Wolters, Maik H.

Conference Paper

Forecasting under Model Uncertainty

Beiträge zur Jahrestagung des Vereins für Socialpolitik 2011: Die Ordnung der Weltwirtschaft: Lektionen aus der Krise - Session: Forecasting Methods, No. G17-V2

Provided in Cooperation with:
Verein für Socialpolitik / German Economic Association

Suggested Citation: Wolters, Maik H. (2011) : Forecasting under Model Uncertainty, Beiträge zur Jahrestagung des Vereins für Socialpolitik 2011: Die Ordnung der Weltwirtschaft: Lektionen aus der Krise - Session: Forecasting Methods, No. G17-V2, ZBW - Deutsche Zentralbibliothek für Wirtschaftswissenschaften, Leibniz-Informationszentrum Wirtschaft

This Version is available at:
https://hdl.handle.net/10419/48723

Forecasting under Model Uncertainty

Maik H. Wolters *

Goethe University Frankfurt

February 4, 2011

Abstract

This paper investigates the accuracy of point and density forecasts of four dynamic stochastic general equilibrium (DSGE) models for output growth, inflation and the interest rate. The model parameters are estimated and forecasts are derived successively from historical U.S. data vintages synchronized with the Fed's Greenbook projections. In addition, I compute weighted forecasts using simple combination schemes as well as likelihood based methods. While forecasts from structural models fail to forecast large recessions and booms, they are quite accurate during normal times. Model forecasts compare particularly well to nonstructural forecasts and to Greenbook projections for horizons of three quarters ahead and higher. Weighted forecasts are more precise than forecasts from single models. A simple average of forecasts yields an accuracy comparable to the one obtained with state of the art time series methods that can incorporate large datasets. Comparing density forecasts of DSGE models with the actual distribution of observations shows that the models overestimate uncertainty around point forecasts.

Keywords: DSGE, forecasting, model uncertainty, density forecasts, business cycle models

JEL-Codes: C53, E31, E32, E37

*Mailing address: Grüneburgplatz 1, 60323 Frankfurt, Germany; wolters@wiwi.uni-frankfurt.de

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 (Smets and Wouters, 2004; Adolfson, Andersson, Linde, Villani, and Vredin, 2007; Smets and Wouters, 2007; Edge, Kiley, and Laforte, 2010; Wang, 2009; Christoffel, Coenen, and Warne, 2010). While these studies analyse only one model at a time, Wieland and Wolters (2010) compute forecasts from several theory based models for the five most recent U.S. recessions. The advantage of using structural models is that an economically meaningful interpretation of the forecasts can be given. While the forecasting accuracy of structural models is interesting on its own, it is also a test to which extent this class of models explains real world business cycle dynamics. A thorough assessment of different structural models including a comparison to forecasts from sophisticated time series models and to professional forecasts has not been undertaken yet. 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, Hallin, Lippi, and Reichlin, 2003; Marcellino, Stock, and Watson, 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 is perfectly synchronized with the Greenbook and thus the results can also be compared to a best practice benchmark given by the Greenbook projections. 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). The dataset includes data vintages for 145 FOMC meetings between March 1980 and December 2000.

I consider models that cover to some extent the range of closed-economy DSGE models used in academia and at policy institutions. 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 backward looking demand side, while the Phillips curve is derived from overlapping wage contracts. The third model is a medium-scale New Keynesian model as developed in Christiano, Eichenbaum, and Evans (2005). I use the estimated version by Smets and Wouters (2007). The fourth model is a version of the DSGE model by Edge, Kiley, and Laforte (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 the 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 use of such data in combination with structural macroeconomic models (see Giannone, Monti, and Reichlin, 2009). Employing such methods is beyond the scope of this paper. 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 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 the individual model forecasts, I consider several simple and sophisticated model averaging schemes to compute weighted forecasts. For example, Gerard and Nimark (2008) and Bache, Jore, Mitchell, and Vahey (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. The forecasting quality of the structural models is in particular competitive to the Greenbook projections for medium term horizons. For output

growth, several models outperform the Greenbook projections and have an accuracy comparable to the best nonstructural models. Large scale 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 gives more accurate forecasts than a simple univariate autoregressive process. The Greenbook inflation forecast is more accurate than all model forecasts. For the 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. 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. Yet the parameterization is tight enough to yield accurate forecasts.

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 to identify a single model rather than giving weight to several models. The forecasts of a simple average of the forecasts of all models are in many cases most accurate and otherwise only marginally less accurate than weighted forecasts from more sophisticated weighting methods.

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

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.[1] I briefly describe the main features of the models. All models have been applied in Wieland and Wolters (2010) 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. 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

[1] 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, Schorfheide, Smets, and Wouters (2007).

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 so-called FRB/EDO model by Edge, Kiley, and Laforte (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).[2] 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

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

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, Litterman, and Sims, 1984) to shrink the parameters towards zero. The Minnesota prior assumes that the vector of time series is well-described as a collection of independent random walks. I use growth rates or stationary time series and therefore put a prior assumption of a zero coefficient on the first lag of the dependent variable instead of a one. Therefore, all parameters are assumed to be normally distributed with mean zero. The prior variance of the parameters decreases with the lag length. While larger Bayesian VARs and specifications with the level of output and prices can potentially increase the accuracy of foreacsts, I use a version that uses the same variables as the 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. 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. 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 number of parameters in the Bayesian VAR can vary from 3 shock variances to 39 parameters depending on the significance of the four lags of each variable in each of the three equations. 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.[3].

For the priors, I use the ones in the original research referenced in Table 1. Except for the model by Fuhrer & Moore, 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 derive a first order approximation of the solution. All other models are linearized.

[3]To be sure, I approximate maximum likelihood estimation by defining wide uniform priors for all parameters and use then the same Bayesian estimation algorithm as for the other models. Therefore, exactly the same statistics are derived for all models which is important for the computation of weighted forecasts in section 6

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	3-39	39	3: output growth, inflation, interest rate	Doan et al. (1984)

Notes: Type: short classification of the models according to the main modelling 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.

3 A real-time dataset

I use the real-time dataset described in Faust and Wright (2009).[4] The dataset is prepared by the Federal Reserve staff to compute the Greenbook forecasts. The data is perfectly synchronized with the Greenbook and contains historical samples, i.e. data vintages, of 109 variables as observed at the time the Greenbook was published. In addition, it contains nowcasts and forecasts up to five quarters for all variables. The dataset contains data vintages for 145 FOMC meetings from March 1980 to December 2000, while the different data series start in 1960.[5] While 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

[4]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.

[5]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 Wright (2009).

the FRB/EDO model are not available in the data set. Therefore I add the necessary real-time data series from the Federal Reserve Bank of St. Louis' Alfred database and also the accordant nowcasts from the Greenbook. To each data vintage I add only observations that would have been available at the Greenbook publication date.

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. Therefore a window of 80 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.

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.

While 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 other two models are fit to seven and eleven time series, respectively. 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.[6]

[6]Output is in real terms available in the data set 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

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 order approximation of the solution is derived. The other models are already linearized before solving

the aging of the population that cannot be captured by the models. I remove these following Francis and Ramey (1995) by computing deviations of the hours per capita series from its low frequent HP-filtered trend with a parameter of 16000. The realtime characteristic of the data remains unaffected by this procedure. For the FRB/EDO model nominal time series except for output growth are used. Growth rates are computed for 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.

them.[7] 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 both 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. maximum likelihood estimation is basically Bayesian estimation with uniform or uninformative priors. 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)[8]. 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 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. Reestimating the models for all 145 available data vintages would be computationally too intensive. Finally, the mean parameters can be computed from the posterior distribution of β^m.

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

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

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 ($s = 1, \ldots, 10000$) 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.

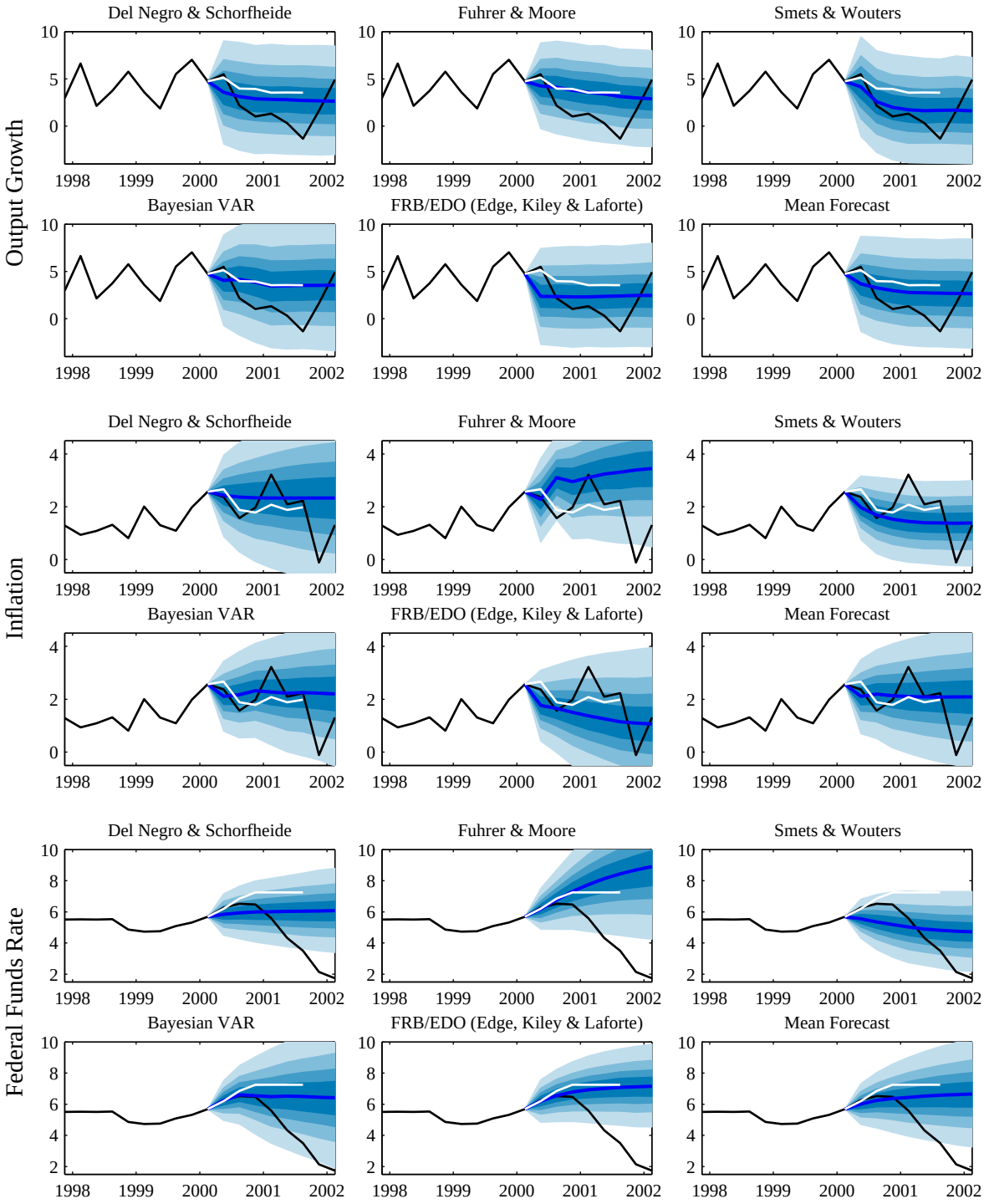
The different steps to compute forecasts are:

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.

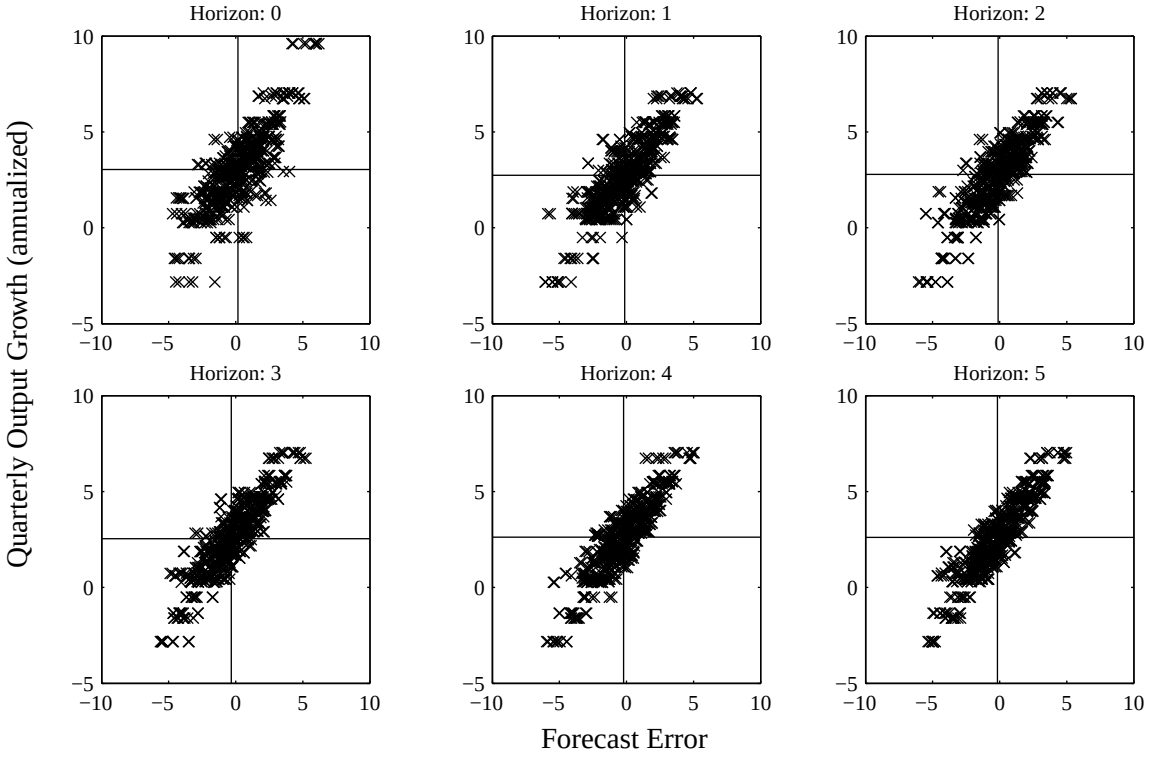
Figure 1 shows as an example forecasts for output growth, inflation and the federal funds rate derived from data vintage May 12, 2000. The black line shows real-time data until the forecast start and revised data afterwards. I plot the 0.05, 0.15, 0.25, 0.35 and 0.65, 0.75, 0.85 and 0.95 percentiles 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. 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. The average of the five 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.

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 and the Bayesian VAR forecast.

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 and the Bayesian VAR 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.

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 well 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 the models including the Bayesian VAR predict a quick return back to the steady state growth rate. Even if one of the models would imply more persistence, it is unlikely to capture the length of recessions accurately as these are rare events with few data points so that their implied persistence cannot be captured

precisely when estimating a model. Each recessions might be caused by different exogenous reasons and therefore there is no information in previous data samples that can be used to forecast the length of such a recession in the future. While the point forecasts cannot predict a recession, the possibility that a large deviation from steady state values occurs is captured by the wide confidence bands. Once the turning point of a recession has been reached, all models predict the economic recovery back to the balanced growth path well. Recoveries in this data sample are quick with little persistence just like the internal propagation mechanism of the models used in this paper.

5 Forecast Evaluation

Table 2 reports the 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 forecast. 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 the RMSEs for output growth and inflation are directly comparable to Table 2 in Faust and Wright (2009). The reported RMSEs are thus based on 122 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-lgenth of 10, covering a bit more than a year, to account fore serial correlation of forecast errors.

The results for inflation, output growth and the federal funds rate are very different. 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 not translate into a superior forecasting performance at higher horizons. The SW, EDO and BVAR models' forecasts dominate the Greenbook forecast from horizon 1 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 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. 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 an univariate autoregressive forecast.[9] I find that only the SW model's forecasts are more precise than an autoregressive forecast from horizon 2 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 differs statistically significant from the Greenbook RMSE with the SW model's forecasts for horizon 3 being the only exception. The difference in the forecasting accuracy of the models can be traced to the different modelling 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 tightly enough parametrized to yield precise forecasts.
The 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. None of 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

[9] 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.

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	1.75	1.20	1.13	1.24	1.21	1.11	1.09	1.39
1	2.12	**0.95**	1.05	**0.91**	**0.91**	**0.97**	**0.86**	1.20
2	2.01	1.06	1.10	**0.93**	1.00	**0.96**	**0.95**	1.15
3	2.15	**0.99**	1.09	**0.86**•	**0.95**	**0.97**	**0.94**	1.12
4	2.08	1.01	1.05	**0.89**	**0.94**	**0.94**	**0.99**	1.11
5	2.08	1.02	1.05	**0.90**	**0.99**	1.00	**0.97**	1.09
				jump off 0				
1	2.12	**0.95**	1.03	**0.93**	**0.94**	**0.94**	**0.84**	1.07
2	2.01	1.06	1.13	**0.94**	1.00	**0.97**	**0.90**	1.12
3	2.15	1.00	1.12	**0.87**	**0.97**	**0.96**	**0.95**	1.18
4	2.08	1.01	1.08	**0.88**	**0.97**	**0.97**	**0.96**	1.09
5	2.08	1.03	1.06	**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	0.69	1.52●	1.86●	1.48●	1.65●	1.47●	1.34●	1.63●
1	0.79	1.59●	1.80●	1.44●	1.50●	1.45●	1.22●	1.86●
2	0.81	1.38●	1.57●	1.29●	1.59●	1.30●	1.15•	1.92●
3	0.93	1.17•	1.42•	1.20•	1.50●	1.14	1.03	1.84●
4	0.89	1.28•	1.80●	1.29●	1.46●	1.35•	1.08	2.11●
5	1.14	1.24•	1.62●	1.24●	1.33●	1.30	**0.99**	1.83●
				jump off 0				
1	0.79	1.24•	1.61●	1.15•	1.17•	1.25●	1.20●	1.58●
2	0.81	1.25●	1.50●	1.18•	1.16•	1.25●	1.18	1.69●
3	0.93	1.24●	1.27•	1.22●	1.27•	1.15•	1.04	1.66●
4	0.89	1.19●	1.51●	1.20●	1.26•	1.19	1.05	1.91●
5	1.14	1.15•	1.47•	1.21●	1.14	1.19	**0.97**	1.77●

(c) Federal Funds Rate

horizon	GB	DS	FM	SW	EDO	BVAR	best FW	worst FW
				jump off -1				
0	0.11	5.91●	4.84●	4.63●	5.98●	3.57●	-	-
1	0.49	2.13●	1.88●	1.89●	2.39●	1.55●	-	-
2	0.90	1.49•	1.46●	1.37•	1.75●	1.18	-	-
3	1.25	1.19	1.25•	1.10	1.53●	1.01	-	-
4	1.60	1.05	1.22	**0.97**	1.40●	**0.96**	-	-
5	1.90	**0.97**	1.23•	**0.87**	1.29•	**0.92**	-	-
				jump off 0				
1	0.49	1.37•	1.30●	1.19•	1.66●	1.06	-	-
2	0.90	1.18	1.08	1.07	1.53●	**0.96**	-	-
3	1.25	1.02	1.01	**0.95**	1.45●	**0.90**	-	-
4	1.60	**0.95**	1.03	**0.89**	1.38●	**0.88**	-	-
5	1.90	**0.90**	1.08	**0.83**	1.31●	**0.86**	-	-

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 FW: Best performing atheoretical model for the specific horizon considered by Faust & Wright; Worst FW: Worst 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 RMSEs of alternative models 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.

the quality of the Greenbook forecasts. 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 all horizons for the jump of 0 scenario.

The Greenbook projections are conditioned on a hypothetical path of policy. This hypothetical federal funds rate is not meant to be a forecast. Nevertheless, viewing it as a forecast 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 future 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 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. For medium term forecasts, however, the 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. The results might be sensitive to the hypothetical policy path characteristic of the Greenbook projection. If the Fed's staff would compute an unconditional best forecast for the federal funds rate it might as well dominate the model forecasts for all horizons. 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.

Faust and Wright (2009) present a table showing the percentage of forecast periods in which the time series model 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 4 in the Appendix. 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 dominates the model nowcasts. For the other horizons the model forecasts for output growth are as good as the Greenbook forecasts or even better. 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

Density forecasts are useful to show uncertainty around point forecasts. Having estimated the posterior parameter distribution of a certain model, it is straightforward to compute density forecasts that include various sources of uncertainty. One computes forecasts for a large number of draws from the models' posterior parameter distribution to take into account parameter uncertainty. Uncertainty about future realizations of shocks is incorporated by repeatedly drawing from their estimated distribution. However, the largest source of uncertainty - model uncertainty - is ignored. Using only one model to forecast 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 and an unconstrained Bayesian VAR. Computing weighted forecasts is interesting for a second reason: the results in the empirical literature on forecast combination show that combining multiple forecasts increases the forecasting accuracy. Unless one can identify a single model with superior forecasting performance, forecast combinations are useful for diversification reasons as one does not have to rely on the forecast of a single model. 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 structural 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 assigend 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 forecast 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:

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

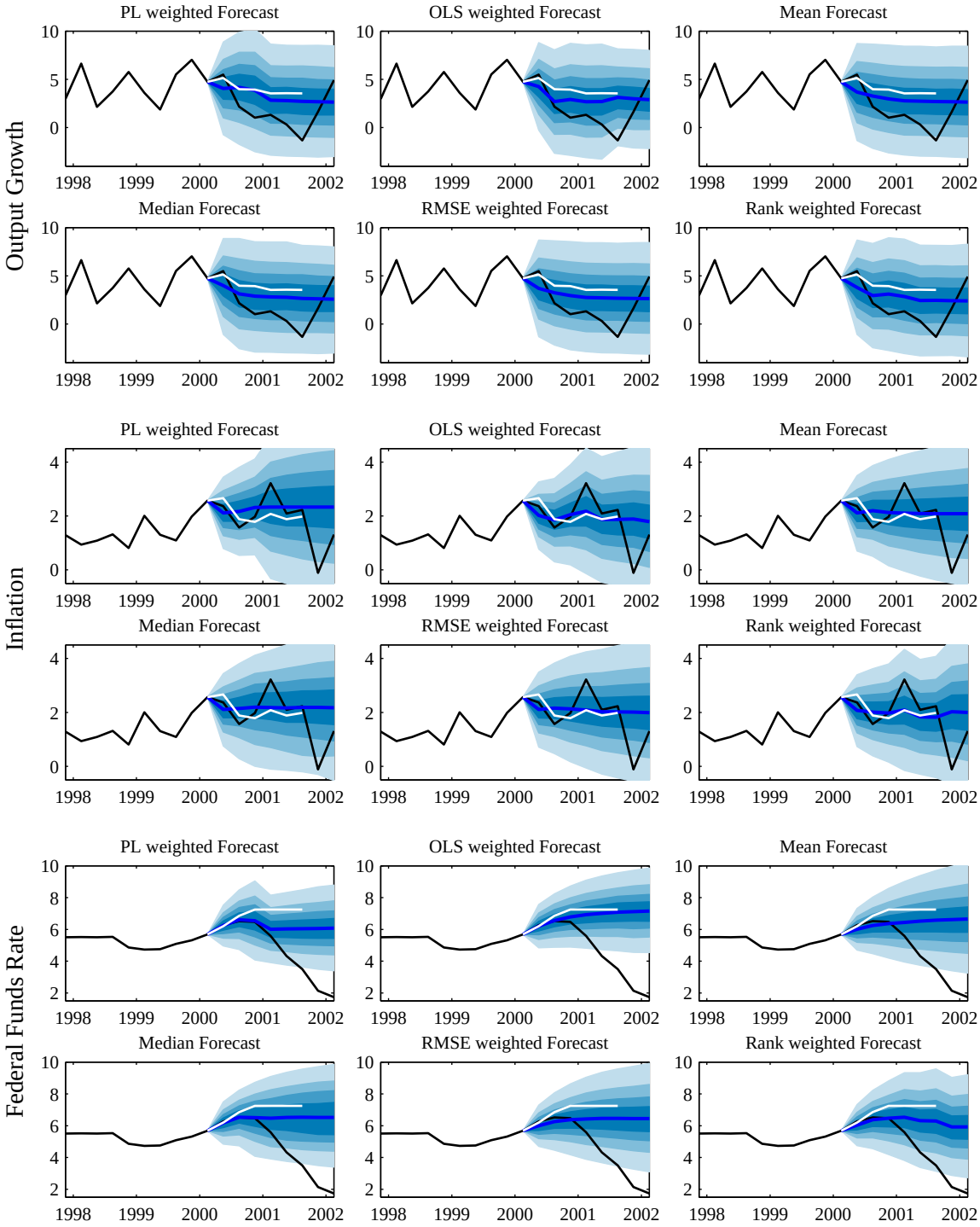
Both methods can assign different weights to forecasts of different variables and the different forecasting horizons.

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 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. Only the weighted forecasts based on the predictive likelihood and on ranked past forecasting performance predict a slight interest rate decrease.

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.

7 Forecast Evaluation of combined forecasts

In Table 3, I report the 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 have in general an accuracy higher than forecasts from most single models. For output growth the Greenbook nowcast is slightly better 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 horizons two and above most weighted forecasts 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. All the differences in output growth forecasting accuracy are statistically insignificant, with the Rank weighted horizon 3 forecast being the only exception.

For the inflation forecast, weighted forecasts increase the forecasting accuracy compared to most single model forecasts. 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. Especially for the medium term forecast it improves upon the best single model forecast. For medium term horizons it is only slightly worse than the Greenbook forecast and the best nonstructural forecast. The 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. An 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 3: 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	1.75	1.17	1.05	1.07	1.06	1.06	1.04	1.11	1.09
1	2.12	**0.93**	**0.90**	**0.89**	**0.86**	**0.86**	**0.87**	**0.91**	**0.86**
2	2.01	1.06	**0.93**	**0.92**	**0.92**	**0.91**	**0.90**	**0.93**	**0.95**
3	2.15	**0.99**	**0.88**	**0.91**	**0.90**	**0.89**	**0.85●**	**0.86•**	**0.94**
4	2.08	1.00	**0.92**	**0.90**	**0.89**	**0.89**	**0.87**	**0.89**	**0.99**
5	2.08	1.02	**0.92**	**0.91**	**0.92**	**0.92**	**0.90**	**0.90**	**0.97**
					jump off 0				
1	2.12	**0.96**	**0.90**	**0.85**	**0.85**	**0.85**	**0.85**	**0.93**	**0.84**
2	2.01	1.01	**0.94**	**0.93**	**0.91**	**0.91•**	**0.90●**	**0.94**	**0.90**
3	2.15	1.02	**0.94**	**0.92**	**0.90**	**0.90**	**0.91**	**0.87**	**0.95**
4	2.08	1.02	**0.93**	**0.92**	**0.90**	**0.90**	**0.89**	**0.88**	**0.96**
5	2.08	1.03	**0.98**	**0.92**	**0.92**	**0.92**	**0.95**	**0.89•**	**0.98**

(b) Inflation

horizon	GB	PL	OLS	Median	Mean	RMSE	Rank	best M	best FW
					jump off -1				
0	0.69	1.52●	1.60●	1.45●	1.44●	1.44●	1.45●	1.47●	1.34
1	0.79	1.58●	1.54●	1.47●	1.43●	1.44●	1.47●	1.44●	1.22
2	0.81	1.37●	1.42●	1.25●	1.23●	1.23●	1.25●	1.29●	1.15
3	0.93	1.17●	1.20●	1.10	1.06	1.07	1.11	1.14	1.03
4	0.89	1.28●	1.32●	1.20•	1.15	1.17	1.20•	1.28●	1.08
5	1.14	1.24●	1.21	1.19•	1.11	1.12	1.16	1.24●	**0.99**
					jump off 0				
1	0.79	1.23●	1.25•	1.16●	1.18●	1.17●	1.17●	1.15●	1.20●
2	0.81	1.24●	1.27●	1.19●	1.16•	1.16•	1.17●	1.16•	1.18
3	0.93	1.23●	1.29●	1.15●	1.09•	1.09•	1.11•	1.15•	1.04
4	0.89	1.18●	1.18•	1.10	1.07	1.07	1.14•	1.19	1.05
5	1.14	1.15●	1.17●	1.12•	1.06	1.06	1.09	1.14	**0.97**

(c) Federal Funds Rate

horizon	GB	PL	OLS	Median	Mean	RMSE	Rank	best M	best FW
					jump off -1				
0	0.11	5.95●	4.45●	3.77●	3.56●	3.49●	3.42●	3.57●	-
1	0.49	2.14●	2.13●	1.65●	1.47●	1.47●	1.45●	1.55●	-
2	0.90	1.49●	1.54●	1.22•	1.14	1.14	1.14	1.18	-
3	1.25	1.19	1.33●	1.01	**0.99**	**0.99**	1.00	1.01	-
4	1.60	1.05	1.26●	**0.95**	**0.94**	**0.94**	**0.97**	**0.96**	-
5	1.90	**0.97**	1.19●	**0.91**	**0.92**	**0.91**	**0.91**	**0.87**	-
					jump off 0				
1	0.49	1.37●	1.63●	1.08	1.01	1.02	1.07	1.06	-
2	0.90	1.18	1.49●	**0.99**	**0.93**	**0.93**	**0.97**	**0.96**	-
3	1.25	1.02	1.29•	**0.89**	**0.86**	**0.87**	**0.94**	**0.90**	-
4	1.60	**0.95**	1.23•	**0.88**	**0.87**	**0.87**	**0.92**	**0.88**	-
5	1.90	**0.90**	1.19•	**0.86**	**0.86**	**0.86**	**0.89**	**0.86**	-

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.

Mean forecast becomes as precise as the best nonstructural forecast. For the jump of 0 scenario all weighted forecasts are more accurate than an 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. Nevertheless, combining this forecast with other less accurate forecasts even improves the forecasting quality: the Mean, RMSE and Rank weighted forecasts are more accurate than the forecasts from the Bayesian VAR. While the Greenbook interest rate path is significantly 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 6 to 8 in the Appendix report as an example model weights for forecasts derived from data vintage May 12, 2000. Wieland and Wolters (2010) report RMSEs for structural forecasts for five different recessions and find that there is no model that consistently outperforms other models. This shows that the forecasting performance of different models relative to each other varies over time. 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 many 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. Table 5 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 data sets. All forecasts based on structural models are especially suited to compute medium term forecasts.

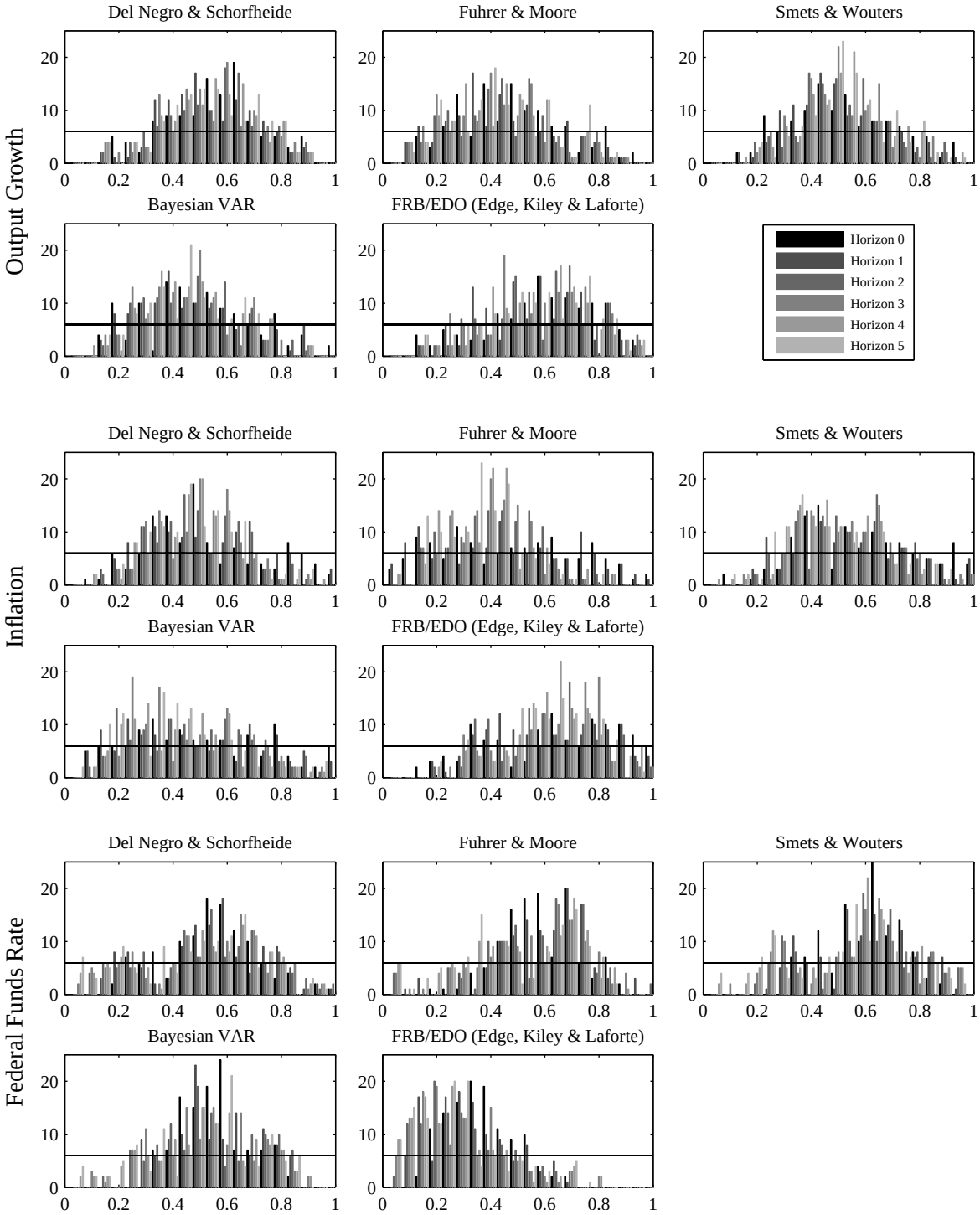
8 Density Forecast Evaluation

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 the forecasts provide a realistic description of actual uncertainty. 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, Gunther, and Tay (1998) and Diebold, Hahn, and Tay (1999). More formally, it is based on the relationship between the 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 $\{y_{t+h}^{obs}\}_{t=1}^{n}$:

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

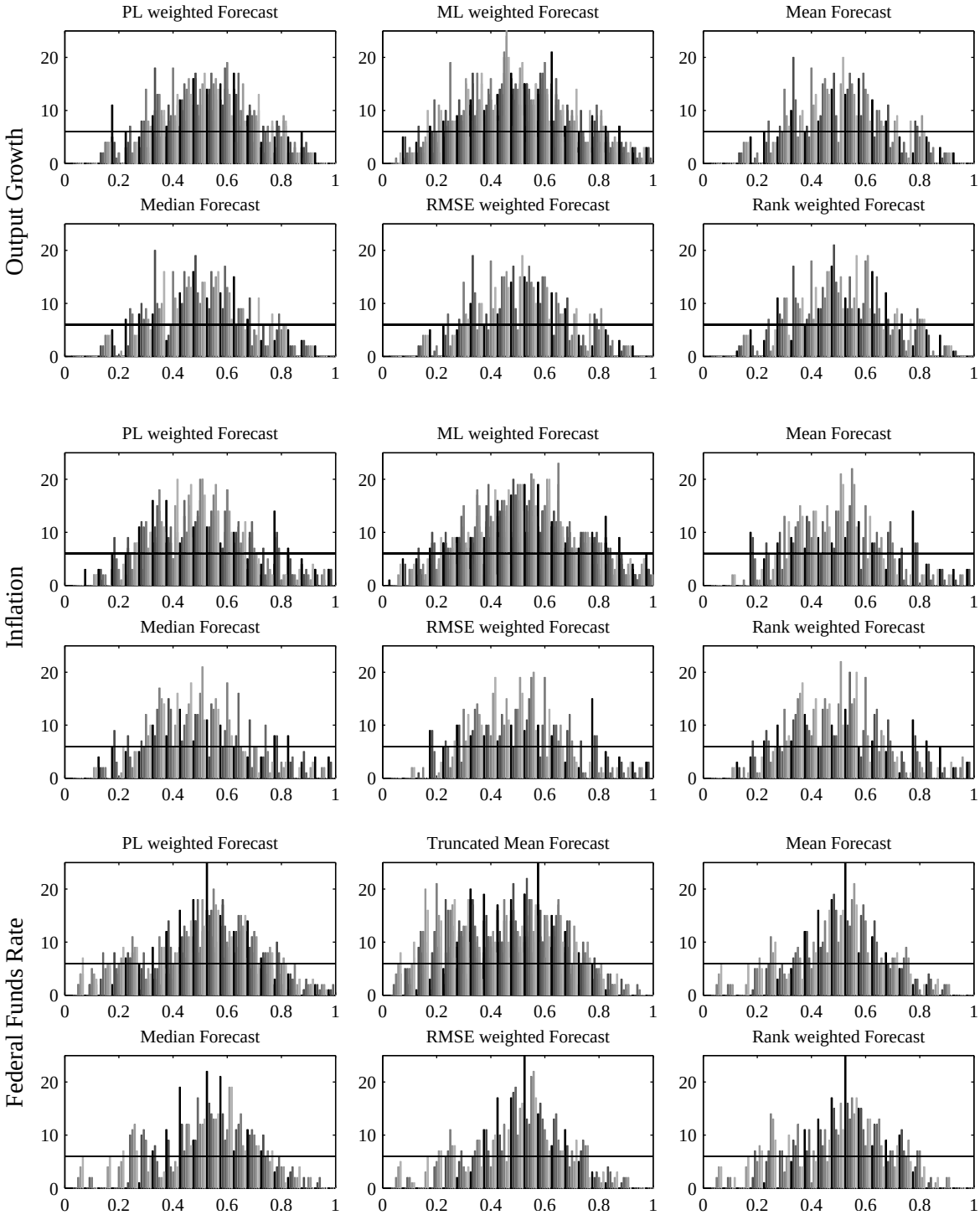
The PIT is the probability implied by the density forecast that a realized data point would be equal or less than what is actually observed. If the sequence of density forecasts is an accurate description of actual uncertainty, the sequence of PITs, $\{z_t\}_{t=1}^{n}$, should be distributed uniformly 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 $n = 122$ 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

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 122 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 122 observations should fall in each probability band.

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.[10]

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 academics 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 data sets. Especially the model by Smets and Wouters (2007) yields relatively precise inflation, output growth and interest rate forecasts. Combining several forecasts can 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.

[10] 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, Kapetanios, Taylor, and Yates, 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.

Large shocks are needed to fit the models to volatile periods of the sample. This is also the reason for their wide confidence bands.

References

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

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., 2010. Forecasting with DSGE models, ECB Working Paper No. 1185.

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. 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, forthcoming.

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 real-time 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., 1995. Measures of per capita hours and their implications for the technology-hours debate, nBER Working Paper 11694.

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.

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.

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

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., 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., 2010. The diversity of forecasts from macroeconomic models of the U.S. economy. Economic Theory forthcoming.

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

Appendix: Additional Results

Table 4: 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	29	34	37	32	38	43	30
1	**52**	45	48	48	**51**	**60**	39
2	48	47	**53**	49	**53**	**58**	37
3	47	43	**59**	**51**	**51**	**57**	42
4	44	45	**52**	48	48	**54**	36
5	45	43	**60**	49	42	**52**	43
				jump off 0			
1	43	**51**	49	48	50	**59**	40
2	48	49	**57**	43	**53**	**55**	41
3	48	47	**55**	48	**52**	**57**	38
4	46	47	**53**	42	**52**	**57**	39
5	43	44	**55**	43	47	49	43

(b) Inflation

horizon	DS	FM	SW	EDO	BVAR	best FW	worst FW
				jump off -1			
0	41	30	41	29	38	37	25
1	29	31	44	38	35	40	21
2	41	38	36	35	39	43	25
3	44	36	33	32	40	44	17
4	43	30	36	31	34	43	11
5	37	31	38	35	35	46	16
				jump off 0			
1	36	35	36	43	36	41	30
2	37	32	40	45	38	40	21
3	42	43	37	38	48	43	20
4	37	26	33	36	38	43	18
5	38	31	31	50	33	48	15

(c) Federal Funds Rate

horizon	DS	FM	SW	EDO	BVAR	best FW	worst FW
				jump off -1			
0	8	13	6	4	13	-	-
1	28	27	22	11	25	-	-
2	45	33	32	18	38	-	-
3	50	34	39	23	45	-	-
4	**56**	31	45	30	48	-	-
5	**60**	34	50	29	**56**	-	-
				jump off 0			
1	33	31	29	23	38	-	-
2	41	35	39	27	50	-	-
3	46	42	48	27	**53**	-	-
4	48	40	**53**	29	**57**	-	-
5	**53**	42	**54**	24	**59**	-	-

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 FW: Best performing atheoretical model for the specific horizon considered by Faust & Wright; Worst FW: Worst 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.

Table 5: 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	36	43	36	40	40	39	38	43
1	**52**	**55**	**55**	**55**	**56**	**55**	**52**	**60**
2	45	**55**	**54**	**57**	**57**	**56**	**53**	**58**
3	47	**57**	**55**	**57**	**58**	**63**	**59**	**57**
4	44	49	**60**	**54**	**54**	**65**	**52**	**54**
5	45	49	**54**	**56**	**55**	**56**	**60**	**52**
				jump off 0				
1	44	**53**	**54**	**57**	**57**	**56**	**51**	**59**
2	46	**54**	**62**	**58**	**61**	**53**	**57**	**55**
3	44	**53**	**55**	**55**	**55**	**56**	**55**	**57**
4	46	**54**	**55**	**53**	**53**	**53**	**53**	**57**
5	43	49	**53**	**53**	**53**	**53**	**55**	49

(b) Inflation

horizon	PL	OLS	Median	Mean	RMSE	Rank	best M	best FW
				jump off -1				
0	39	32	42	37	38	40	41	37
1	33	34	33	38	38	34	44	40
2	41	40	46	43	44	46	41	43
3	44	43	45	49	48	46	44	44
4	43	42	43	44	45	43	43	43
5	37	38	39	43	44	41	38	46
				jump off 0				
1	38	40	37	37	38	39	43	41
2	38	39	41	43	43	45	45	40
3	42	37	43	47	46	50	48	43
4	37	38	39	44	43	42	38	43
5	38	43	35	40	42	43	50	48

(c) Federal Funds Rate

horizon	PL	OLS	Median	Mean	RMSE	Rank	best M	best FW
				jump off -1				
0	10	9	13	12	12	13	13	-
1	29	14	29	29	32	31	28	-
2	43	29	42	41	38	40	45	-
3	50	37	48	**51**	**54**	50	49	-
4	**56**	34	**57**	**56**	**57**	**56**	**56**	-
5	**60**	33	**58**	**61**	**61**	**60**	**60**	-
				jump off 0				
1	31	23	36	38	37	40	38	-
2	43	29	45	48	45	**53**	50	-
3	45	38	**55**	**58**	**57**	**51**	50	-
4	48	38	**59**	**56**	**57**	**54**	**57**	-
5	**53**	33	**60**	**63**	**62**	**53**	**59**	-

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.

Table 6: Combination weights for data vintage May 12, 2000: output growth

model	PL	OLS	Median	Mean	RMSE	Rank
			horizon 0			
DS	0.00	0.00	0.01	0.20	0.19	0.09
FM	0.00	1.00	0.33	0.20	0.21	0.22
SW	0.00	0.00	0.33	0.20	0.19	0.11
EDO	0.00	0.00	0.00	0.20	0.19	0.15
BVAR	1.00	0.00	0.32	0.20	0.22	0.44
			horizon 1			
DS	0.00	0.00	0.98	0.20	0.19	0.11
FM	0.00	0.00	0.00	0.20	0.18	0.09
SW	0.00	0.42	0.00	0.20	0.21	0.44
EDO	0.00	0.45	0.02	0.20	0.21	0.22
BVAR	1.00	0.12	0.00	0.20	0.21	0.15
			horizon 2			
DS	0.00	0.00	0.93	0.20	0.19	0.11
FM	0.00	0.00	0.02	0.20	0.18	0.09
SW	0.00	0.19	0.00	0.20	0.21	0.22
EDO	0.00	0.44	0.05	0.20	0.21	0.15
BVAR	1.00	0.37	0.00	0.20	0.21	0.44
			horizon 3			
DS	1.00	0.00	0.78	0.20	0.19	0.11
FM	0.00	0.00	0.06	0.20	0.18	0.09
SW	0.00	0.19	0.00	0.20	0.21	0.44
EDO	0.00	0.42	0.10	0.20	0.21	0.15
BVAR	0.00	0.38	0.06	0.20	0.21	0.22
			horizon 4			
DS	1.00	0.00	0.75	0.20	0.19	0.09
FM	0.00	0.00	0.09	0.20	0.19	0.11
SW	0.00	0.28	0.00	0.20	0.21	0.44
EDO	0.00	0.37	0.12	0.20	0.20	0.15
BVAR	0.00	0.35	0.04	0.20	0.21	0.22
			horizon 5			
DS	1.00	0.00	0.53	0.20	0.19	0.09
FM	0.00	1.00	0.26	0.20	0.20	0.15
SW	0.00	0.00	0.00	0.20	0.21	0.44
EDO	0.00	0.00	0.15	0.20	0.19	0.11
BVAR	0.00	0.00	0.06	0.20	0.21	0.22

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; BVAR: Bayesian VAR; 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 five model weights sum up to 1.

Table 7: Combination weights for data vintage May 12, 2000: inflation

model	PL	OLS	Median	Mean	RMSE	Rank
			horizon 0			
DS	0.00	0.00	0.00	0.20	0.22	0.15
FM	0.00	0.00	0.11	0.20	0.16	0.09
SW	0.00	0.62	0.05	0.20	0.23	0.44
EDO	0.00	0.00	0.00	0.20	0.18	0.11
BVAR	1.00	0.38	0.84	0.20	0.22	0.22
			horizon 1			
DS	0.00	0.00	0.21	0.20	0.20	0.15
FM	0.00	0.00	0.00	0.20	0.17	0.09
SW	0.00	0.49	0.03	0.20	0.23	0.44
EDO	0.00	0.14	0.00	0.20	0.19	0.11
BVAR	1.00	0.37	0.76	0.20	0.22	0.22
			horizon 2			
DS	0.00	0.00	0.50	0.20	0.20	0.15
FM	0.00	0.30	0.00	0.20	0.19	0.11
SW	0.00	0.35	0.07	0.20	0.22	0.44
EDO	0.00	0.23	0.00	0.20	0.17	0.09
BVAR	1.00	0.11	0.44	0.20	0.22	0.22
			horizon 3			
DS	1.00	0.25	0.44	0.20	0.24	0.44
FM	0.00	0.35	0.00	0.20	0.17	0.09
SW	0.00	0.00	0.10	0.20	0.22	0.22
EDO	0.00	0.39	0.00	0.20	0.17	0.11
BVAR	0.00	0.00	0.46	0.20	0.20	0.15
			horizon 4			
DS	1.00	0.00	0.36	0.20	0.22	0.22
FM	0.00	0.31	0.00	0.20	0.16	0.09
SW	0.00	0.16	0.11	0.20	0.23	0.44
EDO	0.00	0.54	0.00	0.20	0.20	0.15
BVAR	0.00	0.00	0.52	0.20	0.19	0.11
			horizon 5			
DS	1.00	0.00	0.33	0.20	0.22	0.22
FM	0.00	0.33	0.00	0.20	0.16	0.09
SW	0.00	0.15	0.13	0.20	0.23	0.44
EDO	0.00	0.52	0.00	0.20	0.20	0.15
BVAR	0.00	0.00	0.54	0.20	0.18	0.11

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; BVAR: Bayesian VAR; 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 five model weights sum up to 1.

Table 8: Combination weights for data vintage May 12, 2000: Federal Funds Rate

model	PL	OLS	Median	Mean	RMSE	Rank
			horizon 0			
DS	0.00	0.00	0.00	0.20	0.18	0.11
FM	0.00	0.00	0.00	0.20	0.21	0.15
SW	0.00	0.00	0.00	0.20	0.22	0.22
EDO	0.00	1.00	1.00	0.20	0.14	0.09
BVAR	1.00	0.00	0.00	0.20	0.25	0.44
			horizon 1			
DS	0.00	0.00	0.00	0.20	0.18	0.11
FM	0.00	0.00	0.00	0.20	0.23	0.22
SW	0.00	0.00	0.00	0.20	0.20	0.15
EDO	0.00	1.00	1.00	0.20	0.14	0.09
BVAR	1.00	0.00	0.00	0.20	0.24	0.44
			horizon 2			
DS	0.00	0.00	0.03	0.20	0.19	0.11
FM	0.00	0.00	0.00	0.20	0.22	0.22
SW	0.00	0.00	0.00	0.20	0.20	0.15
EDO	0.00	1.00	0.54	0.20	0.15	0.09
BVAR	1.00	0.00	0.43	0.20	0.25	0.44
			horizon 3			
DS	1.00	0.00	0.12	0.20	0.19	0.11
FM	0.00	0.00	0.00	0.20	0.20	0.22
SW	0.00	0.00	0.00	0.20	0.20	0.15
EDO	0.00	1.00	0.38	0.20	0.16	0.09
BVAR	0.00	0.00	0.50	0.20	0.24	0.44
			horizon 4			
DS	1.00	0.00	0.16	0.20	0.21	0.15
FM	0.00	0.00	0.00	0.20	0.18	0.11
SW	0.00	0.00	0.00	0.20	0.21	0.22
EDO	0.00	1.00	0.38	0.20	0.16	0.09
BVAR	0.00	0.00	0.46	0.20	0.23	0.44
			horizon 5			
DS	1.00	0.00	0.22	0.20	0.21	0.15
FM	0.00	0.00	0.00	0.20	0.17	0.09
SW	0.00	0.00	0.00	0.20	0.22	0.22
EDO	0.00	1.00	0.38	0.20	0.17	0.11
BVAR	0.00	0.00	0.40	0.20	0.23	0.44

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; BVAR: Bayesian VAR; 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 five model weights sum up to 1.