x

TODO

(1) Jacob needs to re-edit/revise the entire document.
(2) MDW bio needs to be added, and JMM Bio constructed.
(3) Jacob's NSF grant needs to be added to main doc.
(4) Review references to ensure that they include: article title, journal title, volume number, page numbers, authors, and year of publication.
(5) Budget (one for each school)
(6) Budget justification (One for each school or just one?)
(7) MDW current and Pending support
(8) Facilities, Equipment and Other Resources (one for each school)
(9) Data management plan (pg. II-19)
(10) Make all of this conform to the "collaborative proposals" guidelines reproduced belo.
(11) Get the Fastlane "cover sheet" filled out correctly (and identically in both locations).

4. Collaborative Proposals (pg. II-23

A collaborative proposal is one in which investigators from two or more organizations wish to collaborate on a unified research project. Collaborative proposals may be submitted to NSF in one of two methods: as a single proposal, in which a single award is being requested (with subawards administered by the lead organization); or by simultaneous submission of proposals from different organizations, with each organization requesting a separate award. In either case, the lead organizations proposal must contain all of the requisite sections as a single package to be provided to reviewers (that will happen automatically when procedures below are followed). All collaborative proposals must clearly describe the roles to be played by the other organizations, specify the managerial arrangements, and explain the advantages of the multi-organizational effort within the Project Description. PIs are strongly encouraged to contact the cognizant NSF Program Officer prior to submission of a collaborative proposal.

a. Submission of a collaborative proposal from one organization The single proposal method allows investigators from two or more organizations who have developed an integrated research project to submit a single, focused proposal. A single investigator bears primary responsibility for the administration of the grant and discussions with NSF, and, at the discretion of the organizations involved, investigators from any of the participating organizations may be designated as co-PIs. Please note, however, that if awarded, a single award would be made to the submitting organization, with any collaborators listed as subawards.

If a proposed subaward includes funding to support postdoctoral researchers, the mentoring activities to be provided for such individuals must be incorporated in the supplemental mentoring plan outlined in GPG Chapter II.C.2j.

By submission of the proposal, the organization has determined that the proposed activity is administratively manageable. NSF may request a revised proposal, however, if it considers that the project is so complex that it will be too difficult to review or administer as presented. (See GPG Chapter II.C.2g.(vi)(e) for additional instructions on preparation of this type of proposal.)

PROJECT SUMMARY

While generally seen as the highest validity check of statistical models and theory, political scientists rarely make scientific predictions about the future. Empirical models are seldom applied to out-of-sample data much less used to make informed predictions about future outcomes. Rather, researchers focus on developing and validating theories that explain past events.

In part, the lack of emphasis on prediction results from the fact that it is very difficult to make accurate predictions about complex social phenomena. Yet, research in political science and other fields could gain immensely in its policy relevance if predictions were both more common and more accurate. Improved forecasting of important political events would make political science research more germane to policy-makers and the general public who are less interested in explaining the past than anticipating or altering the future. From a scientific standpoint, greater attention to forecasting would facilitate more stringent validation of both theoretical and statistical models. Truly causal models will always perform better in forecasts and out-of-sample predictions than models that are solely based on prior odds. We propose research that will make it easier to find the "best" of several individual forecast models and also increase the accuracy of predictions by combining different forecast models.

We are proposing to extend a promising statistical method – ensemble Bayesian model averaging (EBMA) – and develop software that will aid researchers across disciplines to make more accurate forecasts. This project will build on work done in the fields of meteorology, fluid dynamics, 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 will include accessible explanations of the method and real-world social science applications.

In essence, EBMA makes more accurate predictions possible by pooling predictions from multiple forecast models. The weight assigned to each component is calibrated via the predictive performance of the individual models in some training period. The component models can be diverse. They need not share covariates, functional forms, or error structures. Indeed, the ensemble components may not even be statistical models, but may be predictions generated by agent based models, stochastic simulations, or subject matter experts. The advantage of the EBMA approach is that it pools information from heterogenous sources to improve the forecast accuracy of outcomes determined by complex processes. In practice, the method provides superior predictive power relative to any single component model. Further, each component model's weight can be interpreted as the posterior probability that this particular model reflects the "true" data generating process and thus facilitates meaningful model comparison.

Intellectual merit: This project will address several substantive questions in political science with a focus on the field of international relations, where policy makers have a particular demand for improved forecasting. We will also further develop statistical techniques that will improve the capabilities of researchers across disciplines to produce more accurate forecasts of future events. The methodological advancement made in this project will be available to the larger scholarly community and influence research agendas in multiple fields. EBMA has received considerable attention in the fields of statistics, meteorology, and (to a lesser extent) economics. Yet, it has not been advanced in the methodological directions we are proposing here. Additionally, our work will make EBMA easily available to a wider audience of researchers across the physical and social sciences. This is particularly important as existing research projects and software packages are narrowly tailored to the needs of weather forecasting.

Broader impacts: The principal investigators are active members of the research community at the intersection of statistics and the social sciences. Their work is widely read in political science and other disciplines. Further, at least two graduate student in political science will be included in the research and will gain experience in large-scale research projects.

PROJECT DESCRIPTION

While generally seen as the highest validity check of statistical models and theory, political scientists rarely make scientific predictions about the future. Empirical models are seldom applied to out-of-sample data much less used to make informed predictions about future outcomes. Rather, researchers focus on developing and validating theories that explain past events.

In part, the lack of emphasis on prediction results from the fact that it is very difficult to make accurate predictions about complex social phenomena. Yet, research in political science and other fields could gain immensely in its policy relevance if predictions were both more common and more accurate. Improved forecasting of important political events would make political science research more germane to policy-makers and the general public who are less interested in explaining the past than anticipating or altering the future. From a scientific standpoint, greater attention to forecasting would facilitate more stringent validation of both theoretical and statistical models. Truly causal models will always perform better in forecasts and out-of-sample predictions than models that are solely based on prior odds. We propose research that will make it easier to find the "best" of several individual forecast models and also increase the accuracy of predictions by combining different forecast models.

We are proposing to extend a promising statistical method – ensemble Bayesian model averaging (EBMA) – and develop software that will aid researchers across disciplines to make more accurate forecasts. This project will build on work done in the fields of meteorology, fluid dynamics, 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 will include accessible explanations of the method and real-world social science applications.

First, we briefly review existing political science research aimed at forecasting. We then present the EBMA method. Next we present results from empirical applications of the method to the areas of insurgency events on the Pacific Rim, U.S. presidential elections, and voting on the U.S. Supreme Court. We discuss our proposed research agenda and conclude with a discussion of results from prior NSF supported research.

1. DYNAMIC FORECASTING IN POLITICAL SCIENCE

Although forecasting is a rare exercise in political science, there are an increasing number of exceptions. In most cases, "forecasts" are conceptualized as an exercise in which the predicted values of a dependent variable are calculated based on a specific statistical model and then compared with observed values (e.g., Hildebrand, Laing and Rosenthal 1976). In many instances, this becomes nothing more than an analysis of residuals. In others, the focus is on randomly selecting subsets of the data to be excluded during model development for subsequent cross-validation.

However, there is also a more limited tradition of making political predictions about things that have not yet occurred, in the sense that the *Old Farmer's Almanac*, published continuously since the late 18th Century, predicts the weather for the coming year. An early proponent of using statistical models to make such predictions in the realm of international relations (IR) was Stephen Andriole, a research director at ARPA in the late 1970s (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), as and Vincent (1980). In addition, a few efforts began to generate forecasts of domestic conflict (e.g., Gurr and Lichbach 1986).

In the 1990s, scholars of American electoral politics began making predictions of voting patterns in presidential elections (Campbell and Wink 1990; Campbell 1992). One of the most famous early models of presidential vote share forecasts was established by economist Ray C. Fair Fair (1978).[1] As is discussed at more length below, predicting American presidential and congressional elections has subsequently developed into a regular exercise. Moreover, similar efforts have recently been made to forecast election outcomes in France (e.g., Jerome, Jerome and Lewis-Beck 1999) and the United Kingdom (e.g., Whiteley 2005).[2]

Indeed, there is evidence of increasing interest in prediction across a wide array of contexts in IR including: Krause (1997), Davies and Gurr (1998), Pevehouse and Goldstein (1999), Schrodt and Gerner (2000), King and Zeng (2001), O'Brien (2002), de Mesquita (2002), Fearon and Laitin (2003), De Marchi, Gelpi and Grynaviski (2004), Ruger et al. (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).[3] Recently, a special issue of *Conflict Management and Peace Science* on predicting conflict in the field of IR exemplifies the growing emphasis on forecasting within that subfield (c.f., Schneider, Gleditsch and Carey 2011; Mesquita 2011; Brandt, Freeman and Schrodt 2011).

While efforts to predict future outcomes are uncommon in political science, research to combine forecasts are almost non-existant. To our knowledge, the only example of researchers using multiple forecasts to make more accurate predictions is the PollyVote project (c.f. Graefe et al. 2010), which uses predictions from multiple sources to more forecast the American presidential elections.

Ensemble BMA (EBMA) lets the researcher combine the information from multiple forecast models into a single predictive model with increased accuracy. The paucity of research applying ensemble forecasting methods is unfortunate as they have been shown to significantly 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 (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 incorporation of more information than any one theoretical or statistical model is likely to include in isolation. In the next section, we briefly discuss the advantages of ensemble forecasting methods and then present the details of the EBMA method we propose to extend for social science research with support from the NSF.

2. Ensemble Bayesian model averaging

As the previous section has shown, while predictive models are still underutilized, an increasing number of scholars have developed forecasting models for specific research domains. As the number of forecasting models proliferate, however, there is a growing need to develop 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, as we show in

[1] The most recent models by Fair predicted the presidential election in 2008 Fair (2009) and were updated to include predictions for 2010 and 2012 Fair (2010).

[2] See Lewis-Beck (2005) for a more in-depth discussion of election forecasting across multiple countries.

[3] 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).

our examples below, specific models tend to be more sensitive to unusual events or particular data issues than ensemble methods.

The research proposed here will aid the newfound focus on making predictions about the future by incorporating and advancing recent work in statistics and other fields on how to meaningfully integrate multiple predictions into one improved forecast. In particular, we propose adapting an ensemble method first developed for application to the most developed and complex prediction models in existence – weather forecasting models. To generate predictive distributions of outcomes (e.g., temperature or wind speed), weather researchers apply ensemble models to forecasts generated from a variety of models (Raftery et al. 2005). State-of-the-art ensemble forecasts aggregate multiple runs of (often multiple) weather prediction models into a single forecast.

The EBMA method we plan to extend, first proposed by Raftery et al. (2005), pools across various forecasts, while meaningfully incorporating our uncertainty about which is the "best" model. No particular model or forecasting method can fully encapsulate the true data generating process for complex social phenomena. Rather, various research teams or statistical techniques will reflect different facets of reality. EBMA aims to collect *all* of the insight from multiple forecasting models in a coherent manner. The aim is not to choose some "best" model, but rather 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; Clyde and George 2004), which has received considerable attention in the field of statistics and has been shown to have good performance in a variety of settings (Raftery and Zheng 2003). BMA itself 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 intuition.** The basic BMA approach to forecasting is as follows. Assume we have some quantity of interest in the future to forecast, y^*, based on previously collected training data 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 $p(y^T|M_k)$. The outcome of interest is distributed $p(y^*|M_k)$. Applying Bayes Rule, we get that

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

and the marginal predictive PDF for y^* is

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

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 make a deterministic estimate as the weighted predictions of the components, denoted

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

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 countries S in time periods T, which we refer to as the training period. Using the same statistical model (or general technique in the case of subject-expert predictions), we then generate a forecast, f_k, for observations S in future time periods T^*.[4]

Let us assume that we have K forecasting models predicting outcome events y. Each component forecast, f_k, is associated with a component PDF, $g_k(y|f_k)$, which may be the original predictive PDF from the forecast model or a bias corrected forecast. These components are the conditional PDFs of some outcome y given the kth forecast, f_k, conditional on it being the "best" forecast in the ensemble. For example, the posterior PDF of an outcome y_{st} for some country $s \in S$ in period $t \in T$ given the forecast f_{kst} from model k is $g_k(y_{st}|f_{kst})$. This assumes that $P(M_k|y_{ST}) \equiv w_k = 1$, or that the posterior odds of model k is unity.

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

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

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 event y_{st^*} in country $s \in S$ at time $t^* \in T^*$ will then be

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

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, $g_k(y|f_k) = N(a_{k0} + a_{k1}f_k, \sigma^2)$. Using (4) above, the EBMA PDF is then

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

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 in a specified country or region. Thus, none of the distributional assumptions used in past work are appropriate in this context. Fortunately, it is possible to extend 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

[4]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.

assume that f_k is the forecast after the adjustment for bias reduction. Therefore, let $f'_k \in [0,1]$ be the forecast on the predicted probability scale and

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

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.[5]

The logistic model for the outcome variables is [6]

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

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

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

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 in the ensemble and require no data from additional components. For model k, these parameters can be estimated using the traditional ordinary least squares or logistic regression where y is the dependent variable and the covariate list includes only f_k. We emphasize that these parameters should *only* be estimated using forecasts generated from observations contained in the training data.

The difficulty is in estimating the weighting parameters, $w_k \quad \forall \quad k \quad \in [1, 2, \ldots, K]$. One approach, we propose to implement with NSF support, would be to place priors on all parameters and conduct a fully Bayesian analysis with Markov chain Monte Carlo 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 to estimate model weights.

With the standard assumptions of independence in forecast errors across countries and time-periods, the log-likelihood function for the full EBMA model (10) can be written

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

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}. An alternative interpretation is that z_{kst} represents the probability that model

[5] The purpose of this transformation is to "dampen" the effect of extreme observations in the conditional PDF, $g_k(y|f_k)$, and reduce over-fitting.

[6] Likewise, logit $P(y=0|f_k) \equiv \log \frac{P(y=0|f_k)}{P(y=1|f_k)}$.

k is the best model for predicting observation y_{st}. The E step for the full EBMA model in (10) 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 straight forward to derive the maximizing value for the model weights. Thus, the M step estimates these w_k using the current estimates of z_{kst} (in this case $\hat{z}_{kst}^{(j+1)}$) 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.[7]

The E and M steps are iterated until the improvement in the log-likelihood is no larger than some pre-defined tolerance. 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 is therefore sensitive to initial conditions.[8] In the examples below, we begin with the assumption that all models are equally important, $w_k = \frac{1}{K} \quad \forall \quad k \in [1, \ldots, K]$.

2.6. **Ensemble prediction.** With these parameter estimates completed, it is now possible to generate ensemble forecasts. If our forecasts, f_k, are generated from a statistical model, we now generate a new set of predictions 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 empirical example applies the EBMA methodology to data collected for the Integrated Crisis Early Warning Systems (ICEWS) project, sponsored by the Defense Advanced Research Projects Agency (DARPA). The basic task of the ICEWS project is train models on data on five dependent variables for 29 countries for every month from 1997 through the present and to then make accurate predictions about expected crisis events in the next three months.

For purposes of demonstration we focus on only one of these outcomes – violent insurgency.

---

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

[8] In future research, we propose to explore these convergence issues more fully with special attention paid to comparison with fully Bayesian implementations.

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.

The bulk of this data is gleaned from natural language processing of a continuously updated harvest of news stories, primarily taken from Lexus/Nexus and factiva archives. These are processed with a version of the TABARI processor for events developed by Philip Schrodt and colleagues in the context of the event data project and are described in more detail elsewhere (see `http://eventdata.psu.edu/`). These data are augmented with a variety of covariates. In particular, we use attributes (coded on a monthly or yearly basis) from the Polity, MAR, and World Bank datasets. We also include information about: the election cycles (if any) in each of the countries, events in neighboring countries, and the length of shared borders.

3.1.1. *Component Models.* In the remainder of this sub-section we apply the EBMA approach introduced above to make out-of-sample predictions for the occurrence of insurgency in these 29 countries. We fit three exemplar models using data for the in-sample period ranging from January 1999 to December 2008. We then make out-of-sample prediction for the period from January 2009 to December 2010 using the component models and the EBMA forecast. To provide variation in the complexity (as well as accuracy) of the component models, we included the following as input models:

- **SAE**: SAE is one of the models developed to produce forecasts of conflictual events as part of the ICEWS project and was designed by the SAE team. The model is a generalized linear model using 27 different independent variables (CITATION NEEDED). All of the variables are taken from basic event stream data provided through the ICEWS project.
- **LMER**: This is a generalized linear mixed effects model including five covariates and random country-level intercepts. The list of covariates includes: the *executive constrain* and *competitiveness of participation* variables from the Polity IV dataset (Marshall, Jaggers and Gurr 2009), *population size*, *proximity to election*,[10] and a *spatial lag* that reflects recent occurrences of insurgencies in the countries' geographic neighbors.[11]
- **GLM**: For the purposes of demonstrating the properties of the EBMA method, we also include a very crude logistic model. The model includes *population size* and *GDP growth*, both independent variables are lagged three months.

3.1.2. *Results.* Table 1 shows the results of the individual models, as well as the EBMA forecasts generated for the in-sample time period. The first column shows the weights that the EBMA model associates with 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 Equation 8, while predictor term corresponds to a_{k1}. AUC is the area under the Receiver-Operating Characteristic (ROC) curve. The advantage of using ROC curves in evaluating forecasts is that it produces an evaluation of the correctly predicted events at each possible cutoff point for making a positive prediction. A value of 1 of the AUC score would mean that all observations were predicted correctly at all possible cutoff points (King and Zeng 2001).

We compare the models using three other metrics. The proportional reduction in error (PRE) is the percentage increase of correctly predicted observations relative to some pre-defined base model. The base model in this case is a predicting "no insurgencies" for all observations. Insurgencies are rare events. Thus, predicting a zero for all observations leads to a 89% correct prediction rate. The Brier score is the average squared deviation of the prediction from the true

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

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

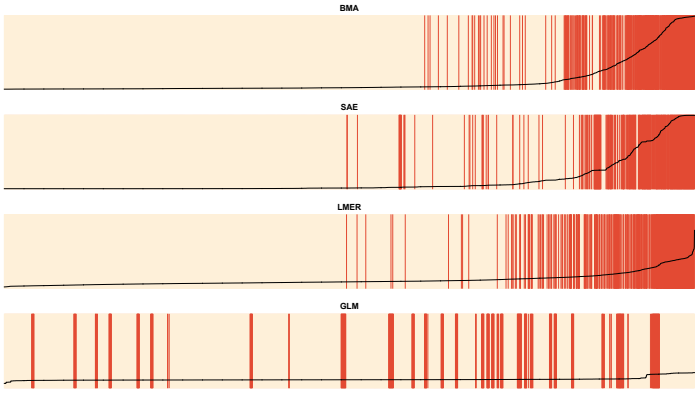
h!

TABLE 1. In-sample results: Estimated model weights, parameters, and fit statistics for EBMA deterministic forecast and all component model forecasts of insurgency in 29 countries of the Pacific rim.

	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							

event, thus a lower score corresponds to higher accuracy of forecasts (Brier 1950). Finally, we calculate the percentage of observations that each model would predict correctly using a 0.5 threshold on the predicted probability scale.

FIGURE 1. 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.



There are two important 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 highest model weight (0.57) is assigned to the SAE model, which appears to be the best (or tied for the best) on all of our fit statistics. Meanwhile, the lowest weight (0.00) is assigned to the rudimentary GLM model.

Figure 1 shows the separation plots for the EBMA model as well as all the individual components. The plots can be interpreted as follows. 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 model for each observation. Actual occurrences of insurgencies are colored red. Figure 1 shows visually that the GLM model perform very poorly, whereas of the SAE model is the best component. More importantly, the overall best performance is 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

TABLE 2. Out-of-sample results: Fit statistics for EBMA deterministic forecast and all component model forecasts of insurgency in 29 countries of the Pacific rim.

	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				

smaller area under the ROC curve than the LMER models, it outperforms all component models on any of the other fit statistics. In particular, the EBMA model has by far the highest PRE at 0.18. Since it is possible to predict 89.22% of these observations correctly by simply guessing that there is no insurgency, an 18% reduction of error relative to the baseline model is quite substantial.

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.

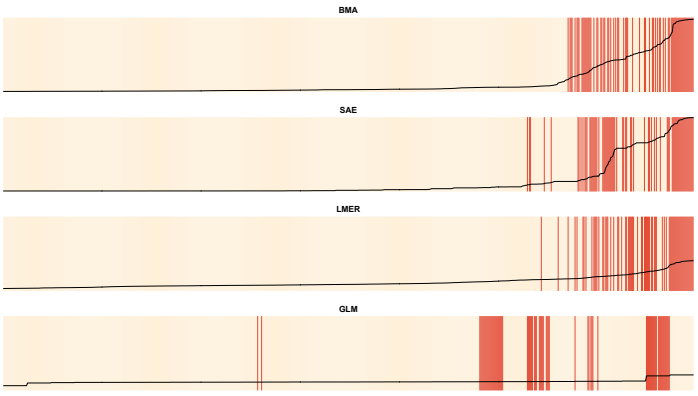


Figure 2 shows the separation plots for the components as well as the EBMA forecasts. The EBMA model performs better than any of the individual components, with very high predicted probabilities for the majority of actual events.

Taking all the evaluation statistics together, as well as the visual evidence, we can conclude that the EBMA model leads to a substantial improvement in out-of-sample forecasts relative to its components, even in datasets with rare events and when individual models are already performing very well.

3.2. **Application to US presidential election forecasts.** For the past several presidential elections, a number of research teams have developed forecasting models. For example, before the most recent 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 and evaluations were published in a subsequent issue of the journal.[12] 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).

[12]In 1999, an entire issue of the *International Journal of Forecasting* was dedicated to the task of predicting presidential elections (Brown and Chappell 1999).

3.2.1. *Component Models.* In the rest of this sub-section, 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**: Joseph Campbell's "Trial-Heat and Economy Model" (Campbell 2008).
- **Lewis-Beck**: Lewis-Beck and Tien's "Jobs Model Forecast" (Lewis-Beck and Tien 2008),
- **Erikson**: Erikson and Wlezien's "Leading Economic Indicators and Poll" forecast,[13]
- **Fair**: Fair's presidential vote-share model,[14]
- **Hibbs**: Hibbs' "Bread and Peace Model" (Hibbs 2000),
- **Abramowitz**: and, the "Time-for-Change Model" created by Abramowitz (2008).

With the exception of the Hibbs forecast, all of these models are simple linear regressions. In all cases, the dependent variable is the share of the two-party vote received by the candidate from the incumbent party.

The data to replicate the models by Abramowitz (2008); Campbell (2008); Erikson and Wlezien (2008); Lewis-Beck and Tien (2008) were provided to us in personal correspondence with the respective authors. To replicate the Hibbs (2000) and Fair (2010) model the data were downloaded from the personal websites of Ray C. Fair (Fair 2011) and Douglas Hibbs Hibbs (2011).

3.2.2. *Results.* Rather than choosing a single training period as in the insurgency analysis above, here we generate sequential predictions. For each year from 1976 to 2008 we use all available prior data to fit the component model.[15] 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 prediction). For example, to generate prediction for the 1988 election we used the in-sample performance of each component for the 1952-1984 period to estimate model weights.[16]

Table 3 provides the results of the analysis for the 2008 and 2004 elections. The table shows the out-of-sample prediction errors, calculated as $y_{predicted} - y_{observed}$, for each component model and the EBMA forecast. Table 3 also 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. A visual representation of the of the weighted component predictive PDFs and the resulting predictive EBMA PDF is shown in Figure 3.

Examining Table 3, it is clear that there is not a clean relationship between in-sample model performance and model weights as there was in the insurgency example above. For instance, the model weights for the EW model in 2008 is 0.20 even though it has nearly the highest RMSE and MAE of any component model. This is because many of the forecasts are highly correlated.

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

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

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

[16]Because of the paucity of data, we constrain the predictor, denoted a_{1k} above, to one for all models.

TABLE 3. Prediction errors, model weights, and in-sample fit statistics for component and EBMA forecasts of the 2008 and 2004 elections. Models are sequentially fit using all available data from prior elections.

	2008 Election				*2004 Election*			
	Pred. Error	Weights	RMSE	MAE	Pred. Error	Weights	RMSE	MAE
Campbell	5.92	0.34	1.69	1.36	0.11	0.29	1.75	1.45
Lewis-Beck	-1.95	0.05	1.64	1.40	0.30	0.00	1.70	1.49
Erikson	-0.11	0.20	2.81	2.18	3.70	0.21	2.72	2.06
Fair	-1.49	0.00	2.23	1.83	3.41	0.10	2.12	1.71
Hibbs	-0.93	0.34	1.93	1.33	1.50	0.40	1.95	1.32
Abramowitz	-1.88	0.08	1.57	1.30	1.92	0.00	1.55	1.25
EBMA	-0.29		1.43	1.08	1.59		1.49	1.09

[17] 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 the Campbell and Lewis-Beck/Tien forecasts, the Abramowitz forecast provides limited additional information. Thus, the bulk of the model weight within this cluster of predictions is assigned to the Campbell model, which is generally the best predictor.

With these two 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 both the RMSE and MAE statistics for each model as well as the percentage of observations that fall within the 67% and 95% predictive credible intervals for each model. The last column shows the continuous ranked probability scores (CRPS) for each probabilistic forecast. The CRPS is the integral of the Brier score at all possible threshold values (Hersbach 2000). It is a richer evaluation when considering predictive distributions rather than point estimates and scores closer to 0 indicate superior predictive power.

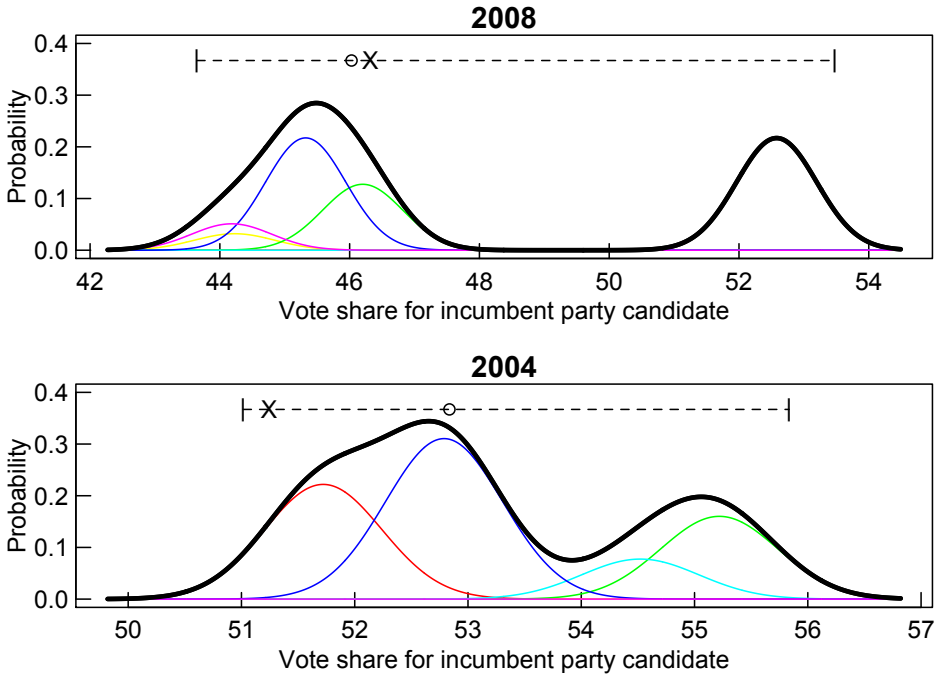
For our purposes here, the main result apparent in Table 4 is that the EBMA models again outperform all component models. The first two columns of Table 4 show that this is true in terms of predicted error (RMSE and MAE). Moreover, CRPS statistics show that this is true for the overal predictive PDFs and not just for the deterministic forecasts.

Just as important, however, the coverage results show much better calibration of EBMA forecasts relative to the component models. For instance, the observed outcome falls within the 95% predictive credible interval for the Campbell model only four out of eleven times. In general, since the EBMA forecasts is not as sensitive to particular data issues or events, it is less likely to produce wildly incorrect predictions.

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

	C	LBT	EW	F	H	A
Campbell	1.00					
Lewis-Beck	0.93	1.00				
Erikson	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

FIGURE 3. Weighted predictive PDF's for component models (colored lines) and the full EBMA predictive PDF (thick black line). The "O" shows the deterministic EBMA prediction and the "X" shows the actual observed value for the given year. The dashed line is the 95% credible interval for the EBMA predictive PDF.



An example of the kinds of problems that may arise through reliance on a single model is seen in the Campbell model. From 1952 to 2004, this model was one of the strongest performers. Indeed, this was the most accurate forecast of the 2004 election. However, as a result of the particularly late timing of the Republican Convention in the 2008 election,[18] it was the only model to forecast a victory for Republican John McCain. By relying on a wider array of data sources and methodologies, the EBMA method allows us easily "hedge" our bets and reduce the likelihood of such large misses.

3.3. **Application to Supreme Court Forecasting Project.** Our final application of EBMA is a re-analysis of forecasts made for the Supreme Court Forecasting Project (Ruger et al. 2004; Martin et al. 2004).[19] This example highlights the ability of EBMA to handle forecasts not only generated by statistical models, but also ones provided by classification trees, subject experts, or 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

[18]One of the only two variables in the Campbell model comes from polling data measured in early September. The model also relies on 2nd quarter GDP growth, which dramatically fell as the election approached due to the financial crisis. The net result of these two factors is that the Campbell model's prediction for 2008 was one of the largest mis-predictions in the dataset.

[19]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`.

TABLE 4. Fit statistics, observed coverage probabilities, and continuous ranked probability scores (CRPS) for sequentially generated predictions of presidential elections from 1976-2008.

	RMSE	MAE	67% Coverage	95% Coverage	CRPS
C	2.35	1.65	0.22	0.44	1.45
LBT	1.81	1.52	0.11	0.33	1.29
EW	2.41	1.88	0.22	0.33	1.68
F	2.51	2.16	0.00	0.11	1.91
H	2.20	1.60	0.22	0.44	1.38
A	1.84	1.62	0.11	0.22	1.39
EBMA	1.63	1.26	0.67	0.78	0.95

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. Thus, it is not possible to calibrate each expert and treat them all as individual "models." Instead, we treat all of the expert opinions as part of a single forecasting effort. We coded the expert forecast to be the mean expert forecast. 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).[20] We then made EBMA forecasts for the remaining 296 cases with 2002 docket numbers.

3.3.1. *Results.* Table 5 shows the component weights for the two forecasts and the out-of-sample fit statistics for the classification trees, subject experts, and EBMA forecasts. Figure **??** provides the separation plots. 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 out-performs both the statistical model and the combined subject experts. In addition, the EBMA forecast offers a significantly better PRE.[21]

TABLE 5. Out-of-sample results: Fit statistics for EBMA deterministic forecast and all component model forecasts of Supreme Court votes on cases in the 2002-2003 session with 2002 docket numbers.

	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 statistical models and subject experts for making predictions. Models that use quantifiable measurements and widely available (if sometimes crude) data to make comprehensive predictions can often make

[20]For reasons of space, we do not present the in-sample results here. These are available upon request.

[21]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.

egregious errors in particular cases. In some cases outcomes may be determined by forces invisible to the statistical model but obvious to experts familiar with the case. Subject experts, on the other hand, can often become too focused on minutia and miss larger (if more subtle) trends in the data that are easily recognizable 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 possible.

4. Proposed research

In our work so far, we have extended prior research to make EBMA applicable to binary as well as continuous dependent variables in political science. While, as the above examples demonstrate, the current state of the method already increases the accuracy of predictions significantly, with support from the NSF we plan to expand this research in the following ways.

(1) Extend the above into a fully Bayesian framework. Markov chain Monte Carlo estimation of EBMA models promises to more efficiently handle a wider variety of predictive distributions and will provide additional information regarding our uncertainty about model weights and within-model variances (Vrugt, Diks and Clark 2008).
(2) The current versions of EBMA estimates model weights exclusively based on the point predictions of the 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 no reflect the uncertainty associated with each model's predictions. Applying both Bayesian and bootstrap methods, we intend to incorporate the entire probability distribution of each of the individual predictive models to calibrate model weights. By using this additional information about the component models in the EBMA process, we hope to even further increase the performance of the ensemble forecasts.
(3) As a result of our research we will develop open-source software that will be available for general use. Specifically, we intend to develop an R package that will implement 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 available in the 'ensembleBMA' package. The development and free distribution of this package will allow researchers in political science and other fields to easily employ the EBMA method proposed here.
(4) As we specify in our data management plan, all the data used in this project will be publicly available for use by other researchers.

5. Results from Prior NSF Support

6. Results from Prior NSF Support

Results from Prior NSF Results during the 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 10/01/2008–09/30/2011
(2) 0631531 ($400,000) Longitudinal Network Modeling of International Relations Data 11/15/2006–10/31/2009
(3) 0433927 ($650,000; PI's sub $150,000) The Dynamics of Civil War Outcomes: Bosnia and the North Caucasus 10/01/2004–09/30/2008

(4) 0417559 ($150,000) Network Modeling of International Peace and Trade Data 10/01/2004–09/30/2006
(5) 0631531 () Longitudinal Network Modeling of International Relations Data

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 # 5 and below findings and impacts of the project are provided:

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 international relations data. We are developing new methods that are able to account for these dependencies. We have also explored several other approaches to modeling the longitudinal dependencies, one which models the temporal correlations directly, treating them directly as a network and a second which 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 have shown that the regularized estimates from the hierarchical model provide better out of sample predictive performance of longitudinal trade data than existing methods.

Broader Impacts: We have presented the research at a number of conferences and in departmental seminars, in the fields of statistics and biostatistics, as well as 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 undertaking network analysis. In addition we have constructed a database of trade and conflict that can be accessed by this suite of software.
In addition, within one year the project 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, including data on conflicts, trade, currency exchange, membership in IGOs and others
(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. Developed an extension of the multivariate and matrix-variate normal distribution appropriate for modeling multiway array data.

Furthermore, 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:

- Michael D. Ward, Randolph M. Siverson, and Xun Cao. Disputes, Democracies, and Dependencies: A Re-examination of the Kantian Peace. AMERICAN JOURNAL OF POLITICAL SCIENCE, Volume 51, no. 3 (July), pp. 583-601, 2007.

- Michael D. Ward and Peter D. Hoff. Analyzing Dependencies in Geo-Politics and Geo-Economics. In Jacques Fontanel & Manas Chatterji (editors) CONTRIBUTIONS TO CONFLICT MANAGEMENT, PEACE ECONOMICS, AND DEVELOPMENT, VOLUME 6, WAR, PEACE AND SECURITY, Emerald Publishing, pp. 133-160.
- Pavel Krivitsky, Mark Handcock, Adrian Raftery, Peter Hoff, "Representing degree distributions, homophily and clustering in social networks with latent cluster models", Social Networks, p. , vol. , (2007). Submitted,
- Hoff, Peter, "Modeling homophily and stochastic equivalence in symmetric relational data", Advances in Neural Information Processing Systems 20, p. 667, vol. , (2008). Published,
- Hoff, Peter, "Simulation of the matrix Bingham-von Mises-Fisher distribution, with applications to multivariate and relational data", Journal of Computational and Graphical Statistics, p. , vol. , (2008). Submitted,
- Hoff, Peter, "A hierarchical eigenmodel for pooled covariance estimation", Journal of the Royal Statistical Society, Series B, p. , vol. , (2008). Submitted,
- Michael D. Ward. Statistical Analysis of International Interdependencies. THE INTERNATIONAL STUDIES ENCYCLOPAEDIC COMPENDIUM: SCIENTIFIC STUDIES OF INTERNATIONAL PROCESSES, Volume 10, edited by Paul F. Diehl and James D. Morrow, pp. 6615-6628, 2010.
- Kristin M. Bakke, Xun Cao, John O'Loughlin, & Michael D. Ward. Social Distance in Bosnia and the North Caucasus Region of Russia: Inter- and intra-ethnic attitudes and identities. NATIONS AND NATIONALISM, in press, 2009, Volume 15, Issue 2, pages 227-253.
- Peter Hoff, "Hierarchical multilinear models for multiway data", Computational Statistics and Data Analysis, p. , vol. , (2010).
- Peter Hoff and Xiaoyue Niu, "A covariance regression model", Statistica Sinica, p. , vol. , (2010). Submitted, Books or Other One-time Publications
- John S. Ahlquist, "Building and Using Strategic Capacity: Labor Union Confederations and Economic Policy", (2008). Thesis, Published Bibliography: Doctoral Dissertation, University of Washington, July
- Adrian E. Raftery and Michael D. Ward STATISTICAL METHODOLOGY: SPECIAL ISSUE ON STATISTICAL METHODS FOR THE SOCIAL SCIENCES. Elsevier, Volume 8, 2010.

REFERENCES

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

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

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

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

Bartels, Larry M. and John Zaller. 2001. "Presidential Vote Models: A Recount." *PS: Political Science and Politics* 34(01):9–20.

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

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

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

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

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

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

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

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

Clyde, Merlise. 2003. Model averaging. In *Subjective and Objective Bayesian Statistics*, ed. S. James Press. 2nd ed. Hoboken, NJ: Wiley-Interscience chapter Chap. 13, pp. 320–335.

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

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

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

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

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

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

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

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

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

Fair, Ray C. 2009. "Presidential and Congressional Vote-Share Equations." *American Journal of Political Science* 53(1):55–72.

Fair, Ray C. 2010. "Presidential and Congressional Vote-Share Equations: November 2010 Update.".
URL: *http://fairmodel.econ.yale.edu/RAYFAIR/PDF/2010C.pdf*

Fair, Ray C. 2011. "Vote-Share Equations: November 2010 Update.".
URL: *http://fairmodel.econ.yale.edu/vote2012/index2.htm*

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

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

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

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

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

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

Graefe, Andreas, Aldfred G. Cuzan, Randal J. Jones and J. Scott Armstrong. 2010. "Combining Forecasts for U.S. Presidential Elections: The PollyVote.".

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

Hamill, Thomas S., Jeffrey S. Whitaker and X. Wei. 2004. "Ensemble re-forecasting: Improving medium-range forecast skill using retrospective forecasts." *Monthly Weather Review* (132):1434 – 1447.

Hersbach, Hans. 2000. "Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems." *Weather and Forecasting* 15(5):559–570.

Hibbs, Douglas. 2011. "The Bread and Peace Model Applied to the 2008 US Presidential Election.".
URL: *http://douglas-hibbs.com/Election2008/2008Election-MainPage.htm*

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

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

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

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

Imai, Kosuke and Gary King. 2004. "Did Illegal Overseas Absentee Ballots Decide the 2000 US Presidential Election?" *Perspectives on Politics* 2(03):537–549.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Raftery, Adrian E., Tilmann Gneiting, Fadoua Balabdaoui and Michael Polakowski. 2005. "Using Bayesian model averaging to calibrate forecast ensembles." *Monthly Weather Review* 133(5):1155–1174.

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

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

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

Schrodt, Philip A. and Deborah J. Gerner. 2000. "Using Cluster Analysis to Derive Early Warning Indicators for Political Change in the Middle East, 1979-1996." *American Political Science*

Review 94(4):803–818.

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

Sloughter, J. McLean, Adrian E. Raftery, TTilmann Gneiting and Cchris Fraley. 2007. "Probabilistic quantitative precipitation forecasting using Bayesian model averaging." *Monthly Weather Review* 135(9):3209–3220.

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

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

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

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

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