AUTHOR QUERY FORM

ELSEVIER	**Journal:** International Journal of Forecasting **Article Number:** 4203	**Please e-mail or fax your responses and any corrections to:** **E-mail: corrections.eseo@elsevier.river-valley.com** **Fax: +44 1392 285879**

Dear Author,

Please check your proof carefully and mark all corrections at the appropriate place in the proof (e.g., by using on-screen annotation in the PDF file) or compile them in a separate list. Note: if you opt to annotate the file with software other than Adobe Reader then please also highlight the appropriate place in the PDF file. To ensure fast publication of your paper please return your corrections within 48 hours.

For correction or revision of any artwork, please consult http://www.elsevier.com/artworkinstructions.

Location in article	**Query / Remark click on the Q link to go** **Please insert your reply or correction at the corresponding line in the proof**
Q1	Please confirm that given names and surnames have been identified correctly.
Q2	An extra closing parenthesis in Table 1 is deleted. Please check, and correct if necessary.
Q3	Color statement has been added to the caption(s) of Fig. 3. Please check, and correct if necessary.
Q4	Please update the status of publication for this reference.
	Please check this box or indicate your approval if you have no corrections to make to the PDF file ☐

Thank you for your assistance.

INTFOR: 4203 Model 3Gdc pp. 1–13 (col. fig: NIL)

ARTICLE IN PRESS

International Journal of Forecasting xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Calibrating ensemble forecasting models with sparse data in the social sciences

Jacob M. Montgomery[a,*], Florian M. Hollenbach[b], Michael D. Ward[b]

[a] Department of Political Science, Washington University in St. Louis, Campus Box 1063, One Brookings Drive, St. Louis, MO, 63130-4899, USA

[b] Department of Political Science, Duke University, Perkins Hall 326 Box 90204, Durham, NC, 27707-4330, USA

ARTICLE INFO

Keywords:
Bayesian methods
Election forecasting
Labour market forecasting
Calibration
Ensembles

ABSTRACT

We consider ensemble Bayesian model averaging (EBMA) in the context of small-*n* prediction tasks in the presence of large numbers of component models. With large numbers of observations for calibrating 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. 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, which improves its performance. Specifically, we show that this solution improves the accuracy of EBMA forecasts in predicting the 2012 US presidential election and the US unemployment rate.

1. Introduction

Although producing accurate predictions of future events is not the primary goal in most of the social sciences, recent years have witnessed the spread~~ing~~ of systematic forecasting from more traditional topics, such as GDP growth and unemployment, to many new domains, including elections (e.g., Linzer, 2013), political instability (e.g., Goldstone et al., 2010), and mass killings (Ulfelder, 2012). Several factors have motivated 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, & Bakke, 2010). In addition, the forecasting of important political, economic, and social events is of great interest to policy-makers and the public.

With the proliferation of forecasting efforts, however, comes a need for sensible methods of aggregating and utilizing the various scholarly efforts. One attractive solution to this problem is to combine prediction models and create an ensemble forecast. Combining forecasts reduces the 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 ensemble predictions to be more accurate than any individual component model (Armstrong, 2001; Bates & Granger, 1969; Raftery, Gneiting, Balabdaoui, & Polakowski, 2005).

One approach to combining multiple forecasts is ensemble Bayesian model averaging (EBMA), a method which was first proposed by Raftery et al. (2005) for combining weather forecasts, and was introduced for the social sciences by Montgomery, Hollenbach, and Ward (2012a). EBMA combines multiple forecasts using a finite mixture model that generates a weighted predictive probability

* Corresponding author. Tel.: +1 314 935 9106.
E-mail address: jacob.montgomery@wustl.edu (J.M. Montgomery).

http://dx.doi.org/10.1016/j.ijforecast.2014.08.001
0169-2070/© 2014 Published by Elsevier B.V. on behalf of International Institute of Forecasters.

density function (PDF). EBMA mixture models collate the good parts of existing forecasting models, while avoiding either over-fitting to past observations or over-estimating our level of certainty about forecasts of future events. The hope is that integrating the knowledge and implied uncertainty of a variety of approaches into a combined predictive PDF will result in more accurate and better calibrated forecasts.

In this article, we present several adjustments to the basic EBMA model, as specified by Montgomery et al. (2012a), that can aid applied researchers in creating ensemble forecasts in the presence of the sorts of data-quality challenges that are common in real-world social science settings. Specifically, we show that 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 misweighting of components resulting from either strong or poor performances in the limited calibration period. After discussing the data-quality challenges which are commonly experienced in ensemble forecasting, we introduce the basic EBMA model and outline modifications 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 use the method to predict the 2012 US presidential election and the US unemployment rate.

2. Ensemble prediction with sparse data and multiple forecasts

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

In recent years, the advantages of ensembles have become particularly prominent in the machine-learning and nonparametric statistics community (Hastie, Tibshirani, & 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 & Schapire, 1997; Friedman, 2001), bagging (Breiman, 1996), random forests (Breiman, 2001), and related techniques (e.g., Chipman, George, & McCulloch, 2010) for aggregating so-called "weak learners". These approaches to classification and prediction have been advertised as the "best off-the-shelf classifier[s] in the world" (Zhu, Zou, Rosset, & Hastie, 2009, 350), 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. Not all guesses are equally informative, and naïve approaches to the collation of forecasts risk 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 a weighted combination of predictive PDFs. Rather than selecting some particular "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 that are different from those of other component models).

In recent years, variants of the EBMA method have been applied to subjects as diverse as inflation (e.g., Wright, 2009), stock prices (e.g., Billio, Casarin, Ravazzolo, & Van Dijk, 2011), economic growth and policymaking (e.g., Billio, Casarin, Ravazzolo, & Van Dijk, 2010), and climatology (Min, Simonis, & Hense, 2007). Though EBMA is already in use, research aiming to improve upon the basic model is ongoing. It has, for instance, been adjusted to handle missing data (Fraley, Raftery, & Gneiting, 2010; McCandless, Haupt, & Young, 2011) and incorporate spatial information (Feldman, 2012). Other recent innovations include that of Möller, Lenkoski, and Thorarinsdottir (in press), who use multiple EBMA models predicting univariate outcomes to create joint predictive distributions of multiple (correlated) dependent variables. All in all, the promise of ensemble forecasting via EBMA has led to multiple efforts to refine the method, although this work has taken place primarily outside of the social sciences.

In this article, we focus on difficulties in calibrating accurate EBMA forecasting models in the context of data-quality challenges, which are especially common in (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 the measurement of outcomes is fairly precise and data are abundant. For instance, predicting water surface temperatures in 200 locations across just five days provides 1000 observations from 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 that of dimensionality. Prediction tasks often involve many forecasts predicting a few outcomes, or even just one. For example, in the field of economics, a wide variety of consulting firms, banks, and international organizations provide forecasts for various economic quantities, such as the unemployment rate, GDP growth, and inflation. Indeed, the Federal Open Market Committee (FOMC) of the US Federal Reserve Board generates over a dozen forecasts for key economic indicators.[1]

A final issue is the inconsistency of issuing forecasts. Given the lengthy time periods that are often involved, there are likely to be many missing forecasts in any time window which contains a modestly large number of observations. 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

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

Table 1
Pre-election forecasts of the percentage of the two-party vote that will go to the incumbent party in US 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

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

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.

Predicting US presidential elections is, perhaps, the quintessential forecasting task that combines all of these iss Table 1 lists nearly all scholarly forecasts that produced more than one true out-of-sample forecast for elections in the 20th century prior to the 2012 election.[2] In this instance, we have only five observations on which to calibrate an ensemble model, while we have nine forecasting models.[3] Moreover, several of the individual forecasts are missing for a significant portion of this period. The forecast of Cuzàn, for instance, is missing for 60% of the elections in this dataset.[4]

While these data issues are particularly egregious for US presidential elections, they are endemic to the social sciences and are far from benign. Important events involving economic growth, inflation rates, monetary policy decisions, political elections, and civil conflict are often measured infrequently (annually or quarterly), which fundamentally limits the numbers of observations available, even when they are measured for multiple countries. In addition, these data issues are likely to arise in any instance where predictions are made for a single unit observed over a modest time period, which includes making predictions for trade regimes, supply chains, commodity prices, and more.

As we demonstrate below, calibrating large ensemble models on sparse (and even incomplete) data leads to the misspecification of EBMA model weights and worsened out-of-sample predictive performances. In essence, when there are many forecasting models and few observations on which to evaluate them, the probability that some forecasting model will perform unusually well or poorly due to mere *chance* increases markedly. Under these circumstances, ensemble methods in general, and EBMA models in particular, will mis-estimate model weights: too much weight will be put on poor models that were *randomly* correct, while generally accurate models that were wrong for the small calibration sample will be similarly undervalued.

In light of these difficulties, below we propose several extensions to the baseline EBMA algorithm introduced by Montgomery et al. (2012a), and explore the effects of these modifications on the method's predictive performance. To begin with, we discuss a method proposed by Fraley et al. (2010) for the efficient handling of missing data. Furthermore, we add a simple and intuitive "wisdom of crowds" parameter, which ensures that model weights are distributed more evenly.

3. EBMA for sparse data

As its name suggests, EBMA is descended from the Bayesian model averaging (BMA) methodology (cf., Clyde & George, 2004; Hoeting, Madigan, Raftery, & Volinsky, 1999; Raftery, 1995), which was first introduced to political science by Bartels (1997) and has been applied in a number of contexts (e.g., Imai & King, 2004; Montgomery & Nyhan, 2010). A more detailed discussion of the basic EBMA model extended here by ided in Montgomery et al. (2012a).

3.1. Baseline EBMA model

Assume that the researcher is interested in predicting event $\mathbf{y}^{t^*}$ for some future time period $t^* \in T^*$, termed the test period below. In addition, we have a number of different out-of-sample forecasts for similar events $\mathbf{y}^t$ in some past period $t \in T$, which we term the calibration period. The different predictions were generated from K forecasting models or teams, $M_1, M_2, \ldots, M_K$. As we show in our examples below, these predictions might originate from the insights and intuitions of individual subject-experts, traditional statistical models, non-linear classification trees, neural networks, agent-based models, or anything in between. Indeed, there is no restriction at all on the kind of forecasting method that can be incorporated into the ensemble, so long as it offers a prediction for a sufficiently large subset of the calibration sample.

For each forecast we have a prior probability distribution $M_k \sim \pi(M_k)$, and 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

[2] See, for example: Abramowitz (2008); Campbell (2008b); Cuzán and Bundrick (2004, 2008); Erikson and Wlezien (2008); Fair (2009, 2011); Graefe, Cuzán, Jones, and Armstrong (2010); Hibbs (2012a); Holbrook (2008); and Lockerbie (2008). A recent symposium in *PS: Political Science & Politics* presents and summarizes attempts by a variety of scholars to predict the 2012 US 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, & Ward, 2012b). However, the strength of EBMA is greatest when the model is calibrated on true out-of-sample forecasts, as we do here.

[3] The out-of-sample predictions for these models were collected from the individual journal articles, personal websites and symposia introductions. When multiple forecasts were made, we used the authors' preferred forecast, or took the mean if no preference was given (Abramowitz, 2000, 2012; Campbell, 2000, 2001, 2004, 2005, 2008a, 2012; Campbell & Garand, 2000; Cuzán, 2012; Erikson & Wlezien, 2012; Fair, 2012; Hibbs, 1992, 2000, 2004, 2012b; Holbrook, 1996, 2012; Lewis-Beck & Tien, 1996, 2012; Lockerbie, 2012; Wlezien & Erikson, 1996).

[4] Cuzàn's prediction for 2004 stems from the FISCAL model published prior to the 2004 election by Cuzán and Bundrick (2004), while the 2008 prediction comes from the FPRIME short model presented in advance of the election (Cuzán & Bundrick, 2008). However, the two models are quite similar in their composition.

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 $p(\mathbf{y}^{t^*}) = \sum_{k=1}^{K} p(\mathbf{y}^{t^*}|M_k)p(M_k|\mathbf{y}^t)$. Thus, the prediction via EBMA is a weighted average of the component PDFs, and 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 fit on data from the training period T'. The assumption is that, for each component model, the model-specific parameters (if any) will be estimated based on observations in the training period T'. Each component model then generates out-of-sample predictions for the calibration period T, which represents the time period used to calculate model weights. It is then possible to generate true ensemble out-of-sample forecasts ($\mathbf{f}_k^{t^*}$) for observations in the test period $t^* \in T^*$.

The use of these three distinct time periods makes it possible to calibrate the EBMA model on the component models' out-of-sample predictive power, thus implicitly penalizing overly-complex "garbage can" models. Thus, 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. Model weights are calculated based purely on their out-of-sample predictive performance in the calibration period. (However, as we show below, this same feature leads to mis-specification of the model weights when data are sparse in the calibration period.)

Let $g_k(\mathbf{y}|\mathbf{f}_k^{s|t,t^*})$ represent the predictive PDF of component k, which may be either 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 $p(y|f_1^{t^*}, \ldots, f_K^{t^*}) = \sum_{k=1}^{K} w_k g_k(y|f_k^{t^*})$.[5] 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 across components.[6] 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 $\mathbf{w} = (w_1, \ldots, w_k)$.

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 the model weights $\mathbf{w}$ and the common variance parameter σ^2. We estimate these using maximum likelihood methods (Raftery et al., 2005). The log-likelihood function,

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

cannot be maximized analytically. Instead, we follow Raftery et al. (2005) and use an EM algorithm to calibrate the weights, an approach which is made possible by recognizing the EBMA here as a finite mixture model (McLachlan & Peel, 2000). Specifically, the unobserved quantities z_k^t are introduced, which represent 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

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

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 Eq. (1).

Using the estimates of $z_k^{s|t}$, one can then calculate the maximizing value (the M step) for the component weights 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 common variance term is estimated as

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

The E and M steps are iterated until the improvement in the log-likelihood is no larger than some predefined tolerance level. 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 weight component models incorrectly in a way that worsens the out-of-sample performance due to unusually poor or strong predictive performances in the limited calibration sample. With small samples, random chance alone can lead some specific models to perform quite well, even if their long-term predictive power is quite weak. This is especially true when the short calibration period is combined

[5] 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 leads to a risk of over-fitting and reduced predictive performance. We therefore use a simpler formulation.

[6] The model presented here assumes that the errors are distributed normally and with a common variance term. However, neither of these assumptions is strictly necessary. EBMA has been extended to handle multiple distributional forms, including a zero-inflated gamma, discrete quantitative outcomes, and binary outcomes (Montgomery et al., 2012a; Sloughter, Gneiting, & Raftery, 2010; Sloughter, Raftery, Gneiting, & Fraley, 2007). Likewise, the method allows for the estimation of distinct variance parameters, although this is not usually advantageous for limited samples in practice.

with missing observations in the component model predictions. (Adjustments to the baseline model to accommodate missing components are provided in Appendix A.) To ameliorate these problems, we therefore introduce a "wisdom of crowds" parameter, which serves to distribute model weights more evenly than they would be otherwise. As we show below, in practice this can significantly improve the out-of-sample performance of ensemble models in the context of sparse data.

The "wisdom of crowds" parameter, $c \in [0, 1]$, is introducted to improve the performance of EBMA in the context of sparse data, and reflects our prior belief that all models should receive some weight, but not necessarily be equal. We rescale z_k^t to have a minimum value of $\frac{c}{K}$. This states that there is at least a $\frac{c}{K}$ probability that observation t is represented correctly 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 Eq. (4) above with

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

Conceptually, c resembles a shrinkage or regularization prior in a fully Bayesian setup which serves to prevent over-fitting the available training data. Indeed, in a proper Bayesian framework, several classes of priors could serve the same purpose.[7]

Note that when $c = 1$, all models are considered to be 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 where $c = 1$.[8] Likewise, the EBMA discussed by Montgomery et al. (2012a) 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 the out-of-sample predictive performance in the context of the data-quality challenges that are common to social science applications. In particular, it is designed to address the poor weight calibrations that are common 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 mis-estimation of component model weights and worsen the predictive performance of EBMA.

To justify these claims, we present the results of a simulation study involving our modified EBMA algorithm and two empirical applications of the adjusted method. We begin with a simulation that illustrates the reduced predictive performance of the baseline EBMA model in the circumstances described above, together with the improvements that result from our proposed modification. We then apply our method, first, to the prediction of the 2012 US presidential election, and, second, to the prediction of the US unemployment rate.[9]

4.1. Simulation study

In this section, we conduct a simulation study of the adjusted EBMA model proposed above. This study serve two purposes. First, it demonstrates the challenges that are inherent in ensemble forecasting when the calibration samples are small and the number of forecasting models is large.[10] Second, it explores the extent to which our modified EBMA algorithm ameliorates these difficulties. In addition, we also 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 from precisely the mixture distribution shown in Eq. (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 weights 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 $K = 3$.[11] This ensures that the model weights always sum to 1, but that there is still some heterogeneity. We varied the size of the calibration sample (n_T), the number of component forecasts (K), and the wisdom of crowds parameter (c). The c parameter is used only for model estimation, and plays no role in the creation of the simulated data itself.

For each simulation, we generate component forecasts for both the calibration and test periods. We fit an EBMA model as specified above to the calibration sample data only, 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.

7 A helpful reviewer noted that a symmetric Dirichlet prior on the model weights might serve the same purpose, but may prove to be more general. While we believe that this approach appears plausible, we leave this proposed variant to future research.

8 The mean or median would be equivalent, depending on whether the posterior mean or median is used to make a point prediction.

9 All files needed for these analyses will be made available online at the time of publication. Our modified EBMA algorithm will be released as part of the R package 'EBMAforecast', which is hosted online with the Comprehensive R Archiving Network (CRAN) (Montgomery, Hollenbach, & Ward, 2013).

10 To reduce the parameter space for these simulations, we limit ourselves here to exploring the roles of calibration sample sizes and numbers of component forecasts. We do not consider issues of missingness.

11 Implicitly, this simulation framework states that the first model will generally be the most "correct", as it has the highest concentration parameter of 10. In models where $K > 3$, on the other hand, the additional models will become increasingly uninformative about the outcome as their weight is decreased. However, as a helpful reviewer noted, adding hundreds of models that essentially *never* provide correct answers would lead to a different simulation outcome. In this hypothetical case, where many models are by definition inappropriate and uninformative, giving more models more weight will provide an inferior solution.

Table 2
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

We begin by examining the accuracy of the baseline EBMA ($c = 0$) predictive PDF shown in Eq. (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.[12] The CRPS has been used widely to evaluate forecasts of continuous outcomes, and its many advantages as a proper scoring rule have been discussed elsewhere (Brandt, Freeman, & Schrodt, 2011; Gneiting & Raftery, 2007; Hersbach, 2000). 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 n_{T*}, with smaller numbers indicating better forecast performances.[13] (The mathematical details behind the CRPS can be found in Appendix B.)

Fig. 1 shows the out-of-sample performance of the EBMA method, as measured by CRPS, against the ratio of the number of component models included to the size of the calibration period (i.e., $\frac{K}{n_T}$). As one can see, the performance of the EBMA model depends significantly on this ratio. As the number of component models included increases or the calibration sample size decreases, CRPS rises; that is, the predictive performance of the ensembles decreases as a function of $\frac{K}{n_T}$. Note that, in this figure, $c = 0$ for all models.

Thinking of each parameter in isolation, Fig. 1 shows that 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 estimated correctly in order to make accurate predictions increases, the quality of the forecast goes down. Furthermore, CRPS is a decreasing function of n_T; i.e., the performance of the baseline EBMA model improves as the calibration sample grows.[14]

The remaining question, then, is: to what extent does adding the "wisdom of crowds" parameter to the baseline model improve the performance? To answer this question, we examined the out-of-sample predictive performance of EBMA for differing values of c. Fig. 2 shows a smoothed plot of the median CRPS for differing values of the ratio $\frac{K}{n_T}$ and c in our simulation. Darker gray tones indicate higher CRPS levels. The plot 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 when $\frac{K}{n_T}$ is large.

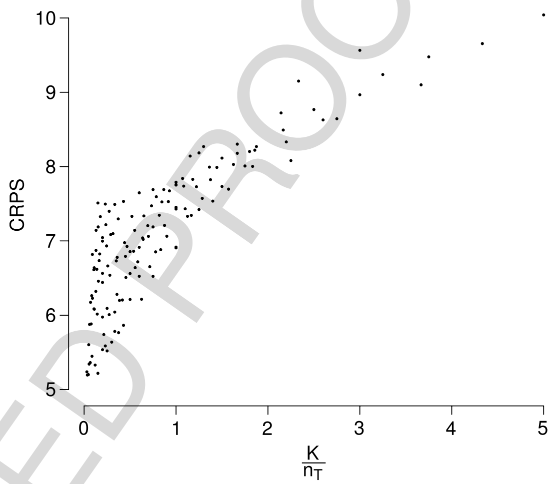

Fig. 1. CRPS with varying numbers of models and observations in the calibration period. See Table 2 for a full list of the parameter values considered in our simulations. These results represent all simulations where $c = 0$, which is the baseline EBMA model. Note that the EBMA model performs significantly less well when the ratio of component forecasts to the sample size in the calibration period $\frac{K}{n_T}$ is high.

There are two aspects of Fig. 2 that are particularly salient. First, note that the addition of c to the base EBMA model does not uniformly improve the out-of-sample performance. When the calibration sample is modestly large and there are few models to calibrate, the addition of c uniformly worsens the model performance. However, with small calibration samples and modestly large numbers of component models, the addition of c improves the 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. While this is far from being a complete analysis of the simulated data, it does serve the limited purpose of demonstrating that, in some circumstances, the "wisdom of crowds" parameter aids prediction.

Generally speaking, in choosing c, the researcher faces a trade-off between the possible overweighting of certain models based on their performances in short calibration periods and moving too far in the direction of a simple average. In general, a relatively small c seems to improve the

12 Note that it is only possible to use the CRPS because we are comparing multiple EBMA models, for which we have the entire predictive PDF. For the applied examples below, we only have point estimates from the component models, and cannot evaluate their individual performances with CRPS.

13 Technically, it ranges from 0 to 1 for each of the n_{T*} observations in the test sample.

14 Examining each parameter in isolation supports this claim, although the relationship is highly interactive (results not shown).

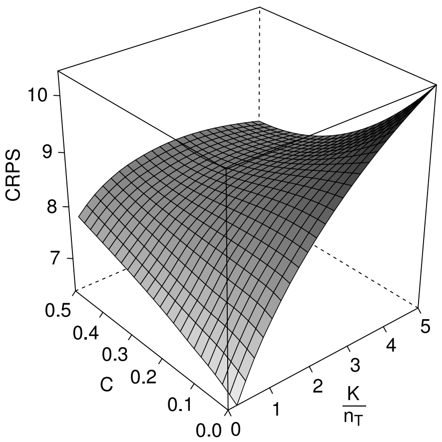

Fig. 2. CRPS with the "wisdom of crowds" parameter. The figure shows a smoothed relationship between the median CRPS for differing values of the ratio $\frac{K}{n_T}$ and c. Darker gray tones indicate higher CRPS levels. Note that while CRPS generally still increases with higher values of $\frac{K}{n_T}$, including a "wisdom of crowds" parameter within the EBMA algorithm can help when $\frac{K}{n_T}$ is large.

performance. Based on our examination of the broader set of simulations, we generally recommend the selection of values of $c \in [0, 0.1]$.[15] Higher settings of c will be preferred as the ratio $\frac{K}{n_T}$ goes up. The simulations favor smaller values of c as the ratio of K to n_T decreases. Our experience, based on the simulations, as well as on cross-validation studies during the applications presented below, suggests that a c value of 0.05 is a good default choice, although the "best" value of c may vary from application to application. Many researchers may consider it preferable 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 an examination of the way in which these methods work in two areas that exemplify forecasting in the social sciences. The first is predicting the vote for the incumbent-party candidate in United States presidential elections, and the second is the prediction of unemployment in the United States. Both areas have well developed forecasting traditions in the scholarly and policy communities.

4.2. United States presidential elections

We now turn to the task of combining expert predictions of presidential elections.[16] 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.

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 3. 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 given such a low weight because out-of-sample predictions were only available for 2004 and 2008.

Table 3
Model weights and in-sample fit statistics for the EBMA model of US 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

Fig. 3 shows the posterior predictive distribution for the 2008 election (top) and, based on the forecasts from each of the component models, the 2012 election. The component models' predictive distributions are shown in color (scaled by their respective weights), while the EBMA predictive distribution is shown in black. The bold dashed line displays the point prediction for the EBMA model. The vertical dotted line depicts the actual election results in 2008 and 2012.[17]

The EBMA model ($c = 0.05$) predicted Obama receiving 50.3% of the two-party vote share in 2012. This resulted in a reasonably large absolute error of 1.7%.[18] Fig. 3 shows that the EBMA prediction performs better than the majority of component forecasts for the 2012 election, with only three providing more accurate estimates. However, we believe that it is important to note that 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 relative to other methods of aggregating forecasts for the 2012 election. Table 4 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 $c = 0.1$. While the differences in model weights are relatively small, the EBMA predictions for 2012 were 49% and 50.1% for the EBMA models with $c = 0$ and $c = 0.1$ respectively—both of which are considerably worse than the prediction with $c = 0.05$.

The EBMA model with a "wisdom of crowds" parameter of 0.05 also did considerably better than naïve approaches to aggregation. The mean of the component models' predictions was 49.9%, and the median prediction was 49.5%. In essence, while we believe that prediction methods should be evaluated on more than one observation, this

15 As we show in our empirical examples below, even very small settings of c can improve out-of-sample predictive performance when there are many models.

16 See also Montgomery et al. (2012b).

17 We emphasize again that these are the predictive PDFs generated, not by the component models, according to $g_k(y|f_k^{t*}) = N(f_k^{t*}, \sigma^2)$ in Eq. (1).

18 The EBMA model assigned a considerable amount of weight to models predicting a Romney *victory*, especially the models of Hibbs (2012b) and Lewis-Beck and Tien (2012).

Please cite this article in press as: Montgomery, J. M., et al., Calibrating ensemble forecasting models with sparse data in the social sciences. International Journal of Forecasting (2014), http://dx.doi.org/10.1016/j.ijforecast.2014.08.001

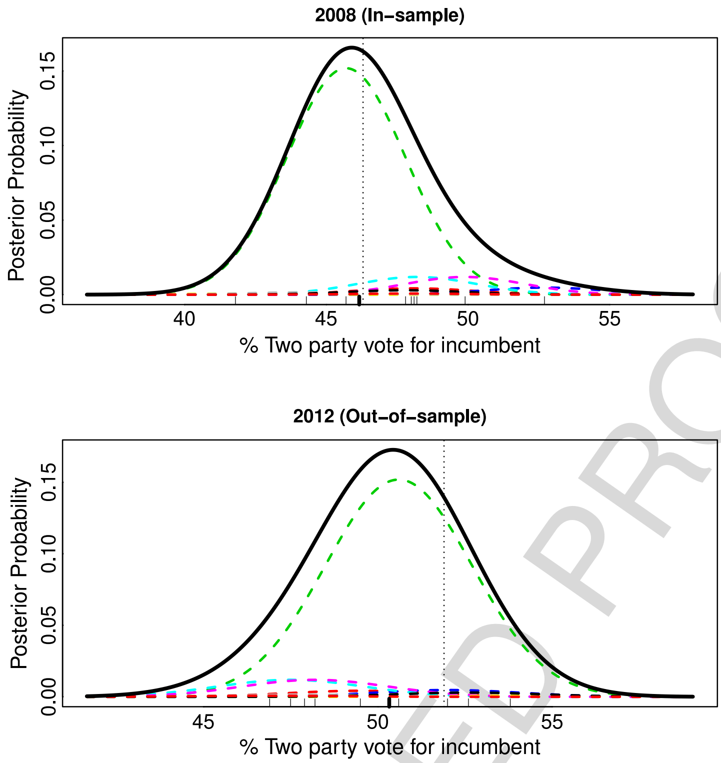

Fig. 3. Predictive ensemble PDFs of incumbent-part vote shares in US presidential elections. The figure shows the density functions for each of the component models in different colors and scaled by their respective weights. The black curve is the density of the EBMA prediction, with the bold dashed line indicating the EBMA point prediction. The vertical dashed lines show the actual results of the 2008 and 2012 elections. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 4
Comparing model results for US 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

example again indicates the utility of EBMA in forecasting tasks and, in particular, the improvements that the "wisdom of crowds" parameter offer for out-of-sample predictions in the context of sparse data.

4.3. Quarterly unemployment in the United States

Forecasting macroeconomic variables is a quite common exercise in the fields of economics and statistics. Policymakers and businesses are both enormously interested in the calculation of accurate forecasts of economic variables. These forecasts are often created using a wide variety of statistical models; however, professional forecasts are often based only on expert knowledge.[19] 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 typical forecasting models has increased considerably over time. In particular, nonlinear dynamic models have gained prominence, including threshold autoregressive models, Markov switching autoregressive models and smooth transition autoregressions (e.g., Elliott & Timmermann, 2008). More recently, forecasters have added Bayesian VAR models and state space models to their arsenal (De Gooijer & Hyndman, 2006).

Unsurprisingly, given the large number of forecasts being produced on an ongoing basis, scholars have attempted to improve the predictive accuracy by combining forecasts (e.g., Bates & Granger, 1969; Elliott & Timmermann, 2008). Recently, EBMA and related Bayesian model averaging methods have been employed successfully for creating ensemble forecasts of various macroeconomic indicators, such as inflation and exchange rates (e.g., Wright, 2008, 2009).

Policymakers have also 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*

19 For a more comprehensive overview of the forecasting of economic variables, see De Gooijer and Hyndman (2006) and Elliott and Timmermann (2008). See also Brandt (2012) and Brandt and Williams (2007) for work on forecasting multiple time series.

Please cite this article in press as: Montgomery, J. M., et al., Calibrating ensemble forecasting models with sparse data in the social sciences. International Journal of Forecasting (2014), http://dx.doi.org/10.1016/j.ijforecast.2014.08.001

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 United States, including the unemployment rate, inflation, and GDP growth.[20] In the first month of every quarter, a survey is sent to selected forecasters, who return it 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. A significant amount of attention is given to the average (usually the median) reported forecast.

This plethora of predictions seems ideal for applying EBMA. Nonetheless, it is plagued by the issues 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 good test bed for the adjusted EBMA model presented 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., the 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, the average number of teams making a prediction in any quarter for four quarters into the future is quite small, and the majority of the observations for any given quarter are missing.[21]

To provide a meaningful benchmark for our adjusted EBMA model, we also include in the ensemble 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).[22]

Taking the *SPF* and Green Book unemployment forecasts, we calibrate an ensemble model for each period t, using forecaster performance over the last ten quarters. Only forecasts that had made predictions for at least five of these ten quarters were included in the ensemble. Thus, the EBMA model uses only 144 models out of a possible 293 forecasting models that made predictions during the period studied. 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 until the fourth quarter of 2007.

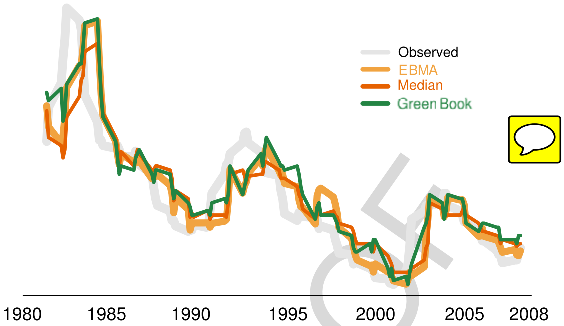

Fig. 4. Observed and forecasted US unemployment (1981–2007).

One approach to evaluating the performance of EBMA is to compare its predictive accuracy to those from 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.[23]

Fig. 4 shows a visual representation of the Green Book, median SPF and EBMA (with $c = 0.05$) forecasts over time, together with the true unemployment rate. As was noted above, and as is clearly visible, the SPF and Green Book forecasts are quite similar. Baghestani (2008) states that the Green Book forecast is slightly biased toward overpredicting the unemployment rate. In some periods EBMA is able to correct this bias; however, given the similarity of the component models, the improvement in that direction is rather small. In general, however, it is clearly visible that the EBMA forecast is closer to the actual rate than the median SPF or the Green Book forecast.

Table 5 formally compares these models to EBMA models with $c = 0, 0.05, 0.1$, and 1 respectively. To do this, we focus on eight model fit indices that are available in the literature (Brandt et al., 2011)[24]: the 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 naïve model, which simply predicts the future rate of unemployment as being the same as the current rate of unemployment. Further details for these metrics are shown in Appendix B.

The values in bold in each column of Table 5 indicate the model that performed "best" as measured by each metric. With one exception (the Green Book outperforms the ensemble by 0.01 with the RMSE), the EBMA model outperforms both the Green Book forecasts and the unweighted

20 The *SPF* was first administered in 1968 by the American Statistical Association and the National Bureau of Economic Research (NBER). Since 1990, however, it has been being run by the Federal Reserve Bank of Philadelphia. http://www.phil.frb.org/research-and-data/real-time-center/survey-of-professional-forecasters/.

21 On average, only 8.4% of all teams make a forecast in any one quarter.

22 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 revisions when better data become available. Thus, when evaluating forecasts, it is important to know whether the predictions are being compared to the outcome data for each quarter available at the time, or whether the revised and most recent data are 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. Here, we have decided 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.

23 Note that the EBMA model is calculated on only the subset of forecasts that have made enough recent predictions to enable the calibration of model weights. Thus, the median forecast and the ensemble forecast will not be the same even when $c = 1$.

24 We are unable to use CRPS here because we do not have the predictive PDFs for the Green Book, SPF mean, or SPF median models.

Table 5
Comparing adjusted EBMA models with Green Book, median, and mean forecasts of US 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

Note: Definitions of the model fit statistics are provided in the Appendix. The model(s) with the lowest scores for each metric are shown in bold. Differences between model performances may not be obvious due to rounding.

mean and median forecasts 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 equal best) on six of these eight metrics.

We now turn to the evaluation of the performance of the ensemble relative to its 144 component forecasts. It is important to note that many of these forecasters only 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 only calculate these fit indices for $n_k \forall k \in [1, K]$. According to this measure, the EBMA model performs very well. Table 6 provides a summary of these results. The rows of the table show the numbers of metrics according to which EBMA outperforms the component models, while the columns show the numbers of forecasts made by these models. The values in each cell of the table are the proportions of component models in each column that fall into each row (the columns each sum to unity). For instance, the top-left cell represents cases where EBMA is better on at least seven out of eight metrics for component models which make between one and ten predictions. Approximately 60% of the models that make 5–~~10~~ predictions fall into this category.

Notably, the superiority of EBMA to its components is experienced 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. In addition, when the number of forecasts is low, it is likely that a given model will receive less weight than it "deserves" given the model's performance.[25] However, Table 6 shows that, across a large number of forecasts, EBMA significantly outperforms its components. That is, when moving to the right in the table, the values in the lower cells decrease. It is also worth noting that only six of the total of 144 components outperform EBMA on every metric.

Table 6
Comparing the predictive accuraciwa of EBMA and component models with eight metrics.

Number of metrics on which EBMA performed better	Number of predictions made			
	5–10	11–30	31–60	>60
7–8	0.59	0.74	0.68	0.75
5–6	0.12	0.09	0.18	0.25
2–4	0.06	0.08	0.05	0.00
0–1	0.22	0.09	0.09	0.00
Number of components	49	65	22	8

The rows of the table show the numbers of metrics according to which EBMA outperforms the components, while columns show the numbers of forecasts made by these models. The values in each cell of the table are the proportion of component models which fall into that category (the columns sum to unity). Note that EBMA performs very well against its components, especially those that make many predictions.

5. Discussion

Ensemble Bayesian model averaging is a principled way of combining forecasts ~~to~~ in order to improve the prediction accuracy. However, the calibration of such models in the social sciences is often hindered by issues with the quality and availability of data. First, the number of forecasting models in many forecasting exercises is large, yet the number of observations on which the EBMA model can be calibrated is small. This creates problems for the estimation of model weights, as it is likely that overly-high weights will be assigned to models that perform 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 may be different for different models. In the standard EBMA model introduced by Montgomery et al. (2012a), missing observations in component model predictions are not allowed.

In this article, we address both of these issues in order to make EBMA more applicable for researchers and those making predictions in the social sciences. After reviewing the standard EBMA framework, we proceeded to introduce a "wisdom of crowds" parameter into the model, which forces EBMA to put some given minimal weight on all component models. Adding this constant assists in the calibration of EBMA when the number of observations in the calibration period is small.

After explaining our adjustments, we illustrated the advantages of the adjusted model via simulations. We then applied the modified EBMA model in two prediction exercises. We used the out-of-sample forecasts of nine prediction models of presidential elections from 1992 to 2008 to calibrate an ensemble model, then used this model to make an informed prediction for the 2012 elections based on a weighted combination of the component predictions. This example neatly illustrates the common difficulties facing forecasters in the social sciences, and provides an illustrative example for applied researchers in future. In a second

25 See Appendix A for a discussion of how EBMA handles missing component forecasts.

example, we used EBMA to combine predictions of the unemployment rate in the US from the Survey of Professional Forecasters and the Green Book. As we have shown, even when large numbers of forecasts are missing for any given quarter, EBMA generally outperforms the Green Book, SPF component models, and the median and mean SPF forecasts.

A comprehensive approach to the data problems raised in Section 2 would be to estimate the "wisdom of crowds" parameter within the EBMA algorithm specifically for each forecasting application. So far we have refrained from doing this, as we are concerned about the numbers of parameters being estimated on relatively small numbers of observations (i.e., limited degrees of freedom). In addition, simple solutions have failed so far, because of issues of identifiability. Future extensions of this model should aim to adjust the EBMA algorithm further so as to make the estimation of c possible.

In addition, future research should investigate alternative imputation techniques within the EBMA algorithm for handling missing data. The approach presented in Appendix A follows Fraley et al. (2010). This is an improvement, in that it allows for the inclusion of models with missing predictions, but at the same time components with missingness are severely down-weighted. Moreover, the algorithm shown in Appendix A was developed in the context of meteorological sciences, where weather stations may fail to report observations randomly. In the social sciences, however, different approaches may be more appropriate. One possible direction would be to implement the imputation of missing observations within the EBMA framework via copula methods.

Ensembles are a useful approach to the aggregation of predictive information. Even a decade ago there was not a lot of predictive information which needed aggregation in the social sciences. However, with the advent of newer approaches, more widely available data, and predictive heuristics, the arena of predictive forecasts has expanded considerably. The EBMA approach is now not only feasible in the social sciences, but also increasingly advantageous.

Appendix A. EM-algorithm for missing data

To accommodate missing values within the EBMA procedure, we follow Fraley et al. (2010) and modify the EM algorithm as follows. Define

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

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

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

The M steps in Eqs. (4) and (5) are likewise replaced with

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

and

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

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 US 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:

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

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.[26]

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 predictions from the naïve forecast, r_i, that serves as a baseline for comparison. In the example in the main text, this naïve model is simply the lagged observation. We can therefore define $b_i \equiv |r_i - y_i|$.[27]

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

$$\text{MAE} = \frac{\sum_1^n e_i}{n}$$

$$\text{RMSE} = \sqrt{\frac{\sum_1^n e_i^2}{n}}$$

$$\text{MAD} = \text{med}(\mathbf{e})$$

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

27 See Brandt et al. (2011) for an additional discussion of comparative fit metrics.

$$\text{RMSLE} = \sqrt{\frac{\sum_{1}^{n} (\ln(f_i + 1) - \ln(y_i + 1))^2}{n}}$$

$$\text{MAPE} = \frac{\sum_{1}^{n} a_i}{n}$$

$$\text{MEAPE} = \text{med}(\mathbf{a})$$

$$\text{MRAE} = \text{med}\left(\frac{e_1}{b_1}, \ldots, \frac{e_n}{b_n}\right)$$

$$\text{PW} = \frac{\sum_{1}^{n} I(e_i > b_i)}{n} \times 100.$$

References

Abramowitz, A. I. (2000). Bill and Al's excellent adventure: forecasting the 1996 presidential election. In J. E. Campbell, & J. C. Garand (Eds.), *Before the vote: forecasting American national elections* (pp. 47–56). Thousand Oaks: Sage Publications, (Chapter 2).

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

Abramowitz, A. 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. S. (2001). Combining forecasts. In J. S. Armstrong (Ed.), *Principles of forecasting: a handbook for researchers and practitioners*. Norwell, MA: Kluwer Academic Publishers.

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

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

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

Billio, M., Casarin, R., Ravazzolo, F., & Van Dijk, H. K. (2010). Combining predictive densities using Bayesian filtering with applications to US economics data. Norges Bank working paper. http://ssrn.com/abstract=1735421 (accessed June 1, 2011).

Billio, M., Casarin, R., Ravazzolo, F., & Van Dijk, H. K. (2011). Bayesian combinations of stock price predictions with an application to the Amsterdam exchange index. Tinbergen institute discussion paper no. 2011-082/4. http://www.tinbergen.nl/discussionpapers/11082.pdf (accessed June 1, 2011).

Brandt, P. T. (2012). MSBVAR: Bayesian estimators and inferences for VAR models. Comprehensive R-Archive Network.

Brandt, P. T., Freeman, J. R., & Schrodt, P. A. (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).

Brandt, P. T., & Williams, J. T. (2007). *Multiple time series models*. Thousand Oaks, CA: Sage Publications.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *26*, 123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

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

Campbell, J. E. (2000). Polls and votes: the trial-heat presidential election forecasting model, certainty, and political campaigns. In J. E. Campbell, & J. C. Garand (Eds.), *Before the vote: forecasting American national elections* (pp. 17–46). Thousand Oaks: Sage Publications, (Chapter 1).

Campbell, J. E. (2001). Taking stock of the forecasts of the 2000 presidential election. *American Politics Research*, *29*(3), 275–278.

Campbell, J. E. (2004). Introduction—the 2004 presidential election forecasts. *PS: Political Science & Politics*, *37*(4), 733–735.

Campbell, J. E. (2005). Introduction—assessments of the 2004 presidential vote forecasts. *PS: Political Science & Politics*, *38*(1), 23–24.

Campbell, J. E. (2008a). Editor's introduction: forecasting the 2008 national elections. *PS: Political Science & Politics*, *41*(4), 679–681.

Campbell, J. E. (2008b). The trial-heat forecast of the 2008 presidential vote: performance and value considerations in an open-seat election. *PS: Political Science & Politics*, *41*(4), 697–701.

Campbell, J. E. (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, J. E., & Garand, J. C. (2000). *Before the vote: forecasting American national elections*. Thousand Oaks: Sage Publications.

Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, *4*(1), 266–298.

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

Croushore, D., & Stark, T. (2001). A real-time data set for macroeconomists. *Journal of Econometrics*, *105*(1), 111–130.

Cuzán, A. G. (2012). Forecasting the 2012 presidential election with the fiscal model. *PS: Political Science & Politics*, *45*(4), 648–650.

Cuzán, A. G., & Bundrick, C. M. (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.

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

De Gooijer, J. G., & Hyndman, R. J. (2006). 25 years of time series forecasting. *International Journal of Forecasting*, *22*(3), 443–473.

Elliott, G., & Timmermann, A. (2008). Economic forecasting. *Journal of Economic Literature*, *46*(1), 3–56.

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

Erikson, R. S., & Wlezien, C. (2012). The objective and subjective economy and the presidential vote. *PS: Political Science & Politics*, *45*(4), 620–624.

Fair, R. C. (2009). Presidential and congressional vote-share equations. *American Journal of Political Science*, *53*(1), 55–72.

Fair, R. C. (2011). Vote-share equations: November 2010 update. Working paper, Yale University. http://fairmodel.econ.yale.edu/vote2012/index2.htm (accessed March 07, 2011).

Fair, R. C. (2012). Personal website. http://fairmodel.econ.yale.edu/ (accessed January 27, 2013).

Feldman, K. (2012). *Statistical postprocessing of ensemble forecasts for temperature: the importance of spatial modeling*. (Master's thesis), Ruprecht-Karls-Universitat Heidelberg.

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

Freund, Y., & Schapire, R. (1997). A decision–theoretic generalization of online learning and an application to boosting. *Journal of Computer and System Sciences*, *55*, 119–139.

Friedman, J. (2001). Greedy function approximation: a gradient boosing machine. *Annals of Statistics*, *29*, 1189–1232.

Galton, F. (1907). Vox populi. *Nature*, *75*(1949), 450–451.

Gneiting, T., Balabdaoui, F., & Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *69*(2), 243–268.

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359–378.

Goldstone, J. A., Bates, R. H., Epstein, D. L., Gurr, T. R., Lustik, M. B., Marshall, M. G., et al. (2010). A global model for forecasting political instability. *American Journal of Political Science*, *54*(1), 190–208.

Graefe, A., Cuzán, A. G., Jones, R. J., & Armstrong, J. S. (2010). Combining forecasts for US presidential elections: the PollyVote. Working paper. http://dl.dropbox.com/u/3662406/Articles/Graefe_et_al_Combining.pdf (accessed May 15, 2011).

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. New York, NY: Springer.

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, *15*(5), 559–570.

Hibbs, D. A. (1992). The 1992 presidential election: Clinton is the probable winner, but only narrowly, getting just over 51% of the two-party vote. http://www.douglas-hibbs.com/Elections2004-00-96-92/forecast92.pdf (accessed January 27, 2013).

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

Hibbs, D. 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 January 27, 2013).

Hibbs, D. A. (2012a). 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, D. A. (2012b). Obama's reelection prospects under bread and peace voting in the 2012 US presidential election. *PS: Political Science & Politics*, *45*(4), 635–639.

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

Holbrook, T. M. (1996). Reading the political tea leaves: a forecasting model of contemporary presidential elections. *American Politics Research*, *24*(4), 506–519.

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

Holbrook, T. M. (2012). Incumbency, national conditions, and the 2012 presidential election. *PS: Political Science & Politics*, *45*(4), 640–643.

Imai, K., & King, G. (2004). Did illegal overseas absentee ballots decide the 2000 US presidential election? *Perspectives on Politics*, *2*(3), 537–549.

Lewis-Beck, M. S., & Tien, C. (1996). The future in forecasting: prospective presidential models. *American Politics Research*, *24*(4), 468–491.

Lewis-Beck, M. S., & Tien, C. (2012). Election forecasting for turbulent times. *PS: Political Science & Politics*, *45*(4), 625–629.

Linzer, D. (2013). Dynamic Bayesian forecasting of presidential elections in the states. *Journal of the American Statistical Association*, *108*(501), 124–134.

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

Lockerbie, B. (2012). Economic expectations and election outcomes: the presidency and the house in 2012. *PS: Political Science & Politics*, *45*(4), 644–647.

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

McLachlan, G., & Peel, D. (2000). *Finite mixture models*. New York, NY: John Wiley & Sons, Ltd..

Min, S.-K., Simonis, D., & Hense, A. (2007). Probabilistic climate change predictions applying Bayesian model averaging. *Philosophical Transactions of the Royal Society of London A (Mathematical and Physical Sciences)*, *365*(1857), 2103–2116.

Möller, A., Lenkoski, A., & Thorarinsdottir, T. L. (2013). Multivariate probabilistic forecasting using ensemble Bayesian model averaging and copulas. *Quarterly Journal of the Royal Meteorological Society*, (in press).

Montgomery, J. M., Hollenbach, F., & Ward, M. D. (2012a). Improving predictions using ensemble Bayesian model averaging. *Political Analysis*, *20*(3), 271–291.

Montgomery, J. M., Hollenbach, F., & Ward, M. D. (2012b). Ensemble predictions of the 2012 US presidential election. *PS: Political Science & Politics*, *45*(4), 651–654.

Montgomery, J. M., Hollenbach, F., & Ward, M. D. (2013). Ensemble BMA forecasting. http://cran.r-project.org/web/packages/EBMAforecast/EBMAforecast.pdf.

Montgomery, J. M., & Nyhan, B. (2010). Bayesian model averaging: theoretical developments and practical applications. *Political Analysis*, *18*(2), 245–270.

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

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

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

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

Surowiecki, J. (2004). *The wisdom of crowds: why the many are smarter than the few and how collective wisdom shapes business, economies, society, and nations*. New York: Doubleday.

Ulfelder, J. (2012). Forecasting onset of mass killings. *Paper presented at the annual northeast political methodology meeting at New York University*.

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

Wlezien, C., & Erikson, R. S. (1996). Temporal horizons and presidential election forecasts. *American Politics Research*, *24*(4), 492–505.

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

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

Zhu, J., Zou, H., Rosset, S., & Hastie, T. (2009). Multi-class adaboost. *Statistics and its Interface*, *2*, 349–360.

Please cite this article in press as: Montgomery, J. M., et al., Calibrating ensemble forecasting models with sparse data in the social sciences. International Journal of Forecasting (2014), http://dx.doi.org/10.1016/j.ijforecast.2014.08.001