Real-Time Forecast Averaging with ALFRED

Chanont Banternghansa and Michael W. McCracken

This paper presents empirical evidence on the efficacy of forecast averaging using the ALFRED (ArchivaL Federal Reserve Economic Data) real-time database. The authors consider averages over a variety of bivariate vector autoregressive models. These models are distinguished from one another based on at least one of the following factors: (i) the choice of variables used as predictors, (ii) the number of lags, (iii) use of all available data or only data after the Great Moderation, (iv) the observation window used to estimate the model parameters and construct averaging weights, and (v) the use of either iterated multistep or direct multistep methods for forecast horizons greater than one. A variety of averaging methods are considered. The results indicate that the benefits of model averaging relative to Bayesian information criterion-based model selection are highly dependent on the class of models averaged The authors provide a novel decomposition of the forecast improvements that allows determination of the most (and least) helpful types of averaging methods and models averaged across. (JEL E52, E58, C53)

Federal Reserve Bank of St. Louis *Review*, January/February 2011, *93*(1), pp. 49-66.

This paper provides evidence on the ability of various forms of forecast averaging to improve the real-time forecast accuracy of monthly bivariate vector autoregressive (VAR) forecasts of headline and core consumer price index (CPI)-based inflation, growth in industrial production (IP), and the unemployment rate. We consider a range of approaches to averaging forecasts obtained by a variety of primitive methods for managing the estimation of each bivariate VAR model. The averaging methods include equally weighted averages, medians, mean square error (MSE)-weighted averages, Bayesian model averages based on a Bayesian information criterion (BIC) approximation, and averages based on the top 10 percent of models that have performed best historically. For each averaging approach, we construct forecasts of each variable using real-time data from the ALFRED (ArchivaL Federal Reserve Economic Data) database. We compare our model-averaging results with those obtained with BIC-based model selection.

Model averaging for forecasting is nothing new. An abundance of evidence suggests that model averaging can improve forecast accuracy relative to model selection. Empirical examples of this evidence include, but are certainly not limited to, Stock and Watson (2004), Kapetanios, Labhard, and Price (2008), and Kascha and Ravazzolo (2010). Theoretical results include Hansen (2008), Elliott and Timmermann (2004), Clark and McCracken (2008), and many others.

In some instances (e.g., Clark and McCracken, 2010, and Faust and Wright, 2009), forecasting with model averages accounts for the real-time nature of the data. Even so, such examples are the exception and not the norm. Here we use the

Chanont Banternghansa was a research assistant and Michael W. McCracken is a research officer and economist at the Federal Reserve Bank of St. Louis.

© 2011, The Federal Reserve Bank of St. Louis. The views expressed in this article are those of the author(s) and do not necessarily reflect the views of the Federal Reserve System, the Board of Governors, or the regional Federal Reserve Banks. Articles may be reprinted, reproduced, published, distributed, displayed, and transmitted in their entirety if copyright notice, author name(s), and full citation are included. Abstracts, synopses, and other derivative works may be made only with prior written permission of the Federal Reserve Bank of St. Louis.

ALFRED database to mimic the type of data that would have been accessible to forecasters at each point in time as they construct their monthly forecasts. Using real-time data is important because it accounts for the fact that economic data are often subject to revision and hence the actual value of a variable may change across forecast origins. In addition, using real-time data accounts for the fact that most macroeconomic data become available only after a substantial lag and, moreover, these time lags can vary widely across variables from as short as a week (for employment figures) to as long as two months (for trade data). Finally, by using the ALFRED database as the universe of potential predictors, we allow for the availability of new series across time and existing series that are sometimes discontinued.

In accordance with the literature, our results indicate that model averaging can—but does not always—improve forecast accuracy relative to the more-standard BIC-based approach to model selection. Put differently, model averaging per se is not a panacea for improving forecast accuracy. Improvements from model averaging depend critically on the type of models averaged across. Preselecting which primitive models should be used in the averaging process appears to offer some advantage. For example, when forecasting core CPI-based inflation there appear to be substantial gains in forecast accuracy at all horizons when averaging over only those models estimated with a rolling observation window of fixed size rather than a recursive, expanding observation window. In contrast, we find improved IP forecasting accuracy when averaging over only those models estimated with a recursive window rather than a rolling window of observations.

With these two examples in mind, we provide a novel decomposition of the relative root mean square error (RMSE) improvements for each dependent variable at each forecast horizon, which allows us to determine which primitive model types and model-averaging techniques are, on average, most (and least) beneficial. In some, though not all, instances our decomposition meshes well with the permutations of types of models and types of averaging procedures that produce the most accurate forecasts.

The remainder of the paper proceeds as follows. The next section describes the real-time data used in our analysis. We then provide a synopsis of the primitive models we average over, followed by a section describing the types of model averaging we consider. Finally, we present our results on forecast accuracy, our decomposition, and our conclusions.

DATA

We obtained our data from the ALFRED database maintained by the Federal Reserve Bank of St. Louis. This database consists of collections of vintages of data for each variable—that is, vintages that vary across time as either new data are released or existing data are revised by the relevant statistical agency. Using this database ensures that at each monthly forecast origin we are using only data that were available as of the date of the forecast origin. We therefore define "real-time" forecasting as using any data available by the end of the month from which we are forecasting.[1]

Choosing the end of a month as the forecast origin is nontrivial. Nearly all monthly macroeconomic data are released after the end of the month the data reference. A model needing data for January 1996 must therefore be constructed after that month has ended. If we choose the first day of February 1996 as our forecast origin, the forecast would be very timely but there would be almost no data for January to use, thus reducing the accuracy of the forecast. On the other hand, if we choose the first day of May as our forecast origin, all the data for January would be available but the forecast would be very outdated. As a middle ground we choose the end of the month following the most recent data vintage as the relevant forecast origin. For example, this implies that one-step-ahead forecasts, constructed using January 1996 vintage data, made at the end of

[1] The ALFRED database (http://alfred.stlouisfed.org/) allows retrieval of vintage versions of economic data available on specific dates in history. In general, economic data for past observation periods are revised as more accurate estimates become available. Currently, vintage data are available for 24,293 series in 14 categories.

February, will be forecasts of data associated with February 1996.[2]

Our analysis uses a total of 238 unique monthly macroeconomic series from the ALFRED database. Of these 238 series, 67 are available for the January 1996 vintage data. As we progress across time, we allow the number of variables to increase or decrease with data availability. For example, the number of series available more than doubles in November 1996. By the end of our forecasting exercise in December 2008 a total of 193 series are used either as dependent variables or as predictors. This is less than the total number of variables because 45 series were discontinued or did not have enough observations at some point in time to adequately estimate either the model parameters or model-averaging weights.[3] There are 29 output and production series; 8 income, outlays, and savings series; 40 labor market series; 52 monetary aggregate and reserve series; 35 exchange rate series; 38 financial market and interest rate series; 34 price series; and 2 survey series. The detailed list is available from the authors on request.

For brevity, in our forecasting exercise we focus exclusively on forecasting four of the most publicly visible nominal and real monthly frequency variables: headline and core CPI-based inflation, IP growth, and the unemployment rate. Specifically, at each forecast origin starting in February of 1996, we construct forecasts of three variables: headline CPI-based inflation, IP growth, and the unemployment rate. We begin forecasting core CPI-based inflation using December 1996 vintage data—the first available vintage for this series. For each of the four variables we construct h = 1-, 3-, 6-, 12- and 24-month-ahead forecasts. For unemployment, the target variable being forecast is y_{t+h}, the unemployment rate at the forecast horizon h. For CPI and IP, the target variable being forecast is the average annualized monthly rate of growth over the forecast horizon and hence interpretation of the target variable varies with the forecast horizon. More precisely, if we let y_t denote the time t log difference in, say, headline CPI, the target variable being forecast at horizon h is

$$y_{t+h}^{(h)} = \left(\frac{1200}{h}\right)\sum_{i=1}^{h} y_{t+i}.$$

In constructing our forecast errors, we use the third release (or, equivalently, the second revision) of the variable as the realized value of our target variable. In total, because December 2008 is the final vintage used to evaluate our forecasts, for each model we have roughly 155 1-month-ahead forecast errors that we use to measure accuracy. This number shrinks to 151, 145, 133, and 109 for the 3-, 6-, 12-, and 24-month-ahead forecasts, respectively. Following Marcellino, Stock, and Watson (2006), each variable is transformed to ensure stationarity using differences or log differences. For the dependent variables, we treat the unemployment rate as stationary in levels but treat headline CPI, core CPI, and IP as stationary in log-first differences. These transformations are made across all vintages uniformly. We do not allow for differences in the type of transformation across vintages. After transforming the variables we then check for outliers, defined as observations greater than six times the interquartile range. The outliers are replaced with the mean of the series (without the outlier) from the relevant vintage. This replacement is done vintage by vintage and hence the outlier detection is not influenced by observations not available at each forecast origin. Note that across the forecasting period the CPI and IP indices have been periodically renormalized so that the units of measurement are not the same across all vintages. To avoid mixing and matching, we renormalized each vintage relative to the December 2008 vintage.

METHODS

In this section we describe the primitive models over which we average. All models have one thing in common: They all take the form of

[2] Giannone, Reichlin, and Small (2008) refer to this type of forecast as a "nowcast."

[3] In our analysis, we set a few basic rules for inclusion of variables: (i) We do not use seasonally unadjusted data when the seasonally adjusted version is available, (ii) we do not use regional data for our analysis, (iii) we omit a variable if fewer than 10 years of data are available for estimating the model parameters, and (iv) we omit a variable if we do not have at least 24 pseudo out-of-sample forecast errors to calculate the MSE-weighted forecasts.

an OLS-estimated bivariate VAR in the variable to be predicted and one additional predictor (see the section "Iterated Multistep and Direct Multistep Forecasts" for a caveat). Otherwise, all the primitive models differ by at least one of six features: (i) the series from the ALFRED database used as an additional predictor, (ii) the number of lags of the dependent variable used as a predictor, (iii) the number of lags of the additional predictor used, (iv) whether the model is estimated using all available data (i.e., the recursive scheme) or a moving window of observations (i.e., the rolling scheme), (v) whether the model is estimated using only post-Great Moderation data or data as far back as available for that vintage, or (vi) for forecast horizons greater than one step ahead, whether iterated multistep (IMS) or direct multistep (DMS) methods are used to create our primitive forecasts.

Predictors

As noted previously, we use the ALFRED database for our real-time forecasting exercise. In particular, we treat it as the universe of potential variables that could be used as a predictor for any one of our four dependent variables. Since the number of variables in ALFRED changes across forecast origins, the number of primitive models over which we average changes across forecast origins. At the beginning of our sample, January 1996, we have a total of only 66 potential predictors for each dependent variable. At the last potential forecast origin, November 2008, we have a total of 192 potential predictors for 1-step-ahead forecasts. While the number of predictors typically grows—sometimes dramatically, as for November 1996—in a few instances the number of predictors falls as various variables are discontinued or dropped because of insufficient data.[4]

Full Sample and Great Moderation Sample

For each model, we estimate the regression parameters using one of two subsets of data. In the first, the full sample, we use all available data in that vintage. While the date of the first observation varies across individual variables, many date back to as early as January 1959. In the second, the post sample, we restrict attention to only those data available starting in January 1983, roughly the time frame for the start of the Great Moderation. Note that for the post sample, this implies that for each vintage used for estimation, any pre-1983 observations are discarded.

We consider both subsets of data because there is considerable evidence, including that in D'Agostino, Giannone, and Surico (2007), that the predictability of many macroeconomic variables has changed since the onset of the Great Moderation. Even so, there is a trade-off. Using less information to estimate model parameters may generate estimates that are more likely to be unbiased because older data come from a different macroeconomic regime, but less information also can decrease the precision of the estimates. In practice, this trade-off may favor using more (or less) data to estimate parameters due to a bias-variance trade-off.

Recursive and Rolling Windows

For each model, and conditional on whether we use the full or post sample, we estimate the bivariate VAR using one of two observation windows. In the recursive scheme, we estimate the model by OLS using all available data. Hence as we move forward from one month to the next, we use one more observation to estimate the model parameters. In the rolling scheme, we estimate the model by OLS using only the past 10 years of available data. Hence when using the rolling scheme, as we move forward from one month to the next we use the same number of observations to estimate the model parameters.

In some ways, our decision to consider two subsets of data (full vs. post) and two types of observation windows (recursive vs. rolling) may seem redundant. We view the two choices, however, as distinct but related. In the former, we essentially assume a discrete break in 1983 and see how doing so helps forecast accuracy. For the latter, we assume a somewhat smoother sequence of breaks. Since we are unsure which is the proper way to manage forecasting in the presence of

[4] See footnote 3 for more detail.

uncertain forms of potential structural change, we consider both. See Clark and McCracken (2010) for further discussion on this issue.

Iterated Multistep and Direct Multistep Forecasts

For each permutation of predictor, sample, and observation window, we estimate our bivariate VAR forecasting model using two different methods: the textbook method that induces an IMS forecast and the somewhat easier-to-implement method of DMS forecasting. The following text provides a brief description of each approach.

Let y_t denote either the time t level of the unemployment rate or the time t log-first difference of headline or core CPI or IP. In addition, recall that the target variable to be forecast at forecast horizon h is

$$y_{t+h}^{(h)} = \left(\tfrac{1200}{h}\right)\sum_{i=1}^{h} y_{t+i}$$

for the CPI and IP indices but is simply y_{t+h} for unemployment. For the IMS forecasting approach, at each forecast origin t we first use OLS to estimate the bivariate VAR model,

$$(1)\quad \begin{pmatrix} y_t \\ x_t \end{pmatrix} = \begin{pmatrix} \alpha_{y,0} \\ \alpha_{x,0} \end{pmatrix} + A(L)\begin{pmatrix} y_{t-1} \\ x_{t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{y,t} \\ \varepsilon_{x,t} \end{pmatrix},$$

where $A(L)$ denotes a lag operator of appropriate dimension for the given number of lags used in both the y and x equations. With the regression parameter estimates in hand, the recursive nature of the VAR is used to generate a sequence of 1- through h-step-ahead forecasts $\hat{y}_{t+i}$ 1 = 1,…,h. For the unemployment rate, $\hat{y}_{t+h}$ is the resulting forecast of our target variable. For the other dependent variables, we follow Marcellino, Stock, and Watson (2006) and define our h-step-ahead IMS forecast as

$$\left(\tfrac{1200}{h}\right)\sum_{i=1}^{h} \hat{y}_{t+i}.$$

Note that for each forecast horizon, the same parameter estimates are used to construct the forecasts.

For the DMS forecasting approach, a distinct model is estimated separately for each forecast horizon h. For the unemployment rate and a fixed value of h, this model takes the form

$$(2)\quad y_{t+h} = \alpha_{y,0} + A_y(L)y_t + A_x(L)x_t + \varepsilon_{y,t+h},$$

where $A_y(L)$ and $A_x(L)$ denote lag operators of appropriate dimension for the given number of lags used for y and x, respectively. For each separate forecast horizon the forecast is defined as

$$\hat{y}_{t+h} = \hat{\alpha}_{y,0} + \hat{A}_y(L)y_t + \hat{A}_x(L)x_t.$$

For the CPI and IP indices, the model takes the slightly different form of

$$(3)\quad y_{t+h}^{(h)} = \alpha_{y,0} + A_y(L)y_t + A_x(L)x_t + \varepsilon_{y,t+h}.$$

For each separate forecast horizon the forecast is similarly defined as

$$\hat{y}_{t+h}^{(h)} = \hat{\alpha}_{y,0} + \hat{A}_y(L)y_t + \hat{A}_x(L)x_t.$$

Note that in each of the above examples, the parameter estimates from these models vary with the forecast horizon.

Lags

Each of the IMS and DMS specifications requires choosing the number of lags of y and x to use as predictors. The textbook approach would be to use a model-selection procedure such as BIC. Such a choice, however, contrasts with our goal of providing evidence on the benefits of model averaging relative to model-selection techniques. In addition, because of the considerable evidence suggesting a change in the degree of persistence in inflation (e.g., Levin and Piger, 2006), one might consider the possibility that the lag order structure of the model, for inflation in particular, has changed over time. We therefore consider all 144 permutations of up to 12 lags of either the y or x variable.

AVERAGING METHODS

After considering all the permutations of model elements discussed above, for each variable we have 76,128 1-month-ahead forecasting models estimated in January 1996 and 221,280

1-month-ahead forecasting models estimated in November 2008.[5] With this rich collection of individual forecasting models as building blocks, we consider a range of approaches to model averaging with an eye toward determining which types of model averaging are most useful and moreover, which types of primitive models are the most useful for averaging over.

Simple Model Averages

Our first set of model averages is the simplest. We consider the equally weighted average and the median forecast from among these models. While these methods are not statistically exciting, substantial evidence suggests that simple forms of model averaging can perform quite well (e.g., Smith and Wallis, 2009). Note that this form of model averaging implies model weights invariant to the forecast horizon.

Weighted Model Averages: Inverse Mean Square Error Weights

We then consider two distinct forms of weighted model averaging. In the first, we follow Stock and Watson (2004) (among others) and consider relative inverse mean square forecast error (MSE)-based weights to combine our models. The intuition is that if historical evidence suggests some models are more accurate than others, it may be beneficial to give those particular models more weight. Computationally, if $MSE_{i,t,h}$ denotes the known MSE associated with individual model i at forecast origin t associated with a sequence of past h-step-ahead forecast errors, the weight given to model i is

$$\frac{MSE_{i,t,h}^{-1}}{\sum_{j=1}^{N_t} MSE_{j,t,h}^{-1}},$$

where $j = 1,\ldots,N_t$ denotes an index of all the available primitive models at forecast origin t.

In our application, for the relevant vintages of data needed to estimate a particular model at forecast origin t, we conduct a pseudo out-of-sample forecasting exercise to generate these MSEs. The particulars of the exercise depend on whether (i) the full or post sample and (ii) the recursive or rolling scheme are used to construct our forecasts. If the recursive (rolling) scheme is used for the model forecast, then the recursive (rolling) scheme is used for the pseudo out-of-sample forecasts used to construct the model weights. If the full sample is used, the first pseudo out-of-sample forecast is based on parameters estimated using data from January 1960 to December 1969 and iterates forward until the availability of real-time data, at time t, is insufficient to calculate a forecast error using the third release of the relevant dependent variable. If the post sample is used, the first pseudo out-of-sample forecast is based on parameters estimated using data from January 1984 to December 1993 and iterates forward as discussed. Since our forecasting exercise starts in January 1996, this implies that the model weights constructed with the full sample are estimated based on an average MSE that uses many more squared forecast errors than those constructed with the post sample.

Weighted Model Averages: Bayesian Weights

We also consider an approximate Bayesian model-averaging strategy in which we calculate a posterior probability from prior probabilities and marginal likelihoods for each model, with each model assigned the same prior probability. Following Garratt, Koop, and Vahey (2006), the marginal likelihood of a given model is approximated using its BIC. In our analysis, for each vintage we estimate each model using the relevant subset of the available data (i.e., the full or post sample) and, based on the subsequent residuals, calculate the value of the BIC. Computationally, if we let $BIC_{i,t,h}$ denote the value of the BIC associated with the residuals from individual model i at forecast origin t, the weight given to model i is

$$\frac{\exp\left(-0.5*BIC_{i,t,h}\right)}{\sum_{j=1}^{N_t}\exp\left(-0.5*BIC_{j,t,h}\right)}.$$

For the IMS models the BIC is constructed in the typical fashion using equation (1), which implic-

[5] The number of models not only changes across forecast origins but also varies slightly across forecast horizons due to data availability. See footnote 3.

itly assumes that the residuals are serially uncorrelated. For the DMS models, however, we know that when $h > 1$ the residuals from equation (2) are not serially uncorrelated and hence the typical formulation is invalid.[6] For simplicity, we use the standard BIC formula regardless.

Weighted Model Averages with Trimming

In addition to the previously described weighted forecasts that average across all models, we also considered a variant that filters out the models considered "less accurate" by some metric and averages over only those remaining. Specifically, at each forecast origin t we follow Aiolfi and Timmermann (2006) and Clark and McCracken (2010) by calculating a top 10 percent MSE-weighted and a top 10 percent BIC-weighted average constructed using only the top 10 percent of the available models. For the top 10 percent MSE models this is done by averaging over only the models with the lowest 10 percent of pseudo out-of-sample MSEs based on the data available as of the forecast origin. Similarly, for the top 10 percent BIC models this is done by averaging over only the models with the lowest 10 percent of values of BIC based on the data available as of the forecast origin.

Benchmark Forecast

In reporting our results it is useful to gain some perspective on the magnitude of the benefits of model averaging. Doing so requires choosing a baseline for comparison. Since our goal is to observe the benefits of model averaging relative to model selection, using a fixed autoregressive model with known lags is insufficient. Not only does that baseline fail to capture the time-varying nature of model selection in a real-time forecast setting, in many cases it does not even serve as a particularly difficult benchmark to "beat." For example, we could have used the standard random walk benchmark but, as seen below, while this is a strong benchmark for the unemployment rate, it is a horrible benchmark for IP and both CPI indices.

Instead, we use the recursively estimated, IMS, BIC-selected forecast estimated over the full sample as our benchmark. At each forecast origin t this entails calculating the value of the BIC for each IMS model from equation (1), estimated by (i) using the full sample, separately across all possible lag permutations and choices of additional predictor, and (ii) then choosing the model with the lowest BIC as the model that is used to construct the forecast. The reason for our selection is that this particular BIC-selected forecast is the conventional methodology that a textbook in time-series econometrics would suggest. For completeness, we also report the relative RMSEs associated with the random walk model.

Before we proceed, it is important to clarify two things about our "benchmark model." First, it is chosen in real time in the sense that at each forecast origin we use only the vintage of data available at that forecast origin.[7] In particular, we use only the vintage of data available at the time the forecast is constructed to compute the value of the BIC for each possible model. Second, across time there is no single benchmark model. That is, as we proceed across forecast origins, it is possible for the model with the smallest value of BIC to change. This can occur for any number of reasons: the presence of unmodeled structural change, revisions in the data across vintages, or even changes in the collection of models considered as the universe of variables in ALFRED expands or contracts across time. Because of this possibility, the benchmark model is not so much a "model" as it is a forecasting method.

Summary of Methods

For each variable and each horizon, we consider six different forms of model averaging: average, median, (inverse) MSE-weighted, BIC-weighted, top 10 percent (inverse) MSE-weighted, and top 10 percent BIC-weighted. Each form of

[6] See Hansen (2010) for a discussion of how this affects the definition of BIC.

[7] Recall that the phrase "full sample" is intended to denote that for a given forecast origin the entirety of the corresponding vintage of data is used for estimation. This is in contrast to the phrase "post sample," which uses only the portion of the corresponding vintage that coincides with the Great Moderation.

Table 1

RMSEs of Out-of-Sample Forecasts of Nominal Variables

	Forecast horizon				
Variables	1 month	3 month	6 month	12 month	24 month
Headline CPI					
BIC, recursive, IMS, full*	3.560	**2.741**	1.622	1.146	0.800
Random walk	1.151	1.519	2.298	2.851	4.234
Median	0.995	1.018	**0.990**	0.934	0.760
Average, all forecasts	0.995	1.021	1.008	0.952	0.816
MSE weight, all forecasts	0.995	1.018	0.993	0.934	0.778
MSE weight, top 10%	1.000	1.030	1.006	0.943	0.821
BIC weight, all forecasts	**0.994**	1.024	1.007	0.946	0.781
BIC weight, top 10%	**0.994**	1.023	0.996	**0.913**	**0.667**
Core CPI					
BIC, recursive, IMS, full*	1.233	0.805	0.606	0.586	0.591
Random walk	1.198	1.580	1.858	1.855	1.954
Median	0.938	0.942	0.899	0.867	0.827
Average, all forecasts	**0.931**	0.962	0.944	0.944	0.967
MSE weight, all forecasts	0.934	0.938	**0.884**	**0.840**	0.827
MSE weight, top 10%	0.958	0.949	0.893	0.848	0.834
BIC weight, all forecasts	0.936	0.946	0.918	0.918	0.922
BIC weight, top 10%	0.946	**0.932**	0.888	0.842	**0.810**

NOTE: *Values associated with the first row in each panel are RMSEs. The remaining values are ratios of RMSEs relative to that of the first row. For each forecast horizon, the best relative RMSE is shown in bold type. BIC, Bayesian information criterion; CPI, consumer price index; full, full sample (all available data in that vintage); IMS, iterated multistep; MSE, mean square error. See text for details.

averaging is then applied separately to several distinct classes of models, which are indexed by their type of construction using (i) the full and/or post samples, (ii) the recursive and/or rolling schemes, and (iii) the IMS and/or DMS approaches to forecasting. Note that since we allow for averaging over, for example, models estimated using either the recursive or rolling schemes, there are $3^3 = 27$ model classes that we consider. In all, this gives us $6 \times 3^3 = 162$ distinct permutations of forms of model averaging and the types of models that are averaged over.

RESULTS

In this section, we discuss our results on the benefits of using forecast averaging as a tool for improving forecast accuracy. For brevity, however, we do not present the tables associated with all 162 model-averaging and model class variants. Instead, Tables 1 and 2 present results for each type of model averaging when we average over *all* models. Table 1 presents results for headline and core CPI-based inflation and Table 2 presents results for growth in IP and the unemployment rate. The values in the first row of each panel of these tables are the RMSEs associated with the benchmark model chosen using BIC at each forecast origin. The remaining values in each panel are relative RMSEs. Values greater than 1 favor the benchmark model, while values less than 1 favor the form of model averaging denoted in the first column. For each forecast horizon, the best relative RMSE is shown in bold type.

Table 2

RMSEs of Out-of-Sample Forecasts of Real Variables

Variables	Forecast horizon				
	1 month	3 month	6 month	12 month	24 month
Industrial production					
BIC, recursive, IMS, full*	9.951	6.229	5.050	4.136	**2.837**
Random walk	1.125	1.263	1.313	1.561	2.281
Median	**0.985**	0.972	0.986	**0.994**	1.070
Average, all forecasts	**0.985**	**0.970**	**0.981**	**0.994**	1.055
MSE weight, all forecasts	0.986	**0.970**	0.982	0.996	1.056
MSE weight, top 10%	0.988	0.974	0.982	1.011	1.067
BIC weight, all forecasts	0.987	0.972	0.983	1.001	1.072
BIC weight, top 10%	0.989	0.990	1.005	1.023	1.094
Unemployment rate					
BIC, recursive, IMS, full*	0.167	0.301	0.463	0.696	1.023
Random walk	0.995	0.958	0.947	0.982	1.031
Median	0.936	0.882	0.864	0.922	0.936
Average, all forecasts	0.935	0.877	0.857	0.917	0.922
MSE weight, all forecasts	0.935	0.877	0.856	0.916	0.926
MSE weight, top 10%	**0.922**	**0.865**	**0.836**	**0.910**	0.969
BIC weight, all forecasts	0.935	0.879	0.862	0.918	**0.919**
BIC weight, top 10%	0.938	0.895	0.889	0.953	0.985

NOTE: *Values associated with the first row in each panel are RMSEs. The remaining values are ratios of RMSEs relative to that of the first row. For each forecast horizon, the best relative RMSE is shown in bold type. BIC, Bayesian information criterion; CPI, consumer price index; full, full sample (all available data in that vintage); IMS, iterated multistep; MSE, mean square error. See text for details.

Root Mean Square Errors of Nominal Variables

The first panel of Table 1 provides the results of forecasts for headline CPI-based inflation averaged across all models. At the three shortest horizons there are few, if any, advantages to forecast averaging across all models in terms of RMSEs. When averaging over all the primitive models, the benchmark is either better than model averaging or only marginally worse. However, as the forecast horizon increases to 12 months, model averaging improves accuracy by roughly 5 percent and at the longest horizon, forecast averaging improves accuracy by as much as 30 percent. In each of these latter horizons, the top 10 percent BIC-weighted forecasts yielded the lowest RMSEs.[8]

The second panel of Table 1 provides the results for core CPI-based inflation. In contrast to the results for headline inflation, consistent improvements are noted at all horizons for model averaging across all models. At the shortest horizons, the gains were on the order of a modest 5 percent, but as the horizon increases the improvements rise to about 15 percent. Across all horizons, no single averaging approach consistently gives the greatest improvements: The average, MSE-weighted, and top 10 percent BIC-weighted forecasts each perform best in at least one horizon.

[8] We do not test for statistical significance in our results because there is no known method for doing so when the baseline model is allowed to change across time and the competing model forecast is not based on a model per se but is instead an average across many models.

Root Mean Square Errors of Real Variables

The first and second panels of Table 2 parallel those in Table 1 in terms of the benefits of model averaging. As for headline CPI, model averaging across all models provides little to no improvement relative to model selection when forecasting IP growth at the shortest horizons. In fact, model averaging typically is worse than model selection at the longest horizons with losses of roughly 5 percent.

But again, in contrast to the results in the first panel, model averaging across all models consistently improves forecast accuracy relative to model selection when forecasting the unemployment rate. Each model-averaging procedure improves forecast accuracy at every horizon. Somewhat surprisingly, the improvements are (inverse) U shaped: The improvements in RMSE are roughly 7 percent at the shortest and longest horizons but are closer to 12 percent at the intermediate horizons. Across all but the longest horizon, the top 10 percent MSE-weighted forecast has the largest improvements relative to the benchmark. At the longest horizon the BIC-weighted average performs best.

Decomposition Regression Analysis

Tables 1 and 2 show that while model averaging can improve forecast accuracy, it does not always do so relative to our model selection-based benchmark. Moreover, when model averaging does provide improvements, the best form of model averaging varies across both dependent variables and forecast horizons. Finally, though obviously not apparent in Tables 1 and 2 (which present results averaged over *all* the primitive models), comparable conclusions can be reached if we report all the remaining permutations of types of model averaging and model classes for each dependent variable and each horizon.

Even so, it may be that on average across all these permutations, some simple patterns emerge that could help in identifying the best types of model averaging and the classes of models that should be averaged over. To parse out such effects we estimate a regression in which we use dummy variables for the types of model averaging and model classes as predictors for the corresponding relative RMSEs. Specifically, for each dependent variable and each forecast horizon, we use OLS to estimate the following regression:

$$\begin{aligned} RMSE_i^h - 1 &= \alpha_1 \text{DMS} + \alpha_2 \text{Post} + \alpha_3 \text{Roll} \\ &+ \beta_1 \text{IMS/DMS} + \beta_2 \text{Full/Post} + \beta_3 \text{Rec/Roll} \\ &+ \gamma_1 \text{Equal} + \gamma_2 \text{Weight} + \gamma_3 \text{Top 10\%} + \gamma_4 \text{MSE} + \varepsilon_i^h, \end{aligned} \tag{4}$$

where $RMSE_i^h$ is the relative RMSE of permutation $i = 1,\ldots,162$ and *Rec* and *Roll* denote the recursive and rolling window schemes, respectively. By subtracting 1 the coefficients are more easily interpreted as indicating percent improvement (a negative coefficient) or percent deterioration (a positive coefficient) relative to our benchmark.

The α coefficients in equation (4) are associated with variables that indicate how an individual forecast is made: *DMS* takes the value 1 if only DMS models are included and 0 otherwise, *Post* takes the value 1 if only Great Moderation data are used and 0 otherwise, and *Roll* takes the value 1 if only a rolling window of observations is used to estimate the model parameters and 0 otherwise. The β coefficients are associated with the different combinations of the α coefficients: *IMS*/*DMS* takes the value 1 if the weighted forecast combines both IMS and DMS forecasts and 0 otherwise, *Full*/*Post* takes the value 1 if the weighted forecast combines both the full and post samples and 0 otherwise, and *Rec*/*Roll* takes the value 1 if the weighted forecast combines both recursive and rolling estimation schemes and 0 otherwise. The γ coefficients are associated with how the weighted forecasts are constructed: *Equal* takes the value 1 if either the average or median averaging methods are used and 0 otherwise, *Weight* takes the value 1 if the models are weighted unequally and 0 otherwise, *Top 10%* takes the value 1 if the averaging uses only the top 10 percent of forecasts and 0 otherwise, and *MSE* takes the value 1 if MSE-based weights are used and 0 otherwise.

Results for Nominal Variables

Table 3 shows decomposition results for both headline and core CPI-based inflation. In each

Table 3

Decomposition Regression of Nominal Variables

	Forecast horizon				
Variables	1 month	3 month	6 month	12 month	24 month
Headline CPI					
DMS	0.000	–0.004	–0.001	–0.016*	0.014
DMS/IMS	0.000	–0.001	0.001	–0.005	–0.000
Post	–0.011***	–0.027***	–0.049***	–0.066***	–0.147***
Post/Full	–0.009***	–0.019***	–0.035***	–0.050***	–0.113***
Roll	–0.004***	–0.045***	–0.075***	–0.086***	–0.141***
Rec/Roll	–0.012***	–0.031***	–0.050***	–0.078***	–0.177***
Equal	0.018***	0.074***	0.088***	0.082***	0.087**
Weighted	–0.001	–0.002	–0.000	0.001	–0.017
Top 10%	0.000	–0.002	–0.011*	–0.020**	–0.047***
MSE	–0.001	0.002	–0.006	–0.022***	–0.011
N	162	162	162	162	162
Core CPI					
DMS	–0.000	–0.010*	–0.034***	–0.085***	–0.152***
DMS/IMS	–0.000	–0.004	–0.011	–0.025	–0.044
Post	0.001	–0.041***	–0.075***	–0.130***	–0.230***
Post/Full	–0.002	–0.034***	–0.060***	–0.102***	–0.176***
Roll	–0.002*	–0.069***	–0.134***	–0.236***	–0.414***
Rec/Roll	–0.013***	–0.056***	–0.105***	–0.182***	–0.317***
Equal	–0.048***	0.046***	0.096***	0.215***	0.439***
Weighted	–0.004***	–0.013**	–0.011	–0.015	–0.027
Top 10%	0.009***	–0.011**	–0.025**	–0.049***	–0.073***
MSE	0.004**	0.000	–0.020**	–0.040**	–0.041
N	162	162	162	162	162

NOTE: Each column in each panel provides the coefficients associated with a distinct OLS-estimated version of equation (4). *, **, and *** denote statistical significance at the 10 percent, 5 percent, and 1 percent levels, respectively. BIC, Bayesian information criterion; CPI, consumer price index; DMS, direct multistep; Full, full sample (all available data in that vintage); IMS, iterated multistep; MSE, mean square error; Post, only data that occur starting in January 1983; Rec, recursive window scheme; Roll, rolling window scheme. See text for details.

panel, the first six rows relate to the selection of models to average over and the next four rows relate to the type of averaging method. We begin by studying panel 1 (that associated with headline CPI-based inflation).

In the first two rows of panel 1 (those associated with averaging over DMS models, IMS models, or both), there appears to be little statistically significant advantage to any of these particular forecasting methods. The sole exception is at the 12-month horizon, where DMS models appear to be favored. The results are stronger for the choice of data used to estimate the models. Across all horizons, the use of only Great Moderation data to estimate the models appears to be a significant advantage: Not only are the coefficients on post samples significantly different from 0 and negative, they also are more negative than the coefficients associated with averaging over both the full and post samples. The results for the choice of sampling scheme are a bit more muddled but still instructive. At the shortest and longest horizons, combining the recursive and rolling schemes—as suggested by Clark and McCracken (2008)—appears to offer the most advantage in terms of reducing RMSEs. At the other horizons, using the rolling scheme tends to be the best choice.

In the next four rows of panel 1, results for the type of averaging method clearly indicate that the simple equally weighted averaging methods perform significantly worse than the benchmark. At all horizons the coefficient associated with *Equal* is positive and different from 0. Unfortunately, the remaining three rows are not as easy to interpret. While the *MSE*, *Weight*, and *Top 10%* coefficients are typically negative—suggesting that a top 10 percent MSE-weighted average might be best—the coefficients are statistically significant only in a few instances at the longer horizons.

The results in panel 2 (those associated with core inflation) are similar to those for headline inflation, with a few specific differences. The evidence in favor of using the DMS approach to forecasting is stronger at all horizons and significantly so. Again, for all but the shortest horizon, the evidence favors using only Great Moderation data to estimate the model parameters. Similarly, using the rolling scheme or a combination of the rolling and recursive schemes is the preferred approach.

In the seventh through ninth rows of panel 2, the results for the type of averaging method are much sharper than those for headline inflation. In all but the shortest horizons, the simple equally weighted averaging methods perform significantly worse than the benchmark. But at the 1-month horizon, it appears that a simple averaging method does provide significant gains in forecast accuracy and, moreover, those gains are larger than when some form of weighting is used. For horizons longer than 1 month, the coefficients on *Top 10%* are all significantly negative, which along with the negative *MSE* and *Weight* coefficients suggests that a top 10 percent MSE-weighted average might be best.

Results for Real Variables

The results for the real variables (Table 4), particularly those for IP, are quite different from those for the nominal variables. A quick glance at the first six rows of panel 1 indicates quite clearly that the preferred model types for averaging are now IMS forecasting models estimated recursively using the full sample—a sharp contrast to the type of models chosen for both headline and core CPI-based inflation. Moreover, in the next four rows of panel 1, it appears that while some evidence favors MSE weighting relative to BIC weighting, the majority of the evidence suggests even better results would be obtained using the simple equally weighted averages rather than a weighted or top 10 percent weighted average.

The results in panel 2 (those associated with the unemployment rate) are less clear cut than those for IP and even those for headline and core CPI-based inflation. At the 3- and 6-month horizons, the DMS approach to forecasting appears to perform best but at the longest horizon the IMS appears to perform best. Similarly, at the intermediate horizons, using the post (Great Moderation) sample appears to perform best but at the longest horizon the full sample appears to perform best. And while the rolling scheme or a combination of the recursive and rolling schemes tends to perform best at the shortest horizons, the recursive

Table 4

Decomposition Regression of Real Variables

	Forecast horizon				
Variables	1 month	3 month	6 month	12 month	24 month
Industrial production					
DMS	–0.000	0.003***	0.013***	0.024***	0.025***
DMS/IMS	–0.000	0.000	0.004**	0.001	–0.003
Post	0.000	0.000	0.006***	0.006**	–0.001
Post/Full	–0.000	–0.000	0.004**	0.004	–0.001
Roll	0.012***	0.013***	0.049***	0.089***	0.147***
Rec/Roll	0.004***	0.006***	0.025***	0.050***	0.085***
Equal	–0.018***	–0.034***	–0.047***	–0.053***	–0.009**
Weighted	0.001***	0.002***	0.000	–0.000	0.003
Top 10%	0.003***	0.007***	0.007***	0.002	0.003
MSE	–0.002***	–0.006***	–0.003*	0.002	–0.006
N	162	162	162	162	162
Unemployment rate					
DMS	0.000	–0.006***	–0.016***	0.004	0.077***
DMS/IMS	0.000	–0.002	–0.004*	–0.002	0.001
Post	0.001	–0.009***	–0.010***	–0.009***	0.026***
Post/Full	–0.001	–0.007***	–0.008***	–0.009***	0.014
Roll	–0.014***	–0.002	0.013***	0.054***	0.159***
Rec/Roll	–0.009***	–0.004**	0.004	0.023***	0.071***
Equal	–0.054***	–0.107***	–0.127***	–0.088***	–0.151***
Weighted	0.003***	0.003*	0.003	0.002	0.001
Top 10%	–0.007***	–0.005**	–0.006*	0.001	0.034***
MSE	–0.008***	–0.009***	–0.014***	–0.010***	–0.006
N	162	162	162	162	162

NOTE: Each column in each panel provides the coefficients associated with a distinct OLS-estimated version of equation (4). *, **, and *** denote statistical significance at the 10 percent, 5 percent, and 1 percent levels, respectively. BIC, Bayesian information criterion; CPI, consumer price index; DMS, direct multistep; Full, full sample (all available data in that vintage); IMS, iterated multistep; MSE, mean square error; Post, only data that occur starting in January 1983; Rec, recursive window scheme; Roll, rolling window scheme. See text for details.

scheme clearly tends to dominate at the 6-month and longer horizons. Finally, as for IP, it appears that some evidence favors MSE weighting relative to BIC weighting, but the majority of the evidence suggests even better results would come from using one of the equally weighted averages rather than a weighted or top 10 percent weighted average.

Rankings

Tables 3 and 4 give some indication of which model-averaging types should be used and which model classes should be averaged over. However, we emphasize that these results are indicators of average treatment effects across all 162 permutations of averages and model classes. They do not necessarily indicate which permutations actually do perform best. Tables 5 and 6 provide a brief description of the permutations that perform best. In particular, we list the 10 best-performing permutations of averaging methods and model classes and their respective relative RMSEs for each variable and each of the 1-, 3-, and 12-month horizons.[9] In addition, we provide the five worst-performing permutations for the sake of comparison.

The first panel of Table 5 provides the rankings for headline CPI-based inflation. There are several striking features. In line with the results from Table 1, at the 1- and 3-month horizons there are few, if any, gains to model averaging irrelevant of model class. But as the horizon increases to 12 months, gains of roughly 10 percent are available when top 10 percent weighted averages are used; these gains are consistent with the decomposition results from Table 3. In addition, across all horizons, the 10 best-performing permutations use either the rolling scheme or a combination of the rolling and recursive schemes. In contrast, the five worst-performing permutations exclusively use the recursive scheme. Finally, as suggested in Table 3, all but one of the five worst-performing permutations use the simple equally weighted averaging schemes.

The second panel of Table 5 (that associated with core inflation) offers a slightly different picture of the benefits of model averaging relative to model selection. In particular, as in Table 1, model averaging is consistently beneficial at all horizons provided the right permutations of model averages and model classes are used. The 10 best-performing permutations outperform the benchmark by roughly 7 percent at the shortest horizon and by as much as 25 percent at the longest horizon. On the other hand, the 5 worst-performing permutations outperform the benchmark at the 1-month horizon but not at the 3- and 12-month horizons.

Interestingly, the types of model averages that perform best and worst for core inflation coincide nicely with the results in Table 3. At the shortest horizon, equally weighted averages tend to perform best but as the horizon increases, the top 10 percent weighted averages begin to dominate. In general, the class of models to average over also coincides with the results in Table 3: The 10 best-performing permutations are dominated by DMS forecasting models estimated over the post (Great Moderation) sample or an average of the post and full samples, using the rolling scheme or a combination of the rolling and recursive schemes. One result that does not coincide is at the 12-month horizon, where the BIC-weighted averages appear to perform best while the results in Table 3 suggest the MSE-weighted average would perform better.

The first panel of Table 6 provides the rankings for IP growth. As in Table 2, the advantages to model averaging relative to model selection, while feasible, are not particularly large, with a maximum of only 5 percent at the 12-month horizon. As indicated in the decomposition (see Table 4), the equally weighted averages seem to perform best at the 1-month horizon but as the horizon increases to 3 months, top 10 percent weighted averages appear to gain some traction among the best-performing averaging methods—a sharp contrast to the decomposition. Apparently part of the problem is that many of the worst-performing models are also top 10 percent weighted averages; hence, in averaging across all permutations, the decomposition indicates the equally weighted averages should perform better. One point that clearly matches our decomposition is the choice of sampling scheme: Nearly

[9] We present these three horizons for brevity. A complete set of results is available from the authors on request.

Table 5

Ranking of Model Averagings for Nominal Variables

	Forecast horizon					
Ranking	1 month	Relative RMSE	3 month	Relative RMSE	12 month	Relative RMSE
CPI						
1	BIC-DMS/IMS-Post-Rec/Roll	0.993	Top 10% BIC-DMS/IMS-Post-Roll	1.005	Top 10% MSE-DMS-Full-Roll	0.894
2	BIC-IMS-Post-Rec/Roll	0.993	Top 10% BIC-DMS/IMS-Full-Roll	1.005	Top 10% BIC-IMS-Post-Rec/Roll	0.899
3	BIC-DMS-Post-Rec/Roll	0.993	Top 10% BIC-DMS/IMS-Full/Post-Roll	1.005	Top 10% BIC-DMS/IMS-Post-Rec/Roll	0.900
4	Simple Avg.-IMS-Post-Rec/Roll	0.994	Top 10% BIC-IMS-Full-Roll	1.005	Top 10% MSE-DMS-Full-Rec/Roll	0.902
5	Simple Avg.-DMS-Post-Rec/Roll	0.994	Top 10% BIC-IMS-Post-Roll	1.005	Top 10% BIC-IMS-Full/Post-Rec/Roll	0.911
6	Simple Avg.-DMS/IMS-Post-Rec/Roll	0.994	Top 10% BIC-IMS-Full/Post-Roll	1.005	Top 10% BIC-DMS/IMS-Full/Post-Rec/Roll	0.913
7	BIC-DMS/IMS-Full/Post-Rec/Roll	0.994	Top 10% BIC-DMS-Post-Roll	1.006	Top 10% MSE-DMS/IMS-Full-Roll	0.915
8	BIC-IMS-Full/Post-Rec/Roll	0.994	Top 10% BIC-DMS-Full-Roll	1.006	Top 10% MSE-DMS/IMS-Full-Rec/Roll	0.919
9	BIC-DMS-Full/Post-Rec/Roll	0.994	Top 10% BIC-DMS-Full/Post-Roll	1.006	MSE-DMS-Full-Roll	0.919
10	Top 10% BIC-DMS-Post-Rec/Roll	0.994	Median-DMS-Full/Post-Roll	1.006	BIC-DMS-Post-Rec/Roll	0.921
⋮						
158	Simple Avg.-IMS-Full-Rec	1.032	BIC-IMS-Full-Rec	1.113	Median-DMS/IMS-Full-Rec	1.195
159	Simple Avg.-DMS-Full-Rec	1.032	Median-DMS-Full-Rec	1.113	Median-DMS-Full-Rec	1.196
160	Median-DMS-Full-Rec	1.040	Simple Avg.-IMS-Full-Rec	1.114	Simple Avg.-DMS/IMS-Full-Rec	1.197
161	Median-DMS/IMS-Full-Rec	1.040	Median-DMS/IMS-Full-Rec	1.117	Median-IMS-Full-Rec	1.201
162	Median-IMS-Full-Rec	1.040	Median-IMS-Full-Rec	1.120	Simple Avg.-DMS-Full-Rec	1.212
Core CPI						
1	Simple Avg.-DMS-Full-Rec/Roll	0.929	Top 10% BIC-DMS-Full-Roll	0.928	Top 10% BIC-DMS-Post-Roll	0.761
2	Simple Avg.-DMS/IMS-Full-Rec/Roll	0.929	Top 10% BIC-DMS-Post-Roll	0.928	Top 10% BIC-DMS-Full-Roll	0.761
3	Simple Avg.-IMS-Full-Rec/Roll	0.929	Top 10% BIC-DMS-Full/Post-Roll	0.928	Top 10% BIC-DMS-Full/Post-Roll	0.761
4	MSE-DMS-Full-Rec/Roll	0.930	MSE-IMS-Post-Roll	0.929	Top 10% BIC-DMS-Full-Rec/Roll	0.769
5	MSE-DMS/IMS-Full-Rec/Roll	0.930	MSE-IMS-Full/Post-Roll	0.929	Top 10% BIC-DMS-Full/Post-Rec/Roll	0.780
6	MSE-IMS-Full-Rec/Roll	0.930	Top 10% BIC-DMS/IMS-Full-Rec/Roll	0.930	Top 10% BIC-DMS-Post-Rec/Roll	0.783
7	Simple Avg.-DMS-Full/Post-Rec/Roll	0.931	Top 10% BIC-DMS-Full-Rec/Roll	0.930	MSE-DMS-Post-Roll	0.797
8	Simple Avg.-DMS/IMS-Full/Post-Rec/Roll	0.931	Top 10% BIC-DMS-Full/Post-Rec/Roll	0.931	MSE-DMS-Full/Post-Roll	0.798
9	Simple Avg.-IMS-Full/Post-Rec/Roll	0.931	MSE-DMS/IMS-Post-Roll	0.931	Top 10% MSE-DMS-Full-Roll	0.799
10	Top 10% MSE-DMS-Full-Rec/Roll	0.932	MSE-DMS/IMS-Full/Post-Roll	0.931	BIC-DMS-Full/Post-Roll	0.801
⋮						
158	Top 10% MSE-DMS/IMS-Post-Roll	0.966	BIC-DMS/IMS-Full-Rec	1.105	Median-IMS-Full-Rec	1.397
159	Simple Avg.-IMS-Full-Rec	0.967	Median-IMS-Full-Rec	1.108	Top 10% BIC-DMS/IMS-Full-Rec	1.436
160	Simple Avg.-DMS/IMS-Full-Rec	0.967	BIC-IMS-Full-Rec	1.115	BIC-DMS/IMS-Full-Rec	1.478
161	Simple Avg.-DMS-Full-Rec	0.967	Simple Avg.-DMS/IMS-Full-Rec	1.116	BIC-IMS-Full-Rec	1.484
162	Top 10% MSE-IMS-Post-Roll	0.967	Simple Avg.-IMS-Full-Rec	1.133	Simple Avg.-IMS-Full-Rec	1.523

NOTE: The values indicate the relative RMSEs of the model-averaging permutation in each row relative to the baseline. Avg., averaging; BIC, Bayesian information criterion; CPI, consumer price index; DMS, direct multistep; Full, full sample (all available data in that vintage); IMS, iterated multistep; MSE, mean square error; Post, only data that occur starting in January 1983; Rec, recursive window scheme; Roll, rolling window scheme. See text for details.

Table 6

Ranking of Model Averagings for Real Variables

	Forecast horizon					
Ranking	1 month	Relative RMSE	3 month	Relative RMSE	12 month	Relative RMSE
Industrial production						
1	Median-IMS-Full/Post-Rec	0.978	Top 10% BIC-DMS-Full-Rec	0.962	BIC-IMS-Full-Rec	0.949
2	Median-DMS-Full/Post-Rec	0.978	Top 10% MSE-IMS-Full/Post-Rec	0.964	Simple Avg.-IMS-Full-Rec	0.949
3	Median-DMS-Full/Post-Rec	0.978	Top 10% MSE-DMS-Full-Rec	0.964	BIC-DMS/IMS-Full-Rec	0.949
4	Simple Avg.-DMS-Full/Post-Rec	0.981	Top 10% BIC-IMS-Full-Rec	0.964	Top 10% BIC-DMS/IMS-Full-Rec	0.949
5	Simple Avg.-DMS/IMS-Full/Post-Rec	0.981	Top 10% MSE-DMS/IMS-Full-Rec	0.965	Median-IMS-Full-Rec	0.950
6	Simple Avg.-IMS-Full/Post-Rec	0.981	Simple Avg.-IMS-Post-Rec	0.965	MSE-IMS-Full-Rec	0.950
7	Top 10% BIC-DMS-Full-Rec	0.981	Median-IMS-Post-Rec	0.965	MSE-DMS/IMS-Full-Rec	0.951
8	Top 10% BIC-IMS-Full-Rec	0.981	Top 10% MSE-IMS-Full-Rec	0.965	Simple Avg.-DMS/IMS-Full-Rec	0.952
9	Top 10% BIC-DMS/IMS-Full-Rec	0.981	Top 10% MSE-DMS/IMS-Full/Post-Rec	0.965	Median-DMS/IMS-Full-Rec	0.952
10	BIC-DMS/IMS-Full/Post-Rec	0.981	MSE-IMS-Post-Rec	0.965	Simple Avg.-IMS-Full/Post-Rec	0.954
⋮						
158	Top 10% BIC-IMS-Full-Roll	1.001	Top 10% BIC-IMS-Full-Roll	0.996	MSE-DMS-Full-Roll	1.088
159	Top 10% BIC-DMS-Full-Roll	1.001	Top 10% BIC-IMS-Post-Roll	0.996	Median-DMS-Full/Post-Roll	1.090
160	Top 10% BIC-DMS-Full/Post-Roll	1.001	Top 10% MSE-DMS-Post-Roll	0.997	Median-DMS-Post-Roll	1.090
161	Top 10% BIC-IMS-Post-Roll	1.001	Top 10% BIC-DMS-Full/Post-Rec/Roll	0.997	Median-DMS-Full-Roll	1.090
162	Top 10% BIC-IMS-Full/Post-Roll	1.001	Top 10% BIC-DMS-Post-Rec/Roll	0.998	Top 10% MSE-DMS-Full-Roll	1.093
Unemployment rate						
1	Top 10% MSE-IMS-Post-Roll	0.905	Top 10% MSE-DMS-Post-Rec/Roll	0.848	Top 10% MSE-DMS-Post-Rec	0.855
2	Top 10% MSE-DMS/IMS-Post-Roll	0.905	Top 10% MSE-DMS-Post-Rec	0.849	Top 10% MSE-DMS-Post-Rec/Roll	0.860
3	Top 10% MSE-DMS-Post-Roll	0.905	Top 10% MSE-DMS-Post-Roll	0.851	Top 10% BIC-DMS-Full-Rec	0.863
4	Top 10% MSE-IMS-Full/Post-Roll	0.913	Top 10% MSE-DMS/IMS-Post-Rec/Roll	0.851	Top 10% BIC-DMS-Post-Rec	0.867
5	Top 10% MSE-IMS-Post-Rec/Roll	0.913	Top 10% MSE-DMS/IMS-Post-Rec	0.851	Top 10% MSE-DMS-Full/Post-Rec	0.867
6	Top 10% MSE-DMS/IMS-Full/Post-Roll	0.913	Top 10% MSE-DMS/IMS-Post-Roll	0.853	Top 10% BIC-DMS-Full/Post-Rec	0.874
7	Top 10% MSE-DMS/IMS-Post-Rec/Roll	0.914	Top 10% MSE-IMS-Post-Rec	0.854	Top 10% MSE-DMS-Full/Post-Rec/Roll	0.875
8	Top 10% MSE-DMS-Full/Post-Roll	0.914	Top 10% MSE-IMS-Post-Rec/Roll	0.854	Top 10% MSE-DMS/IMS-Post-Rec	0.877
9	Top 10% MSE-DMS-Post-Rec/Roll	0.915	Top 10% MSE-IMS-Post-Roll	0.856	Top 10% MSE-DMS/IMS-Full/Post-Rec	0.892
10	Top 10% MSE-IMS-Full/Post-Rec/Roll	0.921	Top 10% MSE-DMS-Full/Post-Roll	0.863	Top 10% BIC-DMS/IMS-Full-Rec	0.892
⋮						
158	Simple Avg.-DMS-Post-Rec	0.951	Top 10% BIC-IMS-Post-Roll	0.899	MSE-DMS-Full-Roll	0.987
159	Simple Avg.-IMS-Post-Rec	0.951	Top 10% BIC-IMS-Full-Roll	0.899	Median-DMS-Full-Roll	0.989
160	Median-DMS-Post-Rec	0.952	Top 10% BIC-IMS-Full/Post-Roll	0.899	Median-DMS-Full/Post-Roll	0.989
161	Median-DMS/IMS-Post-Rec	0.952	Median-DMS-Full-Rec	0.899	Median-DMS-Post-Roll	0.989
162	Median-IMS-Post-Rec	0.952	Top 10% BIC-IMS-Full-Rec/Roll	0.900	Top 10% MSE-DMS-Full-Roll	0.995

NOTE: The values indicate the relative RMSEs of the model-averaging permutation in each row relative to the baseline. Avg., averaging; BIC, Bayesian information criterion; CPI, consumer price index; DMS, direct multistep; Full, full sample (all available data in that vintage); IMS, iterated multistep; MSE, mean square error; Post, only data that occur starting in January 1983; Rec, recursive window scheme; Roll, rolling window scheme. See text for details.

all the best-performing permutations average across models estimated with the recursive scheme while all the worst-performing permutations average across models estimated with the rolling scheme.

In the second panel of Table 6 (that associated with forecasts of the unemployment rate), a few things are immediately apparent. First, model averaging uniformly improves forecast accuracy across all horizons and all permutations. In fact, at the 3-month horizon, the worst-performing model average provides an improvement of 10 percent relative to the benchmark. Also, across all horizons the 10 best-performing types of model averaging are of the top 10 percent form. This is in sharp contrast with the decomposition results, which predicted that the equally weighted averages tended to perform best. Even so, as for the decomposition shown in Table 4, it appears that at the shortest horizon the rolling scheme appears to perform best but as the horizon increases the recursive scheme becomes preferred. At the 12-month horizon, the 5 worst-performing permutations use the rolling scheme.

CONCLUSION

We use the ALFRED real-time database to provide empirical evidence on the real-time benefits of model averaging monthly-frequency forecasts of headline and core CPI-based inflation, growth in IP, and the unemployment rate. Our results support those discussed in much of the literature on forecasting: Model averaging typically improves forecast accuracy relative to a benchmark chosen using model selection. Even so, we emphasize a different point that is typically glossed over in the literature on forecast averaging: The choice of models averaged across can greatly influence the efficacy of the averaging methods.

This of course raises the question of how to choose the correct class of models to average across. Based upon a novel decomposition of the benefits of forecast averaging relative to using model-selection methods, a few rules of thumb seem evident. First, DMS forecasting models estimated over the post (Great Moderation) sample (or an average of the post and full samples) using the rolling scheme (or a combination of the rolling and recursive schemes) seem to perform best for forecasting either headline or core CPI-based inflation. Second, averaging over models estimated using the recursive scheme (or an average of the rolling and recursive) seems to perform best for forecasting either IP growth or the unemployment rate. Third, the top 10 percent averaging approach frequently provides the best improvements in forecast accuracy, but it is not immune to poor performance relative to equally weighted averages because past model performance does not always ensure future model performance.

REFERENCES

Aiolfi, Marco and Timmermann, Allan. "Persistence in Forecasting Performance and Conditional Combination Strategies." *Journal of Econometrics*, 2006, *135*(1-2), pp. 31-53.

Clark, Todd E. and McCracken, Michael W. "Improving Forecast Accuracy by Combining Recursive and Rolling Forecasts." *International Economic Review*, 2008, *50*(2), pp. 363-95.

Clark, Todd E. and McCracken, Michael W. "Averaging Forecasts from VARs with Uncertain Instabilities." *Journal of Applied Econometrics*, January/February 2010, *25*(1), pp. 5-29.

D'Agostino, Antonello; Giannone, Domenico and Surico, Paolo. "(Un)Predictability and Macroeconomic Stability." Working Paper Series No. 605, European Central Bank, April 2006; www.ecb.int/pub/pdf/scpwps/ecbwp605.pdf.

Elliott, Graham and Timmermann, Allan. "Optimal Forecast Combinations Under General Loss Functions and Forecast Error Distributions." *Journal of Econometrics*, September 2004, *122*(1), pp. 47-79.

Faust, Jon and Wright, Jonathan H. "Comparing Greenbook and Reduced Form Forecasts Using a Large Realtime Dataset." *Journal of Business and Economic Statistics*, October 2009, *27*(4), pp. 468-79.

Garratt, Anthony; Koop, Gary and Vahey, Shaun P. "Forecasting Substantial Data Revisions in the Presence of Model Uncertainty." *Economic Journal*, 2008, *118*(53), pp. 1128-44.

Giannone, Domenico; Reichlin, Lucrezia and Small, David. "Nowcasting: The Real Time Informational Content of Macroeconomic Data Releases." *Journal of Monetary Economics*, May 2008, *55*(4), pp. 665-76.

Hansen, Bruce E. "Least-Squares Forecast Averaging." *Journal of Econometrics*, 2008, *146*(2), pp. 342-50.

Hansen, Bruce E. "Multi-Step Forecast Model Selection." Presented at the 20th Annual Meetings of the Midwest Econometrics Group, October 1-2, 2010, Olin Business School, Washington University in St. Louis; http://apps.olin.wustl.edu/MEGConference/Files/pdf/2010/61.pdf.

Kapetanios, George; Labhard, Vincent and Price, Simon. "Forecasting Using Bayesian and Information Theoretic Model Averaging: An Application to U.K. inflation." *Journal of Business and Economic Statistics*, January 2008, *26*, pp. 33-41.

Kascha, Christian and Ravazzolo, Francesco. "Combining Inflation Density Forecasts." *Journal of Forecasting*, 2010, *29*(1-2), pp. 231-50.

Levin, Andrew T. and Piger, Jeremy M. "Is Inflation Persistence Intrinsic in Industrial Economies?" Working Paper No. 2002-0231; Federal Reserve Bank of St. Louis; October 2002, revised November 2003; http://research.stlouisfed.org/wp/2002/2002-023.pdf.

Marcellino, Massimiliano; Stock, James H. and Watson, Mark W. "A Comparison of Direct and Iterated AR Methods for Forecasting Macroeconomic Time Series." *Journal of Econometrics*, November-December 2006, *135*(1-2), pp. 499-526.

Smith, Jeremy and Wallis, Kenneth F. "A Simple Explanation of the Forecast Combination Puzzle." *Oxford Bulletin of Economics and Statistics*, June 2009, *71*(3), pp. 331-55.

Stock, James H. and Watson, Mark. "Combination Forecasts of Output Growth in a Seven-Country Data Set." *Journal of Forecasting*, September 2004, *23*(6), pp. 405-30.