

Comparing Greenbook and Reduced Form Forecasts using a Large Realtime Dataset

Jon Faust*

Jonathan H. Wright†

*Department of Economics, Johns Hopkins University, Baltimore MD 21218 and Division of International Finance, Federal Reserve Board, Washington DC 20551, phone: (202) 452-2328, e-mail: faustj@jhu.edu

†Division of Monetary Affairs, Federal Reserve Board, Washington DC 20551, phone: (202) 452-3605, e-mail: jonathan.h.wright@frb.gov.

Abstract: Many recent papers have found that atheoretical forecasting methods using many predictors give better predictions for key macroeconomic variables than various small-model methods. The practical relevance of these results is open to question, however, because these papers generally use *ex-post* revised data not available to forecasters and because no comparison is made to best actual practice. We provide some evidence on both of these points using a new large dataset of vintage data synchronized with the Fed's Greenbook forecast. This dataset consists of a large number of variables as observed at the time of each Greenbook forecast since 1979. We compare realtime, large-dataset predictions to both simple univariate methods and to the Greenbook forecast. For inflation we find that univariate methods are dominated by the best atheoretical large dataset methods and that these, in turn, are dominated by Greenbook. For GDP growth, in contrast, we find that once one takes account of Greenbook's advantage in evaluating the current state of the economy, neither large dataset methods nor the Greenbook process offers much advantage over a univariate autoregressive forecast.

1. Introduction

In recent years, researchers have investigated many different atheoretical ways of forecasting an economic time series using a large number of predictors—say, 40 or more (e.g., Bernanke, Boivin and Elias, 2005; Boivin and Ng, 2006; Forni, Hallin, Lippi and Reichlin, 2005; Giannone Reichlin and Sala, 2004; Stock and Watson, 1999, 2002, 2003, 2005). When the number of predictors is large relative to the available sample size, one must constrain the estimation of the forecasting model in some way in order to avoid the perils of overfitting, and the many methods differ mainly in how they do so. A consistent result in this work is that these large-dataset methods outperform various naive and semi-sophisticated benchmarks including random walk forecasts, simple univariate time series models, and, in some cases, simple models motivated by economic theory.

While these pioneering results are tantalizing, their importance for practical forecasting is difficult to assess for two related reasons. First, arguably none of the benchmarks used in this research is used for practical forecasting. While the simple benchmarks are probably the right starting point for assessing new methods, ultimately we care whether the new methods outperform standard practice or best practice in forecasting. Second, comparison to real-world forecasting methods is complicated by the issue of data revisions. Many macro time series are heavily revised through time. Real-world forecasts use the noisy early vintages available in realtime; large-dataset forecasting research has been conducted almost exclusively with a single vintage of revised data.

Bernanke and Boivin (2003) provide a notable exception. They use a realtime dataset to assess various large dataset forecasting methods and they include a comparison to the Fed’s Greenbook forecast. Two results are especially interesting and provocative. First, factor models generally predict less well in their 78-variable realtime dataset than using the 215 series of fully-revised data used by Stock and Watson (2002). This appears to owe mainly to the specific variables in, and the larger size of the Stock and Watson dataset, because the factor models also predict less well in the revised version of the 78-

variable dataset. Second, the Federal Reserve’s Greenbook forecast seems to outperform all the other methods considered—regardless of the dataset employed or whether revised data are used.

In this paper, we take up some of the questions raised by Bernanke and Boivin (2003), using a unique set of vintage data associated with the Greenbook forecast, which is prepared for each FOMC meeting. The dataset has a snapshot of a large number of macroeconomic time series as they existed at the time of 145 Greenbooks between March 1980 and December 2000. These data allow us to create realtime large dataset forecasts using information sets that are precisely synchronized with the Greenbook. Using these data, we compare four classes of forecasting models: i) Greenbook, ii) univariate benchmarks, iii) large dataset, model averaging, iv) large dataset, factor methods.

While the comparisons among the various time series methods are of interest, we have several reasons to be especially interested in comparing all the methods to Greenbook. First, the Greenbook forecast plays an important role in the setting of U.S. monetary policy, making the quality of this forecast directly interesting. Second, much earlier work suggests that the Greenbook forecast has generally been at or near the frontier of best performance in forecasting. Romer and Romer (2000) found that Greenbook outperformed private sector surveys; and Sims (2002) found generally favorable results, especially for Greenbook inflation forecasts; Bernanke and Boivin (2003) reach the same conclusion. Based on these results we take Greenbook to be a decent proxy for best practice, allowing us to investigate whether any of these methods fares well relative to a measure of best actual practice. Finally, the Greenbook is a subjective forecast based on an immense range of information processed through an economics-influenced subjective filter (see, e.g., Reifschneider, Stockton and Wilcox (1997)). Thus, we can view the large-dataset-forecast versus Greenbook comparison as a test of the atheoretical use of a large dataset versus sophisticated use of an immense dataset.

Our unique dataset, allows us to add an important dimension to comparisons that

have been done to date. Sims (2002) has suggested that the good properties of Greenbook might flow from the great effort the Fed makes to evaluate the current and recent past state of the economy at the time of the forecast—sometimes called nowcasting and backcasting. For example, by mirroring key elements of the data construction machinery of the Bureau of Economic Analysis, the Fed staff forms a relatively precise estimate of what BEA will announce for the previous quarter’s GDP even before it is announced. Further, the staff adjusts the estimate of the current state for certain large transitory events that have already been observed, such as dock strikes or hurricanes. A better estimate of the current state might translate into forecasting advantages over moderate horizons. To put it most starkly, one might conjecture that the Fed’s forecasting advantage stems purely from measuring the current state, with little or no advantage over atheoretical methods in projecting what the current state implies for the future.

We assess this conjecture in a simple way. Consider the vintage data available in quarter t , so that the released values of many data series end in quarter $t - 1$ or $t - 2$. Our dataset includes a Fed forecast of each of the many variables, so we can append the Fed forecast to the actual data, and bring the data all up to quarter $t - 1$, or to t , or to $t + j$. We can then use this updated dataset to form the various time series forecasts. In short, we give the time series models the benefit of the Fed’s perspective on the current state and see if any forecasting advantage remains for Greenbook. We call the point to which we update the data the “jumping-off point” for the time series methods. We consider jumping-off points from quarter $t - 1$ through $t + 3$.

For inflation, for all jumping-off points, we find that Greenbook dominates the best large model methods, which, in turn, dominate the univariate benchmarks. For inflation: Greenbook beats large beats small at all horizons and jumping-off points. Our results are very different for output. We find that Greenbook has an advantage in forecasting the current quarter, but that once we move the jumping-off point up to the current quarter, the best atheoretical methods and Greenbook perform comparably. Indeed, no

method clearly dominates the naive benchmark of a univariate autoregression. For GDP at all horizons so long as the jumping-off point is the current quarter or later: nothing beats a univariate AR. We believe that the differential predictability of inflation and output growth is an important stylized fact of the U.S. case. We discuss some possible implications in the conclusion.

We also find some general patterns in performance among the various large model methods. The model averaging methods, in which we take an average of forecasts from a large number of simple bivariate models, are often best and are never far from the best methods. The various factor model approaches almost never outperform the model averaging methods and the performance of the factor methods is much less consistent, and some factor methods sometimes perform very poorly. One way of viewing this result is that among the large dataset methods, those that are *minimally multivariate*—each underlying model in the averaging methods is bivariate—perform most robustly.

From the outset, we recognize that a paper paying great attention to realtime issues with large numbers of variables and examining a wide range of models has the potential to be both long and tedious. To reduce this burden, many details of our data and methods and many supplementary results have been relegated to a web appendix available at <http://e105.org/faustj/download/faustWrightGBApp.pdf>.

The plan for the remainder of this paper is as follows. Sections 2 and 3 describe our realtime dataset and methods, respectively. Section 4 contains the main results; section 5 addresses a number of additional topics. Section 6 concludes.

2. The Realtime Dataset

Before each FOMC meeting, the Federal Reserve staff prepares a briefing document called the Greenbook, which contains a staff forecast of the macroeconomy. While the Greenbook forecast is subjective, a large-scale econometric model has long been one tool used in the Greenbook process. Since 1996, the FRB/US model has been used.

Before that, a model called variously the FMP (Federal Reserve-MIT-Penn) or MPS (MIT-Penn-SSRC) was employed. Since 1979, the model database has been archived electronically on the date that the Greenbook is published, a few days before the FOMC meeting. These archived databases contain data on many variables of which some (but not all) are forecasted in the Greenbook. The databases contain the historical sample of each variable in the model as observed at that time, and forecasts. While the forecasts are generated from the econometric model, for variables reported in the Greenbook, an add-factor is included to force the forecasts to match those in the Greenbook. Thus these archives comprise a large vintage dataset that is perfectly synchronized with the Greenbook forecast itself. Our datasets include, and we use, forecasts for all variables. Only a subset of the variables we use is actually reported in Greenbook. Strictly then, forecasts for the remaining variables are best thought of as model forecasts produced as an input to, or by-product of, the Greenbook forecasting process.

We extracted 145 vintages of data from these electronic archives covering Greenbooks from March 1980 to December 2000. The data used in this paper stop in 2000 because it is Federal Reserve policy that the Greenbook forecast cannot be released until 5 years after the forecast date. A few vintages are lost or never existed, including those for early 1996 during the transition from the MPS model to the FRB/US model. The maximum forecast horizon varies from one Greenbook to the next, but we considered only Greenbooks for which the forecast horizon goes out to at least 5 quarters. The publication dates for our 145 vintages of data are listed in the web appendix.

Ideally, we might like to have the same forecast variables in each vintage. Unfortunately, the variables available in any given vintage vary a good bit. There is a large break in the list of available variables at the time of the major model revision in 1996. Finally, the amount of historical data for some series varies a bit across vintages. Further, the nature of the model databases presents an additional variable selection problem. Many variables in the databases are constructed from other variables based on identities, other

transforms of included variables, and slightly different versions of included variables (e.g., foreign sector variables on both the NIPA and BOP basis).

We extracted 109 macro time series, from these databases. We omitted variables that are constructed from other variables and different definitions of the same variable. From each vintage, we keep any of these 109 that have historical data back to 1960Q2. We augment each vintage with the dividend-exclusive returns on the S&P500 stock index. To each vintage we added only observations that would have been available at the Greenbook publication date. These data are of course not subject to revision. Our resulting vintage databases have relatively extensive coverage of income and spending, prices, interest rates, and employment, as well as several data series on stocks of durables, exchange rates and foreign output, prices and interest rates (see the list in the Appendix).

In total, the number of series in each vintage ranges from 47 to 80, with an average of 67. This is a large number of predictors, although it falls short of the very large datasets of more than 100 variables that are used in some papers.

In the end, we have 145 vintages containing a varying list of variables, each of which has history back to 1960Q2 and for which we have 5 quarter forecasts. The vast majority of our predictors are available out to the Greenbook forecast horizon. As is usual, we transformed many of the series so that the transformed series is arguably stationary. Some series are kept in levels (no transformation), others are in differences and others are in log differences.

The series that we forecast in this paper are the quarterly inflation rate as measured by the GNP/GDP deflator and the quarterly real growth rate (GNP/GDP). As is standard, we use GNP during and before 1991, and GDP subsequently. The inflation and growth rates are computed as $400 \log(x_t/x_{t-1})$ where x is the price or output series.

Forecast errors are calculated as forecast value minus actual, but for variables that are repeatedly and indefinitely revised with evolving definitions, an issue arises as to what to treat as the actual. For the national income and product accounts (NIPA),

the source of our forecast variables, the first data release (known as the *advance* release) comes out about one month after the end of the quarter to which the data refer. The data are then revised in the *preliminary* and *final* releases, which incorporate more source data, and are released about two and three months after the end of the quarter to which the data refer, respectively. The data are then revised repeatedly in through annual and then benchmark revisions, with the latter incorporating conceptual and definitional changes. It makes little sense to evaluate whether Greenbook or time series models predict definitional changes; the Fed explicitly does not attempt to do so. Thus, we follow Tulip (2005) in measuring actual realized inflation and growth by the data as recorded in the realtime dataset of the Federal Reserve Bank of Philadelphia two quarters after the quarter to which the data refer. This will typically, but not always, correspond to the NIPA ‘final’ release. The web appendix reports results for two alternative definitions of actuals, and shows the same broad results we emphasize below.

3. Methods

3.1 The forecasting models

We construct forecasts using 11 time series models taken from the literature. Call the variable to be forecast y_t and a collection of potential predictors $\{x_{it}\}_{i=1}^n$. When forming a forecast in quarter T we observe data that were available in quarter T . We consider forecasts for y_{T+h} , $h = 0, \dots, 5$.

The variable y_t is either the annualized quarterly inflation rate or the annualized quarterly growth rate of output. For inflation, we report results for both forecasts that impose a unit root in y_t and forecasts that do not (that is the models are posed in terms of either inflation or the change in inflation). Several authors have argued that treating inflation as nonstationary, perhaps reflecting infrequent shifts in the Fed’s implicit inflation target, may improve forecasting. For output growth, we report results only for forecasts that treat the growth rate of output as stationary.

We briefly describe here our baseline set of 11 models taken from the literature (Table 1). For a more complete description of the models, see the original citations.

1. A random walk model (RW). This takes y_{T-1} as the forecast for y_{T+h} , $h = 0, \dots, 5$. This is close to, but not quite the same as, the forecast for inflation considered by Atkeson and Ohanian (2001), who take $\frac{1}{4}\sum_{j=0}^3 y_{T-j}$ as the forecast for y_{T+h} .
2. Recursive autoregression (RAR). We estimate $y_t = \rho_0 + \sum_{j=1}^p \rho_j y_{t-j} + \varepsilon_t$ (we use $p = 4$). The h -period forecast is constructed by recursively iterating the one-step forecast forward.
3. Direct forecast from autoregression (DAR). For each h , we estimate $y_{t+h} = \rho_0 + \sum_{j=1}^p \rho_j y_{t+1-j} + \varepsilon_t$ (we use $p = 4$). Each h -step forecast is a one-step forecast from the model for the appropriate h . The RAR forecast will asymptotically outperform the direct model if the AR(4) model is correctly specified, but the direct forecast may be more robust to misspecification, as discussed by Marcellino, Stock and Watson (2006).
4. An unobserved component stochastic volatility model (SV). The model is univariate: $y_t = \tau_t + \eta_t^T$ and $\tau_t = \tau_{t-1} + \eta_t^P$ where η_t^T is $iidN(0, \sigma_{T,t}^2)$, η_t^P is $iidN(0, \sigma_{P,t}^2)$, $\log(\sigma_{T,t}^2) = \log(\sigma_{T,t-1}^2) + \psi_{1,t}$, $\log(\sigma_{P,t}^2) = \log(\sigma_{P,t-1}^2) + \psi_{2,t}$ and $(\psi_{1,t}, \psi_{2,t})'$ is $iidN(0, I_2)$. The model is estimated by Markov Chain Monte Carlo. The forecast of y_{T+h} is the filtered estimate of τ_T . Stock and Watson (2007) find that this model provides good forecasts for inflation.
5. Equal-weighted averaging (EWA). We first estimate and forecast using n simple models, each of the form $y_{t+h} = \rho_0 + \sum_{j=1}^p \rho_j y_{t+1-j} + \beta_i x_{it} + \varepsilon_{it}$ for $i = 1, \dots, n$ (we use $p = 4$). Letting \hat{y}_{T+h}^i be the forecast of y_{T+h} from the i th model, the EWA forecast is $n^{-1}\sum_{i=1}^n \hat{y}_{T+h}^i$. This method was first proposed by Bates and Granger (1969) and its surprising empirical success is part of the folklore of forecasting. Stock and Watson (2003) among others find continuing support for the folklore.
6. Bayesian model averaging (BMA). In this method, described in more detail by

Wright (2003), we assign a prior over the parameters of the n models used in EWA, just described; and a flat prior that each model is equally likely to be true. The prior for the model parameters follows Fernandez, Ley and Steel (2001). Write each model as $y_{t+h} = \lambda_i' w_{it} + \varepsilon_{it}$, where $\varepsilon_{it} \sim N(0, \sigma^2)$, let the prior for λ_i conditional on σ be $N(\bar{\lambda}, \phi(\sigma^2 \sum_{t=1}^T w_{it} w_{it}')^{-1})$ and the marginal prior for σ be proportional to $1/\sigma$. The models are then estimated and the forecast from each is evaluated at the posterior mean for the parameters. Finally, these n forecasts are then combined in a weighted average with weights determined by the posterior probability that each model is correct. The prior has a hyperparameter, ϕ , that determines how much the model weights are likely to vary from equal weighting. Our baseline results set $\phi = 2$.

The theoretical justification of this method relies on strictly exogenous regressors and iid errors—assumptions that are patently false in our application. Earlier work (Koop and Potter (2003) and Wright (2003)) shows that the method works well in cases like the one at hand, however, and we simply view BMA as a pragmatic shrinkage device.

7. Factor augmented autoregression (FAA). For each h , we estimate $y_{t+h} = \rho_0 + \sum_{j=1}^p \rho_j y_{t+1-j} + \sum_{i=1}^m \gamma_i z_{it} + \varepsilon_t$ where $\{z_{it}\}_{i=1}^m$ are the first m principal components of $\{x_{it}\}_{i=1}^n$. The predictors are first standardized to have mean zero and unit variance. We use $p = 4$, $m = 3$. The forecasts are then constructed as in the direct AR forecast.

8. Factor augmented vector autoregression (FAV). This uses the VAR $\xi_t = \mu_0 + \sum_{j=1}^{\bar{p}} \mu_j \xi_{t-j} + \varepsilon_t$, where $\xi_t = (y_t, z_{1t}, z_{2t}, \dots, z_{mt})'$. We set $\bar{p} = 1$ and $m = 3$. The model can be estimated and iterated forward to provide a forecast of y_{T+h} . This method was proposed by Bernanke, Boivin and Elias (2005).

9. An integrated factor augmented VAR (IFV). This is just as in the factor augmented VAR except that the forecast variable is differenced before estimation/forecasting, that is, $\xi_t = (\Delta y_t, z_{1t}, z_{2t}, \dots, z_{mt})'$. We set $\bar{p} = 1$ and $m = 3$. The model is estimated and iterated forward to provide a forecast of y_{T+h} .

10. Dynamic factor model (DF). We use the model described in detail by Forni, Hallin, Lippi and Reichlin (2005). We take 3 dynamic factors and 15 static factors, estimating the spectral density matrix of the data using a Bartlett window with truncation lag equal to the square root of the sample size. We are grateful to Mario Forni for providing us with the code for implementing this procedure.

11. Factor-spanned variable selection (FVS). This uses the $A(j)$ test of Bai and Ng (2006) to select a single variable from the set of possible predictors. This test identifies which variable is closest to the first 3 principal components of $\{x_{it}\}_{i=1}^n$. The forecast is then made by augmenting the direct autoregression with $p = 4$ with this chosen predictor. Note that unlike in the factor-augmented autoregression, there is no parameter estimation error in the extra variable that is being added to the autoregression. Otherwise, there is parameter estimation error that is asymptotically negligible only if the temporal dimension is large relative to the number of predictors. Armah and Swanson (2007) evaluate this forecasting method in a large realtime dataset and find favorable results.

The RW, SV and IFV methods impose a unit root in the variable being forecast, and are used for predicting inflation, but not output growth. We leave consideration of nonlinear models to future work, but note that Marcellino (2006) has found that some nonlinearity can be helpful for forecasting inflation, though not growth.

Finally, we note that throughout this paper we consider forecasting of one-quarter growth or one-quarter inflation, h quarters hence. Many authors instead consider the prediction of cumulative growth or inflation from quarter $t - 1$ to quarter $t + h$, or four-quarter growth or inflation ending h quarters hence. Since one of our main purposes is to assess the relative information content of Greenbook forecasts at different horizons, it seemed best to report results in terms of one-quarter growth at different horizons. The other measures confound short- and longer-term predictive ability.

Each model is re-estimated and forecasted on each available vintage using data from 1960Q2 to the most recently available and possibly supplemented by the Fed forecast to bring all relevant data off to the jumping-off point.

3.2 *Bootstrap p-values*

We use the RMSPE as the primary criterion for judging forecast quality, and want to evaluate the statistical significance of differences in the RMSPE of Greenbook and the time series models. The appropriate p -values depend on whether the forecasting models being compared are nested or non-nested under the null of equal predictive accuracy. We think it is appropriate to assume that the time series forecasts are not nested in the Greenbook. Thus, we base our comparison on the Diebold-Mariano (DM) statistic, and assume that the conventional DM assumptions are satisfied so that this statistic will be asymptotically normal. Under these assumptions, we conjecture that bootstrap p -values based on a suitable re-sampling of the forecast errors should also be valid to first order and may have better small sample properties. Corradi and Swanson (2007) find that a residual-based bootstrap works well in Monte-Carlo simulations. We report bootstrap p -values based on resampling blocks of forecast errors using the moving-blocks re-sampling scheme of Künsch (1989) and Liu and Singh (1992). The block length is 10, corresponding to a span of a bit more than one year of forecasts. The bootstrap p -values come from inverting the percentile confidence interval for the DM-statistic (and so are percentile- t -style p -values). The asymptotic p -values for the DM test using Newey-West standard errors with a lag length of 10 (not shown) are generally similar to those from the bootstrap. The fact that we have multiple data vintages does not pose any special problems for our bootstrapping scheme because it resamples from forecast errors. Note that our method would not be appropriate for comparing pairs of time series models that are nested under the null hypothesis of equal forecast accuracy. See Clark and McCracken (2001, 2007) and Corradi and Swanson (2006) for further discussion of these

issues.

4. Results

We report the RMSPE for Greenbook and the other 11 models in Table 2. Note that the first column gives the level of the RMSPE for Greenbook; all other columns report the RMSPE for an alternative model relative to Greenbook—values less than one mean the alternative forecast is more accurate than Greenbook. Jumping-off points and forecast horizons are indexed relative to the quarter in which the forecast is made, which is labeled time zero. We consider jumping-off points from minus 1, one quarter before the forecast quarter, through 3. If the jumping-off point is, e.g., 2, we have brought all data up to two quarters after the forecast quarter using the Fed forecast, then the time series forecasts start at horizon 3.

The period from 1979-1983 was especially volatile in the U.S. economy, containing the sharpest and deepest recessions in the post-war era, and the period since 1982 has seen the Great Moderation with the economy quieter than in the pre-1979 period. Tulip (2005) has documented that the Greenbook forecast errors for output were largest early in the sample period. Because of the particular turbulence of the early 1980s, our baseline results are for the period since 1984.

4.1 Greenbook versus atheoretical models large and small

The results for inflation and output growth are strikingly different. For inflation, Greenbook dominates all the time series methods at nearly all forecasting horizons and jumping-off points. The Greenbook RMSPE is typically 10 to 40 percent smaller than those of the other models and in some cases we can reject the hypothesis of equal forecast accuracy of Greenbook and the other methods. The main exception to this general result is that as we move the jumping-off point out in time, the Bayesian model averaging and integrated factor augmented VAR are about as accurate as Greenbook.

For output growth, Greenbook is a good deal better than the atheoretical methods for jumping-off point -1 at horizon zero. This is consistent with the view that the Fed usefully exploits a great deal more information about the current state of the economy than is used in the time series models. After horizon zero, however, the advantage largely evaporates. One can see this in two ways. First, if we keep the jumping-off point at $t - 1$, the relative RMSPEs are close to one as we consider horizons beyond zero. Thus, the Greenbook advantage at time zero does not translate into substantial forecasting gains at other horizons. Second, if we move the jumping-off point out even one quarter, the Greenbook advantage at all remaining horizons disappears and is perhaps even reversed. That is, for several time series methods the point estimate of the relative RMSPE is less than one and we cannot reject the hypothesis that the method is at least as good as Greenbook.

4.2 Comparison of atheoretical methods

We are also interested in evaluating the relative merits of large dataset methods versus univariate methods. For inflation, among the univariate forecasts, the RAR and SV forecasts generally seem to do best (Table 2). However, BMA does a good bit better than any of the univariate inflation forecasts, and generally gives the smallest RMSPE among all the atheoretical inflation forecasts considered. The FAV and DF forecasts are generally somewhat *less* accurate predictors of inflation than the best univariate methods.

For growth, both univariate methods and both model averaging methods produce essentially equivalent results. Some of the factor models perform much worse, especially for the early jumping-off points. Among the factor models, only the FVS performs similarly to the univariate and averaging methods.

4.3 Robustness

Our goal in this paper is not to select a best method, but to see if any generalities arise in comparing the four classes of model. The introduction and sections 4.1 and 4.2 lay

out the generalities we seek to emphasize. We explored a wide range of variations on our exercise to see if these general patterns are robust. Here we review these. Full results for all combinations of the choices described here are in the web appendix.

To document that the Table 2 results are not greatly sensitive to outliers, we show a less sensitive metric—the percentage of forecast periods in which the time series model is more accurate than Greenbook (Table 3). Numbers greater than 50 percent favor the atheoretical forecast and are in bold. The picture that emerges is much like that from Table 2: for inflation, Greenbook generally does better than the time series forecasts a good bit more than half the time. For growth, Greenbook does better than the time series methods well over than half the time at horizon zero, but at all other horizons Greenbook and time series models seem about equally accurate.

Several of the forecasting models have model selection parameters—lag lengths and numbers of factors—that are fixed in the baseline results. We also examined selecting lag lengths and numbers of factors using the Bayes information criterion (BIC). The forecast accuracy was generally a bit worse under BIC than under our choice of fixed parameters, but the general pattern of results remained.

The BMA approach also has a hyperparameter in the prior, and we explored a range of values. The key parameter, ϕ above, determines how far the weights are likely to vary from equal. Moving ϕ either direction from our baseline value of 2, taken from other work, tends to worsen performance. Thus, for example, BMAs superiority to EWA on forecasting inflation is a bit sensitive to the choice of hyperparameter.

As noted above, because of the turbulence of the early 1980s, our baseline results are for the period since 1984. Results for the full sample period show larger RMSPEs for all the forecast methods but the patterns we emphasize remain.

Our measure of inflation is from the GDP deflator. We also considered using CPI inflation, which is essentially unrevised, but for seasonal factors. The full analysis with the CPI data also shows the patterns emphasized here. The main difference for the CPI

is that the GB accuracy on the current quarter is much higher. This appears to owe to the fact that the CPI is a monthly series and hard data is available for part of the quarter when many Greenbook forecasts are made.

Finally, as noted above, there is no unambiguous concept of ‘actual’ data for series that are continuously revised. We computed full results for two alternative ‘actual’ concepts and verified that the patterns emphasized here remain unchanged.

4.4 Within-quarter information

In Table 2, we can see that the Greenbook does quite well in measuring the current state of the economy, that is in forecasting quarter t from the $t - 1$ jumping-off point. This presumably owes in part to the special efforts of the Fed to exploit incoming weekly and monthly macroeconomic data releases. When the Greenbook forecast is made after the middle of the second month of the quarter, the staff will know and take account of the values of the key monthly indicators from the first month of the quarter. The time series models that we consider do not have this information.

It is not our intention to test whether some mixed-frequency, atheoretical method could perform comparably to Greenbook in assessing the current state. A number of recent papers including Giannone, Reichlin and Small (2007) and Aruoba, Diebold and Scotti (2007) have proposed powerful methods for measuring the current state of the economy from the realtime monitoring of incoming data releases. We, however, are mainly interested in how the atheoretical models perform in forecasting the future when given a good estimate of the current state, from Greenbook or any other source.

In any case, it is of some interest to assess the importance of within quarter information in our results. To do so, we calculated the correlation between (a) the current-quarter inflation forecast errors for the final vintage in each quarter using each of our methods and (b) a measure of economic *news* about the current quarter that has already arrived at the time of the forecast. As our news measure we use the unexpected component of

the CPI or nonfarm payrolls announcement for the first month in the quarter, with the expectation measured by the Money Market Services (MMS) survey. In almost all cases, the nonfarm payrolls and CPI announcements for the first month of the quarter came out before the last Greenbook in that quarter, and so this news was in the Fed's information set. The correlation between the CPI surprise and the inflation forecast error averages about 0.17 across the time series models, but is about 0.04 and statistically insignificant for Greenbook. For output growth, using the payrolls surprise, the comparable numbers are 0.27 and 0.12. These results are consistent with the view that Greenbook effectively incorporates within-quarter information that is missing from our time series forecasts.

4.5 Comments

One striking observation from Table 2 is how small the Greenbook prediction errors for inflation are relative to any of the other forecasting procedures. This is consistent with Ang, Bekaert and Wei (2007) who find that private sector surveys outperform both univariate and multivariate time series forecasts of inflation. Our results sharpen those of Ang, Bekaert and Wei in showing that the advantage of the subjective methods over atheoretical methods in forecasting inflation remains even when the atheoretical methods take advantage of the subjective assessment up through 3 quarters in the future. Whether this advantage of Greenbook stems from access to a greater range of information or a more sophisticated use of the available information, or both, remains an open question.

It is worth noting that there appear to be quite strong seasonal patterns in many vintages of deflator inflation data, despite the fact that these data are seasonally adjusted. We experimented with adding deterministic seasonal dummies to each of the forecasting methods, but found that this nearly always worsened forecasting performance.

The decomposition of the inflation RMSPEs into bias and variance components (Table 5 in web appendix) sheds some light on the performance of the of the inflation models. The models generally have upward bias. The bias is worse for the stationary

univariate and factor model time series forecasts than it is for the Greenbook forecasts, the random walk, IFV and SV forecasts, and the predictions based on Bayesian model averaging. The relative lack of bias in models that impose a unit root in inflation might owe to the robustifying role of taking first differences in the presence of structural breaks (Clements and Hendry (1999)).

It would be natural to compare the Greenbook forecasts in this dataset with those from private sector forecasts, or to include such forecasts in the inputs to the time series models. Romer and Romer (2000) provided some evidence on this. Reifschneider and Tulip (2007) compare Greenbook and other government and private sector forecasts, finding that the historical errors from these alternative forecast errors are large, but of comparable magnitude. However, it would be difficult to do these comparisons with the same precision in the timing of information as we have in this paper.

5. Additional topics

5.1 *Comparison with Atkeson and Ohanian*

In contrast to our results, Atkeson and Ohanian (2001) compared the RMSPE of Greenbook and random walk forecasts of inflation, and found that they were roughly equal, concluding that there was no incremental information in the Greenbook. Their conclusion was based on 13 observations taken from the last Greenbook forecast in each year from 1983 to 1995 inclusive. They compared the Greenbook projection of GNP/GDP deflator inflation over the subsequent four quarters with a random walk forecast computed as the value of inflation over the previous four quarters. Thus, their measure of inflation and their version of the random walk forecast are different from ours.

We came very close to replicating their results by using their definitions and limiting our focus to the 13 observations they studied. We find that the RMSPE of the Greenbook forecast, relative to that of the random walk, was 0.96. Using their definitions but our full sample, the ratio is more favorable to Greenbook at 0.73.

We examine the sampling properties of the estimate based on 13 observations using the same bootstrap procedure used above. Figure 1 shows the bootstrap approximation to the sampling distribution of the ratio of RMSPEs. While this is centered around 1, in line with the Atkeson and Ohanian result, the ratio is quite imprecisely estimated. A 95 percent confidence interval for this ratio would span from 0.67 to 1.28, which includes the estimates for this ratio that we obtain with our much larger sample size. Thus, the good performance of the random walk 13 observations is entirely consistent with our finding that Greenbook dominates the random walk forecast. The issue is the low power of out-of-sample forecast accuracy tests, as discussed by Kilian and Inoue (2004) and Clark and McCracken (2006).

5.2 Comparison to work using ignoring realtime issues

All of the results presented so far use our realtime dataset. For comparison purposes with earlier work that uses a single vintage of revised data, we follow Bernanke and Boivin (2003) in repeating our exercise using a single vintage of data as observed in the Greenbook dated December 14, 2000. At the time this work was done, this was the last vintage that was outside of the five-year window during which the data are confidential.

For the time series models, we take the data from the December 2000 Greenbook vintage. We estimate and forecast the time series models using data from 1960Q2 up through some quarter T ; increment T and repeat; T goes from 1984Q1 to 2000Q3. Of course, the Greenbook forecast remains the realtime forecast as used throughout the paper. Given that the time series models are using data concepts as they stood in December 2000, we must define the ‘actuals’ differently from in the vintage work. For actuals, we use the data as observed in the December 2000 Greenbook.

The RMSPEs for the time series models obtained using *ex-post* revised data (Table 4; upper panels) are uniformly smaller than their counterparts using realtime data (Table 2). Thus, ignoring realtime issues gives an optimistic view of the precision that might

be obtained in real time. Notice, however, that the relative performance of the large-dataset methods, Greenbook and univariate forecasts is very similar to what we found using realtime data. For example, the model averaging methods perform among the best; the factor models sometimes perform quite badly.

This raises a bit of a puzzle. Both in the work of Bernanke and Boivin and in our work, factor models sometimes fare quite poorly. Stock and Watson (2002, 2005) however find that the factor methods perform quite well in forecasting inflation. From Table 4, we see that the poor performance of factor models forecasting inflation in our earlier results is not primarily due to the use of realtime data. There are two additional factors that may explain the difference. Stock and Watson use many more predictors—215 in the 2002 work; 135 in the 2005 work—than we can use given the limits of our vintage data; further the sample periods studied differ.

To shed some further light on this issue, we re-did our analysis using the single vintage in the Stock and Watson dataset of 135 predictors (lower panels of Table 4). Using the 1984–2000 sample used throughout the paper, the factor models continue to perform quite poorly in some cases. When we repeat the Table 4 comparison using our full sample from 1980–2000, we find that the factor models continue to perform poorly using our dataset of less than 100 variables, but perform much better using the Stock-Watson 135 predictors than our smaller dataset (see the web appendix). Thus, the combination of including the pre-1984 period and using many more predictors greatly improves the performance of the factor models. Of course, the 1980-1984 period saw dramatic disinflation. It may be that something in the very large dataset of Stock and Watson is allowing the factor models to better track this disinflation. This does not appear to carry over to the relatively stable period since 1984. It suggests that factor models may be more sensitive than model averaging methods to the sample period, and the specific number and mix of variables that are included in the dataset. Perhaps this apparent lack of robustness is because factor model methods pick out the factors that

explain most of the multi-variate variation in the entire dataset, without reference to the target forecast variable. Information that explains only a small share of the variation of the entire dataset but that is useful forecasting the target variable will be missed by the factor model methods.

5.3 *The Greenbook as a conditional forecast*

The Greenbook projection is conditioned on a hypothetical path of policy that is counterfactual in the sense that it is not supposed to be a forecast of policy (see Faust and Wright (2008)). Nearly all work on assessing the information content of central bank forecasts ignores their conditional nature and assesses them as though they were unconditional forecasts. We have done so here, thereby implicitly assessing their properties when viewed as though they were unconditional forecasts. However it should be noted that Greenbook could appear better—or worse—if we were to take proper account of its conditional nature. Faust and Wright propose a method for backing out an implied unconditional forecast based on comparing the hypothetical path of policy with the path implied by money market futures rates, under the assumption that the latter is the unconditional expectation of policy. In future work, we intend to include this implied unconditional counterpart of the Greenbook in the forecast evaluation exercise.

6. Conclusions

We compare the forecast accuracy of four classes of forecasting models using carefully synchronized realtime or vintage data: Greenbook, univariate time series models that are atheoretical from an economic perspective, and two families of large, atheoretical time series models. We focus on predicting inflation and output growth. The goal is to see how modern time series models perform versus Greenbook, which we take as a proxy for best current practice.

We sharpen earlier results in this area by paying careful attention to realtime data issues using set of archived datasets recently uncovered at the Federal Reserve. Further,

we add a new element in assessing whether the well-documented advantage of Greenbook relative to time series models is due mainly to the Fed’s ability to assess the current and recent past state of the economy—that is, due to nowcasting and backcasting. We give the time series models the benefit of the Fed’s assessment by extending the realtime data using the Fed’s forecast. The time series forecasts ‘jump off’ from the Fed’s assessment. Effectively, we are asking whether there is some point in the forecast horizon at which one could switch over from the Greenbook forecast to a time series model without loss, or perhaps with some gain.

We are not so much interested in picking a winner among the many methods; rather, we are looking for broad patterns in predictability that might guide future research both in macroeconomics and in time series. We find several such results. First, we again find that Greenbook is a very good forecast for both output growth and inflation.

Second, the predictability of inflation and output growth are very different. Elements of this result have been noted before but we considerably strengthen the result. Given a good estimate of the current state of output growth, neither of the multivariate atheoretical approaches nor the sophisticated Greenbook improve on a simple univariate forecast. For inflation in contrast, large time series models improve on univariate models, and Greenbook improves on those. Even after giving the time series models the Fed’s forecast for three quarters into the future, Greenbook still outperforms all the time series methods for forecasting further into the future. We believe that this dramatic difference in predictability is an important stylized fact for both macroeconomists and macropolicymakers. It would be interesting to see if this difference in forecastability in output and inflation can be obtained in standard DSGE models. Perhaps some features of the low frequency properties of inflation (structural breaks or near-I(1) behavior) makes it much harder to forecast with time series methods than with a judgmental approach.

Third, among large model methods, averaging of the forecasts of many simple, bivariate models is often the best method in our results; it is never very far from the best.

The factor model approaches seldom outperform model averaging and are considerably less consistent: in some cases the factor methods perform very badly.

Finally, while we find some general patterns, our results are for one relatively brief (about 20 year) period in a single economy. Patterns seen in one case should, we emphasize, be interpreted with caution.

Acknowledgments

We thank Douglas Battenberg, Jean Boivin, Bryan Campbell, Frank Diebold, Mike McCracken, Athanasios Orphanides, Lucrezia Reichlin, Dave Reifschneider, Chris Sims, Jim Stock, Norm Swanson, Mark Watson and three anonymous referees for helpful comments. All remaining errors are our own. The views expressed in this paper are solely the responsibility of the authors and should not be interpreted as reflecting the views of the Board of Governors of the Federal Reserve System or of any other employee of the Federal Reserve System.

Appendix: Data series summary

The following variables are included in our vintage dataset. Definitions of some variables and availability in the vintage databases varies through time. Additionally, slight variations on the variables listed here are sometimes present. For a complete listing, see the web appendix. The data are available at <http://e105.org/faustj/papgbts.php>. All variables are transformed into change in natural logarithm form unless otherwise noted in parenthesis: LN is natural logarithm, FD is first difference, and level means no transformation.

Income and spending: GNP or GDP real and nominal (DLN); Personal consumption spending real and nominal ; Consumer spending on durable goods real and nominal ; Final sales real and nominal ; Personal saving; Disposable Income real and nominal ; Consumption energy sector; Output per hour; Consumer sentiment (level);

Government spending real and nominal ; Corporate profits; Gross private domestic investment; Spending on business fixed investment real and nominal ; Spending on producers' durable equipment real and nominal ; Spending on producers' structures real and nominal ; Inventory investment real and nominal (FD); Inventory investment manufacturing and trade, real (FD); Residential construction spending real and nominal ;

Stocks: Consumer durables, real; Consumer durables ex autos, real; Autos; Motor vehicles and parts, real; Nonfarm inventories, real; Nonfarm inventories ex manufacturing and trade; Nonfarm nondurable inventories, real; Nonfarm nonretail inventories, real; Nonfarm retail durable inventories, real; Nonfarm nondurable inventories, real; Housing; Net stock, producers' durable equipment, real; Net stock, producers' structures, real.

Employment: Civilian employment; Nonfarm business employment; Employment of nonfarm proprietors; Civilian labor force; Labor force participation rate (level); Civilian unemployment rate (level); Hours, employees nonfarm business sector; Hours, household and institutions sector; Nonfarm business sector workweek (LN).

Prices and wages: CPI; CPI ex food and energy; GNP or GDP deflator; Consumption deflator; Core PCE price index; Producer Price Index; Durable goods consumption deflator; Commodity price industrial materials; Average price per barrel of imported oil; Wholesale price index, fuels; Import price, petroleum products; Price deflator, crude energy consumption; Wholesale price, petroleum products; Employee compensation per hour; Compensation of employees.

Returns: Federal funds rate (level); 10-year Treasury yield (level); 3-month Treasury yield (level); Moody's seasoned AAA yield (level); Moody's BAA corporate yield (level); Mortgage rate (level); Dividend-price ratio (Standard and Poors) (LN).

Banking: M1; M2; Nonborrowed reserves; Currency plus travelers checks; Total reserves; Commercial and industrial loans at banks.

Foreign Sector: Exports real and nominal ; Imports real and nominal ; Net exports (FD); Current account balance (FD); Merchandise trade balance (FD); Merchandise ex-

ports; Foreign exchange rate index, bilateral weights; Foreign exchange rate index, multilateral weights; G10 real exchange rate; G18 exchange rate, real and nominal ; Foreign GNP index (bilateral weights) real and nominal ; Foreign short-term interest rate (level); Foreign CPI.

References

- Ang, A., G. Bekaert and M. Wei (2007): Do Macro Variables, Asset Markets or Surveys Forecast Inflation Better?, *Journal of Monetary Economics*, 54, pp.1163-1212.
- Armah, N.A. and N.R. Swanson (2007): Seeing Inside the Black Box: Using Diffusion Index Methodology to Construct Factor Proxies in Largescale Macroeconomic Time Series Environments, working paper.
- Aruoba, S.B., F.X. Diebold and C. Scotti (2007): Real-Time Measurement of Business Conditions, working paper.
- Atkeson, A. and L.E. Ohanian (2001): Are Phillips Curves Useful for Forecasting Inflation?, *Federal Reserve Bank of Minneapolis Quarterly Review*, 25, pp.2-11.
- Bai, J. and S. Ng (2006): Evaluating Latent and Observed Factors in Macroeconomics and Finance, *Journal of Econometrics*, 113, pp.507-537.
- Bates, J.M. and C.W.J. Granger (1969): The Combination of Forecasts, *Operations Research Quarterly*, 20, pp.451-468.
- Bernanke, B.S. and J. Boivin (2003): Monetary Policy in a Data-Rich Environment, *Journal of Monetary Economics*, 50, pp.525-546.
- Bernanke, B.S., J. Boivin and P. Eliasz (2005): Measuring Monetary Policy: A Factor Augmented Vector Autoregressive (FAVAR) Approach, *Quarterly Journal of Economics*, 120, pp.387-422.

- Boivin, J. and S. Ng (2006): Are More Data Always Better for Factor Analysis?, *Journal of Econometrics*, 132, pp.169-194.
- Clark, T.E. and M. McCracken (2001): Tests of Equal Forecast Accuracy and Encompassing for Nested Models, *Journal of Econometrics*, 105, pp.85-110.
- Clark, T.E. and M. McCracken (2006): The Predictive Content of the Output Gap for Inflation: Resolving In-Sample and Out-of-Sample Evidence, *Journal of Money, Credit and Banking*, 38, pp.1127-1148.
- Clark, T.E. and M. McCracken (2007): Tests of Equal Predictive Ability with Real-Time Data, working paper.
- Clements, M. and D.F. Hendry (1999): Forecasting Non-stationary Economic Time Series, MIT Press, Cambridge.
- Corradi, V. and N.R. Swanson (2006): Predictive Density Evaluation, in “Handbook of Economic Forecasting,” C.W..J. Granger, G. Elliott and A. Timmerman (eds.), Elsevier, Amsterdam.
- Corradi, V. and N.R. Swanson (2007): Nonparametric Bootstrap Procedures for Predictive Inference Based on Recursive Estimation Schemes, *International Economic Review*, 48, pp.67-109.
- D’Agostino, A. and D. Giannone (2006): Comparing Alternative Predictors Based on Large Panel Factor Models, European Central Bank Working Paper 680.
- Diebold, F.X. and R.S. Mariano (1995): Comparing Predictive Accuracy, *Journal of Business and Economic Statistics*, 13, pp.253-263.
- Faust, J. and J.H. Wright (2008): Efficient Forecast Tests for Conditional Policy Forecasts, *Journal of Econometrics*, forthcoming.

- Fernandez, C., E. Ley and M.F.J. Steel (2001): Model Uncertainty in Cross-Country Growth Regressions, *Journal of Applied Econometrics*, 16, pp.563-576.
- Forni, M., M. Hallin, M. Lippi and L. Reichlin (2005): The Generalized Dynamic Factor Model: One-Sided Estimation and Forecasting, *Journal of the American Statistical Association*, 100, pp.830-840.
- Giannone, D., L. Reichlin and L. Sala (2004): Monetary Policy in Real Time, in M. Gertler and K. Rogoff (eds.), NBER Macroeconomics Annual, 2004, MIT Press, Cambridge.
- Giannone, D., L. Reichlin and D. Small (2007): Nowcasting GDP: The Real Time Informational Content of Macroeconomic Data Releases, working paper.
- Inoue, A. and L. Kilian (2004): In-Sample or Out-of-Sample Tests of Predictability: Which One Should We Use, *Econometric Reviews*, 23, pp.371-402.
- Koop, G. and S. Potter (2003): Forecasting in Large Macroeconomic Panels Using Bayesian Model Averaging, Federal Reserve Bank of New York Staff Report 163.
- Künsch (1989): The Jackknife and the Bootstrap for General Stationary Observations, *Annals of Statistics*, 17, pp.1217-1241.
- Liu, R. and K. Singh (1992): Moving Blocks Jackknife and Bootstrap Capture Weak Dependence, in *Exploring the Limits of the Bootstrap* (R. Lepage and L. Billard, eds.), Wiley, New York.
- Marcellino, M., J.H. Stock and M.W. Watson (2006): A Comparison of Direct and Iterated Multistep AR methods for Forecasting Macroeconomic Time Series, *Journal of Econometrics*, 135, pp.499-526.
- Marcellino, M. (2006): A Benchmark for Models of Growth and Inflation, working paper.

- Reifschneider, D.L., D.J. Stockton and D.W. Wilcox (1997): Econometric Models and the Monetary Policy Process, *Carnegie Rochester Series on Public Policy*, 47, pp.1-37.
- Reifschneider, D.L. and P. Tulip (2007): Gauging the Uncertainty of the Economic Outlook from Historical Forecasting Errors, Finance and Economics Discussion Series 2007-60, Federal Reserve Board.
- Romer, C.D. and D.H. Romer (2000): Federal Reserve Information and the Behavior of Interest Rates, *American Economic Review*, 90, pp.429-457.
- Sims, C.A. (2002): The Role of Models and Probabilities in the Monetary Policy Process, *Brookings Papers on Economic Activity*, 2, pp.1-40.
- Stock, J.H. and M.W. Watson (1999): Forecasting Inflation, *Journal of Monetary Economics*, 44, pp.293-335.
- Stock, J.H. and M.W. Watson (2002): Forecasting Using Principal Components from a Large Number of Predictors, *Journal of the American Statistical Association*, 97, pp.1167-1179.
- Stock, J.H. and M.W. Watson (2003): Forecasting Output and Inflation: The Role of Asset Prices, *Journal of Economic Literature*, 41, pp.788-829.
- Stock, J.H. and M.W. Watson (2005): An Empirical Comparison of Methods for Forecasting Using Many Predictors, working paper.
- Stock, J.H. and M.W. Watson (2007): Has Inflation Become Harder to Forecast?, *Journal of Money, Credit and Banking*, 39, pp.3-34.
- Tulip, P. (2005b): Has Output Become More Predictable? Changes in Greenbook Forecast Accuracy, Finance and Economics Discussion Series 2005-31, Federal Reserve

Board.

Wright, J.H. (2003): Forecasting U.S. Inflation by Bayesian Model Averaging, International Finance Discussion Paper 780, Federal Reserve Board.

Table 1: Model summary

Code	Description	Citation
GB	Greenbook	Reifschneider, et al. (1997)
RW	Random walk model	
RAR	Recursive autoregression	
DAR	Direct forecast from autoregression	Marcellino, et al. (2006)
SV	Unobserved component, stochastic volatility	Stock and Watson (2007)
EWA	Equal weighted averaging	Stock and Watson (2003)
BMA	Bayesian model averaging	Wright (2003)
FAA	Factor augmented autoregression	
FAV	Factor augmented vector autoregression	Bernanke, et al. (2005)
IFV	Integrated factor augmented VAR	
DF	Dynamic factor	Forni, et al. (2005).
FVS	Factor-spanned variable selection	Bai and Ng (2006)

Note: The citation is to the paper we took our version of the model from, which is not necessarily the seminal paper on the generic model type. See the other citations in the text.

Table 2a. Greenbook RMSPE and relative RMSPE of time series models: deflator inflation, 1984–2000

hor	GB	RW	RAR	DAR	SV	EWA	BMA	FAA	FAV	IFV	DF	FVS
jump off -1												
0	0.69	1.63•	1.39•	1.39•	1.42•	1.37•	1.34•	1.45•	1.58•	1.56•	1.62•	1.41•
1	0.79	1.66•	1.42•	1.42•	1.38•	1.37•	1.22•	1.32•	1.42•	1.26•	1.86•	1.46•
2	0.81	1.51•	1.31•	1.30•	1.25•	1.25•	1.15•	1.19•	1.31•	1.17•	1.92•	1.31•
3	0.93	1.16	1.24•	1.24•	1.12	1.20•	1.03	1.24•	1.34•	1.08	1.84•	1.23•
4	0.89	1.30•	1.41•	1.43•	1.20	1.36•	1.08	1.28•	1.49•	1.22•	2.11•	1.42•
5	1.14	1.28	1.32•	1.35•	1.12	1.29•	0.99	1.22•	1.31•	1.06	1.83•	1.35•
jump off 0												
1	0.79	1.20•	1.28•	1.28•	1.22•	1.26•	1.22•	1.32•	1.37•	1.26•	1.58•	1.28•
2	0.81	1.31•	1.27•	1.27•	1.20•	1.23•	1.18	1.19•	1.32•	1.20•	1.69•	1.27•
3	0.93	1.22•	1.22•	1.21•	1.12	1.17•	1.07	1.21•	1.26•	1.04	1.66•	1.22•
4	0.89	1.08	1.30•	1.32•	1.09	1.26•	1.05	1.25•	1.37•	1.09	1.91•	1.30•
5	1.14	1.07	1.26•	1.28•	1.05	1.22•	0.97	1.12	1.26	1.02	1.77•	1.28•
jump off 1												
2	0.81	1.14•	1.18•	1.18•	1.14•	1.16•	1.16•	1.16•	1.21•	1.16	1.43•	1.19•
3	0.93	1.23•	1.20•	1.19•	1.14	1.17•	1.07	1.15•	1.17	1.01	1.50•	1.23•
4	0.89	1.26•	1.28•	1.30•	1.15	1.25•	1.13	1.23•	1.26•	1.07	1.71•	1.32•
5	1.14	1.02	1.18	1.19	1.02	1.15	0.96	1.13•	1.19	0.98	1.60•	1.22
jump off 2												
3	0.93	1.05	1.14•	1.14•	1.07	1.12•	1.09	1.14•	1.11	0.99	1.28•	1.14•
4	0.89	1.19•	1.25•	1.27•	1.14•	1.23•	1.16	1.21•	1.21•	1.03	1.48•	1.28•
5	1.14	1.06	1.18	1.19	1.03	1.16	1.01	1.13•	1.15•	0.98	1.39•	1.20•
jump off 3												
4	0.89	1.07	1.17•	1.17•	1.09	1.16•	1.11	1.15•	1.11•	0.97	1.29•	1.19•
5	1.14	1.07	1.14•	1.14•	1.04	1.12	1.01	1.11•	1.10•	0.95	1.24•	1.16•

Note: See notes at end of Table 2

Table 2b. Greenbook RMSPE and relative RMSPE of time series models: output growth, 1984–2000

hor	GB	RAR	DAR	EWA	BMA	FAA	FAV	DF	FVS
jump off -1									
0	1.75	1.09	1.09	1.09	1.10	1.30	1.39	1.32	1.24
1	2.12	0.87	0.86	0.87	0.93	1.13	1.20	1.15	0.93
2	2.01	0.98	0.97	0.95	0.98	1.10	1.15	1.07	1.00
3	2.15	0.97	0.95	0.94	1.02	1.10	1.12	1.02	1.02
4	2.08	0.99	1.00	0.99	1.05	1.08	1.11	1.06	1.02
5	2.08	1.02	0.98	0.97	1.00	1.00	1.09	1.03	0.98
jump off 0									
1	2.12	0.84	0.84	0.84	0.85	0.96	1.07	0.99	0.85
2	2.01	0.94	0.93	0.91	0.90	1.04	1.12	1.09	0.95
3	2.15	0.96	0.95	0.95	0.99	1.10	1.18	1.04	0.96
4	2.08	0.98	0.97	0.96	1.02	1.05	1.09	1.07	0.96
5	2.08	1.02	0.99	0.98	1.04	1.03	1.11	1.09	0.98
jump off 1									
2	2.01	0.94	0.94	0.91	0.90	0.94	1.05	0.99	0.97
3	2.15	0.97	0.96	0.96	0.95	1.07	1.13	1.05	0.97
4	2.08	0.98	0.98	0.97	0.99	1.09	1.15	1.04	0.99
5	2.08	1.02	1.01	1.01	1.07	1.06	1.10	1.13	1.02
jump off 2									
3	2.15	0.96	0.96	0.95	0.92	0.97	1.01	0.98	0.96
4	2.08	0.97	0.98	0.97	0.94	1.05	1.08	1.02	0.97
5	2.08	1.02	1.02	1.01	1.03	1.11	1.11	1.09	1.03
jump off 3									
4	2.08	0.98	0.98	0.95	0.93	0.95	1.01	0.97	0.94
5	2.08	1.00	1.01	1.00	0.96	1.06	1.07	1.06	1.00

Notes: The second column gives the RMSPE for Greenbook. Each subsequent column reports the ratio of the RMSPE for an alternative forecast to that of Greenbook; values less than one mean the alternative model has smaller RMSPE and are in bold. Data for the alternative forecasts are brought up to the indicated jumping off point using the Fed forecast. The symbols •, •, • indicate that the associated DM statistic is significantly different from one at the 1, 5, or 10 percent level, respectively, based on the bootstrap described in the text.

Table 3a. Percentage of periods alternative forecast better than GB: deflator inflation, 1984–2000

hor	RW	RAR	DAR	SV	EWA	BMA	FAA	FAV	IFV	DF	FVS
jump off -1											
0	37	32	32	34	29	32	30	25	31	34	31
1	25	27	31	34	33	31	33	32	40	21	30
2	38	36	34	35	36	35	43	34	43	25	34
3	42	39	34	37	35	36	39	39	44	17	34
4	30	25	25	43	25	37	30	32	39	11	26
5	35	23	21	42	22	46	34	27	34	16	25
jump off 0											
1	32	32	32	34	32	30	32	35	41	30	30
2	34	34	34	39	33	33	39	34	40	21	31
3	37	37	33	41	34	34	32	37	43	20	34
4	33	31	33	41	30	35	40	35	43	18	32
5	34	28	27	48	24	44	35	27	34	15	26
jump off 1											
2	38	34	34	41	36	36	39	36	42	24	36
3	33	31	30	39	33	34	34	31	36	20	31
4	33	30	31	39	30	32	33	34	41	17	31
5	34	34	31	36	30	43	34	27	41	21	32
jump off 2											
3	34	34	34	40	36	34	43	42	48	29	34
4	29	31	31	31	32	32	33	36	37	25	32
5	36	30	30	41	28	34	32	28	44	22	29
jump off 3											
4	33	33	33	36	34	34	35	40	46	30	35
5	34	31	30	37	29	37	30	36	44	28	30

Note: See notes at end of Table 3

Table 3b. Percentage of periods alternative forecast better than GB: output growth, 1984–2000

hor	RAR	DAR	EWA	BMA	FAA	FAV	DF	FVS
jump off -1								
0	39	39	40	43	36	36	36	30
1	57	60	52	48	41	39	45	52
2	56	55	58	57	45	37	52	55
3	56	57	54	49	45	42	52	48
4	54	45	46	45	36	36	46	44
5	45	48	52	43	48	43	46	49
jump off 0								
1	59	59	57	54	45	45	40	54
2	54	55	55	54	47	41	44	53
3	55	57	55	52	39	38	47	54
4	53	56	54	48	41	39	43	57
5	44	43	49	43	43	43	48	44
jump off 1								
2	56	56	59	60	50	47	52	51
3	55	57	54	50	41	41	40	53
4	53	52	52	50	43	39	43	54
5	46	47	48	43	43	42	44	48
jump off 2								
3	52	52	49	58	43	46	49	56
4	52	53	51	52	43	43	46	52
5	45	50	45	43	39	37	38	50
jump off 3								
4	54	54	57	57	45	48	48	57
5	48	51	46	45	39	42	45	47

Notes: Each column after the first reports the percentage of forecast periods in which the alternative forecast error is smaller in absolute value than the GB error. Entries greater than 50 percent indicate that the alternative forecast is better more than half the time and are in bold. See also notes to Table 2.

Table 4a. RMSPE of Greenbook and relative RMSPE of time series forecasts using single data vintage: deflator inflation, 1984–2000

hor	GB	RW	RAR	DAR	SV	EWA	BMA	FAA	FAV	IFV	DF	FVS
our final vintage												
0	0.59	1.53	1.46	1.46	1.47	1.45	1.38	1.68	1.88	1.77	1.78	1.44
1	0.69	1.40	1.27	1.27	1.19	1.22	1.16	1.19	1.35	1.28	1.80	1.23
2	0.75	1.20	1.21	1.23	1.08	1.17	1.04	1.16	1.22	1.13	2.03	1.22
3	0.82	1.11	1.22	1.28	1.06	1.17	0.93	1.08	1.17	1.07	2.02	1.31
4	0.90	1.20	1.28	1.36	1.08	1.23	0.94	1.02	1.24	1.13	2.07	1.36
5	0.99	1.11	1.23	1.33	0.98	1.18	0.90	1.02	1.16	1.05	1.99	1.35
Stock-Watson data												
0	0.59	1.53	1.46	1.46	1.47	1.43	1.42	1.59	1.79	1.84	2.36	1.51
1	0.69	1.40	1.27	1.27	1.19	1.22	1.11	1.18	1.27	1.28	2.02	1.34
2	0.75	1.20	1.21	1.23	1.08	1.18	1.04	1.28	1.30	1.29	1.84	1.21
3	0.82	1.11	1.22	1.28	1.06	1.20	1.02	1.25	1.26	1.24	1.74	1.29
4	0.90	1.20	1.28	1.36	1.08	1.27	1.10	1.26	1.28	1.24	1.71	1.34
5	0.99	1.11	1.23	1.33	0.98	1.23	1.11	1.30	1.28	1.21	1.68	1.28

Note: See notes at end of Table 4.

Table 4b. RMSPE of Greenbook and relative RMSPE of time series forecasts using single data vintage: output growth, 1984–2000

hor	GB	RAR	DAR	EWA	BMA	FAA	FAV	DF	FVS
our final vintage									
0	1.84	1.62	1.62	1.53	1.49	1.48	1.57	1.55	1.75
1	2.33	0.84	0.85	0.84	0.88	1.03	1.09	0.93	0.85
2	2.10	0.97	0.99	1.00	1.02	1.21	1.25	1.18	1.00
3	2.26	0.91	0.93	0.94	0.97	1.11	1.13	1.05	1.00
4	2.22	0.94	0.96	0.98	1.00	1.13	1.17	1.10	0.96
5	2.25	0.93	0.94	0.95	0.94	1.04	1.14	1.02	0.98
Stock-Watson data									
0	1.84	1.62	1.62	1.55	1.50	1.41	1.68	1.48	1.67
1	2.33	0.84	0.85	0.85	0.94	1.04	1.17	0.98	0.92
2	2.10	0.97	0.99	0.99	1.04	1.04	1.34	1.09	1.04
3	2.26	0.91	0.93	0.93	0.95	1.00	1.12	0.95	0.97
4	2.22	0.94	0.96	0.97	0.97	1.06	1.14	1.03	1.07
5	2.25	0.93	0.94	0.95	0.94	0.96	1.04	1.05	1.01

Notes: The forecasts are pseudo-out-of-sample based on a single vintage of data. The single recent vintage is the final vintage from our data, Dec. 2000. Stock-Watson uses the Stock Watson (2003) data. The ‘final’ used in calculating forecast errors for this table is from the Dec. 2000 vintage. The GB values are RMSPEs; values for other models are relative to GB; < 1 means alternative model has smaller RMSPE and are in bold.

Fig. 1: Bootstrap pdf for the ratio of GB to RW RMSPEs in Atkeson–Ohanian sample

