Staff Memo

Improving and evaluating short term forecasts at the Norges Bank

Hilde C. Bjørnland, Anne Sofie Jore, Christie Smith and Leif Anders Thorsrud, Norges Bank Monetary Policy

%NB% NORGES BANK

Improving and evaluating short term forecasts at the Norges Bank

Hilde C. Bjørnland, Anne Sofie Jore, Christie Smith and Leif Anders Thorsrud[*]

September 1, 2008

Abstract

In 2006 the Norges Bank initiated a project to improve its short term forecasts. The current phase of the project is tasked with developing a system that provides model-based forecasts for gross domestic product and consumer price inflation excluding taxes and energy prices, for each policy round. Forecasts are recursively evaluated from 1999Q2 to the current quarter. The performance of the models over this period is then used to derive weights that are used to combine the forecasts. Our results indicate that model combination improves upon the forecasts from individual models. The ultimate goal of this phase is to provide density forecasts for GDP and CPI, to provide a formal characterisation of the uncertainty that surrounds the central forecasts. The system is flexible enough to incorporate new models as they are developed, refreshing the system of models.

1 Introduction

Norges Bank's projections form an important basis for the conduct of monetary policy. The forecasting work involves the use of the macromodel NEMO[1], but the model is primarily suited for medium and long-term projections. Projections for the coming few quarters are largely based on current statistics, information from Norges Bank's regional network and forecasts obtained from a number of statistical and econometric models. The published projections are the result of an overall assessment based on both models and judgement.

---

[1]See Brubakk, Husebø, Maih, Olsen, and Østnor (2006) for details

In 2006 the Norges Bank embarked upon a project to improve its short-term forecasting using statistical and econometric models. Over the first two years, the project was directed at improving human capital in the Norges Bank, developing a greater understanding of the models that are in common use for short-term forecasting.

At the beginning of 2008 a new phase of the project was initiated, aimed at producing model-based short term forecasts (up to a year) for the Norwegian economy for each round of monetary policy discussions. The project has three main objectives:

1. Evaluate the short-term forecasts obtained from different models;

2. Improve forecasts; and

3. Provide a better characterisation of uncertainty.

Policy-making entails evaluating the future trajectory of the economy, and making policy decisions to influence that trajectory in favourable directions. The forward-looking nature of these policy decisions means that macroeconomic forecasting is a critical component underlying decisions. By evaluating models on a recurrent basis, the short-term forecasting project helps policy-makers to identify which forecasting models they should pay most attention to, assisting them to make good monetary policy decisions.

Developing empirical models to describe and forecast the behaviour of the economy is subject to many important decisions that can have a material impact on the output – e.g. forecasts – of the models. These decisions include the choice of the data set, the transformations applied to the data, the sample period used to estimate the parameters of the model, the choice of estimation techniques, the dynamic specification of the model, and so on.

A common research strategy is to make choices to test down to a single model specification. However, the model ultimately arrived at will most likely diverge from the true but unknown process that drives the behaviour of the economy. Settling on a single model also disregards all of the other possible models that might be nearly as good as (possibly even better than) the model that was ultimately chosen. If these other models have different implications, such as different forecasts, then one may mis-characterise the central location of the forecasts and also mis-characterise the uncertainty around the forecasts. The sequential testing involved in selection can also distort inference making it difficult to know whether variables have been correctly included or excluding from the set of regressors used to forecast the variables of interest (See for example Bancroft 1944, Bock, Yancey, and Judge 1973 and Raftery 1995).

In recent years it has become increasingly common to adopt an alternative research strategy, which emphasises the *combination* of models or forecasts. Rather than arrive at a single specification, one entertains a wide variety of models and then weights together the output from these models in a sensible

manner. By entertaining a variety of models one can develop a better appreciation of the range of views that could be supported by formal models, and a better appreciation of which outcomes are most likely.

An important part of the Norges Bank's short term forecasting project (colloquially referred to as the 'Nowcasting' project) is to evaluate the forecasts obtained from a broad spectrum of models, and to consider whether model or forecast combination improves short-term forecasts of the Norwegian economy. At the heart of the evaluation and combination of short-term forecasts at the Norges Bank is a set of programs collectively referred to as SAM – the System for Averaging Models. The staff forecasts from SAM impose additional judgment on the benchmark forecasts from NEMO, reflecting alternative information (such as is provided by the Norges Bank's regional network of business contacts), alternative views regarding important macroeconomic drivers, and other models that are not incorporated into SAM.

The rest of this paper has five sections. In section 2 we discuss the motivation for the modelling/forecasting approach from a theoretical and empirical perspective. Section 3 discusses how the forecasts are produced and provides a preliminary discussion of the point forecasts provided by the Nowcasting project. To whet the appetite, it appears that forecast combination does improve upon single model forecasts, and can improve the Norges Bank's own forecasts. In section 4 we discuss work in progress and describe some of the issues that arise with respect to density combination. Section 5 concludes. An extensive appendix is included at the end of the paper, containing a forecasting report with all the forecasts. These forecasts were produced in the first week of June 2008. The report illustrates the information that is currently produced and presented to the monetary policy decision-makers. This appendix forms the basis for the discussion in section 3.

2   Model/forecast evaluation and combination

As discussed in the introduction, the Nowcasting project aims to address three related concerns: evaluating models, improving forecasts, and characterising uncertainty. With respect to the first objective it is immediately obvious that forecast performance needs to be tracked to enable policy-makers to determine how much attention should be paid to different forecasts. In this section we will argue that there are both *theoretical and empirical* reasons for taking a model combination approach to achieve these goals.

2.1   Theoretical reasons for model combination

Model or forecast combination has a long history, dating back at least to Bates and Granger (1969). Bates and Granger consider a situation where a decision-

maker is presented with multiple forecasts and must then decide what to do. In the context of trying to forecast accurately, Bates and Granger show that, ex post, a weighted average of two unbiased forecasts will always have a variance that is less than or equal to the lowest variance of the individual forecasts. A priori, the optimal (constant) weights are not known, and so Bates and Granger consider and evaluate a variety of different schemes for deriving weights recursively. The model combination approach is clearly helpful in evaluating models, because weights (or probabilities) are assigned to the various models, providing a clear signal of performance. Such weights can also be used to rank the models, as we do in our report.

In an excellent survey, Timmerman (2006) emphasizes at least three main reasons for why forecast combinations may produce better forecast on average than methods based on the ex-ante best individual forecasting model. First, forecast combination can be motivated by a simple portfolio diversification (hedging) argument. Suppose the policymaker is faced with many different forecasting models, but can not observe the information sets underlying each of the the individual forecasts. It this situation, it is not possible to pool all the relevant information sets and construct a large model that nests each of the individual forecasting models. Instead, the best way to exploit the information behind the different forecasts is to combine the forecasts. This is discussed in more detail in Huang and Lee (2008). They show that combining forecasts, often with near-equal weights, is frequently superior even when it is feasible to combine information.[2] Hendry and Clements (2002) also show why combining forecasts adds value, and can even dominate the best individual device.

A second rationale for combination, see for example the references in Timmerman (2006), is that there may be unknown instabilities (structural breaks) that sometimes favour one model over another. Some models may adapt to breaks quickly while others may have parameters that will only adjust slowly to the post structural breaks. By combining forecasts from different models, the decision maker may obtain forecasts that are more robust to these instabilities, than if they had chosen a single model.

Related to the above argument, a third motivation is that forecast combination may be desirable as individual forecasting models may be subject to misidentifications bias that are unknown to the model operators. If these models are subject to different biases; combining forecasts may average out the biases, improving forecast accuracy; see the references in Timmerman (2006) once more. Hence, even if the combined forecast may not always be superior, model combination is preferable as it will ensure against selecting a bad model.

---

[2]On the other hand, Diebold and Lopez (1996) note that it is always optimal to combine information sets rather than forecasts if such combination can be done costlessly. In practice, however, it is often infeasible to combine information sets, and the combined information set may not be amenable to usual analytical techniques (for example, if the time dimension is shorter than the cross-sectional dimension).

4

2.2 Empirical experience with model combination

Empirical experience often, though not always, supports the theoretical results that imply model combination will improve forecast performance.[3] Timmerman (2006), with suitable caveats, provides a broad characterisation of empirical results and suggests the following: i) simple combination schemes are hard to beat, and the failure of more elaborate combination schemes is often attributed to the difficulty in estimating model weights; ii) forecasts based on the model with the single-best in-sample performance often perform poorly out-of-sample; iii) shrinkage towards equal weights often improves forecast performance; yet iv) some time-variation or adaptive adjustment of the weights may improve forecast performance.

Following on from earlier papers, Makridakis and Hibon (2000) develop a forecasting competition (known as M3) to evaluate the performance of different forecasting techniques. In their summary of the M3 forecasting competition, Makridakis and Hibon (2000) draw four main conclusions. They note that: a) complex statistical models are often not superior in terms of their forecasts; b) the relative rank of the models depends on the metric used to assess performance; c) on average, combination is superior to individual forecasts; and d) the accuracy of the forecast methods depends on the forecast horizon.

Specific examples and counter-examples of model performance are easy to come by. Koop and Potter (2004) find that there are appreciable gains from using model averaging of factor models to forecast quarterly US GDP and inflation, relative to using single models. Clark and McCracken (2008) argue that combining real-time point forecasts from vector autoregressions of output, prices and interest rates improves point forecast accuracy in the presence of uncertain model instabilities. Stock and Watson (2004) show that for G-7 countries, combining forecasts generally results in robust predictions of GDP growth, and sometimes improves markedly on simple autoregressive benchmarks and dynamic factor model forecasts. Marcellino (2002) suggests that on average, across a wide variety of European economic series, a linear combination of forecasts or a combination using MSE (mean square error) weights works very well. However, he notes that linear and nonlinear models can do better for specific series, and suggests that careful forecast selection may be preferable. Kapetanios, Labhard, and Price (2008), in their discussion of the Bank of England's suite of forecasting models, also find that two non-linear models out-perform their combination forecasts for UK GDP growth (though their performance for inflation is less convincing).

Marcellino's two conclusions echo the theoretical perspective provided by Raftery and Zheng (2003): the long run performance of Bayesian model averaging is good, yet there may be specific instances where model selection is superior. However, the difficulty is always in knowing when such a situation occurs, and at the

---

[3]We concentrate on examples using time series data. See for instance Fernandez, Ley, and Steel (2001) for application to panel data.

very least model performance needs to be reported and evaluated. Raftery and Zheng (2003) provide a number of references that show that the out-of-sample performance of Bayesian model averaging is superior to selection methods. Nevertheless, there are still advocates of model selection, see for example the edited volume Campos, Ericsson, and Hendry (2005).

2.3   Model combination and uncertainty

The main ingredient of the evaluation and combination of short-term forecasts at the Norges Bank is a set of programs collectively referred to as SAM – the System for Averaging Models. These programs are written in Gauss and Matlab, which are matrix programming languages commonly used for computation in economics and other fields. The software is used to provide a quasi-real-time out-of-sample forecasting exercise to establish the performance of the various models and combination schemes. In what sense this is an out-of-sample and quasi-real-time exercise will be clarified in Section 3 below.

The forecasts of the SAM system are combined using univariate, horizon-specific weights. In principle it is possible to use weights derived from multivariate measures of fit (such as the log-likelihood of a model for example), but because not all models forecast all variables it was decided to use univariate weights. Currently the weights are based on root mean squared forecast errors (RMSFE). The RMSFE can be a interpreted as a measure of the differences between the values predicted by a model and the values actually observed at a later stage. Hence, it is the amount by which a forecast differs from the true value of the quantity being forecasted. As Timmerman (2006) notes, equal weights will be appropriate when models have equal forecast error variance, but this is not the case in the suite of models investigated in the short-term forecasting project, making it most reasonable to adapt the model weights based on fit to the data.

Makridakis and Hibon (2000) note that the performance of models varies by forecasting horizon, and the SAM analysis is consistent with this conclusion. For example, a model may forecast well for a 1-step ahead horizon, but may be much worse at forecasting 4-steps ahead. The model weights in SAM have been made horizon-specific for this reason. Thus far the weighting scheme has focussed on root mean squared forecast error, but we are well aware that the relative performance of different models hinges on the method used to evaluate forecasts. We have therefore started producing density forecasts and explored various ways to weight these together. An example of how one can use the densities to evaluate the uncertainty of the forecasts is given in Section 3. The ultimate goal is to provide model-based forecast densities, i.e. fan charts, of possible forecast outcomes detailing the probabilities of future events for all models. Like the central forecasts, these fan charts will also need to be evaluated. Clements (2004), Elder, Kapetanios, Taylor, and Yates (2005), Hall and Mitchell (2007), Eklund and Karlsson (2007) and Gerard and Nimark (2008) provide examples of such density evaluations. Given that we are interested in particular

model properties, e.g. the (density) forecasts rather than the models per se, a sensible approach is to average across the models – integrating out the unknowns – concentrating solely on the things that are of inherent interest.

Model and forecast combination fits quite neatly into a Bayesian modelling strategy. Bordley (1982) casts forecasts combination as a Bayesian problem, where the forecasts are observed (and hence may be regarded as 'data') and the decision-maker's goal is to forecast an uncertain variable conditional on the known forecasts. In this framework there is a family of models that are combined, and the combination weights are calculated as the (posterior) probabilities based on model performance (i.e. fit to the data). These probabilities can be thought of as degrees of belief in the various models, as influenced by the data that has actually occurred. Bayes rule enables one to update the probabilities attached to the different models in light of new data. The implications of the combined model, such as forecasts, can be calculated as a weighted average of the individual models. By considering a family of models, the decision-maker develops a view of the quantities of interest (say forecasts) that is more robust to model uncertainty.

A full Bayesian analysis, while appealing from a conceptual point of view, may be computationally intractable because of the number of possible models that one could theoretically search over.[4] When the model space is smaller, it is possible to use other methods to characterise the model probabilities. For example, one could use use Bayesian (BIC) information criteria to compare alternative models.[5] Information criteria penalize models with additional parameters, hence the BIC model order selection criteria are based on parsimony.

Define $BIC(m) = n\ln(\hat{\sigma}^2) + k(m)ln(n)$ for a model with normally distributed errors, where $\hat{\sigma}^2$ is the estimated variance of the errors, n is the sample size, and $k(m)$ is a penalty weight for model m that depends on the number of parameters. Traditionally information criteria have been used to *select* a model from a family of candidates, but Kass and Raftery (1995) note that the BIC asymptotically approaches the Bayes factor and can be used to approximate the posterior model probabilities if a 'unit information prior' is used. Bayes factors can be calculated for non-nested models and provide a way of characterising the weight of evidence *in favour* of a hypothesis.

In practice weights for each model are estimated by calculating the likelihood and penalising highly parameterised models more heavily. Similar combination schemes can be derived from other information criteria, such as the Akaike

---

[4]The computational difficulty posed by the large model space has been addressed in a number of different ways. When the model space is too large to fully explore, a modern approach is to use Monte Carlo Markov chain (MCMC) methods to sample from the model space. By judiciously choosing the transition function for the Markov chain, one can ensure that the posterior probabilities of the models will coincide with the stationary distribution of the Markov chain and the posterior probabilities can therefore be approximated by sampling from a long realisation of the chain. See Gelman, Carlin, Stern, and Rubin (2003) for a discussion of MCMC methods for Bayesian analysis.

[5]The BIC is sometimes also named the Schwarz Information Criterion.

information criterion (AIC), or indeed simply by using mean-squared errors (omitting penalty terms) to weight the models. MSE-based combination was one of the schemes first investigated in Bates and Granger (1969). Burnham and Anderson (2004) discuss how the AIC (or its small-sample variant AICc) can be used as an asymptotically unbiased measure of Kullback-Leibler information given repeated sampling.[6].

2.4 Forecasting/combination schemes at peer central banks

A number of central banks have developed forecasting systems to meet similar forecasting objectives. A few of these are briefly described below.

The Bank of England has used a 'suite' of models for policy purposes for a considerable period of time.[7] However, a 2003 review by Adrian Pagan suggested that the suite of models was not particularly well-adapted to forecasting inflation and GDP. In response to this criticism, the Bank of England subsequently developed a suite of forecasting models and a system to evaluate their forecast performance. The near-term forecasts from the suite of models are distinct from those produced by BEQM, the Bank of England's dynamic stochastic general equilibrium model, which underpin the fan charts reported in the Bank of England's Inflation Reports.

The suite of models allows for a greater range of modelling techniques and data to be used in the forecasting process. The suite includes linear and non-linear univariate models, vector autoregressive (VAR) models of various specifications, Bayesian VARs, factor models and time-varying coefficient models.

The Bank of England applies forecast combination techniques to the forecasts from the suite of models, see for example Kapetanios, Labhard, and Price (2005) and particularly Kapetanios, Labhard, and Price (2008). Their preferred combination scheme is derived from AIC-weights, though they also report results for equal-weighted averages. (They note that the performance of BIC-weighted averages is similar to the AIC-weighted averages.) Their results indicate individual models find it difficult to beat the forecasts from a simple benchmark autoregressive model. However, combined forecasts frequently out-perform the benchmark and exhibit similar performance to the benchmark even when beaten.

The Riksbank has also taken a multi-model approach to near-term forecasting, see Andersson and Löf (2007) and Andersson, Karlsson, and Svensson (2007). The Riksbank's suite of models incorporates bivariate and multivariate VARs, Bayesian VARS, VARs that also incorporate factors (which summarise the broad

---

[6]Hansen (2008) advocates forecast weights obtained from the Mallows criterion. The Mallows criterion is asymptotically equivalent to the AIC, but the Mallows uses the mean squared prediction error as a measure of divergence from the truth rather than the Kullback-Leibler information criteria (Kobayashi and Sakata 1990)

[7]See `http://www.bankofengland.co.uk/publications/other/beqm/models00.pdf` for a description of some of these models.

comovements in a large number of data series), and indicator models (where the indicators have shorter publication lags than the variables of actual interest). The Riksbank has also explored forecast combination methods; see for example Andersson and Karlsson (2007).

The Reserve Bank of New Zealand uses a suite of statistical models as a cross-check on the central projection provided by the forecasting and modelling teams of the Economics Department. The suite of models includes several factor models, Bayesian VARs, an average of indicator models, and a weighted combination of VAR forecasts. Together with forecasts obtained from the private sector, the suite of models is used to illustrate the uncertainty around the central projection. The point forecasting performance of the models (both root mean squared error and bias) is routinely reported, but densities are not provided or evaluated.

The Bank of Canada is another central bank that is using multiple models for near-term forecasting. The Projection Coordination and Model Development division of the Research Department is responsible for developing and maintaining a suite of macroeconomic models used to prepare economic projections of the Canadian economy. Coletti and Murchison (2002) provide a description of the multi-model approach to policy-making employed at the bank of Canada. A multi-model approach has also been adopted by the Bank of Canada. The Bank of Canada's suite of models includes single equation and indicator models, multi-equation reduced form models and medium-sized dynamic general equilibrium models. The extent to which formal model or forecast combination techniques is employed is difficult to determine from published material.

The model suites described above exhibit considerable overlap with the models employed by the Norges Bank. An exception to this statement is that the Norges Bank does not currently employ non-linear models (such as the regime-switching models and threshold autoregressions employed by the Bank of England), but does include models motivated by formal theory, such as the dynamic stochastic general equilibrium model described in section 3.

Although there are a number of central banks that have explicitly adopted multi-model and forecast/model combination approaches to short-term forecasting, the evaluation of these models universally hinges on an evaluation of the point forecasts (typically through mean squared forecast error). The Bank of England has evaluated the density forecasts of its central projections (the fan charts of the inflation report),[8] but this work does not yet seem to have spilled over into the evaluation of multiple short-term forecasting models.

3   Preliminary discussion of results

In this section we describe the main results based on experience with using SAM to forecast the short term (the horizon is up to a year) so far. The Tables and

---

[8]See Elder, Kapetanios, Taylor, and Yates (2005) and Clements (2004).

Figures referred to here can be found in the forecasting report we attach in the appendix. All the forecast were made in the first week of June.

In genuine real-time forecast evaluations one uses information available at time t, which we will denote I_t, to provide a forecast for a vector of variables X_{t+h} where h is the forecast horizon. One then evaluates the forecast against the actual realisation. For the Nowcasting project the vector of variables X_{t+h} consists of the year-on-year (yoy) inflation rate for the consumer price index excluding taxes and energy (CPIATE), and the year-on-year growth rate for gross domestic product (GDP) in Mainland Noway (YMN).

In a *real-time* forecasting exercise, the decision-maker uses exactly the information available at time t to forecast future variables. Some historical data are, however, subject to revision. For example, the characterisation of YMN at 1999Q4 has evolved over time as Statistics Norway uses more data or new methods to update their calculation of GDP. An important source of revision for seasonally adjusted data is an updated view as to the seasonal factors.

For CPI, the real time issue is not important as CPI is rarely revised. For GDP on the other hand, the Nowcasting project performs a *quasi*-real-time forecasting exercise in the sense that we use the *latest* vintage of data to perform our real-time forecasting exercise. A full real-time dataset, which would detail what vintages of data are available at each particular point in time, is currently unavailable. The Norges Bank is developing a real-time database for national accounts data, and eventually the Nowcasting evaluation will be performed using this data set.

We conduct an out-of-sample forecasting exercise by using information up to time t (but from the *final* vintage data) to forecast variables four quarters ahead, i.e. we wish to forecast $X_{t+1}, X_{t+2}, X_{t+3}, X_{t+4}$. With quarterly data, the maximum forecasting horizon for the Nowcasting project is thus 1 year ahead. Conditioning on information up to and including 1999Q1, the models are used to produce forecasts for 1999Q2 to 2000Q1. The conditioning information set is then extended one period forward (to 1999Q2), and forecasts are then made for CPI-ATE and YMN for 1999Q3 to 2000Q2. The conditioning information set is then extended another period, and the whole process is repeated until all available information has been used, to provide the current forecast conditioning on currently available information.

As of 2008Q2, the actual data between 1999Q2 and 2008Q1 are used to evaluate the average performance of the h-step forecasts, $h = 1, \ldots, 4$, from the different models. Currently, the SAM project evaluates the point forecasts for each horizon, which indicate the central expected location of yoy CPI-ATE inflation and yoy YMN growth at the four horizons. The primary evaluation criteria has been the root mean squared forecast errors (RMSFEs), though clearly there are a variety of other criteria that could also be employed.[9]

---

[9]In fact, it turns out that once we have trimmed the model to a smaller sample, using equal weights give very similar results to using weights based on RMSFE

To sum up, SAM refers to the forecasts from the different models that are weighted by how accurately they have forecasted economic developments one to four quarters ahead. The weights of the different models are calculated as an average of the weights throughout the evaluation period 1999Q2-2008Q1. The models that provide the best forecasts since 1999 are assigned the highest weight. For further details we refer to Section 2 of the forecasting report attached in the appendix.

The period used for the out-of-sample evaluation roughly coincides with the shift towards inflation targeting in Norway. This evaluation period is neither exceptionally long nor exceptionally short. It is an open question whether using a different period would have a marked impact on the forecast performance. It is also well known that estimating model weights can be difficult, and altering the sample period will undoubtedly alter the weights that would be attached to different models. One could make a case that one should be more concerned about the recent forecasting performance, but schemes that discount data from the distant past have not generally been very successful (Timmerman 2006).

A key element of the SAM programs is that it manages the input of data and the output of results as gracefully as possibly, with a minimum of intervention from the model users. The Norges Bank uses Fame as its primary data warehouse, and software has been written for Gauss and Matlab that can download data from the database automatically, seamlessly integrating the two. There is also scope to store the output from SAM back in a Fame database, so that further analysis or evaluation can be conducted at a later date.

The presentation of output has also been automated. Reports and presentations for the policy rounds are produced in pdf format using LaTeX and charts are automatically incorporated into these documents, as shown in the report attached in the appendix.

3.1 Models

To provide the forecasts, we have developed a series of new models. The models have been developed to produce accurate forecasts up to one year ahead. A brief description of the estimated, empirical models included in SAM is provided below. For more details on the models refer to the appendix of the attached report.

The following types of models are included in SAM:

1. *AutoRegressive Integrated Moving Average (ARIMA) models.* These models use historical variations in a single time series to provide forecasts. The forecasts can be regarded as an advanced form of trend extension.

2. *Monthly Indicator (MI) models.* The monthly indicator models predict GDP based on many monthly indicators (that are averaged up to a quar-

terly frequency). We use variables such as manufacturing production, employment, retail sales, hotel occupancy rates and building starts.

3. *Factor models.* Factor models are estimated using large data sets. Based on correlation between the different variables, the data sets are reduced to a few common factors. These factors are then used in various equations to provide forecasts of economic developments. We have different models (denoted *FM* and *FQ* respectively) using monthly and quarterly data-sets.[10]

4. *Bivariate Vector AutoRegressive (2VAR) models.* The 2VARs are based on statistical relationships between GDP and a series of quarterly indicator variables. We estimate a 2VAR for each explanatory variable.

5. *VAR with money (VARm) models.* The VARm models predict inflation using monetary aggregates as explanatory variables.

6. *Vector AutoRegressive (VAR) models.* The VAR models are based on statistical relationships between GDP, interest rates and inflation. These tri-variate models take into account that there may be co-movement between these variables. All the variables are a function of lagged values of itself and the other variables.

7. *Bayesian VAR (BVAR models).* The BVARs impose a priori views that influence the parameter values. These 'priors' can be based on data or insight from economic theory. Such a priori information can reduce the uncertainty surrounding the parameter estimates and these models have often proved to have better forecasting properties than the classic VAR models. In addition to GDP, interest rates and inflation, the BVAR models also include exchange rates, oil prices and a number of country-specific variables.

8. *Term structure models (Tstruc).* Tstruc models use information about the term structure of interest rates in the Norwegian money market and in the interest rate swap market in order to forecast GDP. We look at the difference between short-term interest rates (3-month) and long-term rates (up to 5 years).

9. *Error correction model (EMOD).* EMOD contains econometric equations for prices, wages, the interest rate, the krone exchange rate, productivity and mainland GDP.[11]

10. *Dynamic stochastic general equilibrium model (DSGE).* The DSGE model is a New Keynesian small open economy model. A version of NEMO is documented in Staff Memo 2006/6. The DSGE model is estimated and embodies the most important channels between monetary policy and key macroeconomic variables.

---

[10] Aastveit and Trovik (2007) describe the factor model based on monthly data (FM) for projecting developments in mainland GDP.

[11] The model is documented in Akram (2008).

Within each model class, there are several variants with different specifications. In total, SAM contains about 80-100 models. The combination forecasts are a weighted average of the forecasts from about 10 percent of the models.

3.2   Short term forecasting report

The forecasts and their derivation are described in the report attached in the appendix. The report consists of six sections. Section 1 provides a brief summary of the results. The SAM forecasts for core inflation (measured as CPIATE, year on year growth rates) and Mainland GDP (year on year growth rates) for the next four quarters are shown in Table 1.

Section 2 describes the procedure used to derive the forecast. Figures 1 and 2 then graph respectively the forecast for CPIATE and Mainland GDP. The left frames of these figures show the most recent SAM forecast (made at time t) while the right frames compares the most recent SAM forecast with previous forecasts (made at time $t - 1$, $t - 2$, etc.). Compared to previous forecasts, inflation forecast have been revised up slightly since last period $(t - 1)$, but slightly down compared to the forecast made two periods ago $(t - 2)$. On average, though, they are unchanged. For Mainland GDP, the forecasts have pointed to a slowdown for some periods now, but the last two forecast $(t - 1)$, and $(t - 2)$ seem to have captured the downturn quite accurately.

Section 3 of the appendix then looks in depth into the inflation forecast. Figure 3 shows the forecast for inflation based on SAM (red line) together with the forecast from the eight best models at each horizon.[12] The figures emphasize that the quarterly factor model (FQ) and VAR4[13] are among the best model for all horizons. The DSGE model turns out to be among the eight best models in predicting inflation at a 3 quarter horizon. The figures also illustrate that the spread of the inflation forecasts increases with the different horizons, indicating increased uncertainty about the more distant future. For example, for horizon 1 (2008Q2), the eight best models forecast inflation to be in the range $2.2 - 2.4$ percent yoy, while for horizon 4 (2009Q1) inflation is forecast to be between 1.8 and 3.4. Note that the composition of models may vary from horizon to horizon.

The uncertainty of the forecasts is illustrated further in Figure 4 and 5, where we have calculated model-based forecast densities of possible forecast outcomes detailing the probabilities of future events. Rather than looking at the probability bands around the forecast, we investigate the probability that the forecast of inflation will exceed 2.5 percent (the mandated inflation target). Figure 4 and 5 graph actual inflation (blue) together with the probabilities based on the forecast made two and four quarters before the actual data was published. Hence,

---

[12]These models are picked based on RMSFE. The model with the lowest RMSFE is ordered left in each frames.

[13]VAR4 is a VAR model consisting of GDP, inflation and interest rates, estimated using four lags.

the forecast at the end of the period represent the most recent out of sample forecast. We report probabilities calculated by two different density combination weights, the log score and the RMSFE. (The log score weights are discussed in more depth in the next section.) Note that these probabilities are based on the information from the VAR models only, as we have only considered density combinations from these models so far. The figures emphasize that the probability that inflation will exceed the target in two quarters (2008Q3) is relatively high (0.6), while in four quarters (2009Q1) the probability is just below 0.5.

Figures 6-8 detail the forecasts made at different points in time, which underlies the out-of-sample forecast just presented. In particular, Figure 6 shows what the forecast for inflation would have been one to four quarters ahead if the short-term projections had been based on SAM in the period 1999Q2 - 2008Q1. In constructing these forecast, we use the eight best performing models, as measured by RMSFE.

Figure 7 graphs the standard deviations for SAM forecast errors compared to some of the individual models at the one-to four-quarter horizon for inflation. SAM has substantially lower standard deviations than individual models for all horizons. This shows the benefit of averaging the forecasts from several models. However, Figure 7 shows that the standard deviations of the factor model are almost as low as those of SAM. Factor models contain a very large number of explanatory variables, and the common factors can be interpreted as an averaging of large amounts of information. Finally, Figure 8 graphs the bias (mean forecast error) for inflation for all four horizons. Compared to SAM, most models have a positive bias, indicating that they have, on average, over-predicted inflation somewhat during the evaluation period.

In Section 4, the analysis is repeated for GDP. That is, Figure 9 shows the forecast for Mainland GDP based on SAM (red line) together with the forecast from the eight best models at each horizon. According to the SAM forecast, growth in mainland GDP slows later in the year, but picks up again somewhat in 2009Q1. Due to a relatively large amount of models used for predicting GDP, we have, for simplicity, averaged the best performing VAR models (denoted *VAR-SAM*). Forecasts from VAR-SAM, term structure models, and the BVARs turn out to be among the best performing models. Again, the spread in the forecasts is greater at longer horizons, reflecting increased uncertainty at longer horizons. The uncertainty of the forecast are illustrated further in Figures 10 and 11, where we investigate the probability of GDP-growth falling below 2.5 percent. These probabilities have increased in recent quarters, but from very low levels.

Finally, Figure 11 shows SAM-based historical forecasts for mainland GDP growth one to four quarters ahead together with actual developments. RMSFE and bias comparing some selected models with SAM are graphed in Figures 14 and 15 respectively. Again, SAM has substantially lower standard deviations than all individual models, except VAR-SAM at horizon three and four. However, note that VAR-SAM is already an average of VAR models. Regarding the bias, SAM has bias closer to zero than many individual models. However, most

models, including SAM, have under-predicted GDP somewhat in the evaluation period.

Section 5 provides more details and technical comments. First we provide a comparison of the SAM forecast with Norges Bank's official short term forecasts over the period 1999Q1-2008Q1 in Table 4. The results emphasize that SAM forecast outperforms Norges Banks official short term forecast in terms of RMSFE. The gain increases with the horizons and illustrates the usefulness in averaging short term forecasts.

We then look in more detail into the performance of the individual forecasting models in figures 15-24. Figures 19 and 24 require somewhat more explanation. They show, respectively, the performance of the best forecasting models over time for CPI and GDP. The graphs are 3-dimensional and should be read as follows. Along the horizontal axis we display the name of the eight best forecasting models, with the model ranked best (number one) placed to the far right. Moving left we find the name of the model that is ranked as number two etc. The ranking is based on the average of the whole forecasting period; 1999-2008. Moving back into the graph, we investigate how these eight models have performed in the past. That is, we ask to what extent the models that are ranked from one to eight would be ranked similarly had the forecasting period ended in, say 2001 instead of 2008? If one observes a flat (visualized as blue) landscape, the models have been ranked among the best models at all time. A 'mountain' (visualized as red), on the other hand, indicates that the models have been ranked worse earlier in the period (that is, they are ranked with a higher number, as can be seen along the vertical axis).

Finally, Section 6 gives a detailed description of the models used in SAM in this policy round.

Overall, the results using SAM indicate that there are clear advantages to averaging forecasts from several individual models when making predictions in the short term. In addition, the development of new types of models has illustrated more clearly the uncertainty around the central projections. In section 4 we explore this issue in more detail by among other discussing how one can average forecast based on uncertainty of point forecasts (density combinations). We emphasize that this is work in progress.

4   Work in Progress

The Norges Bank's Nowcasting project has a number of areas in which it needs to be extended: 1) the development of density forecasts; 2) evaluation of the models using a real-time data set; and 3) an expansion of SAM to forecast other variables of interest to Norges Bank (such as consumption, investment and components of inflation).

Providing density forecasts is at the core of the Nowcasting project's third objective: providing a better characterisation of the uncertainty that surrounds the Norges Bank's short term forecasts.[14] With a characterisation of the probability density function at hand, a decision maker is better able to evaluate any expected 'loss', irrespective of the shape of the decision maker's loss function. Thus, there is no longer any need to concentrate solely on quadratic loss and mean squared forecast errors to evaluate forecasts.

4.1 Evaluation of densities

Two of the most natural ways of evaluating density forecasts are by considering their log score or indeed simply their score, where the score is the probability that a particular realisation would have arisen given the density function specified by a model or some combination of models. As noted by Hall and Mitchell (2007), maximising the log score amounts to minimising the Kullback-Leibler distance between the models and the true but unknown density. A highly convenient feature of this minimisation is that the true density does not need to be known or even inferred.

As discussed in Hoeting, Madigan, Raftery, and Volinsky (1999), the log score is a combined measure of bias and 'calibration'. The former reflects whether a probability density function is correctly centred, whereas calibration refers to the spread of the distribution. Distributions that spread probability mass broadly will have low probability density function values over the entire support. Conversely, a density that is not very dispersed may also have a low value at a given realisation of a random variable if the probability mass is centred on the wrong location. The log score thus reflects both the degree of bias and the degree of calibration. Ideally we wish to have densities that concentrate probability mass on the correct location.

4.2 Combination of densities

Genest and Zidek (1986) summarize the literature up to the mid-1980s on the combination of a group of experts opinions, where these opinions are cast in probabilistic terms. Many interesting issues are picked up in this article and in the comments from the discussants, some of quite a philosophical nature about the meaning of probability and the possibility for group comparisons of such probabilities. In a situation where a group of experts (models in our context) are reporting to an external decision-maker (e.g. the policy-makers), Genest and Zidek suggest that a Bayesian framework "seems to be the only acceptable methods for accumulating the information that lies behind consultants'

---

[14]Wallis (2005) notes that combining forecast *intervals* will not generally lead to intervals with the desired probability. This problem is circumvented by calculating intervals from the combined densities.

conflicting advice." Much of the article is concerned with assessing whether various combination schemes satisfy axioms; these axioms generally relate to whether common properties of the individual densities also aggregate up into the combination density.

The so-called linear opinion pool (or finite mixture density) is perhaps the easiest way to combine density forecasts. At each point in the support of the distribution, the linear opinion pool density is simply a weighted average of the various individual densities, i.e.

$$p(y_t) = \sum_{i=1}^{N} w_i p_i(y_t) \tag{1}$$

where $p(y_t)$ is the combined density, w_i is the weight of the ith density, and $p_i(y_t)$ is one of N individual densities that we wish to combine. Provided the weights are non-negative and sum to one the combined density will be everywhere non-negative and will integrate to one, and will thus be a proper density. Hall and Mitchell (2008) note that, although the combination of forecasts will ex post improve the accuracy of forecasts, a combined density typically has increased variance relative to the average variance from the individual densities. Given that bias and calibration are both important for density forecasts, as discussed above, the increased variance in the combined density may actually be a positive feature: if two densities had the same variances but were centred on different locations then one might reasonably want the combined density to exhibit larger variance than the individual densities.

The interesting part of the linear opinion pool is considering how to weight the individual densities. Simply equally weighting all densities is clearly feasible, though if there are poor models in the set under consideration this may adversely affect the properties of the density combination. Implicitly, Bayesian model averaging uses the linear opinion pool in combining densities from different models, with the weights attached to the densities corresponding to the posterior probabilities of the models. These probabilities can be thought of as the probability that a model from the set under consideration has the smallest Kullback-Leibler divergence among the set of models. These weights will vary over time as new data cause the probabilities to be revised.

Hall and Mitchell (2007) suggest optimising the weights of the combined density to minimise the Kullback-Leibler distance. This is equivalent to maximising the likelihood with respect to the weights, treating them as additional parameters. (Hall and Mitchell liken this density combination to a Markov-switching model with an unobserved state dictating which density is operative.) An advantage of this approach is that the true unknown density is a constant in the maximisation problem and can be ignored.[15]

---

[15] Hall and Mitchell also consider similar alternatives based on goodness of fit tests and using assumptions about the functional form of the true unknown density.

In essence, Hall and Mitchell suggest choosing the weights $w*$:

$$w* \quad = \quad \arg\max_{w} \frac{1}{T} \sum_{t=1}^{T} \ln(\sum_{i=1}^{N} w_i P_{it}(y_t)) \tag{2}$$

$$w = (w_i, w_2, ..., w_N)'$$

Here there are T periods in the evaluation period; in our analysis T grows as the evaluation period is recursively extended. $p_{it}(y_t)$ is the score of the data realisation y_t from the evaluation period for the ith density, where there are N such densities. The inner summation thus corresponds to a candidate combined density, from a linear opinion pool with weights w. To reiterate, the combination density is a linear combination of the *unlogged* individual densities, but the weights are optimised with respect to the *log* score.

This evaluation presumes that the data are independent and identically distributed. Given that our forecasts are for multi-step forecasts that overlap the dependence in the forecasts should probably be taken into account when estimating the weights.

A full Bayesian analysis of the kind outlined by Bordley (1982) imposes great demands, in terms of knowing the full joint distribution of the variable of interest conditional on the forecasts from the experts. DeGroot and Mortera (1991) consider optimising the weights of a linear opinion pool in a Bayesian framework with an explicit loss function. These weights depend only on the marginal and pairwise distributions of the forecasts. Both the full Bayesian analysis, and indeed even the work of DeGroot and Mortera, illustrates that ideally one needs to understand whether or not the forecasts are independent. There are also substantive issues surrounding 'groups' of forecasts, which is akin to concerns about whether priors should be completely independent across models or whether similar models should in some sense 'share' in some common allocation of prior probability.

The linear opinion pool is by no means the only possible means of combining densities. Timmerman (2006) notes that it is possibly to combine densities geometrically. A logarithmic opinion pool can be represented as:

$$p(y_t) = k\Pi_{i=1}^{N} p_i(y_t)^{w_i} \tag{3}$$

where k is a normalising constant to ensure the density integrates to one. Interestingly, geometric combination typically results in a unimodal distribution and is less dispersed than the linear opinion pool (Genest and Zidek 1986), which may or may not be desirable. If normal distributions are combined using the log score then the combined distribution will also be normal. see (Winkler 1981). Again, whether this is a desirable feature is open to question. By way of contrast, using a linear opinion pool may result in a multi-modal mixture distribution. Explaining and justifying multi-modal distributions to policy-makers and the general public may be rather difficult to say the least.

In principle there are two main ways to provide forecast densities: analytically or through simulation. The Nowcasting project has to be flexible enough to incorporate both methods: both the DSGE model and the DSGE-VAR are estimated by simulation methods and thus use Markov chains to provide forecast distributions. In contrast, densities from most of the VARs are derived under assumptions of normality. Since the simulated distributions are unlikely to be normally distributed, the combined density forecast will also likely be non-normal, possibly skewed or fat-tailed.

It should be noted that the combined density, even with weights chosen optimally according to the Kullback-Leibler divergence, will not necessarily improve markedly upon the individual densities. The empirical example in Hall and Mitchell (2007) shows that combining the National Institute of Economic and Social Research's density forecast of inflation with the Bank of England's density forecast is only marginally better, in terms of recursive out-of-sample log score, than the Bank of England's density by itself. Once again, the evaluation of the densities needs to be performed before such conclusions can be drawn, and thus the *evaluation* of forecasting models is an important motivation for the Norges Bank's own short term forecasting project. Precisely how forecast densities should be combined remains, however, an open question.

5 Conclusion

Combination methods have gained grounds in the forecast literature. There is by now a body of empirical evidence suggesting that forecast combinations produce better forecasts on average than alternative forecasts from a single model. This memo has added further evidence to this conclusion. By developing a System for Averaging Models (SAM), we have shown that there there are clear advantages to averaging forecasts from several individual models when predicting GDP growth and inflation in Norway in the short term (up to a year).

Prior to combining these forecasts, we have, with the aid of colleagues in Norges Bank, developed a series of new forecasting models including factor models, term structure models, bayesian vector autoregressive models, equilibrium correction models and dynamic stochastic general equilibrium models, to name a few. The individual and combined forecasts are collected and discussed in an attached forecasting report. The report illustrates the information that is currently produced and presented to the monetary policy decision-makers.

In evaluating and combining the models we have focused on point forecasts. The suite of models developed illustrates the uncertainty around the central projections. So far, densities are not routinely provided. We have, however, illustrated the uncertainty of the forecasts from the VAR models by calculating model-based forecast densities of possible forecast outcomes. Rather than looking at the probability bands around the forecast, we investigate the probability

of a future events (like what is the probability that the forecast of inflation will exceed the mandated inflation target). In our work ahead, we will work on averaging models based on both point forecasts and density combinations. Based on the experience so far, we will also consider whether other methods of averaging forecasts can provide more precise forecasts.

References

Aastveit, K. A. and T. G. Trovik (2007). Nowcasting Norwegian GDP: The role of asset prices in a small open economy. Working Paper 2007/9, Norges Bank.

Akram, F. (2008). An econometric model of mainland norway: Emod. Mimeo, Norges Bank.

Andersson, M., G. Karlsson, and J. Svensson (2007). An evaluation of the Riksbank's forecasting performance. *Sveriges Riksbank Economic Review 3*, 59–75.

Andersson, M. and S. Karlsson (2007). Bayesian forecast combination for VAR models. *Riksbank Research Paper 216*.

Andersson, M. and M. Löf (2007). The Riksbank's new indicator procedures. *Sveriges Riksbank Economic Review 1*, 76–95.

Bancroft, T. (1944). On biases in estimation due to the use of preliminary tests of significance. *The Annals of Mathematical Statistics 15*(2), 190–204.

Bates, J. M. and C. W. J. Granger (1969). The combination of forecasts. *Operational Research Quarterly 20*(4), 451–468.

Bock, M. E., T. A. Yancey, and G. G. Judge (1973). The statistical consequences of preliminary test estimators in regression. *Journal of the American Statistical Association 68*(341), 109–116.

Bordley, R. F. (1982). The combination of forecasts: A Bayesian approach. *The Journal of the Operational Research Society 33*(2), 171–174.

Brubakk, L., T. A. Husebø, J. Maih, K. Olsen, and M. Østnor (2006). Finding nemo: Documentation of the norwegian economy model. Staff memo 2006/6, Norges Bank.

Burnham, K. P. and D. R. Anderson (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods and Research 33*(2), 261–304.

Campos, J., N. R. Ericsson, and D. F. Hendry (Eds.) (2005). *General-to-Specific Modelling*. Cheltenham: Edward Elgar.

Clark, T. E. and M. W. McCracken (2008). Averaging forecasts from vars with uncertain instabilities. *Journal of Applied Econometrics*. (Fortcoming).

Clements, M. P. (2004). Evaluating the Bank of England density forecasts of inflation. *Economic Journal 114*(498), 844–866.

Coletti, D. and S. Murchison (2002). Models in policy-making. *Bank of Canada Review*, 19–26.

DeGroot, M. H. and J. Mortera (1991). Optimal linear opinion pools. *Management Science 37*(5), 546–558.

Diebold, F. X. and J. Lopez (1996). Forecast evaluation and combination. In G. Maddala and C. Rao (Eds.), *Statistical Methods in Finance, Handbook of Statistics*, Volume 14, pp. 241–268. Amsterdam: Elsevier.

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

Elder, R., G. Kapetanios, T. Taylor, and T. Yates (2005). Assessing the MPC's fan charts. *Bank of England Quarterly Bulletin 45*(3), 326–348.

Fernandez, C., E. Ley, and M. F. J. Steel (2001, September-October). Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics 16*(5), 563–76.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2003). *Bayesian Data Analysis* (2nd ed.). Boca Raton, Fl.: Chapman and Hall/CRC.

Genest, C. and J. V. Zidek (1986). Combining probability distributions: A critique and an annotated bibliography. *Statistical Science 1*(1), 114–148.

Gerard, H. and C. Nimark (2008). Combining multivariate density forecasts using predictive criteria. Research Discussion Paper 2008-02, Reserve Bank of Australia.

Hall, S. G. and J. Mitchell (2007). Combining density forecasts. *International Journal of Forecasting 23*, 1–13.

Hall, S. G. and J. Mitchell (2008). Recent developments in density forecasting. In T. Mills and K. Patterson (Eds.), *Palgrave Handbook of Econometrics*, Volume II. Basingstoke, Hampshire: Palgrave Macmillan. forthcoming.

Hansen, B. D. (2008). Least squares forecast averaging. *Journal of Econometrics*. forthcoming.

Hendry, D. F. and M. P. Clements (2002). Pooling of forecasts. *Econometrics Journal 5*, 1–26.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky (1999). Bayesian model averaging: A tutorial. *Statistical Science 14*(4), 382–417.

Huang, H. and T.-H. Lee (2008). To combine forecasts or to combine information? Mimeo, University of California Riverside.

Kapetanios, G., V. Labhard, and S. Price (2005). Forecasting using Bayesian and information theoretic averaging: An application to UK inflation. Technical Report 268, Bank of England.

Kapetanios, G., V. Labhard, and S. Price (2008). Forecast combination and the Bank of England's suite of statistical forecasting models. *Economic Modelling 25*, 772–792.

Kass, R. E. and A. E. Raftery (1995). Bayes factors. *Journal of the American Statistical Association 90*(430), 773–795.

Kobayashi, M. and S. Sakata (1990). Mallows' c_p criterion and unbiasedness of model selection. *Journal of Econometrics 45*, 385–395.

Koop, G. and S. Potter (2004). Forecasting in dynamic factor models using Bayesian model averaging. *Econometrics Journal 7*(2), 550–65.

Makridakis, S. and M. Hibon (2000). The M3 competition: Results, conclusions, and implications. *International Journal of Forecasting 16*, 451–476.

Marcellino, M. (2002). Forecast pooling for short time series of macroeconomic variables. Working Papers 212, IGIER (Innocenzo Gasparini Institute for Economic Research), Bocconi University.

Raftery, A. E. (1995). Bayesian model selection in social research. In P. V. Marsden (Ed.), *Sociological Methodology 1995*. Oxford: Basil Blackwell.

Raftery, A. E. and Y. Zheng (2003, December). Frequentist model average estimators: Discussion: Performance of Bayesian model averaging. *Journal of the American Statistical Association 98*(464), 931–38.

Stock, J. H. and M. W. Watson (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting 23*, 405–430.

Timmerman, A. (2006). Forecast combinations. In G. Elliott, C. W. J. Granger, and A. Timmerman (Eds.), *Handbook of Economic Forecasting*, Volume 1. Amsterdam: Elsevier.

Wallis, K. S. (2005). Combining density and interval forecasts: A modest proposal. *Oxford Bulletin of Economics and Statistics 67*, 983–994.

Winkler, R. (1981). Combining probability distributions from dependent information sources. *Management Science 27*, 479–488.

Appendix

As summarised in section 3, the material below illustrates the kind of information provided to senior staff by the 'nowcasting team'. The report contains a mix of forecasting and evaluation information.

Hilde C. Bjørnland

Anne S. Jore

Christie Smith

Leif A. Thorsrud

Notes:
All the forecasts were produced between 3rd. and 5th. of June 2008.

1 Summary

- *NOTE: This is an example of a forecasting report. The forecast referred to as SAM are from a new **S**ystem for **A**veraging **M**odels.*

- Inflation projections using SAM indicate that core inflation (CPIATE, year on year (yoy) growth rates) will increase slightly throughout the year, to 2.3 percent in Q2 2008, 2.5 percent in Q3 2008, and reach 2.75 percent in Q4 2008, where it remains at the year ending in Q1 2009.

- Of the eight best performing models (that SAM is averaging over), all but one model predict that core inflation will exceed 2.5 percent in a year.

- Projections for (mainland) GDP using SAM signal lower growth rates ahead than we have seen the last two years. Growth rates are, however, still predicted to be moderate throughout the year, so that by the end of the year, mainland GDP is expected to increase by 2.6 percent (year on year growth rates). The growth rates are predicted to pick up in Q1 2009, to an annual level of 3.9 by Q1 2009.

- Of the eight best performing models (that SAM is averaging over), five predict that GDP will grow by less than the SAM average (or just at the average) of 3.9 pct. in the year to 2009Q1, while three predict GDP will grow by more than 3.9 pct.

Table 1: Forecasts using SAM

Variable	Q2-2008	Q3-2008	Q4-2008	Q1-2009
Inflation, CPIATE, yoy growth	2.29	2.51	2.76	2.75
	Q2-2008	Q3-2008	Q4-2008	Q1-2009
Mainland GDP, yoy growth	3.9	2.9	2.6	3.9

Note: **SAM**: A System for Averaging Models.

Contents

List of Tables

List of Figures

2   Forecasts

Below we explain how we have constructed the SAM forecast. The purpose of model averaging is to address model uncertainty by averaging across models according to their fit to data. Various studies, including Stock and Watson (2003) and Clark and McCracken (2008), have shown that combining multiple forecasts can lead to increased forecast accuracy, in particular in the presence of uncertain model instabilities.

In constructing the forecast for CPI and GDP, we perform a 'historical evaluation', where we assess the forecast accuracy of various models. Each model gets a weight according to how well it has done it in the past. The forecast for the next year is then a weighted average of the forecasts from the various models. This is explained in detail below:

1. All models are first estimated up to 1999Q1. Forecast are made for 1999Q2 (horizon 1), 1999Q3 (horizon 2), 1999Q4 (horizon 3) and 2000Q1 (horizon 4).

2. The forecasts are stored and compared with actual data. The forecast error and the standard deviation of the forecast error, i.e. the root mean squared forecast error (RMSFE) are computed. This is done for each horizon

3. We extend the estimation one period forward (1999Q2), and redo 1-2 until, ultimately, the final forecast for 2008Q1 are made (which is the last period we have observed data).

4. For *each* horizon, weights are then given to the different models according to the RMSFE. That is, those that have the lowest RMSFE gets the highest weights, while those that have the highest RMSFE gets the lowest weight. We do this for each period, so that for horizon 1, all model have obtained 36 weights (1999Q2-2008Q1), for horizon 2, all models have obtained 35 weights (1999Q3-2008Q1), etc.

5. The forecast from all models are then averaged according to their weights. The weights are updated recursively as we move forward in time and will sum to one in each period.

6. We refer to the forecast from the weighted average as SAM.

7. For the forecasts out of sample, we multiply the individual model forecast with their last observed weights (which is is an average of the whole historical evaluation period).

8. The forecast based on SAM is an average of the eight best performing models (10 percent of all models).

Figure 1: *Core inflation forecast*

SAM

SAM at different times
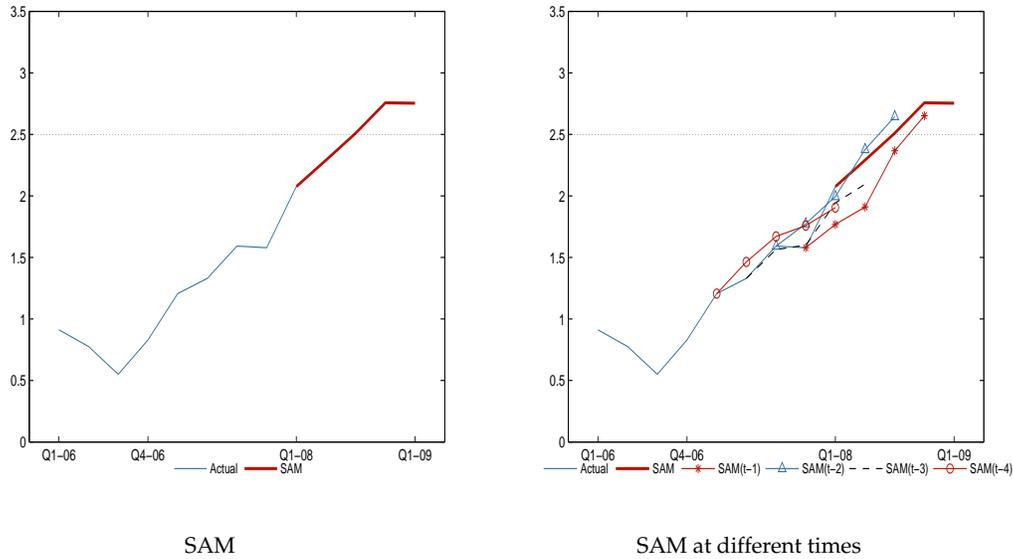
Figure 2: *Mainland GDP growth forecast*

SAM

SAM at different times
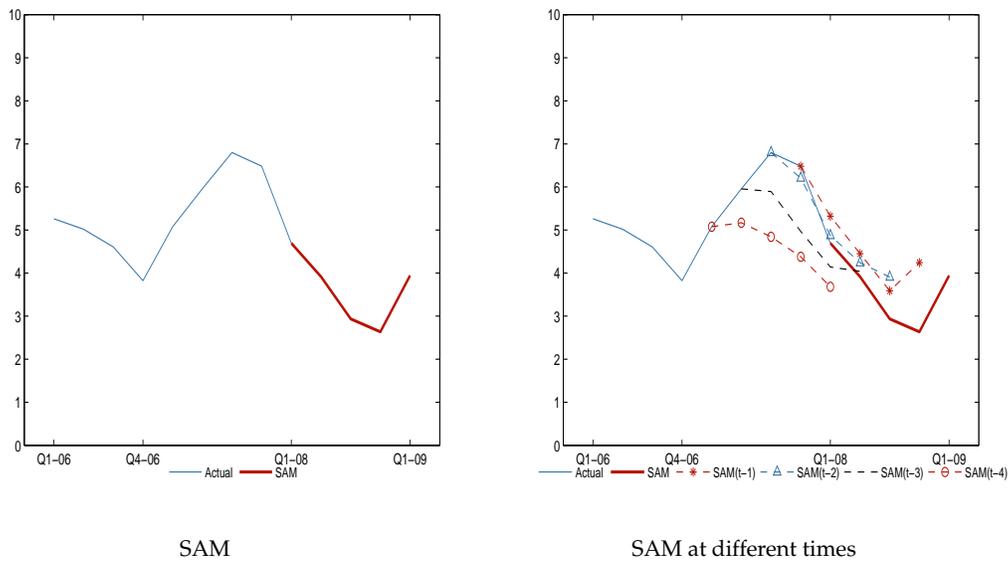
Note: Figure 1 graphs core inflation measured as CPIATE, year on year (yoy) growth rates, together with the forecast based on SAM. The left frame shows the most recent SAM forecast (red) together with actual inflation (blue). The right frame compares the most recent forecast from SAM with previous forecasts, where SAM (*t-i*) is the forecast path *i* periods before the current SAM forecast. Figure 2 shows the yoy growth rates in mainland GDP, (s.a. data), together with the forecast based on SAM. Again, the left frame shows the current SAM forecast (red) together with actual GDP growth (blue), while the right frame compares the most recent forecast from SAM with previous forecasts.

3 Inflation in depth

Figure 3: *Inflation forecast: Best performing models*

Horizon 1

Horizon 2
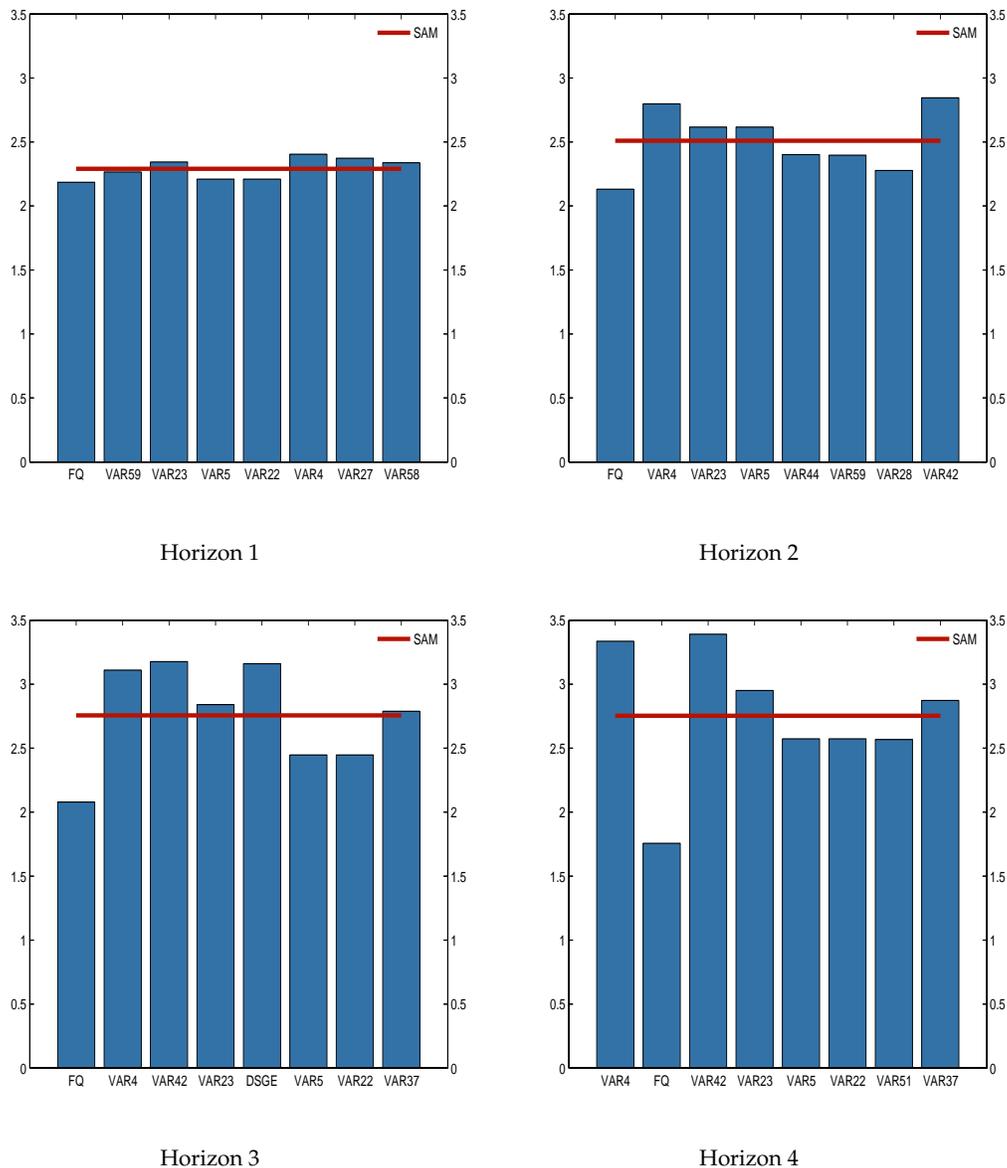
Horizon 3

Horizon 4

Note: The graphs compares SAM forecast for inflation (CPIATE, yoy growth), (red line) with the forecast from the eight best performing models (that SAM is averaging over) at the different horizons.

Figure 4: *What is the probability of inflation exceeding 2.5 percent? Horizon 2*
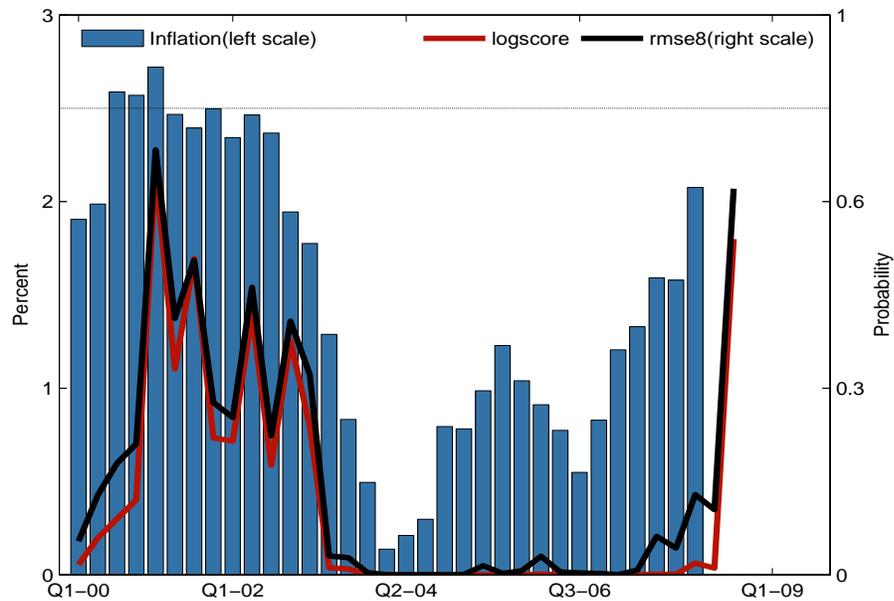
Figure 5: *What is the probability of inflation exceeding 2.5 percent? Horizon 4*
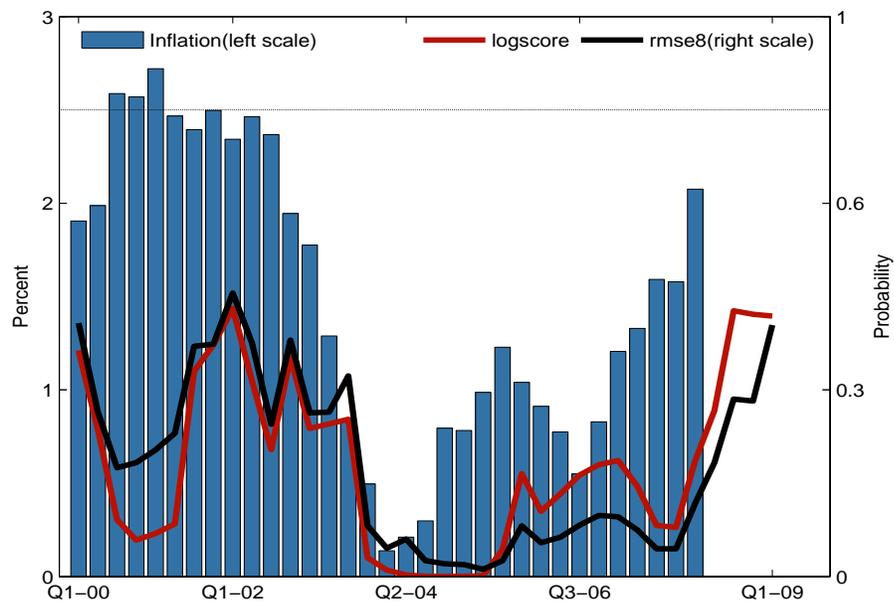
Note: The probabilities are based on on the forecast for inflation (CPIATE, yoy growth) using an average over the VAR models only.

Figure 6: *SAM-inflation forecasts at different points in time*

Horizon 1

Horizon 2

Horizon 3

Horizon 4
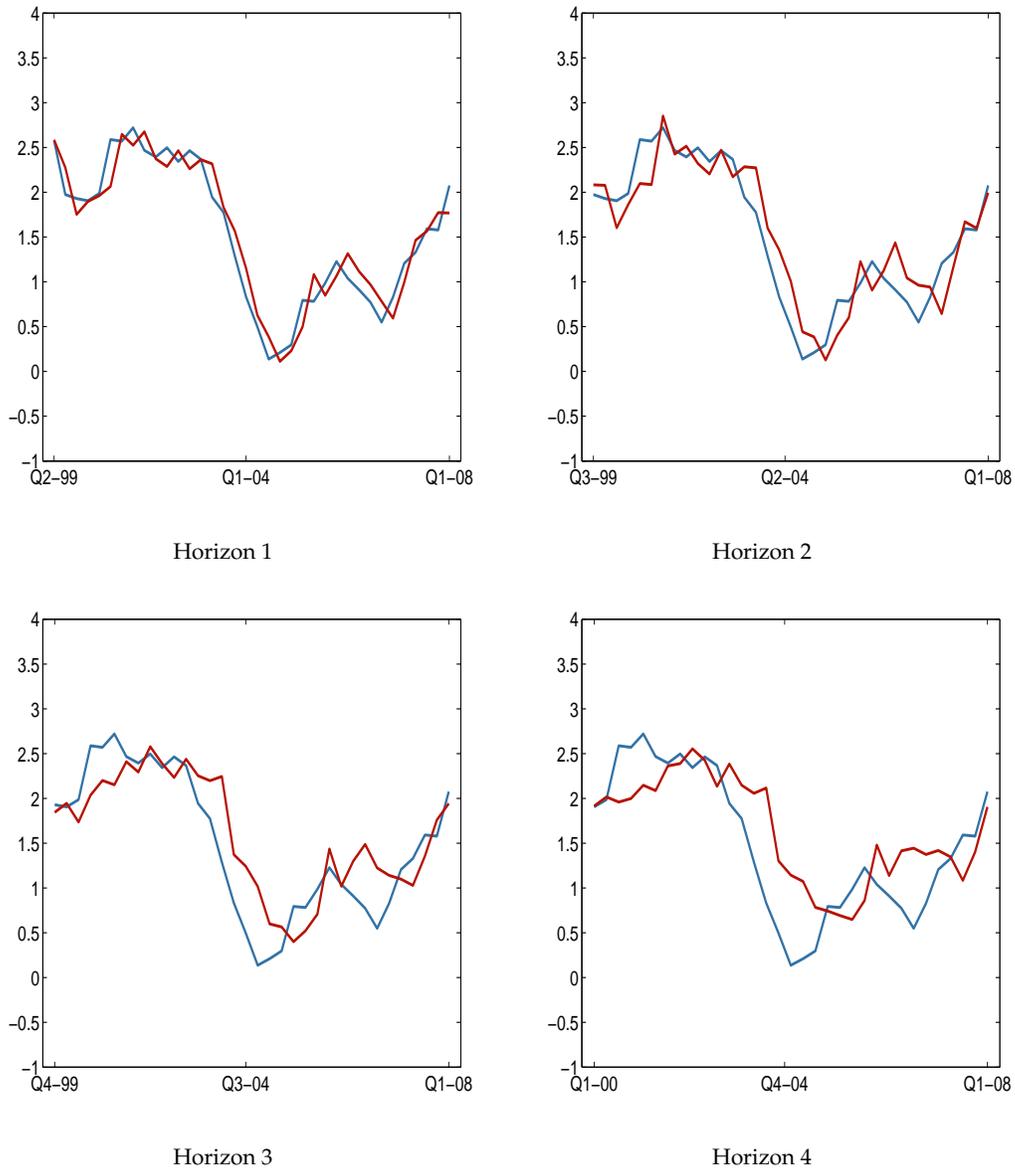
Note: The graphs compares actual inflation (CPIATE, yoy growth, blue line) with SAM forecast for inflation (red line) for the period 1999Q2-2008Q1. SAM forecasts are constructed based on the eight best performing models.

Figure 7: *RMSFE for inflation*

RMSFE
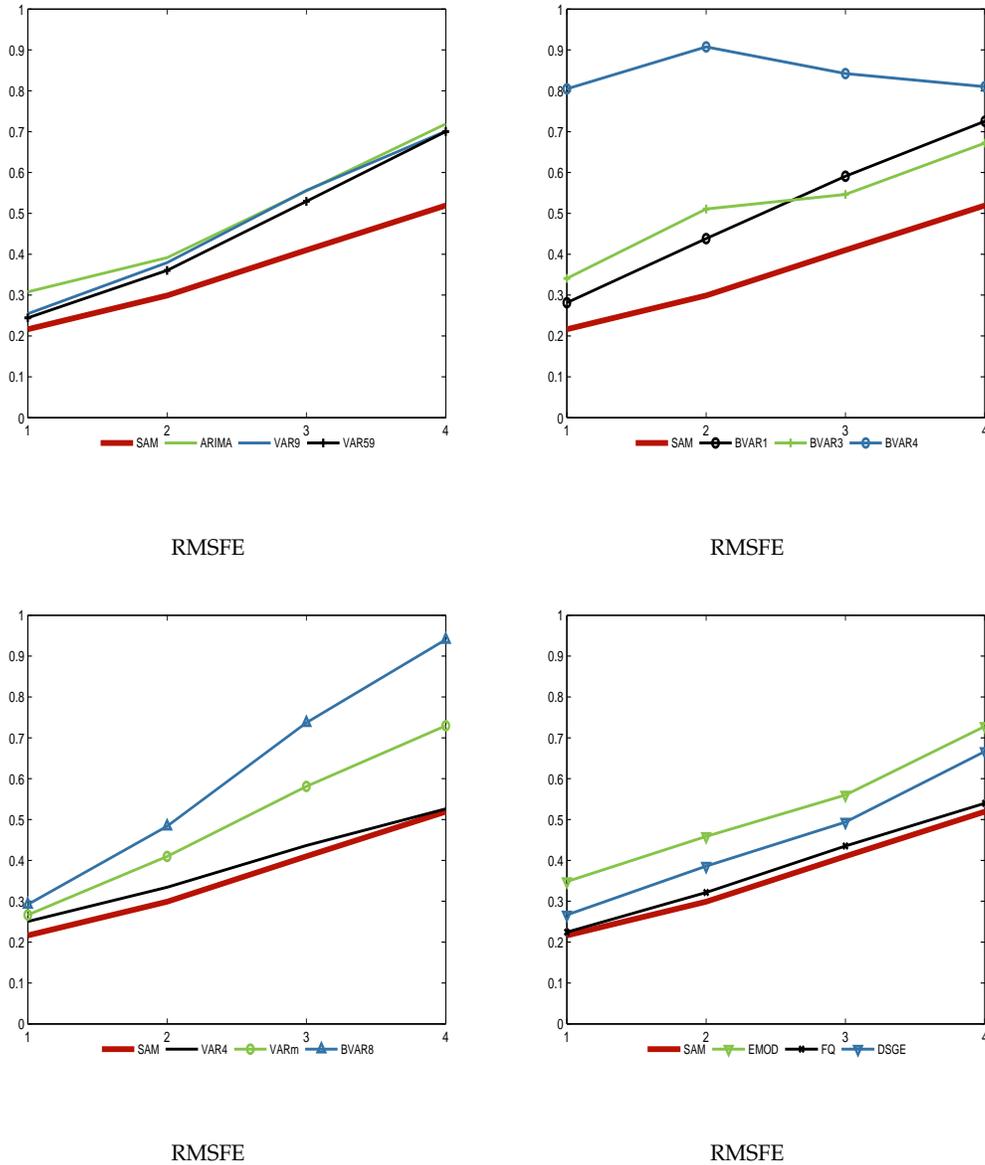
RMSFE

RMSFE

RMSFE

Note: Root mean square forecast error (RMSFE) for SAM forecast (red line) is present in each panel. The horizontal axis give the different forecast horizons.

Figure 8: *Bias for inflation*
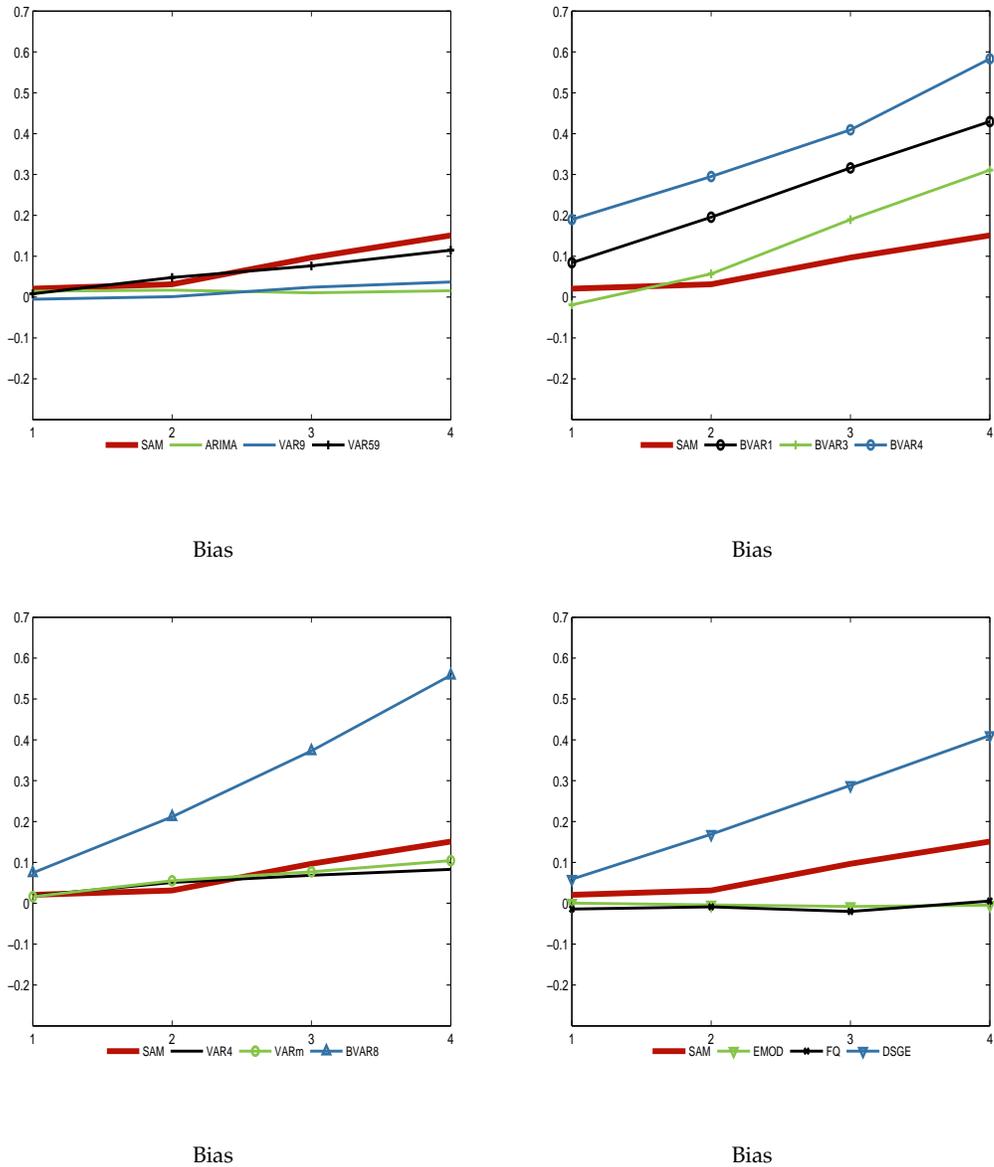
Bias

Bias

Bias

Bias

Note: Bias is the average forecast failure at each horizon. Bias for SAM forecast (red line) is present in each panel. The horizontal axis give the different forecast horizons.

4   GDP in depth

Figure 9: *GDP forecast: Best performing models*

Horizon 1

Horizon 2
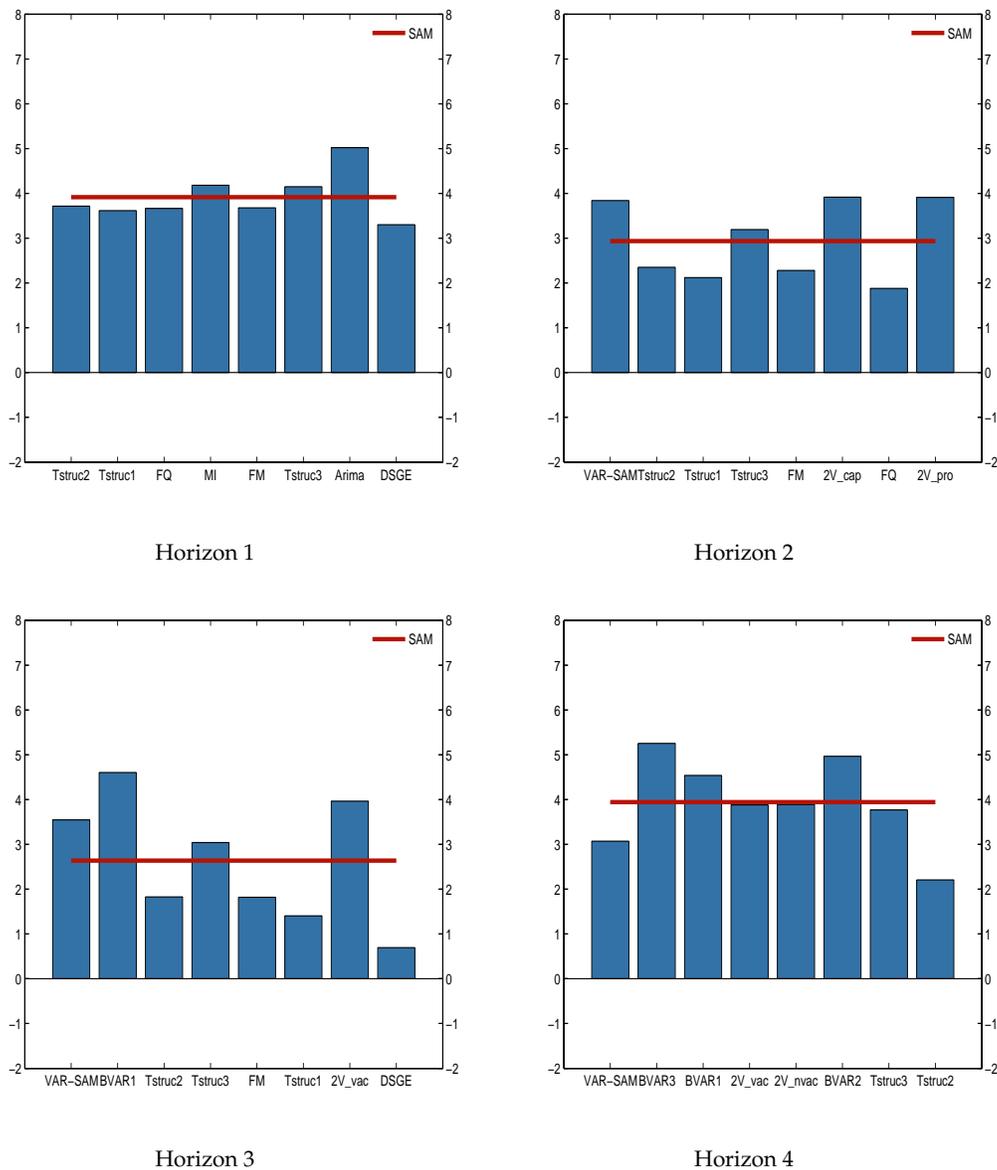
Horizon 3

Horizon 4

Note: The graphs compares SAM forecast for yoy growth rates in mainland GDP, (red line) with the forecast from the eight best performing models (that SAM is averaging over) at the different horizons.

Figure 10: *What is the probability of GDP-growth falling below 2.5 percent: Horizon 2*
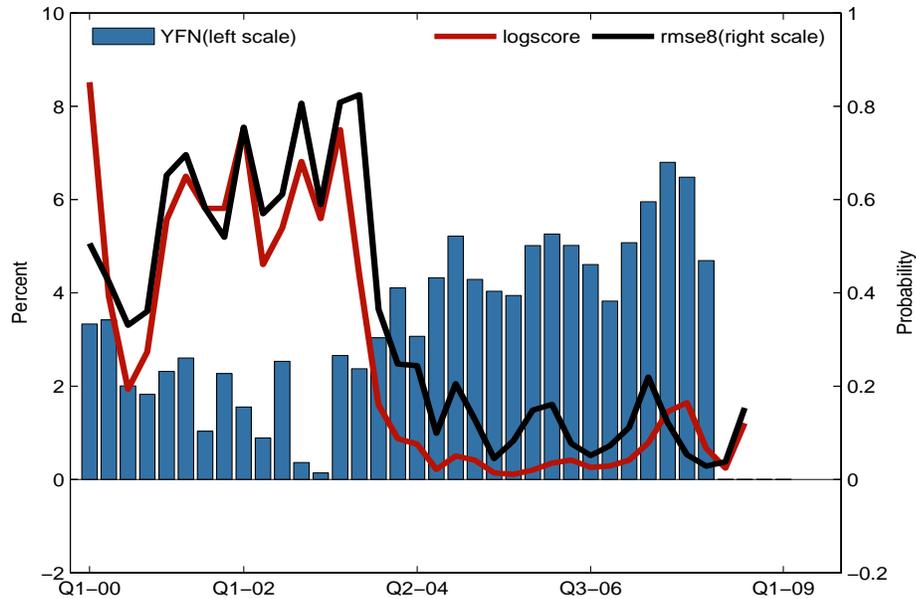
Figure 11: *What is the probability of GDP-growth falling below 2.5 percent: Horizon 4*
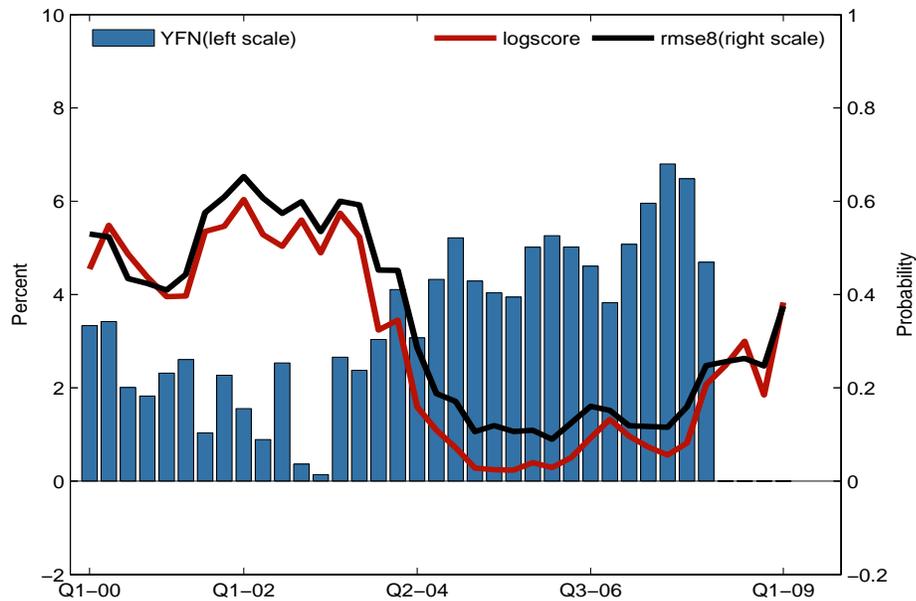
Note: The probabilities are based on the forecast for mainland GDP (yoy changes), using an average over the SAM models.

Figure 12: *SAM-GDP forecasts at different points in time*

Horizon 1

Horizon 2

Horizon 3

Horizon 4
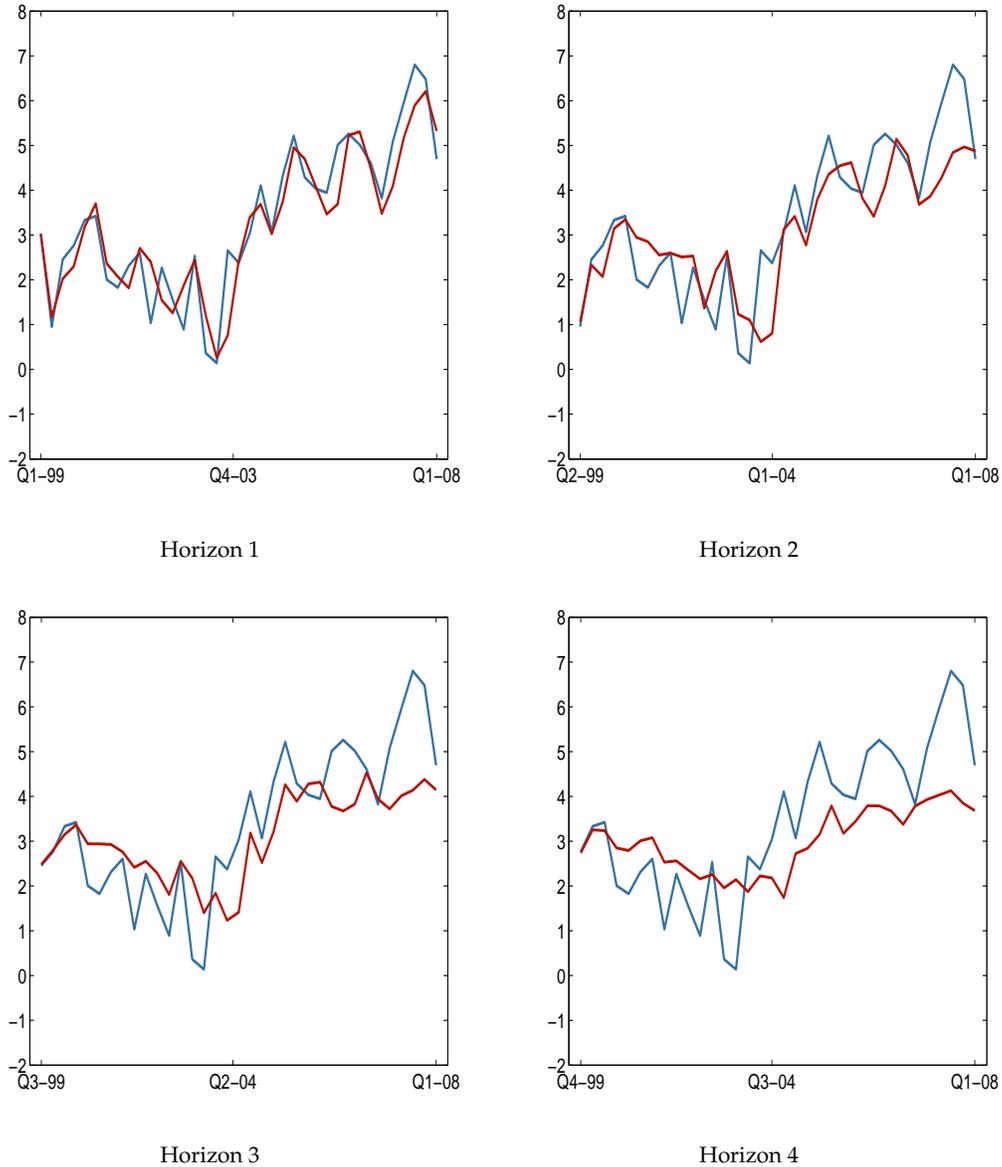
Note: The graphs compares actual yoy growth rates in mainland GDP (blue line) with SAM forecasts for GDP (red line) for the period 1999Q2-2008Q1. SAM forecasts are constructed based on the eight best performing models.

Figure 13: *RMSFE for GDP*
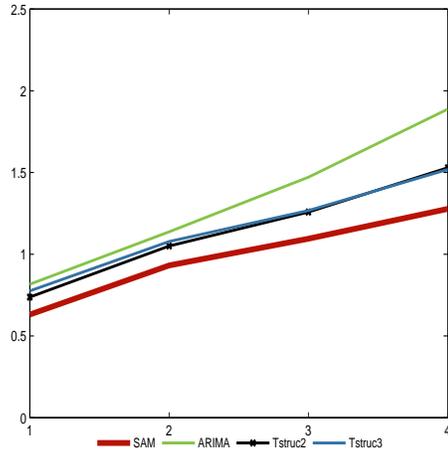
RMSFE
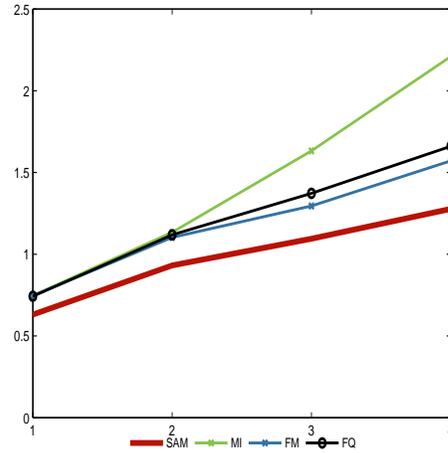
RMSFE
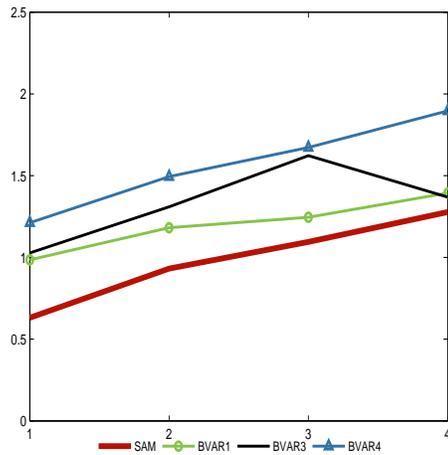
RMSFE
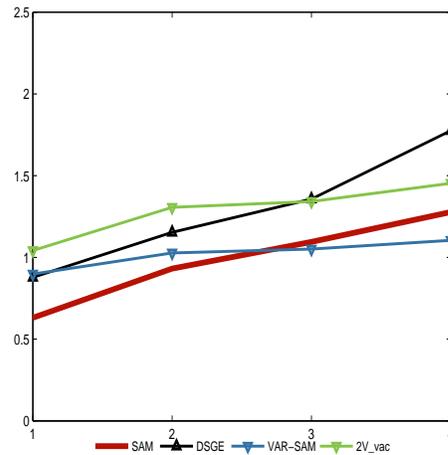
RMSFE

Note: RMSFE for SAM forecast (red line) is present in each panel. The horizontal axis give the different forecast horizons.

Figure 14: *Bias for GDP*
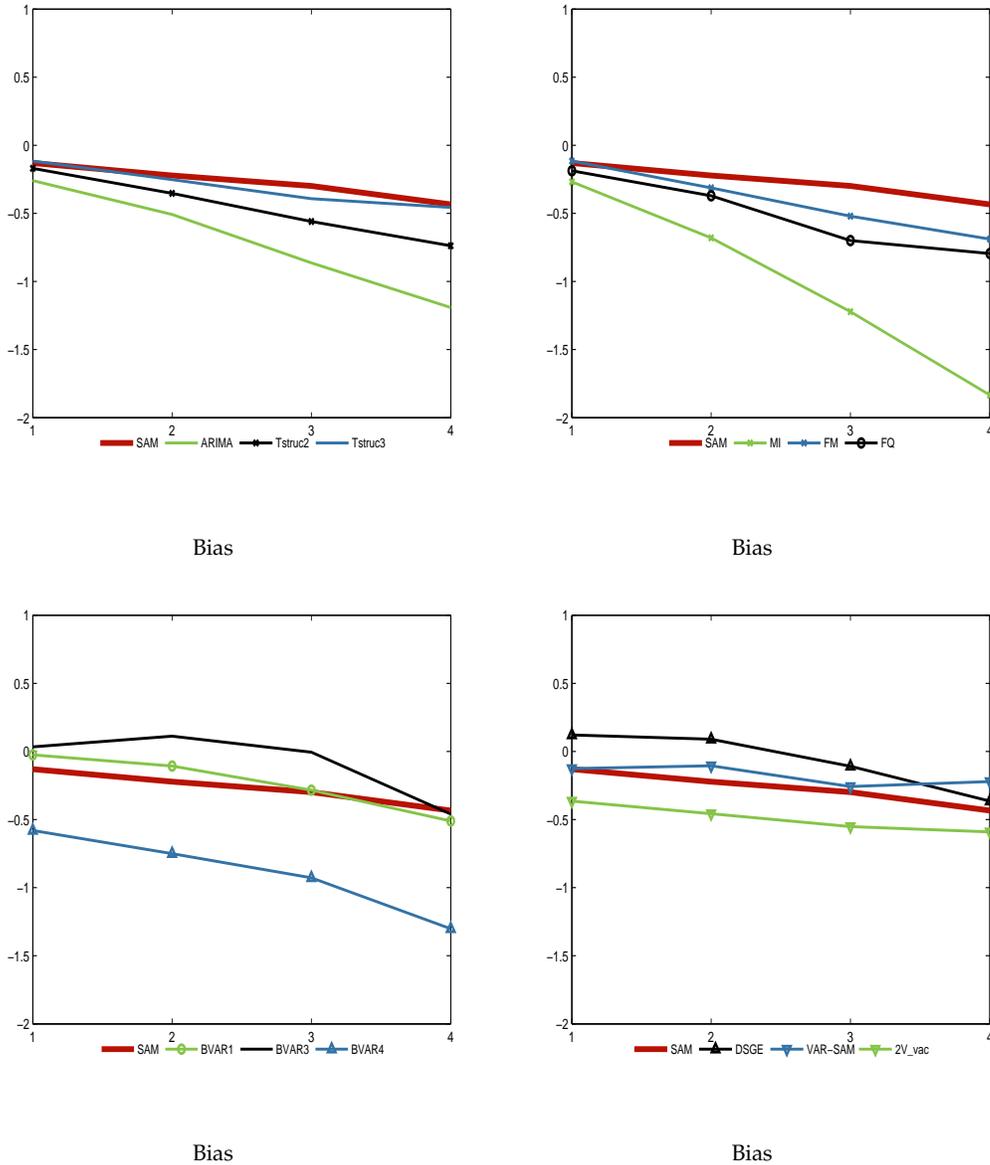
Bias

Bias

Bias

Bias

Note: Bias is the average forecast failure at each horizon. Bias for SAM forecast (red line) is present in each panel. The horizontal axis give the different forecast horizons.

5 Details and technical comments

Details on inflation forecast

Table 2: Inflation forecasts

	Q2.08	Q3.08	Q4.08	Q1.09
SAM	2.29	2.51	2.76	2.75
SAM(t-1)	1.91	2.37	2.65	
SAM(t-2)	2.38	2.64		
SAM(t-3)	2.10			

Note: The maximum SAM forecast horizon is four quarters. SAM (*t-i*) is the forecast path *i* periods before the current SAM forecast.

Table 3: Best models for inflation

	1st	2nd	3rd	4th	5th	6th	7th	8th
Q2.08	FQ	VAR59	VAR23	VAR5	VAR22	VAR4	VAR27	VAR58
val	2.18	2.27	2.34	2.21	2.21	2.40	2.37	2.34
Q3.08	FQ	VAR4	VAR23	VAR5	VAR44	VAR59	VAR28	VAR42
val	2.13	2.80	2.62	2.62	2.40	2.40	2.28	2.84
Q4.08	FQ	VAR4	VAR42	VAR23	DSGE	VAR5	VAR22	VAR37
val	2.08	3.11	3.18	2.84	3.16	2.45	2.45	2.79
Q1.09	VAR4	FQ	VAR42	VAR23	VAR5	VAR22	VAR51	VAR37
val	3.34	1.76	3.39	2.95	2.57	2.57	2.57	2.87

Note:

SAM versus Norges Bank's forecast

Below we compare the RMSFE for SAM with Norges Bank's (NB) official forecast from 1999Q1-2008Q1. The NB forecasts are collected from various monetary policy reports. We assume that Norges Bank's one step ahead forecast (Horizon 1) is made in the same quarter that it is forecasting, giving Norges Bank a small advantage over the other models. In particular,

- The forecasts for Q1 are taken from MPR1 (published end of March), when CPI is known for January and February, giving Norges Bank two months information **advantage** relatively to SAM (in the period 2001 to 2004, the monetary policy reports are published before CPI is known for February, giving Norges Bank only one month information advantage).

- The forecasts for Q2 are taken from MPR2 (published end of June), when CPI is known for April and May, giving Norges Bank two months information **advantage** relatively to SAM.

- Regarding the forecast for Q3, for the period 1999-2000, the forecasts are taken from the monetary policy report published end of September, when CPI is known for July and August, giving Norges Bank two months information advantage. However, from 2001, Norges Bank no longer publish any forecast in the third quarter. Most of the forecasts for Q3 are therefore taken from MPR2 (published end of June), giving Norges Bank instead one month information **disadvantage** relatively to SAM (since June figures for CPI were yet not known).

- The forecasts for Q4 are taken from MPR3 (published end of October), when no CPI information for Q4 is known. Hence, in Q4, Norges Bank has no advantage over SAM (except in the brief period 1999-2000, when the report is published in December when two months of CPI is known).

This gives Norges Bank an advantage over SAM in two of the quarters (Q1 and Q2) and an disadvantage in one quarter (Q3). For Q4, the information content is about the same. Note also that for the period 1999-2000, Norges Bank's forecast are for CPI and not CPIATE. This may give Norges Bank a disadvantage if CPI tracks systematically above or below CPIATE in this period. This is examined in the table below.

- When comparing the whole forecasting period; 1999Q2-2008Q1, SAM outperforms NB in terms of RMSFE at all the horizons.

- For the period 2001Q1-2008Q1, NB outperforms SAM (marginally) at horizon 1, but then SAM outperforms Norges Bank at horizons 2-4.

Table 4: RMSFE for inflation: Comparing Norges Bank and SAM

Model		Horizon 1	Horizon 2	Horizon 3	Horizon 4
1999Q2-2008Q1					
	SAM	0.22	0.30	0.41	0.52
	NB	0.37	0.38	0.56	0.74
Model		**Horizon 1**	**Horizon 2**	**Horizon 3**	**Horizon 4**
2001Q1-2008Q1					
	SAM	0.21	0.30	0.42	0.53
	NB	0.18	0.36	0.59	0.78

Figure 15: *Actual inflation vs. forecast (1999Q1-2008Q1). Horizon 1*
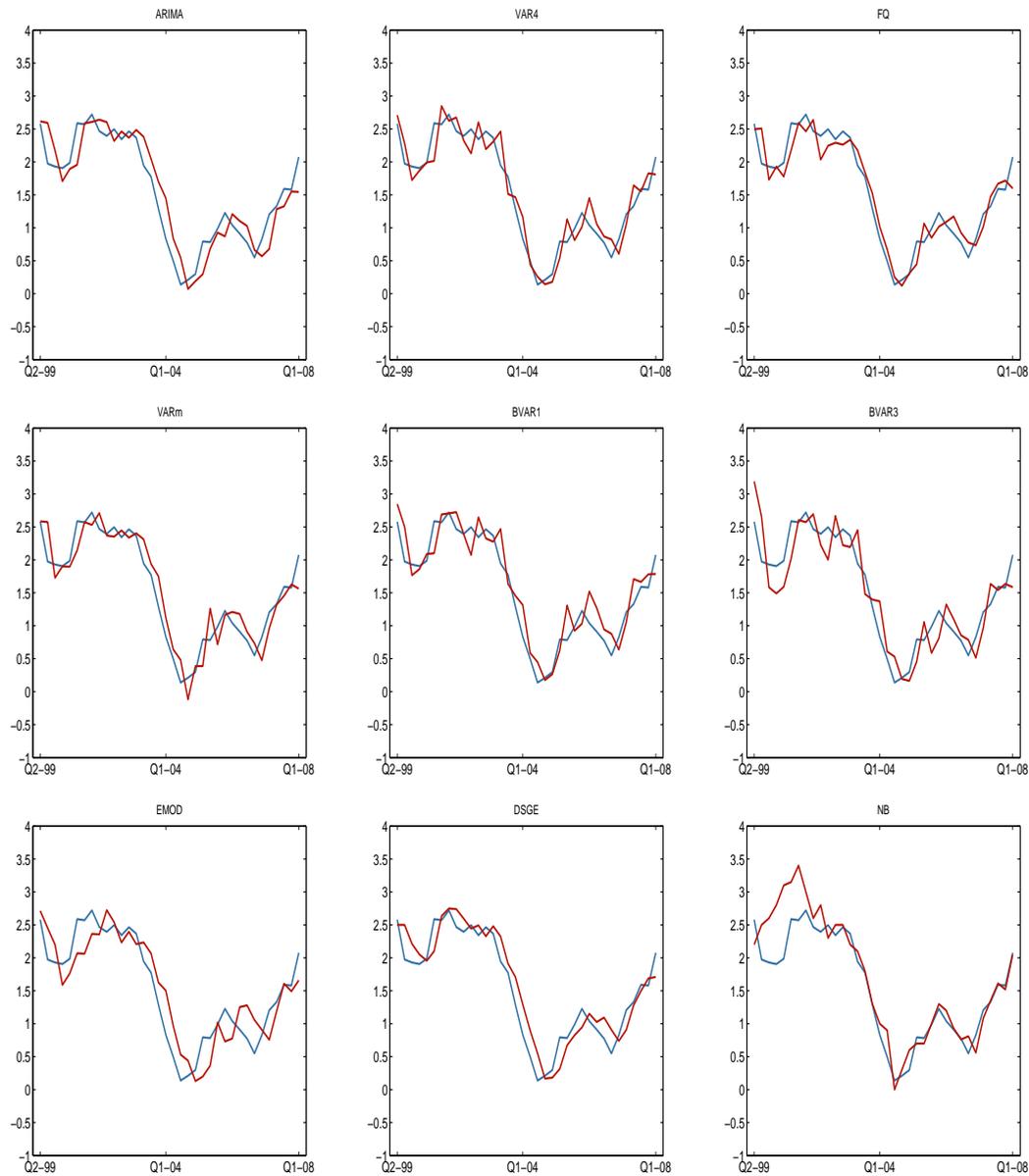
Note: Inflation (CPIATE, yoy growth). Actual (blue) and forecast (red). NB refers to Norges Bank's official forecast collected from various monetary policy reports.

Figure 16: *Actual inflation vs. forecast (1999Q2-2008Q1). Horizon 2*
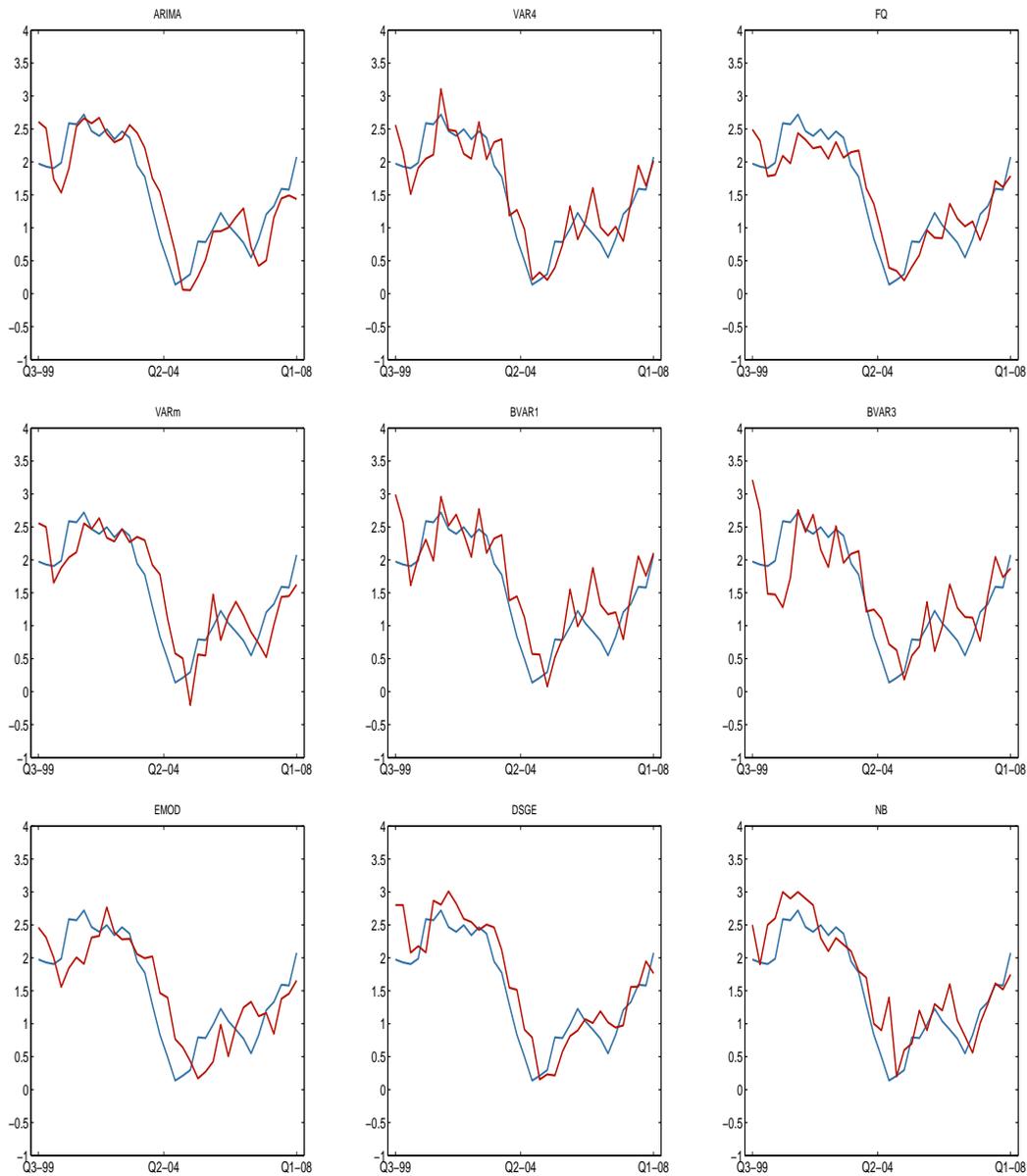
Note: Inflation (CPIATE, yoy growth). Actual (blue) and forecast (red). NB refers to Norges Bank's official forecast collected from various monetary policy reports.

Figure 17: *Actual inflation vs. forecast (1999Q2-2008Q1). Horizon 3*
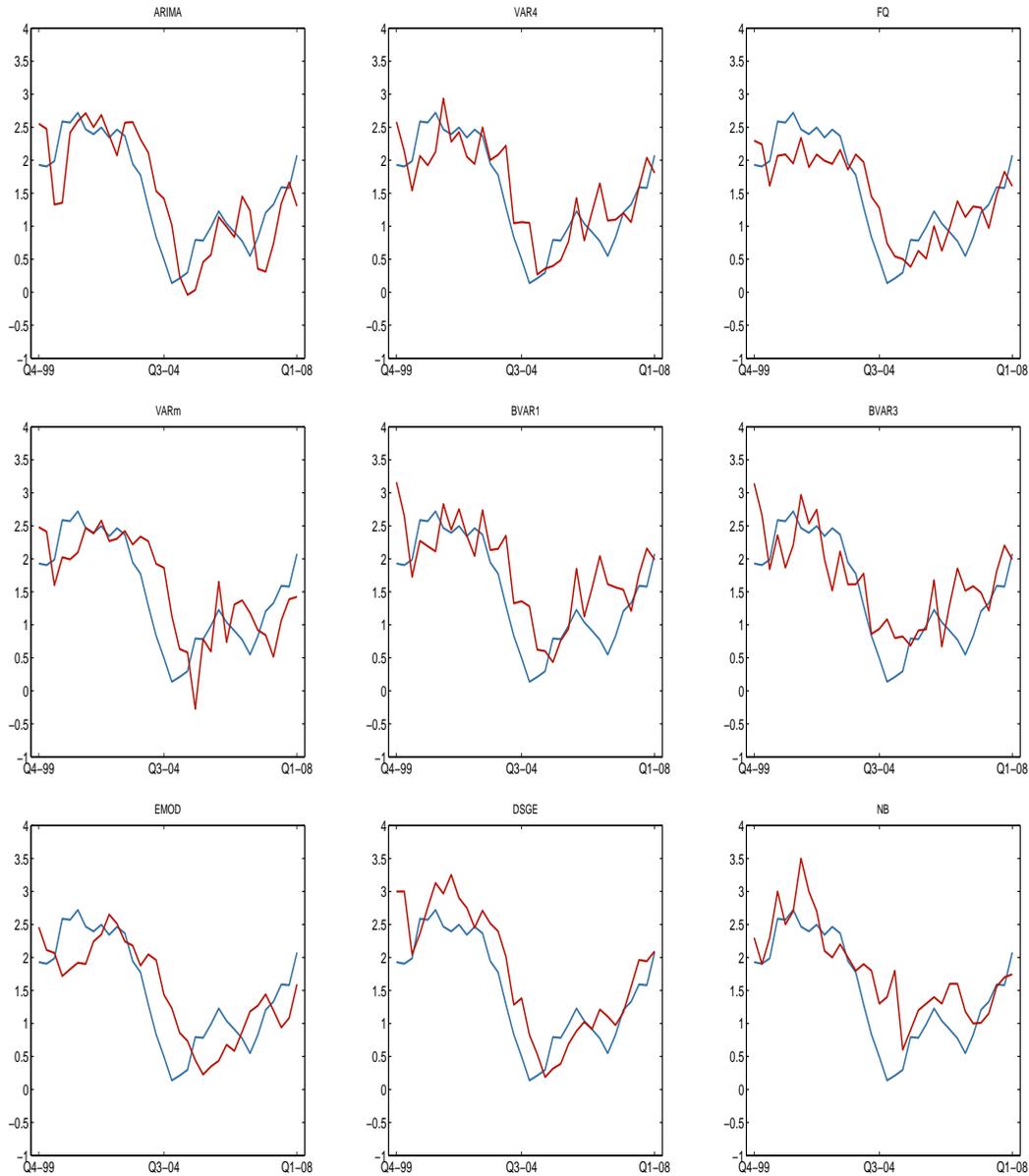
Note: Inflation (CPIATE, yoy growth). NB refers to Norges Bank's official forecast collected from various monetary policy reports.

Figure 18: *Actual inflation vs. forecast (1999Q2-2008Q1). Horizon 4*
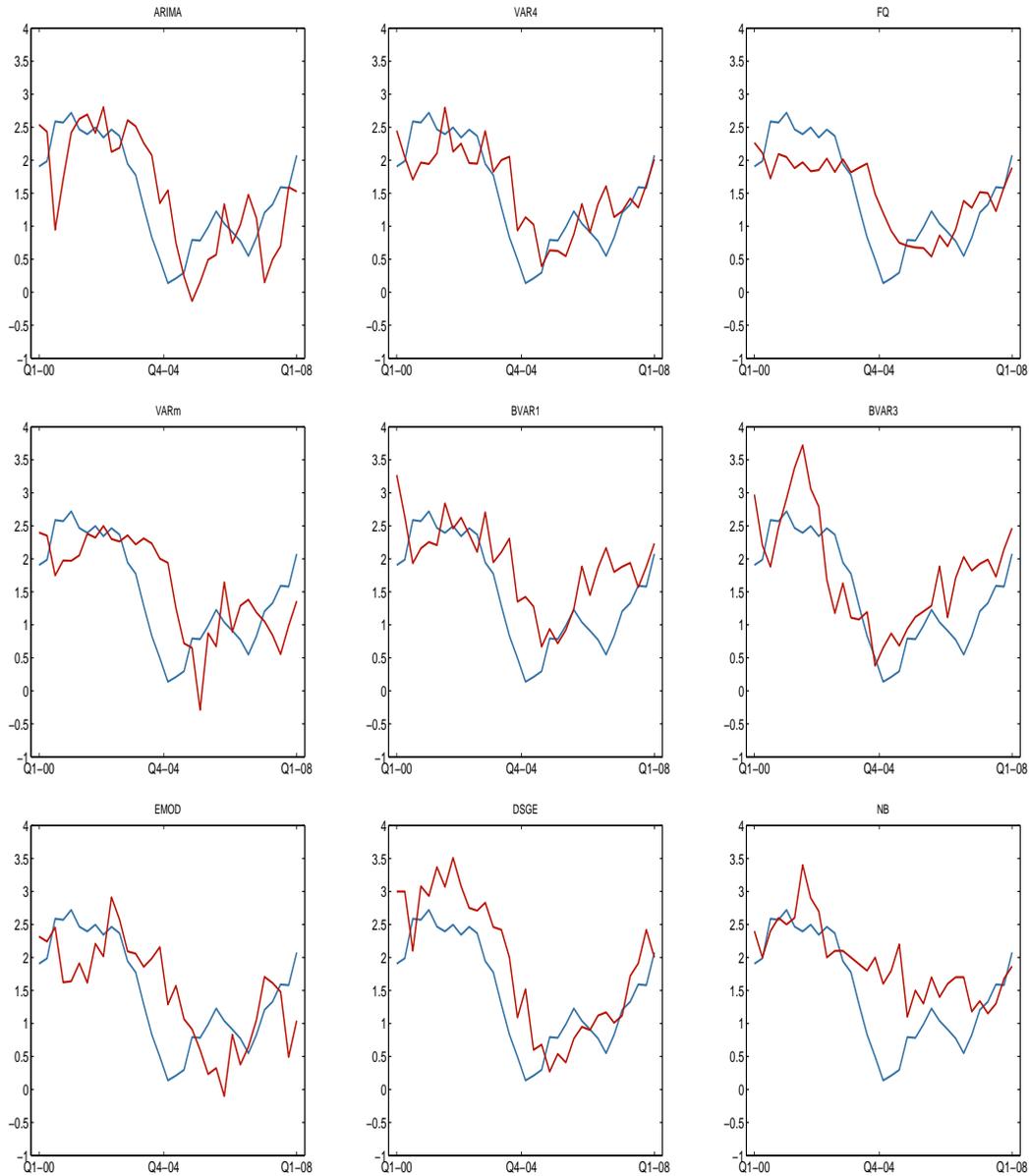
Note: Inflation (CPIATE, yoy growth). Actual (blue) and forecast (red). NB refers to Norges Bank's official forecast collected from various monetary policy reports.

Figure 19: *Rank of 8 best models predicting inflation*
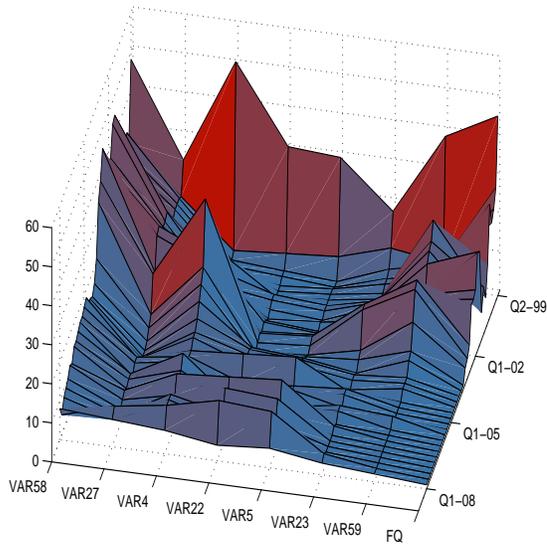
Horizon 1
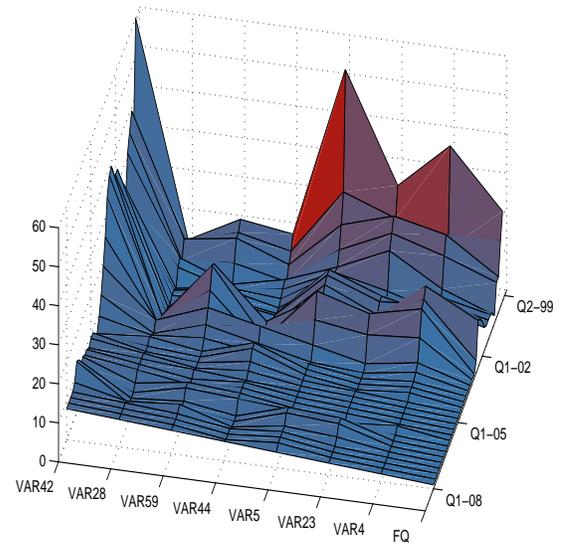
Horizon 2
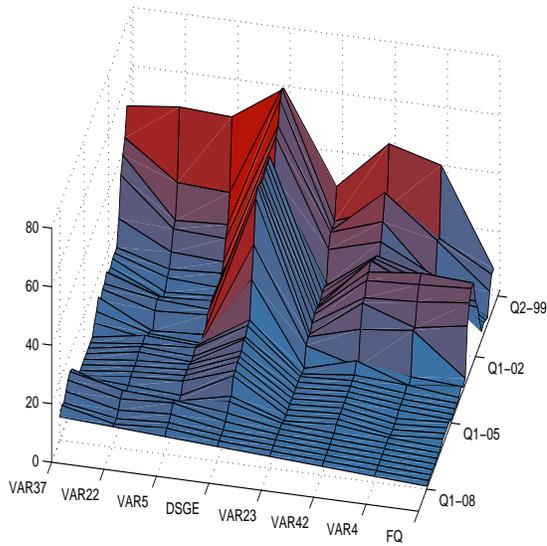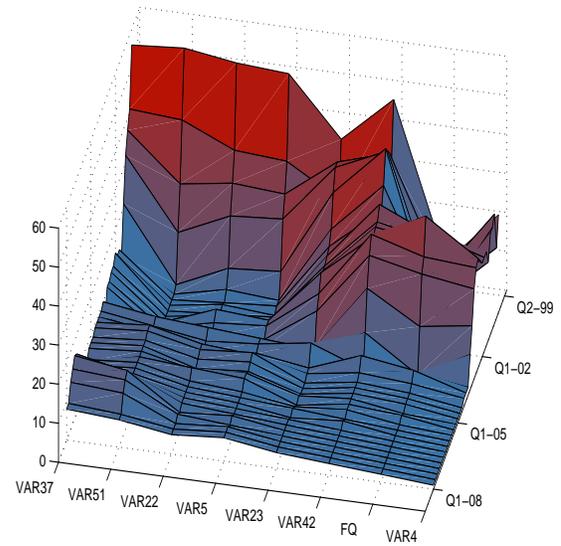
Horizon 3

Horizon 4

Note: The graphs compares the rank of the 8 best models over time.

Details on GDP forecast

Table 5: GDP forecasts

	Q2.08	Q3.08	Q4.08	Q1.09
SAM	3.91	2.94	2.64	3.94
SAM(t-1)	4.45	3.59	4.24	
SAM(t-2)	4.24	3.91		
SAM(t-3)	4.04			

Note: The maximum SAM forecast horizon is four quarters. SAM (*t-i*) is the forecast path *i* periods before the current SAM forecast.

Table 6: Best models for GDP

	1st	2nd	3rd	4th	5th	6th	7th	8th
Q2.08	Tstruc2	Tstruc1	FQ	MI	FM	Tstruc3	Arima	DSGE
val	3.71	3.61	3.66	4.18	3.68	4.15	5.02	3.30
Q3.08	VAR-SAM	Tstruc2	Tstruc1	Tstruc3	FM	2V_cap	FQ	2V_pro
val	3.84	2.35	2.12	3.19	2.28	3.92	1.88	3.91
Q4.08	VAR-SAM	BVAR1	Tstruc2	Tstruc3	FM	Tstruc1	2V_vac	DSGE
val	3.55	4.60	1.83	3.04	1.82	1.40	3.96	0.69
Q1.09	VAR-SAM	BVAR3	BVAR1	2V_vac	2V_nvac	BVAR2	Tstruc3	Tstruc2
val	3.07	5.25	4.54	3.88	3.89	4.97	3.77	2.21

Notes:

Figure 20: *Actual GDP vs. forecast (1999Q2-2008Q1). Horizon 1*
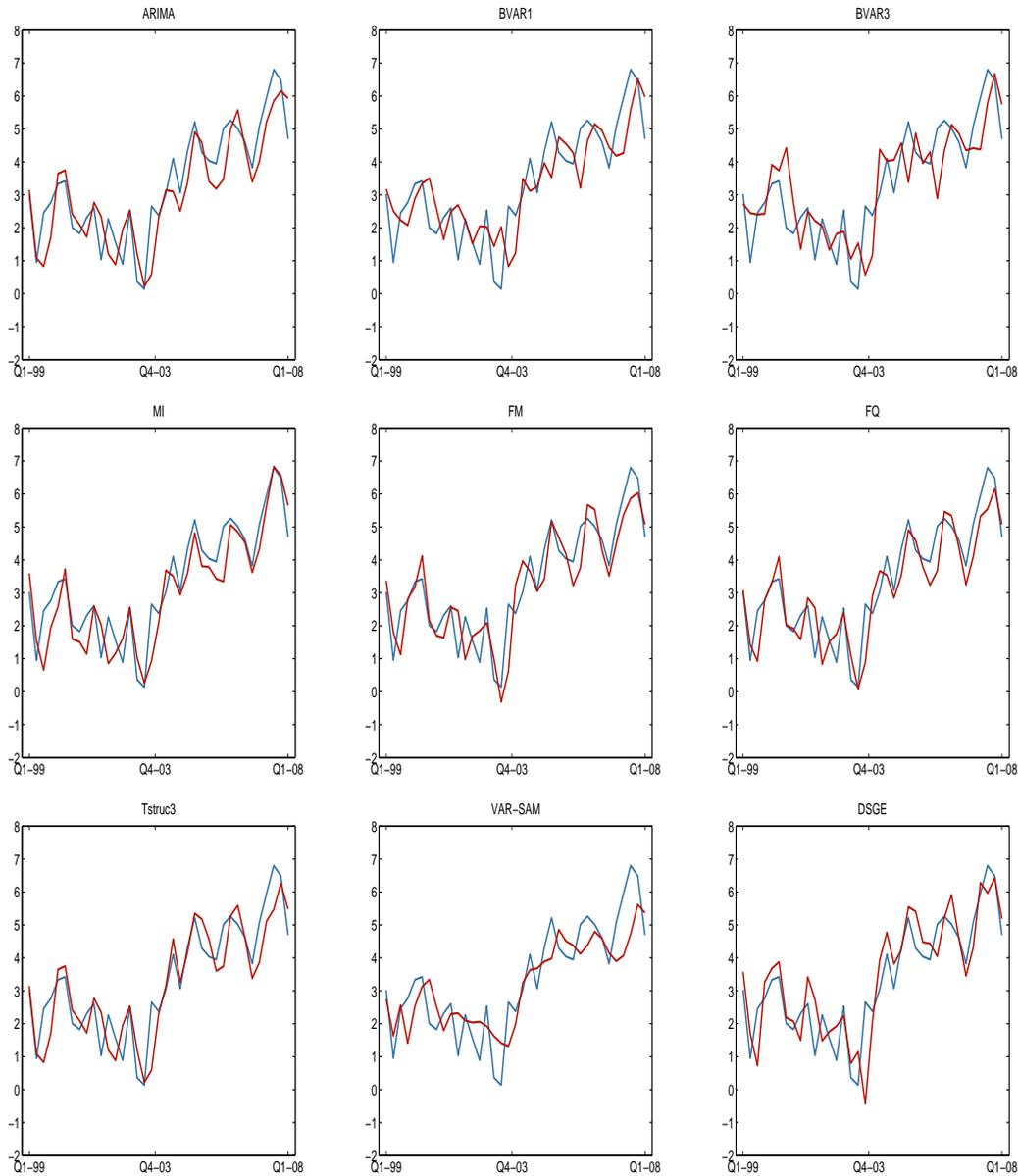
Note: Mainland GDP (year on year growth rates). Actual (blue) and forecast (red)

Figure 21: *Actual GDP vs. forecast (1999Q2-2008Q1). Horizon 2*
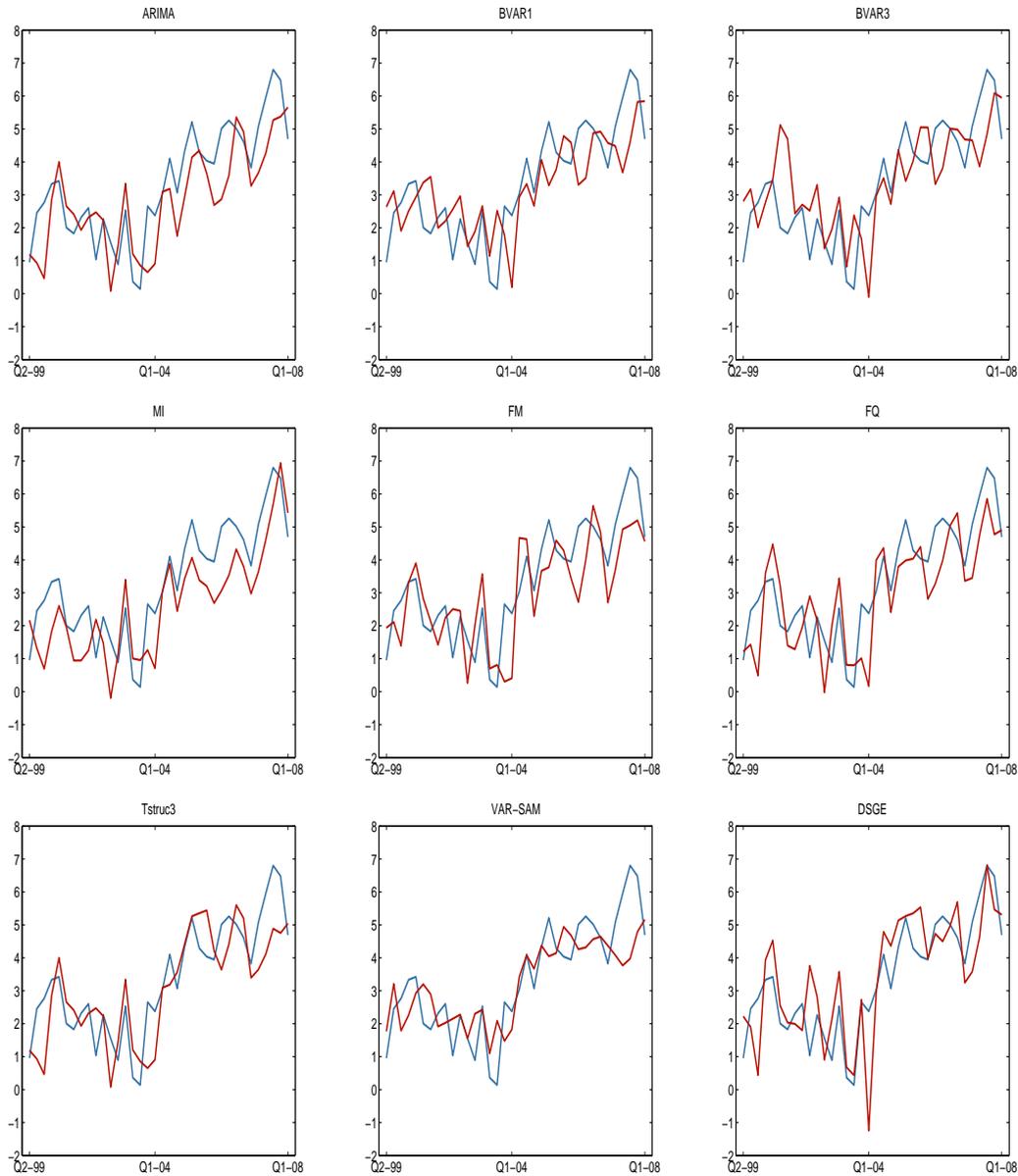
Note: Mainland GDP (year on year growth rates). Actual (blue) and forecast (red)

Figure 22: *Actual GDP vs. forecast (1999Q2-2008Q1). Horizon 3*
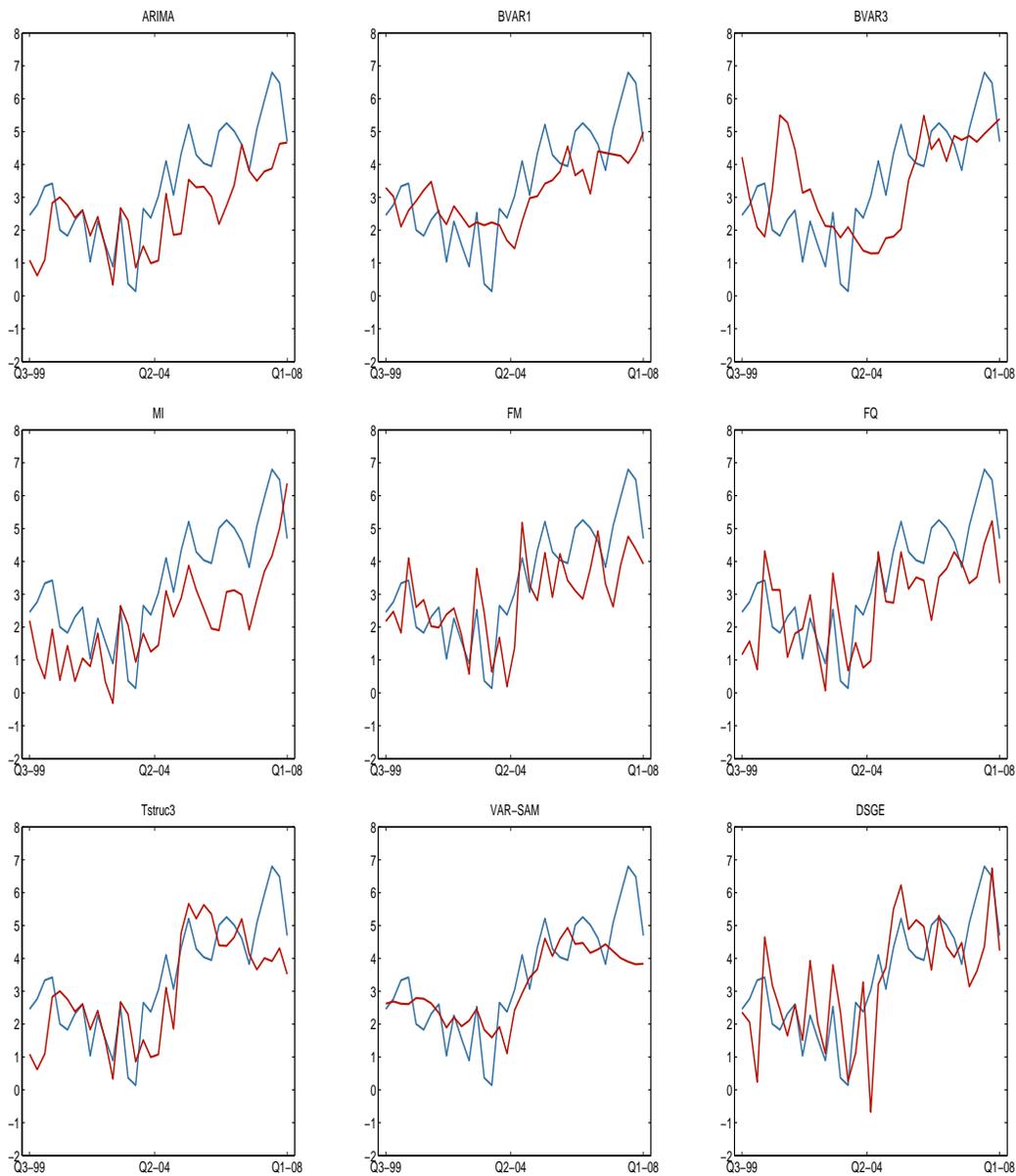
Note: Mainland GDP (year on year growth rates). Actual (blue) and forecast (red)

Figure 23: *Actual GDP vs. forecast (1999Q2-2008Q1). Horizon 4*
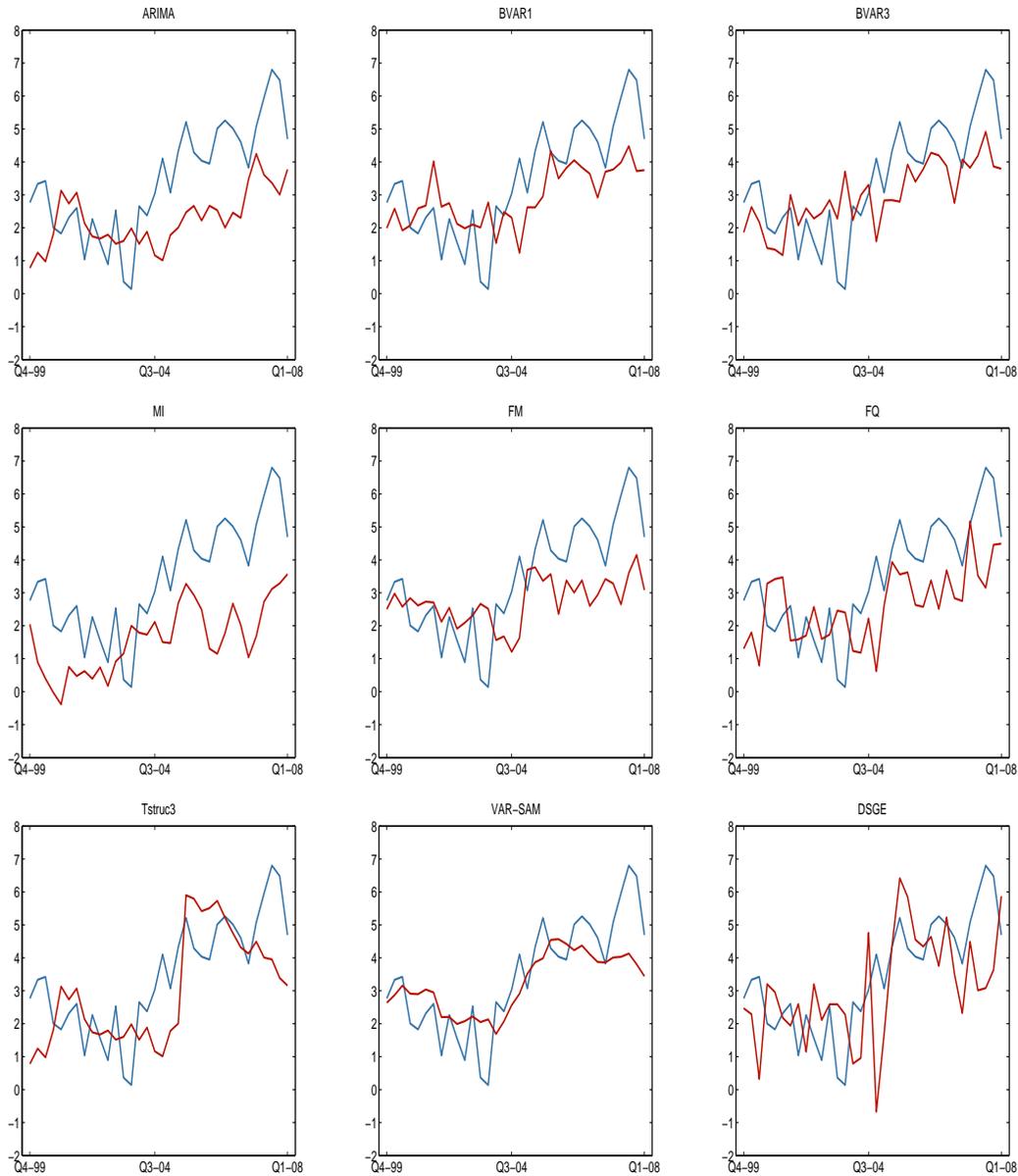
Note: Mainland GDP (year on year growth rates). Actual (blue) and forecast (red)

Figure 24: *Rank of 8 best models predicting GDP*
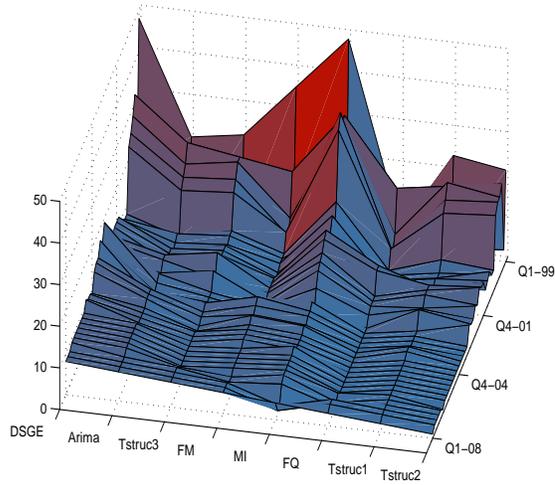
Horizon 1
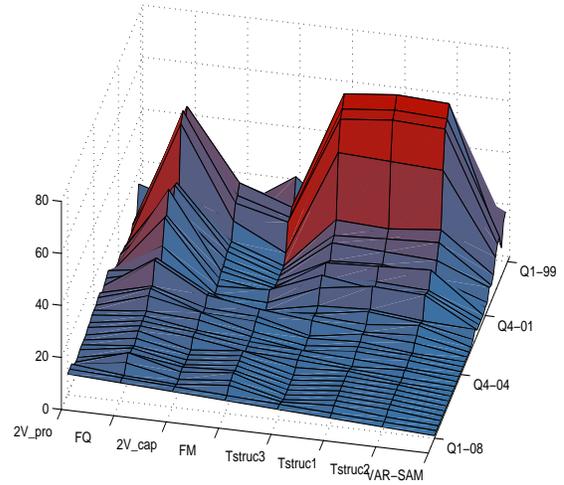
Horizon 2
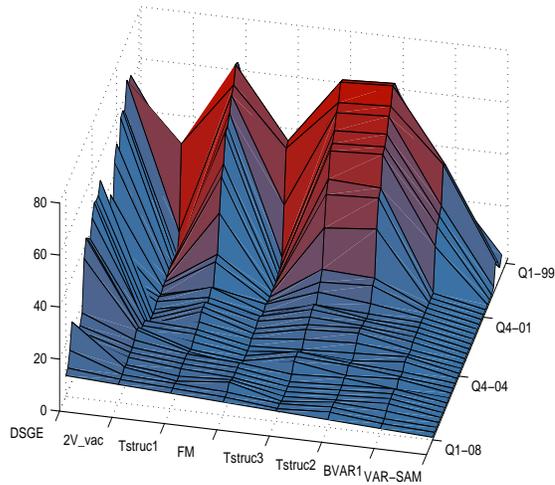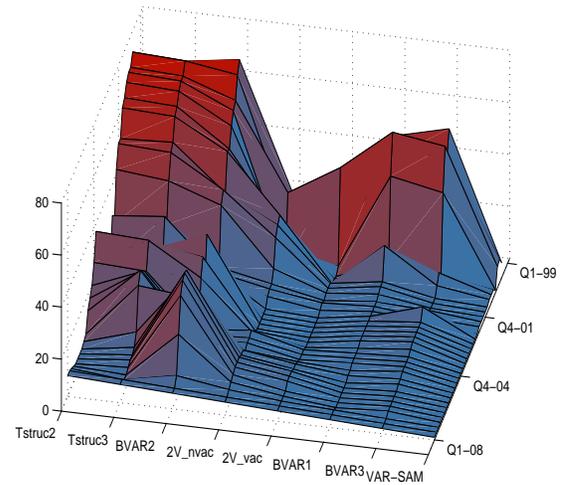
Horizon 3

Horizon 4

Note: The graphs compares the rank of the 8 best models over time.

6   Appendix: Brief explanation of models

(Number of such models in parenthesis)

1. Autoregressive (**AR**) model (*10*)

2. Monthly indicator (**MI**) model

3. Bivariate Vector Autoregressive (**2VAR**) models *(58)*

4. Vector autoregressive model (**VARm**) with money

5. Vector autoregressive (**VAR**) models (*54*)

6. Bayesian Vector autoregressive (**BVAR**) models (*10*)

7. Factor (**FM**) models (*2*)

8. Term structure models (**Tstruc**) models (*3*)

9. Vector equilibrium correction model (e.g. **EMOD**)

10. Dynamic Stochastic General Equilibrium (DSGE) model (e.g. **NEMO**)

1. Autoregressive (AR) models

- AR(p) models explain a series based on p past observations (lags) of the same variable

- ARMA(p,q) models explain a series base on p lags of the same variable but assume that the error term is a moving average with q elements

- ARIMA(p,d,q) models difference the series d times to eliminate trends, and then apply ARMA(p,q) models to the differenced series

- AR, ARMA, and ARIMA models are all linear models; the behaviour of the series is a linear function of past observations

2. Monthly indicator (MI) models

- Predict GDP based on many monthly indicators (that are averaged up to a quarterly frequency).

- Short term monthly indicator statistics are available before GDP figures are published a good two months after the end of the quarter.

- The indicators that covary with GDP and that cover the entire or parts of a non-published quarter in the national accounts can be used to project GDP for that quarter.

- We use an indicator model for GDP that contains effects of manufacturing production, employment, retail sales, hotel statistics and building starts.

- In order to forecast mainland GDP ahead, the explanatory variables in the indicator model are projected using ARIMA models.

3. Factor models

- **Factor model with monthly data (FM**)

    - Estimated on monthly data (financial prices, labour data, industrial production, etc)
    - The factors (principal components) summarise the common variation in the series
    - Each monthly series reflects the common factors and an idiosyncratic component
    - The appropriate *monthly* factors are used as regressors to forecast *quarterly* GDP
    - Forecasts of future GDP growth reflect the dynamics estimated for the factors

- **Factor model with quarterly data (FQ)**

    - Factor model 'imported' from RBNZ to predict both GDP and inflation
    - Estimated using principal components on quarterly data (completely different data set to previous factor model; includes survey data)
    - Uses a 'direct' forecasting method (eg GDP at time $t+h$ is regressed against factors from data at time t)
    - Different model variants
        * Different numbers of lagged factors
        * Different factors estimated by using subsets of data (factors not always better from larger data sets; subsets determined by the correlation between indicators & GDP)

4. Bivariate VAR (2VAR) models

- VAR(p) models are the vector equivalent of AR(p) models; they explain the vector based on p lags of all of the variables in the vector

- 2V is a bivariate VAR (VAR with two variables). Predict GDP based on quarterly (indicator) variables. Estimate a bivariate VAR for each explanatory variable.

- p is chosen to be between 1 and 4.

    - 2V_cap (*Degree of cap. utilization*)
    - 2V_pro (*Industrial production*)
    - 2V_vac (*Stock of vacancies, percentage change*)
    - 2V_nvac (*New vacancies, percentage change*)
    - 2V_bts (*Industrial confidence indicator*)
    - 2V_bts (*Industrial confidence indicator*)
    - 2V_ordm (*New orders; Manufacturing*)
    - 2V_ordc (*New orders; Construction*)
    - 2V_is (*Indicator of resource shortage*)
    - 2V_hti (*Household trend indicator; various*)
    - 2V_hs (*Employment in services*)
    - ...

5. VAR with money (VARm) models

- VAR(p) models predicting inflation where monetary aggregates are used as explanatory variables.

- The VARm is an average over one univariate model and two bivariate VARs.

- The univariate model is an autoregressive model with seasonal dummies.

- The bivariate VARs are estimated on year to year changes and contain prices ex energy and taxes and short-term interest rates and money (M2), respectively.

- All models are estimated on moving data windows.

6. VAR models

- Estimate GDP, inflation and interest rates jointly using a

Table 7: VAR models in SAM

Model	Description	Model	Description
1	AR(2)	33	VAR(1) rolling even shorter
2	AR(BIC)	34	VAR(1) rolling even shorter
3	rolling window AR(2)	35	VAR(BIC) rolling even shorter
4	VAR(4)	36	VAR(1) Inflation detrended Rolling VAR
5	VAR(BIC)	37	BiVAR(2): (ex interest rate)
6	VAR(4): rolling	38	BiVAR(1): rolling: (ex interest rate)
7	VAR(BIC): rolling	39	BiVAR(2): rolling: (ex GDP/inflation)
8	AR(BIC): rolling	40	AR(3)
9	DVAR(4)	41	AR(3) rolling
10	DVAR(BIC)	42	VAR(3)
11	DVAR(4): rolling	43	VAR(3) rolling
12	DVAR(BIC): rolling	44	DVAR(3)
13	VAR(4) Inflation detrended	45	DVAR(3): rolling
14	VAR(BIC) Inflation detrended	46	VAR(2) Inflation detrended
15	VAR(BIC) rolling even shorter	47	VAR(3) Inflation detrended
16	VAR(BIC) rolling even shorter	48	VAR(2) Inflation detrended Rolling VAR
17	VAR(BIC) rolling even shorter	49	VAR(3) Inflation detrended Rolling VAR
18	VAR(4) Inflation detrended Rolling VAR	50	VAR(BIC) Inflation detrended Rolling VAR
19	AR(1)	51	BiVAR(1): (ex interest rate)
20	AR(4)	52	BiVAR(3): (ex interest rate)
21	rolling window AR(1)	53	BiVAR(4): (ex interest rate)
22	VAR(1)	54	BiVAR(2): rolling : (ex interest rate)
23	VAR(2)	55	BiVAR(3): rolling: (ex interest rate)
24	VAR(1): rolling	56	BiVAR(4): rolling: (ex interest rate)
25	VAR(2): rolling	57	BiVAR(1): (ex GDP/inflation)
26	AR(4): rolling	58	BiVAR(3): (ex GDP/inflation)
27	DVAR(1)	59	BiVAR(4): (ex GDP/inflation)
28	DVAR(2)	60	BiVAR(2): (ex GDP/inflation)
29	DVAR(1): rolling	61	BiVAR(2): rolling: (ex GDP/inflation)
30	DVAR(2): rolling	62	BiVAR(3): rolling: (ex GDP/inflation)
31	VAR(2) Inflation detrended	63	BiVAR(4): rolling : (ex GDP/inflation)
32	VAR(1) rolling even shorter	64	BiVAR(1): rolling: (ex GDP/inflation)

Note: The number refers to the different VAR models used in the report.

7. Bayesian VAR (BVAR) models

BVAR models are VAR models that are estimated using Bayesian techniques

- In Bayesian analysis the econometrician has to specify 'prior' beliefs about the parameters

- 'Prior' beliefs are combined with the data to form a 'posterior' (post-data) view of the parameters

- When data samples are large the influence of the prior beliefs will diminish and the data will guide the parameter values;

- In small samples prior beliefs may help to guide the parameter estimates towards sensible values, assisting forecasting

- Uses either a *direct* forecasting method (eg GDP at time $t+h$ is regressed against factors from data at time t), or an iterative forecasting method, (where the one period ahead model is iterated forward for the desired number of forecast).

- BVAR models in SAM

  - BVAR1 (2 variables: GDP, inflation) *Direct forecast*
  - BVAR2 (4 variables: GDP, inflation, interest rates and **excahnge rate**) *Direct forecast*
  - BVAR3 (2 variables (BVAR1) plus exogenous variables) *Direct forecast*
  - BVAR4 (4 variables (BVAR2) plus exogenous variables) *Direct forecast*
  - BVAR8 (4 variables: GDP, inflation, interest rates and trade weighted **terms of trade**) *Direct forecast*
  - BVAR9 (4 variables (BVAR8) plus exogenous variables) *Direct forecast*

8. Term structure (Tstruc) models

We construct three models for GDP using term structure information:

- An unrestricted VAR model for the 3-month yield, the yield spread (5-years minus 3-months) and GDP (Trsruct1).

- Two affine models for forecasting GDP.

  - An affine model builds a relationship between yields of different maturities (in our case 3-months, 1-year, 2-years, 3-years, 4 years, 5-years).

  - More precisely, it is a structural VAR model where parameters are restricted to ensure there is no possibility of arbitrage between yields of different maturities.

  - The affine model provides both short run and long run forecasts (being a structural VAR model), resulting in two sets of forecasts (Tstruc2, Tstruc3).

- Note that, due to data limitations, the term structure models only produce forecast from 2004 and onwards. Prior to 2004, we use forecast from the ARIMA model.

9. (Modified) Emod

- An econometric (equilibrium correction) model of 13 (*not-seasonally adjusted*) macro variables; specification derived from data

- CPI, GDP, other domestic variables, auxiliary equations for variables such as foreign prices, interest rates, oil price

- Sample period beginning 1982Q4/2001Q1 (latter date reflecting change in monetary policy regime)

- Real-time forecast evaluation problematic given the regime shift; performance depends on auxiliary equations

  - Missing forecasts (for the short series) approximated with an AR(2)
  - Emod forecasts for all variables from 2003Q4

10. DSGE model

New Keynesian small open economy model: version of NEMO documented in Staff Memo 2006/6

- Estimated using Bayesian maximum likelihood on unadjusted data for

    - Mainland GDP growth
    - Consumption growth
    - Investment growth
    - Export growth
    - Employment
    - Inflation (CPI-ATE)
    - Imported inflation
    - Real wage growth
    - Real exchange rate (I44)
    - Nominal interest rate

    Sample period: 1987Q1–1998Q4 (extended recursively until 2007Q4)
    Steady-state levels equal to recursively updated means of the variables

Acknowledgement

The forecast in this report are technical forecast. There is no judgment in the averaging of the forecast. The forecasts may not necessarily represent Norges Bank's view.

Table 8: Nowcasting schedule

Date	May/June 2008	September 2008	December 2008	February 2009

Note: The dates are tentative.