CENTRE FOR APPLIED MACROECONOMIC RESEARCH (CAMAR)

CAMAR Working Paper Series
No 2/2010

Does forecast combination improve Norges Bank inflation forecasts?

Hilde C. Bjørnland, Karsten Gerdrup, Anne Sofie Jore,
Christie Smith, Leif Anders Thorsrud

BI NORWEGIAN SCHOOL OF MANAGEMENT

Does forecast combination improve Norges Bank inflation forecasts?*

HILDE C. BJØRNLAND† KARSTEN GERDRUP

Norwegian School of Management Norges Bank

ANNE SOFIE JORE CHRISTIE SMITH

Norges Bank Reserve Bank of New Zealand

LEIF ANDERS THORSRUD

Norwegian School of Management

December 15, 2010

Abstract

We develop a system that provides model-based forecasts for inflation in Norway. We recursively evaluate quasi out-of-sample forecasts from a large suite of models from 1999 to 2009. The performance of the models are then used to derive quasi real time weights that are used to combine the forecasts. Our results indicate that a combination forecast improves upon the point forecasts from individual models. Furthermore, a combination forecast out-performs Norges Bank's own point forecast for inflation. The beneficial results are obtained using a trimmed weighted average. Some degree of trimming is required for the combination forecasts to out-perform the judgmental forecasts from the policymaker.

JEL-codes: E52, E37 E47.
Keywords: Forecasting, forecast combination, model versus judgment.

*The authors thank two anonymous referees, Gernot Doppelhofer, James Mitchell and participants at seminars in Norges Bank, at the Norwegian School of Economics and Business Administration and at the 30th International Symposium on forecasting in San Diego. The usual disclaimer applies. The views expressed in this paper are those of the authors and should not be attributed to Norges Bank nor to the Reserve Bank of New Zealand.

†*Norwegian School of Management (BI), NO-0442 Oslo. Norway. (e-mail. hilde.c.bjornland@bi.no)*

1 Introduction

Policy-making entails evaluating the future trajectory of the economy, and making policy decisions to influence that trajectory in favorable directions. The forward-looking nature of these policy decisions means that macroeconomic forecasting is a critical component underlying decisions.

Developing empirical models to describe and forecast the behaviour of the economy is, however, 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 et al. (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.

There are at least three reasons why forecast combinations may be preferable to methods based on the ex-ante best individual forecasting model, see Bates and Granger (1969) for a seminal paper and Timmerman (2006) for a survey. First, forecast combination can be motivated by a simple portfolio diversification argument. By combining models based on different information sets, the combined forecast may be more accurate than that from a single model (trying to incorporate all the information), see e.g. Huang and Lee (2008) and Hendry and Clements (2002). Second, there may be unknown instabilities (structural breaks) that favour one model over another at different points in time, though which model is superior is unknown. By combining forecasts, one may obtain forecasts that are more robust to these instabilities, see e.g. Clark and McCracken (2008) and Jore et al. (2007). Finally, forecast combination may be desirable when models are subject to omitted variable bias. Combining forecasts may average out these unknown biases, particularly if they are idiosyncratic. Hence, even though the combined forecast may not always be superior, model combination may be preferable as it will ensure against selecting a single bad model.

In this paper we conduct a quasi out-of-sample forecasting evaluation of core inflation in Norway (consumer prices excluding taxes and energy, CPIATE) obtained from a broad spectrum of models. Our main objective is to consider whether forecast combination improves upon the forecasts from individual models. We focus on CPIATE as it has been the key underlying inflation measure that has been used to guide Norges Bank's

inflation targeting regime since 2001.[1] We then perform a quasi real time forecasting exercise to see whether forecast combination is superior to Norges Bank's own forecasts. Although there are a number of central banks that have explicitly adopted multi-model and model/forecast combination approaches to short-term forecasting, there has been little formal evaluation of whether model/forecast combination in fact out-performs official central bank forecasts (with the recent exception of Adolfson et al. (2007)), once the preferred forecasts and judgment are taken into account. It turns out that forecast combination does improve upon single model forecasts. Furthermore, model combination out-performs Norges Bank's own point forecast for inflation. The suite of models allows for a greater range of modelling techniques and data to be used in the forecasting process.

The rest of this paper is organized as follows. In Section 2 we discuss the modelling/forecasting approach adopted in peer central banks. Section 3 describes how our forecasts are produced, while Section 4 presents the results and compares the model forecasts to Norges Bank's official forecasts. Section 5 concludes.

2 Forecasting/combination schemes at peer central banks

A number of central banks have recently adopted multi-model and forecast/model combination approaches to short-term forecasting. For instance, the Bank of England applies combination techniques to the forecasts from the suite of models, see for example Kapetanios et al. (2005) and particularly Kapetanios et al. (2008). The suite includes linear and non-linear univariate models, vector autoregressive (VAR) models of various

[1]Norges Bank's mandate requires it to disregard any direct effects on consumer prices resulting from changes in interest rates, taxes, excise duties and extraordinary temporary disturbances when designing monetary policy.

specifications, Bayesian VARs, factor models and time-varying coefficient models. 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 Adolfson et al. (2007). The Riksbank's suite of models incorporates bivariate and multivariate VARs, Bayesian VARs, VARs that also incorporate factors (which summarise the broad 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) and Adolfson et al. (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 Bank of Canada is another central bank that is using multiple models for near-term forecasting. Coletti and Murchison (2002) provide a description of the multi-model approach to policy-making employed at 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.

Although some of these central banks have adopted model combination approaches to short-term forecasting, there is no consensus with regard to whether model combination in fact out-performs the Central Banks's own projections, once model based forecasts and

judgment are taken into account. This is the issue we examine below when comparing the combined model forecast to Norges Bank's own inflation forecasts.

3 Forecast comparison

Norges Bank's projections form an important basis for the conduct of monetary policy. The forecasting work involves the use of the structural macro-model NEMO,[2] but this model is primarily directed at providing medium and long term projections, embodying the effects of endogenous policy. 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.

A new ingredient in the above forecasting framework is the formal combination of model-based, short-term forecasts. Such a system has been developed in Norges Bank and is collectively referred to as SAM – the System for Averaging Models.[3] The forecasts in 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), but because not all models forecast all variables it was decided to use univariate weights.

To evaluate the forecasts, we use a quasi out-of-sample forecasting approach. This is the general approach to forecast evaluation followed in the literature, see Stock and Watson (2003) for definition, and Clark and McCracken (2008) and Smith and Wallis (2009) for recent applications. The method mimics real-time out of sample forecasting, yet as the final vintage of data is known, we can assess forecast performance. The idea is

[2]See Brubakk et al. (2006) for details.

[3]See Bjørnland et al. (2008) for more details on Norges Bank's forecasting/nowcasting project.

to split the sample into two sub-samples. The first sub-sample is used for estimating the forecasting relationships (but from the *final* vintage of data), while the second is used to evaluate their forecasting performance. In particular, one simulates starting at a given point in time t and performing all model specification and parameter estimation using only the data available at that date. Then one computes the $h - period$ ahead forecast for date $t + h$ and evaluates this with the actual data for period $t + h$. One then moves forward to date $t + 1$. The information set on which the forecast is based is updated and new forecasts are made and evaluated with actual data. This is repeated for all dates in the forecast period.[4] Note that the weights are constructed using the quasi out-of-sample forecast errors that existed (in quasi real time) prior to each forecast origin. This will be described in more detail below.

3.1 Data and forecast horizon

Here we forecast the year-on-year (yoy) inflation rate, measured by CPIATE. As mentioned above, CPIATE has been Norges Bank's core/underlying inflation measure since 2001. We choose to focus on inflation as it is rarely revised, hence there are fewer real time data issues (though of course regressors such as GDP are subject to revision). We will forecast variables up to five steps ahead, $(h = 1, 2, \ldots, 5)$, i.e. we wish to forecast $y_{t+1}, y_{t+2}, y_{t+3}, y_{t+4}, y_{t+5}$, where y is the forecast variable of interest (the inflation rate). With quarterly data, the maximum forecasting horizon is thus 5 quarters ahead.[5]

The models are first estimated up to 1998Q4, and then the estimation window is

[4] It is an open question whether the estimated relationship should also be updated. However, since the evaluation of the combined forecast is based on the evaluation sample (and not using a formal inference procedure), the choice of updating scheme is not material (see Smith and Wallis (2009) for discussions).

[5] The main reason for choosing this forecast horizon is that Norges Bank doesn't use statistical models beyond 4-5 quarters in the forecast process, but instead uses the DSGE model. Hence, since we want to compare staff forecast (i.e. model averaging) with Norges Bank's judgment (official forecast), we decided to keep the forecast horizon the same.

recursively expanded in quasi-real time from 1999Q1-2009Q1 to make forecasts. In particular, conditioning on information up to and including 1998Q4, the models are first used to produce forecasts for 1999Q1 to 2000Q1. The conditioning combination's information set is then extended one period forward (to 1999Q1), and forecasts are made for 1999Q2 to 2000Q2. The whole process is repeated until all available information has been used, to provide the final forecast conditioning on information available up to and including 2009Q1.

Having obtained all the forecasts, the actual data between 1999Q1 and 2009Q1 are then used to evaluate the average performance of the h-step forecasts, $h = 1, \ldots, 5$. The models will be evaluated and ranked in real time prior to each forecast origin. Note that we evaluate the point forecasts for each horizon; the point forecasts represent the central expected location of yoy CPIATE inflation at the five horizons. Makridakis and Hibon (2000) note that the performance of models varies by forecasting horizon. For example, a model may forecast well for a 1-step ahead horizon, but may be much worse, relative to the other models, at forecasting 4 or 5-steps ahead. The model weights conducted here have been made horizon-specific for this reason.

The period used for the out-of-sample evaluation roughly coincides with the shift to 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, see Timmerman (2006).

3.2 Forecast combinations

As mentioned above, we perform a quasi real time forecasting exercise, so that the forecasting models are ranked and given weight in real time. That is, the weights are constructed ex-ante to each forecast origin. We consider two types of weights, both of them common in the literature. The first choice of weights is simply to equally-weight all models that enter the combination. That is, we construct a simple mean of all the forecasts. Timmerman (2006) has noted that equal weights will be appropriate when models have equal forecast error variance.

The models estimated here, however, may not necessarily have equal forecast error variance. As an alternative to the simple mean we therefore also allow for a weighting scheme. In so doing, we follow Bates and Granger (1969), who suggest combining models using weights derived from their sum of squared errors (SSE). These weights will minimise a quadratic loss function based on forecast errors, provided the estimation errors of different models are actually uncorrelated.[6] Using inverse-SSE weights produces the same weights as those derived from the inverse of mean squared errors (MSEs) computed over some recent observed sample:

$$w_i = \frac{\frac{1}{MSE_i}}{\sum_{j=1}^{n} \frac{1}{MSE_j}} \tag{1}$$

Aiolfi and Timmermann (2006) have found that model averaging across a subset of all models does well. In our application one can further improve forecast performance by trimming the model space so as to keep a small subgroup of the best performing models. Hence, we rank and give weights to all models according to forecast performance, so as

[6]The estimation errors from our model suite may, however, not all be uncorrelated.

to keep only a small subgroup of the top performing models. That is, we evaluate using the top 8 models (the 5 percent best performing models)[7] and the top 20 models (the 15 percent best performing models). Since we rank and give weights to all models in quasi real time, the collection of models in the different subgroups will then inevitably vary across time. Combining a limited number of models is quite a common strategy, as discussed in Armstrong (2001). Makridakis and Winkler (1983), for example find that the benefits from adding successively more models diminishes quite rapidly once a small number of models is added together (say five or so).

3.3 Models

To provide the forecasts, we have developed a series of models that produce sensible forecasts up to five quarters ahead. In creating the model suite, we cast our net relatively wide. We include autoregressive integrated moving average (ARIMA) models, a random walk (RW) in mean model, vector autoregressive (VAR) models, Bayesian estimated VAR models, error correction models, factor models and, finally, a dynamic stochastic general equilibrium (DSGE) model. Note that within each model class, there could be several variants with different specifications. In total, we have about 150 models. In the following, we refer to forecasts of inflation measured by annual (four quarter change) of CPIATE. In the discussion below, the models are based on seasonally adjusted quarterly data, unless otherwise stated. All the monthly models use seasonally adjusted data. Note that the forecasts from the monthly models are aggregated to quarterly frequencies.

[7] Using an even smaller model space will give very similar results, but could be interpreted as being model selection rather than model combination.

3.3.1 Autoregressive Integrated Moving Average (ARIMA) models

ARIMA models use historical variations in a single time series to provide forecasts. Generally, the form of the model is given by,

$$y_t = \alpha + \sum_{j=1}^{p} \phi_j y_{t-j} + \sum_{j=0}^{q} \theta_j \varepsilon_{t-j} \tag{2}$$

where y_t is the inflation rate and p and q are the lag order of the autoregressive (AR) and moving average (MA) terms respectively. In practice, univariate representations can often be captured by low-order AR models. We have not found that including MA-terms improves the forecasting properties. Hence, in the current version we only include AR-models.[8] In practice, we have approximately 30 quarterly AR-models, which differ by the number of lags, estimation periods and by the transformation of data (differenced, double-differenced, and trend-adjusted). In addition we estimate two monthly models; (i) An AR-model where the forecasts are first constructed on a monthly basis and then converted to quarterly frequencies (ARm) and (ii) a disaggregate monthly model (ARd) where the forecasts from each component are weighted using the expenditure weights in the CPI to form composite forecasts for CPIATE.

3.3.2 Random Walk (RW) in Mean

Random walks are often hard to beat, in particular for short term forecasts. We therefor include a random walk (RW) in mean into our model space. The forecasts are the mean of a rolling window of monthly data. The mean is updated iteratively over the forecasting horizons.

[8] In the evaluation exercise below, however, we will compare the forecast to some simple benchmarks, including an MA(1) model.

3.3.3 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,

$$X_t = A + \sum_{j=1}^{p} B_j X_{t-j} + \nu_t \tag{3}$$

where X_t is the vector of variables in the model. Building on Clark and McCracken (2008), we estimate a variety of models and lag combinations (based on various information criteria) and different transformations of the variables (such as double-differencing).[9] Bjørnland et al. (2008) give a list of of the alternative VAR models estimated (approximately 80 models), including models where we exclude one variable, effectively estimating bivariate models.

In addition to the tri- or bivariate VAR models reported above, we estimate three additional special VARs: (i) a VAR model that predicts inflation using various measures of monetary (money and interest rates) explanatory variables (VARm), (ii) a model that predicts inflation based on the exchange rate (VARe), and (iii) a monthly bivariate VAR model that that predicts inflation based on registered unemployment (VARu).

3.3.4 Bayesian VAR (BVAR) models

Bayesian methods have proven useful in the estimation of VARs as a way to overcome overfitting. In Bayesian analysis the econometrician has to specify prior beliefs about

[9] Motivated by the difficulties structural breaks present for forecasting, Hendry (2006) suggests to difference the model twice to robustify against deterministic breaks.

the parameters. The prior beliefs are then combined with the data in the VAR to form a posterior view of the parameters. In small samples prior beliefs may help to guide the parameter estimates towards sensible values, which also assists forecasting since the forecasts are nonlinear combinations of the parameters combined with a starting data vector. For the BVAR models specified here we use a direct forecasting method (eg CPIATE at time $t+h$ is regressed against variables at time t).

In addition to GDP, interest rates and inflation, we also include either the exchange rate or the terms of trade as an endogenous variable, thereby allowing for open economy considerations. Furthermore, we include a series of exogenous variables, such as oil prices and a number of foreign country-specific variables. We use a Normal-inverted Wishart conjugate prior for the model parameters, see Kadiyala and Karlsson (1997). Note that the Minnesota prior is a special case in such a framework.[10]

3.3.5 Monthly (FM) and Quartely factor (FQ) models

Factor models are estimated using large quantities of data. The ability to forecast using large data sets implies that the model builder doesn't have to take a strong stand on what series to include in the forecasting model. Once the data set is constructed, 'common factors' are estimated using a procedure of principal components. The common factors will then be linear combinations of all the data in the model that explain the highest proportion of the variance in the data. The factors are used in various equations to provide forecasts of inflation.

We estimate factor models using either monthly or quarterly data.[11]The factor model exploiting monthly data builds on work by Aastveit and Trovik (2007), who apply a

[10]For more details on the BVAR model estimated here, see Ravazzolo (2008).
[11]That is, we do not mix frequencies within the same model.

dynamic factor model to forecast Norwegian GDP. The data set contains 152 variables in total. Of these there are around 45 domestic and international financial series, and 30 price series. The main advantage of the model is its ability to efficiently exploit the ragged edge in the data set, e.g. exploit the extra information of having quarterly information using monthly data.

Our quarterly factor model is similar to that proposed by Stock and Watson (2002). We include over 200 quarterly series. Over half of these are real variables from the national accounts. Around 40 series are financial variables; mostly interest rates, money and exchange rates. However, the models good forecasting power comes from the inclusion of around 40 series of survey data.

The monthly dynamic factor model (FM) uses iterative forecasting techniques for the factors, and conditional on these factors, forecasts CPIATE. The quarterly factor model (QF) uses a direct forecasting method, (e.g. CPIATE at time $t+h$ is regressed against *factors* from data at time t). As noted by Stock and Watson (2002), this reduces the number of parameters that have to be estimated, and thereby also the potential uncertainty in the estimates.

3.3.6 Error correction model (EMOD)

We estimate an econometric (equilibrium correction) model of 13 macro variables; with specification derived from data. We use CPIATE, GDP, other domestic variables, auxiliary equations for variables such as foreign prices, interest rates and oil price. The sample period begins in 1982Q4/2001Q1 (the latter date reflecting changes in monetary policy regimes). The missing forecasts in our evaluation period are approximated with an AR(2). EMOD produces forecasts for all variables from 2003Q4 and onwards.[12]

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

3.3.7 Dynamic stochastic general equilibrium (DSGE) model

The DSGE model is a New Keynesian small open economy model. A version applied to the Norwegian economy is documented in Brubakk et al. (2006). The DSGE model is estimated using Bayesian maximum likelihood on seasonal adjusted data for mainland GDP growth, consumption growth, investment growth, export growth, employment, inflation (CPIATE), imported inflation, real wage growth, the real exchange rate (I44) and the nominal interest rate. The sample period is 1987Q1–1998Q4 (extended recursively until 2009Q1). The steady-state levels are equal to recursively updated means of the variables.

4 Empirical results

In this section we describe the main results. We first look at different ways to combine the point forecasts and then compare the individual forecasts to our preferred combined forecasts. In the end we compare the combined forecasts with Norges Bank's official forecasts, to investigate if our forecasts out-perform Norges Bank's own forecasts.

4.1 The preferred combined forecast

Table 1 shows the RMSEs for a simple AR(4) benchmark model and then displays RMSEs of the combined forecasts relative to the benchmark model's RMSEs. Hence, a value lower than one implies that the combined forecasts improve upon a simple AR(4) model in terms of forecast accuracy. The Table refers to *Top 8* as the weighted average of the eight models that have the lowest RMSE (approximately 5 percent of the models), *Top 20* as the weighted average of the twenty models that have the lowest RMSE (approximately 15

percent of the models), while *Weighted all* is the weighted average of all models. Finally, *Simple mean* refers to the simple mean of the forecasts from all individual models.

Table 1: Root mean square error relative to AR benchmark, (1999-2009).

Forecast horizon	**1-step**	**2-steps**	**3-steps**	**4-steps**	**5-steps**
AR(4) [Absolute RMSE]	[0.252]	[0.400]	[0.590]	[0.753]	[0.854]
RMSE relative to AR(4)					
Top 8	0.190	0.550	0.617	0.668	0.704
Top 20	0.230	0.577	0.640	0.691	0.729
Weighted all	0.612	0.847	0.838	0.862	0.882
Simple mean	0.948	0.964	0.945	0.940	0.946

The table illustrates that combining models improve upon the forecast accuracy of the benchmark AR(4) model. With regard to finding an optimal number of models, the combination based on the smallest sample of models (top 8 models) performs the best (in terms of having the lowest RMSE relative to the AR(4)). By including more models, the forecast accuracy deteriorates slightly, but still a weighted average of all models improve upon an AR(4). Finally, using a simple mean of all the forecasts (i.e. models are not weighted by their performance) provides forecasts that just barely improves on an AR(4). Hence, the model combination *Top 8* minimizes RMSE. In the following we therefore choose the *Top 8* model combination as our benchmark, and refer to this as *SAM* (system for averaging models).

Table 2 compares the RMSE of SAM with some alternative benchmark models, like an AR(2) model, an MA(1) model and a random walk (RW) in mean. The preferred combination out-performs all benchmark models at all horizons. However, while SAM still does much better than both an AR(2) and an MA(1), the gain of using model

combinations relative to a RW is much smaller, particularly at horizon one.

Table 2: Root mean square error of SAM (top 8 models) relative to different benchmarks, (1999-2009).

Forecast horizon	**1-step**	**2-steps**	**3-steps**	**4-steps**	**5-steps**
SAM/Benchmarks					
AR(2)	0.183	0.519	0.602	0.653	0.690
MA(1)	0.236	0.600	0.627	0.627	0.655
RW	0.875	0.838	0.788	0.718	0.637

Having seen that the forecasts from SAM out-performs the forecasts from some simple benchmarks, we next set out to compare the performance of SAM to some of the top performing individual models (not counting the benchmarks). That is, Table 3 displays the RMSE for SAM relative to the top five performing individual models (chosen ex-post) at the one-to five-quarter horizon. Again, a value lower than one implies that the combined forecast (SAM) improves upon the individual models in terms of forecast accuracy.[13] Overall, SAM has slightly lower standard deviations than the individual models for all horizons, except the ARm model at horizon 3. Note although many individual models (like VARu) performs very similar to SAM, these individual models are chosen ex post, hence the policy makers will not know in advance which model to choose. We therefore argue that averaging the forecasts from several models still makes sense.

To conclude, we have found that trimming the model space, and averaging just a small subset (eight models), results in superior forecast performance relative to a simple average of all models. Using a small group of models is consistent with what one might term the folk wisdom provided by Armstrong (2001), who advocates using at least five different

[13]Typically, different models forecast well at different horizons. Hence, we have chosen the best individual model for each forecast horizon, noting that if a model is performing best for more than one horizon, we also include the second best model.

Table 3: Root mean square error of SAM (top 8 models) relative to some individual models, (1999-2009).

Forecast horizon	**1-step**	**2-steps**	**3-steps**	**4-steps**	**5-steps**
SAM/Best models					
VARu	0.843	0.898	0.925	0.851	0.790
FM	0.815	0.756	0.740	0.703	0.670
ARm	0.750	0.791	1.005	0.786	0.878
VAR	0.183	0.585	0.708	0.811	0.878
DVAR	0.187	0.608	0.729	0.810	0.865

Note: (VARu); A VAR model with unemployment as regressor, (FM); A monthly factor model, (ARm); A monthly AR model, (VAR); A trivariate VAR using 4 lags, (DVAR); The differenced VAR model.

forecasting methods when the methods are inexpensive. Makridakis and Winkler (1983) show that, across a wide variety of series, there are diminishing returns from including more and more forecasting methods. In the expert combination literature similar advice prevails: Hogarth (1978) suggests that averaging forecasts from eight to twelve experts will have close to the 'optimal' predictive ability.

If the models are uncorrelated and have equal forecasting accuracy, then simple averages of the forecasts will be optimal, and in many empirical applications equal weighted forecast combinations are hard to improve upon. There are a number of possible explanations why trimming our model space (applying zero weights to all except the best eight models) is better than simply averaging across all models. First, we have many models that are similar in structure or data and, consequently, the forecasts from our suite of models are clearly correlated.[14] Using a simple average of all models may over-weight particular parts of the model space, resulting in insufficient weight being attached to

[14]Our very large model space contrasts with, for example, Makridakis et al. (1982) who combine six of sixteen possible models and find that the combined forecast out-performs all or virtually all individual models.

truly independent information.

Second, there is considerable variation in the forecast performance across our suite of models and at the different horizons. One model could possibly be the best for one horizon but really bad for another horizon. Thus, it is readily apparent that some models are demonstrably inferior to the best individual models, and attaching zero weight to them ensures that they have no influence on the combined forecast.

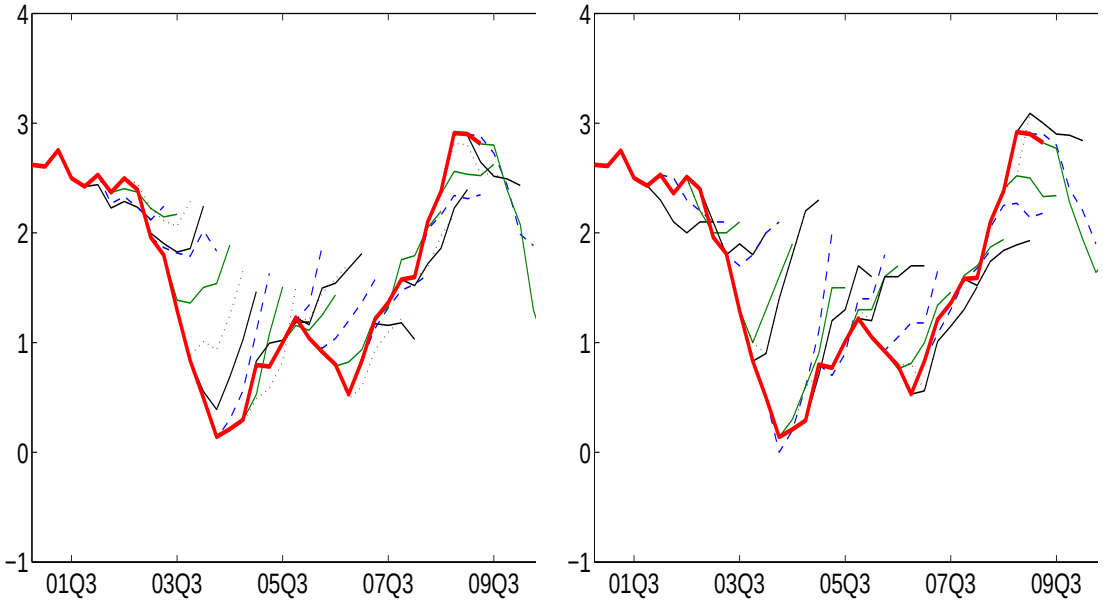
4.2 Do combined forecasts outperforms Norges Bank's official forecasts?

Below we evaluate our 'clean' technical combination of forecasts against Norges Bank's (NB) published forecasts, which do reflect additional judgement. Norges Bank's projections are based on all available information in real time, including forecast from many of the econometric models currently in SAM. The series are collected from various monetary policy reports (MPRs) from 1999 to 2009, see the appendix for more details.

Figure 1 depicts the forecasts from SAM and from Norges Bank against the actual evolution of CPIATE inflation. The left frame graphs the 1-5 step ahead forecast based on SAM. The figure graphs the inflation forecasts for various horizons against the inflation out-turns that ideally would have been predicted. Similarly, the right frame graphs the forecasts made by Norges Bank at various point in time and compares them to actual inflation. Note that for the period 1999-2000, Norges Bank's forecasts are for CPI and not CPIATE. This may give Norges Bank a disadvantage if CPI tracks systematically above or below CPIATE during this period. In the following evaluation we therefore compare the forecasts starting in 2001, when Norges Bank began publishing forecasts for CPIATE. The period 2001-2009 is also of particular interest because this is the official

inflation targeting period.

Figure 1: *Inflation forecasts at different points in time*



(a) SAM (b) Norges Bank

Note: The graphs compare actual inflation (CPIATE, yoy growth, red line) with (left) SAM forecast for inflation (hairy lines) given at different points in time and (right) Norges Banks forecast for inflation (hairy lines) given at different points in time.

Table 4 compares the RMSEs for Norges Bank's projections with the RMSEs from SAM's projections. The results illustrate that for, the period 2001Q1-2009Q1, SAM's forecasts out-perform Norges Bank's forecasts at all horizons. The gain from using SAM also increases with the horizons and illustrates the usefulness in averaging short term forecasts.[15]

Table 5 measures the bias (average forecast error) for each forecast horizon using SAM

[15]One could question whether these results to a large extent are driven by the central bank's failure to forecast the sudden drop in inflation in the first few years of the sample. Splitting the sample in two (2001-2004 and 2005-2009), we still find that SAM out-performs Norges Bank in both periods. However, we also note that Norges Bank has improved its forecast performance relatively to SAM in the later period (2005-2009).

Table 4: RMSE for inflation: Comparing Norges Bank and SAM, (2001-2009).

Model	**1-step**	**2-steps**	**3-steps**	**4-steps**	**5-steps**
SAM	0.05	0.23	0.38	0.52	0.62
NB	0.19	0.35	0.55	0.75	0.89

and Norges Bank. The table shows that both forecasts have positive bias (implying an over-prediction of inflation), but also that the bias from the forecast of Norges Bank is 3-4 times larger than that of SAM at all horizons. Assuming that Norges Bank based their policy decisions on their published inflation forecasts in the period 2001-2009, the bias implies, all other things equal, that interest rates on average would have be higher than if the forecasts from SAM had been used. How much higher obviously depends on the reaction function of Norges Bank and is an issue for further research.

Table 5: BIAS: Comparing Norges Bank and SAM, (2001-2009).

Model	**1-step**	**2-steps**	**3-steps**	**4-steps**	**5-steps**
Bias SAM	0.00	0.04	0.06	0.12	0.13
Bias NB	0.04	0.10	0.21	0.31	0.43

Finally, is the improvement in forecast accuracy reported above statistically significant? To examine this hypothesis, we use the Diebold and Mariano (1995) and West (1996) (DMW henceforth) test statistics. The DMW test statistics measure statistical differences in the forecasting performance of two competing models and can be computed as follows:

$$DMW = \frac{d}{\sqrt{\frac{2\pi \hat{f}(0)}{T}}} \tag{4}$$

where d is the mean of the difference in squared forecast errors between the two models that is compared, and $\hat{f}(0)$ is an estimator of its spectral density at frequency zero. Here we use the standard Newey-West robust estimator of the long run variance of d. Note, however, that Ashley (2003) has argued that more than 100 observations are necessary to establish significant differences in predictive accuracy across models. Hence, with fewer observations, our results should be taken with some caution.[16]

In Table 6 we first present the DMW test statistics for the forecasts to be equally accurate as the benchmark AR forecast, with corresponding p-values. Failure to reject the null hypothesis implies that forecasts do not improve the AR model significantly. We then finally compare the SAM and NB forecasts, to investigate if SAM performs significantly better than the official Norges Bank's forecasts.

Table 6: Diebold-Mariano-West test. P-values in parenthesis, (2001-2009).

Horizon	**SAM vs AR(4)**	**NB vs AR(4)**	**SAM vs NB**
1-step	-5.042 (0.000)	-1.486 (0.069)	-3.813 (0.000)
2-step	-3.409 (0.000)	-0.619 (0.268)	-1.410 (0.079)
3-step	-3.304 (0.000)	-0.423 (0.336)	-1.957 (0.025)
4-step	-3.151 (0.001)	-0.004 (0.498)	-2.309 (0.010)
5-step	-3.295 (0.001)	0.250 (0.599)	-2.807 (0.003)

Comparing the forecasts with the simple benchmark AR(4) model, we find, consistent

[16]The DMW statistics may provide non-normal critical values for asymptotic inference if the two models being compared are nested. However, we do not believe this is a substantial problem in our application.

with the results suggested above, that SAM performs significantly better than the AR (4) model at all horizons (measured at all significance levels).[17] Norges Bank's forecasts, however, only performs significantly better than the benchmark AR model at horizon 1 when measured at the 10 percent level. Finally, when comparing SAM to Norges Bank's official forecasts, SAM performs significantly better than Norges Bank at all horizons measured at the five percent level, with the exception of horizon 2, where the significance level is 10 percent.

4.3 Where can monetary policymakers add value?

Having compared Norges Bank's official forecasts to SAM's forecasts, we now analyze in more detail whether monetary policymakers can add value to the forecast process (analyzed in retrospect). In so doing, we follow Romer and Romer (2008) and test whether policymakers have useful information in the area of forecasting by estimating a regression of the form:

$$X_t = \alpha + \beta_1 S_t + \beta_2 P_t + \varepsilon_t, \tag{5}$$

where X is the realized value of inflation, S is the 'staff' (SAM) forecast and P is the policymaker forecast (published in Norges Bank's official reports). Our main interest is if β_2 is positive and significantly different from zero: Conditional on the SAM forecast, does inflation turn out higher when the policymaker forecast is higher? The results are given in Table 7 where we again focus on the recent inflation targeting period 2001-2009. Table 7 reports OLS estimates, with t-values in parenthesis.[18]

The positive coefficient on β_2 suggests that policymakers in Norges Bank add some

[17] The results also hold when SAM is compared to the MA(1) model. However, SAM only performs

Table 7: Where does Norges Bank's forecast add value (2001-2009)?

Horizon	**Constant**	β_1	β_2	R^2
2001-2008				
1-step	-0.02 (-0.83)	0.99 (18.70)	0.02 (0.41)	0.99
2-step	-0.19 (-2.01)	0.99 (6.33)	0.09 (0.60)	0.94
3-step	-0.38 (-1.89)	1.10 (5.43)	0.09 (0.41)	0.83
4-step	-0.63 (-1.71)	1.10 (5.28)	0.18 (0.71)	0.67
5-step	-0.64 (-1.05)	0.98 (4.59)	0.27 (0.82)	0.51

value to the SAM forecasts.[19] However, since the coefficient is not significantly different from zero, at no horizon do the policymakers have more useful information than the SAM forecast. The OLS estimate suggests that someone having access to both the SAM and the policymaker forecast should put no a weight of close to 1 on the SAM forecasts and a weight of zero on the policymaker forecast.

By now, a fairly large literature has found that the small sample properties of estimates of forecast combination weights can be quite poor, so that equally-weighted averages may dominate the weighted averages based on estimates of optimal weights (see e.g., Diebold and Lopez (1996), Timmerman (2006) and Smith and Wallis (2009). Romer and Romer (2008) basically ignore these issues. As a minimum, we would like to investigate the robustness of the reported combination regression results in Table 7 to alternative specifications. In so doing, we ask if our results are robust to how we have constructed our weighted average (SAM)? That is, if someone has access to the all the models but can only compute a simple average, (i.e. do not know how to rank, weight and trim the

significantly better than the random walk model at horizon 3 and above.

[18]We also tried estimating (5) using weighted least squares (WLS) (using Newey-West standard errors with three lags), to correct for any possible evidence of serial correlation. The results remained robust.

[19]Note that in the period we are examining, Norges Bank only had access to the actual SAM forecasts the last year in the sample. However, many of the individual models present in SAM have been available to the policymakers in some form or the other over the whole period.

model space in real time), should she still put a weight of close to one on the model forecasts and a weight of zero on the policymaker forecast?

No. If we redo the exercise using a simple average of all models (the 'Simple mean' from Table 1), the OLS-estimate on the forecast from the policymaker increases relative to the OLS-estimate on the model forecast. In fact, the results now suggest that one should put a weight of 0.6 on the policymaker forecast and a weight of only 0.4 on the staff (model) forecasts, at all forecast horizons.

This is interesting, as it suggests that the policymaker is able to see through some of the noise in the model space, and improves the forecast relatively to a simple average. However, once we start to weight and trim the model space, the weight on the policymaker's forecasts gradually declines, until we have reduced the model space to our top 8 models, at which point the weight on the policymaker is down to zero. Hence, weighting and trimming the model space seem essential to improving the forecasts.

Finally, how do our results compare with those of Romer and Romer (2008)? By comparing the projections for inflation from the Federal Open Market Committee (FOMC) in the US to the staff forecasts (Green book), they find β_2 to be negative and insignificant. According to Romer and Romer (2008), this implies that the FOMC do not contribute any useful information relative to the staff forecast at all.[20] We argue that our results are in line with their general findings. Norges Bank's official forecasts did not contribute any value added relative to SAM (staff) forecasts when predicting inflation one year ahead. However, as we also pointed out, this depends crucially on whether the model space has been weighted and trimmed according to forecast performance in real time. If the FOMC are faced with multiple forecasts, then they may still contribute with useful information,

[20] Although this may technically be the case, another interpretation could just be that the Policymaker's forecast are negatively correlated with the staff forecast.

using judgment to scale down the model space versus keeping to a simple average.

5 Concluding remarks

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 paper has added further evidence to this conclusion. By developing a System for Averaging Models (SAM), we have shown that there are clear advantages to averaging forecasts from several individual models when predicting inflation in Norway in the short term (up to a year).

Furthermore, the combined forecast clearly out-performs Norges Bank's own forecasts, particularly at longer horizons. In fact, someone having access to both the SAM and the policymaker forecast should put a weight of close to 1 on the SAM forecasts and a weight of zero on the policymaker forecast. However, these beneficial results depend on the degree of trimming of the model space. If a the staff only computes a simple average, (i.e. do not rank, weight and trim the model space in real time), the weight on the the judgmental forecast from the policymaker increases relative to the simple model average. Hence, weighting and trimming the model space seem essential to improving the 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.

Adolfson, M., M. K. Andersson, J. Lindé, M. Villani, and A. Vredin (2007). Bayesian

forecast combination for VAR models. *International Journal of Central Banking 3*, 111–144.

Aiolfi, M. and A. Timmermann (2006). Persistence in forecasting performance and conditional combination strategies. *Journal of Econometrics 135*, 31–53.

Akram, F. (2008). The econometric model of mainland Norway, EMod, on seasonally adjusted data. Mimeo, Norges Bank.

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

Armstrong, J. S. (2001). Combining forecasts. In J. S. Armstrong (Ed.), *Principles of Forecasting: A Handbook for Researchers and Practitioners*, pp. 417–440. Norwell, MA: Kluwer Academic Publishers.

Ashley, R. (2003). Statistically significant forecasting improvements: How much out-of-sample data is likely necessary? *International Journal of Forecasting 13*, 229–239.

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.

Bjørnland, H. C., A. S. Jore, C. Smith, and L. A. Thorsrud (2008). Improving and evaluating short term forecasts at the Norges Bank. Staff memo 2008/4, Norges Bank.

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.

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.

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

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

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.

Diebold, F. X. and R. S. Mariano (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics 13*, 134–144.

Hendry, D. F. (2006). Robustifying forecasts from equilibrium-correction systems. *Journal of Econometrics 135*(1-2), 399–426.

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

Hogarth, R. (1978). A note on aggregating opinions. *Organizational Behaviour and Human Performance 21*, 40–46.

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

Jore, A. S., J. Mitchell, and S. P. Vahey (2007). Combining real-time VAR density forecasts with uncertain instabilities. *Mimeo*. Paper presented at 3rd Annual Workshop on Macroeconomic Forecasting, Analysis and Policy with Data Revision.

Kadiyala, K. R. and S. Karlsson (1997). Numerical methods for estimation and inference in Bayesian VAR-models. *Journal of Applied Econometrics 12*(2), 99–132.

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.

Makridakis, S., A. Andersen, R. Carbone, R. Fildes, M. Hibon, R. Lewandowski, J. Newton, E. Parzen, and R. Winkler (1982). The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *Journal of Forecasting 1*(2), 111–153.

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

Makridakis, S. and R. Winkler (1983). Average of forecasts: Some empirical results. *Management Science 29*, 987–996.

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

Ravazzolo, F. (2008). Unrestricted Bayesian VAR to forecast quarterly Norwegian macroeconomic series. Mimeo, Norges Bank.

Romer, C. D. and D. H. Romer (2008). The FOMC versus the staff: Where can monetary policymakers add value? *American Economic Review: Papers and Proceedings 98*(2), 230–35.

Smith, J. and K. F. Wallis (2009). A Simple Explanation of the Forecast Combination Puzzle. *Oxford Bulletin of Economics and Statistics 71*(3), 331–355.

Stock, J. and M. Watson (2002). Macroeconomic forecasting using diffusion indexes. *Journal of Business and Economic Statistics 20*(2), 147–162.

Stock, J. and M. Watson (2003). *Introduction to Econometrics*. Addison Wesley, Boston, MA.

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

West, K. D. (1996). Asymptotic inference about predictive ability. *Econometrica 64*, 1067–1084.

6 Appendix - Norges Bank's forecast

Norges Bank's one step ahead forecast is made in the same quarter that it is forecasting, implying that they will have observed at least 1 or 2 months of CPIATE when making the one-step ahead forecasts. This gives Norges Bank an advantage in the forecast competition compared to our quarterly models, that use no information within the current quarter. However, for the five monthly models included in our model suite, the information content is the same.[21] On the other hand, Norges Bank has a disadvantage as they publish forecasts only three times a year. Overall, the net benefit is therefore probably

[21]The five models that uses monthly information are a monthly factor (FM) model, two monthly AR models (ARm and ARd), a monthly VAR model with unemployment and a random walk (RW) in mean.

close to zero. The following describes in detail how the forecasts are collected from the various Monetary Policy Reports (MPRs).

- The forecasts for Q1 are taken from MPR1 (published in February/March), when CPI is known for January, giving Norges Bank one month information **advantage** relative to the quarterly models in SAM.

- 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 the quarterly models in SAM.

- Regarding the forecast for Q3, Norges Bank does not publish any forecast in the third quarter. Most of the forecasts for Q3 are therefore taken from MPR2 (published at the end of June), giving Norges Bank one month information **disadvantage** relative to SAM (since June figures for CPI were yet not known).

- The forecasts for Q4 are from MPR3 (published end of October), when no information for Q4 is known. Hence, in Q4, Norges Bank has no advantage over SAM.

To sum up, this gives Norges Bank an advantage over the quarterly models in SAM in two of the quarters (Q1 and Q2) and a disadvantage in one quarter (Q3). For Q4, the information content is about the same. However, compared to the five monthly models also included in SAM, there is no information advantage for Norges Bank.

Another issue is that while there are essentially no revisions to inflation, there are probably revisions to some of the other indicators (GDP growth, the output gap, etc.) used by the Norges Bank in forming its official inflation forecast. Such revisions may put the Norges Bank forecasts at a bit of a disadvantage relative to the time series models that are based entirely on the most recent vintage of data on GDP, etc. However, as

some of the best performing time series models do not use information from GDP at all, we claim that this source of disadvantage is minimal.

Centre for Applied Macroeconomic Research (CAMAR)

The objective of CAMAR is to provide high quality research and analysis into the field of macroeconomics, as well as financial issues.

The research activities of CAMAR will be broad and will encompass all elements pertaining to the analysis of macroeconomic data.

BI Norwegian School of Management
Centre for Applied Macroeconomic Research (CAMAR)
N-0442 Oslo

http://www.bi.no/camar

CAMAR Working Paper Series
ISSN: 1892-2198