Evaluating point and density forecasts of DSGE models

Maik Hendrik Wolters

Goethe University Frankfurt

23. January 2012

Evaluating point and density forecasts of DSGE models

Maik H. Wolters *

Goethe University Frankfurt.

January 23, 2012

Abstract

This paper investigates the accuracy of point and density forecasts of four DSGE models for inflation, output growth and the federal funds rate. Model parameters are estimated and forecasts are derived successively from historical U.S. data vintages synchronized with the Fed's Greenbook projections. Point forecasts of some models are of similar accuracy as the forecasts of nonstructural large dataset methods. Despite their common underlying New Keynesian modeling philosophy, forecasts of different DSGE models turn out to be quite distinct. Weighted forecasts are more precise than forecasts from individual models. The accuracy of a simple average of DSGE model forecasts is comparable to Greenbook projections for medium term horizons. Comparing density forecasts of DSGE models with the actual distribution of observations shows that the models overestimate uncertainty around point forecasts.

Keywords: DSGE models, forecasting, model uncertainty, forecast combination, density forecasts, real-time data, Greenbook

JEL-Codes: C53, E31, E32, E37

*Mailing address: Grüneburgplatz 1, 60323 Frankfurt, Germany; wolters@wiwi.uni-frankfurt.de. An earlier version of this paper circulated under the title "Forecasting under model uncertainty" (first version: July 2010). I am grateful for helpful comments by Tim Oliver Berg, Stefan Gerlach, Sebastian Schmidt, Tomi Wang, Mike Wickens, Volker Wieland as well as seminar participants at Goethe University Frankfurt, Stanford University, the Federal Reserve Bank of Philadelphia, the Federal Reserve Board, the 2010 DIW macroeconometric workshop in Berlin, the 2010 IWH macroeconomic workshop in Halle, the 2010 ZEW macroeconomics conference in Mannheim, the 2010 RGS conference in Bochum, the 2011 SNDE conference in Washington DC, the 2011 Midwest Macro Meeting in Nashville, the 2011 CEF conference in San Francisco, the 2011 Dynare conference in Atlanta and the 2011 annual meeting of the Verein für Socialpolitik in Frankfurt.

1 Introduction

For a long time business cycle models with microeconomic foundations have been calibrated and used for policy simulations while atheoretical time series methods have been used to forecast macroeconomic variables. Recently, several researchers have shown that estimated DSGE models can generate forecasts of reasonable accuracy in comparison to benchmark forecasts including nonstructural time series methods and central bank forecasts (Smets and Wouters, 2004; Adolfson et al., 2007a; Smets and Wouters, 2007; Wang, 2009; Edge et al., 2010; Christoffel et al., 2011).[1] The advantage of using structural models is that an economically meaningful interpretation of the forecasts can be given.[2] While these studies analyse only one model at a time, Wieland and Wolters (2011) compute forecasts from several theory based models. They focus exclusively on periods around business cycle turning points and find that DSGE models as wells as professional forecasters fail to predict recessions. DSGE model forecasts turn out to be quite heterogeneous despite their common underlying New Keynesian modeling philosophy. Although all major central banks have included one or several DSGE models into their forecasting toolkit[3] a thorough assessment of forecasts from different structural models including a comparison to forecasts from sophisticated nonstructural time series methods and to professional forecasts has not been undertaken yet for a long sample. Recent comparison studies of state of the art forecasting methods have been restricted to nonstructural econometric methods (c.f. Stock and Watson, 2002; Bernanke and Boivin, 2003; Forni et al., 2003; Marcellino et al., 2003; Faust and Wright, 2009; Hsiao and Wan, 2010).

In this paper, I carry out a detailed assessment of the forecasting accuracy of a suite of structural models. I use the same sample and real-time dataset as Faust and Wright (2009) who assess the forecasting accuracy of eleven nonstructural models. Therefore, my results are directly comparable to the forecasts from these models. The dataset includes data vintages for 145 FOMC meetings between March 1980 and December 2000 and is perfectly synchronized with the Greenbook. Hence, the results can also be compared to a best practice benchmark given by the Greenbook projections.

[1] Edge and Gürkaynak (2010) study the absolute rather than relative forecasting accuracy of the DSGE model by Smets and Wouters (2007). They find that the forecasts contain little information about actual future macroeconomic dynamics, although statistical and judgmental forecasts perform equally poorly.

[2] Wieland and Wolters (2012) show examples how to structurally interpret forecasts from DSGE models.

[3] Examples include the FRB/EDO-model at the Fed (Edge et al., 2007, 2008, 2010), the New Area-Wide Model and the CMR model at the ECB (Christoffel et al., 2008; Smets et al., 2010), the ToTEM model at the Bank of Canada (Murchison and Rennison, 2006) or the Ramses model at the Riksbank (Adolfson et al., 2007b).

The Greenbook projections are computed by the Federal Reserve's staff before each FOMC meeting and have been found to dominate forecasts from other professional forecasters in terms of forecasting accuracy (Romer and Romer, 2000; Sims, 2002; Bernanke and Boivin, 2003).

I consider models that cover to some extent the range of closed-economy DSGE models that have been used in academia and at policy institutions prior to the recent financial crisis. The first model is a purely forward looking small-scale New Keynesian model with sticky prices that is analysed in detail in Woodford (2003). The second model by Fuhrer (1997) has a mainly backward looking demand side, while the Phillips curve is derived from expectation-based overlapping wage contracts. The third model is a medium-scale New Keynesian model as developed in Christiano et al. (2005). I use the estimated version by Smets and Wouters (2007). The fourth model is a version of the DSGE model by Edge et al. (2007) that features two production sectors with different technology growth rates and is itself an extension of the Christiano, Eichenbaum & Evans model. To determine how much of the forecasting accuracy of these four models is due to the theoretical foundations and what can be attributed to the parsimonious parametrization of these stylized models, I also consider a Bayesian VAR. It is a datadriven nonstructural counterpart to the four DSGE models with a comparably strict parametrization.

The parameters of the models are reestimated on three to eleven time series—as proposed by the original authors—for historical data vintages. Given this estimate, I compute a nowcast and forecasts up to five quarters into the future that take into account information that was actually available at the forecast start. Forecast precision is assessed relative to revised data that became available during the subsequent quarters of the dates to which the forecasts apply.

Good forecasts are in general based on good forecasting methods and an accurate assessment of the current state of the economy. The Fed's great efforts to evaluate the current state of the economy are reflected in the accuracy of the Greenbook nowcasts. Sims (2002) suggests that this accurate data basis is a main reason for the precise Greenbook projections. The Fed's nowcasts exploit high frequency time series with more recent data than quarterly time series. In principle, there are methods available that allow the usage of such data in combination with structural macroeconomic models (see Giannone et al., 2009). Employing such methods is beyond the scope of this paper. To approximate the effect of including more information in nowcasting, I investigate the effect of using Greenbook

nowcasts as a starting point for model-based forecasts by appending them to the actually available data. Thus, the potential informational advantage of the Fed about the current state of the economy is eliminated and a proper comparison of model forecasts with Greenbook projections is possible.

Timmermann (2006) surveys model averaging methods and finds that weighted forecasts from several nonstructural models outperform forecasts from individual models. Combining several models provides a hedge against model uncertainty when it is not possible to identify a single model that consistently dominates the forecasting accuracy of other models. Therefore, in addition to individual model forecasts, I consider several simple and sophisticated model averaging schemes to compute weighted forecasts. For example, Gerard and Nimark (2008) and Bache et al. (2009) take into account forecasting uncertainty due to model uncertainty by combining forecasts from VARs and a single DSGE model. This paper is an extension of their approach to a suite of theory based business cycle models.

The evaluation results of the point forecasts confirm the reasonable forecasting accuracy of DSGE models found in the above mentioned studies. However, the forecasting quality of different DSGE models differs a lot from each other. For output growth, several DSGE models outperform the Greenbook projections and have an accuracy comparable to the best nonstructural models. Large scale DSGE models perform better than small scale models. However, quarterly output growth has little persistence and is thus difficult to forecast in general. Only one of the DSGE models yields more accurate forecasts than a simple univariate autoregressive process. All other DSGE models, the Greenbook projections and nonstructural forecasting methods—including large dataset methods—perform only as well as or even worse than a univariate autoregressive forecast. Greenbook inflation forecast is more accurate than all model forecasts. The best nonstructural large dataset methods dominate the DSGE model inflation forecasts, too. For interest rate projections, the structural models perform worse than a Bayesian VAR probably due to the very simple monetery policy rules imposed in the models. For short term horizons the hypothetical future interest rate path that the Greenbook projections are conditioned on, is more accurate than the DSGE model interest rate forecasts.[4] The forecasting quality of the structural models relative to the Greenbook increases with the forecast horizon for inflation and

[4]Faust and Wright (2009) analyse inflation and output growth forecasts, but not interest rate forecasts. Hence, I cannot compare DSGE model interest rate forecasts to their large set of nonstructural forecasting methods. The comparison is restricted to the Greenbook projections and a BVAR.

the interest rate.

Looking at the whole sample, the forecasts from the model by Smets and Wouters (2007) are in many cases more precise than forecasts from the other models. The model has a rich economic structure and is estimated on more variables than the standard New Keynesian models. However, deviding the evaluation period into subsamples, shows that no DSGE model continuously dominates the other models in terms of forecasting accuracy. The problem of instability in the performance and predictive content of different forecasting methods is well known and surveyed in detail in Rossi (2012). The survey in Timmermann (2006) discusses the usefulness of forecast combination as a hedge against such model instability. One main result is that in practice equal weighted forecast combinations are hard to beat in terms of accuracy. I find that weighted forecasts have a higher accuracy than forecasts from individual models. Combined forecasts based on simple weighting schemes that give significant weight to several models are superior to likelihood based weighting schemes that turn out trying to identify a single best model rather than giving weight to several models. A simple average of the forecasts of all four DSGE models is in many cases most accurate and otherwise only marginally less accurate than weighted forecasts from more sophisticated weighting methods. A simple average of the models' output growth forecasts outperforms the Greenbook projections and the most accurate nonstructural forecasts. While we don't know ex-ante which DSGE model yields the best forecast, using a simple average of all four DSGE models yields an output growth forecast that is as accurate as the best performing DSGE model forecast. Combined inflation forecasts are as accurate as the best performing nonstructural forecasting methods for all horizons and not significantly less accurate than Greenbook projections for four and five quarter ahead forecasts.[5] It is worth mentioning that DSGE model forecasts are based on few empirical time series, while some of the nonstructural forecasts exploit the information content of all 109 variables in the real-time dataset. Combining interest rate forecasts from DSGE models improves the forecasting accuracy. However, a simple BVAR forecast leads to forecasts that are as precise as any considered combination of DSGE model forecasts.

While point forecasts are interesting, economists are concerned about the uncertainty surrounding

[5]The structural interpretability of DSGE model forecasts is not lost by combining forecasts from several models. For example one can use historical decompositions to investigate which structural shocks drive the predicted macroeconomic dynamics (c.f. Wieland and Wolters, 2012). Deviding these shocks into broad categories that are included in all considered models like demand, supply and monetary policy shocks, one can weight the shock contributions of these different categories with the combination weights of the forecasting models to get the overall contribution of each shock category to a combined forecast.

these. Answers to questions like what is the probability of output growth being above 2% while inflation is between 0 and 2% are of interest to policy makers and DSGE models estimated with Bayesian techniques are suitable to answer these questions. However, for those answers to be meaningful, density forecasts of DSGE models should be a realistic description of actual uncertainty. So far the literature on the evaluation of DSGE model forecasts has focused on the evaluation of point forecasts.[6] I derive density forecasts for the DSGE models that take into account parameter uncertainty and uncertainty about economic shocks in the future. I find that all model forecasts overestimate actual uncertainty, i.e. density forecasts are very wide when compared with the actual distribution of data. A reason might be the tight restrictions imposed on the data. If the data rejects these restrictions, large shocks are needed to fit the models to the data resulting in high shock uncertainty (see also Gerard and Nimark, 2008). In a second step, I take into account model uncertainty and compute combined density forecasts using the same model averaging methods as for the point forecasts. This is similar to Gerard and Nimark (2008) who combine density forecasts of a DSGE model, a FAVAR model and a Bayesian VAR. Given the bad performance of individual models' density forecasts, it comes at no surprise that combined denstity forecasts overestimate uncertainty as well.

The remainder of this paper proceeds as follows. Section 2 outlines the different macroeconomic models that are used to compute forecasts. Section 3 gives an overview of the dataset. Section 4 describes the estimation and forecasting methodology. Section 5 evaluates point forecasts from the individual models and compares them to Greenbook projections and nonstructural forecasts. Section 6 describes several model combination schemes. Section 7 provides a comparison of the accuracy of weighted forecasts, individual forecasts, Greenbook projections and nonstructural forecasts. Section 8 evaluates density forecasts of individual models and weighted models. Section 9 summarizes the findings and concludes.

[6]The only exception is work by Herbst and Schorfheide (2011), who evaluate density forecasts of DSGE models. Their paper has been written simultaneously with my paper.

2 Forecasting Models

I consider five different models of the U.S. economy. Four are structural New Keynesian macroeconomic models and one model is a Bayesian VAR. The latter is representative of simple vector autoregression models that are often used to summarize macroeconomic dynamics without imposing strong theoretical restrictions. It is thus the unrestricted counterpart of the three variables output growth, inflation and the federal funds rate that are common to the four structural models. The models are chosen to broadly reflect the variety of DSGE models used in academia and at policy institutions.[7] I briefly describe the main features of the models. All models have been applied in Wieland and Wolters (2011) to compute point forecasts during the last five U.S. recessions.

Small New Keynesian Model estimated by Del Negro & Schorfheide (DS) The New Keynesian model is described, e.g., in Goodfriend and King (1997) and Rotemberg and Woodford (1997). It is often referenced to be the workhorse model in modern monetary economics and a comprehensive analysis is presented in the monograph of Woodford (2003). The model consists of three main equations: an IS curve, a monetary policy rule and a Phillips curve. The expectational IS curve can be derived from the behavior of optimizing and forward looking representative households that have rational expectations. Together with a monetary policy rule it determines aggregate demand. The New Keynesian Phillips curve determines aggregate supply and can be derived from monopolistic firms that face sticky prices. Del Negro and Schorfheide (2004) use Bayesian estimation to fit the model to output growth, inflation and interest rate data. The methodology is reviewed in An and Schorfheide (2007). Wang (2009) shows that the small number of frictions is sufficient to provide reasonable output growth and inflation forecasts.

Small Model with Overlapping Wage Contracts by Fuhrer & Moore (FM) This is a small scale model of the U.S. economy described in Fuhrer (1997). It differs from the New Keynesian model with respect to the degree of forward lookingness and the specification of sticky prices. Aggregate demand is determined by a reduced form backward looking IS curve together with a monetary policy rule.

[7] A comparison to large scale econometric models in the tradition of the Cowles Commission is unfortunately more burdensome. Fair (2007) compares the forecasting accuracy of a large econometric model to a DSGE model by Del Negro et al. (2007).

Aggregate supply is modelled via overlapping wage contracts: agents care about real wage contracts relative to those negotiated in the recent past and those that are expected to be negotiated in the near future (see Fuhrer and Moore, 1995a,b). The aggregate price level is a constant mark-up over the aggregate wage rate. The resulting Phillips curve depends on current and past demand and expectations about future demand. Fuhrer (1997) uses maximum likelihood estimation to parameterize the model. In contrast to all other models in this paper, variables are not defined in percentage deviations from the steady state. While a measurement equation is needed to link output growth via a trend growth rate to the data, inflation and the interest rate are directly defined in the model equations as in the data.

Medium Scale Model by Smets & Wouters (SW) The small New Keynesian model has been extended by Christiano et al. (2005) to fit a high fraction of U.S. business cycle dynamics. It is a closed economy model that incorporates physical capital in the production function and capital formation is endogenized. Labor supply is modelled explicitly. Nominal frictions include sticky prices and wages as well as inflation and wage indexation. Real frictions include consumption habit formation, investment adjustment costs and variable capital utilization. Smets and Wouters (2007) added nonseparable utility and fixed costs in production. They replaced the Dixit-Stiglitz aggregator with the aggregator by Kimball (1995) which leads to a non-constant elasticity of demand. The model includes equations for consumption, investment, price and wage setting as well as several identities. Smets and Wouters (2007) used Bayesian estimation with a complete set of structural shocks to fit the model to seven U.S. time series.

Medium Scale Model by Edge, Kiley & Laforte (FRB/EDO) The FRB/EDO (Estimated Dynamic Optimized) model by Edge et al. (2008) has been developed at the Federal Reserve and also builds on the work by Christiano et al. (2005). It features two production sectors, which differ with respect to the pace of technological progress. This structure can capture the different growth rates and relative prices observed in the data. Accordingly, the expenditure side is disaggregated as well. It is divided into business investment and three categories of household expenditure: consumption of non-durables and services, investment in durable goods and residential investment. The model is able to capture different cyclical properties in these four expenditure categories. As in the Smets & Wouters model all behavioral equations are derived in a completely consistent manner from the optimization problems

of representative households and firms. The model is documented in Edge et al. (2007).[8] To estimate the model using Bayesian techniques, 14 structural shocks are added to the equations and the model is estimated on eleven time series.

Bayesian VAR (BVAR) In addition to the four structural models, I estimate a VAR on output growth, inflation and the federal funds rate using four lags. The VAR is a more general description of the data than the DSGE models as it imposes little restrictions on the data generating process. All variables are treated symmetrically and therefore the VAR incorporates no behavioral interpretations of parameters or equations. Unrestricted VARs are heavily overparametrized and therefore not suitable for forecasting. I therefore use a Minnesota prior (see Doan et al., 1984). This prior implies shrinking the parameters towards zero by assuming that the price level, real output and the interest rate follow independent random walks. The prior variance of the parameters decreases with the lag length. The rationale for this assumption is that short lags contain more information about the dependent variables than long lags. Including more variables into the Bayesian VAR did not help to increase the forecasting accuracy. Therefore, I use a version that uses the same three core variables that are contained in all four DSGE models. This can be helpful to disentangle the importance of theoretical foundations and a parsimonious parametrization for accurate forecasts.

Table 1 summarizes the most important features of the four structural models and the Bayesian VAR.[9] It is apparent that the size of the models differs a lot from each other. Furthermore the number of estimated parameters per equation are different. The FRB/EDO model includes about one parameter per equation implying high cross equation restrictions. The authors added measurement errors to the model to fit it to 11 time series. The Fuhrer & Moore model in contrast has two parameters per equation.

The method of estimating the structural parameters also varies across the models: I adapt the methodology used by the original authors and use maximum likelihood estimation for the Fuhrer & Moore model while Bayesian estimation is used to estimate the other models.For the priors, I use the ones in the original research referenced in Table 1. Except for the model by Fuhrer & Moore,

[8] My version is not able to replicate the figures in the documentation exactly, but is reasonably close.

[9] The number of equations refers to all equations in a model taking into account shock processes, measurement equations and identities. For example the standard New Keynesian model consists of 3 structural equations, 2 shock processes (+1 iid shock) and 3 measurement equations.

Table 1: Model Overview

Type	Eq.	Par.	Est. Par.	Observable Variables	Reference
Small-scale microfounded forward looking New Keynesian Model	8	13	13	3: output growth, inflation, interest rate	Del Negro and Schorfheide (2004)
Small-scale model with overlapping real wage contracts and a backward looking IS curve	10	20	19	3: output growth, inflation, interest rate	Fuhrer (1997)
Medium-scale DSGE-model with many nominal and real frictions as used by policy institutions	27	42	37	7: output growth, consumption growth, investment growth, inflation, wages, hours, interest rate	Smets and Wouters (2007)
Large-scale DSGE-model developed at the Federal Reserve. Two production sectors with different technology growth rates. The demand side is disaggregated into four categories	59	71	51	11: output growth, inflation, interest rate, consumption of nondurables and services, consumption of durables, residential investment, business investment, hours, wages, inflation for consumer nondurables and services, inflation for consumer durables	Edge et al. (2008)
Bayesian VAR with 4 lags; Minnesota priors	3	39	39	3: output growth, inflation, interest rate	Doan et al. (1984)

Notes: Type: short classification of the models according to the main modeling assumptions; Eq.: number of equations including shock processes, measurement equations and identities, but excluding variable definitions and flexible price allocations; Par.: total number of parameters in the model file excluding all auxiliary parameters; Est. Par.: exact number of estimated parameters including shock variances and covariances; Observable Variables: the number and names of the observable variables; Reference: original reference that is closest to the implemented version in this paper.

variables are defined in percentage deviations from steady state and thus measurement equations that include an output growth trend and the steady state of inflation, the interest rate and other observables are needed to link the equations to the data. The FRB/EDO model is implemented nonlinearly and I compute a first order approximation of the solution. All other models are linearized.

3 A real-time dataset

I use the real-time dataset described in Faust and Wright (2009).[10] These historical data vintages have been used by the Federal Reserve staff to compute the Greenbook forecasts. Thus, the dataset is perfectly synchronized with the Greenbook and contains historical samples of 109 variables as observed at the time the Greenbook was published. The dataset contains data vintages for 145 FOMC meetings from March 1980 to December 2000, while the different data series start in 1960.[11] While

[10]The dataset can be downloaded from the website of Jon Faust: http://e105.org/faustj/papgbts.php?d=n. A detailed data appendix is available on the same website.

[11]The dataset ends in 2000 because Greenbook data remains confidential for 5 years after the forecast date. I don't update the data for the additional years that are now available to make the forecasting results directly comparable to Faust and

some of the nonstructural forecasting models considered in Faust and Wright (2009) can process as many data series as available, the structural models considered in this paper use only a small subset of the available time series varying from three to eleven variables to estimate the different models. Still some variables for the FRB/EDO model are not available in the dataset. Therefore, I have added the necessary real-time data series from the Federal Reserve Bank of St. Louis' Alfred database. To each data vintage I add only observations that would have been available at the Greenbook publication date.

The models by Del Negro & Schorfheide, Fuhrer & Moore and the Bayesian VAR are estimated on the three key variables output growth, inflation and the federal funds rate. The Smets & Wouters model is estimated on the three key variables and a wage time series, hours worked, consumption and investment. The FRB/EDO model is estimated on eleven empirical time series: output growth, inflation, the federal funds fate, consumption of non-durables and services, consumption of durables, residential investment, business investment, hours, wages, inflation for consumer nondurables and services and inflation for consumer durables.[12]

There is a trade-off between using a long sample to get precise parameter estimates and for leaving out a fraction of past data that might contain structural breaks. Therefore, I use a moving window of the latest eighty quarterly observations of each data vintage to estimate the models. Aside from structural breaks the high inflation periods of the 70's and 80's influence the estimated inflation steady state which can bias the inflation forecasts of the late 80's and the 90's. A window of eighty observations gives at least the chance of a diminishing effect on the forecasts. The first sample for the FOMC meeting of March 1980 starts in 1960Q1 and ends in 1979Q4, the second sample for the FOMC meeting of April 1980 starts in 1960Q2 and ends in 1980Q1, and this goes on until the last sample for the FOMC meeting of December 2000 that starts in 1980Q4 and ends in 2000Q3.

Wright (2009).

[12]Output is in real terms available in the dataset and growth rates can be computed directly. Consumption, investment and wages are expressed in real terms as defined in the models through division with the output deflator. Growth rates are computed afterwards. Inflation is computed as the first difference of the log output deflator. The nominal interest rate is expressed on a quarterly basis. I compute hours per capita by dividing aggregate hours with civilian employment (16 years and older). The hours per capita series includes low frequent movements in government employment, schooling and the aging of the population that cannot be captured by the models. I remove these following Francis and Ramey (2009) by computing deviations of the hours per capita series from its low frequent (one-sided) HP-filtered trend with a parameter of 16000. The realtime characteristic of the data remains unaffected by this procedure. For the FRB/EDO model growth rates are computed for real output, consumption of non-durables and services, consumption of durables, residential investment and business investment. Inflation of nondurables and services and inflation of durable goods is computed by dividing the accordant nominal and real time series and calculating log first differences.

I forecast annualized quarterly real output growth as measured by the GNP/GDP real growth rate, annualized quarterly inflation as measured by the GNP/GDP deflator and the federal funds rate. GDP data is first released about one month after the end of the quarter to which the data refer, the so-called advance release. These data series are then revised several times at the occasion of the preliminary release, final release, annual revisions and benchmark revisions. I follow Faust and Wright (2009) and use actual realized data as recorded in the data vintage that was released two quarters after the quarter to which the data refer to evaluate the forecasting accuracy. For example, revised data for 1999Q1 is obtained by selecting the entry for 1999Q1 from the data vintage released in 1999Q3. Hence, I do not attempt to forecast annual and benchmark revisions, because the models cannot predict changes in data definitions. The revised data against which the accuracy of forecasts is judged will typically correspond to the final NIPA release.

DSGE model forecasts are compared to Greenbook projections and nonstructural forecasts from Faust and Wright (2009). The dataset by Faust & Wright contains Greenbook nowcasts and forecasts up to five quarters ahead for all variables. However, I have spotted some differences between the Greenbook projections in the dataset by Faust & Wright and the Greenbook projection dataset that is published by the Federal Reserve Bank of Philadelphia.[13] The differences are neglectable for inflation and the interest rate. However, for GDP nowcasts and one quarter ahead forecasts the dataset by Faust & Wright understates the accuracy of the Greenbook projections quite a bit. A detailed comparison of the two datasets is contained in Appendix A. For the evaluation of forecasts I have replaced the Greenbook projections from the Faust & Wright dataset with the Greenbook projections from the historical Greenbook documents published by the Philadelphia Fed.[14]

[13]The Federal Reserve Bank of Philadelphia has made available scanned pdf files of all Greenbook projections from 1966-2005: http://www.philadelphiafed.org/research-and-data/real-time-center/greenbook-data/pdf-data-set.cfm. Previously only Greenbook projections of four of the eight FOMC meetings per year were available as an Excel-file: http://www.philadelphiafed.org/research-and-data/real-time-center/greenbook-data/philadelphia-data-set.cfm. The update allowed a comparison of all Greenbook projections contained in the dataset by Faust and Wright (2009) with the original scanned documents. The assumed interest rate path that the Greenbook projections are conditioned on is also available from the website of the Federal Reserve Bank of Philadelphia: http://www.philadelphiafed.org/research-and-data/real-time-center/greenbook-data/gap-and-financial-data-set.cfm.

[14]Accordingly, I recalculate the root mean squared prediction errors (RMSE) of the nonstructural forecasts that are stated relative to Greenbook RMSEs by Faust and Wright (2009). I multiply the relative RMSEs with the RMSEs of the Greenbook projections by the Faust & Wright dataset and devide by the RMSEs of Greenbook projections of the Philadelphia Fed dataset.

4 Forecasting Methodology

Computing recursive forecasts using structural models and real-time data vintages requires a sequence of steps that are explained in the following. First, the models need to be specified, solved and linked to the empirical data. Second, the data needs to be updated to the current vintage and parameters have to be estimated. Third, density and point forecasts are computed.

Model specification and solution. Each of the models consists of a number of linear or nonlinear equations that determine the dynamics of the endogenous variables. A number of structural shocks is included in each model. Any of the models $m = 1, ..., 4$ can be written as follows:

$$E_t\left[f_m(y_t^m, y_{t+1}^m, y_{t-1}^m, \varepsilon_t^m, \beta^m)\right] = 0 \quad (1)$$

$$E(\varepsilon_t^m) = 0 \quad (2)$$

$$E(\varepsilon_t^m \varepsilon_t^{m\prime}) = \Sigma_\varepsilon^m, \quad (3)$$

where $E_t\left[f_m(.)\right]$ is a system of expectational difference equations, y_t^m is a vector of endogenous variables, ε_t^m a vector of exogenous stochastic shocks, β^m a vector of parameters and Σ_ε^m is the variance-covariance matrix of the exogenous shocks. The parameters and the variance-covariance matrix are either calibrated or estimated or a mixture of both.

A subset of the endogenous variables consists of empirically observable variables $y_t^{m,obs}$. If variables in the models are defined in percentage deviations from steady state then there is a subset of the equations that are so-called measurement equations $f_m^{obs}(.)$. These link the observable variables to the other endogenous variables through the inclusion of steady state values or steady state growth rates. Another possibility is that the observable variables are directly included in the general equations of a model. The latter is the case in the Fuhrer & Moore model. Inflation and the interest rate are included in the model as they appear in the data and are not redefined as deviations from steady states. For the FRB/EDO model, it is assumed that not all observable variables are measured exactly and therefore a set of nonstructural measurement shocks is added to the measurement equations.

The system of equations is solved using a conventional solution method for rational expectations models such as the technique of Blanchard and Kahn. In the case of the FRB/EDO model a first or-

der approximation of the solution is derived. The other models are already linearized before solving them.[15] Given the solution, the following state space representation of the system is derived:

$$y_t^{m,obs} = \Gamma^m \bar{y}^m + \Gamma^m y_t^m + \varepsilon_t^{m,obs}, \quad (4)$$

$$y_t^m = g_y^m(\beta^m) y_{t-1}^m + g_\varepsilon^m(\beta^m)\varepsilon_t^m, \quad (5)$$

$$E(\varepsilon_t^m \varepsilon_t^{m\prime}) = \Sigma_\varepsilon^m \quad (6)$$

The first equation summarizes the measurement equations and shows the link between observable variables and the endogenous model variables via steady state values or deterministic trends $\bar{y}^m$. The matrix Γ^m might include lots of zero entries as not all variables are directly linked to observables. The measurement errors $\varepsilon_t^{m,obs}$ are a subset of the shocks ε_t^m. The second equation constitutes the transition equations including the solution matrices g_y^m and g_ε^m that are nonlinear functions of the structural parameters β^m. The transition equations relate the endogenous variables to their own lags and the vector of exogenous shocks. The third equation denotes the variance-covariance matrix Σ_ε^m.

Estimation. Having solved the model and linked to the data, one needs to update the data before estimating the model. I use for each forecast the 80 most recent observations of the respective historical data vintage that was available at the time of the forecast start. Estimating DSGE models using Bayesian estimation has become a popular approach due to the combination of economic theory which is imposed on the priors and data fit summed up in the posterior estimates. A survey of the methodology is presented in An and Schorfheide (2007). Therefore, I only give a short overview of the algorithm. Due to the nonlinearity in β^m the calculation of the likelihood is not straightforward. The Kalman filter is applied to the state space representation to set up the likelihood function (see e.g. Hamilton, 1994, chapter 13.4)[16]. Since the models considered are stationary, one can initialize the Kalman Filter using the unconditional distribution of the state variables. Combining the likelihood with the priors yields the log posterior kernel $ln\mathscr{L}(\beta^m|y_1^{m,obs},...,y_t^{m,obs}) + lnp(\beta^m)$ that is maximized over β^m using numerical methods to compute the posterior mode. The posterior distribution of the parameters is a complicated nonlinear function of the structural parameters. The Metropolis-Hastings

[15] I use the solution procedure of the Dynare software package. See www.dynare.org and Juillard (1996) for a description.

[16] I consider only unique stable solutions. If the Blanchard-Kahn conditions are violated I set the likelihood equal to zero.

algorithm offers an efficient method to derive the posterior distribution via simulation. Details are provided for example in Schorfheide (2000). I compute 500000 draws from the Metropolis-Hastings algorithm and use the first 25000 of these to calibrate the scale such that an acceptance ratio of 0.3 is achieved. Another 25000 draws are disregarded as a burn in sample. The models are reestimated for the first data vintage of each year. Finally, the mean parameters can be computed from the posterior distribution of β^m.

Forecast computation. Having estimated the different models, forecasts for the horizons $h \in (0,1,2,3,4,5)$ are derived. First, a density forecast is computed and afterwards a point forecast is calculated as the mean of the density forecast. For each parameter a large number of values are drawn from the parameter's posterior distribution. For a random draw s a projection of the observable variables is derived by iterating over the solution matrix $g_y^m(\hat{\beta}^{m,s})$. At each iteration i in addition a vector of shocks $\varepsilon_i^{m,s}$ is drawn from a mean zero normal distribution where the variance is itself a random draw from the posterior distribution of the variance-covariance matrix:

$$y_{t+h}^{s,m,obs} = \Gamma^m \hat{\bar{y}}^{m,s} + \Gamma^m g_y^m(\hat{\beta}^{m,s})^{h+1} y_{t-1}^m + \Gamma^m \sum_{i=0}^{h} g_\varepsilon^m(\hat{\beta}^{m,s})^{(h+1-i)} \varepsilon_i^{m,s} \quad (7)$$

$$\varepsilon_i^{m,s} \sim N(0, \hat{\Sigma}_\varepsilon^{m,s}), \quad (8)$$

where a hat on the structural parameters $\beta^{m,s}$, the variance covariance matrix $\Sigma_\varepsilon^{m,s}$ and the steady state values of observable variables $\bar{y}^{m,s}$ denotes that they are estimated. The reduced form solution matrices g_y^m and g_ε^m are functions of the estimated parameters and change over time as the models are reestimated. The procedure is repeated 10000 times ($s = 1, ..., 10000$) and finally the forecast density is given by the ordered set of forecast draws $y_{t+h}^{s,m,obs}$. The point forecast is given by the mean of the forecast density.

The different steps to compute forecasts are:

1. Model specification: set up a file with the model equations and add measurement equations that link the model to the empirical time series.
2. Solution: solve the model and express it in state space form.
3. Data update: update the data with the current vintage.

4. Estimation: reestimate the model for the first data vintage of each year. Otherwise, use the posterior distribution of the parameters from previous estimation. Add a prior distribution of the model parameters. Estimate the structural parameters by maximizing the posterior kernel. Afterwards simulate the posterior distribution of the parameters using the Metropolis-Hastings algorithm.
5. Density forecast: compute forecast draws by iterating over the solution matrices for different parameter values drawn from the posterior distribution. At each iteration draw a vector of shocks from a mean zero normal distribution with the variance itself being a draw from the posterior distribution. The forecast density is given by the ordered forecast draws.
6. Mean forecast: compute the mean of the forecast density to get the point forecast.
7. Repeat steps 3 to 6 for all data vintages.
8. Repeat steps 1 to 7 for different models, possibly extending the information set by additional variables as required by the respective model.

Figure 1 shows as an example forecasts for output growth, inflation and the federal funds rate derived from data vintage May 12, 2000. The black line shows real-time data until the forecast start and revised data afterwards. I plot the 0.05, 0.15, 0.25, 0.35 and 0.65, 0.75, 0.85 and 0.95 percentiles to graphically represent the density forecasts. The different shades therefore show for 90%, 70%, 50% and 30% probability bands. The line in the middle of the confidence bands shows the mean forecast for each model. The short white line shows the correspondent Greenbook projections. In addition to forecasts from the four DSGE models and the Bayesian VAR, I plot the mean of the four DSGE model forecasts. Data is available until the first quarter of 2000. The current state of the economy in the second quarter of 2000 is estimated using the different models.

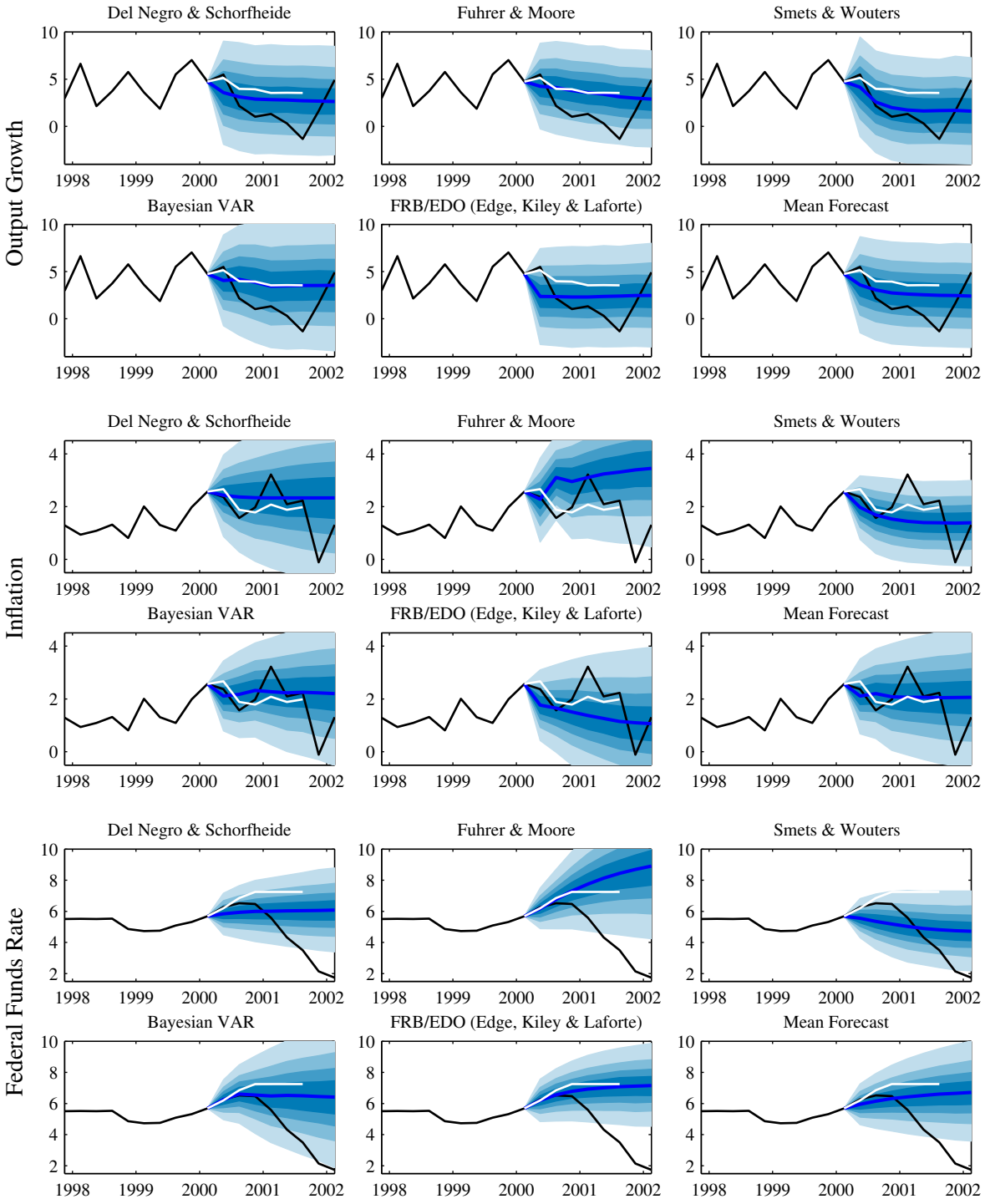
The economy was in a boom in early 2000 and the models broadly predict the return to average growth rates over the next quarters. They are not able to predict the 2001 recession that has been defined by the NBER to take place between the first and the fourth quarter of 2001. Inflation is predicted by the Del Negro & Schorfheide model and the Bayesian VAR to stay on a similar level as in the first quarter of 2000. The Fuhrer & Moore model predicts an increase of the inflation rate. The FRB/EDO and the Smets & Wouters models are able to predict the inflation decrease in the third

quarter of 2000. None of the models is able to predict the short inflation increase in the first quarter of 2001. The interest rate is forecast to increase by the Fuhrer & Moore model, the FRB/EDO model and the Bayesian VAR. It is predicted to stay constant by the Del Negro & Schorfheide model and to decrease by the Smets & Wouters model. These examples already indicate that forecasts from different DSGE models can deviate a lot from each other. The average of the four DSGE model forecasts predicts the interest rate path quite precisely until the end of the year. The decrease in the federal funds rate beginning in 2001 is not captured by the forecasts. This is consistent with the output growth forecasts that miss the recession in 2001 that is in turn a reason for the interest rate cuts.

I plot a figure like this for the forecasts derived from each data vintage. Unfortunately, it is not possible to show all these figures in this paper. However, screening over all the forecasts for the different historical data vintages reveals some notable observations. Structural models and the Bayesian VAR are suited to forecast during normal times. Given small or average exogenous shocks the models give a good view about how the economy will return back to steady state. In contrast, large recessions or booms and the respective turning points are impossible to forecast with these models. Figure 2 plots the forecast errors (outcome minus forecast) of all models on the horizontal axis and the correponding realized output growth rate on the vertical axis. A clear positive relation is visible. When output growth is highly negative the models are not able to forecast such a sharp downturn and thus the forecast error is negative. The models require large exogenous shocks to capture large deviations from the balanced growth path and the steady state inflation and interest rate. This is due to the weak internal propagation mechanism of the models. Therefore, for a given shock all models including the Bayesian VAR predict a quick return back to the steady state growth rate. While the point forecasts cannot predict a recession, the possibility that a large deviation from steady state values occurs is captured by the density forecasts. Once the turning point of a recession has been reached, all models predict the economic recovery back to the balanced growth path well.[17]

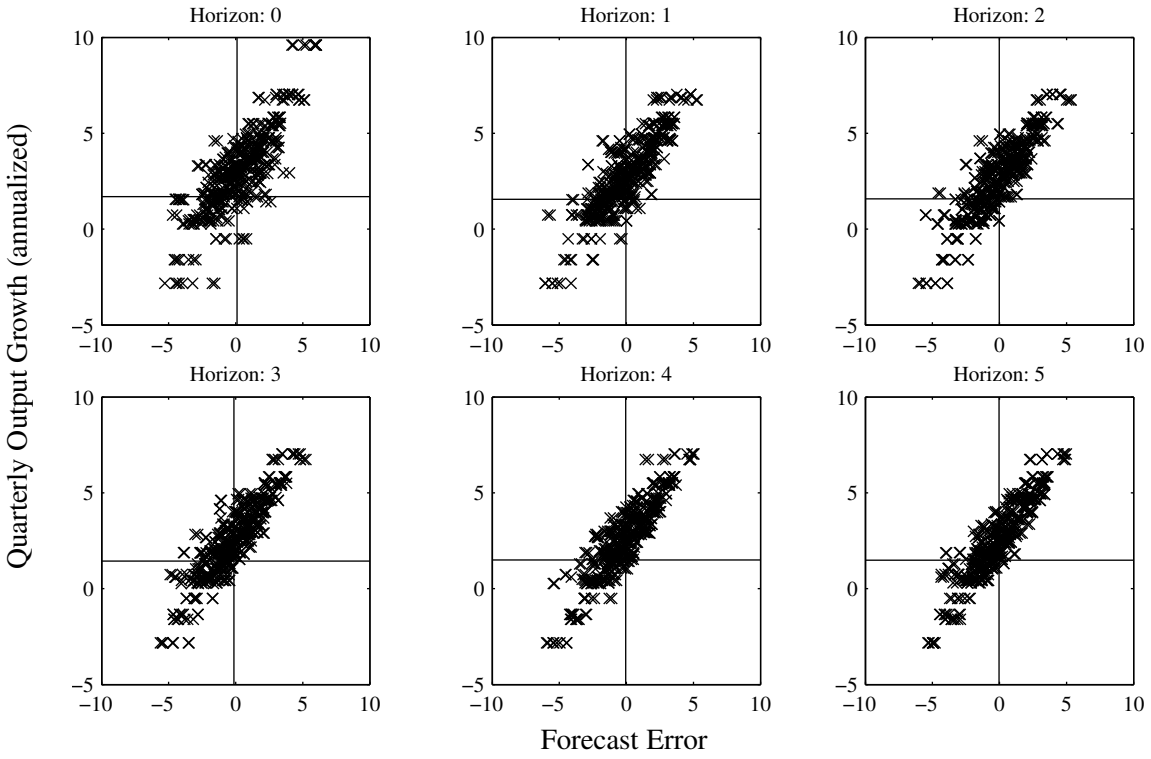
[17]The accuracy of model-based forecasts during recessions and recoveries is analysed more systematically in Wieland and Wolters (2011). They find that models as well as professional forecasters fail to predict recessions, but predict recoveries quite accurately.

Figure 1: Structural Forecasts; Data Vintage May 12, 2000



Notes: the black line shows real-time data until the forecast start and revised data afterwards; the shaded areas show 90% 70%, 50% and 30% confidence bands; the line in the middle of the confidence bands shows the mean forecast for horizons 0 to 7; the short white line shows the Greenbook forecast for horizons 0 to 5. Mean Forecast is the average of the four model forecasts.

Figure 2: Forecast Errors and Output Growth Rates



Notes: the figure shows observed output growth rates and the corresponding forecast errors of the four DSGE models for different forecasting horizons. The horizontal lines show the mean output growth rate and the vertical line the mean forecast errors of all models for each horizon.

5 Forecast Evaluation

Table 2 reports root mean squared prediction errors (RMSE) for output growth, inflation and interest rate forecasts from the Greenbook, the four structural models, the Bayesian VAR and the respective best and worst performing nonstructural model considered by Faust and Wright (2009). The first column gives the RMSE for the Greenbook and all other columns report the RMSE of the specific models relative to the Greenbook RMSE. Values less than one show that a model forecast is more accurate than the corresponding Greenbook projection. The last two columns report the relative RMSEs of the most and the least accurate nonstructural forecasting model from Faust and Wright (2009) for each horizon.

The first six rows in each table show forecasts based on the available data at the starting point of

the forecast. The current state of the economy is not available in the data and therefore needs to be estimated. This nowcast is labeled as a forecast for horizon zero. As the data becomes available with a lag of one quarter, the results are labeled as "jump off -1". In practice, however, there are many data series that are available on a monthly, weekly or daily frequency that can be used to improve current-quarter estimates of GDP. Examples are industrial production, sales, unemployment, opinion surveys, interest rates and other financial prices. This data can be used to improve nowcasts and the Federal Reserve staff and many professional forecasters certainly make use of it. To approximate the effect of using more information in nowcasting, I investigate the effect of using Greenbook nowcasts as a starting point for model-based forecasts regarding future quarters. The results are shown in the last five rows of each table and are labeled as "jump off 0".

I follow Faust and Wright (2009) in leaving out the period from 1980-1983 from the evaluation as this period was very volatile and might bias the assessment of forecasting accuracy for the whole sample. Therefore, the results start in 1984 so that RMSEs for output growth and inflation are directly comparable to Table 2 in Faust and Wright (2009). Reported RMSEs are thus based on 123 forecasts from 1984 to 2000. I evaluate whether the difference of Greenbook RMSEs and model RMSEs is statistically significant based on the Diebold-Mariano statistic (Diebold and Mariano, 1995) using a symmetric loss function. Asymptotic p-values are computed using Newey-West standard errors with a lag-length of 10, covering a bit more than a year, to account fore serial correlation of forecast errors.[18]

Results for inflation, output growth and the federal funds rate are very different from each other. For output growth the Greenbook nowcast is more precise than the model nowcasts. This was expected as the Fed can exploit more information about the current state of the economy. However, this precise estimate of the current state of the economy does translate only into a slightly superior forecasting performance one quarter ahead, but no more accurate forecasts than from different methods for higher horizons. The SW, EDO and BVAR models' forecasts dominate the Greenbook forecast from horizon 3 onwards. The DS model yields a similar forecasting accuracy as the Greenbook. Only the FM model is slightly less accurate than the Greenbook forecast for all horizons. If I include the Greenbook

[18]The significance levels for the nonstructural forecasting models from Faust and Wright (2009) are only indicative. They are based on the Greenbook projections used in Faust and Wright (2009). Using the Greenbook projections from the Philadelphia Fed dataset (see Appendix A for a comparison of the two datasources), it is likely that the Faust & Wright output growth nowcasts are significantly different from the Greenbook nowcast on the 1% level. For the DSGE models differences in the output growth nowcast changed from insignificance to being significant on the 1% level when switching from the Faust & Wright Greenbook projections to the Philadelphia Fed Greenbook projections.

Table 2: Greenbook RMSE and relative RMSE of model forecasts: 1984-2000

(a) Output growth

horizon	GB	DS	FM	SW	EDO	BVAR	best FW	worst FW
				jump off -1				
0	1.52	1.41●	1.32●	1.43●	1.40●	1.29●	1.25	1.60
1	1.85	1.09	1.22	1.05	1.04	1.11	**0.99**	1.38
2	2.01	1.05	1.10	**0.92**	1.01	**0.95**	**0.95**	1.15
3	2.15	**0.99**	1.09	**0.86**•	**0.94**	**0.97**	**0.94**	1.12
4	2.08	1.00	1.05	**0.89**	**0.95**	**0.94**	**0.99**	1.11
5	2.08	1.02	1.06	**0.89**	**0.99**	1.00	**0.97**	1.09
				jump off 0				
1	1.85	1.08	1.19•	1.07	1.07	1.07	**0.96**	1.23
2	2.01	1.06	1.14	**0.92**	1.01	**0.97**	**0.90**	1.12
3	2.15	1.00	1.13	**0.86**•	**0.96**	**0.96**	**0.95**	1.18
4	2.08	1.01	1.08	**0.88**	**0.98**	**0.97**	**0.96**	1.09
5	2.08	1.03	1.08	**0.89**•	1.01	**0.99**	**0.98**	1.11

(b) Inflation

horizon	GB	DS	FM	SW	EDO	BVAR	best FW	worst FW
				jump off -1				
0	0.70	1.51●	1.84●	1.49●	1.65●	1.46●	1.32●	1.61●
1	0.80	1.57●	1.76●	1.42●	1.48●	1.42●	1.20●	1.84●
2	0.82	1.35●	1.54●	1.27●	1.55●	1.28●	1.14•	1.90●
3	0.94	1.16●	1.40●	1.17●	1.46●	1.12	1.02	1.82●
4	0.91	1.24●	1.74●	1.25●	1.43●	1.30•	1.06	2.06●
5	1.15	1.22•	1.59●	1.22●	1.31●	1.27	**0.98**	1.81●
				jump off 0				
1	0.80	1.21●	1.58●	1.12	1.16•	1.23●	1.19●	1.56●
2	0.82	1.25●	1.51●	1.18•	1.15	1.24●	1.17	1.67●
3	0.94	1.24●	1.26●	1.20●	1.23●	1.13	1.03	1.64●
4	0.91	1.16●	1.46●	1.16●	1.22•	1.14	1.03	1.87●
5	1.15	1.13●	1.45●	1.18●	1.11	1.16	**0.96**	1.75●

(c) Federal Funds Rate

horizon	GB	DS	FM	SW	EDO	BVAR	best FW	worst FW
				jump off -1				
0	0.11	6.38●	5.02●	4.95●	6.19●	3.71●	-	-
1	0.52	2.00●	1.75●	1.76●	2.26●	1.46●	-	-
2	0.94	1.42●	1.38●	1.30•	1.69●	1.13	-	-
3	1.29	1.14	1.21	1.06	1.50●	**0.99**	-	-
4	1.64	1.03	1.19•	**0.95**	1.38●	**0.94**	-	-
5	1.93	**0.96**	1.21•	**0.86**	1.29●	**0.91**	-	-
				jump off 0				
1	0.52	1.28•	1.24●	1.11	1.59●	1.03	-	-
2	0.94	1.13	1.03	1.02	1.48●	**0.93**	-	-
3	1.29	**0.99**	**0.97**	**0.92**	1.42●	**0.88**	-	-
4	1.64	**0.94**	1.00	**0.87**	1.36●	**0.86**	-	-
5	1.93	**0.89**	1.06	**0.82**	1.29●	**0.85**	-	-

Notes: GB: Greenbook; DS: Del Negro & Schorfheide; FM: Fuhrer & Moore; SW: Smets & Wouters; EDO: FRB/EDO Model by Edge, Kiley & Laforte; BVAR: Bayesian VAR; Best/Worst FW: Best/Worst performing atheoretical model for the specific horizon considered by Faust & Wright. The first column shows the forecast horizon. The second column shows the Greenbook RMSE. The other columns show RMSEs relative to the Greenbook. Values less than one are in bold and show that a forecast is more accurate than the one by the Greenbook. The symbols ●, •, •, indicate that the relative RMSE is significantly different from one at the 1, 5, or 10% level, respectively. Significance levels are only an approximation for the results in the last two columns as they are based on another Greenbook dataset (see Appendix A, the correct significance level for the output growth nowcast is more likley to be 1% for bestFW and worstFW).

nowcast in the information set used to compute forecasts the results hardly change as quarterly output growth is not very persistent. Viewing the Greenbook as a best practice benchmark, one could be tempted to judge the forecasting ability of the structural models as very good at least for medium term horizons. However, one should keep in mind that quarterly output growth has little persistence and thus is difficult to forecast in general. The reported RMSEs in Faust and Wright (2009) show that none of their nonstructural forecasting methods is more accurate than a univariate autoregressive forecast.[19] I find that only the SW model's forecasts are more precise than an autoregressive forecast from horizon 3 onwards. The forecasting accuracy of the EDO and BVAR model is similar to the autoregressive forecast and the DS and FM forecasts are less precise. In addition, none of the models' RMSEs differ statistically significantly from the Greenbook RMSEs except for the nowcast. The difference in the forecasting accuracy of the models can be traced to the different modeling assumptions. The SW and EDO model have a richer economic structure than the DS and FM model. The BVAR also performs very good as the higher number of lags compared to the other models can catch important business cycle dynamics. Despite this richer structure the SW, EDO and BVAR models are apparently tightly enough parametrized to yield precise forecasts.

Greenbook inflation forecasts are more accurate than all structural as well as all nonstructural inflation forecasts. The structural forecasts have an accuracy in line with the accuracy range of the nonstructural forecasts. However, none of the DSGE models reaches the forecasting quality of the best nonstructural forecasts. Among the DSGE models the DS and SW model show a good forecasting performance. They achieve a forecast of similar accuracy as the BVAR. The EDO model forecasts are somehow less precise and the FM forecasts are relatively imprecise. The forecasting accuracy relative to the Greenbook forecasts improves with increasing horizons for all models. When I add the Greenbook nowcast to the information set of the models, the forecasting accuracy increases, but does not reach the quality of the Greenbook projections.[20] While it is not possible to forecast inflation with DSGE models as precise as the Fed does, the forecasts are reasonable: with the exception of the FM model they are as good or better than a simple autoregressive forecast from horizon 3 onwards and for

[19] Faust and Wright (2009) consider two types of autoregressive forecasts. First, a recursive autoregression, where the h-period ahead forecast is constructed by recursively iterating the one-step ahead forecast forward. Second, they use a direct forecast from the autoregression by regressing h-period ahead output growth values on the autoregressive process. For both types they use four lags and get a similar forecasting accuracy.

[20] Especially, the forecasting accuracy of the FRB/EDO model increases considerably. This model can thus deliver in principle precise forecasts if conditioned on a precise nowcast.

all horizons for the jump of 0 scenario.

Greenbook projections are conditioned on a hypothetical path of policy. This hypothetical federal funds rate is not meant to be a forecast. Nevertheless, treating it as a forecast shows that its accuracy for short horizons is extremely high. Therefore, the Fed might have conditioned the projections on a policy path that is likely to be implemented in the next few quarters and it is reasonable to view this as a forecasting benchmark. Faust and Wright (2009) did not compute interest rate forecasts, so that I cannot compare the structural forecasts to forecasts from their nonstructural time series models. Due to its extremely high accuracy in the short term, the structural forecasts do much worse than the Greenbook for horizons 0 to 3. Hence, for future research it might be of interest to compute DSGE model forecasts conditioned on the Greenbook interest rate path. There is the possibility that the short run accuracy of the assumed interest rate path increases the short-run forecasting accuracy for other variables as well. For medium term forecasts, however, the interest rate forecasting accuracy of the DS, SW and BVAR models dominates the Greenbook path. For short forecasting horizons it is apparent that the BVAR forecasts have a much higher accuracy than the DSGE forecasts. The monetary policy rules in the DSGE models include only few variables and might be too simple. In contrast, the policy rule implicit in the BVAR contains four lags of the interest rate, output growth and the inflation rate. Among the DSGE models the EDO forecasts are very imprecise as they underestimate the level of the interest rate many times. Taking the Greenbook nowcast as given, the forecasting accuracy of the models relative to the Greenbook increases.

Del Negro and Schorfheide (2004) propose to use DSGE models as priors for VARs. They show that the forecasting accuracy of these so-called DSGE-VARs improves relative to a VAR and partly to a BVAR with Minnesota priors. They advocate to use DSGE-VARs for forecasting until structural models are available that have the same forecasting performance. The forecasting results in Table 2 show that at least the SW models' forecasting performance for output growth, inflation and the interest rate is already good enough to be considered for forecasting exercises on its own. Furthermore, the results show that the forecasting quality of different DSGE models shows a high dispersion. Therefore, in practice it might be a good idea to compare the forecasting accuracy of a DSGE model to other DSGE models before using it as a forecasting tool on a regular basis.

Faust and Wright (2009) present a table showing the percentage of forecast periods in which the

nonstructural forecasts are more accurate than the Greenbook. This metric is not as sensitive to outliers as the RMSEs. I compute accordant numbers for the structural forecasts which are shown in Table 1 in Appendix B. A value higher than 50% indicates that the specific forecast was more accurate than the Greenbook forecast for more than half of the sample. The results are similar to the RMSE results: the Greenbook output growth nowcast and one quarter ahead forecasts are more accurate than the DSGE model counterparts. For the other horizons the model forecasts for output growth are as good as the Greenbook forecasts or even better with large differences between the different DSGE models. For inflation the Greenbook forecasts are more accurate than all model forecasts. The interest rate path of the Greenbook is more precise than model forecasts for short horizons, but model forecasts do as well as the Greenbook for medium forecasts with the EDO model being an exception.

6 Model Averaging

The forecast evaluation results in the previous section suggest that the forecasts from the model by Smets and Wouters (2007) are more precise than those of the other considered DSGE models. However, deviding the evaluation sample into three subsamples, that cover about five years each shows that the Smets & Wouters model does not continuously outperform the other DSGE models. Table 3 shows for three subsamples the model with the lowest RMSE for the different forecast horizons and for output growth, inflation and the federal funds rate, respectively. There is no model that continuously performs better than all other models. Even for a given subsample the most precise forecasts for the different variables can be generated by different models. Different frictions in different models seem to be useful for forecasting specific variables in certain periods only, while other frictions are more important for other periods. The problem of instability in the performance and predictive content of different forecasting methods is well known and surveyed in detail in Rossi (2012). Rossi (2012) provides a survey of various contributions that show that combined forecasts are one possibility to overcome such instability in the forecasting performance of nonstructural models. Timmermann (2006) surveys the literature on forecast combinations and concludes that combining multiple forecasts increases the forecasting accuracy in most cases. Unless one can identify a single model with superior forecasting performance, forecast combinations are useful for diversification reasons as one

Table 3: Best performing forecasting models for three subsamples

(a) Output Growth

horizon	1984-1990	1990-1995	1995-2000
0	FM95	FRB/EDO	FM95
1	DS04	FRB/EDO	FM95
2	SW07	FRB/EDO	FM95
3	SW07	FRB/EDO	FM95
4	SW07	SW07	FM95
5	SW07	SW07	FM95

(b) Inflation

horizon	1984-1990	1990-1995	1995-2000
0	FRB/EDO	SW07	FM95
1	FRB/EDO	SW07	FM95
2	SW07	SW07	FM95
3	SW07	SW07	FM95
4	DS04	SW07	FM95
5	DS04	SW07	FM95

(c) Federal Funds Rate

horizon	1984-1990	1990-1995	1995-2000
0	FRB/EDO	SW07	DS04
1	SW07	FM95	DS04
2	SW07	FM95	DS04
3	SW07	FM95	DS04
4	FM95	SW07	DS04
5	SW07	SW07	DS04

Notes: The table shows for each forecast horizon the best performing model in terms of RMSEs for three subsamples. DS04: Del Negro & Schorfheide (2004); FM95: Fuhrer & Moore (Fuhrer, 1995); SW07: Smets & Wouters (2007); FRB/EDO: Edge, Kiley & Laforte (2007).

does not have to rely on the forecast of a single model. I will check in the following whether this also holds for forecast combinations of DSGE models.

Using only one model for forecasting is equivalent to a subjective prior of the forecaster that the specific model is the best representation of the unknown true data generating process. Gerard and Nimark (2008) take into account model uncertainty by combining forecasts from a Bayesian VAR, a FAVAR and a DSGE model. I extend their work to combining forecasts from four DSGE models. I consider several methods to combine forecasts from the set of models: likelihood based weights, relative performance weights based on past RMSEs, a least squares estimator of weights, and non-parametric combination schemes (mean forecast, median forecast and weights based on model ranks reflecting past RMSEs). While many of these methods have been applied to nonstructural forecasts (see Timmermann, 2006, for a survey) there are to my knowledge no applications to a suite of struc-

tural models. From a theoretical point of view likelihood based weights or weights estimated by least squares are appealing. In practice, these estimated weights have the disadvantage that they introduce estimation errors. In the applied literature simple combination schemes like equal-weighting of all models have widely been found to perform better than theoretically optimal combination methods (see e.g. Hsiao and Wan, 2010, for the disconnect of Monte Carlo simulation results and empirical results).

Let I_t^m be the information set of model m at time t including the model equations, parameter estimates and the observable time series of the accordant data vintage. A combined point forecast of models $m = 1, ..., M$ for horizon h denoted as $E[y_{t+h}^{obs}|I_t^1, ..., I_t^M, \omega_{1,h}, ..., \omega_{M,h}]$ can be written as the weighted sum of individual density forecasts $p[y_{t+h}^{obs}|I_t^m]$ with assigned weights $\omega_{m,h}$ divided by the number of draws S:

$$E[y_{t+h}^{obs}|I_t^1, ..., I_t^M, \omega_{1,h}, ..., \omega_{M,h}] = \frac{1}{S}\sum_{m=1}^{M} \omega_{m,h} p[y_{t+h}^{obs}|I_t^m]. \tag{9}$$

I take 10000 draws from each individual forecast and order them in ascending order to get the density forecast for each model. Afterwards I weight each of the 10000 draws for each model with the specific model weights to compute 10000 draws of the combined forecast. This is the weighted or averaged density forecast. The weighted point forecast is computed as the mean of the 10000 draws of the weighted forecast. In the following, I discuss various methods how to choose the weights $\omega_{m,h}$.

A natural way to weight different models in a Bayesian context is to use Bayesian Model Averaging. The marginal likelihood $ML(y_T^{obs}|m)$—with T denoting all observations of a specific historical data sample observed in period t—is computed for each model $m = 1, ..., M$ and posterior probability weights are given by:

$$\omega_m = ML(m|y_T^{obs}) = \frac{ML(y_T^{obs}|m)}{\sum_{m=1}^{M} ML(y_T^{obs}|m)}, \tag{10}$$

where a flat prior belief about model m being the true model is used so that no prior beliefs show up in the formula. This weighting scheme is based on the fit of a model to the observed time series. Unfortunately, posterior probability weights are not comparable for models that are estimated on a different number of time series. A second problem of the posterior probability weights is that over-parameterized models that have an extreme good in-sample fit, but a bad out-of-sample forecasting

accuracy, are assigned high weights. To circumvent these problems Gerard and Nimark (2008) use an out-of-sample weighting scheme based on predictive likelihoods as proposed by Eklund and Karlsson (2007) and Andersson and Karlsson (2007).

Predictive Likelihood (PL) The available data is split into a training sample used to estimate the models and a hold-out sample used to evaluate each model's forecasting performance. The forecasting performance is measured by the predictive likelihood, i.e. the marginal likelihood of the hold-out sample conditional on a specific model. I follow the approach suggested by Andersson and Karlsson (2007) and used by Gerard and Nimark (2008) to compute a series of small hold-out sample predictive likelihoods for each horizon. Equation (11) shows how to compute the predictive likelihood PL of model m for horizon h:

$$PL_h^m = ML(y_{holdout}^{obs}|y_{training}^{obs}) = \prod_{t=l}^{T-h} ML(y_{t+h}^{obs}|y_t^{obs}). \tag{11}$$

Starting with an initial trainings sample of length l, one computes the marginal likelihood for horizon h using the hold-out sample. The training sample is expanded by one observation to $l+1$ and a second maginal likelihood is computed for the hold-out sample that is one observation shorter than the previous one. This continues until the trainings sample has increased to lenght $T-h$ and the hold-out sample has shrinked to length h. To make the results comparable among models, only the three common variables output growth, inflation and the interest rate are considered for the computation of the predictive likelihood. Finally, the predictive likelihood weights are computed by replacing the marginal likelihood in equation (10) with the predictive likelihood:

$$\omega_{m,h} = \frac{PL_h^m}{\sum_{m=1}^{M} PL_h^m}. \tag{12}$$

The predictive likelihood weighting scheme allows for different weights to be assigned to a given model at different forecast horizons.

Ordinary Least Squares Weights (OLS) In model averaging applications of time series models it is common to assume a linear-in-weights model and estimate combination weights by ordinary least

squares (see Timmermann, 2006). I use the forecasts from previous vintages for each model and the accordant data realizations to regress the realizations y_{t+h}^{obs} on the forecasts $E[y_{t+h}^{obs}|I_t^m]$ from the different models via constrained OLS separately for each variable:

$$y_{t+h}^{obs} = \omega_{1,h} E[y_{t+h}^{obs}|I_t^1] + ... + \omega_{M,h} E[y_{t+h}^{obs}|I_t^M] + \varepsilon_{t+h}, \qquad s.t. \sum_{m=1}^{M} \omega_{m,h} = 1. \tag{13}$$

The resulting parameter estimates $\omega_{1,h}, ..., \omega_{M,h}$ are the combination weights. Therefore, the combination weights differ for different horizons and also for the three different variables. I omit an intercept term and restrict the weights to sum to one so that the weights can be interpreted as the fractions the specific models contribute to the weighted forecast. It also ensures that the combined forecast lies inside the range of the individual forecasts.

RMSE based weights (RMSE) There are several ways to compute simple relative performance weights. I consider here weightings based on RMSEs of past forecasts and weights based on the relative past forecasting accuracy by ranking the accuracy of the different models. For the prior case RMSE based weights can be computed by taking forecasts from previous vintages and compute the RMSE for each model. The weights are then calculated by taking the inverse relative RMSE performance:

$$\omega_{m,h} = \frac{(1/RMSE_h^m)}{\sum_{m=1}^{M}(1/RMSE_h^m)}. \tag{14}$$

Rank based weights (Rank) A second possibility to compute relative performance weights is to assign ranks R from 1 to M according to the past forecasting accuracy measured by the RMSEs. This method is similar to the RMSE based weights while being more robust to outliers. The performance rank based weights are computed as follows:

$$\omega_{m,h} = \frac{(1/R_h^m)}{\sum_{m=1}^{M}(1/R_h^m)}. \tag{15}$$

Both methods can assign different weights to forecasts of different variables and different forecasting horizons.

Mean Forecast (Mean) The simplest method to compute a weighted forecast is to give equal weight to each model and simply compute the mean forecast of all models. From a theoretical point of view this approach is not preferable as the weights are purely subjective prior weights implicitly given by the choice of models. However, it has often been found that simple weighting schemes perform well (see e.g. Hsiao and Wan, 2010). A reason is that they give weight to several models instead of choosing one optimal model and are thus robust.

Median Forecast (Median) Another possibility is to choose the median of different model forecasts. I compute the median forecast for each of the ordered draws of all models. This gives the density of the median forecast which is used to compute the mean of all these draws as a point forecast. The approach is similar to taking the mean forecast, but is more robust to outliers. The medians from the ordered forecast draws need not to come from the same model for different slices of the ordered forecast draws. By counting the fraction that the median forecast is generated by a specific model one can compute pseudo weights of the different model forecasts that show the contribution of each model to the final point forecast.

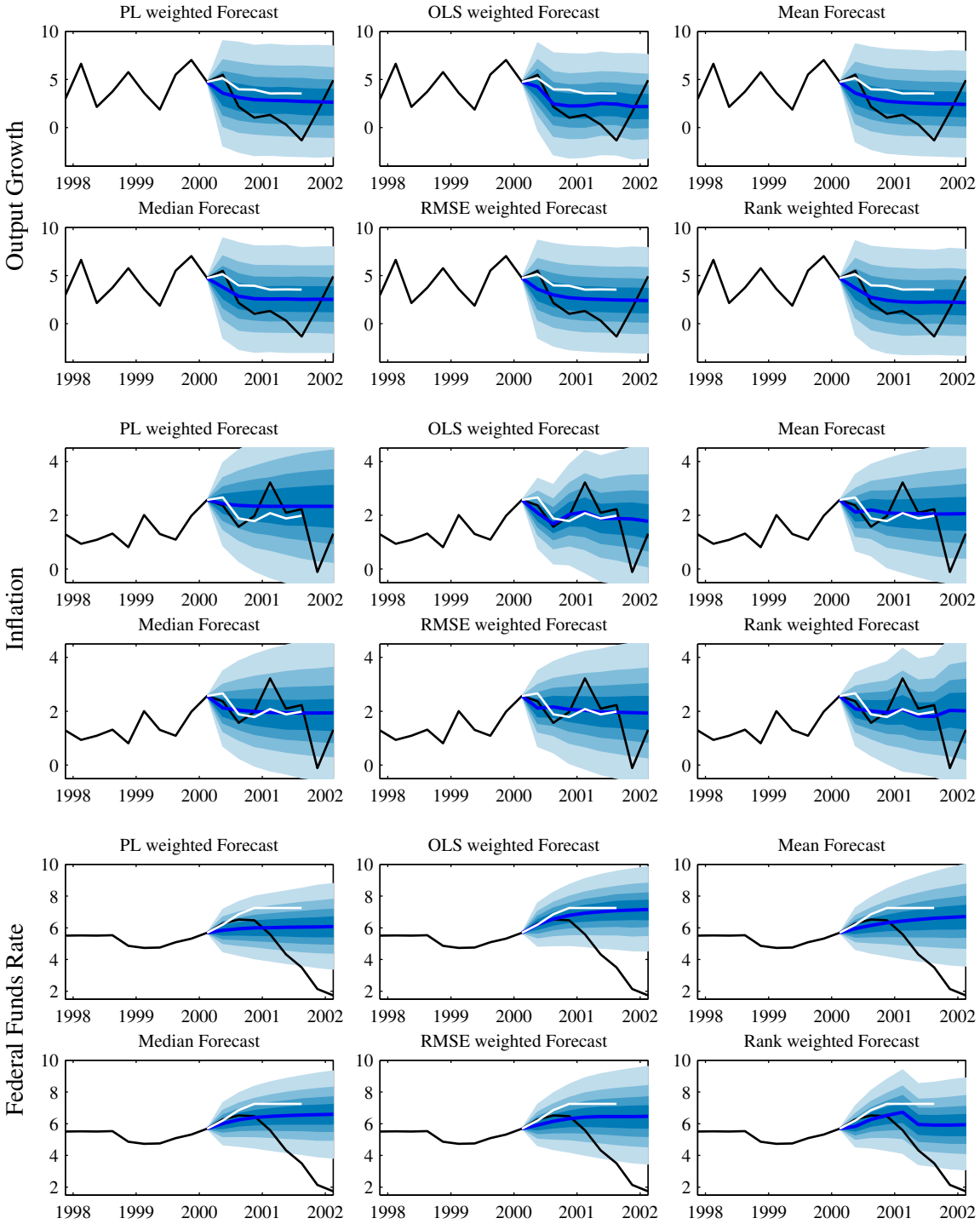
Figure 3 shows as an example weighted forecasts computed for the data vintage of May 12, 2000. In comparison with the individual forecasts in Figure 1 the forecasts are more robust as no outliers are visible. All methods predict a slightly lower output growth path than the Greenbook and a slight decrease of inflation in the current quarter. Afterwards inflation is predicted to remain about constant. For the interest rate forecasts all models predict an increase in the interest rate for the next three to four quarters. Afterwards the interest rate is predicted to remain at roughly six percent.

7 Forecast Evaluation of combined forecasts

In Table 4, I report RMSEs for output growth, inflation and interest rate forecasts from the Greenbook, and RMSEs of the six weighted forecasts relative to the Greenbook RMSE. The second last column shows for comparison the relative RMSEs of the best single model as reported in Table 2 and the last column shows the relative RMSEs of the best nonstructural model for each horizon as computed by Faust and Wright (2009).

For output growth, inflation and the federal funds rate, it is apparent that the weighted forecasts

Figure 3: Weighted Structural Forecasts; Data Vintage May 12, 2000



Notes: the black line shows real-time data until the forecast start and revised data afterwards; the shaded areas show 90% 70%, 50% and 30% confidence bands; the line in the middle of the confidence bands shows the mean forecast for horizons 0 to 7; the short white line shows the Greenbook forecast for horizons 0 to 5.

have in general an accuracy higher than forecasts from most single models. For output growth the Greenbook nowcast is more accurate than all other forecasts, but for all other horizons the weighted model forecasts dominate the Greenbook forecast. The PL weighting scheme is an exception with a forecasting quality not better, but still comparable to the Greenbook. There is not much of a difference between the accuracy of the other combination schemes. The Rank weighted forecast yields the most precise forecasts. Most methods give a similar forecasting accuracy in comparison to the best nonstructural forecasts and for medium forecasts even dominate those. The forecasting accuracy of the Mean and RMSE weighted forecasts is very similar because the weights computed by inverse RMSEs deviate only slightly from equal weights. Using inverse Ranks to compute weights, differentiates more between the different models' past forecasting performance. However, the increase in forecasting accuracy hardly justifies the increased computational efforts compared to the simple mean forecast. Taking the Greenbook nowcast as given does not translate into more accurate forecasts due to the low persistence of output growth data. For the one quarter ahead horizon weighted forecasts are as good as Greenbook projections, while for individual models the forecasting precision is somewhat worse. For horizons three and above most weighted forecasts even dominate RMSEs of a simple autoregressive forecast as reported in Faust and Wright (2009). In contrast, in the case of single model forecasts only the Smets & Wouters model is able to beat the autoregressive forecast. Most of the differences in output growth forecasting accuracy are statistically insignificant except for the nowcast.

For the inflation forecast, weighted forecasts increase the forecasting accuracy compared to most single model forecasts, especially for medium term horizons. However, the performance of the Greenbook forecasts is still the best. The weighting schemes can roughly be devided into two groups: the PL and OLS weighted forecasts are less precise than the Median, Mean, RMSE and Rank weighted forecasts. The simple Mean forecast is most accurate. For medium term horizons it is only slightly worse than the Greenbook forecast and the best nonstructural forecast. The difference is not statistically significant form horizon 3 onwards. Forecasting accuracy relative to the Greenbook increases with increasing horizons for all weighting schemes. This shows that structural forecasts are especially useful for medium term forecasts. A univariate autoregressive forecast is less precise than the weighted forecasts from horizon 2 onwards. Appending the Greenbook nowcast to the information set of the forecasting models increases the forecasting performance of all weighting methods and the

Table 4: Greenbook RMSE and relative RMSE of weighted model forecasts: 1984-2000

(a) Output growth

horizon	GB	PL	OLS	Median	Mean	RMSE	Rank	best M	best FW
					jump off -1				
0	1.52	1.41●	1.26●	1.26●	1.25●	1.24●	1.23●	1.32●	1.25
1	1.85	1.09	1.02	**0.99**	**0.99**	**0.98**	**0.99**	1.04	**0.99**
2	2.01	1.05	**0.99**	**0.92**	**0.92**	**0.92**	**0.91**	1.05	**0.95**
3	2.15	**0.99**	**0.89**	**0.90•**	**0.89•**	**0.89•**	**0.87●**	**0.88**	**0.94**
4	2.08	1.00	**0.90**	**0.90**	**0.90**	**0.89**	**0.86•**	**0.89**	**0.99**
5	2.08	1.02	**0.91**	**0.93**	**0.92**	**0.92**	**0.89**	**0.89**	**0.97**
					jump off 0				
1	1.85	1.08	1.00	**0.97**	**0.97**	**0.97**	**0.98**	1.07	**0.96**
2	2.01	1.06	**0.94**	**0.91•**	**0.91**	**0.91•**	**0.91•**	**0.97**	**0.90**
3	2.15	1.00	**0.92**	**0.89•**	**0.89•**	**0.89•**	**0.88●**	**0.88**	**0.95**
4	2.08	1.01	**0.92**	**0.91**	**0.90**	**0.89**	**0.89**	**0.88**	**0.96**
5	2.08	1.03	**0.95**	**0.93**	**0.92**	**0.92**	**0.94**	**0.89**	**0.98**

(b) Inflation

horizon	GB	PL	OLS	Median	Mean	RMSE	Rank	best M	best FW
					jump off -1				
0	0.70	1.51●	1.68●	1.43●	1.45●	1.45●	1.45●	1.49●	1.32●
1	0.80	1.57●	1.59●	1.47●	1.44●	1.45●	1.47●	1.41●	1.20●
2	0.82	1.35●	1.37●	1.25●	1.22●	1.22●	1.25●	1.26●	1.14•
3	0.94	1.16●	1.25●	1.10	1.06	1.07	1.09	1.16●	1.02
4	0.91	1.24●	1.30●	1.17	1.11	1.114	1.15	1.24•	1.06
5	1.15	1.22•	1.23	1.14	1.08	1.09	1.12	1.22•	**0.98**
					jump off 0				
1	0.80	1.21●	1.22●	1.12	1.15•	1.14•	1.15•	1.12	1.19●
2	0.82	1.25●	1.27●	1.19•	1.16•	1.17•	1.17•	1.16	1.17
3	0.94	1.24●	1.26●	1.13●	1.08	1.09	1.10	1.20●	1.03
4	0.91	1.16●	1.16•	1.06	1.03	1.04	1.09	1.15●	1.03
5	1.15	1.13●	1.12	1.07	1.03	1.04	1.06	1.11	**0.96**

(c) Federal Funds Rate

horizon	GB	PL	OLS	Median	Mean	RMSE	Rank	best M	best FW
					jump off -1				
0	0.11	6.38●	5.42●	4.43●	4.07●	3.92●	3.62●	4.99●	-
1	0.52	2.00●	2.12●	1.63●	1.42●	1.41●	1.37●	1.55●	-
2	0.94	1.42●	1.52●	1.22	1.12	1.11	1.10	1.31•	-
3	1.29	1.14	1.34●	1.02	**0.97**	**0.97**	**0.99**	1.07	-
4	1.64	1.03	1.25●	**0.96**	**0.95**	**0.94**	**0.95**	**0.96**	-
5	1.93	**0.96**	1.17●	**0.92**	**0.93**	**0.91**	**0.91**	**0.86**	-
					jump off 0				
1	0.52	1.28•	1.52●	1.08	1.00	1.00	1.05	1.11	-
2	0.94	1.13	1.45●	**0.99**	**0.91**	**0.91**	**0.95**	1.03	-
3	1.29	**0.99**	1.27•	**0.90**	**0.88**	**0.86**	**0.92**	**0.93**	-
4	1.64	**0.94**	1.22•	**0.89**	**0.90**	**0.87**	**0.91**	**0.89**	-
5	1.93	**0.89**	1.17•	**0.88**	**0.88**	**0.87**	**0.89**	**0.82**	-

Notes: PL: Predictive Likelihood; OLS: Ordinay Least Squares; Median: Median forecast; Mean: Mean forecast; RMSE: weighted by inverse RMSE; Rank: weighted by inverse ranks; best M: best single model forecast; Best FW: Best performing atheoretical model for the specific horizon considered by Faust & Wright; The first column shows the forecast horizon. The second column shows the RMSE for the Greenbook. The other columns show RMSE of alternative forecasts relative to the Greenbook. Values less than one are in bold and show that a forecast is more accurate than the one by the Greenbook. The symbols ●, ●, •, indicate that the relative RMSE is significantly different from one at the 1, 5, or 10% level, respectively. Significance levels are only an approximation for the results in the last column as they are based on another Greenbook dataset (see Appendix A, the correct significance level for the output growth nowcast is more likley to be 1%).

Mean forecast becomes as precise as the best nonstructural forecast. For the jump of 0 scenario all weighted forecasts are more accurate than a univariate autoregressive forecast.

The interest rate forecast results for individual models showed that the Bayesian VAR model performed better than all other models at least for short horizons. Combining the forecasts of all four DSGE models improves the forecasting quality: the Mean, RMSE and Rank weighted forecasts are as accurate as the forecasts from the Bayesian VAR. While the Greenbook interest rate path is more accurate for horizons 0 to 2, the Mean, RMSE and Rank weighted forecasts are more precise for horizons 3 to 5. The relative forecasting accuracy improves with increasing horizons for all weighting schemes. Taking the Greenbook nowcast as given, the accuracy of all weighting schemes increases due to the high persistence of the interest rate. The Mean forecast is as precise as the Greenbook policy path for horizon 1 and dominates it for all other horizons.

Overall it turns out that model combination methods that give weight to several models perform well. Likelihood based weighting methods are preferable in theory, but do not work as well in practice. Differences in predictive likelihoods of different models are so high that at most times all weight is given to a single model. Tables 3 to 5 in Appendix B report as an example model weights for forecasts derived from data vintage May 12, 2000. The forecasting performance of different models relative to each other varies over time (see table 3). Therefore, it is important to choose an average of several models to hedge against inaccurate forecasts of individual models. Combining several models gives a more robust forecast as it prevents against choosing an outlier that produces high forecast errors. Also estimated weights by least squares do not perform as good as simpler combination schemes: restricting the weights to sum to one leads to estimation problems so that in some cases weight is given only to one model. The Median forecast works quite well as it ensures that outliers are not chosen. The best forecasting performance is achieved by the Mean forecast and the RMSE and Rank based weighted forecasts. However, the RMSE weights deviate only slightly from the Mean forecast. The Rank weights take past forecasting performance more into account: this increases the accuracy of the output growth forecast, but does not improve on the Mean forecast for inflation and the interest rate. Therefore, at this stage, one can conclude that a simple Mean forecast is the preferable method. It is very easy to compute as one needs no forecasts and realization from earlier data vintages to calculate model weights and it yields precise forecasts that are quite robust to outliers. Using a simple

mean is also helpful for the structural interpretability of DSGE model forecasts. One can simply sum up the demand shock, supply shock and monetary policy shock contributions to the forecasts of the four different models and devide the overall contributions of these different shocks by the number of models. Table 2 in Appendix B shows the percentage of forecast periods in which the weighted forecasts are more accurate than the Greenbook projections. The results of this robust statistic are very similar to the RMSE results.

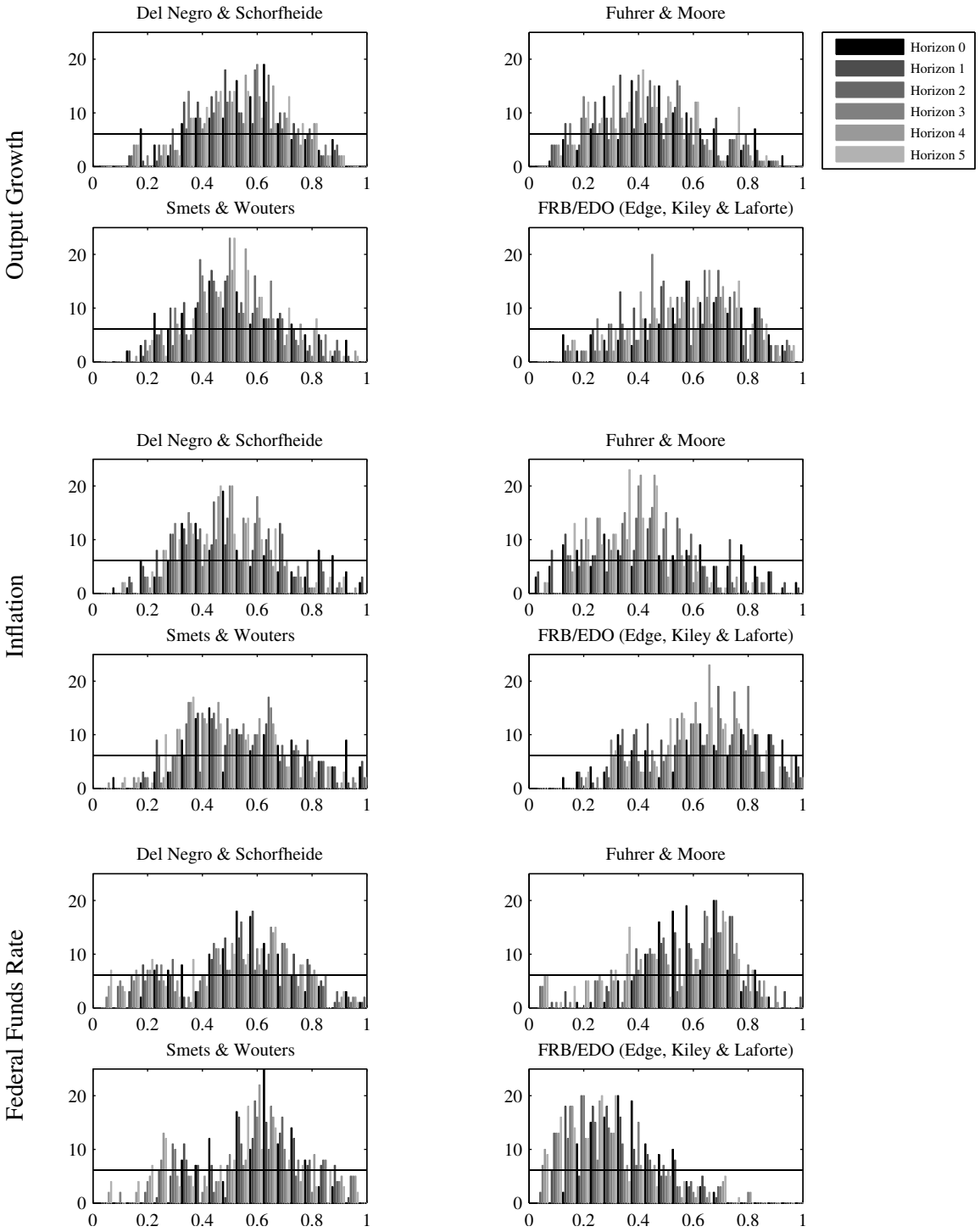
To sum up the point forecast evaluation, the forecasts of the Smets & Wouters model are good. The accuracy of forecasts that give considerable weight to several forecasts is as high as the Smets & Wouters forecast and in most cases even better. The accuracy of the Mean forecast is comparable to nonstructural forecasting methods that can process large datasets. All forecasts based on structural models are especially suited to compute medium term forecasts.

8 Density Forecast Evaluation

Assuming a symmetric loss function, the accuracy of point forecasts can be easily compared by computing RMSEs. Evaluating density forecasts is less straightforward. The true density is never observed. Still one can compare the distribution of observed data with density forecasts to check whether forecasts provide a realistic description of actual uncertainty.[21] I use the following evaluation procedure: I split up the density forecasts into probability bands that each cover 5% of the probability mass. This is similar to disaggregating the fan charts plotted in Figures 1 and 3 further into smaller confidence bands. For each data realization I can check into which of the 20 probability bands of the accordant density forecast it falls. Doing this for all realization and the corresponding density forecasts, 5% of the realizations should be contained in each of the probability bands. Otherwise the density forecasts are not a good characterization of the distribution of the data realizations. In general, if one divides density forecasts into probability bands of equal coverage, data realisations should be uniformly distributed across all probability bands. This is the approach outlined in Diebold et al. (1998) and Diebold et al. (1999). More formally, it is based on the relationship between the

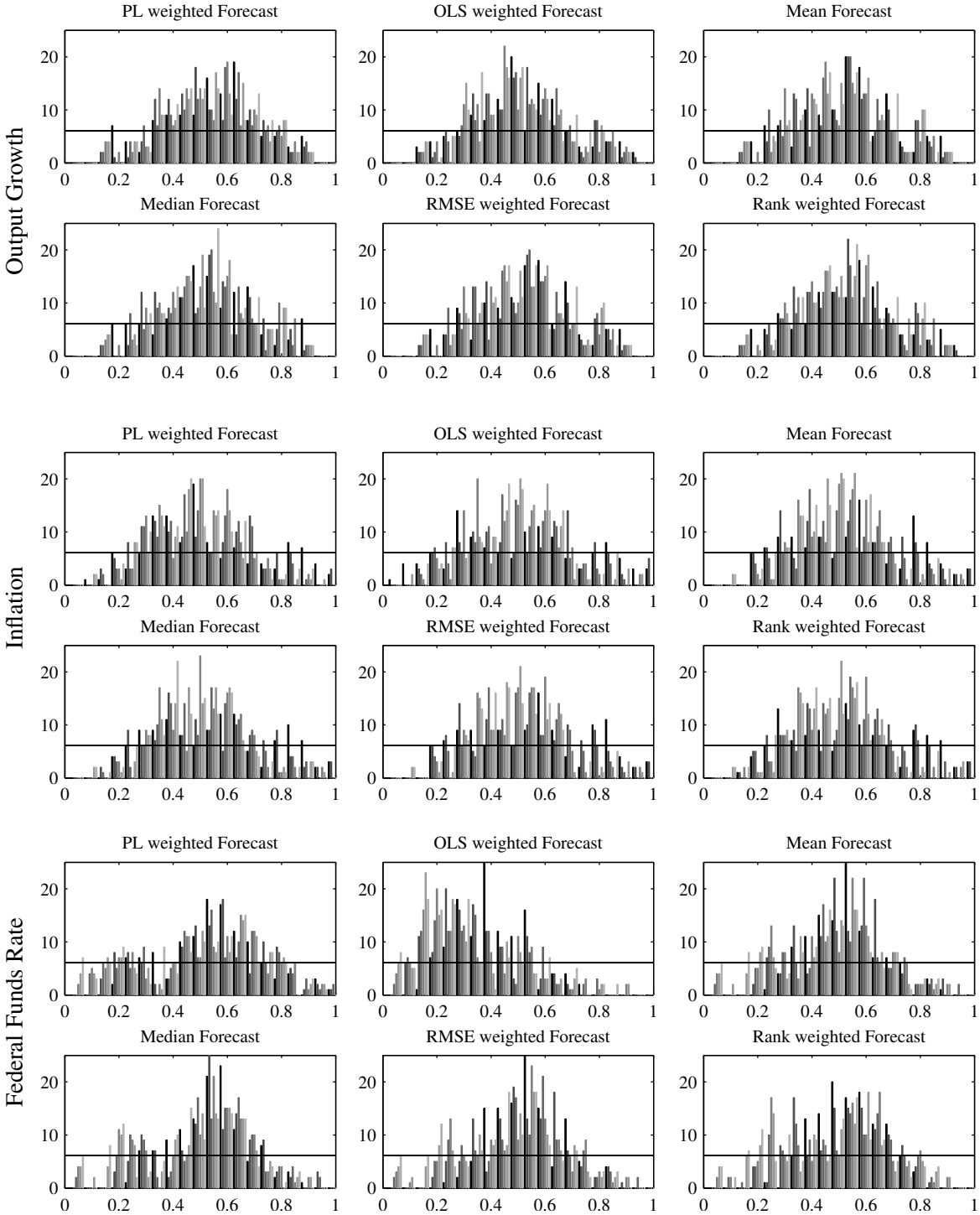
[21] While one can rank the accuracy of density forecasts from different models using the log score, I will check here whether DSGE models yield reasonable forecasts at all, i.e. I will look at the absolute rather tahn relative forecasting performance.

Figure 4: Evaluation of Structural Density Forecasts; 1984 - 2000



Notes: The figures show the distribution of realized data points on the density forecasts. The density forecasts are represented as probability bands each covering 5% of the density. The bars show how many of the realized observations fall in each of the probability bands. If the density forecast is an accurate description of actual uncertainty, than about six of the 123 observations should fall in each probability band.

Figure 5: Evaluation of Structural Density Forecasts; 1984 - 2000



Notes: The figures show the distribution of realized data points on the density forecasts. The density forecasts are represented as probability bands each covering 5% of the density. The bars show how many of the realized observations fall in each of the probability bands. If the density forecast is an accurate description of actual uncertainty, than about six of the 123 observations should fall in each probability band.

data generating process and the sequence of density forecasts via probability integral transforms of the observed data with respect to the density forecasts. The probability integral transform (PIT) is the cumulative density function corresponding to the sequence of n density forecasts $\{p_t(y_{t+h}^{obs})\}_{t=1}^{n}$ evaluated at the corresponding observed data points $\{y_{t+h}^{obs}\}_{t=1}^{n}$:

$$z_t = \int_{-\infty}^{y_{t+h}^{obs}} p_t(u)du, \qquad \text{for} \quad t = 1,...,n. \tag{16}$$

The PIT is the probability implied by the density forecast that a realized data point would be equal or less than what is actually observed. If the sequence of density forecasts is an accurate description of actual uncertainty, the sequence of PITs, $\{z_t\}_{t=1}^{n}$, should be distributed uniformly between zero and one. Figures 4 and 5 presents a visual assessment of the distribution of realized data points on the sequence of PITs that is represented as a histogram of 20 probability bands each covering 5%. There are $n = 123$ forecasts, so that there should be about 6 observations in each of the probability bands if the density forecasts are accurate. This is represented by the horizontal line. The bars shaded in different colors reflect PITs for the different forecasting horizons.

The peak in the middle of the histograms of the output growth forecasts shows that these overestimate actual uncertainty. The histograms for inflation are closer to a uniform distribution, especially for the inflation nowcast. There is only a slight peak in the middle of the distributions and the histogramms for some models cover the entire distribution including the tails. Higher horizon forecasts overestimate actual inflation uncertainty. The density forecasts are imprecise for the federal funds rate. The tails are not covered, especially for short horizons, and thus uncertainty is overestimated by the density forecasts. Gerard and Nimark (2008) give a plausible reason for the overestimation of actual uncertainty by DSGE models. The models impose tight restrictions on the data. If the data rejects these restrictions, large shocks are needed to fit the models to the data resulting in high shock uncertainty. As all individual model forecasts overestimate actual uncertainty it is not possible that the weighted forecasts yield a more realistic assesment of uncertainty. Therefore, the averaged density forecasts overestimate uncertainty as well.[22]

[22] In principle, there are tests available to formally check for a uniform distribution (Berkowitz, 2001). Unfortunately, the results have to be treated with high caution (see Elder et al., 2005; Gerard and Nimark, 2008). As the visual assessement has already shown clear evidence against a uniform distribution of the PITs, I do not use additional formal tests.

Herbst and Schorfheide (2011) also use PITs to evaluate density forecasts of a small New Keynesian model and the Smets & Wouters model. They find that for the Smets & Wouters model the density forecasts for output growth are too wide. The interest rate forecasts of both models are skewed as the federal funds rate was lower than predicted over the evaluation sample. The inflation forecasts of both models and the output forecasts of the small New Keynesian model perform well. The differences to my results might be due to differences in the sample and the evaluation data. Their estimation sample starts in 1984, while I use a rolling window of the most recent 80 quarterly observations. Their evaluation sample ranges from 1997 to 2004, while my evaluation sample starts in 1984 and ends in 2000. Furthermore, they deviate from Smets and Wouters (2007) by calibrating some parameters that have been estimated by Smets & Wouters. Herbst & Schorfheide use final revised data for the evaluation of the forecasts, while I use data that includes initial revisions only. At least in the first half of my estimation and evaluation samples I estimate the models on relatively volatile data, but evaluate the forecasts on data that belong to the great moderation. This puts the DSGE models to a very hard test and might partly explain the overestimation of actual uncertainty. The evaluation sample of Herbst and Schorfheide (2011) is relatively short and includes the high productivity growth period of the late 1990s and only one very mild recession in 2001. They estimate the models on data with low volatility and also evaluate the models on data of low volatility. The inclusion of the recent financial crisis could potentially lead to a reduction in the difference of predicted and actual uncertainty relative to my current results (at least at the left tail of the forecast distribution), but to an underestimation of actual uncertainty in the Herbst & Schorfheide case.[23] I leave it for future research to assess the sensitivity of DSGE models' density forecast performance to the estimation and evaluation sample.

9 Conclusion

During the last decade theory based DSGE models that are consistently derived from microeconomic optimization problems of households and firms have become the workhorse of modern monetary economics. Despite their stylized nature and their reliance on few equations they are widely used in

[23] Wieland and Wolters (2011) evaluate point forecasts of DSGE models for the recent financial crisis (and four other recessions) and find that the DSGE models are not able to forecast the intensity and lengths of such a deep recession. Thus, data realizations will be at left tail of density forecasts for output growth, inflation and the interest rate for 2008 and 2009.

academic work as well as at policy institutions. Computing out of sample forecasts is an ultimate test of the ability of this class of models to explain business cycles. In this paper, I have assessed the accuracy of point and density forecasts of four DSGE models using real-time data. While point forecasts are surprisingly precise, density forecasts have been shown to overestimate actual uncertainty. Point forecasts of some models are comparable to the forecasting accuracy of atheoretical forecasting methods that can process large datasets. Especially the model by Smets and Wouters (2007) yields relatively precise inflation, output growth and interest rate forecasts. However, this superior performance does not hold for a subsample-based evaluation. The forecasts and the forecasting accuracy of different DSGE models is quite distinct and different models perform best for different subsamples. This instability can be overcome by combining forecasts from several models. Weighted forecasts increase the forecasting accuracy. Combination methods that give significant weight to several models are preferable over methods that aim to identify a single best model. The accuracy of a simple mean of model forecasts is hard to beat by other forecast weighting methods. DSGE based forecasts perform particularly well for medium term forecasts in comparison with Greenbook projections and nonstructural forecasts. Structural forecasts perform quite well during normal times, but they are not able to detect large recessions and turning points due to their weak internal propagation meachanism. Large shocks are needed to fit the models to volatile periods of the sample. This is also the reason for their wide confidence bands that overestimate actual uncertainty.

References

Adolfson, M., Andersson, M. K., Lindé, J., Villani, M., Vredin, A., 2007a. Modern forecasting models in action: improving macroeconomic analyses at central banks. International Journal of Central Banking 3(4), 111–144.

Adolfson, M., Laseén, S., Lindé, J., Villani, M., 2007b. Bayesian estimation of an open economy DSGE model with incomplete pass-through. Journal of International Economics 72(2), 481–511.

An, S., Schorfheide, F., 2007. Bayesian analysis of DSGE models. Econometric Reviews 26(2-4), 113–172.

Andersson, M. K., Karlsson, S., 2007. Bayesian forecast combination for VAR models, Sveriges Riksbank Working Paper No 216.

Bache, I. W., Jore, A. S., Mitchell, J., Vahey, S. P., 2009. Combining VAR and DSGE forecast densities, norges Bank Working paper 2009/23.

Berkowitz, J., 2001. Testing density forecasts, with applications to risk management. Journal of Business and Economic Statistics 19(4), 465–474.

Bernanke, B. S., Boivin, J., 2003. Monetary policy in a data-rich environment. Journal of Monetary Economics 50(3), 525–546.

Christiano, L. J., Eichenbaum, M., Evans, C. L., 2005. Nominal rigidities and the dynamic effects of a shock to monetary policy. Journal of Political Economy 113(1), 1–45.

Christoffel, K., Coenen, G., Warne, A., 2008. The new area-wide model of the euro area - a micro-founded open-economy model for forecasting and policy analysis, European Central Bank Working Paper 944.

Christoffel, K., Coenen, G., Warne, A., 2011. Forecasting with DSGE models. In: Clements, M. P., Hendry, D. F. (Eds.), Oxford Handbook on Economic Forecasting. Oxford University Press, USA.

Del Negro, M., Schorfheide, F., 2004. Priors from general equilibrium models for VARS. International Economic Review 45(2), 643–673.

Del Negro, M., Schorfheide, F., Smets, F. R., Wouters, R., 2007. On the fit of new Keynesian models. Journal of Business and Economic Statistics 25, 123–143.

Diebold, F. X., Gunther, T. A., Tay, A. S., 1998. Evaluating density forecasts with applications to financial risk management. International Economic Review 39(4), 863–883.

Diebold, F. X., Hahn, J., Tay, A. S., 1999. Multivariate density forecast evaluation and calibration in financial risk management: High-frequency returns on foreign exchange. Review of Economics and Statistics 81(4), 661–673.

Diebold, F. X., Mariano, R. S., 1995. Comparing predictive accuracy. Journal of Business & Economic Statistics 13, 253–263.

Doan, T., Litterman, R., Sims, C., 1984. Forecasting and conditional projection using realistic prior distributions. Econometric Reviews 3, 1 – 100.

Edge, R., Gürkaynak, R., 2010. How useful are estimated DSGE model forecasts for central bankers? Brookings Papers on Economic Activity 2, 209–244.

Edge, R. M., Kiley, M. T., Laforte, J.-P., 2007. Documentation of the research and statistics divisions estimated DSGE model of the U.S. economy: 2006 version, finance and Economics Discussion Series, Divisions of Research & Statistics and Monetary Affairs, Federal Reserve Board, Washington, D.C.: 2007-53.

Edge, R. M., Kiley, M. T., Laforte, J.-P., 2008. Natural rate measures in an estimated DSGE model of the U.S. economy. Journal of Economic Dynamics and Control 32, 2512–2535.

Edge, R. M., Kiley, M. T., Laforte, J.-P., 2010. A comparison of forecast performance between federal reserve staff forecasts, simple reduced form models and a DSGE model. Journal of Applied Econometrics 25(4), 720–754.

Eklund, J., Karlsson, S., 2007. Forecast combination and model averaging using predictive measures. Econometric Reviews 26(2-4), 329–363.

Elder, R., Kapetanios, G., Taylor, T., Yates, T., 2005. Assessing the MPC's fan charts. Bank of England Quarterly Bulletin Autumn, 326–345.

Fair, R. C., 2007. Evaluating inflation targeting using a macroeconometric model. Economics: The Open-Access, Open-Assessment E-Journal 8.

Faust, J., Wright, J. H., 2009. Comparing Greenbook and reduced form forecasts using a large realtime dataset. Journal of Business and Economic Statistics 27(4), 468–479.

Forni, M., Hallin, M., Lippi, M., Reichlin, L., 2003. Do financial variables help forecasting inflation and real activity in the Euro area? Journal of Monetary Economics 50, 1243–1255.

Francis, N., Ramey, V. A., 2009. Measures of per capita hours and their implications for the technology-hours debate. Journal of Money, Credit and Banking 41(6), 1071–1097.

Fuhrer, J. C., 1997. Inflation/output variance trade-offs and optimal monetary policy. Journal of Money, Credit and Banking 29(2), 214–234.

Fuhrer, J. C., Moore, G., 1995a. Inflation persistence. The Quarterly Journal of Economics 110(1), 127–159.

Fuhrer, J. C., Moore, G., 1995b. Monetary policy trade-offs and the correlation between nominal interest rates and real output. The American Economic Review 85(1), 219–239.

Gerard, H., Nimark, K., 2008. Combing multivariate density forecasts using predictive criteria, Research Discussion Paper 2008-2, Reserve Bank of Australia.

Giannone, D., Monti, F., Reichlin, L., 2009. Incorporating conjunctural analysis in structural models. In: Wieland, V. (Ed.), The Science and Practice of Monetary Policy Today. Springer Science, pp. 41–57.

Goodfriend, M., King, R. G., 1997. The new neoclassical synthesis and the role of monetary policy. In: Bernanke, B. S., Rotemberg, J. J. (Eds.), National Bureau of Economic Research Macroeconomics Annual 1997. MIT Press, Cambridge, MA.

Hamilton, J. D., 1994. Time Series Analysis. Princeton University Press, Princeton, NJ.

Herbst, E., Schorfheide, F., 2011. Evaluating DSGE model forecasts of comovements, manuscript, University of Pennsylvania.

Hsiao, C., Wan, S. K., 2010. Is there an optimal forecast combination?, Working Paper University of Southern California.

Kimball, M., 1995. The quantitative analytics of the basic monetarist model. Journal of Money, Credit and Banking 27(4), 1241–1277.

Marcellino, M., Stock, J., Watson, M., 2003. Macroeconomic forecasting in the Euro area: Country-specific versus area-wide information. European Economic Review 47, 1–18.

Murchison, S., Rennison, A., 2006. ToTEM: The Bank of Canada's new quarterly projection model, Technical Reports 97, Bank of Canada.

Romer, C. D., Romer, D. H., 2000. Federal reserve information and the behavior of interest rates. American Economic Review 90, 429–457.

Rossi, B., 2012. Advances in forecasting under instability. In: Elliott, G., Timmermann, A. (Eds.), Handbook of Economic Forecasting. Vol. 2. Elsevier Publications.

Rotemberg, J. J., Woodford, M., 1997. An optimization-based econometric framework for the evaluation of monetary policy, in B. Bernanke and J. Rotemberg, (eds.), NBER Macroeconomics Annual, The MIT Press.

Schorfheide, F., 2000. Loss function-based evaluation of DSGE models. Journal of Applied Econometrics 15, 645–670.

Sims, C. A., 2002. The role of models and probabilities in the monetary policy process. Brookings Papers on Economic Activity 2, 1–40.

Smets, F., Christoffel, K., Coenen, G., Motto, R., Rostagno, M., 2010. DSGE models and their use at the ECB. Journal of the Spanish Economic Society 1(1), 51–65.

Smets, F., Wouters, R., 2004. Forecasting with a bayesian DSGE model: An application to the Euro area. Journal of Common Market Studies 42(4), 841–867.

Smets, F., Wouters, R., 2007. Shocks and frictions in US business cycles: A Bayesian DSGE approach. The American Economic Review 97(3), 586–606.

Stock, J., Watson, M., 2002. Forecasting using principal components from a large number of predictors. Journal of the American Statistical Association 97, 1167–1179.

Timmermann, A., 2006. Forecast combinations. In: Elliott, G., Granger, C. W. J., Timmermann, A. (Eds.), Handbook of Economic Forecasting. Amsterdam: North Holland, pp. 135–196.

Wang, M.-C., 2009. Comparing the DSGE model with the factor model: An out-of-sample forecasting experiment. Journal of Forecasting 28(2), 167–182.

Wieland, V., Wolters, M. H., 2011. The diversity of forecasts from macroeconomic models of the U.S. economy. Economic Theory 47(2-3), 247–292.

Wieland, V., Wolters, M. H., 2012. Forecasting and policy making. In: Elliott, G., Timmermann, A. (Eds.), Handbook of Economic Forecasting. Vol. 2. Elsevier Publications.

Woodford, M., 2003. Interest and Prices: Foundations of a Theory of Monetary Policy. Princeton University Press.

Appendix A: Greenbook forecasts from different data sources

I use the dataset by Faust and Wright (2009) from Jon Faust's website to compare the Greenbook forecasts reported in this dataset to those reported in a dataset that is available from the Federal Reserve Bank of Philadelphia.[24] Using the same definition of revised data as Faust and Wright (2009) I can replicate the Greenbook RMSEs in the paper exactly.[25] However, when plotting forecast errors, I found some surprisingly large values. For the Greenbook from May 15, 1997, GDP growth has a forecast error (actual minus predicted) of -9.58% for the nowcast and of 11.59% for one quarter ahead. The corresponding annualized GDP growth forecasts are 12.84% for the nowcast and -8.56% for one quarter ahead. Revised data shows entries of 3.25 and 3.03, respectively.

Output growth forecasts

I have compared Greenbook output growth forecasts from the dataset by Faust and Wright (2009) with those that are available from the Federal Reserve Bank of Philadelphia. I found some discrepancy between these series, likely due to the fact that Philadelphia Fed figures are rounded to the first decimal place while the Faust and Wright (2009) series comprise higher-digit precision. Some of the remaining discrepancies might also be attributed to the fact that Faust and Wright (2009) compute approximate growth rates as $400\log(x_t/x_{t-1})$, rather than computing accurate growth rates as $100[(x_t/x_{t-1})^4-1]$. However, some of the discrepancies cannot be explained by mere rounding errors and it remains unclear from where the differences arise. Table 5 lists all discrepancies from 1984-2000, i.e. the sample that is used for the main results of the paper, that exceed 0.2%.

To check how these differences affect the assessment of the forecasting accuracy, I compute RMSEs for the Philadelphia Fed dataset and compare them to the results by Faust and Wright (2009). Table 6 shows that the dataset by Faust and Wright (2009) yields a lower forecasting accuracy for

[24] The dataset used by Faust and Wright (2009) is available on http://e105.org/faustj/download/faustWrightGBTSdata.zip?d=n. Greenbook forecasts are also available from the Federal Reserve Bank of Philadelphia in two datasets: forecasts that are closest to the middle of a quarter are available in an Excel sheet and all other forecasts are available as .pdf documents that contain scanned original Greenbook documents. The URLs are: http://www.philadelphiafed.org/research-and-data/real-time-center/greenbook-data/philadelphia-data-set.cfm and http://www.philadelphiafed.org/research-and-data/real-time-center/greenbook-data/pdf-data-set.cfm.

[25] They use the data vintage that was released two quarters after the quarter to which the data refer to evaluate the forecasting accuracy. For example, revised data for 1999Q1 is obtained by selecting the entry for 1999Q1 from the data vintage released in 1999Q3. They use the Federal Reserve Bank of Philadelphia's real-time dataset. It is available on: http://www.philadelphiafed.org/research-and-data/real-time-center/real-time-data.

Table 5: Greenbook output growth forecasts from two data sources

date	Faust & Wright	Philadelphia Fed	difference
	Horizon: 0		
19840321	7.6591	8.0	-0.3409
19970515	12.8371	1.8	11.0371
19990128	3.6747	2.7	0.9747
	Horizon: 1		
19840321	5.7972	6.0	-0.2028
19881207	5.5942	5.8	-0.2058
19970515	-8.5586	2.5	-11.0586
	Horizon: 5		
19910925	7.1815	3.1	4.0815

Notes: the table shows annualized output growth forecasts from the Greenbook as reported in the dataset used by Faust and Wright (2009) and the Greenbook forecasts available from the Federal Reserve Bank of Philadelphia. The table only reports discrepancies of 0.2% or higher.

horizons 0 and 1 than the dataset by the Philadelphia Fed.

To assess whether these differences might change the interpretation of the main results in Faust and Wright (2009), table 7 shows results for horizons 0 and 1 from table 2 in Faust and Wright (2009) when using the Philadelphia Fed's Greenbook forecasts (Phil Fed) compared to using the dataset from Faust and Wright (F&W). The table shows the RMSEs of the Greenbook in column 2. The other columns report RMSEs of different time series model forecasts relative to the Greenbook RMSE. Values higher than 1 show that forecasts are on average less accurate than the Greenbook projections and values lower than 1 (printed in bold) show that forecasts are on average more accurate than the Greenbook projections. To compute the values for the Philadelphia Fed dataset one thus

Table 6: Greenbook output growth RMSEs based on two different data sources

horizon	Faust & Wright	Philadelphia Fed
0	1.75	1.52
1	2.12	1.85
2	2.01	2.01
3	2.15	2.15
4	2.08	2.08
5	2.08	2.08

Notes: The RMSEs are computed based on all available forecasts in the dataset from Faust and Wright (2009) from 1984-2000 and the corresponding data from the Federal Reserve Bank of Philadelphia.

only needs to multiply for each forecasting model the relative RMSE from Faust and Wright (2009) with the Greenbook RMSE to get the absolute RMSE. Afterwards I divide the absolute RMSE with the Greenbook RMSE from the Philadelphia Fed dataset to get the new relative RMSE. Of course, I cannot say anything about the statistical significance in the difference of the forecasting accuracy of models relative to the Greenbook. Computing relative RMSEs for the DSGE models in the paper for the Faust & Wright Greenbook dataset yields no significant difference of model forecasts to Greenbook projections for horizon 0. However, using the Philadelphia Fed Greenbook projections leads to a significant difference of DSGE model nowcasts and Greenbook nowcasts on the 1% level. For the one quarter ahead forecasts the difference is insignificant for both datasets.

Table 7: Greenbook RMSE and relative RMSE of time series models: 1984-2000

Output growth									
horizon	GB	RAR	DAR	EWA	BMA	FAA	FAV	DF	FVS
				jump off -1					
0 (F&W)	1.75	1.09	1.09	1.09	1.10	1.30	1.39	1.32	1.24
0 (Phil Fed)	1.52	1.25	1.25	1.25	1.27	1.50	1.60	1.52	1.43
1 (F&W)	2.12	**0.87**	**0.86**	**0.87**	**0.93**	1.13	1.20	1.15	0.93
1 (Phil Fed)	1.85	1.00	**0.99**	1.00	1.07	1.29	1.38	1.32	1.07
				jump off 0					
1 (F&W)	2.12	**0.84**	**0.84**	**0.84**	**0.85**	**0.96**	1.07	**0.99**	**0.85**
1 (Phil Fed)	1.85	**0.96**	**0.96**	**0.96**	**0.97**	1.10	1.23	1.13	**0.97**

Notes: GB: Greenbook; RAR: recursive autoregression; DAR: direct forecast from autoregression; EWA: equal-weighted averagin; BMA: Bayesian model averaging; FAA: factor augmented autoregression; FAV: factor augmented vector autoregression; DF: dynamic factor model; FVS: factor-spanned variable selection. A detailed description of the different models is contained in Faust and Wright (2009).

Faust and Wright (2009) stress that the Greenbook nowcast is extremely accurate compared to other methods due to the fact that the Fed makes great efforts to mirrorring key elements of the data construction process of the BEA. Using the Philadelphia Fed's dataset even increases the accuracy of the Greenbook nowcast to a RMSE of 1.52 compared to 1.75 in Faust and Wright (2009). They use a forecast from a simple autoregressive process (RAR and DAR) as a benchmark and find that this is at least as accurate as the other considered time series models. The results in table 7 show that the relative RMSE of the autoregressive process is only 1.25 when using the Philadelphia Fed dataset compared to 1.09 when using the Faust & Wright dataset. This stresses once more the extremely high accuracy of the Greenbook nowcasts.

For the one quarter ahead output growth forecasts the results by Faust and Wright (2009) give

the impression that many simple time series methods yield a better forecast than the Greenbook even though the difference is statistically not significant. Using the Philadelphia Fed dataset makes clear that the best forecasts from time series models are as good as the Greenbook forecasts, but not better. The best relative RMSE is 0.99 for the direct autoregressive forecast. In contrast, using the Faust and Wright (2009) dataset one gets a relative RMSE of 0.86 that might yield the impression that the Greenbook forecast is somewhat worse than the autoregressive forecast.

One of the main results of Faust and Wright (2009) is that conditional on an accurate nowcast no forecast improves upon a simple autoregressive forecast. Looking at the "jump off 0" results in table 7 shows that a better description of this main result is that no method is able to improve upon the Greenbook or the autoregressive forecast that are equally accurate. While Faust and Wright (2009) find a relative RMSE of 0.84 for the autoregressive forecast, with the Philadelphia Fed dataset I find it to be 0.96 which shows again that the accuracy of the autoregressive benchmark is in line with the Greenbook forecast, but not better.

To sum up, the main findings of the paper regarding the output growth forecasts are not affected. However, the dataset by Faust and Wright (2009) understates the Greenbook's forecasting accuracy compared to the autoregressive forecast to some extent.

Inflation forecasts

While I did not find any extremely large forecasting errors for inflation, the inflation forecasts from the two dataset differ from each other. Table 8 shows discrepancies that exceed 0.2% for the sample 1984-2000.[26]

Most of the larger discrepancies originate from Greenbook forecasts from 1996. In 1996 the Fed changed from using the MPS model to the FRB/US model to compute forecasts. However, it is unclear whether this change is a reason for some of the larger discrepancies between the two datasources. As there are no extreme discrepancies as in the case of output growth, all the inflation forecast results in Faust and Wright (2009) are robust to using the dataset from the Federal Reserve Bank of Phildelphia. Table 9 shows RMSE from the two datasets. The differences are small.

[26]The difference of the inflation forecasts between 1980 and 1984 exceeds 0.2% in most cases. The inflation forecasts from the Faust& Wright dataset are for these observations with only few exceptions lower than the forecasts from the Philadelphia Fed dataset. However, these observations are not used for the main results of Faust and Wright (2009).

Table 8: Greenbook inflation forecasts from two data sources

date	Faust & Wright	Philadelphia Fed	difference
	Horizon: 0		
19960626	2.0783	1.6	0.4783
19960918	2.0783	1.7	0.3783
19961212	2.4693	2.2	0.2693
	Horizon: 1		
19911030	3.0529	3.7	-0.6471
19960626	2.5668	2.2	0.3668
19960918	2.5668	2.3	0.2668
	Horizon: 2		
19960626	2.6642	2.4	0.2642
	Horizon: 3		
19860813	2.5668	2.8	-0.2332
19960626	2.7615	2.3	0.4615
19961212	2.2739	2.0	0.2739
	Horizon: 4		
19960626	2.7615	2.3	0.4615
19960918	2.5668	2.3	0.2668
19961212	2.2739	2.0	0.2739
	Horizon: 5		
19960626	2.7615	2.3	0.4615
19960918	2.6642	2.4	0.2642
19961212	2.5668	2.3	0.2668

Notes: the table shows annualized inflation forecasts from the Greenbook as reported in the dataset used by Faust and Wright (2009) and the Greenbook forecasts available from the Federal Reserve Bank of Philadelphia. The table only reports discrepancies of 0.2% or higher.

Table 9: Greenbook inflation RMSEs based on two different data sources

horizon	Faust & Wright	Philadelphia Fed
0	0.69	0.70
1	0.79	0.80
2	0.81	0.82
3	0.93	0.94
4	0.89	0.91
5	1.14	1.15

Notes: The RMSEs are computed based on all available forecasts in the dataset from Faust and Wright (2009) from 1984-2000 and the corresponding data from the Federal Reserve Bank of Philadelphia.

Interest rate path

Finally, I checked whether the assumed future interest rate path differs between the two datasets. The Greenbook projections are conditioned on this hypothetical path of the federal funds rate. I find some differences between the datasets, but these are for the sample from 1984 to 2000 equal or smaller than 25 basis points for the different forecast horizons. It is unlikely that these differences can explain the differences in the inflation and output growth forecasts. Table 10 shows the RMSE of the assumed interest rate paths from the two data sources. The differences are tiny.

Table 10: Federal funds rate RMSEs based on two different data sources

horizon	Faust & Wright	Philadelphia Fed
0	0.11	0.11
1	0.52	0.52
2	0.94	0.94
3	1.30	1.29
4	1.64	1.64
5	1.93	1.93

Notes: The RMSEs are computed based on the hypothetical future federal fudns rate paths available in the dataset from Faust and Wright (2009) from 1984-2000 and the corresponding data from the Federal Reserve Bank of Philadelphia.

Appendix B: Additional Results

Table 11: Percentage of periods alternative forecast better than Greenbook: 1984-2000

(a) Output growth

horizon	DS	FM	SW	EDO	BVAR	best FW	worst FW
				jump off -1			
0	30	36	36	34	38	43	30
1	50	47	47	48	49	**60**	39
2	47	46	**53**	50	**54**	**58**	37
3	46	43	**59**	**51**	**52**	**57**	42
4	43	45	**52**	46	48	**54**	36
5	42	43	**59**	48	43	**52**	43
				jump off 0			
1	43	48	48	48	50	**59**	40
2	50	50	**58**	42	**55**	**55**	41
3	47	48	**57**	49	**52**	**57**	38
4	43	46	**54**	42	**51**	**57**	39
5	44	43	**54**	43	46	49	43

(b) Inflation

horizon	DS	FM	SW	EDO	BVAR	best FW	worst FW
				jump off -1			
0	43	31	43	30	37	37	25
1	30	30	42	38	34	40	21
2	41	35	37	35	37	38	25
3	46	38	37	30	40	39	17
4	43	30	36	31	34	43	11
5	34	30	38	34	33	46	16
				jump off 0			
1	37	31	36	44	37	41	30
2	37	33	39	46	37	40	21
3	41	42	39	39	46	43	20
4	39	25	32	39	43	43	18
5	39	30	33	**53**	34	48	15

(c) Federal Funds Rate

horizon	DS	FM	SW	EDO	BVAR	best FW	worst FW
				jump off -1			
0	8	12	7	5	11	-	-
1	29	27	21	11	25	-	-
2	44	32	32	18	39	-	-
3	49	34	40	24	44	-	-
4	**56**	31	46	30	49	-	-
5	**59**	34	50	29	**55**	-	-
				jump off 0			
1	33	30	30	24	39	-	-
2	42	36	40	27	49	-	-
3	48	41	48	26	**55**	-	-
4	48	39	**53**	28	**58**	-	-
5	**52**	42	**54**	25	**60**	-	-

Notes: The first column shows the forecast horizon. The other columns show the percentage of forecast periods in which forecast errors of specific models are smaller in absolute value than the Greenbook forecast error. Entries greater than 50 percent indicate that the alternative forecast is better more than half the time and are in bold. The results for bestFW and worstFW are only indicative (and probably too high for output growth horizons 0 and 1) as they are based on another Greenbook dataset (see Appendix A).

Table 12: Percentage of periods weighted forecast better than Greenbook: 1984-2000

(a) Output growth

horizon	PL	OLS	Median	Mean	RMSE	Rank	best M	best FW
				jump off -1				
0	30	45	37	37	36	39	36	43
1	50	**57**	**56**	**58**	**57**	**56**	50	**60**
2	47	**57**	**58**	**61**	**59**	**57**	**53**	**58**
3	46	**55**	**57**	**58**	**57**	**57**	**59**	**57**
4	43	**52**	**56**	**52**	**51**	**60**	**52**	**54**
5	42	50	**51**	**53**	**52**	**54**	**59**	**52**
				jump off 0				
1	43	**57**	**59**	**59**	**60**	**53**	48	**59**
2	50	**56**	**64**	**63**	**59**	**56**	**58**	**55**
3	47	**54**	**55**	**57**	**57**	**56**	**57**	**57**
4	46	**51**	**52**	**51**	49	**55**	**54**	**57**
5	44	48	**55**	**60**	**60**	49	**54**	49

(b) Inflation

horizon	PL	OLS	Median	Mean	RMSE	Rank	best M	best FW
				jump off -1				
0	43	34	42	37	37	39	44	37
1	30	30	35	35	34	37	42	40
2	41	39	44	45	44	43	41	43
3	46	43	49	46	48	47	46	44
4	43	43	44	49	48	46	43	43
5	34	33	41	43	42	41	37	46
				jump off 0				
1	37	39	41	39	38	38	44	41
2	37	38	43	42	42	43	46	40
3	41	36	45	46	49	47	42	43
4	39	44	41	46	43	43	39	43
5	39	46	40	39	41	43	**54**	48

(c) Federal Funds Rate

horizon	PL	OLS	Median	Mean	RMSE	Rank	best M	best FW
				jump off -1				
0	8	7	11	11	11	11	12	-
1	29	14	32	30	28	25	29	-
2	44	30	39	40	39	35	44	-
3	49	34	**51**	**51**	47	43	49	-
4	**56**	33	**52**	**52**	**52**	50	**55**	-
5	**59**	35	**59**	**57**	**59**	**58**	**59**	-
				jump off 0				
1	33	25	38	38	42	38	33	-
2	42	30	43	44	45	40	41	-
3	48	39	**52**	**55**	**53**	44	47	-
4	48	35	**55**	**54**	**56**	50	**53**	-
5	**52**	32	**57**	**59**	**64**	**51**	**54**	-

Notes: PL: Predictive Likelihood; OLS: Ordinay Least Squares; Median: Median forecast; Mean: Mean forecast; RMSE: weighted by inverse RMSE; Rank: weighted by inverse ranks; best M: best single model forecast; Best FW: Best performing atheoretical model for the specific horizon considered by Faust & Wright; The first column shows the forecast horizon. The other columns show the percentage of forecast periods in which forecast errors of specific models are smaller in absolute value than the Greenbook forecast error. Entries greater than 50 percent indicate that the alternative forecast is better more than half the time and are in bold. The results for bestFW are only indicative (and probably too high for output growth horizons 0 and 1) as they are based on another Greenbook dataset (see Appendix A).

Table 13: Combination weights for data vintage May 12, 2000: output growth

model	PL	OLS	Median	Mean	RMSE	Rank
			horizon 0			
DS	1.00	0.00	0.50	0.25	0.24	0.12
FM	0.00	1.00	0.21	0.25	0.26	0.48
SW	0.00	0.00	0.29	0.25	0.24	0.16
EDO	0.00	0.00	0.00	0.25	0.26	0.24
			horizon 1			
DS	1.00	0.00	0.48	0.25	0.24	0.16
FM	0.00	0.00	0.02	0.25	0.23	0.12
SW	0.00	0.50	0.50	0.25	0.27	0.24
EDO	0.00	0.50	0.00	0.25	0.26	0.48
			horizon 2			
DS	1.00	0.00	0.00	0.25	0.24	0.16
FM	0.00	0.06	0.50	0.25	0.23	0.12
SW	0.00	0.44	0.49	0.25	0.27	0.24
EDO	0.00	0.50	0.01	0.25	0.26	0.48
			horizon 3			
DS	1.00	0.00	0.00	0.25	0.24	0.16
FM	0.00	0.12	0.50	0.25	0.23	0.12
SW	0.00	0.36	0.47	0.25	0.26	0.48
EDO	0.00	0.52	0.03	0.25	0.27	0.24
			horizon 4			
DS	1.00	0.00	0.00	0.25	0.24	0.12
FM	0.00	0.29	0.50	0.25	0.24	0.16
SW	0.00	0.22	0.45	0.25	0.26	0.48
EDO	0.00	0.49	0.05	0.25	0.26	0.24
			horizon 5			
DS	1.00	0.00	0.00	0.25	0.23	0.12
FM	0.00	0.41	0.50	0.25	0.26	0.24
SW	0.00	0.34	0.38	0.25	0.27	0.48
EDO	0.00	0.25	0.12	0.25	0.24	0.16

Notes: PL: Predictive Likelihood; OLS: Ordinay Least Squares; Median: Median forecast; Mean: Mean forecast; RMSE: weighted by inverse RMSE; Rank: weighted by inverse ranks; DS: Del Negro & Schorfheide; FM: Fuhrer & Moore; SW: Smets & Wouters; EDO: FRB/EDO Model by Edge, Kiley & Laforte; The first column shows the model name and the rows show the weight of each model for the different combination schemes. For each horizon, the four model weights sum up to 1.

Table 14: Combination weights for data vintage May 12, 2000: inflation

model	PL	OLS	Median	Mean	RMSE	Rank
			horizon 0			
DS	1.00	0.21	0.01	0.25	0.28	0.24
FM	0.00	0.05	0.50	0.25	0.20	0.12
SW	0.00	0.74	0.49	0.25	0.29	0.48
EDO	0.00	0.00	0.00	0.25	0.23	0.16
			horizon 1			
DS	1.00	0.00	0.29	0.25	0.26	0.24
FM	0.00	0.00	0.21	0.25	0.22	0.12
SW	0.00	0.88	0.50	0.25	0.28	0.48
EDO	0.00	0.12	0.00	0.25	0.24	0.16
			horizon 2			
DS	1.00	0.00	0.26	0.25	0.26	0.24
FM	0.00	0.35	0.24	0.25	0.24	0.16
SW	0.00	0.40	0.50	0.25	0.28	0.48
EDO	0.00	0.25	0.00	0.25	0.22	0.12
			horizon 3			
DS	1.00	0.16	0.29	0.25	0.29	0.48
FM	0.00	0.34	0.21	0.25	0.21	0.12
SW	0.00	0.09	0.50	0.25	0.28	0.24
EDO	0.00	0.41	0.00	0.25	0.22	0.16
			horizon 4			
DS	1.00	0.00	0.30	0.25	0.28	0.24
FM	0.00	0.30	0.20	0.25	0.19	0.12
SW	0.00	0.16	0.50	0.25	0.28	0.48
EDO	0.00	0.54	0.00	0.25	0.25	0.16
			horizon 5			
DS	1.00	0.00	0.33	0.25	0.27	0.24
FM	0.00	0.32	0.17	0.25	0.20	0.12
SW	0.00	0.16	0.50	0.25	0.28	0.48
EDO	0.00	0.52	0.00	0.25	0.25	0.16

Notes: PL: Predictive Likelihood; OLS: Ordinay Least Squares; Median: Median forecast; Mean: Mean forecast; RMSE: weighted by inverse RMSE; Rank: weighted by inverse ranks; DS: Del Negro & Schorfheide; FM: Fuhrer & Moore; SW: Smets & Wouters; EDO: FRB/EDO Model by Edge, Kiley & Laforte; The first column shows the model name and the rows show the weight of each model for the different combination schemes. For each horizon, the four model weights sum up to 1.

Table 15: Combination weights for data vintage May 12, 2000: Federal Funds Rate

model	PL	OLS	Median	Mean	RMSE	Rank
			horizon 0			
DS	1.00	0.00	0.44	0.25	0.24	0.16
FM	0.00	0.00	0.50	0.25	0.28	0.24
SW	0.00	0.00	0.06	0.25	0.30	0.48
EDO	0.00	1.00	0.00	0.25	0.18	0.12
			horizon 1			
DS	1.00	0.00	0.02	0.25	0.24	0.16
FM	0.00	0.00	0.50	0.25	0.31	0.48
SW	0.00	0.00	0.48	0.25	0.27	0.24
EDO	0.00	1.00	0.00	0.25	0.18	0.12
			horizon 2			
DS	1.00	0.00	0.03	0.25	0.25	0.16
FM	0.00	0.00	0.50	0.25	0.29	0.48
SW	0.00	0.00	0.47	0.25	0.26	0.24
EDO	0.00	1.00	0.00	0.25	0.20	0.12
			horizon 3			
DS	1.00	0.00	0.03	0.25	0.26	0.16
FM	0.00	0.00	0.50	0.25	0.27	0.48
SW	0.00	0.00	0.47	0.25	0.27	0.24
EDO	0.00	1.00	0.00	0.25	0.20	0.12
			horizon 4			
DS	1.00	0.00	0.04	0.25	0.27	0.24
FM	0.00	0.00	0.50	0.25	0.25	0.16
SW	0.00	0.00	0.46	0.25	0.27	0.48
EDO	0.00	1.00	0.00	0.25	0.21	0.12
			horizon 5			
DS	1.00	0.00	0.04	0.25	0.27	0.24
FM	0.00	0.00	0.50	0.25	0.22	0.12
SW	0.00	0.00	0.46	0.25	0.29	0.48
EDO	0.00	1.00	0.00	0.25	0.22	0.16

Notes: PL: Predictive Likelihood; OLS: Ordinay Least Squares; Median: Median forecast; Mean: Mean forecast; RMSE: weighted by inverse RMSE; Rank: weighted by inverse ranks; DS: Del Negro & Schorfheide; FM: Fuhrer & Moore; SW: Smets & Wouters; EDO: FRB/EDO Model by Edge, Kiley & Laforte; The first column shows the model name and the rows show the weight of each model for the different combination schemes. For each horizon, the four model weights sum up to 1.