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

January 17, 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 **?** 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 in the context of ensemble prediction and provide a solution that diminishes undesirable outcomes by introducing a "wisdom of the crowds" parameter to the standard EBMA framework. 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., **?**), political instability (e.g., **?**), and mass killings (**?**). 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. 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 model in isolation can provide. Across subject domains, scholars have shown ensemble predictions to be more accurate than any individual component model and less likely to make dramatically incorrect predictions (**???**).

One promising approach to combining multiple forecasts is ensemble Bayesian model averaging (EBMA). This method was first proposed by **?** to combine weather forecasts and was introduced to the social sciences by **?**. 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 or over-estimating our certainty about the future. 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 **?** 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., **?**). Among the more famous demonstrations of this phenomenon is a competition to guess the weight of an ox at the West Of England Fat Stock and Poultry Exhibition reported by **?**. Galton 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. This observation is related to Condorcet's jury theorem that calculates the probability of a jury reaching a correct verdict will be greater than even the most accurate of judges.

In recent years, the advantages of ensembles have come to play a particularly prominent role in the machine learning and nonparametric statistics community (**?**). 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 (**??**), bagging (**?**), random forests (**?**), and related techniques (e.g., **?**) 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" (**?**), 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., **?**). Not all guesses are equally informative, and naive approaches to collating forecasts risks both overvaluing wild guesses and undervaluing unusual that are nonetheless sometimes correct.

The particular ensemble method we are extending is ensemble Bayesian model averaging (EBMA). First proposed by **?**, 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 two aspects its past predictive accuracy and uniqueness.

FLORIAN: THIS PARAGRAPH NEEDS AN UPDATE. FIND RECENT WORK, AND TRY TO ILLUSTRATE THAT EBMA IS A BURGEONING AREA OF RESEARCH. In recent years, variants of the EBMA method have been applied to subjects as diverse as inflation (**???**), stock prices (**?**), economic growth and policymaking (**??**), exchange rates (**?**), industrial production (**?**), ice formation (**?**), visibility (**?**), water catchment streamflow (**?**), climatology (**???**), and hydrology (**?**). Indeed, research is underway to extend the method to handle missing data (**??**) as well as calibrate model weights on non-likelihood criteria (e.g., **?**).

Although there are clear advantages to combining forecasts via EBMA, several challenges common to the social sciences pose a particular challenge to more widespread adoption of the method. 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 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.

A second, and related, issue is dimensionality. Many prediction tasks 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 multiple 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 of multiple key economic indicators for the next three years.[1]

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. Moreover, forecasts have tended 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. Moreover, they are not 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. We therefore ... TRANSITION NEEDED.

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

3 EBMA for sparse data

As its name suggests, EBMA descends from Bayesian model averaging (BMA) methodology (c.f., **?????**), which was first introduced to political science by **?** and has been applied in a number of contexts (e.g., **???**). **?** provide a more in-depth discussion of BMA and its applications in political science. A more detailed discussion of the basic EBMA model extended here is provided in **?**.

3.1 Baseline EBMA model

Assume we have some quantity of interest to forecast, $\mathbf{y}^{t^*}$, in future period $t^* \in T^*$. Further assume that we have extant forecasts for events $\mathbf{y}^t$ for some past period $t \in T$ that were generated from K forecasting models or teams, $M_1, M_2, \ldots, M_K$, for which 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 $p(M_k|\mathbf{y}^t) = p(\mathbf{y}^t|M_k)\pi(M_k)/\sum_{k=1}^{K} p(\mathbf{y}^t|M_k)\pi(M_k)$, and the marginal predictive PDF is $p(\mathbf{y}^{t^*}) = \sum_{k=1}^{K} p(\mathbf{y}^{t^*}|M_k)p(M_k|\mathbf{y}^t)$. This can be viewed as the weighted average of the component PDFs where the weights are determined by each model's performance within the already-observed period T.

The EBMA procedure assumes K forecasting models throughout the training (T') calibration (T) and test (T^*) periods. The component models are created based on data from the training period T'. The component model predictions for the calibration period T are then generated out-of-sample. The goal is to estimate the parameters for the ensemble prediction model using $\mathbf{f}_k^t$ for period T. It is then possible to generate true ensemble forecasts ($\mathbf{f}_k^{t^*}$) for observations in the test period $t^* \in T^*$. This approach allows us to weight models based on their 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 the 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 $p(y|f_1^{t^*}, \ldots, f_K^{t^*}) = \sum_{k=1}^{K} w_k g_k(y|f_k^{t^*})$.[2] 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

FLORIAN: THIS SECTION NEEDS TO BE CHANGED A BIT SO IT IS NOT A DIRECT COPY OF THE FIRST ARTICLE. DONT CHANGE THE MATH, JUST THE WORDS.

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 (**?**). 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. **?** propose an EM algorithm which expresses EBMA as a finite mixture model (**??**). We introduce the unobserved quantities z_k^t, which represents the probability that observation y^t is "best" predicted by model k. The E step involves calculating estimates for

[2]Past applications have statistically post-processed the predictions for out-of-sample bias reduction and treated these adjusted predictions as a component model. **?** 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.

these unobserved quantities using the formula

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

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). Assuming these estimates of $z_k^{s|t}$ are correct, it is then straightforward to derive the maximizing value for the model weights. Thus, the M step estimates these as

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

where n represents the number of observations in the calibration dataset. Finally,

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

The E and M steps are iterated until the improvement in the log-likelihood is no larger than some pre-defined tolerance. We initiate the algorithm with the assumption that all models are equally likely, $w_k = \frac{1}{K} \; \forall \; k \in [1, \ldots, K]$ 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.[3]

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

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

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

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

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 $c = 1$.[4] Likewise, the general EBMA discussed in **?** 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 common when the size of the calibration sample is small, there are missing components in the calibration sample, and the number of component forecasts for which weights must be estimated is large. Above, we claim that this increases the miss-estimation of component model weights and decrease the predictive performance of EBMA.

To justify these claims, in this section 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.

[4]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.[5] 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). This last is used only for model estimation, and plays no role in the creation of the simulated data itself.

[5]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 model and compare then them 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) . We focus here on the CRPS measure because it is awesome. FLORIAN: BRIEF DESCRIPTION OF CRPSIN MAIN TEXT HERE. MATHEMATICAL DETAILS IN THE APPENDIX. CITE RELEVANT WORK. EXPLAIN INTUITION OF WHY IT IS BEST.

Figure 1: A plot to show that error is a function of K and N_T

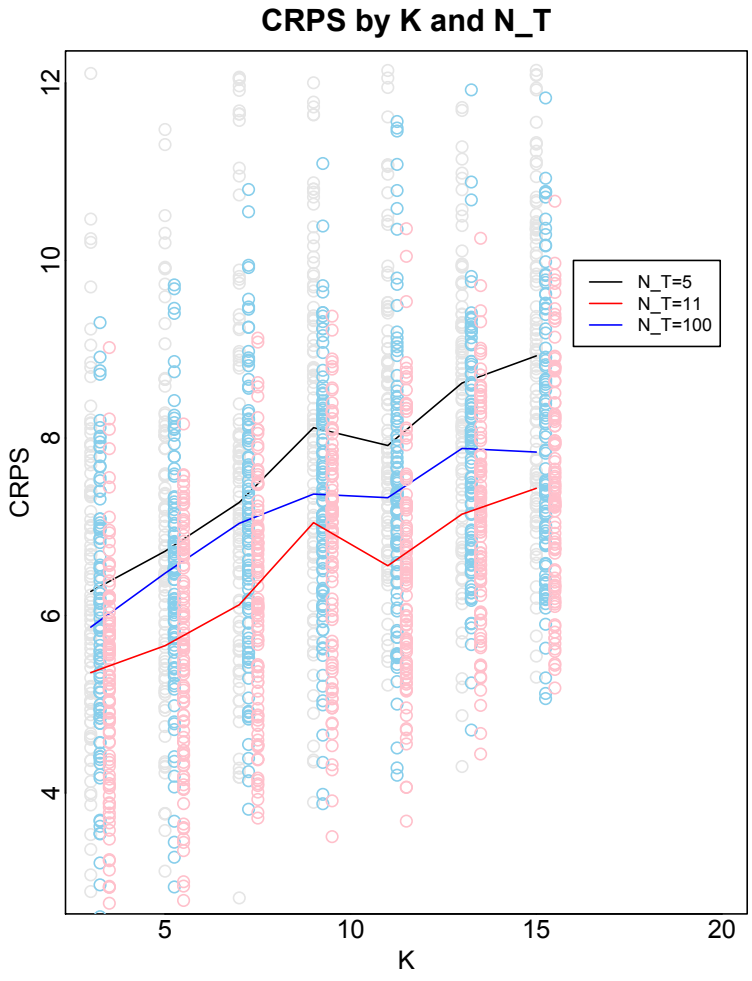
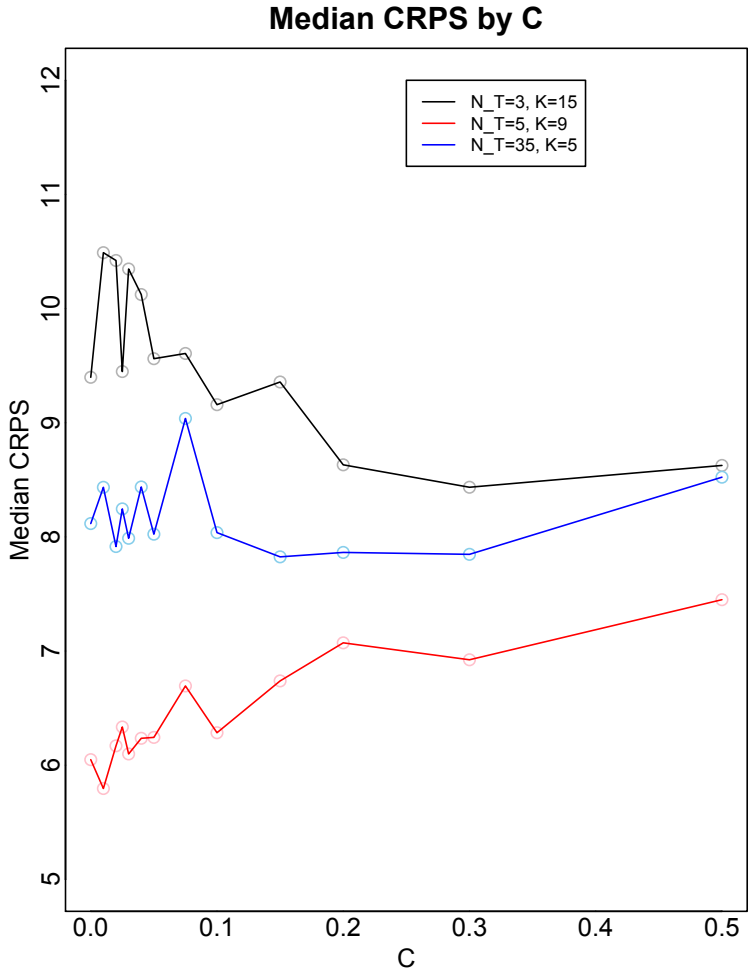


Figure 1 shows how the out-of-sample performance of the EBMA method, as measured by CRPS, depends significantly on the number of forecast models included as well as the calibration sample size. First, note the upward slope of each line, indicating that the predictive power of the model decreases as the 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, Figure 1 shows out-of-sample CRPS when $n_T = 5$ (black), $n_T = 11$ (blue), and $n_T = 100$ (red). For all values of k, the CRPS is a decreasing function of N_T. This illustrates that the performance of the baseline EBMA model improves as the calibration sample grows.

Figure 2: A plot to show that c helps in some situations



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 N_T, K, and c. For the purposes of clarity, we focus on only three subsets from the simulations: $n_T = 3, K = 15$, $n_T = 5, K = 9$, $n_T = 35, K = 5$. 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 modeldoes not uniformly aid in out-of-sample performance. When the calibration sample is modestly large ($N_T = 35$) 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 appears to aid 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 therefore recommend the selection of values of $c \in [0, 0.1]$. The simulations favor smaller values of c as the ratio of N_T to K increases. As a default choice, we recommend $c = 0.05$. Researchers may also 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. Calculating accurate forecasts of economic variables is a necessity for policy makers and

businesses. These forecasts are created using a wide variety of statistical models.[6] The majority of scholars employ time-series models, with the most commonly applied statistical method being 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 (**??**). More recently, forecasters have introduced Bayesian VAR models and state-space models to their arsenal (**??**).

Unsurprisingly, given the large number of ongoing forecasts, scholars have attempted to improve predictive accuracy by combining forecasts (**???**). Recently, EBMA and related Bayesian model averaging methods have been successfully employed to create ensemble forecasts of various macroeconomic indicators including inflation (**??**), GDP (**?**), stock prices (**?**), and exchange rates (**?**).

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.[7] In the first month of every quarter, a survey is sent to selected forecasters 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.

[6] For a more comprehensive overview on foresting of economic variables and time-series forecasting see **?** and **?**.

[7] The *SPF* was first administered in 1968 by the American Statistical Association and the National 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/`

Here 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.[8]

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).[9]

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 sequestered for five years, we generate forecasts beginning in the third quarter of 1983 and running through the fourth quarter of 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.[10]

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

[9]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 **?** 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 it should make a difference in our case, 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. However, in a future version of this paper we will replicate this analysis using real-time data to evaluate the forecasts.

[10]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 $c = 1$.

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

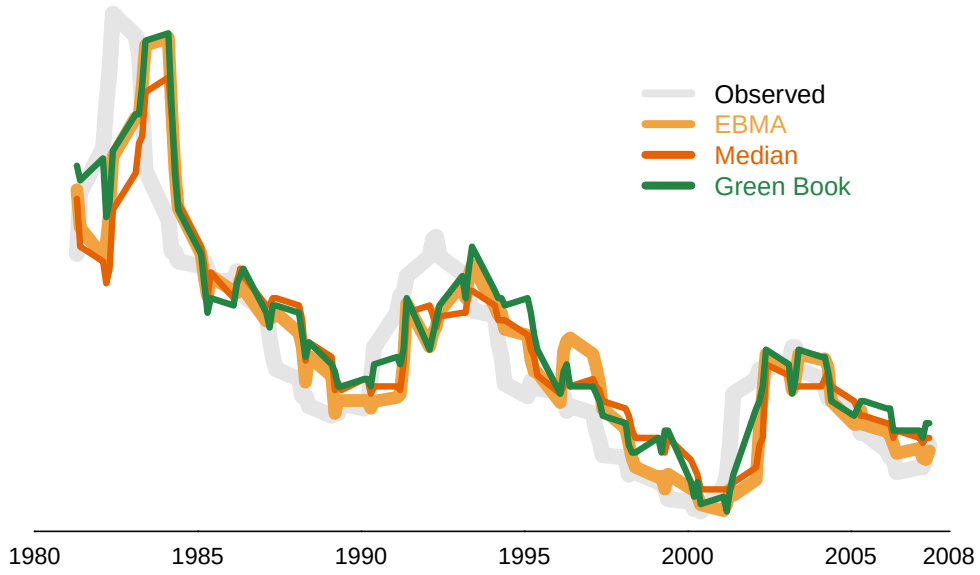


Figure 3 shows a visual representation of the Greenbook, median SPF and the EBMA (with $c = 0.05$) forecasts over time, as well as the true unemployment rate. As was noted above and is clearly visible, the SPF and Greenbook forecasts are quite similar. **?** noted that the Greenbook forecast is slightly 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.

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 (**?**).

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

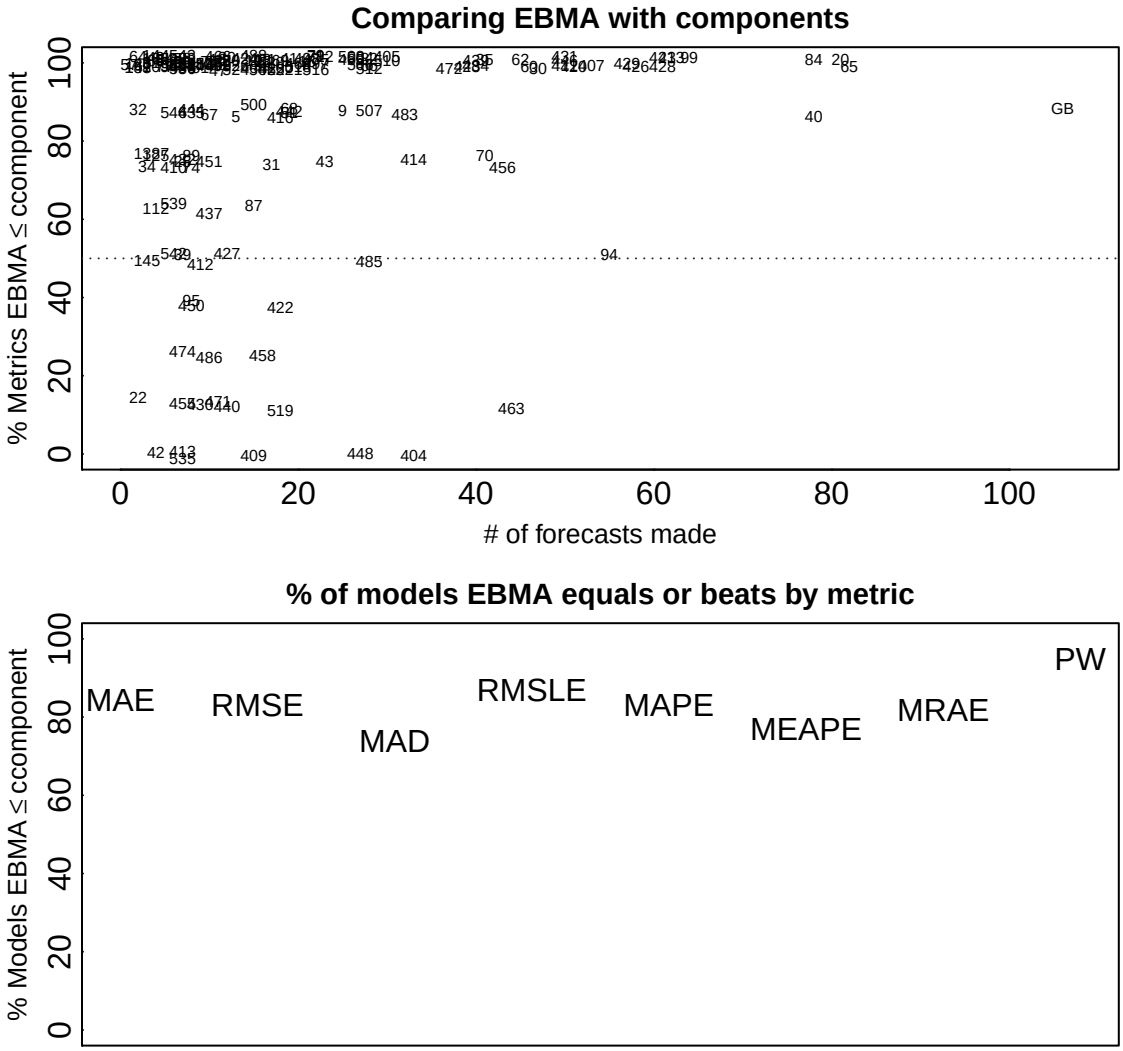
Definitions of model fit statistics are provided in the Appendix. The model with the lowest score for each metric are shown in bold. Differences between model performance may not be obvious due to rounding.

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 confirms 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 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. Figure 4 provides a summary of these results. The top panel shows the percentage of metrics by which EBMA outperforms each component. The bottom panel shows the percentage of component models that EBMA "beats" as measured by each metric.

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

Figure 4: Comparing predictive accuracy of EBMA and component models with eight metrics



The top panel plots EBMA's relative performance, as measured by eight forecasting metrics, to each of its components against the number of forecasts generated by the component models. The bottom panel shows the percentage of component models that EBMA matches or outperforms as measured by each metric. Details on the eight forecasting metrics are shown in the Appendix. The Green Book forecast is labeled (GB) and is located in the upper left portion of the top panel).

is low it is likely that a given model received less weight than it "deserves" given the model's performance.[11] However, across a large number of forecasts, EBMA significantly outperforms any of its components, including the Green Book (GB). It is also worth noting that only 6 out of the

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

total 163 components outperforms EBMA on every metric.

4.3 U.S. presidential elections

We now turn to the task of combining expert predictions of U.S. presidential election.[12] 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 **??** represents nearly the entirety of scholarly forecasts which produced more than one out-of-sample forecast for elections in the 20th century prior to the 2012 election.[13] 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.[14]

Using the forecasts shown in Table 1, we fit an EBMA model with $c = 0.05$. The model weights and in-sample fit statistics for the ensemble and its components are shown in Table **??**. As can be seen, the EBMA model assigns the majority of weight to the Abramowitz model with the model by Campbell 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.

Figure **??** shows the posterior predictive distribution for the 2008 election (top) and, based on

[12] See also (**?**).

[13] See, for example **??????????**. 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 (**?**). However, the strength of EBMA is greatest when the model is calibrated on true out-of-sample forecasts as we do here.

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

Table 3: Pre-election forecasts of the percent of the two-party vote going to the incumbent party in U.S. Presidential elections

	F	A	C	H	LBRT	L	Hol	EW	Cuz
1992	55.7	46.3	49.7	48.9	47.3				
1996	49.5	57.0	55.5	53.5	53.3		57.2	55.6	
2000	50.8	53.2	52.8	54.8	55.4	60.3	60.3	55.2	
2004	57.5	53.7	52.8	53.2	49.9	57.6	55.8	52.9	51.1
2008	48.1	45.7	52.7	48.5	43.4	41.8	44.3	47.8	48.1

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. Data were taken from the collation presented at `http://fivethirtyeight.blogs.nytimes.com/2012/03/26/models-based-on-fundamentals-have-failed-at-predicting-presidential-elections/`. FLORIAN: WE NEED TO VERIFY ALL OF THESE NUMBERS AND ADD CITATIONS TO THE RELEVANT WORK ... FOOTNOTE?

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. Vertical dashes indicate the point prediction of each model (bold dash for the EBMA model). The vertical dashed line in the top panel depicts the actual election result in 2008.

FLORIAN: WE NEED TO UPDATE THIS TO SHOW THE ACTUAL RESULTS. TRY TO ADD A DISCUSSION HERE OF HOW GOOD THE METHOD IS. I SUPPOSE WE SHOULD RUN THE REGULAR EBMA WITHOUT C AND SEE HOW WE DO IN COMPARISON. TRY AND PUT SOME LIPSTICK ON THAT PIG, AND THEN WE'LL SEE WHERE WE ARE AT.

5 Discussion

FLORIAN: CAN YOU ADD SOMETHING IN HERE ABOUT WHY WE DON'T ESTIMATE C AND HOW WE MIGHT DO THAT IN THE FUTURE?

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

Table 4: Model weights and in-sample fit statistics for EBMA model of U.S. Presidential Elections (1992-2008)

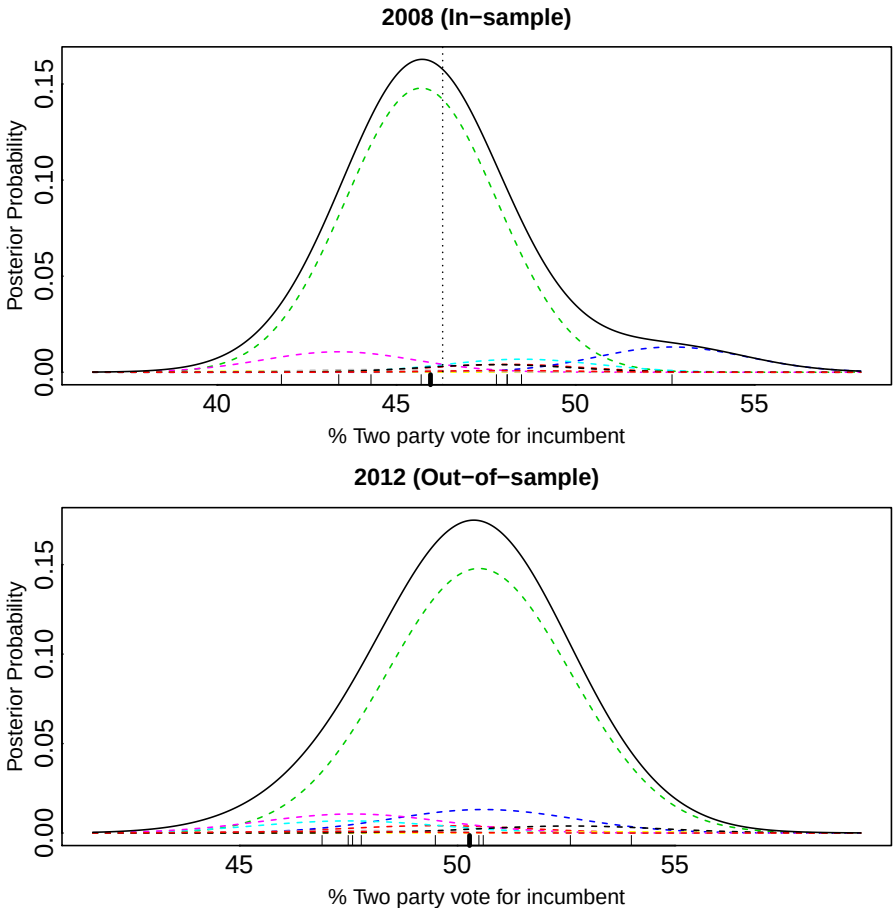
	EBMA Weight	RMSE	MAE
EBMA		1.92	1.56
Fair	0.02	5.53	4.58
Abramowitz	0.78	2.02	1.72
Campbell	0.07	3.46	2.88
Hibbs	0.04	2.68	2.44
Lewis-Beck, Rice, and Tien	0.06	2.78	2.28
Lockerbie	0.00	7.33	6.97
Holbrook	0.01	5.73	4.77
Erikson and Wlezien	0.02	2.74	2.25
Cuzàn	0.00	1.27	0.95

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. This is especially true should the EBMA model be calibrated on in-sample forecasts of the component models. 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 **?** 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 applied 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

Figure 5: 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 point predictions of the individual models are depicted by smalls vertical dashes. The black curve is the density of the EBMA prediction, with the bold dash indicating the EBMA point prediction. For 2008 the vertical dashed line shows the actual result.

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.

FLORIAN ... FUTURE WORK ? MAYBE FOCUS ON BETTER WAYS OF HANDLING MISSING DATA.

References

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

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.

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. 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.

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. 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.

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).

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

Fraley, Chris, Adrian E. Raftery 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 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.

Hibbs, Douglas A. 2012. "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`.

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

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

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

Imai, Kosuke and 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).

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.

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.

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

Montgomery, Jacob M., Florian Hollenbach and Michael D. Ward. 2012*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.

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.

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 **?** 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

FLORIAN – CAN YOU ADD THE MATHEMATICAL DETAILS OF CRPS HERE. DONT GET CARRIED AWAY

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 $b_i \equiv |r_i - y_i|$.[15]

Donating the median of some vector $\mathbf{x}$ as $med(\mathbf{x})$, and the standard indicator function as $I(.)$, we define the following heuristic statistics:

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

[15] See **?** for additional discussion of comparative fit metrics.