods to handle the so-called missing data issue. As an example, prediction markets on U.S. national elections have existed only since 1988, yet the in-sample training period fo our presidential vote example begins in 1952. Moreover, any newly generated forecasting model may not necessarily cascade back over all prior time periods due to data constraints or other common issues. Finally, subject-matter experts may make predictions that are useful for a selection of observations, but not have knowledge about a complete set of events. Nonetheless, it should be possible to combine forecasts in a principled fashion that allows for missingness over part (or even most) of the in-sample training dataset. We plan to conduct basic research, simulation work, and applied case-studies to explore methods for generating ensembles in these settings.

Alternative model weights: Currently, 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 receive less posterior weight.

A related issue is that, as currently constructed, EBMA makes no 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. We plan to incorporate several proposed methods for penalizing complexity into the EBMA method (c.f., Pitt, Myung and Zhang 2002; Pitt and Myung 2002; Spiegelhalter et al. 2002).

Software: We will develop open-source software that will be publicly available. Specifically, we will produce an R package that implements both the binary outcome version we have already developed and the additional extensions just discussed. Moreover, the package will provide a more flexible interface for users interested in ensemble forecasting outside of the weather prediction community than is currently available.

Our specific goals for the software package include: S4 compliance, computationally efficient internal functions written either in C or Java, handling of multiple outcome distributional assumptions and priors with a small set of user functions, Gaussian copula techniques for handling missing data (Hoff 2007), customizable data visualizations to facilitate EBMA and component model comparisons, exemplar datasets and vignettes based on real-world applications of the method to the social and physical sciences, and built-in convergence diagnostics for all Bayesian methods compatible with the 'coda' and 'boa' packages in R. At the completion of the project, we will submit an article to the *Journal of Statistical Software* both explaining the technical details of the package and providing tutorials illustrating its available features.

Dissemination of findings: Beyond providing the software and its attendant documentation, we will disseminate the results of our research in three ways. First, we will submit articles explaining the mathematical and technical details of EBMA to journals in political methodology, economics, and applied statistics. Second, we plan to develop and publish at least two extended applications of the method to topics in political science. Our particular focus here will be: (1) improved forecasting of international crisis events; (2) election forecasting with the aim of collaborating with other scholars to generate ensemble predictions for the 2012 U.S. elections. Third, we hope to offer a workshop on forecasting political outcomes during the first day of the 2012 meeting of the Political Methods Conference (jointly hosted by UNC-Chapel Hill and Duke University).

Collaboration and work-plan: As PI, Ward will take the lead role in the project, working with Montgomery to identify achievable production goals and deadlines, guiding basic research, and disseminating findings through publications and presentations. Montgomery will take a leading role in software development and graduate student supervision. The collaborative venture will synergistically combine the resources of two programs with strong political methods programs that each place particular emphasis on Bayesian approaches. The basic timeline is:

01/2012 – 06/2012: In this stage, we will conduct basic research into MCMC estimation of EBMA and prior structures that penalize model complexity and ensure that posterior estimates reflect uncertainty in component forecasts, data vintage, and missingness.

07/2012 – 06/2013: In this stage, we will develop, test, and document the software. This will also involve gathering exemplar forecasting datasets for inclusion in package vignettes. These examples will be the basis for subsequent applied research and publication. In this period, we will also focus on evaluating the numerous forecast model comparison metrics discussed above. *07/2013 – 12/2013*: This stage will focus on: (1) preparing software and documentation for public dissemination, improving user interfaces, and increasing the computational efficiency of internal functions, and (2) revision and submission of results for publication.

5. Results from Prior NSF Support

5.1. **NSF Grants Received by Michael D. Ward During Previous 5 Years.**

(1) 0827016 ($749,970; PI's sub $150,000) AOC: The Dynamics of Secessionist Regions: Eurasian Unrecognized Quasi-States after Kosovo's Independence 10/01/08 – 09/30/11
(2) 0631531 ($400,000) Longitudinal Network Modeling of IR Data 11/15/06 – 10/31/09
(3) 0433927 ($650,000; PI's sub $150,000) The Dynamics of Civil War Outcomes: Bosnia and the North Caucasus 10/01/04 – 09/30/08
(4) 0417559 ($150,000) Network Modeling of Intl. Peace and Trade Data 10/01/04 – 09/30/06

Only one of these grants is still open (# 1), but these funds have only recently (Spring 2011) been transferred to Duke University. The most relevant grant is # 2, which is discussed below.

Summary of Findings: One of our primary findings is that standard hazard regression methods for longitudinal relational data, using variants of proportional hazards models, are unable to properly account for temporal or relational dependence in IR data. We have explored several approaches to modeling the longitudinal dependencies. One models temporal correlations directly as a network. A second uses the temporal evolution of the latent network as a means of imparting dynamic structure into the estimation of the network parameters. A final approach, which we are completing now, involves modeling a separate time-series regression for each pair of countries, but using a special array-variate hierarchical model to allow for similarity in trade patterns across groups of countries. We show that the regularized estimates from this hierarchical model outperform existing methods in out-of-sample prediction of longitudinal trade data.

Broader Impacts: We have presented the research at a number of conferences and in-departmental seminars in the fields of statistics, biostatistics, political science, and geography. We have created the first installment of a series of open-source software packages for the analysis of relational data. These packages are widely used in the social science community conducting network analyses. Finally, we constructed a database of trade and conflict that can be accessed by our software.

Within the first year we completed the following: (1) trained a graduate student in the analysis of longitudinal relational data; (2) constructed and analyzed databases on longitudinal international relations data; (3) developed new statistical methodologies for multivariate and longitudinal data; (4) conducted basic research in the area of multivariate statistical models, including methods related to copula modeling and reduced-rank matrix models; (5) conducted basic research into the analysis of array data, such as longitudinal trade data, using random-effects versions of multiway array methods such as PARAFAC, and developed an extension of the multivariate and matrix-variate normal distribution appropriate for modeling multiway array data.

Finally, two students learned how to gather and organize data, write technical documents, and perform independent research. Both students completed their Ph.D. in the summer of 2010. One student (John Ahlquist) received two national awards for his dissertation.

Publications Resulting from Award:

- 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.

- Ward, Michael D. and Peter D. Hoff. 2008. Analyzing Dependencies in Geo-Politics and Geo-Economics. In *Contributions to Conflict Management, Peace Economics, and Development, Volume 6, War, Peace and Security*, ed. Jacques Fontanel and Manas Chatterji. Bingley, UK: Emerald Publishing, pp. 133-160.
- Krivitsky, Pavel, Mark Handcock, Adrian Raftery and Peter Hoff. 2009. "Representing Degree Distributions, Homophily and Clustering in Social Networks with Latent Cluster Models." *Social Networks* 31(3):204-213.
- Hoff, Peter. 2008. "Modeling Homophily and Stochastic Equivalence in Symmetric Relational Data." *Advances in Neural Information Processing Systems* 20: 667-674.
- Hoff, Peter. 2009. "Simulation of the Matrix Bingham-von-Mises-Fisher Distribution, With Applications to Multivariate and Relational Data." *Journal of Computational and Graphical Statistics* 18(2):438–456.
- Hoff, Peter. 2009. "A Hierarchical Eigenmodel for Pooled Covariance Estimation." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71(5):971–992.
- Ward, Michael D. 2010. Statistical Analysis of International Interdependencies. In *The International Studies Encyclopedic Compendium: Scientific Studies of International Processes, Volume 10*, ed. Paul F. Diehl and James D. Morrow, pp. 6615-6628.
- Bakke, Kristin M., Xun Cao, John O'Loughlin and Michael D. Ward. 2009. "Social Distance in Bosnia and the North Caucasus Region of Russia: Inter- and Intra-ethnic Attitudes and Identities." *Nations and Nationalism* 15(2):227-253.
- Hoff, Peter. 2011. "Hierarchical multilinear models for multiway data." *Computational Statistics and Data Analysis* 55(1):530–543.
- Hoff, Peter and Xiaoyue Niu. 2010. "A covariance regression model." *Statistica Sinica* Submitted.
- Ahlquist, John S. 2008. "Building and Using Strategic Capacity: Labor Union Confederations and Economic Policy." Doctoral Dissertation, University of Washington.
- Raftery, Adrian E. and Michael D. Ward. 2010. *Statistical Methodology: Special Issue on Statistical Methods for the Social Sciences*. 7(3):173-444.

5.2. NSF Grants Received by Jacob M. Montgomery.

(1) SES-1023762 ($12,000) Doctoral Dissertation Research in Political Science: The Causes, Consequences, and Measurement of Perceived Political Control, 09/10-08/11.

Summary of Findings: I developed, evaluated, and validated a measure of perceived political control. I show that the scale is distinct from extant measures, and has superior explanatory power for predicting important behaviors in the future.

*Broader Impacts:*I have presented the research at a conference and have developed an online platform for applying computer-based psychological testing techniques for more efficient administration of large scales in online surveys.

Publications:

- Montgomery, Jacob M. 2011."An Evolutionary Theory of Democracy: Dynamic Evolutionary Models of American Party Competition with an Empirical Application to the Case of Abortion Policy from 1972-2010." Doctoral Dissertation, Duke University.