Combining forecasts: An application to U.S. Presidential Elections

Andreas Graefe, Karlsruhe Institute of Technology
J. Scott Armstrong, The Wharton School, University of Pennsylvania
Randall J. Jones, Jr., University of Central Oklahoma
Alfred G. Cuzán, University of West Florida

March 16, 2010

Abstract. Prior research suggested that accuracy gains from combining forecasts are particularly high if one uses forecasts from different methods that draw upon different data. We tested this assumption by combining forecasts from three component methods (polls, econometric models, and experts) used for predicting the five U.S. presidential elections between 1992 and 2008. The gains from combining were substantially larger than the 12% reduction obtained from a prior meta-analysis. Mean error reduction from combining within component methods ranged from 18% to 21%. Combining across component methods yielded further error reductions, ranging from 37% to 40%. Compared to the typical component forecast, combining reduced forecast error by 40% and performed about as well as the best component forecast. Average error reduction compared to the worst component forecast was 59%. Compared to the typical uncombined individual forecast, our combining procedure yielded error reductions ranging from 42% to 50%. Combining is probably the most cost efficient way to improve forecast accuracy and to prevent large errors.

Acknowledgments. We thank all the authors that replied to our request to verify whether we cited them correctly.

It is well known that combining forecasts improves accuracy. The error derived from a simple average of a number of individual forecasts will always be at least as accurate as the error of a randomly drawn individual forecast (i.e., the typical forecast). If the true value lies in-between a range of individual forecasts – a situation often referred to as *bracketing* – the simple average will always be more accurate than the typical forecast. Error reductions through combining are particularly high if correlations among the errors of the individual forecasts are low and if the individual forecasts are about equally accurate.

Combining applies to all types of estimation problems. Imagine two people estimating Beethoven's year of birth, which was 1770. If one guesses 1750 and the other one 1740, they yield errors of 20 and 30, respectively. Thus, the typical error of an individual estimate is 25. Since no bracketing occurs, the typical individual estimate is as accurate as the average of both estimates, which is 1745. Now, imagine in a situation in which both individual estimates are equally accurate but lie on either side of the truth. If one guesses 1750 and the other one 1800, they again yield errors of 20 and 30, and a typical error of 25. However, due to bracketing, the average estimate (i.e., 1775) misses the true value only by 5 – an error reduction of 80% compared to the typical individual estimate.

Combining forecasts can help even when one knows in advance which method is most accurate. Again, this is demonstrable in the case of bracketing. Suppose that method A estimated that Beethoven was born in 1760, and method B that he was born in 1799. Although incurring an error three times larger than method A, averaging method B with method A yields a reduction in error. Many empirical studies have found that bracketing is common. Armstrong (2001) conducted a meta-analysis of 30 studies and estimated that, on average, the combined forecast yielded a 12% error reduction compared to the error of the typical forecast; the reductions of forecast error ranged from 3 to 24%. In addition, the combined forecasts were often more accurate than the most accurate component.

Combining is applicable to almost all situations in the management sciences. The only exceptions would be where one has strong evidence on which method is best and where the likelihood of bracketing is very low. However, in most real world situations, analysts do not know about the accuracy of individual sources at the time they make a forecast. Little work has been done to examine *ex ante* conditions for when combining is most beneficial.

Despite its usefulness and ease of use, combining is not used as often as it should in practice. We briefly describe reasons that hinder the application of combining. Then we show that, under nearly ideal conditions, the benefits of combing are large, much larger than previously estimated.

Barriers to the use of combining in practice

Lack of knowledge about the research on combining remains as a major barrier for the use of combining in practice.

People do not believe combining helps

Larrick and Soll (2006) showed for a range of conditions that many people did not understand the benefits of combining forecasts. In one experiment with 145 MBA students (Exp. 1), the majority of participants (57%) thought that the combined forecasts from two judges would perform no better than the average judge – which is the worst possible performance of combining. Of these participants, the vast majority (95%) thought that the combined forecast would perform equal to the average judge. Similar results were obtained from a second experiment with 263 MBA students, in which estimates from two judges were presented either sequentially or simultaneously (Exp. 2): 84% of the participants who believed that combining would not outperform the average judge expected the combined forecast to perform equal to the average judge. However, participants showed some ability to recognize the benefits of combining if the judges' estimates were presented simultaneously. In a third experiment, the authors presented participants with contextual – instead of numerical – information about judges' estimates and accuracy (Exp 3; N=149 participants). Although contextual information helped participants to identify occurrences of bracketing, still almost half of them (45%) expected combining to perform equal (32%) or worse (13%) than the average judge.

People are persuaded by complexity

Hogarth (in press) reported results from four case studies showing that simple models can predict complex problems better than more complex ones: (1) simple statistical aggregation of available data is superior to judgment of clinical psychologists; (2) simple models (such as extrapolation) are more accurate than statistically sophisticated models when making out-of-sample predictions; (3) weighting variables equally often yields more accurate predictions than estimating optimal weights from historical data; (4) decisions can sometimes be improved by ignoring relevant information. For each case, Hogarth described how the findings were met with skepticism by other researchers. People had difficulties to accept the idea that simple methods can outperform more complex ones. There was a strong belief that complex models are necessary to solve complex problems. Similarly, people might perceive the principle of combining as "too easy to be true".

Extreme forecasts build reputation and attract attention

In analyzing GDP and inflation forecasts for the G7 economies within 1990 and 2005, Batchelor (2007) found systematic biases in the forecasts of private sector forecasters. The number of available forecasts ranged from 14 to 30 for each country. While the accuracy of individual forecasters did not differ, forecasters showed persistent individual biases towards either optimism or pessimism, even for short forecast horizons. Batchelor (2007) concluded that forecasters use such a strategy to differentiate their product and to build a reputation as optimists or pessimists. Another reason might be that extreme forecasts usually gain more attention and the media is more likely to report them. Forecasters do not want their individual forecasts to get lost in the crowd.

Combining does not allow forecasters to select forecasts that suit their biases

Based on the findings from his meta-analysis of 30 studies, Armstrong (2001) recommended combining forecasts mechanically, according to a predetermined procedure. A general rule is to weight forecasts equally, unless there is strong prior evidence that supports differential weights. In practice, managers often use unaided judgment to assign differential weights to individual forecasts. Such an *informal* approach to combining is likely to be harmful as managers can select a forecast that suits their biases. For example, 70 % of the 96 respondents in a survey of sales forecasting practices in U.S. corporations said that they preferred to underforecast sales and 15 % said they preferred to overforecast; the rest expressed no preference (Sanders & Manrodt 1994).

People mistakenly believe they can identify the most accurate forecast

Soll and Larrick (2009) conducted a series of four experiments to examine the strategies people use when making decisions. In each experiment, participants made initial individual estimates before they were provided with estimates from an advisor. Then, participates made their final individual estimates. The experiment design varied in the task type, the task-related expertise of participants, and feedback conditions. Participants were students (Exp. 1, 2, & 4) and midcareer executives representing 24 nationalities (Exp. 3). Participants had to estimate the mean annual salary of alumni of 25 U.S. business schools (Exp. 1 & 2), answer five country-specific questions (Exp. 3), or ten trivia questions (Exp. 4). Across the four experiments, participants tended to rely on a single source of information when making their final decision: they stayed with their initial estimate 35% of the times and fully relied on advice 10% of the times. Thus, participants tried to identify the most accurate estimate in 45% of all cases. In only 20% of all cases, participants combined their initial individual estimate and the advice, although this strategy that would have increased decision accuracy. Compared to participants' observed decision strategies, averaging one's initial estimate and the advice would have increased decision accuracy by 6% (Exp. 1), 10% (Exp. 2), and 8% (Exp. 3). Only in Experiment 4, where participants had good feedback about the expertise of judges and variation in judges' expertise was high, participants showed some ability to identify valuable judges. Participants' final estimates were slightly more accurate (1%) than averaging.

Yaniv and Milyavsky (2007) reported results from a similar study (Exp. 1), in which 160 undergraduates answered 24 questions about the dates of historical events. After providing initial individual estimates, participants received two, four, or eight estimates from others as advice. Then, participants revealed their final individual estimates. In each of the three conditions, averaging all opinions (i.e. participants' own initial opinion and the advice) outperformed the revising strategy of the participants. Overall, averaging would have yielded an error reduction of 13% compared to participants' final individual estimates. Calculating the median would have reduced error by 17%.

Ex ante conditions for when combining is most useful

The benefits from combining are expected to be highest if one uses forecasts from different valid methods that draw upon different data (Armstrong 2001). The goal of this approach is to 'create' an environment that meets the conditions of combining. The underlying assumption is that forecasts from different methods that use different data are likely to lead to bracketing and low correlations of errors. In such situations, the gains from combining can be expected to substantially exceed the 12% error reduction reported in Armstrong (2001).

Few studies directly analyzed this *ex ante* condition. Most of the studies analyzed in Armstrong (2001) were based on combining only two methods, and most of the combinations were based on similar methods (e.g., only judgmental forecasts).

An exception is the study conducted by Batchelor and Dua (1995), who analyzed combinations of forecasts made by 22 U.S. economic forecasters. These forecasts differed in their underlying economic theories (e.g., Keynesian, Monetarism, or Supply Side) and methods (e.g., judgment, econometric modeling, or time series analysis). Based on forecasts from a panel of U.S. economists, the authors found that the extent and probability of error reduction through combining were higher if the differences in the underlying theory or method of the component forecasts individual forecasts differed in their underlying theory or method increased. For example, when combining real GNP forecasts of two forecasters, combining the 5% of forecasts that were most similar in their underlying theory reduced the error of the typical forecast by 11%. By comparison, combining the 5% of forecasts that were most diverse in their underlying theory yielded an error reduction of 23%. Similar effects were obtained regarding the underlying forecasting methods. Error reduction from combining the forecasts derived from most similar methods was 2%, compared to 21% for combinations of forecasts derived from most diverse methods.

Herzog and Hertwig (2009) showed that single individuals can improve their decision-making by thinking in *ways that encourage bracketing*. In this experiment, 101 student participants were asked to revise their initial date-estimates of 40 historical events by using a consider-the-opposite strategy. Compared to the participants' initial estimates, the average of the first and the second average reduced error by 4.1 percentage points. By comparison, for the control groups that were given no specific instructions for revising their initial estimate, error reduction was only 0.3 percentage points. Thus, participants' estimates were more accurate when they were asked to use different information and assumptions when making the second estimate. However, two heads were still better than one: combining an individual's first and second estimate did not achieve the gains that could be reached by combining the initial estimates from two different individuals (7.1 percentage points).

Vul and Pashler (2008) suggested another approach that enables individual decision-makers to harness the benefits of combining. In this experiment, 428 participants provided initial estimates for eight factual questions. Then, half of the participants were asked to provide a second estimate immediately after completing the questionnaire. The other half provided their second estimate three weeks later, which was assumed to increase the independence of estimates. As expected, error reduction of the average of both estimates compared to the initial estimate was higher in the delayed conditions (95% confidence interval: 11.6% to 20.4%) than in the immediate condition (2.5% to 10.4%). However, similar to the results by Herzog and Hertwig (2009), combining within individuals was inferior to combining two estimates from different participants.

Vul and Pashler (2008) also plotted the errors for combinations of a varying number of individual estimates. Error reduction increased as more estimate were combined, although at a slower rate. The study by Yaniv and Milyavsky (2007, Exp. 1) revealed similar findings. In the first condition with three estimates (i.e., participants' initial estimate and two sources of advice), averaging yielded an error reduction of 9% compared to participants' final individual estimate. In the second condition with five estimates (four sources of advice), error reduction through averaging was 15%. Additional advice did not improve the gains from combining. In the third condition with nine estimates (eight sources of advice), error reduction was also 15%. This is consistent with Armstrong (2001), who reported on studies that showed the marginal gains from combining to decrease when the number of component forecasts exceeds five.

Election forecasting methods

Election forecasting provides an ideal opportunity for testing the *ex ante* condition of combining because (1) there are several valid approaches that are commonly used for predicting election outcomes and (2) it is unclear which of these methods is most accurate. While the media most commonly cites *polls* or *experts*, economists and political scientists have been developing *econometric models* to forecast election outcomes far in advance. Each method uses a different approach and draws upon different data.

Polls

Campaign – or trial heat – polls reveal voter support for candidates in an election. Typically, voters are asked which candidate they would support if the election were held today. Thus, polls do not provide predictions but rather snapshots of current opinion. Nonetheless, the media and the public routinely interpret polls as forecasts and project the results to Election Day. Polling results have also been used as predictor variables in econometric election forecasting models (e.g., Campbell 2008).

Experts

Another way to make a forecast is to ask political experts to make a prediction of who will win. Experts are commonly assumed to have experience in reading polling results, assessing the effect of campaigns, and estimating the effects of recent or expected events.

Models

For three decades now, economists and political scientists have used econometric models to estimate the impact of certain variables on the outcome of U.S. presidential elections. Then, the researchers would often use these models to provide a forecast of the two-party vote received by the incumbent party candidate in the next election. As summarized by Jones and Cuzán (2008), most models include between two and five variables and typically include an indicator of economic conditions and a measure of public opinion.

Combining procedure

We analyzed ex ante conditions for combining *within* and *across* forecasts from polls, experts, and models. Predictions from polls and models were available for the five presidential elections held between 1992 and 2008. Expert forecasts were available for only the two most recent elections.

Within component combining

We combined polls by calculating rolling averages of three most recent polls. In case more than three polls were published per day, all polls of that day were averaged. That way, we obtained one *poll average* forecast per day. Over all five elections, we used the results from 773 polls: 93 in 1992, 84 in 1996, 329 in 2000, 112 in 2004 and 155 in 2008.

We formed a panel of experts and contacted them periodically for their estimates of the incumbent's share of the two-party vote on Election Day. Most experts were from the ranks of academia, though some were at Washington think tanks, in the media, or former politicos. We deliberately excluded election forecasters who developed econometric models, because that method is represented as a separate component. The number of respondents in the three surveys conducted in 2004 ranged from 12 to 17. For the four surveys in 2008, the number respondents ranged from 9 to 13. Our *experts'* forecast is the simple average of individual expert forecasts.

Model averages were recalculated whenever new individual model forecasts became available. For the 2008 election, we averaged forecasts from 16 quantitative models. Ten models were used in 2004. For the retrospective analyses of the elections in 1992, 1996, and 2000, we included the forecasts of 4, 8, and 9 models, respectively. Forecasts for most models were available in July and August, and several were updated in mid- or late-October, as revised data became available.

Across component combining

For each day in the forecast horizon, we also calculated simple averages across the component forecasts (i.e., poll average, experts, and model average).

Expected accuracy gains from combining

The conditions for combining seem to be favorable in election forecasting. For combining within components, the number of forecasts was usually well above two. For combining across components, we used three forecasts that were themselves based on combinations and differed in their underlying method and data. Thus, we expected combining to yield accuracy gains beyond the 12% reported in prior research (Armstrong 2001).

Combining: within polls

The information generated by any single poll is unreliable for forecasting the outcome of the election, especially early in the campaign. Polls, conducted by reputable survey organizations at about the same time, often reveal considerable variation in results. Common explanations for these discrepancies are sampling error, non-response bias, or response bias.

Prior research has shown that combining can mitigate the problem of the variation among polls. Gott and Colley (2008) found that the median poll from the last 30 days prior to Election Day correctly forecasted the winner of every state but one in the 2004 U.S. Presidential Election. This performance was repeated in 2008, when the median polls again missed only one state.

Combining: within experts

Expert forecasts have been shown to be of some value in political forecasting. For predicting the outcome of two controversial ballot measures, Lemert (1986) found politicians' predictions to be more accurate than predictions made by members of the mass public.

Because our panelists did not meet in person, the possibility of nonrandom bias due to the influence of strong personalities or individual status was eliminated. Furthermore, different experts can be expected to use different approaches and rely on different data sources when generating their forecasts.

Combining: within models

Most econometric models are usually able to predict the correct election winner far in advance. However, the individual track record in predicting the actual vote shares of these models is mixed and forecast errors for a single model can vary widely across elections. In such a situation, it is difficult to identify the most accurate model.

Since many of the models use regression techniques and include varying combinations of the same explanatory variables (Jones and Cuzán 2008), it is unlikely that the models' individual errors are uncorrelated and randomly distributed.

But combining can reduce error even when using forecasts from similar models. Bartels and Zaller (2001) used data from the 13 U.S. presidential elections from 1948 to 1996 and varying combinations of six economic and three political variables to estimate 48 regression models. The authors then analyzed the performance of these 48 models for the 2000 election. Using their data, we calculated an error of 3.0 percentage points for the typical model. By comparison, the average forecast of all models was off by 2.5 percentage points, which refers to an error reduction of 17%. Compared to the model that performed best historically, the simple average reduced error by 25%.

Combining: across components

While each of the three component methods can be expected to produce valid forecasts, a priori it is unclear which method is most accurate. Thus, combining is useful for reducing uncertainty about the accuracy of the typical component forecast and to avoid large errors. In addition, averaging forecasts across components provides almost[1] ideal conditions for combining, as each component uses a different method and draws upon different data. In such situations, the gains from combining compared to the typical forecast can be expected to be particularly high and might even be more accurate than the most accurate component forecast.

Results

All forecasts refer to the popular vote share of the candidate from the incumbent party. We used the absolute error (i.e., the deviation of the predicted from the actual vote share) as a measure of accuracy. For the five U.S. presidential elections from 1992 to 2008, we calculated daily forecasts for each of the 93 days prior to Election Day. Thus, we obtained 465 daily forecasts for polls and models and 186 daily forecasts for experts, which were available only for the two elections in 2004 and 2008.

–– Table 1 about here ––

Accuracy gains from combining within components

Table 1 shows the mean error reduction (MER), achieved through combining within components over the whole forecast horizon, compared to the typical individual forecast. MER from combining within polls was 18%. Gains from combining within models and experts were higher (21%).

Accuracy gains from combining across components

Accuracy gains from combining across components were analyzed compared to the each component method and to the typical, best, and worst component forecast.

Combined forecasts versus component methods

For the following analysis, we used the combined forecasts of polls and models for the three elections from 1992 to 2000. For the two elections in 2004 and 2008, we used the combination of all three component methods.

Table 1 shows the results. Across the whole forecast horizon, MER of the combined forecast compared to each component method was about equal. On average, the combined forecast was 40% more accurate than the model average, 38% more accurate than the poll average, and 37% more

[1] Armstrong (2001) recommended using five or more components.

accurate than the combined experts.

Combined forecast versus the typical, best, and worst component forecast

Table 1 shows the MER of the combined forecast compared to the typical, best, and worst component forecast. On average, the error of the combined forecast was 41% lower than the error of the typical component forecast. As shown in Figure 2, often, the combined forecasts were more accurate than even the best component forecast, especially early in the forecast horizon. On average, the combined forecast performed slightly better than the best component forecast (MER: 3%). Compared to the worst component forecast, MER was large (59%). Error reduction compared to the worst component forecast shows how combining reduces risk. Given that the three component forecasts can all be considered valid forecasting methods – and already incorporate substantial accuracy gains compared to raw individual forecasts – combining is highly valuable for avoiding large errors.

–– Figure 1 about here ––

Accuracy gains from combining within and across components

Table 1 shows the MER of the combined forecast compared to the typical (uncombined) individual forecast. Gains in accuracy were highest compared to the typical individual model forecast (MER: 50%), followed by the typical individual poll (47%), and the typical individual expert (42%).

Conclusion

Averaging within and across forecasts from three component methods improved the accuracy of election forecasts. Average gains from within component combining ranged from 18% to 21%. Combining across components yielded forecasts that were, on average, 37% to 40% more accurate than the three component forecasts. Compared to the typical component forecast, combining reduced forecast error by 40%. In addition, the combined forecast performed slightly better than the best component forecast. Average error reduction compared to the worst component forecast was as large as 59%. Compared to the typical uncombined individual forecast, our combining procedure yielded mean error reductions ranging from 42% to 50%.

The achieved gains in accuracy were substantially larger than the 12% error reduction reported in prior research (Armstrong 2001). We expect that this due to the favorable conditions for combining in forecasting U.S. presidential elections. There are a number of different valid forecasting methods that use different data sources. This allows for combining forecasts within and across component methods.

Combining should be applicable to predicting other elections and, more generally, can be applied in many other contexts, as well. Given the methods available to forecasters, combining is an effective way to improve forecast accuracy and to prevent large errors. One should not put much faith in forecasts based on one method and one data set.

References

Armstrong, J. S. (2001). Combining forecasts. In: J. S. Armstrong (Ed.), *Principles of Forecasting: A Handbook for Researchers and Practitioners*, Norwell: Kluwer, pp.417-439.

Bartels, L. M. & Zaller, J. (2001). Presidential vote models: A recount. *PS: Political Science & Politics*, 34, 9-20.

Batchelor, R. (2007). Bias in macroeconomic forecasts, *International Journal of Forecasting*, 23, 189-203.

Batchelor, R. & Dua, P. (1995). Forecaster diversity and the benefits of combining forecasts. *Management Science*, 41, 68-75.

Campbell, J. 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, 697-701.

Gott , J. R. & Colley, W. N. (2008). Median statistics in polling. *Mathematical and Computer Modelling,* 48, 1396-1408.

Herzog, S. M. & Hertwig, R. (2009). The wisdom of many in one mind. Improving individual judgments with dialectical bootstrapping. *Psychological Science*, 20, 231-237.

Hogarth, R. (in press). When simple is hard to accept. In P. M. Todd, G. Gigerenzer, & The ABC Research Group (Eds.), *Ecological rationality: Intelligence in the world*. Oxford: Oxford University Press.

Jones, R. J. & Cuzán, A. G. (2008). Forecasting U.S. presidential elections: A brief review. *Foresight – The International Journal of Applied Forecasting,* Summer 2008, 29-34.

Larrick, R. P. & Soll, J. B. (2006). Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Science,* 52**,** 111-127.

Lemert, J. B. (1986). Picking the winners: Politician vs. voter predictions of two controversial ballot measures. *Public Opinion Quarterly*, 50, 208-221.

Sanders, N. R. & Manrodt, K. B. (1994). Forecasting practices in U.S. corporations: Survey results. *Interfaces*, 24, 92-100.

Soll, J. B. & Larrick, R. P. (2009). Strategies for revising judgment: How (and how well) people use others' opinions, *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35, 780-805.

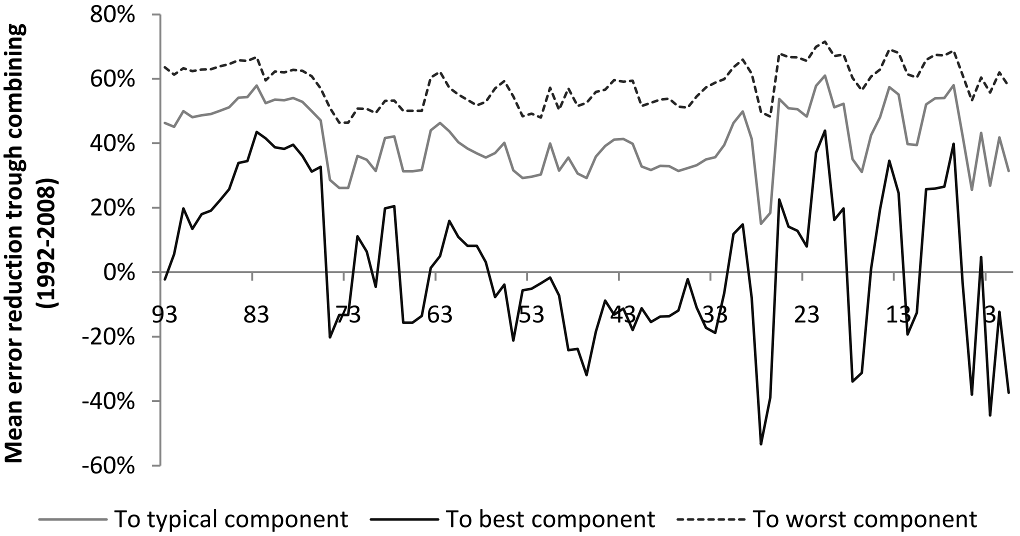
Vul, E. & Pashler, H. (2008). Measuring the crowd within: probabilistic representations within individuals. *Psychological Science*, 19, 645-647.

Yaniv, I. & Milyavsky, M. (2007). Using advice from multiple sources to revise and improve judgments. *Organizational Behavior and Human Decision Processes*, 103, 104-120.

Table 1: Accuracy gains from combining

	Mean error reduction (in %)
Within component combining, compared to	
Typical individual poll	18
Typical individual model	21
Typical individual expert	21
Across component combining, compared to	
Poll average	38
Model average	40
Experts	37
Typical component	41
Best component	3
Worst component	59
Within and across combining, compared to	
Typical individual poll	47
Typical individual model	50
Typical individual expert	42

Figure 1: Mean error reduction from across component combining, compared to typical, best, and worst component forecast


Mean error reduction trough combining (1992-2008)
To typical component
To best component
To worst component