Say Yes to the Guess: Tailoring Elegant Ensembles on a Tight (Data) Budget*

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

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

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

February 9, 2013

*Prepared for the 2012 Annual Meeting of the American Political Science Association, August 30 - September 2, New Orleans, Louisiana. This work was partially supported by the Information Processing Technology Office of the Defense Advanced Research Projects Agency via a holding grant to the Lockheed Martin Corporation, Contract FA8650-07-C-7749. The current support is partially from the Office of Naval Research via ONR contract N00014-12-C-0066 to Lockheed Martin's Advanced Technology Laboratories.

Say Yes to the Guess:
Tailoring Elegant Ensembles on a Tight (Data) Budget

Jacob M. Montgomery, Florian M. Hollenbach, and Michael D. Ward

Abstract

We consider ensemble Bayesian model averaging (EBMA) in the context of small-n prediction tasks in the presence of a large number of component models. With a large number of observations to calibrate ensembles, relatively small numbers of component forecasts, and low rates of missingness, the standard approach to calibrating forecasting ensembles introduced by Raftery et al. (2005) performs well. However, data in the social sciences generally do not fulfill these requirements. The number of outcomes predicted tends to be small, the number of forecasting models in the literature can be large, and missing predictions for component models are neither random nor rare. In these circumstances, EBMA models may miss-weight components, undermining the advantages of the ensemble approach to prediction. In this article, we explore these issues and introduce a "wisdom of the crowds" parameter to the standard EBMA framework that improves its predictive performance. We show that this solution improves predictive accuracy of EBMA forecasts in both political and economic applications.

1 Introduction

Although accurate prediction of future events is not the primary goal for most social sciences, recent years have witnessed spreading of systematic forecasting from more traditional topics such as GDP growth and unemployment to many new domains including elections (e.g., Linzer Forthcoming), political instability (e.g., Goldstone et al. 2010), and mass killings (Ulfelder 2012). Several factors motivate this trend. To begin with, testing predictions about future events against observed outcomes is seen as a stringent validity check of statistical and theoretical models (Ward, Greenhill and Bakke 2010). In addition, forecasting of important political, economic, and social events is of great interest to policymakers and the public.

With the proliferation of forecasting efforts, however, comes a need for sensible methods to aggregate and utilize the various scholarly efforts. One attractive solutions to this problem is to combine prediction models and create an ensemble forecast. Combining forecasts reduces reliance on any single data source or methodology, and allows for the incorporation of more information than any one model can provide in isolation. Across subject domains, scholars have shown ensem-

ble predictions to be more accurate than any individual component model and less likely to make dramatically incorrect predictions (Bates and Granger 1969; Armstrong 2001; Raftery et al. 2005).

One promising approach to combining multiple forecasts is ensemble Bayesian model averaging (EBMA). This method was first proposed by Raftery et al. (2005) to combine weather forecasts and was introduced to the social sciences by Montgomery, Hollenbach and Ward (2012*b*). EBMA combines multiple forecasts using a finite mixture model that generates a weighted predictive probability density function (PDF). EBMA mixture models seek to collate the good parts of existing forecasting models, while avoiding over-fitting to past observations and over-estimating our certainty about forecasts of future events. The hope is for greater accuracy as both the knowledge and implied uncertainty of a variety of approaches are integrated into a combined predictive PDF.

In this article, we present several adjustments to the basic EBMA model as specified in Montgomery, Hollenbach and Ward (2012*b*) that can aid applied researchers to create ensemble forecasts in the presence of data-quality challenges common in real-world social science settings. Specifically, we show EBMA can be adjusted to accommodate small calibration samples, large numbers of candidate components, and missing forecasts. We propose an alteration to the basic model to hedge against the miss-weighting of components resulting from either strong or poor performance in the limited calibration period. After discussing the data-quality challenges commonly experienced in ensemble forecasting, we introduce the basic EBMA model and outline modifications to the model for small samples and missing components in Section 3. In Section 4, we demonstrate how our adjustment to the basic EBMA model improves out-of-sample forecasts in a simulation study and apply the method to predict the U.S. unemployment rate and the 2012 U.S. presidential election.

2 Ensemble prediction with sparse data and multiple forecasts

The concept of ensemble forecasting builds on the basic notion that combining multiple points of view leads to a more accurate picture of reality (c.f., Surowiecki 2004). Among the more famous

demonstrations of this phenomenon was a competition to guess the weight of an ox at the West Of England Fat Stock and Poultry Exhibition. Galton (1907) famously demonstrated that, while individual entrants were often wildly inaccurate, aggregating the "wisdom of crowds" by using the average guess resulted in a remarkably accurate estimate.[1]

In recent years, the advantages of ensembles have come to play a particularly prominent role in the machine learning and nonparametric statistics community (Hastie, Tibshirani and Friedman 2009). A wide range of approaches including neural nets, additive regression trees, and K nearest neighbors fall under the general umbrella of ensemble approaches. Of particular relevance is the success of boosting (Freund and Schapire 1997; Friedman 2001), bagging (Breiman 1996), random forests (Breiman 2001), and related techniques (e.g., Chipman, George and McCulloch 2010) to aggregate so-called "weak learners." These approaches to classification and prediction have been advertised as the "best off-the-shelf classifier[s] in the world" (Breiman 1996), and are equally powerful in prediction tasks.

While the advantages of collating information from multiple sources are manifold, it is nevertheless false to assume that more is always better (c.f., Page 2011). Not all guesses are equally informative, and naive approaches to collating forecasts risks both overvaluing wild guesses and undervaluing unusual forecasts that are nonetheless sometimes correct. The particular ensemble method we are extending is ensemble Bayesian model averaging (EBMA). First proposed by Raftery et al. (2005), EBMA pools forecasts as weighted combination of predictive probability distribution functions (PDFs). Rather that selection some "best model," EBMA collects *all* of the insights from multiple forecasting efforts in a coherent manner via statistical post processing. The weight assigned to each component forecast, reflects both its past predictive accuracy and its uniqueness (i.e., the degree to which it makes predictions different from other component models).

In recent years, variants of the EBMA method have been applied to subjects as diverse as inflation (Wright 2009; Koop and Korobilis 2009; Gneiting and Thorarinsdottir 2010), stock prices

[1]This observation is also implicit in Condorcet's jury theorem that calculates the probability of a jury reaching a correct verdict, based on independent juror opinions with little expertise, will always be greater than even the most accurate of expert judges.

(Billio et al. 2011), economic growth and policymaking (Brock, Durlauf and West 2007; Billio et al. 2010), exchange rates (Wright 2008), industrial production (Feldkircher 2011), ice formation (Berrocal et al. 2010), visibility (Chmielecki and Raftery 2010), water catchment streamflow (Huisman et al. 2009), climatology (Min and Hense 2006; Min, Simonis and Hense 2007; Smith et al. 2009), and hydrology (Zhang, Srinivasan and Bosch 2009).

While already in wide use, research to improve upon the basic EBMA model is ongoing. It has been adjusted to handle missing data (Fraley, Raftery and Gneiting 2010; McCandless, Haupt and Young 2011), incorporate spatial information (Feldman 2012) and calibrate model weights on non-likelihood criteria (e.g., Vrugt et al. 2006). Other recent innovations include Möller, Lenkoski and Thorarinsdottir (2013), who take multiple EBMA models predicting univariate outcomes to create joint predictive distributions of multiple (correlated) dependent variables, and Rings et al. (2012), who allow for time-specific variances using particle filtering methods. All in all, the promise of ensemble forecasting via EBMA has lead to multiple efforts to refine the method in fields outside of the social sciences.

In this paper, however, we focus on difficulties in calibrating accurate EBMA forecasting models in the context of data-quality challenges especially common (although not limited to) social science applications. To begin with, the amount and quality of data for calibrating ensembles is far from ideal. EBMA was first developed for use in weather forecasting where measurement of outcomes is fairly precise and data are abundant. Predicting, for instance, water surface temperatures in 200 locations across just five days provides 1,000 observations by which model weights can be calibrated. In contrast, forecasting quarterly GDP growth in the United States for five *years* provides only 20 data points.

A second, and related, issue is dimensionality. Prediction tasks often involve many forecasts predicting few, or even just one, outcome. For example, in the field of economics, a wide variety of consulting firms, banks, and international organizations provide forecast for various economic quantities, such as the unemployment, GDP growth, and inflation. Indeed, the Federal Open Market Committee (FOMC) of the U.S. Federal Reserve Board itself generates over a dozen forecasts

for key economic indicators.[2]

A final issue is the inconsistency with which forecasts are issued. Given the lengthy time periods often involved, there are likely to be many missing forecasts in any given time window. Moreover, we cannot assume that forecasts for any time period from a specific model or team are missing at random. Particularly, unsuccessful forecasts may be suppressed and some forecasting efforts are only active for short time-periods due to poor performance. In addition, forecasts tend to accumulate with more potential components being available for more proximate time periods.

While particularly egregious for specific application (c.f., presidential election forecasting), these data issues are endemic to the social sciences and are far from benign. As we demonstrate below, calibrating large ensemble models on sparse (and even incomplete) data leads to miss-specification of EBMA model weights and decreased out-of-sample predictive performance. In light of these difficulties, below we introduce several extensions to the baseline EBMA algorithm introduced in Montgomery, Hollenbach and Ward (2012*b*), and explore the effect of these modifications on the method's predictive performance.

3 EBMA for sparse data

As its name suggests, EBMA descends from the Bayesian model averaging (BMA) methodology (c.f., Madigan and Raftery 1994; Raftery 1995; Hoeting et al. 1999; Clyde 2003; Clyde and George 2004), which 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; **?**). A more detailed discussion of the basic EBMA model extended here is provided in Montgomery, Hollenbach and Ward (2012*b*).

[2]For a recent sample of these forecasts, see: `http://1.usa.gov/zjyisV`.

3.1 Baseline EBMA model

Assume the researcher is interested in predicting event $\mathbf{y}^{t^*}$ for some future time point $t^* \in T^*$. In addition, we have a number of different forecasts for this event $\mathbf{y}^t$ for a some past observations in period $t \in T$. The different predictions were generated from K forecasting models or teams, $M_1, M_2, \ldots, M_K$. For each forecast we have a prior probability distribution $M_k \sim \pi(M_k)$. The PDF for $\mathbf{y}^t$ is denoted $p(\mathbf{y}^t|M_k)$. Under this model, the predictive PDF for the quantity of interest is $p(\mathbf{y}^{t^*}|M_k)$, the conditional probability for each model is therefore $p(M_k|\mathbf{y}^t) = p(\mathbf{y}^t|M_k)\pi(M_k)/\sum_{k=1}^{K} p(\mathbf{y}^t|M_k)\pi(M_k)$, and the marginal predictive PDF is

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

The prediction via EBMA is thus a weighted average of the component PDFs. The weight for each model is based on its predictive performance on past observations in period T.

The general EBMA procedure assumes K forecasting models throughout the training (T') calibration (T) and test (T^*) periods. Each component model is fitted based on data from the training period T'. The parameter estimation based on training period T' allows us to then generate out-of-sample predicitions for each component model for the calibration period T. It is then possible to generate true ensemble forecasts ($\mathbf{f}_k^{t^*}$) for observations in the test period $t^* \in T^*$. Using three distinct time periods makes it possible to calibrate the EBMA model on the components models' out-of-sample predictive power, thus implicitly penalizing overly-complex "garbage can" models. One of the distinct advantages of EBMA is that it does not require researchers to develop metrics to penalize component forecasts for complexity or even to have access to the details of the component forecasting methods themselves.

Let $g_k(\mathbf{y}|\mathbf{f}_k^{s|t,t^*})$ represent the predictive PDF of component k, which may be the original prediction from the forecast model or some bias-corrected forecast. The EBMA PDF is a finite mixture of the K component PDFs, denoted $p(\mathbf{y}|\mathbf{f}_1^{s|t}, \ldots, \mathbf{f}_K^{s|t}) = \sum_{k=1}^{K} w_k g_k(\mathbf{y}|\mathbf{f}_k^{s|t})$, where $w_k \in [0, 1]$ are model probabilities, $p(M_k|\mathbf{y}^t)$, and $\sum_{k=1}^{K} w_k = 1$. The ensemble predictive PDF with this

notation is then $p(y|f_1^{t^*}, \ldots, f_K^{t^*}) = \sum_{k=1}^{K} w_k g_k(y|f_k^{t^*})$.[3] For the applications below, we assume $g_k(\mathbf{y}|\mathbf{f}_k^{s|t}) = N(\mathbf{f}_k^t, \sigma^2)$, where σ is a common variance component across components. Thus, the ultimate predictive distribution for some observation y^{t^*} is

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

This is a weighted mixture of K normal distributions each with means determined by $\mathbf{f}^{t^*}$ and scaled by the model weights $\mathbf{w}$.

3.2 Model estimation

Since the component model forecasts, $f_1^t, \ldots, f_k^t$, are pre-determined, the EBMA model is fully specified by estimating model weights, $w_1, \ldots, w_k$ and the common variance parameter σ^2. We estimate these using maximum likelihood methods (Raftery et al. 2005). The log likelihood function,

$$\mathcal{L}(w_1, \ldots, w_k, \sigma^2) = \sum_t \log \left(\sum_{k=1}^{K} w_k N(f_k^t, \sigma^2) \right), \tag{2}$$

cannot be maximized analytically. Instead we follow Raftery et al. (2005) and use an EM algorithm to calibrate the weights associated with the maximum likelihood. Here EBMA is expressed as a finite mixture model (McLachlan and Peel 2000; Imai and Tingley 2012). To do so the unobserved quantities z_k^t are introduced. z_k^t is the probability that observation y^t is "best" predicted by model k. These unobserved quantities are estimated (E-step) in the algorithm using the formula

$$\hat{z}_k^{(j+1)t} = \frac{\hat{w}_k^{(j)} p^{(j)}(y|f_k^t)}{\sum_{k=1}^{K} \hat{w}_k^{(j)} p^{(j)}(y|f_k^t)}, \tag{3}$$

[3]Past applications have statistically post-processed the predictions for out-of-sample bias reduction and treated these adjusted predictions as a component model. Raftery et al. (2005) propose approximating the conditional PDF as a normal distribution centered at a linear transformation of the individual forecast, $g_k(\mathbf{y}|\mathbf{f}_k^{s|t}) = N(a_{k0} + a_{k1}\mathbf{f}_k^t, \sigma^2)$. However, in the presence of sparse data, including the additional $\mathbf{a}$ parameters risks over-fitting and reduced predictive performance. We therefore use a simpler formulation.

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

Note that $w_k^{(j)}$ is the estimate of w_k in the jth iteration and $p^{(j)}(.)$ is shown in (1). Using the estimates of $z_k^{s|t}$, one can then easily calculate the maximizing value for the component weights. The M step of the algorithm therefore estimates these as

$$\hat{w}_k^{(j+1)} = \frac{1}{n}\sum_t \hat{z}_k^{(j+1)t}, \tag{4}$$

where n represents the number of observations in the calibration dataset. Finally, the variances are estimated based on

$$\hat{\sigma}^{2(j+1)} = \frac{1}{n}\sum_t\sum_{k=1}^{K} \hat{z}_k^{(j+1)t}(y - f_k^t)^2. \tag{5}$$

The E and M steps are iterated until the improvement in the log-likelihood is no larger than some pre-defined tolerance. The algorithm is started with the assumption that all component models are equally likely to be the best forecast, i.e. $w_k = \frac{1}{K}\ \forall\ k \in [1, \ldots, K]$ and $\sigma^2 = 1$.

3.3 Adjustments for sparse data

When ensembles are calibrated on very few observations, there is an increased chance that EBMA may miss-weight component models in a way that reduces out-of-sample performance due to unusually poor or strong predictive performance in the limited calibration sample. This is especially true when the short calibration period is combined with missing observations in component model predictions.[4]

To improve the performance of EBMA in the context of sparse data, we propose a "wisdom of crowds" parameter, $c \in [0, 1]$, that reflects our prior belief that all models should receive some, but not necessarily equal, weight. We rescale z_k^t to have a minimum value $\frac{c}{K}$. This states that there is, at a minimum, a $\frac{c}{K}$ probability that observation t is correctly represented by each model k. Since $\sum_{k=1}^{K} z_k^t = 1$, this implies that $z_k^t \in [\frac{c}{K}, (1-c)]$. To achieve this, we replace Equation 4 above with

[4]Adjustments to the baseline model to accomodate missing components is provided in Appendix A.

$$\hat{z}_k^{(j+1)t} = \frac{c}{K} + (1-c)\frac{\hat{w}_k^{(j)} p^{(j)}(y|f_k^t)}{\sum_{k=1}^{K} \hat{w}_k^{(j)} p^{(j)}(y|f_k^t)}. \quad (6)$$

Note that when $c = 1$, all models are considered equally informative about the outcome and $w_k = \frac{1}{K} \forall K$. Thus, we see that the arithmetic mean or median of component forecasts for time period t represents a special case of EBMA where $c = 1$.[5] Likewise, the general EBMA discussed in Montgomery, Hollenbach and Ward (2012*b*) represents a special case of this more general model where $c = 0$.

4 Simulations and Applications

The introduction of the "wisdom of crowds" parameter to the base EBMA model is designed to improve out-of-sample predictive performance in the context of data-quality challenges common to social science applications. In particular, it is designed to address poor weight calibrations that are likely when the size of the calibration sample is small, when component models with missing predictions in the calibration sample are included, and when the number of component forecasts for which weights must be estimated is large. We argue that these issues increases the miss-estimation of component model weights and decrease the predictive performance of EBMA.

To justify these claims, we present the results of a simulation study of our modified EBMA algorithm and two empirical applications of the modified method. We begin with a simulation that illustrates the reduced predictive performance of the baseline EBMA model in the circumstances described above and illustrates the improvements that result from our proposed modification. We then apply our method to, first, the prediction of the US unemployment rate, and, second, to the prediction of the 2012 US presidential election.

[5]The mean or median would be equivalent depending on if the posterior mean or median is used to make a point prediction.

Table 1: Parameters for simulation

Parameter	Meaning	Values
n_T	Sample size in calibration period T	3-15,20,25,35,45,55,65,85,100
n_{T^*}	Sample size in test period T^*	250
K	# of component forecasts	3,5,7,9,11,13,15
σ^2	Common variance component	1
α	Weight concentration parameter	$(10, 5, 3, \frac{1}{K-3})$
c	Wisdom of crowds parameter	0,0.01,0.02,0.03,0.04,0.05,0.075,0.1 0.15,0.2,0.3,0.5
M	Simulations at each setting	100

4.1 Simulation study

In this section, we conduct a simulated study of the adjusted EBMA model proposed above. These simulations serve two purposes. First, they demonstrate the challenges presented to ensemble forecasting when calibration samples are small and the number of forecasting models are large.[6] Second, it explores the extent to which our modified EBMA algorithm ameliorates these difficulties. In addition, we provide some guidance regarding the selection of c.

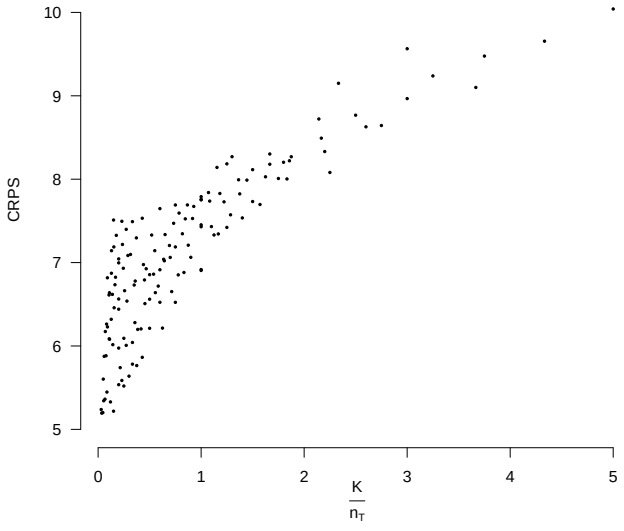
The simulations are designed to reflect the "best possible" world for the baseline EBMA model. The distribution of outcomes is drawn precisely from the mixture distribution shown in Equation (1) where $\sigma^2 = 1$ and the individual component forecasts are drawn from the multivariate normal distribution $N(\mathbf{0}_K, \mathbf{I}_K)$. Moreover, we assume that the true data generating process, both in-sample and out-of sample, involves *only* the K forecasting models which are themselves estimated with perfect precision. The "true" model weight for each simulation are drawn from a Dirichlet distribution with K categories and concentration parameter $\alpha = (10, 5, 3, \frac{1}{K-3})$ when $K > 3$ and $\alpha = (10, 5, 3)$ when $K = 3$. This ensures that the model weights always sum to 1, but that there is still some heterogeneity in the true model weights. We varied the size of the calibration sample (n_T), the number of component forecasts (K), and the wisdom of crowds paramter (c). The c parameter is used only for model estimation, and plays no role in the creation of the simulated data itself.

[6]To reduce the parameter space for these simulations, we limit ourselves here to exploring the roll of calibration sample sizes and number of component forecasts. We do not consider issues of missingness.

For each simulation, we generate component forecasts for both the calibration and test period. We fit an EBMA model as specified above to the calibration sample data only. We then generate out-of-sample predictions for the 250 observations in the test period using the fitted EBMA model and compare the forecasts to the true values from the simulated data.

We begin by examining the accuracy of the baseline EBMA ($c = 0$) predictive PDF shown in Equation 1 for different values of K (the number of components) and n_T (the calibration sample size). To evaluate the forecasts we focus here on the continuous rank probability score (CRPS) for several reasons. The CRPS has been widely used to evaluate forecasts of continuos outcomes and its many advantages as a proper scoring rule have been discussed elsewhere (Hersbach 2000; Gneiting and Raftery 2007; Gneiting, Balabdaoui and Raftery 2007; Brandt, Freeman and Schrodt 2011). One of the main advantages of the CRPS over other scoring rules is that it can be interpreted as the integral over all possible Brier scores (Brier 1950) and takes into account the uncertainty of forecasts (i.e., the predictive distributions rather than the point prediction in isolation). The CRPS ranges from 0 to 1 with smaller numbers indicating a better forecast performance.[7]

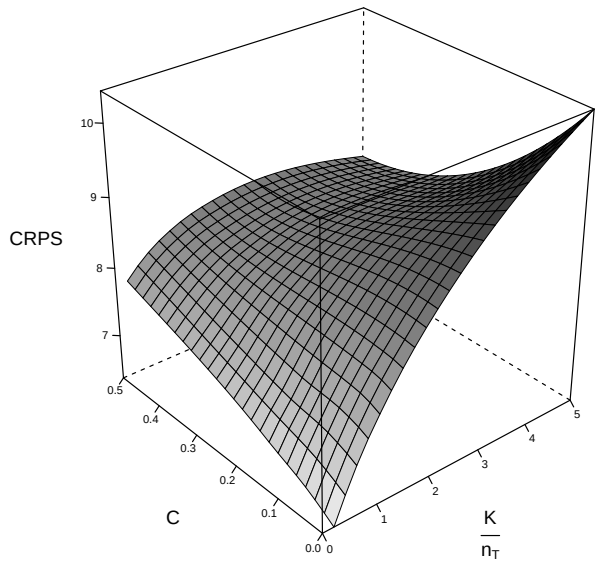
Figure 1: CRPS with varying number of models and observations in calibration period



[7]The mathematical details behind the CRPS can be found in the Appendix.

Figure 1 shows how the out-of-sample performance of the EBMA method, as measured by CRPS against the ratio of the number of component models included and the size of the calibration period. As one can see, the performance of the EBMA model depends significantly on this ratio. As the number of forecast models included as components increases, or the calibration sample size decreases, CRPS rises. The predictive power of the EBMA model decreases as the number of components in the true data generating process increases. That is, as the number of model parameters that *must* be correctly estimated to make accurate predictions increases, the quality of the forecast goes down. Second, CRPS is a decreasing function of n_T, i.e. the performance of the baseline EBMA model improves as the calibration sample grows.

Figure 2: CRPS with "Wisdom of Crowds" Parameter



The remaining question, then, is to what degree adding the "wisdom of crowds" parameter to the baseline model improves performance. To answer, we examined the out-of-sample predictive performance of EBMA for differing values of c.

Figure 2 shows the median CRPS recorded for differing values of the ratio $\frac{K}{n_T}$, and c in our

simulation. Darker gray tones depict higher CRPS levels. The 3D plot clearly shows that while CRPS generally still increases with higher values of $\frac{K}{n_T}$, including a "wisdom of crowds" parameter within the EBMA algorithm can help.

While this is far from a complete analysis of the simulated data, it does serve the limited purposes of demonstrating that, in some circumstances, the "wisdom of crowds" parameter aids prediction under some circumstances.

There are two aspects of Figure **??** that are particularly salient. First, note that the addition of c to the base EBMA model does not uniformly aid in out-of-sample performance. When the calibration sample is modestly large and there are few models to calibrate, the addition of c uniformly decreases model performance. However, with small calibration samples and modestly large numbers of component models, the addition of c aids predictive performance.

Second, the relationship between c and CRPS is non-monotonic. CRPS decreases for small to modest values of c, but eventually begins to rise.

Based on our examination of the broader set of simulations, we generally recommend the selection of values of $c \in [0, 0.1]$. The simulations favor small values of c as the ratio of K to n_T increases. As a starting value, we recommend $c = 0.05$, however the "best" value of c may value from application to application. In our experience based on the simulations, as well as cross-validation studies during the applications presented below, we feel a c value at 0.05 is the best choice. Generally speaking, in choosing c the researcher faces a trade-off between possible overweighting certain models based on the performance in short calibration periods and moving too far into the direction of a simple average. A relatively small c seems to generally aid performance. It may be preferable to many researchers to choose the value of c based on a k-fold cross-validation study of the calibration sample.

Bearing these results in mind, we now turn to examining how these methods work in two areas that exemplify forecasting in the social sciences. The first is the prediction of unemployment in the United States, and the second in the area of predicting the vote for the incumbent party in U.S. presidential elections. Both areas have well developed forecasting traditions in the scholarly and

policy community.

4.2 Quarterly unemployment in the United States

Forecasting macroeconomic variables is a quite common exercise in the field of economics and statistics. Policy makers and businesses both have enormous interest in the calculation of accurate forecasts of economic variables. These forecasts are often created using a wide variety of statistical models, however often professional forecast are also based on expert knowledge.[8] The majority of scholars employ time-series models, most commonly applying autoregressive integrated moving average (ARIMA) and vector autoregressive (VAR) models. The sophistication and complexity of forecasting models has increased considerably over time. In particular, non-linear dynamic models have gained prominence including threshold autoregressive models, Markov switching autoregressive models and smooth transition autoregression (Elliott and Timmermann 2008; Montgomery et al. 1998). More recently, forecasters have added Bayesian VAR models and state-space models to their arsenal (De Goijer and Hyndman 2006; Elliott and Timmermann 2008).

Unsurprisingly, given the large number of ongoing forecasts, scholars have attempted to improve predictive accuracy by combining forecasts (Bates and Granger 1969; Palm and Zellner 1992; Elliott and Timmermann 2008). Recently, EBMA and related Bayesian model averaging methods have been successfully employed to create ensemble forecasts of various macroeconomic indicators including inflation (Koop and Korobilis 2009; Wright 2009), GDP (Billio et al. 2010), stock prices (Billio et al. 2011), and exchange rates (Wright 2008).

Policy makers too have come to rely on ensemble forecasts of a sort. The desire to aggregate the collective wisdom of multiple forecasting teams is apparent in the *Survey of Professional Forecasters (SPF)* published by the *Federal Reserve Bank of Philadelphia*. The *SPF* includes forecasts for a large number of macroeconomic variables in the U.S., including the unemployment rate, inflation, and GDP growth.[9] In the first month of every quarter, a survey is sent to selected fore-

[8]For a more comprehensive overview on foresting of economic variables and time-series forecasting see Elliott and Timmermann (2008) and De Goijer and Hyndman (2006).

[9]The *SPF* was first administered in 1968 by the American Statistical Association and the Na-

casters and is returned by the middle of the second month of the quarter. Forecasts are made for the current quarter as well as several quarters into the future.

This plethora of predictions seems ideal for applying EBMA. Nonetheless, it is plagued by the same issues as discussed in Section 2. Even with quarterly measures, there are relatively few observations, many forecasting teams, and a significant number of missing observations. This setting, therefore, provides a test bed for the adjusted EBMA model discussed above.

We focus on forecasting the civilian unemployment rate (UNEMP) as published by the *SPF*. For this application, we selected the forecast horizon to be four quarters into the future, i.e. predictions made in the first quarter of 2002 are for the first quarter of 2003 and so on. In total, the *SPF* data on unemployment contains forecasts by 569 different teams. However, for any quarter, the average number of forecast teams making a prediction for four quarters into the future is quite small and the majority of observations for any given quarter are missing.[10]

To provide a meaningful benchmark, we also include the "Green Book" forecasts produced by the Federal Reserve. These forecasts are made by the research staff of the Board of Governors and are handed out prior to meetings of the Federal Reserve Open Market Committee (FOMC).[11]

Taking the *SPF* and Green Book unemployment forecasts, we calibrate an ensemble model for each period t, using forecaster performance over the past ten quarters. Only forecasts that had made predictions for five of these quarters were included in the ensemble. Thus, the EBMA model uses only 163 models out of a possible 293 forecasting models that made predictions during the period we study. Due to missing data early in the time series, and the fact that Green Book forecasts are

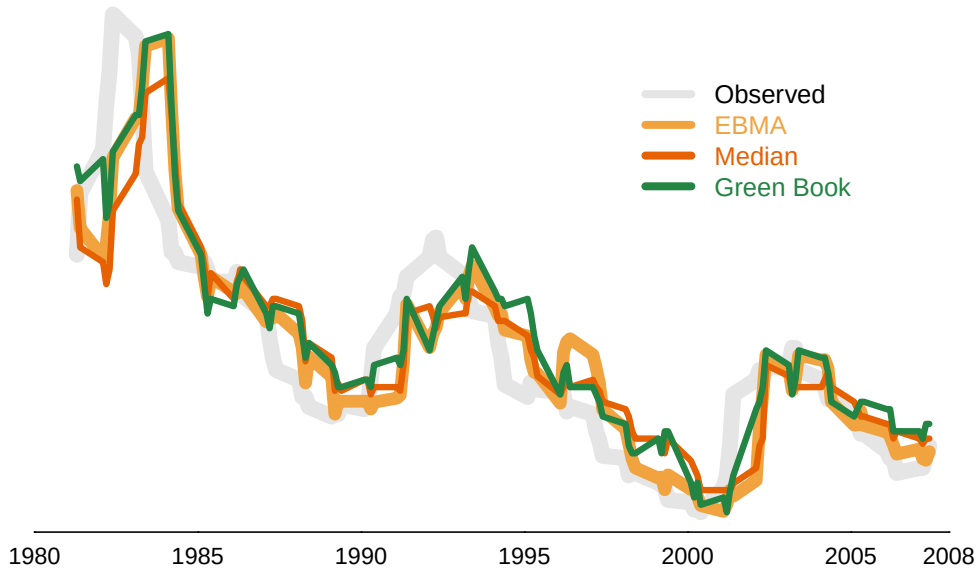
tional Bureau of Economic Research (NBER). Since 1990, however, it is run by the Federal Reserve Bank of Philadelphia. `http://www.phil.frb.org/research-and-data/real-time-center/survey-of-professional-forecasters/`

[10] On average only 8.4 per cent of all teams make a forecast for any one quarter.

[11] One issue with forecast evaluation in many domains in economics is that the macroeconomic data (i.e. our "true observations") are revised regularly. The unemployment rate for a given quarter at that time is generally an estimate that is subject to revision when better data becomes available. When evaluating forecasts, it is thus important whether predictions are compared to the outcome data for each quarter available at the time or whether the revised and most recent data is used. As Croushore and Stark (2001) describe, depending on the forecast exercise, it can make a difference whether the forecast models are evaluated using "real-time" (original estimate) or the "latest available" (revised) data. We have decided here to use the "latest available" data and do not believe that this choice affects our results, as all predictions are evaluated against the same data and EBMA is a mixture of the component forecast models. Thus the component models and our benchmark model are estimated and evaluated on the same data.

sequestered for five years, we generate forecasts beginning in the third quarter of 1983 and running through the fourth quarter of 2007.

Figure 3: Observed and forecasted U.S. unemployment (1981-2007)



One approach to evaluating the performance of EBMA is to compare its predictive accuracy to that made by other systematic forecasting efforts and methods of generating ensemble predictions. Specifically, we compare EBMA's ($c = 0.05$) predictive accuracy to (1) the Green Book, (2) the median forecaster prediction and (3) the mean forecaster prediction.[12]

Figure 3 shows a visual representation of the Greenbook, median SPF and the EBMA (with $c = 0.05$) forecasts over time, as well as the true unemployment rate. As was noted above and is clearly visible, the SPF and Greenbook forecasts are quite similar. Baghestani (2008) noted that the Greenbook forecast is slightly biased to over predict the unemployment rate. In some periods EBMA is able to correct this bias, however given the similarity of component models, the improvement in that direction is rather small. In general, however, it is easily visible that the EBMA forecast is closer to the actual rate than the median SPF or the Green Book forecast.

[12]Note that the EBMA model is calculated on only a the subset of forecasts that have made a sufficiently large number of recent predictions to calibrate model weights. Thus, the median forecast and the ensemble forecast will not be the same even when $c = 1$.

Table 2: Comparing adjusted EBMA models with Green Book, median, and mean forecasts of U.S. Unemployment (1981-2007)

	MAE	RMSE	MAD	RMSLE	MAPE	MEAPE	MRAE	PW
EBMA (c=0)	0.54	0.74	0.37	0.093	8.37	6.49	**0.73**	**27.36**
EBMA (c=0.05)	**0.54**	0.74	**0.37**	**0.093**	**8.33**	**6.30**	0.75	**27.36**
EBMA (c=0.1)	0.54	0.74	0.35	0.093	8.40	6.44	0.76	28.30
EBMA (c=1)	0.61	0.80	0.46	0.102	9.72	8.92	0.95	46.23
Green Book	0.57	**0.73**	0.43	0.093	9.37	8.81	1.00	45.28
Forecast Median	0.62	0.81	0.47	0.103	9.83	8.87	0.98	47.17
Forecast Mean	0.61	0.80	0.46	0.102	9.71	9.06	0.93	46.23

Definitions of model fit statistics are provided in the Appendix. The model with the lowest score for each metric are shown in bold. Differences between model performance may not be obvious due to rounding.

Table 2 formally compares these baseline models using all eight of the metrics to EBMA models with c =0, 0.05, 0.1, and 1 respectively. To do this, we focus on eight model fit indices available in the literature. The eight metrics we use are mean absolute error (MAE), root mean squared error (RMSE), median absolute deviation (MAD), root mean squared logarithmic error (RMSLE), mean absolute percentage error (MAPE), median absolute percentage error (MEAPE), median relative absolute error (MRAE) and percent worse (PW). The latter two metrics are measured relative to a naive model, simply predicting the future rate of unemployment as being the same as the current rate of unemployment. Further details for these metrics are shown in the Appendix (Brandt, Freeman and Schrodt 2011).

The bolded cells in each column of Table 2 indicate the model that performed "best" as measured by each metric. With one exception, (the Green Book outperforms the ensemble by 0.01 on RMSE), the EBMA model outperforms both the Green Book forecast and the unweighted mean and median forecast on every metric. Moreover, these results confirm that the c parameter is best set to a small number. In general, the model with $c = 0.05$ performs best (or is tied for best) on six out of eight of these metrics.

We now turn to evaluating the performance of the ensemble relative to its 163 component forecasts. It is important to note that many of these forecasters make predictions in a relatively

Table 3: Comparing predictive accuracy of EBMA and component models with eight metrics

	1 –10	11-30	31–60	> 60
7 – 8	0.20	0.33	0.10	0.04
5 – 6	0.04	0.04	0.03	0.01
2 – 4	0.02	0.03	0.01	0.00
0 – 1	0.08	0.04	0.01	0.00

The table shows EBMA's relative performance, as measured by eight forecasting metrics, to each of its components against the number of forecasts generated by the individual component models. Rows show the number of metrics EBMA scores better on than the respective component models, while the columns show the number of predictions made by these models.

small subset of cases. That is, each model k offers forecasts for only a subset of cases $n_k \subset n$. To create a fair comparison, therefore, we calculate these fit indices only for $n_k \forall k \in [1, K]$. By this measure, the EBMA model performs very well. Table 4.2 provides a summary of these results. The rows of the table show the number of metrics by which EBMA outperforms components in each column, while columns show the number of forecasts made by these models. The table is filled with the share of the total number of component models falling into this cell. Thus for the top left cell EBMA was better on at least 7 out of eight metrics for models making between 1 and 10 predictions, approximately 20% of models belong in this category.

Notably, the relative superiority of EBMA to its components is somewhat less for components that provide few forecasts. This reflects the fact that with so many forecasts, some are likely to be more accurate than the ensemble by chance alone. Additionally, when the number of forecasts is low it is likely that a given model received less weight than it "deserves" given the model's performance.[13] As can be seen in Table 4.2 however, across a large number of forecasts, EBMA significantly outperforms any of its components, i.e. when moving to the right in the table the values in the lower cells decrease. It is also worth noting that only 6 out of the total 163 components outperforms EBMA on every metric.

[13]See the Appendix for a discussion of how EBMA handles missing component forecasts.

4.3 U.S. presidential elections

We now turn to the task of combining expert predictions of U.S. presidential elections.[14] This example provides a clear illustration of the difficulties of creating ensemble forecasts in the social sciences and allows us to further illustrate the advantages of generating predictive PDFs when focusing on a limited number of important events.

Predicting U.S. presidential elections is, perhaps, the quintissential forecasting task that combines all of the issues discussed in Section 2. Table 4 represents nearly the entirety of scholarly forecasts which produced more than one true out-of-sample forecast for elections in the 20th century prior to the 2012 election.[15] In this instance, we have only five observations by which to calibrate an ensemble model, while we have nine forecasting models. Moreover, several of the individual forecasts are missing for a significant portion of the data. The forecast of Cuzàn, for instance, is missing for 60% of the elections in this dataset.[16]

Table 4: Pre-election forecasts of the percent of the two-party vote going to the incumbent party in U.S. Presidential elections

	Year	F	A	C	H	LBRT	L	Hol	EW	Cuz
1	1992	55.70	46.30	47.10	48.90					
2	1996	49.50	56.80	58.10	53.50	54.80		57.20	57.20	
3	2000	50.80	53.20	52.80	53.80	55.40	60.30	60.30	55.20	
4	2004	57.50	53.70	53.80	53.20	49.90	57.60	54.50	52.30	52.80
5	2008	48.10	45.70	52.70	48.20	49.90	41.80	44.30	47.80	48.00

Forecasts were published prior to each election by **F**air, **A**bramowitz, **C**ampbell, **H**ibbs, **L**ewis-**B**eck and **R**ice (1992), Lewis-Beck and **T**ien (1996-2008), **L**ockerbie, **Hol**brook, **E**rikson and **W**lezien and **Cuz**àn and Bundrick.

Using the forecasts shown in Table 1,[17] we fit an EBMA model with $c = 0.05$. The model

[14] See also (Montgomery, Hollenbach and Ward 2012*a*).

[15] See, for example Fair (2009, 2011); Abramowitz (2008); Campbell (2008*b*); Cuzàn and Bundrick (2004, 2008); Hibbs (2012*a*); Lockerbie (2008); Erikson and Wlezien (2008); Graefe et al. (2010); Holbrook (2008). A recent symposium in *PS: Political Science & Politics* presents and summarizes attempts by a variety of scholars to predict the 2012 U.S. presidential election. In a symposium contribution, we use the in-sample fitted values of the election forecasting models to calibrate the EBMA model (Montgomery, Hollenbach and Ward 2012*a*). However, the strength of EBMA is greatest when the model is calibrated on true out-of-sample forecasts as we do here.

[16] The predictions by Cuzàn for 2004 stems from the FISCAL model published prior to the 2004 election by Cuzàn and Bundrick (2004), while the 2008 prediction comes from the FPRIME short model presented in advance of the election (Cuzàn and Bundrick 2008). However, both models are quite similar in their composition.

[17] The out-of-sample predictions for these models were collected from the individual journal articles, personal web-

weights and in-sample fit statistics for the ensemble and its components are shown in Table 5. As can be seen, the EBMA model assigns the majority of weight to the Abramowitz model with the model by Hibbs receiving the second largest weight. These weights are based on the performance of each model in forecasting the incumbent vote share in the presidential elections between 1992 and 2008. The Cuzàn and Bundrick model is weighted to such a small degree because only out-of-sample predictions for 2004 and 2008 were available here.

Table 5: Model weights and in-sample fit statistics for EBMA model of U.S. Presidential Elections (1992-2008)

	EBMA Weight	RMSE	MAE
EBMA		1.92	1.49
Fair	0.02	5.53	4.58
Abramowitz	0.80	1.98	1.68
Campbell	0.02	3.63	3.08
Hibbs	0.06	2.31	2.18
Lewis-Beck, Rice, and Tien	0.06	2.87	2.16
Lockerbie	0.00	7.33	6.97
Holbrook	0.01	5.50	4.45
Erikson and Wlezien	0.02	2.90	2.50
Cuzàn	0.00	1.65	1.65

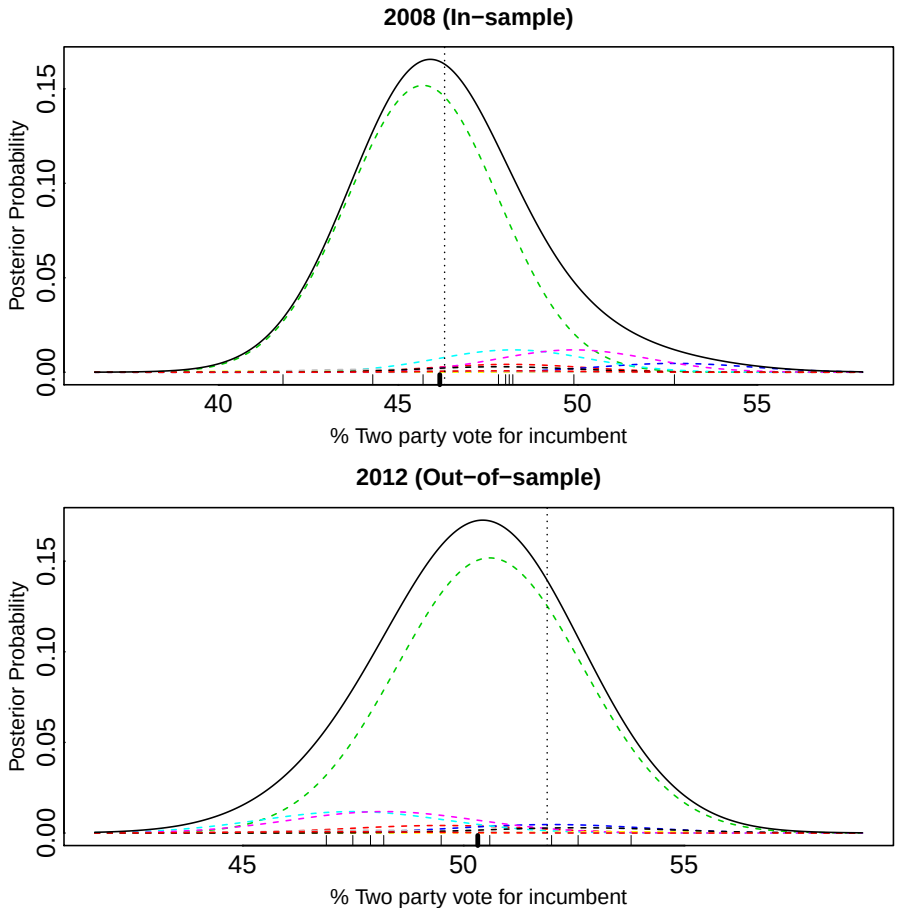
Figure 4 shows the posterior predictive distribution for the 2008 election (top) and, based on the forecasts from each of the component models, the 2012 election. Component models predictive distributions are shown in color (scaled by their respective weight), while the EBMA predictive distribution is shown in black, the bold dash displays the point prediction for the EBMA model. The vertical dashed line depicts the actual election results in 2008 and 2012.

The EBMA model ($c = 0.05$) predicted 50.3% of the two-party vote share for Obama. This resulted in a reasonably large absolute error of 1.7%.[18] Figure 4 shows that the EBMA prediction

sites and symposia introductions. When multiple forecasts were made we used the authors' preferred forecast, or took the mean if no preference was given (Hibbs 1992; Holbrook 1996; Lewis-Beck and Tien 1996; Wlezien and Erikson 1996; Abramowitz 2000; Hibbs 2000; Campbell and Garand 2000; Campbell 2000, 2001; Hibbs 2004; Campbell 2004, 2005, 2008*a*; Abramowitz 2012; Campbell 2012; Cuzàn 2012; Erikson and Wlezien 2012; Fair 2012; Hibbs 2012*b*; Holbrook 2012; Lewis-Beck and Tien 2012; Lockerbie 2012).

[18]The EBMA model assigned considerable weight to models predicting a Romney *victory*, especially Hibbs (2012*b*)

Figure 4: Predictive ensemble PDFs of incumbent-part vote share in U.S. Presidential Elections



The figure shows the density functions for each of the component models in different colors and scaled by their respective weight. The black curve is the density of the EBMA prediction, with the bold dash indicating the EBMA point prediction. The vertical dashed lines show the actual result of the 2008 and 2012 elections.

performs better than the majority of component forecasts for the 2012 election, with only three providing a more accurate estimate. However, we believe it is important to note that this it is difficult to evaluate EBMA against its components using just one out-of-sample observation.

A more important comparison is to examine how EBMA performs versus other methods of aggregating forecasts for the 2012 election. Table 6 shows the point predictions and absolute errors associated with the simple arithmetic mean, the median, and EBMA models fit with $c = 0$, $c = 0.05$, and $c = 0.1$with $c = 0$. While the differences in model weights are relatively small, the

and Lewis-Beck and Tien (2012).

Table 6: Comparing Model Results for U.S. Presidential Elections 2012

	Mean	Median	EBMA ($c = 0$)	EBMA ($c = 0.05$)	EBMA ($c = 0.1$)
2012 prediction	49.9	49.5	49	50.3	50.1
Absolute Error	2	2.4	2.9	1.6	1.8

EBMA prediction for 2012 was 49% and 50.1% for and EBMA model with $c = 0$ and $c = 0.1$ respectively – both considerably worse than the prediction with $c = 0.05$.

EBMA with a "wisdom of crowds" parameter of 0.05 also did considerably better relative to naive approaches to aggregation. The simple arithmetic average of the component models' predictions was 49.9% and the median prediction was 49.5%. In essence, while we believe prediction methods should be evaluated on more than one observation, this example again signifies the utility of EBMA in forecasting tasks and in particular the improvements the "wisdom of crowds" parameter offers to out-of-sample predictions in the context of sparse data.

5 Discussion

Ensemble Bayesian model averaging is a principled way of combining forecasts to improve prediction accuracy. However, the calibration of such models in the social sciences is often hindered by the quality as well as availability of data. For one, in many forecasting exercises the number of forecasting models is large, yet the number of observations on which the EBMA model can be trained is small. This creates problems for the estimation of model weights, as it is likely that overly high weights are assigned to models that are performing well over this particular period. Second, many predictive models do not provide forecasts for all observations in the sample, as some forecasts may be missing or the time-periods for which forecasts were made are different for different models. In the standard EBMA model introduced in Montgomery, Hollenbach and Ward (2012*b*) missing observations in component model predictions are not allowed.

In this article, we address both of these issues to make EBMA more applicable for researchers

and predictioneers in the social sciences. After reviewing the standard EBMA framework, we proceed introduce a "wisdom of the crowds" parameter into the model, which forces EBMA to put some minimal weight on all component models. Adding this constant aids the calibration of EBMA when the number of observations in the calibration period is small.

After explaining our adjustments, we illustrated its advantages via simulation. We then apply the adjusted EBMA model in two prediction exercises. We use ensemble Bayesian model averaging to combine predictions of the unemployment rate in the US from the Survey of Professional Forecasters as well as the Green Book. As we show, even when a large number of forecasts is missing for any given quarter, EBMA generally outperforms the Green Book, SPF component models, as well as the median and mean SPF forecast.

In a second example, we use the out-of-sample forecasts of nine prediction models of presidential elections from 1992 to 2008 to calibrate an ensemble model. We use the model calibrated to make an informed prediction for the 2012 elections based on a weighted combination of the component model predictions for 2012. This example neatly illustrates the common difficulties facing forecasters in the social science, and provides an illustrative example for applied researchers going forward.

A comprehensive approach to the data problems raise in section 2 would be to estimate the "wisdom of crowds" parameter within the EBMA algorithm specifically for each forecasting application. So far we refrain from doing as we are concerned with the number of parameters being estimated on relatively small numbers of observations (i.e. limited degrees of freedom). In addition, simple solutions have so far failed because of issues of identifiability. In future work we plan to rewrite the EBMA algorithm further to make estimation of the c possible. In addition we plan to implement imputation techniques within the EBMA algorith to handle missing data. While our current algorithm follows Fraley, Raftery and Gneiting (2010) and allows for the inclusion of models with missing predictions, those are severely down weighted based on the number of missing observations. The algorithm employed here is driven by the application in meteorological sciences where weather stations may fail to report observations randomly. In the social sciences however

different approaches may be more appropriate. In the future we thus plan to implement imputation of missing observations via copula methods within the EBMA framework.

References

Abramowitz, Alan I. 2000. Bill and Al's Excellent Adventure: Forecasting the 1996 Presidential Election. In *Before the Vote: Forecasting American National Elections*, ed. James E. Campbell and James C. Garand. Thousand Oaks: Sage Publications chapter 2, pp. 47–56.

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

Abramowitz, Alan I. 2012. "Forecasting in a Polarized Era: The Time for Change Model and the 2012 Presidential Election." *PS: Political Science & Politics* 45(4):618–619.

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.

Baghestani, Hamid. 2008. "Federal Reserve versus private information: Who is the best unemployment rate predictor?" *Journal of Policy Modeling* 30(1):101–110.

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

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

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

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

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

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

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

Breiman, L. 1996. "Bagging predictors." *Machine Learning* 26:123–140.

Breiman, L. 2001. "Random forests." *Machine Learning* 45:5–32.

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

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

Campbell, James E. 2000. Polls and Votes: The Trial-Heat Presidential Election Forecasting Model, Certainty, and Political Campaigns. In *Before the Vote: Forecasting American National Elections*, ed. James E. Campbell and James C. Garand. Thousand Oaks: Sage Publications chapter 1, pp. 17–46.

Campbell, James E. 2001. "Taking Stock of the Forecasts of the 2000 Presidential Election." *American Politics Research* 29(3):275–278.

Campbell, James E. 2004. "Introduction—The 2004 Presidential Election Forecasts." *PS: Political Science & Politics* 37(4):733 – 735.

Campbell, James E. 2005. "Introduction—Assessments of the 2004 Presidential Vote Forecasts." *PS: Political Science & Politics* 38(1):23–24.

Campbell, James E. 2008*a*. "Editor's Introduction: Forecasting the 2008 National Elections." *PS: Political Science & Politics* 41(4):679–81.

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

Campbell, James E. 2012. "Forecasting the Presidential and Congressional Elections of 2012: The Trial-Heat and the Seats-in-Trouble Models." *PS: Political Science & Politics* 45(4):630–634.

Campbell, James E. and James C. Garand. 2000. *Before the Vote: Forecasting American National Elections*. Thousand Oaks: Sage Publications.

Chipman, Hugh A., Edward I. George and Robert E. McCulloch. 2010. "BART: Bayesian Additive Regression Trees." *The Annals of Applied Statistics* 4(1):266–298.

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

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

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

Croushore, Dean and Tom Stark. 2001. "A Real-Time Data Set for Macroeconomists." *Journal of Econometrics* 105(1):111–130.

Cuzàn, Alfred G. 2012. "Forecasting the 2012 Presidential Election with the Fiscal Model." *PS: Political Science & Politics* 45(4):648–650.

Cuzàn, Alfred G. and Charles M. Bundrick. 2004. "Fiscal Effects on Presidential Elections: A Forecast for 2004." Paper prepared for presentation at the American Political Science Association, Chicago `http://uwf.edu/govt/facultyforums/documents/fiscaleffectsprselect2004.pdf`.

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.

De Goijer, Jan G. and Rob J. Hyndman. 2006. "25 years of time series forecasting." *International Journal of Forecasting* 22(3):443–473.

Elliott, Graham and Allan Timmermann. 2008. "Economic Forecasting." *Journal of Economic Literature* 46(1):3–56.

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

Erikson, Robert S. and Christopher Wlezien. 2012. "The Objective and Subjective Economy and the Presidential Vote." *PS: Political Science & Politics* 45(4):620–624.

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

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

Fair, Ray C. 2012. "Personal Website." `http://fairmodel.econ.yale.edu/` (accessed january 27, 2013).

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

Feldman, Kira. 2012. Statistical Postprocessing of Ensemble Forecasts for Temperature: The Importance of Spatial Modeling. Master's thesis Ruprecht-Karls-Universitat Heidelberg.

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

Freund, Y. and R.E. Schapire. 1997. "A decision-theoretic generalization of online learning and an application to boosting." *J. Comput. System Sci.* 55:119–139.

Friedman, J.H. 2001. "Greedy function approximation: A gradient boosing machine." *Ann. Statist* 29:1189–1232.

Galton, Francis. 1907. "Vox populi." *Nature* 75(1949):450–451.

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

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

Gneiting, Tilmann, Fadoua Balabdaoui and Adrian E. Raftery. 2007. "Probabilistic Forecasts, Calibration and Sharpness." *Journal of the Royal Statistical Society. Series B (Methodological)* 69(2):243–268.

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

Goldstone, Jack A., Robert H. Bates, David L. Epstein, Ted Robert Gurr, Michael B. Lustik, Monty G. Marshall, Jay Ulfelder and Mark Woodward. 2010. "A Global Model for Forecasting Political Instability." *American Journal of Political Science* 54(1):190–208.

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

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

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

Hibbs, Douglas A. 1992. "THE 1992 PRESIDENTIAL ELECTION: Clinton is the Probable Winner, but Only Narrowly, Getting Just over 51 Percent of the Two-Party Vote." `http://www.douglas-hibbs.com/Elections2004-00-96-92/forecast92.pdf` (accessed Jan. 27, 2013).

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

Hibbs, Douglas A. 2004. "Implications of the 'Bread and Peace Model' of US Presidential Voting for the 2004 Election Outcome." http://www.douglas-hibbs.com/Elections2004-00-96-92/election2004.pdf (accessed Jan. 27, 2013).

Hibbs, Douglas A. 2012*a*. "Obama's Re-election Prospects under 'Bread and Peace/ Voting in the 2012 US Presidential Election." `http://www.douglas-hibbs.com/HibbsArticles/HIBBS_OBAMA-REELECT-31July2012r1.pdf`.

Hibbs, Douglas A. 2012*b*. "Obama's Reelection Prospects under "Bread and Peace" Voting in the 2012 US Presidential Election." *PS: Political Science & Politics* 45(4):635–639.

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. 1996. "Reading the Political Tea Leaves : A Forecasting Model of Contemporary Presidential Elections." *American Politics Research* 24(4):506–519.

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

Holbrook, Thomas M. 2012. "Incumbency, National Conditions, and the 2012 Presidential Election." *PS: Political Science & Politics* 45(4):640–643.

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

Imai, Kosuke and Dustin Tingley. 2012. "A Statistical Method for Empirical Testing of Competing Theories." *American Journal of Political Science* 56(1):218–236.

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

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

Lewis-Beck, Michael S. and Charles Tien. 1996. "The Future in Forecasting : Prospective Presidential Models." *American Politics Research* 24(4):468–491.

Lewis-Beck, Michael S. and Charles Tien. 2012. "Election Forecasting for Turbulent Times." *PS: Political Science & Politics* 45(4):625–629.

Linzer, Drew. Forthcoming. "Dynamic Bayesian Forecasting of Presidential Elections in the States." *Journal of the American Statistical Association* .

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

Lockerbie, Brad. 2012. "Economic Expectations and Election Outcomes: The Presidency and the House in 2012." *PS: Political Science & Politics* 45(4):644–647.

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.

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

McLachlan, Geoffrey and David Peel. 2000. *Finite Mixture Models*. New York, NY: John Wiley & Sons, Ltd.

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

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

Möller, Anette, Alex Lenkoski and Thordis L. Thorarinsdottir. 2013. "Multivariate Probabilistic Forecasting Using Ensemble Bayesian Model Averaging and Copulas." *Quarterly Journal of the Royal Meteorological Society* Forthcoming.

Montgomery, Alan L., Victor Zarnowitz, Ruey Tsay and Tiaom George. 1998. "Forecasting the U.S. Unemployement Rate." *Journal of the American Statistical Association* 93(442):478–493.

Montgomery, Jacob M., Florian Hollenbach and Michael D. Ward. 2012*a*. "Ensemble Predictions of the 2012 US Presidential Election." *PS: Political Science & Politics* 45(4):651–654.

Montgomery, Jacob M., Florian Hollenbach and Michael D. Ward. 2012*b*. "Improving Predictions Using Ensemble Bayesian Model Averaging." *Political Analysis* 20(3):271–291.

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

Palm, Franz C. and Arnold Zellner. 1992. "To Combine of Not to Combine? Issues of Combining Forecasts." *Journal of Forecasting* 11(8):687–701.

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.

Rings, Joerg, Jasper A. Vrugt, Gerrit Schoups, Johan A. Huisman and Harry Vereecken. 2012. "Bayesian Model Averaging Using Particle Filtering and Gaussian Mixture Modeling: Theory, Concepts, and Simulation Experiments." *Water Resources Research* 48(5):1–12.

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

Surowiecki, J. 2004. "The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business." *Economies, Societies and Nations* .

Ulfelder, Jay. 2012. "Forecasting Onset of Mass Killings." Paper presented at the annual Northeast Political Methodology Meeting at New York University .

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

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

Wlezien, Christopher and Robert S. Erikson. 1996. "Temporal Horizons and Presidential Election Forecasts." *American Politics Research* 24(4):492–505.

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

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

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

Appendix A: EM-Algorithm for missing data

To accommodate missing values in component models prediction within the EBMA procedure we follow Fraley, Raftery and Gneiting (2010) and modify the EM algorithm as follows. Define

$$\mathcal{A}^t = \{i|\text{ensemble member i available at time t}\},$$

which is simply the indicators of the list of components that provide forecasts for observation y^t. For convenience, define $\tilde{z}_k^{(j+1)t} \equiv \sum_{k \in A^t} \hat{w}_k^{(j)} p^{(j)}(y|f_k^t) / \sum_{k \in A^t} w_k^{(j)}$. Equation 3 above is then replaced with

$$\hat{\tilde{z}}_k^{(j+1)t} = \begin{cases} \hat{w}_k^{(j)} p^{(j)}(y|f_k^t)/\tilde{z}_k^{(j+1)t} \text{ if } k \in \mathcal{A}^t \\ 0 \text{ if } k \notin \mathcal{A}^t \end{cases} \tag{7}$$

The M steps in Equations 4 and 5 are likewise replaced with

$$\hat{w}_k^{(j+1)} = \frac{\sum_t \hat{\tilde{z}}_k^{(j+1)t}}{\sum_t \sum_{k=1}^{K} \hat{\tilde{z}}_k^{(j+1)t}} \tag{8}$$

and

$$\hat{\sigma}^{2(j+1)} = \frac{\sum_t \sum_{k=1}^{K} \hat{\tilde{z}}_k^{(j+1)t} (y - f_k^t)^2}{\sum_t \sum_{k=1}^{K} \hat{\tilde{z}}_k^{(j+1)t}}. \tag{9}$$

In essence, the likelihood is renormalized given the missing ensemble observations prior to maximization. Using the adjustments above, the EBMA algorithm now allows for missing observations in the component predictions.

Appendix B: Predictive Metrics

Let x be some prediction of an event, for example a prediction model for the U.S. presidential election. Now let $p(x)$ denote the PDF associated with forecast x and x_a be the actual observed values. The continuous rank probability score CRPS for forecast x and outcome y is then:

$$CRPS = CRPS(P, y) = \int_{-\infty}^{\infty} [P(x) - P_a(x)]^2 dx \tag{10}$$

where P and P_a are cumulative distribution functions, such that:

$$P(x) = \int_{-\infty}^{x} p(y)dy \tag{11}$$

and

$$P_a(x) = H(x - x_a) \tag{12}$$

H(x) denotes the Heaviside function where $H(x) = 0$ for $x < 0$ and $H(x) = 1$ for $x \geq 0$ (Hersbach 2000). The CRPS ranges from zero to one, with the best forecast models scoring closer to zero.[19]

Denote the forecast of observation i as f_i and the observed outcome as y_i. We define the *absolute error* as $e_i \equiv |f_i - y_i|$ and the *absolute percentage error* as $a_i \equiv e_i/|y_i| \times 100$. Finally, for each observation we have prediction from naive forecast, r_i, that serves as a baseline for comparison. In the example is the main text, this naive model is simply the lagged observation. We can therefore define $b_i \equiv |r_i - y_i|$.[20]

Denoting the median of some vector $\mathbf{x}$ as $med(\mathbf{x})$, and the standard indicator function as $I(.)$,

[19]The notation here is borrowed from Hersbach (2000), and Gneiting, Balabdaoui and Raftery (2007).

[20]See Brandt, Freeman and Schrodt (2011) for additional discussion of comparative fit metrics.

we define the following heuristic statistics:

$$
\begin{aligned}
\text{MAE} &= \frac{\sum_1^n e_i}{n} \\
\text{RMSE} &= \sqrt{\frac{\sum_1^n e_i^2}{n}} \\
\text{MAD} &= \text{med}(\mathbf{e}) \\
\text{RMSLE} &= \sqrt{\frac{\sum_1^n \left(\ln(f_i+1) - \ln(y_i+1)\right)^2}{n}} \\
\text{MAPE} &= \frac{\sum_1^n a_i}{n} \\
\text{MEAPE} &= \text{med}(\mathbf{a}) \\
\text{MRAE} &= \text{med}\left(\frac{e_1}{b_1}, \ldots, \frac{e_n}{b_n}\right) \\
\text{PW} &= \frac{\sum_1^n I(e_i > b_i)}{n} \times 100
\end{aligned}
$$