Evaluating ensemble density combination – forecasting GDP and inflation*

Karsten R. Gerdrup[**,a], Anne Sofie Jore[a], Christie Smith[b], Leif Anders Thorsrud[a]

[a] *Economics Department, Norges Bank, Bankplassen 2, 0151 Oslo, Norway.*
[b] *Economics Department, Reserve Bank of New Zealand, PO Box 2498, Wellington 6011, New Zealand.*

Abstract

Forecast combination has become popular in central banks as a means to improve forecasts and to alleviate the risk of selecting poor models. However, if a model suite is populated with many similar models, then the weight attached to other independent models may be lower than warranted by their performance. One way to mitigate this problem is to group similar models into distinct 'ensembles'. Using the original suite of models in Norges Bank's system for averaging models (SAM), we evaluate whether forecast performance can be improved by combining ensemble densities, rather than combining individual model densities directly. We evaluate performance both in terms of point forecasts and density forecasts, and test whether the densities are well-calibrated. We find encouraging results for combining ensembles.

Key words: Density combination; model combination; clustering; ensemble density; pits;
JEL: C32, C52, C53, E52

*We thank Christian Kascha, Francesco Ravazzolo and Hilde C. Bjørnland for helpful comments, and James Mitchell for providing codes for *pits* tests.

**Corresponding author: karsten.gerdrup@norges-bank.no; Ph +47 2231 6440; Fax +47 2242 4062.

1. Introduction

In this paper, we investigate a two stage approach to combine forecast densities. In the first step, density forecasts for Norwegian inflation and GDP from a number of models are combined into 'ensembles'. The models in each of the ensembles have some common element, for example the same information set or model structure, which might result in correlated forecasts. In the next step, we combine the ensemble density forecasts. We find that the out-of-sample performance of the ensemble density combination is better than combining all the density forecasts in one step.

Forecast combination can be motivated by several reasons. For example, unknown instabilities might favour one model over another at different points in time (Clark and McCracken (2007) and Jore et al. (2009)). Forecasting models may also be subject to idiosyncratic biases, and combining forecasts may help to average out these unknown biases. Forecast combination has also proven to be successful empirically: see the summaries of the M-competitions in Makridakis et al. (1982), Makridakis et al. (1993), and Makridakis and Hibon (2000). See also Timmermann (2006) for a thorough discussion of theoretical and empirical motivations for combining forecasts.

The literature on forecast combination builds to a large extent on Bates and Granger (1969), where a main conclusion is that a combination of two forecasts can yield lower mean-square forecasts error than either of the original forecasts when optimal weights are used. Forecast combination is expected to yield the largest improvements if the sets of forecasts contain truly independent information. Estimating optimal Bates-Granger weights is usually undesirable because of the large number of models relative to the evaluation period, making it infeasible to estimate the full covariance matrix of forecasts. A simple average of point forecasts has empirically been found to be effective, the so-called "forecast combination puzzle", see Stock and Watson (2004). Because of the effect of finite-sample error in estimating the weights, a recommendation from Wallis and Smith (2009) is to ignore the covariance between forecast errors and base the estimates on mean squared forecast errors alone if estimated weights are to be used. Aiolfi and Timmermann (2006) find, however, that forecasting performance can be improved by first sorting models into clusters based on their past performance, second by pooling forecasts within each cluster, and third by estimating optimal weights on these clusters (followed by shrinkage towards equal weights).

Our paper is close to Aiolfi and Timmermann (2006) in the sense that we combine models in more than one stage. However, our focus is mainly on density forecasting since central banks often communicate through fan charts.[1] Furthermore, we are particularly interested in the case of a model suite which is populated by a wide range of model types which are typically considered by central banks (survey-based models, univariate autoregressions, factor models, dynamic stochastic general equilibrium model etc.) when they form their views on the future trajectory of the economy. Instead of grouping models according to past forecast performance, we group models that share the same information set or model structure. The idea is that the lower the degree of information overlap, the more useful is a combination

[1] Our paper is also close to Bache et al. (2009a) which combine forecast densities from different specifications of Vector Autoregressions (VARs) and a Dynamic Stochastic General Equilibrium Model (DSGE). A result from this paper is that the DSGE is poorly calibrated and that it receives a low weight in the combination. Bache et al. (2009b) finds that a combination of several DSGE models with different break dates outperforms autoregressive benchmarks.

of forecasts likely to be. Clemen (1987) argues that: "In the case of aggregating economic forecasts, say, one might like to select a group of forecasters who exhibit as little dependence as possible. Such a group would be a heterogeneous collection of forecasting approaches, theoretical persuasions, and different views of available data."

In the following sections, we first briefly describe the model suite, and discuss the rationale for combining ensemble densities. Then we describe the weighting schemes used. We evaluate the out-of-sample performance of the ensemble density combination with that of some other main alternatives. We use several tests based on probability integral transforms to assess if the composite densities are well specified. Following Bjørnland et al. (2009a) we investigate whether the ensemble combination forecasts perform better than Norges Bank's own forecasts for inflation.

2. Model suite and ensembling

2.1. Model suite in Norges Bank's system for averaging models (SAM)

SAM is used to provide model-based density forecasts for GDP Mainland-Norway and CPIATE (consumer prices adjusted for taxes and without energy) in the monetary policy process. The models in SAM vary both in terms of structure and information set. For GDP there are 236 models, and for CPIATE there are 165. A large number of models forecast both GDP and inflation, but many models are specific to each of the two target variables.

Table 1: Models for forecasting GDP Mainland-Norway

Ensemble	Description	No of models	Models' ratio
eRegN	Regional network model	1	0.004
eTstruc	Term structure models	4	0.017
eMI	Monthly indicator models	2	0.008
eFM	Factor models	2	0.008
eEmod	Macro model (VECM)	1	0.004
eDSGE	Macro model (DSGE)	1	0.004
eBVAR	Bayesian VARs	10	0.042
eUniv	Univariate autoregressive models (ARs)	38	0.160
eVAR2	VARs with GDP and inflation	36	0.152
eVAR3	VARs with GDP, inflation and interest rate	72	0.304
eTNSG	Bivariate VARs with household surveys	6	0.025
eBuild	Bivariate VARs with building and construction	10	0.042
eOrd	Bivariate VARs with orders to manufacturing	4	0.017
eEmpl	Bivariate VARs with employment data	10	0.042
eMny	Bivariate VARs with money and credit	7	0.030
eBTS	Bivariate VARs with Business Tendency Survey	33	0.139
Sum		237	1

For GDP, SAM has a family of bivariate VAR models with variable indicators as explanatory variables. The indicators encompass building and construction, industrial orders, employment developments, monetary aggregates, household surveys and business tendency survey. One model uses information from Norges Bank's own regional network on expected

Table 2: Models for forecasting CPIATE

Ensemble	Description	No of models	Models' ratio
eDisAgg	ARs for CPI-disaggregates	1	0.006
eMth	Monthly VARs	3	0.018
eFM	Factor models	3	0.018
eEmod	Macro model (VECM)	1	0.006
eDSGE	Macro model (DSGE)	1	0.006
eBVAR	Bayesian VARs	10	0.060
eUniv	Univariate autoregressive models (ARs)	39	0.234
eVAR2	VARs with GDP and inflation	36	0.216
eVAR3	VARs with GDP, inflation and interest rate	72	0.431
eMny	VAR with GDP and money	1	0.006
Sum		167	1

production.[2] The model suite for forecasting GDP also includes factor models which extract factors from monthly or quarterly information, term structure models, and two monthly indicator models which use monthly manufacturing production, employment, retail sales, hotel statistics and building starts as explanatory variables. For CPIATE we have some models which forecast monthly inflation, eg. a model which uses the 14 components of CPI to forecast CPIATE. The monthly forecasts are aggregated to quarterly frequencies and evaluated against the quarterly models.

For some model classes, we have several variants with different specifications because we do not know the true data generating process. We have for example 36 variants of a bivariate VAR with GDP and inflation as explanatory variables, which involve different degrees of differencing, detrending, number of lags, and estimation periods. Since the information set and model class is the same in all these models, there is a risk that several models will provide highly correlated forecasts.

The number of bivariate VARs with GDP and different indicators as explanatory variables are in total 70. Some of these indicators may provide different signals about the future trajectory of the economy, but some indicators may be highly correlated, like the different sub-indices of the Business Tendency Survey (in total 33 different indicators). The different groups of models according to our grouping are listed in table 1 and 2. The different model groups are as mentioned labeled 'ensembles'.

Models can be combined in many ways. If we use equal weights, then each model would get a weight of 0.004 and 0.006 for GDP and inflation, respectively. The average would, however, be heavily influenced by some types of models or information sets. One way of looking at this is to calculate the ratio of models within each class or ensemble relative to the total number of models, as done in table 1 and 2. The tables show that using equal weights would mean that the average would be heavily influenced by AR models, VARs using GDP and inflation as explanatory variables, and VARs using GDP, inflation and interest rate as explanatory variables. When forecasting GDP, the average would also be heavily influenced by the bivariate VARs with Business Tendency Survey data as explanatory variables. These models would drown out signals from many other types of models, for example from the two macro models, or information sets, for example from the regional network.

[2] See Brekke and Halvorsen 2009 for more information on the regional network.

Using weights that adapt to past forecast performance (for example inverse-MSE weights or score-based weights, see section 3) would not remove this problem because similar models could potentially still get a higher weight than warranted if they are many. The method proposed by Winkler (1981) to take account of the covariance between the forecast errors is infeasible in our model suite because of near-singularity in the variance-covariance matrix.

Trimming the model suite (for example down to eight models as in the version of SAM until September 2009) would also not solve this problem, but could potentially exacerbate it since we could end up with having models from only one model class or ensemble in the combination and thus models that are not really independent.

2.2. Why ensembling might be desirable

The rationale behind the pre-grouping models and combining them in two steps can be described with a simple, stylised example. Suppose you have two models that have statistics (eg scores) that are used to compute weights; higher values of the statistic are better. Let M1 denote model 1 and M2 denote model 2. Denote the scores of the models with q1 and q2 respectively.

Model	Statistic/score
M1	q1=0.9
M2	q2=0.1

Given these statistics, a fairly natural way to compute a weight for M1 is $0.9 = 0.9/(0.1+0.9)$. The weight on M2 is simply the complement $0.1 = 0.1/(0.9 + 0.1)$.

Suppose that you repeated M2 99 times and computed the weights again based on these 101 models, i.e you want to combine forecasts from all the models.

Model	Statistic/score
M1	0.9
M2	0.1
M3	0.1
⋮	⋮
M101	0.1

The weight on M1 will now be: $0.0826 = 0.9/(0.9+100\times0.1)$, and the remaining M2,. . . ,M101 models will each have weight $0.0092 = .1/(0.9+100\times0.1)$. Since these latter models are really just a single model, forecasts from this model would receive a weight of 0.92 in the combined forecast. This simple example demonstrates that adding in many similar models does reduce the probability attached to other models, even if the other models are demonstrably better in terms of their performance.

2.3. Remedy: ensembling or prior weights

Suppose we succeed in collecting the similar models in ensembles by combining. In this case, we treat the single model M1 as an ensemble. The second ensemble has 100 models that are equal, and the average statistic of this ensemble is of course 0.1. The first ensemble's

statistic is 0.9. When we use these ensemble statistics to derive weights, we are back with the original model space and hence the same weights as if we had averaged only the two models M1 and M2.[3]

The use of ensembles is not the only way one might seek to address the problem outlined above. An alternative approach would be to provide explicit priors on the models, explicitly assigning a given amount of prior probability to a class of models and spreading that probability across the members of the class. Suppose that prior probability of 0.5 is attached to model M1 and prior probability of 0.5 is 'diluted' over models M2,...,M101 (so-called dilution priors, see George 1999). Then the weight on M1 would be $(0.5 \times 0.9)/(0.5 \times 0.9 + (0.5/100) \times 0.1 \times 100) = 0.9$. This is the same weight as would have been applied to M1 when it is only combined with M2 (and M3,...,M101 are excluded).

Yet another approach, as described by Stock and Watson (2006), is to use factor models to extract common factors from the forecasts provided by the suite of models. One advantage of ensembling over the factor model approach is that we are able to relate developments in the grand ensemble forecast to developments in the data more easily, and this also helps to motivate the forecasts for policy-makers.

3. Weighting schemes

3.1. Type of aggregation

We have to determine both the type of aggregation (or pooling) and the construction of the weights before we can combine models.

The simplest aggregation method is a so-called linear opinion pool in which the combined density is a linear combination of the densities from the individual models or ensembles.:

$$P(y_t) = \sum_{i=1}^{n} w_i P_i(y_t) \tag{1}$$

where $P(y_t)$ is the combination density, $P_i(y_t)$ is a density obtained from the i^{th} model or ensemble and w_i is the weight on the i^{th} model, with $w_i \in [0, 1]$ and $\sum_{i=1}^{n} w_i = 1$. This combined density is clearly non-negative and integrates to one.

An alternative method is the log opinion pool, which is a geometrically weighted average of the individual models or ensemble densities:

$$P(y_t) = K \cdot \Pi_{i=1}^{n} P_i(y_t)^{w_i} \tag{2}$$

where K is a constant to ensure that the log opinion pool integrates to 1. The two combination schemes have different properties, but no scheme is obviously superior to the other. In general, log opinion pools are less dispersed than linear opinion pools, and are typically uni-modal (see Genest and Zidek (1986)). Linear opinion pools, on the other hand, may be multi-modal. The geometric weighting of a log opinion pool means that the combination will have zero probability mass in a region if a single density says that the region has zero probability (contributing to its reduced dispersion). This property means that a single density can have a material impact on the combination density, and explains why it is necessary to

[3] An simpler solution in this stylised example is of course to trim the model suite down to two models again if the two distinct models can be identified.

be careful about which models are included in a log opinion pool. A log opinion pool will be normally distributed if the individual densities underpinning it are normally distributed; see the appendix in Kascha and Ravazzolo (2008). Combining normally distributed probability distributions in a linear opinion pool results in a normal mixture distribution.

3.2. Construction of weights

We consider three types of weights, all of them common in the literature. The simplest choice of weights is simply to equally-weight all of the models or ensembles that enter the combination. Equally-weighted combinations have been found to be surprisingly effective, at least for point forecasting, see Clemen (1989). Bates and Granger (1969) propose another alternative, 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 that the estimation errors of different models are actually uncorrelated. One of the objectives with grouping models is to generate ensembles that are less correlated than the individual models are. 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:

$$w_i = \frac{\frac{1}{MSE_i}}{\sum_{j=1}^{n} \frac{1}{MSE_j}} \tag{3}$$

We also consider the so-called log score weights. The weights use (loosely speeking) the *probability* that a model or ensemble could have generated the evaluation data to help form the weights and can be specified as:

$$w_i = \frac{\exp\left(\log(P_i(\underline{y}))\right)}{\sum_{j=1}^{n} \exp\left(\log(P_j(\underline{y}))\right)} \tag{4}$$

where $\underline{y} = (y_1, \ldots, y_T)'$, and $\log(P_i(\underline{y})) = \sum_{t=1}^{T} \log(P_i(y_t))$.

Bjørnland et al. (2009b) demonstrates that the score for a model with normally distributed errors is a transformation of the MSE, modified by the sample size and the unknown variance. The variance parameter and the sample size are crucial, since the weights from the log score approach will depend on whether the variation of the realized errors was big relative to the variation expected on average. As the sample size increases the MSE and unknown variance should coincide. If two models have different variance then eventually the model with smaller variance will accumulate greater and greater weight; ultimately all weight will be put on the best model or ensemble in the suite. As Hall and Mitchell (2005) observe, even if the suite of models contained the true data generating process it would not receive a weight of 1 given a finite sample of data. When the weights are based on log scores, asymptotically the best model in the suite of models will receive weight of 1 even if it is not the truth. However, Bates-Granger weights would not put all weight on the correct model even asymptotically, which can be seen from equation (3).

4. Evaluation of the ensemble density combination

4.1. The experiment

The experiment in this section is to compare the ensemble density combination (also referred to as 'grand ensemble' in the following) with alternative weighting schemes: 1) combination of all models, 2) combination of the eight best models (as in the previous version of

SAM until September 2009), and 3) selection strategy where we try to pick the 'best' model. We look at two measures of quasi out-of-sample forecasts performance: root mean squared forecast errors (RMSFE) and logarithmic scores.

The models are first estimated up to 1998Q4, and then the estimation window is recursively expanded in quasi-real time.[4] The estimation of the models and the evaluation of forecasts are based on the most recent vintage of data, since Norwegian real-time data are still in the process of being compiled and implemented in SAM.

The models 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 and models may be good at predicting only one of the variables, we have chosen to use univariate weights.

We have made the following choices regarding weighting schemes for the grand ensemble:

1. Linear pool and inverse-MSE weights for weighting the ensembles.
2. Linear pool and score-based weights for weighting models within each ensemble.

First, linear pool and inverse-MSE weights is a very common weighting scheme in forecast combination literature. More importantly, we also wanted to make incremental changes to our combination scheme relative to the previous version of SAM, SAM8, which used this weighting scheme. This means that each ensemble density forecast receive a weight depending on past out-of-sample point forecast performance in the grand ensemble.[5] In the individual model combination case, we also use a linear pool and inverse-MSE weighting scheme. In the selection case, models are ranked according to their MSE, and the best model is picked in quasi-real time.

Second, we use a linear pool and score-based weights for individual models within each ensemble.[6] Score-based weights adapt quickly to changes in models' performance as indicated above, and the weights can be considered as approximations to the models' posterior probabilities.[7] The weight on each individual model will thus depend on each model's past out-of-sample forecast performance relative to other models in the ensemble.

4.2. Recursive out-of-sample forecasts

The figure panel 1 show recursive inflation forecasts together with actual inflation (solid lines) for the different weighting schemes. Recursive forecasts for GDP Mainland-Norway together with the vintage for GDP that was published in May 2009 are depicted in figure panel 2.

[4] It is comparatively uncommon, given data publication lags for gross domestic product, for a data set to be balanced. When the data set is unbalanced the analysis proceeds as if the same unbalancedness prevails each time the model is estimated. For example, at the beginning of the recursive estimation some data points in 1999Q1 will be treated as known, and this unbalancedness will be repeated as the data sample is recursively expanded.

[5] The weights take into account publication lags, only using data that would be known at a given date.

[6] We have one more degree of freedom in choosing weighting scheme in the grand ensemble case than in the individual model combination case because we also can choose the within-ensemble weighting scheme. A priori one could argue that a comparison of the grand ensemble with individual model combinations or SAM8 is unfair because the comparison may be influenced by within-ensemble weights. We will later see, however, that the performance of the grand ensemble is not sensitive to the choice of within-ensemble weighting scheme.

[7] Hall and Mitchell (2008) suggests minimising the Kullback-Leibler information divergence by using a linear pool and score-based weights.

Forecasts for both inflation and GDP tend to return towards a historical mean in all weighting schemes. This tendency seems somewhat stronger in the individual model combination case when forecasting GDP, in particular in the high growth period from 2004 to 2007. The reason for this is that the individual model combination is dominated by quarterly AR-models, bivariate VARs and tri-variate VARs, which in our case have a strong tendency of returning towards a historical mean.

It is not clear from these figures whether one combination scheme is superior to the others. None of the combination schemes were good at predicting the downturn in GDP in 2003 or the sharp decline in 2008Q4-2009Q1.

Selection often produces forecasts that jump from one horizon to the next because a different model is selected for the different horizons. Selection also lead to high forecast errors for some recursive forecasts because the model suite does not include models that are superior to the others over time.

All combination schemes forecasted that inflation would continue to fall quite rapidly in the second half of 1999 and first half of 2000, and did not forecast the rise in inflation in 2001. The grand ensemble seems to have been better at forecasting the upturn in inflation in 2005 and 2007 than the individual model combination, but not necessarily better than SAM8 or selection. We can also see that the weighting schemes give very different forecasts for inflation in the second half of 2009 and the beginning of 2010. The ensemble forecast higher inflation than the other alternatives. However, the jury is still out for these forecasts.

4.3. Evaluation of results

The performance of the weighting schemes – selection, SAM8, individual models, and grand ensemble – are reported in table 3 for inflation and table 4 for GDP. The grand ensemble has better performance than all the other weighting schemes when forecasting CPIATE for the horizons 2-5, both in terms of RMSFE and logarithmic scores. However, for the first horizon, it has the same RMSFE as SAM8 and somewhat lower logarithmic score than SAM8.

The results are a little more mixed when forecasting GDP. The grand ensemble has about the same, or somewhat better, performance than the other weighting schemes in terms of RMSFE for the first four horizons. Using selection when forecasting GDP would produce poor density forecasts at medium-term horizons (3-5 quarters ahead), since the logarithmic score at those horizons are lower than the other weighting schemes. The same is true when forecasting inflation 4-5 quarter ahead.

It is worth noting that constructing a grand ensemble (which include all models) typically is superior to weighting all the individual models together, which seem to suggest that the grand ensemble provide a better hedge against uncertain instabilities. This is certainly the case when the objective is to minimise RMSFE, both when forecasting CPIATE and GDP. However, the performance of the great ensemble is about the same as the individual model combination scheme when evaluating GDP in terms of logarithmic scores.

The performance over time of the different weighting schemes 1-4 quarter ahead, are illustrated in the figure panels 3-4. The performance of the grand ensemble is more stable over time than the other weighting schemes, and produces quite early in the evaluation period relatively low forecast errors and good density forecasts (i.e. relatively high logarithmic scores). Selecting the best model in real time lead at some points in time, and for some forecast horizons, to high forecast errors and poor density forecasts. SAM8 has at some horizons about the same performance in terms of point forecasts as the grand ensemble both when forecasting CPIATE and GDP, at least at the end of the evaluation period, suggesting

that SAM8 has improved more than the grand ensemble after starting out with higher forecast errors at the beginning of the evaluation period. The results are more mixed for SAM8 when considering logarithmic scores, depending on the variable of interest (CPIATE or GDP) or horizon.

It is not obvious from table 3 and 4 to what extent differences in RMSFE and logarithmic scores are large or not. In the next two sections, we investigate this further, first by calculating the weights that we would have attached to the different schemes if we treated them as different models or forecasters, and second by illustrating and testing the probability integral transforms (pits) of the weighting schemes.

4.4. Weights derived from combining the weighting schemes

The results for weights are reported in table 5 for inflation and table 6 for GDP. Both inverse-MSE weights and score-based weights are reported.

Generally, score-based weights discriminate much more between the different weighting schemes than MSE-based weights. The latter attach nearly close to equal weights to the different weighting schemes. Score-based weights would, in the case of CPIATE, put all weight on SAM8 on the first horizon, and most weight on the grand ensemble for the other horizons. When forecasting GDP, score-based weights would put almost all weight on selection and SAM8 on the first horizon, which indicate that the model suite include a few models with good performance for that horizon. The advantage of combining many models becomes clearer on horizon 2-5 when we use score-based weights to forecast GDP. Most weight is on the grand ensemble and the individual model combination.

4.5. Testing the pits

The pits summarize the properties of the densities, and may help us to judge whether the densities are biased in a particular direction, and whether the width of the densities has been roughly correct on average. The pits are the ex ante inverse predictive cumulative distribution evaluated at the ex post actual observations, see Geweke and Amisano (2008). The pits of a forecasting model should have a standard uniform distribution if the model is correctly specified.

The figure panels 5 and 6 depict the pits of the different weighting schemes when forecasting CPIATE and GDP, respectively, by showing histograms with bars for each percentile.

A density forecast is optimal if the density is correctly specified (correctly conditionally calibrated). The pits can be used to test the density forecast. A density is correctly specified if the pits are uniform and, for one-step ahead forecasts, independently and identically distributed. Accordingly, we may test for uniformity and independence at the end of the evaluation period. Several candidate tests exists, but few offer a composite test of uniformity and independence together, as would be appropriate for one-step ahead forecasts. In general, tests for uniformity are not independent of possible dependence and vice versa. Since the appropriateness of the tests are uncertain, we conduct several different tests. See Hall and Mitchell (2004) for elaboration and description of different tests.

We use a test of uniformity of the pits proposed by Berkowitz (2001). The Berkowitz test works with the inverse normal cumulative density function transformation of the pits. Then we can test for normality instead of uniformity. For 1-step ahead forecasts, the null hypothesis is that the transformed pits are identically and independently normally distributed, iid N(0,1). The test statistics is χ^2 with three degrees of freedom. For longer horizons, we do not test for independence. In these cases, the null hypothesis is that the transformed pits are

identically, normally distributed, N(0,1). The test statistics is χ^2 with two degrees of freedom. Other tests of uniformity are the Anderson-Darling (AD) test (see Noceti et al. (2003)) and a Pearson chi-squared test suggested by Wallis (2003). Independence of the pits are tested by a Ljung-Box test, based on autocorrelation coefficients up to four for one-step ahead forecasts. For forecast horizons h>1, we test for autocorrelation at lags equal to or greater than h.

Table 7 and 8 show the p-values of the test for the different weighting schemes for horizons 1 to 5, for CPIATE and GDP, respectively. P-values equal to or higher than 0.05 mean that we can not reject the hypothesis that the combination is correctly calibrated at a 95% significance level.

For CPIATE, the grand ensemble passes all tests for all horizons, except the Berkowitz test for one-step ahead forecasts. The frequency of not being able to reject the null across all horizons and all tests is 0.95. SAM8 and individual model combination also seem to be fairly well calibrated, with frequencies of non-rejections at 0.90. Selection stands out with a frequency of non-rejections of 0.40.

When forecasting GDP, all combinations appear to be well calibrated. We only get rejections of the hypothesis that the densities are well calibrated for the grand ensemble and individual model combination. This occurs for the Berkowitz test for both models for the one-step ahead forecast and for the grand ensemble for the second-step ahead forecast. This suggests that we have one or a few models in the suite with well-calibrated densities for short-term forecasting.

4.6. Sensitivity analysis

We have one more degree of freedom in choosing weighting scheme in the grand ensemble case than in the individual model combination case because we also can choose the within-ensemble weighting scheme. A priori one might think that the comparison of the grand ensemble with individual model combinations or SAM8 is unfair because the comparison may be influenced by the within-ensemble weights. In table 9 and 10 we have summarized RMSFE's and logarithmic scores for CPIATE and GDP, respectively, when we use different schemes, both within-ensemble and between ensembles. The different combination schemes differ first by the weighting scheme within each ensemble, either inverse MSE-weights, score-based weights, or equal weights. Linear pooling is used for the three different within-ensemble weighting schemes. Then for each within-ensemble weighting scheme, we present six between-ensemble weighting schemes: two pooling methods (linear and logarithmic) and three different weights (inverse MSE-weights, score-based weights, and equal weights).

The main result from these tables are that the performance of the grand ensemble in terms of RMSFE are not sensitive to the choice of within-ensemble weights. The alternatives 'Mse Linear Mse' and 'Equal Linear Mse' have relative RMSFE's close to 1, and the performance is therefore quite similar to the benchmark alternative ('Score Linear Mse') used in this paper.

More important is the weighting scheme used between ensembles. It is for example possible to reduce the RMSFE when forecasting CPIATE by about 40% at horizon 1 and 5% at horizon 2 if we use a logarithmic pool instead of a linear pool, see 'Score Log Mse', 'Mse Log Mse' and 'Equal Log Mse' in table 9. Equal weights between the ensembles have about the same performance in terms of RMSFE when forecasting CPIATE as inverse MSE-weights as long as we use a logarithmic pool, see alternatives 'Score Log Equal', 'Mse Log Equal' and 'Equal Log Equal'.

Average logarithmic scores when forecasting CPIATE can also differ, and may change both when we change the within-ensemble weighting schemes and the between-ensemble weighting

schemes. Based on a judgement of relative logarithmic scores, it seems like a good idea to use score-based weights within ensembles and use a logarithmic pool when weighting the ensemble forecasts, see alternatives 'Score Log Mse', 'Score Log Score' and 'Score Log Equal' in table 9.

Furthermore, it is possible to reduce the RMSFE when forecasting GDP by 12% at horizon 1 by using score-based weights between the ensembles (regardless of the within-ensemble weighting scheme), see table 10. It is also possible to increase relative scores by using score-based weights between the ensembles regardless of use of within-ensemble weighting scheme or use of pool to combine forecasts between ensembles.

5. Comparison with Norges Bank's own forecasts for CPIATE

In this section we compare the grand ensemble forecasts for inflation with those of Norges Bank, both point forecasts and density forecasts. We reduce the evaluation period to the period Q1 2001 - Q1 2009 because Norges Bank only forecasted CPI and not CPIATE prior to that period. Norges Bank published three reports (Inflation Reports/Monetary Policy Reports) per year in that period, whereas we make at least four forecasts per year. The forecasts for Q1 are taken from IR1/MPR1. Since these reports were published at the end of March, Norges Bank could utilize information on prices in both January and February (in the period 2001 to 2004, however, only January was known). The forecasts for Q2 are taken from IR2/MPR2 (published at the end of June). CPIATE for April and May is therefore known before the reports are published. Norges Bank did not publish one-step ahead forecasts for Q3 in the evaluation period we are considering. Most of the forecasts for Q3 are therefore taken from IR2/MPR2. This means that forecasts for Q3 are two-step ahead forecasts made in IR2/MPR2. This gives Norges Bank an information disadvantage relative to the grand ensemble. The forecasts for Q4 are from IR3/MPR3 which was published at the end of October when no information for Q4 is know.

The grand ensemble appears to perform better than Norges Bank's own forecasts, thus confirming the results of Bjørnland et al. (2009a), that forecast combinations improve performance. The upper panel in figure 7 shows that the grand ensemble has more precise point forecasts. This is particularly due to some large errors that Norges Bank encountered when forecasting inflation in 2003 and 2004. Norges Bank forecasted that inflation would return quite quickly to more normal levels after falling rapidly from the end of 2002, see figure 8. After that period, Norges Bank's forecast errors have been relatively low, leading to a falling RMSFE.

The gain from using a grand ensemble seems to be particularly large when evaluating density forecasts. The lower panel in figure 7 illustrates that the logarithmic scores of Norges Bank's own forecasts have deteriorated compared to the grand ensemble. This would suggests that the grand ensemble densities are much closer to the true, but unknown, density.

6. Conclusion

In this paper, we investigate a two stage approach to combine forecast densities. In the first step, density forecasts for Norwegian inflation and GDP from a number of models are combined into 'ensembles'. The models in each of the ensembles have some common element, for example the same information set or model structure, which might result in correlated forecasts. In the next step, we combine the ensemble density forecasts. We find that the out-of-sample performance of the ensemble density combination is better than combining all the density forecasts in one step.

We evaluate forecasting performance using quasi out-of-sample root mean square forecast errors and logarithmic scores, and find encouraging results for the ensemble density combination approach. Overall, the ensemble density combination outperforms alternative weighting schemes when forecasting CPIATE, both in terms of point forecasts and density forecasts. Selecting, ex ante, the best model in quasi real-time, produces poorly specified densities when forecasting CPIATE. Overall, the ensemble density combination forecasts for GDP performs at least as good as alternative weighting schemes, both in terms of point forecasts and density forecasts.

References

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

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

Akram, Q. F., 2008. The econometric model of mainland norway EMod on seasonally adjusted data. Mimeo, Norges Bank.

Ang, A., Piazzesi, M., Wei, M., 2006. What does the yield curve tell us about GDP growth. Journal of Econometrics 131 (1-2), 359–403.

Bache, I. W., Jore, A. S., Mitchell, J., Vahey, S. P., 2009a. Combining VAR and DSGE forecast densities. Mimeo, Norges Bank.

Bache, I. W., Mitchell, J., Ravazzolo, F., Vahey, S. P., 2009b. Macro modeling with many models. Working Paper 15, Norges Bank.

Bates, J., Granger, C., 1969. The combination of forecasts. Operations Research Quarterly 20 (4), 451–468.

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

Bjørnland, H. C., Gerdrup, K., Jore, A. S., Smith, C., Thorsrud, L. A., 2009a. Does forecast combination improve Norges Bank inflation forecasts? Working Paper 1, Norges Bank.

Bjørnland, H. C., Gerdrup, K., Jore, A. S., Smith, C., Thorsrud, L. A., 2009b. There is more than one weight to skin a cat: Combining densities at Norges Bank.Paper submitted to North American Journal of Economics and Finance March 24 2009.

Brekke, H., Halvorsen, K. W., 2009. Regionalt nettverk - tidlig og godt bilde av norsk økonomi. Penger og Kreditt 2, Norges Bank, forthcoming in Economic Bulletin 1/2010.

Brubakk, L., Husebø, T. A., Maih, J., Olsen, K., Østnor, M., 2006. Finding NEMO: Documentation of the Norwegian economy model. Staff Memo 2006/6, Norges Bank.

Clark, T. E., McCracken, M. W., 2007. Forecasting with small macroeconomic vars in the presence of instabilities. Federal Reserve Board Finance and Economics Discussion Series (41), 5–31.

Clemen, R. T., 1987. Combining Overlapping information. Management Science 33 (3), 373–380.

Clemen, R. T., 1989. Combining forecasts: A review and annotated bibliography. International Journal of Forecasting 5, 559–583.

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

George, E. I., 1999. Discussion of "Model Averaging and Model Search" by Merlise A. Clyde. In: Bernardo, J., Berger, J., Dawid, A., Smith, A. (Eds.), Bayesian Statistics 6. Oxford University Press, Oxford.

Geweke, J., Amisano, G., 2008. Comparing and evaluating bayesian predictive distributions of asset returns. Working Paper Series 969, European Central Bank, forthcoming in International Journal of Forecasting.

Hall, S. G., Mitchell, J., 2005. Evaluating, comparing and combining density forecasts using the KLIC with an application to the Bank of England and NIESR 'Fan' charts of inflation. Oxford Bulletin of Economics and Statistics 67, 995–1033, supplement.

Hall, S. G., Mitchell, J., 2008. Recent developments in density forecasting. In: Mills, T., Patterson, K. (Eds.), Palgrave Handbook of Econometrics. Vol. 2. Palgrave Macmillan, Basingstoke, Hampshire, forthcoming.

Jore, A.-S., Mitchell, J., Vahey, S. P., 2009. Combining forecast densities from VARs with uncertain instabilities. NIESR Discussion Paper No. 303, fortcoming Journal of Applied Econometrics.

Kascha, C., Ravazzolo, F., 2008. Combining inflation density forecasts. Staff Working Paper 22, Norges Bank, forthcoming in Journal of Forecasting.

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

Makridakis, S., Chatfield, C., Hibon, M., Lawrence, M., Mills, T., Ord, K., Simmons, L. F., 1993. The m2-competition: A real-time judgmentally based forecasting study. International Journal of Forecasting 9 (1), 5–22.

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

Matheson, T., 2006. Factor model forecasts for New Zealand. International Journal of Central Banking 2 (2), 169–237.
URL http://www.ijcb.org/journal/ijcb06q2a6.htm

Noceti, P., Smith, J., Hodges, S., 2003. An evaluation of tests of distributional forecasts. Journal of Forecasting 22 (6-7), 447–455.

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

Stock, J. H., Watson, M. W., 2006. Forecasting with many predictors. In: Elliott, G., Granger, C. W. J., Timmermann, A. (Eds.), Handook of Economic Forecasting. Elsevier, Amsterdam, pp. 515–534.

Timmermann, A., 2006. Forecast combinations. In: Elliott, G., Granger, C. W. J., Timmermann, A. (Eds.), Handbook of Economic Forecasting. Vol. 1. Elsevier, Amsterdam, pp. 136–96.

Wallis, K. F., 2003. Chi-squared tests of interval and density forecasts, and the bank of england's fan charts. International Journal of Forecasting 19 (3), 165–175.

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

Winkler, R. L., 1981. Combining probability distributions from dependent information sources. Management Science 27 (4), p479 – 488.

A. Tables

Table 3: RMSFE and average logarithmic score for selection, SAM8, individual models, and grand ensemble. CPIATE

RMSFE

	hor=1	**hor=2**	**hor=3**	**hor=4**	**hor=5**
Selection	0.15	0.33	0.52	0.77	1.01
SAM8	0.11	0.32	0.52	0.67	0.83
Individual models	0.18	0.38	0.57	0.72	0.83
Grand ensemble	0.11	0.27	0.45	0.64	0.77

Average logarithmic score

	hor=1	**hor=2**	**hor=3**	**hor=4**	**hor=5**
Selection	-0.33	-0.21	-0.57	-1.06	-1.04
SAM8	1.24	-0.04	-0.47	-0.79	-0.96
Individual models	0.76	-0.34	-0.63	-0.79	-0.90
Grand ensemble	1.12	-0.02	-0.44	-0.63	-0.77

Table 4: RMSFE and average logarithmic score for selection, SAM8, individual models, and grand ensemble. GDP

RMSFE

	hor=1	**hor=2**	**hor=3**	**hor=4**	**hor=5**
Selection	0.87	1.41	1.60	1.74	1.81
SAM8	0.85	1.25	1.48	1.67	1.76
Individual models	1.03	1.35	1.62	1.82	1.95
Grand ensemble	0.85	1.13	1.40	1.68	1.83

Average logarithmic score

	hor=1	**hor=2**	**hor=3**	**hor=4**	**hor=5**
Selection	-1.32	-1.51	-1.75	-1.88	-1.91
SAM8	-1.31	-1.50	-1.64	-1.72	-1.75
Individual models	-1.48	-1.54	-1.61	-1.64	-1.65
Grand ensemble	-1.35	-1.49	-1.60	-1.66	-1.66

Table 5: Weights for a combination of selection, SAM8, individual models, and grand ensemble, at the end of the evaluation period (2009Q1). CPIATE

Inverse MSE-weights

	hor=1	**hor=2**	**hor=3**	**hor=4**	**hor=5**
Selection	0.18	0.23	0.24	0.21	0.17
SAM8	0.33	0.24	0.24	0.27	0.26
Individual models	0.12	0.17	0.20	0.23	0.26
Grand ensemble	0.36	0.35	0.32	0.29	0.30

Score-based weights

	hor=1	**hor=2**	**hor=3**	**hor=4**	**hor=5**
Selection	0.00	0.00	0.00	0.00	0.00
SAM8	1.00	0.25	0.19	0.00	0.00
Individual models	0.00	0.00	0.00	0.00	0.00
Grand ensemble	0.00	0.75	0.80	1.00	1.00

Table 6: Weights for a combination of selection, SAM8, individual models, and grand ensemble, at the end of the evaluation period (2009Q1) GDP

Inverse MSE-weights

	hor=1	**hor=2**	**hor=3**	**hor=4**	**hor=5**
Selection	0.26	0.20	0.23	0.25	0.26
SAM8	0.27	0.26	0.26	0.27	0.27
Individual models	0.19	0.22	0.22	0.22	0.22
Grand ensemble	0.27	0.32	0.29	0.26	0.25

Score-based weights

	hor=1	**hor=2**	**hor=3**	**hor=4**	**hor=5**
Selection	0.41	0.21	0.00	0.00	0.00
SAM8	0.51	0.30	0.11	0.03	0.01
Individual models	0.00	0.06	0.36	0.67	0.53
Grand ensemble	0.08	0.43	0.53	0.30	0.46

Table 7: Pits tests for evaluating density forecasts for CPIATE (p-values)

		Berkowitz	χ^2	Ljung-Box	Anderson-Darling
horizon=1	Selection	0.00	**0.17**	**0.06**	0.01
	SAM8	**0.33**	**0.77**	**0.31**	**0.79**
	Individual models	0.03	**0.06**	**0.08**	**0.20**
	Grand ensemble	0.02	**0.63**	**0.15**	**0.37**
horizon=2	Selection	0.00	**0.27**	**0.22**	**0.06**
	SAM8	**0.64**	**0.78**	**0.47**	**0.89**
	Individual models	0.03	**0.37**	**0.78**	**0.43**
	Grand ensemble	**0.05**	**0.54**	**0.78**	**0.41**
horizon=3	Selection	0.00	0.01	**0.75**	0.00
	SAM8	**0.86**	**0.88**	**0.96**	**0.90**
	Individual models	**0.12**	**0.70**	**0.99**	**0.69**
	Grand ensemble	**0.31**	**0.84**	**0.97**	**0.75**
horizon=4	Selection	0.00	0.01	**0.88**	0.00
	SAM8	**0.46**	**0.77**	**0.91**	**0.51**
	Individual models	**0.15**	**0.38**	**0.96**	**0.62**
	Grand ensemble	**0.66**	**0.38**	**0.91**	**0.90**
horizon=5	Selection	0.00	0.02	**0.88**	0.00
	SAM8	0.02	**0.49**	**0.86**	0.03
	Individual models	**0.15**	**0.28**	**0.85**	**0.58**
	Grand ensemble	**0.64**	**0.80**	**0.88**	**0.72**

Note: For 1-step ahead forecasts, the null hypothesis in the Berkowitz test is that the inverse normal cumulative distribution function transformed pits are identically and independently normally distributed, iid N(0,1). For longer horizons, the null hypothesis is that the transformed pits are identically, normally distributed, N(0,1). χ^2 *is the Pearson chi-squared test of uniformity of the pits histogram in eigth equiprobable classes. Ljung-Box is a test for independence of the pits for 1-step ahead forecasts. For longer horizons, the Ljung-Box test is modified such that it test for independence at lags greater than or equal to the horizon. The Anderson-Darling test is a test for uniformity of the pits, with the small-sample (simulated) p-values computed assuming independence of the pits.*

Table 8: Pits tests for evaluating density forecasts for GDP (p-values)

		Berkowitz	χ^2	Ljung-Box	Anderson-Darling
horizon=1	Selection	**0.08**	**0.30**	**0.87**	**0.21**
	SAM8	**0.05**	**0.17**	**0.91**	**0.36**
	Individual models	0.03	**0.49**	**0.14**	**0.54**
	Grand ensemble	0.01	**0.05**	**0.36**	**0.20**
horizon=2	Selection	**0.58**	**0.90**	**0.89**	**0.30**
	SAM8	**0.47**	**0.17**	**0.94**	**0.74**
	Individual models	**0.47**	**0.15**	**0.76**	**0.72**
	Grand ensemble	0.04	**0.30**	**0.87**	**0.29**
horizon=3	Selection	**0.07**	**0.11**	**0.98**	**0.13**
	SAM8	**0.76**	**0.21**	**0.83**	**0.66**
	Individual models	**0.91**	**0.50**	**0.84**	**0.93**
	Grand ensemble	**0.21**	**0.46**	**0.85**	**0.47**
horizon=4	Selection	**0.06**	**0.11**	**0.27**	**0.33**
	SAM8	**0.90**	**0.82**	**0.89**	**0.97**
	Individual models	**0.92**	**0.38**	**0.99**	**0.82**
	Grand ensemble	**0.55**	**0.72**	**0.99**	**0.65**
horizon=5	Selection	**0.11**	**0.19**	**0.94**	**0.27**
	SAM8	**0.72**	**0.64**	**0.95**	**0.54**
	Individual models	**0.75**	**0.40**	**1.00**	**0.40**
	Grand ensemble	**0.80**	**0.22**	**1.00**	**0.56**

Note: For 1-step ahead forecasts, the null hypothesis in the Berkowitz test is that the inverse normal cumulative distribution function transformed pits are identically and independently normally distributed, iid N(0,1). For longer horizons, the null hypothesis is that the transformed pits are identically, normally distributed, N(0,1). χ^2 *is the Pearson chi-squared test of uniformity of the pits histogram in eigth equiprobable classes. Ljung-Box is a test for independence of the pits for 1-step ahead forecasts. For longer horizons, the Ljung-Box test is modified such that it test for independence at lags greater than or equal to the horizon. The Anderson-Darling test is a test for uniformity of the pits, with the small-sample (simulated) p-values computed assuming independence of the pits.*

Table 9: Sensitivity analysis: Performance of alternative weighting schemes relative to grand ensemble. CPI-ATE

Relative RMSFE (numbers lower than one mean better performance than the benchmark)

	hor=1	**hor=2**	**hor=3**	**hor=4**	**hor=5**
Score Linear Mse	1.00	1.00	1.00	1.00	1.00
Score Linear Score	1.23	1.05	1.09	1.17	1.19
Score Linear Equal	1.54	1.03	0.99	0.96	0.96
Score Log Mse	0.60	0.94	0.97	1.02	1.06
Score Log Score	1.04	1.02	1.05	1.16	1.19
Score Log Equal	0.60	0.94	0.97	1.01	1.05
Mse Linear Mse	0.97	1.08	1.02	0.99	0.99
Mse Linear Score	0.96	1.10	1.15	1.15	1.13
Mse Linear Equal	1.35	1.11	1.02	0.94	0.92
Mse Log Mse	0.60	0.98	0.98	1.00	1.04
Mse Log Score	0.68	1.04	1.10	1.12	1.13
Mse Log Equal	0.58	0.98	0.96	0.96	0.99
Equal Linear Mse	1.01	1.08	1.01	0.99	0.98
Equal Linear Score	1.38	1.09	1.14	1.15	1.15
Equal Linear Equal	1.50	1.12	1.00	0.93	0.92
Equal Log Mse	0.60	0.99	0.99	1.01	1.05
Equal Log Score	0.95	1.05	1.10	1.13	1.14
Equal Log Equal	0.60	0.99	0.97	0.97	1.00

Average logarithmic score of combination minus average logarithmic score in benchmark (positive numbers mean better performance than the benchmark)

	hor=1	**hor=2**	**hor=3**	**hor=4**	**hor=5**
Score Linear Mse	0.00	0.00	0.00	0.00	0.00
Score Linear Score	-0.41	0.09	0.15	0.13	0.21
Score Linear Equal	-0.58	-0.09	-0.04	-0.01	0.02
Score Log Mse	-0.62	0.20	0.22	0.18	0.15
Score Log Score	-1.21	0.09	0.26	0.18	0.28
Score Log Equal	-0.40	0.15	0.21	0.21	0.23
Mse Linear Mse	0.04	-0.11	-0.09	-0.15	-0.17
Mse Linear Score	0.02	-0.04	-0.08	-0.15	-0.12
Mse Linear Equal	-0.34	-0.19	-0.16	-0.19	-0.17
Mse Log Mse	-0.64	0.11	0.12	0.05	-0.00
Mse Log Score	-0.76	-0.04	-0.05	-0.09	-0.04
Mse Log Equal	-0.29	0.04	0.09	0.06	0.04
Equal Linear Mse	0.01	-0.11	-0.10	-0.15	-0.17
Equal Linear Score	-0.29	-0.05	-0.08	-0.14	-0.12
Equal Linear Equal	-0.50	-0.20	-0.16	-0.19	-0.18
Equal Log Mse	-0.62	0.11	0.12	0.06	0.00
Equal Log Score	-0.96	-0.05	-0.04	-0.07	-0.04
Equal Log Equal	-0.35	0.04	0.09	0.06	0.04

Note: The different combination schemes differ first by the weighting scheme within each ensemble, either inverse MSE-weights, score-based weights, or equal weights (first entry in column 1). Linear pooling is used for the three different within-ensemble weigthing schemes. Then for each within-ensemble weighting scheme, we present six alternative between-ensemble weighting schemes: two pooling methods (linear and logarithmic, see second entry in column 1) and three different weights (inverse MSE-weights, score-based weights, and equal weights, see third entry in column 1).

Table 10: Sensitivity analysis: Performance of alternative weighting schemes relative to grand ensemble. GDP Mainland-Norway

Relative RMSFE (numbers lower than one mean better performance than the benchmark)

	hor=1	**hor=2**	**hor=3**	**hor=4**	**hor=5**
Score Linear Mse	1.00	1.00	1.00	1.00	1.00
Score Linear Score	0.87	1.03	1.04	1.10	1.11
Score Linear Equal	1.03	1.02	1.01	1.01	1.00
Score Log Mse	0.97	1.00	0.99	0.99	1.02
Score Log Score	0.88	1.03	1.04	1.10	1.11
Score Log Equal	1.00	1.03	1.02	1.01	1.02
Mse Linear Mse	1.03	1.03	1.02	1.01	1.01
Mse Linear Score	0.88	1.02	1.06	1.11	1.11
Mse Linear Equal	1.06	1.05	1.03	1.01	1.01
Mse Log Mse	0.99	1.02	1.01	1.00	1.02
Mse Log Score	0.88	1.01	1.05	1.11	1.12
Mse Log Equal	1.01	1.05	1.03	1.01	1.02
Equal Linear Mse	1.03	1.03	1.02	1.01	1.01
Equal Linear Score	0.87	1.01	1.07	1.11	1.11
Equal Linear Equal	1.06	1.05	1.03	1.01	1.01
Equal Log Mse	0.99	1.02	1.00	1.00	1.02
Equal Log Score	0.88	1.01	1.06	1.11	1.12
Equal Log Equal	1.01	1.05	1.03	1.01	1.02

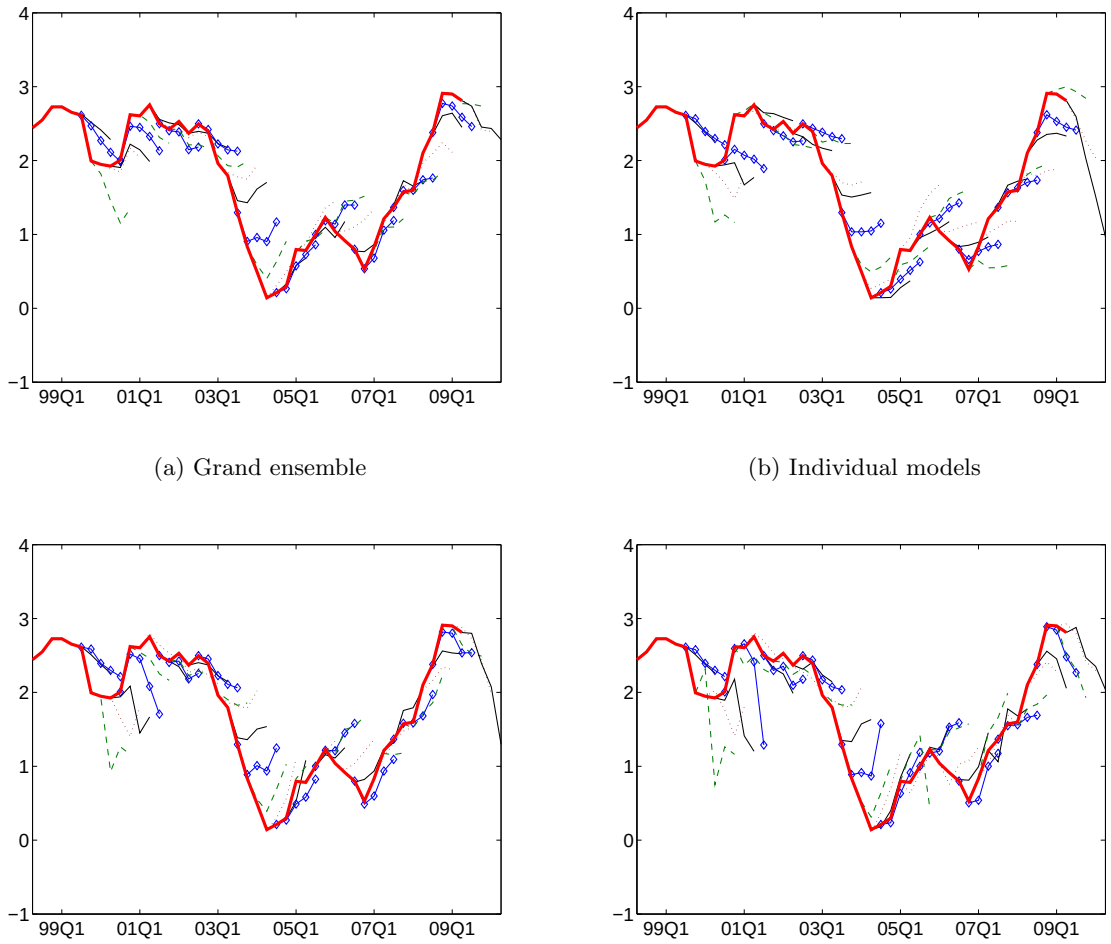
Average logarithmic score of combination minus average logarithmic score in benchmark (positive numbers mean better performance than the benchmark)

	hor=1	**hor=2**	**hor=3**	**hor=4**	**hor=5**
Score Linear Mse	0.00	0.00	0.00	0.00	0.00
Score Linear Score	0.14	0.07	0.10	0.10	0.11
Score Linear Equal	-0.03	-0.02	-0.02	-0.02	-0.01
Score Log Mse	0.09	0.08	0.06	0.00	0.05
Score Log Score	0.18	0.09	0.12	0.12	0.13
Score Log Equal	0.06	0.06	0.05	0.01	0.05
Mse Linear Mse	-0.02	-0.03	-0.03	-0.03	-0.03
Mse Linear Score	0.14	0.06	0.09	0.08	0.11
Mse Linear Equal	-0.05	-0.05	-0.04	-0.05	-0.04
Mse Log Mse	0.07	0.05	0.03	-0.03	0.02
Mse Log Score	0.18	0.06	0.10	0.10	0.13
Mse Log Equal	0.04	0.04	0.03	-0.01	0.02
Equal Linear Mse	-0.02	-0.03	-0.02	-0.03	-0.04
Equal Linear Score	0.14	0.06	0.09	0.09	0.11
Equal Linear Equal	-0.05	-0.05	-0.04	-0.05	-0.05
Equal Log Mse	0.07	0.05	0.03	-0.03	0.02
Equal Log Score	0.18	0.06	0.10	0.10	0.13
Equal Log Equal	0.04	0.04	0.03	-0.02	0.02

Note: The different combination schemes differ first by the weighting scheme within each ensemble, either inverse MSE-weights, score-based weights, or equal weights (first entry in column 1). Linear pooling is used for the three different within-ensemble weigthing schemes. Then for each within-ensemble weighting scheme, we present six alternative between-ensemble weighting schemes: two pooling methods (linear and logarithmic, see second entry in column 1) and three different weights (inverse MSE-weights, score-based weights, and equal weights, see third entry in column 1).

B. Figures

Figure 1: Recursive quasi-out-of-sample forecasts. 4-quarter growth. Per cent. CPIATE



(a) Grand ensemble

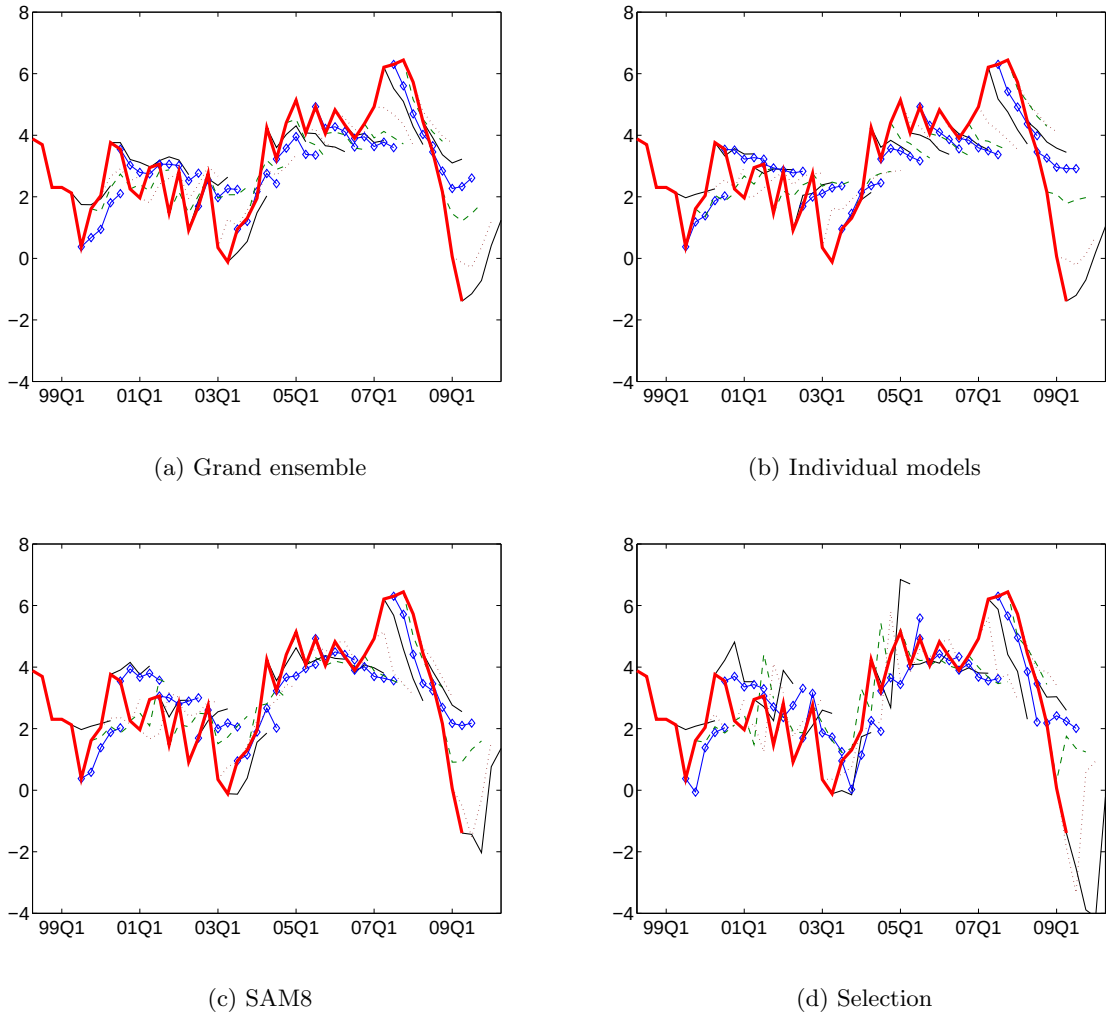
(b) Individual models

(c) SAM8

(d) Selection

Note: The figures show recursive out-of-sample forecasts for the next 1-5 quarters, for the grand ensemble, individual models, SAM8, and selection. Estimation and forecasts are done in quasi real-time, using the latest vintage of data.

Figure 2: Recursive quasi-out-of-sample forecasts. 4-quarter growth. Per cent. GDP



(a) Grand ensemble

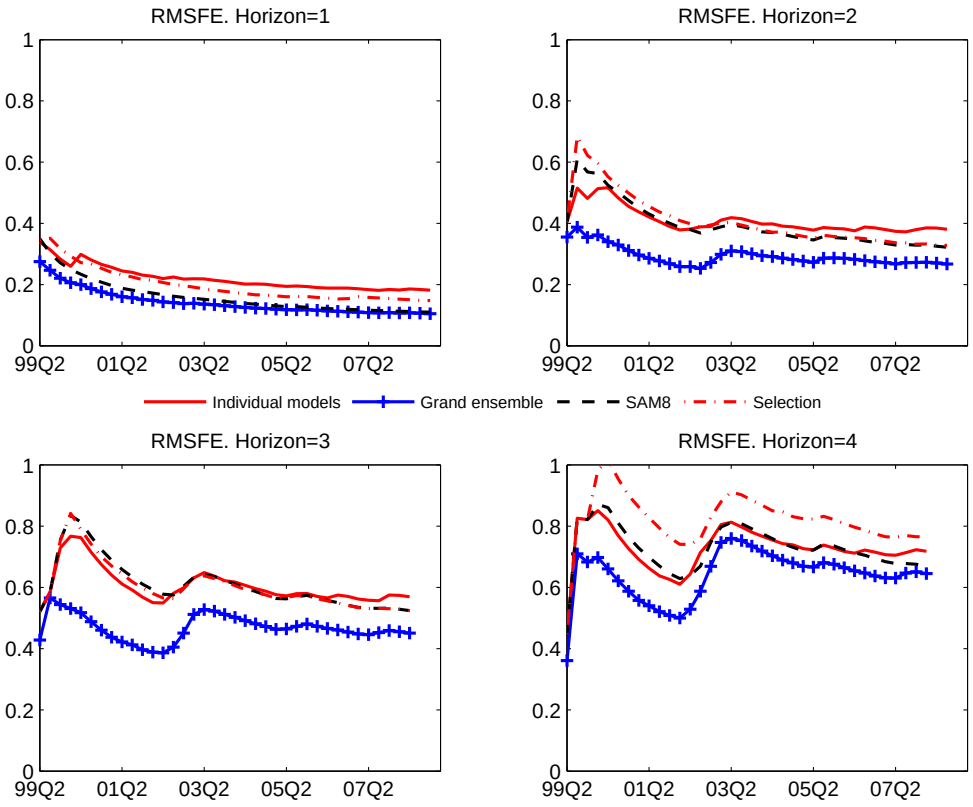
(b) Individual models

(c) SAM8

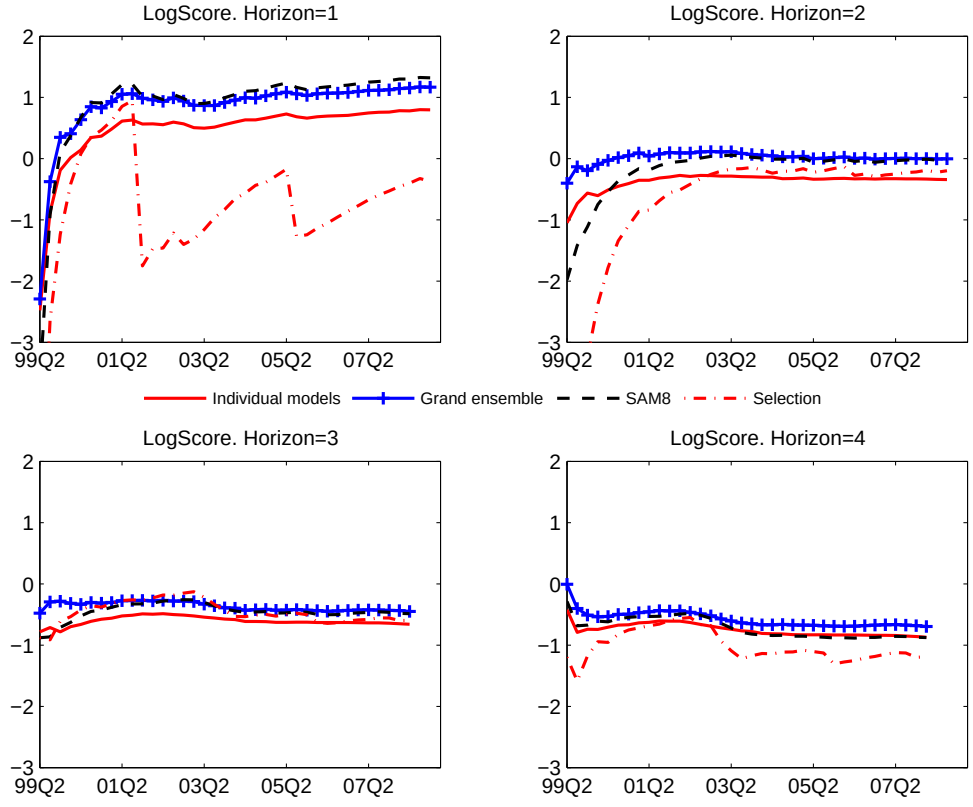
(d) Selection

Note: The figures show recursive out-of-sample forecasts for the next 1-5 quarters, for the grand ensemble, individual models, SAM8, and selection. Estimation and forecasts are done in quasi real-time, using the latest vintage of data.

Figure 3: Performance of grand ensemble, individual model combination, SAM8, and selection. Expanding window. CPIATE

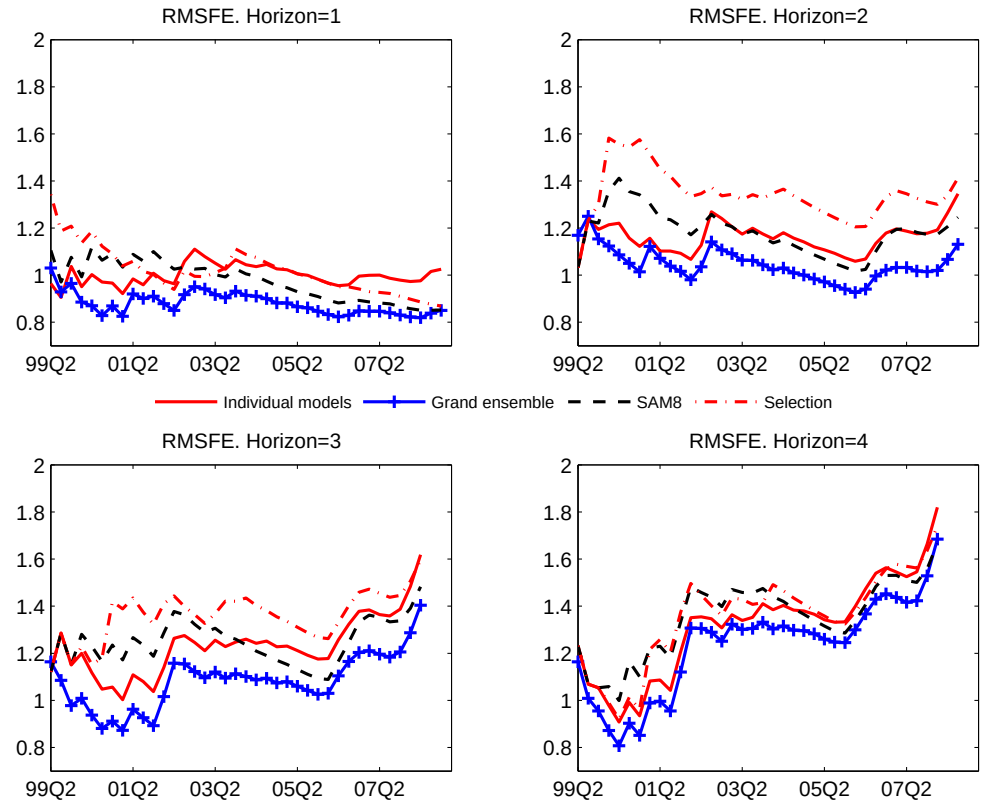


(a) **RMSFE**

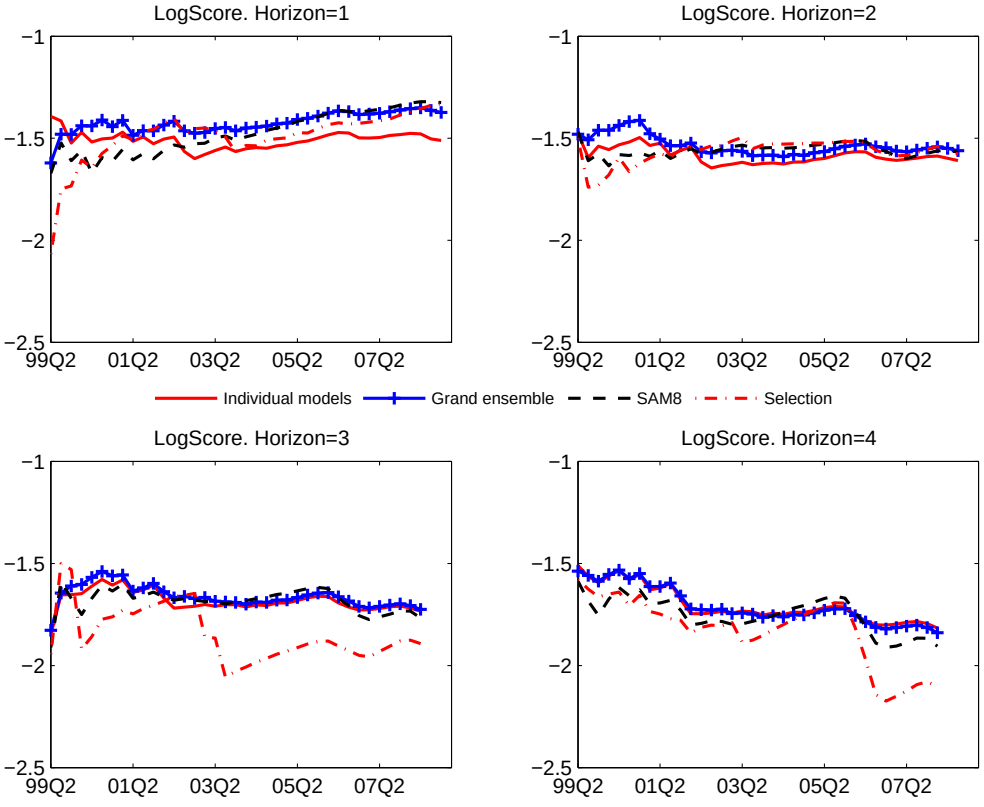


(b) **Average logarithmic score**

Figure 4: Performance of grand ensemble, individual model combination, SAM8, and selection. Expanding window. GDP

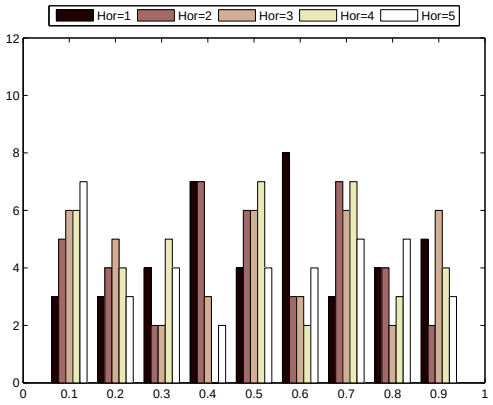


(a) **RMSFE**

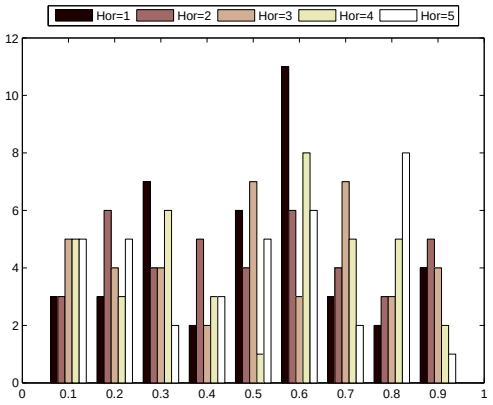


(b) **Average logarithmic score**

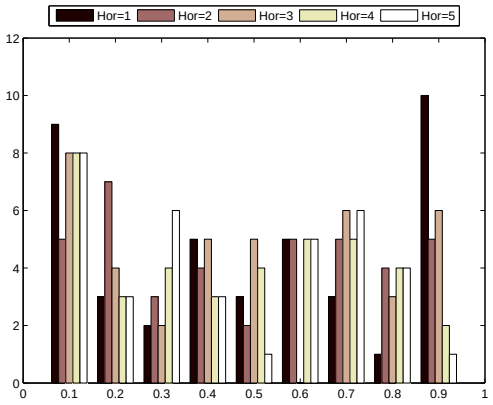
Figure 5: Probability integral transforms. Horizon=1-5. CPIATE



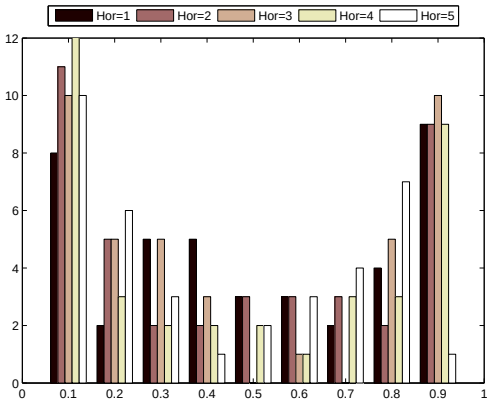
(a) Grand ensemble



(b) Individual models



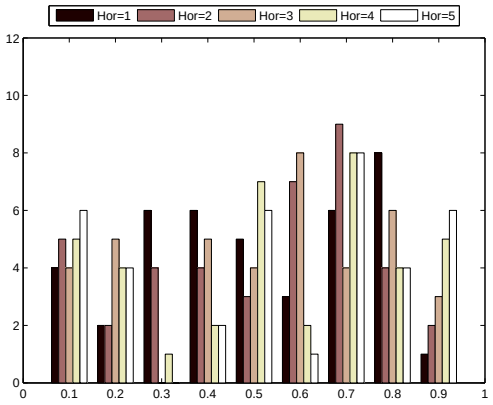
(c) SAM8



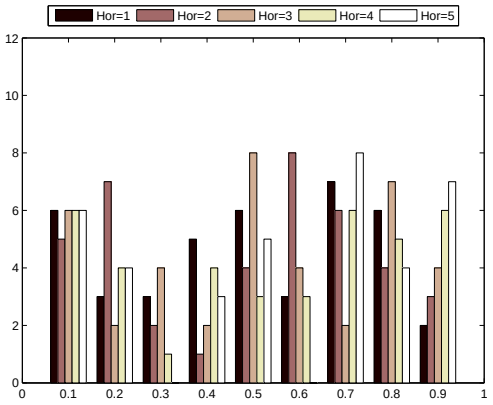
(d) Selection

Note: The pits are the ex ante inverse predictive cumulative distribution evaluated at the ex post actual observations. The pits of a forecasting model should have a standard uniform distribution if the model is correctly specified.

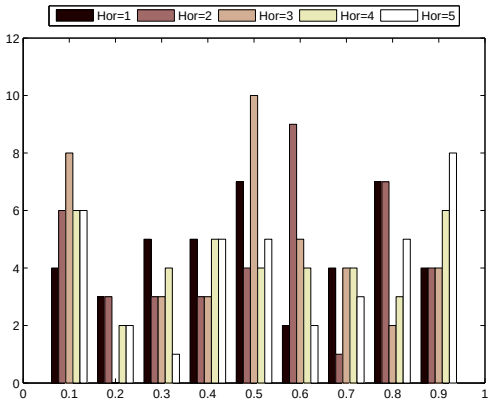
Figure 6: Probability integral transforms. Horizon=1-5. YFN



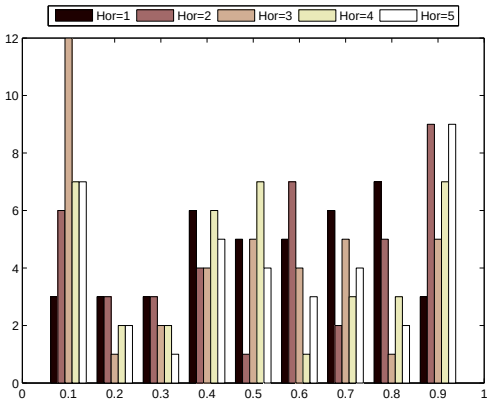
(a) Grand ensemble



(b) Individual models



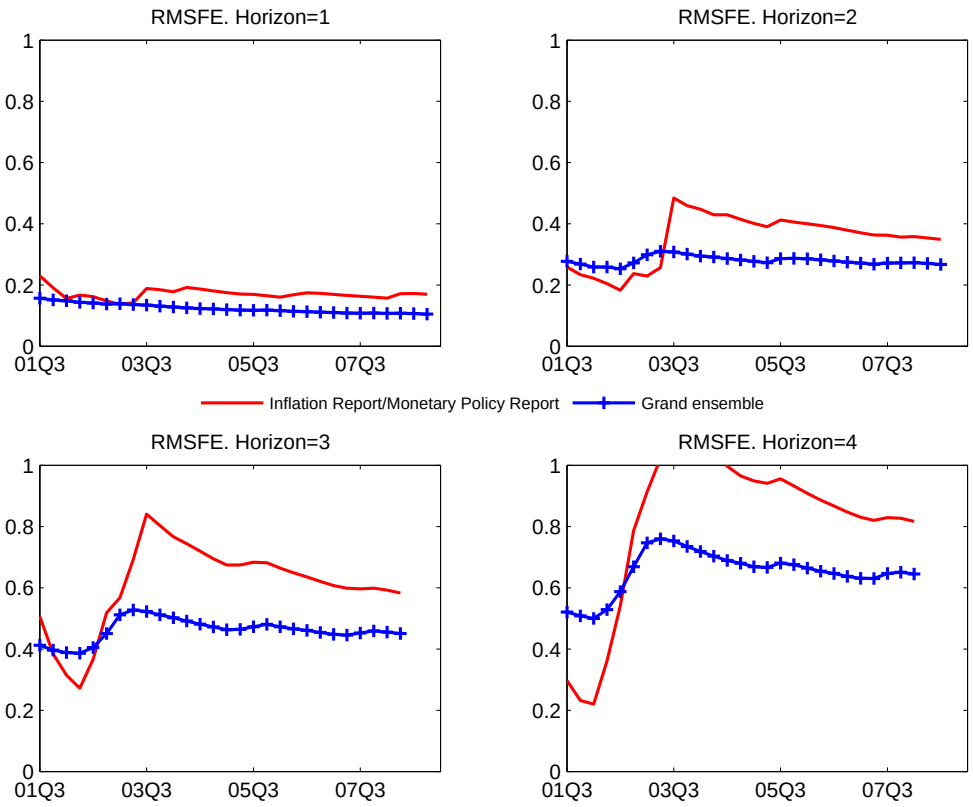
(c) SAM8



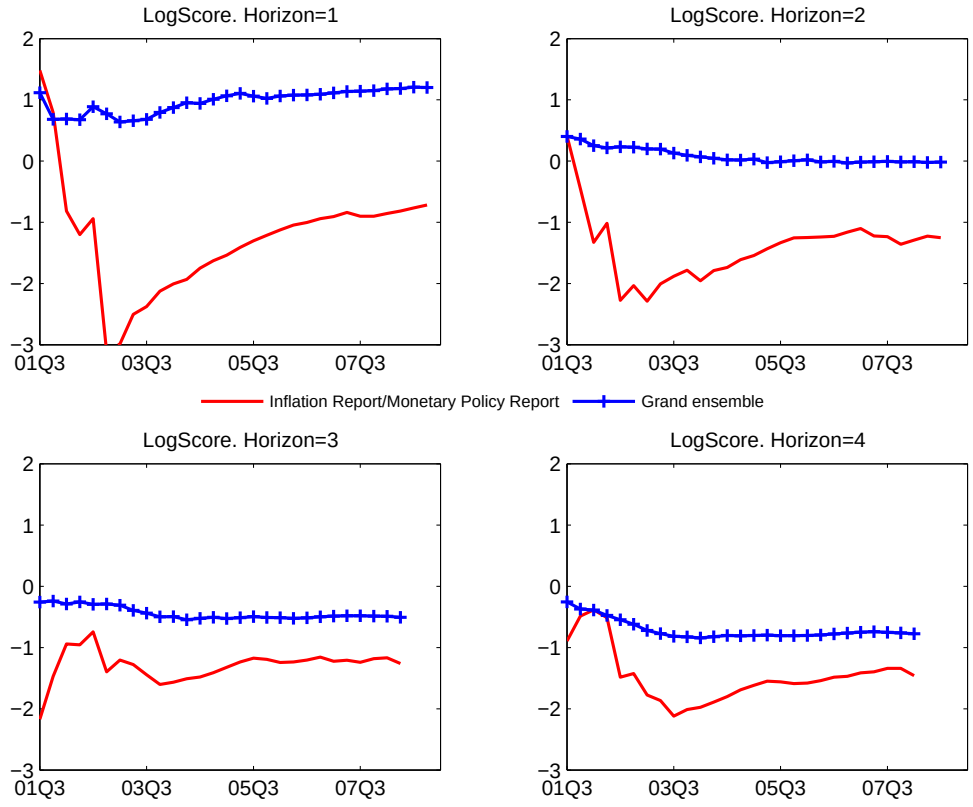
(d) Selection

Note: The pits are the ex ante inverse predictive cumulative distribution evaluated at the ex post actual observations. The pits of a forecasting model should have a standard uniform distribution if the model is correctly specified.

Figure 7: Performance of grand ensemble and Norges Bank. Expanding window. CPIATE

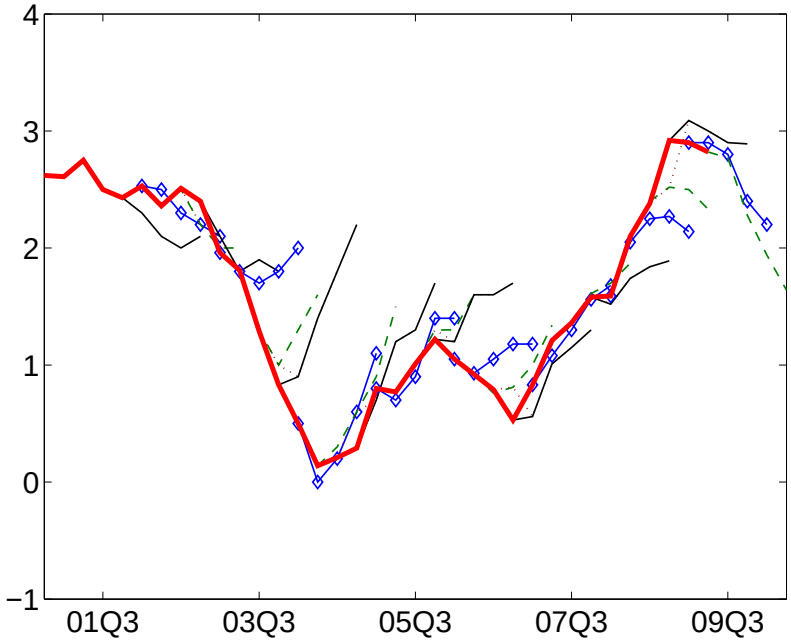


(a) **RMSFE**



(b) **Average logarithmic score**

Figure 8: Recursive quasi-out-of-sample forecasts for MPR. 4-quarter growth. Per cent. CPIATE



Note: The figure shows recursive out-of-sample forecasts the next 1-5 quarters from various Inflation Reports/Monetary Policy Reports.

C. Model suite

C.0.1. 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,

$$y_t = \alpha + \sum_{j=1}^{p} \phi_j y_{t-j} + \sum_{j=0}^{q} \theta_j \varepsilon_{t-j} \tag{5}$$

where y_t is the variable of interest 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, and we have not found that including MA-terms improves the forecasting properties. We have therefore chosen to include only AR-models:

- 36 quarterly AR-models for GDP and inflation, respectively, estimated from 1981, which differ first by the number of lags (one to four), then by the number of estimation periods (whole period, a short rolling window of 20 quarters, and a longer rolling window of 40 quarters), and third by the transformation of data (differenced, double-differenced, and trend-adjusted). The models are included in the ensemble eUniv.
- One quarterly AR-model for GDP (with two lags) and for inflation (one lag), respectively, estimated from 1990. The models are included in the ensemble eUniv.
- One quarterly AR-model for GDP estimated from 1981 where the number of lags (up to four) is determined by a Bayesian information criteria (BIC). The model is included in the ensemble eUniv.
- One AR-model for CPIATE with four lags estimated on monthly data from 1990. Forecasts are done on a monthly basis and converted to quarterly frequencies. The model is included in the ensemble eUniv.
- One model for CPIATE which first forecast 12 main components of CPI and two components covering energy prices using AR-models. The models for the components are estimated on monthly, unadjusted data from 1991. Forecasts from each component are weighted using the consumer weights in CPI to form composite forecasts for CPIATE. The forecasts are converted to quarterly frequencies. The model is included in the ensemble eDisAgg.

C.0.2. Random Walk in Mean

Forecast monthly inflation measured by CPIATE. Monthly forecasts are aggregated to quarterly frequencies. The forecasts are the mean of a rolling window of monthly data. The mean is updated iteratively over the forecasting horizons. The model is estimated on data from 1990.

C.0.3. Vector AutoRegressive (VAR) models

The VAR models are based on statistical relationships between GDP (and/or inflation) and other explanatory variables. All the variables are a function of lagged values of itself and the other variables,

$$X_t = A + \sum_{j=1}^{p} B_j X_{t-j} + \nu_t \tag{6}$$

where X_t is the vector of variables in the model. We use an iterative forecasting method. SAM includes the following VARs:

- Quarterly VARs with different combinations of GDP, inflation and interest rate as explanatory variables, see table 11 for more details.
- Bivariate VARs with CPIATE and different indicators as explanatory variables, see table 12 for more details.
- Quarterly, bivariate VARs with GDP and different indicators of economic activity as explanatory variables, see table 13 for more details.
- Structural VARs with GDP and different monetary aggregates as explanatory variables, se table 14 for more details.

Table 11: VARs with different combinations of GDP, inflation and interest rate as explanatory variables in the model suite in SAM

Description	**Incl. in ensemble**
36 bivariate VARs with GDP and inflation as explanatory variables, s.a., growth	eVAR2
36 bivariate VARs with GDP and interest rate as explanatory variables, s.a., growth	eVAR3
36 bivariate VARs with inflation and interest rate as explanatory variables, s.a., growth	eVAR3
36 tri-variate VARs with GDP, inflation and interest rate as explanatory variables, s.a., growth	eVAR3

Note: The sources of the data is Statistics Norway and Norges Bank. GDP is for Mainland-Norway, inflation is measured by CPIATE and interest rate by 3-month money market rate. The models differ first by the number of lags (one to four), then by the number of estimation periods (whole period, a short rolling window of 30 quarters, and a longer rolling window of 40 quarters), and third by the transformation of data (differenced, double-differenced, and trend-adjusted). The models are estimated on data from 1981 for the longest samples.

Table 12: Bivariate VARs with inflation and different indicators as explanatory variables in the model suite in SAM

Model	**Description of indicator**	**Incl. in ensemble**
mVARurr	Registered unemployment, s.a., monthly frequency, growth. Model estimated from 1992	eMth
mVARi44	Import weighted krone exchange rate (I44) (unadjusted), monthly frequency, growth. Model estimated from 1992.	eMth
VARm	bivariate VAR with inflation and monetary aggregate (s.a.) as explanatory variables, growth.	eMny

Note: The sources of the data is Statistics Norway and Norges Bank. The frequency of the data is monthly. Inflation is measured by CPIATE (s.a.). Seasonally adjusted data is denoted s.a.

Table 13. Bivariate VARs with GDP and different indicators as explanatory variables in the model suite in SAM

Model	**Description of indicator**	**Incl. in ensemble**
X-0	Manufacturing, change in total production since previous quarter, diffusion index, adjusted, level.	eBTS
N-0	Manufacturing, average employment, change since previous quarter, diffusion index, s.a., level.	eBTS
OrdreD-0	Manufacturing, orders from domestic market, change since previous quarter, diffusion index, s.a., level.	eBTS
OrderF-0	Manufacturing, orders from export market, change since previous quarter, diffusion index, s.a., level.	eBTS
OrderTot-0	Manufacturing, total orders, change since previous quarter, diffusion index, s.a., level.	eBTS
PrisD-0	Manufacturing, prices at domestic markets, change since previous quarter, diffusion index, s.a., level.	eBTS
PrisF-0	Manufacturing, prices at export markets, change since previous quarter, diffusion index, s.a., level.	eBTS
X-1	Manufacturing, expected change in total production next quarter, diffusion index, s.a., level.	eBTS
Xinput-1	Intermediate goods, expected change in total production next quarter,diffusion, index, s.a., level.	eBTS
Xinvest-1	Capital goods, expected change in total production next quarter, diffusion index, s.a., level.	eBTS
Xconsum-1	Consumer goods, expected change in total production next quarter, diffusion index, s.a., level.	eBTS
N-1	Manufacturing, average employment, expected change next quarter, diffusion index, s.a., level.	eBTS
Ninput-1	Intermediate goods, average employment, expected change next quarter, diffusion index, s.a., level.	eBTS
Ninvest-1	Capital goods, average employment, expected change next quarter, diffusion index, s.a., level.	eBTS
Nconsum-1	Consumer goods, average employment, expected change next quarter, diffusion index, s.a., level.	eBTS
OrdreD-1	Manufacturing, orders from domestic market, expected change next quarter, diffusion index, s.a., level.	eBTS
OrdreF-1	Manufacturing, orders from export market, expected change next quarter, diffusion index, s.a., level.	eBTS
OrdreTot-1	Manufacturing, total orders, expected change next quarter, diffusion index, s.a., level.	eBTS
PrisD-1	Manufacturing, prices at domestic markets, expected change next quarter, diffusion index, s.a., level.	eBTS
PrisF-1	Manufacturing, prices at export markets, expected change next quarter, diffusion index, s.a., level.	eBTS
Generell-1	Manufacturing, general judgement of the outlooks for the next quarter, diffusion index, s.a., level.	
Kaputn	Manufacturing, utilization of capacity at the end of the quarter with current level of production, s.a., weighted average, percent.	eBTS
Kaputn-0	Manufacturing, average utilization of capacity, change since previous quarter, diffusion index, s.a., level.	eBTS
Kaputn-1	Manufacturing, average utilization of capacity, expected change next quarter, diffusion index, s.a., level.	eBTS
Ressurs	Manufacturing, indicator for bottlenecks in production, s.a., percent.	eBTS
Konj-EU	Manufacturing, industrial confidence indicator, leading indicator for production (EU definitions), based on diffusion indices, indices, s.a.	eBTS
Lag-com	Manufacturing, assessment of inventories of raw materials by end of quarter, diffusion index, s.a., level.	eBTS
Lag-pr	Manufacturing, assessment of inventories of own products by end of quarter,	

	diffusion index, s.a., level.	eBTS
L-c-pr	Manufacturing, assessment of inventories of raw materials etc. compared to current production, diffusion index, s.a., level.	eBTS
L-v-o	Manufacturing, assessment of inventories of own products compared to value of sales, diffusion index, s.a., level.	eBTS
Invest	Manufacturing, assessment of whether the enterprise consider changes in the plans for gross capital investments, diffusion index, s.a., level.	eBTS
Ord-prod	Manufacturing, assessment of stocks of orders compared to current level of production at end of quarter, diffusion index, s.a., level.	eBTS
SKI-s	Manufacturing, industrial confidence indicator, leading indicator for production (EU definitions), based on net numbers, s.a., percent, level.	eBTS
Indikator	Survey of consumer confidence, overall assessment of the economic situation, index, s.a., level. Source: TNS Gallup	eTNSG
Landet_0_s	Survey of consumer confidence, assessment of economic developments in Norway last year, index, s.a., level. Source: TNS Gallup	eTNSG
Store_s	Survey of consumer confidence, assessment of the current situation and whether to buy durables, index, s.a., level. Source: TNS Gallup	eTNSG
Landet_1	Survey of consumer confidence, expected development in the Norwegian economy next year, index, unadjusted, level. Source: TNS Gallup	eTNSG
Egen_0	Survey of consumer confidence, assessment of development in own economic situation last year, index, unadjusted, level. Source: TNS Gallup	eTNSG
Egen_1	Survey of consumer confidence, expected development in own economic situation next year, index, unadjusted, level. Source: TNS Gallup	eTNSG
ORtot	Stock of orders, building and construction, value index, s.a., growth.	eBuild
ORbol	Stock of orders, residential buildings, value index, s.a., growth.	eBuild
ORabygg	Stock of orders, non-residential buildings, value index, s.a., growth.	eBuild
Lager	Total inventories in manufacturing, volume index, s.a., growth.	eOrd
	New orders, manufacturing, domestic, index, s.a., growth.	eOrd
ORi	Stock of orders, manufacturing, domestic, index, s.a., growth.	eOrd
AIi	Estimated fixed investment in electric supply, in millions of NOK, s.a., growth.	eOrd
ORanlegg	Stock of orders, civil engineering works, value index, unadjusted, growth.	eBuild
OTanlegg	New orders, civil engineering works, value index, s.a., growth.	eBuild
OTba	New orders, building and construction, value index, index, s.a., growth.	eBuild
OTbygg	New orders, total building, value index, index, s.a., growth.	eBuild
OTbol	New orders, residential buildings, value index, index, s.a., growth.	eBuild
OTabygg	New orders, non-residential buildings, value index, index, s.a., growth.	eBuild
ORbygg	Stock of orders, total building, value index, index, s.a., growth.	eBuild
Behning	Number of vacant positions, s.a., growth. Source: NAV	eEmpl
tilgang	Number of new vacant positions, s.a., growth. Source: NAV	eEmpl
ukeverk	Number of weeks worked, three-month moving average converted to quarterly numbers, s.a., growth, growth. Source: AKU	eEmpl
syss	Total employment, unadjusted, growth. Source: AKU	eEmpl
sba	Number of employed aged 15-74 in building and construction, unadjusted, growth. Source: AKU	eEmpl
stransp	Number of employed aged 15-74 in tranportation, unadjusted, growth. Source: AKU	eEmpl
svare	Number of employed aged 15-74 in retail trade etc., unadjusted, growth. Source: AKU	eEmpl
sfintj	Number of employed aged 15-74 in financial industry, unadjusted, growth. Source: AKU	eEmpl
stjen	Number of employed aged 15-74 in other services, unadjusted, growth. Source: AKU	eEmpl
sindu	Number of employed aged 15-74 in manufacturing, unadjusted, growth.Source: AKU	eEmpl
K2real	Domestic credit (C2) to general public deflated by CPIATE, s.a., growth.	eMny
K2hus	Domestic credit to households deflated by CPIATE, s.a., growth.	eMny
M2nonfin	Domestic credit to non-financial enterprises deflated by CPIATE, s.a., growth.	eMny
K2Non-Housing	Domestic credit to general public minus domestic credit to households, deflated by CPIATE, s.a., growth.	eMny
RegnX-1	Expected production next 6 months from Norges Bank's regional network, s.a., growth. Due to data limitations, the model only produce forecasts from 2005 and onwards. Prior to 2005, we use forecasts from a VAR where	

	we replace RegnX-1 with the variable X-1.	eRegN

Note: The source of the data is Statistics Norway unless otherwise stated. NAV is The Norwegian Labour and Welfare Administration. AKU is Statistics Norway's Labour Force Survey. The Business Tendency Survey is carried out by Statistics Norway. The frequency of the data is quarterly. GDP is for Mainland-Norway. Seasonally adjusted or trend adjusted data is denoted s.a.

Table 14: Structural VARs with GDP and different monetary aggregates as explanatory variables in the model suite in SAM

Model	**Description of indicator**	**Incl. in ensemble**
MnyM1	Bivariate VAR with narrow monetary aggregate (M1) and GDP as explanatory variables.	eMny
MnyM2	Bivariate VAR with monetary aggregate (M2) and GDP as explanatory variables.	eMny
MnyM12	Tri-variate, structural VAR with M1, M2 and GDP as explanatory variables.	eMny

Note: The sources of the data is Statistics Norway and Norges Bank. GDP is for Mainland-Norway. The models are estimated on quarterly data from 1995. All the data have been seasonally adjusted data. Due to data limitations, the models only produce forecasts from 2001 and onwards. Prior to 2001, we use forecasts from an AR(2)-model in the evaluation.

C.0.4. Bayesian VAR (BVAR) models

Bayesian methods have proven useful in the estimation of VARs. In Bayesian analysis the econometrician has to specify prior beliefs about the parameters. The prior beliefs are then combined with the data in the VAR to form a posterior view of the parameters. We use a direct forecasting method (eg CPIATE at time $t+h$ is now regressed against *factors* from data at time t). Table 14 gives an overview over the BVARs.

Table 15: Bayesian VARs in the model suite in SAM

Model	**Description**	**Incl. in ensemble**
BVAR1	Bivariate VAR with GDP and inflation as explanatory variables.	eBVAR
BVAR2	BVAR with GDP, inflation, interest rate and exchange rate as explanatory variables.	eBVAR
BVAR3	BVAR1 plus exogenous variables (OILGAS, M2, GDPw, INFLw, exchange rate, interest rate, Iw).	eBVAR
BVAR4	BVAR2 plus exogenous variables (M2, OILGAS, GDPw, INFLw, Iw).	eBVAR
BVAR5	BVAR1 with other priors	eBVAR
BVAR6	BVAR2 with other priors	eBVAR
BVAR7	BVAR3 with other priors	eBVAR
BVAR8	BVAR with GDP, inflation, interest rate and terms of trade as explanatory variables.	eBVAR
BVAR9	BVAR8 plus exogenous variables (M2, OILGAS, GDPw, INFLw, Iw)	eBVAR
BVAR10	BVAR8 with other priors	eBVAR

Note: The sources of the data is Statistics Norway and Norges Bank. GDP is for Mainland-Norway, inflation is measured by CPIATE and interest rate by 3-month money market rate. Quarterly frequencies. M2 is money supply, OILGAS is the average price of oil and gas exports, exchange rate is the trade-weighted krone exchange rate (KKI), GDPw is a weighted index of GDP among 25 trading partners, INFLw is is a weighted index of CPI among 25 trading partners, and Iw is a weighted index of 3-month money market rate (US, UK, Sweden and Euro).

C.0.5. Monthly indicator models

The monthly indicator models forecast GDP based on many monthly indicators (that are averaged up to a quarterly frequency). The models are estimated using OLS (ordinary least squares). In order to forecast GDP Mainland-Norway, the explanatory variables in the indicator models are projected using AR-models. The explanatory variables in the two indicator models are: manufacturing production, employment, retail sales, hotel statistics and building starts. The models are included in the ensemble eMI.

C.0.6. Monthly (FM) and Quarterly factor (FQ) 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. Our factor models for predicting inflation builds on Matheson (2006) and use either a monthly or a quarterly data-set. In one factor model for inflation, we also forecast monthly inflation and convert the forecast to quarterly frequencies. The factor models are estimated using principal components on the data data, including survey data.[8] We use a direct forecasting method (eg CPIATE at time $t+h$ is now regressed against *factors* from data at time t). The models are included in the ensemble eFM.

C.0.7. 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, 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.[9] The model is included in the ensemble eEmod.

C.0.8. 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 2008Q1). The steady-state levels are equal to recursively updated means of the variables. The model is included in the ensemble eDSGE.

C.0.9. Term Structure Models

We construct four models for GDP Mainland-Norway using term structure information, two unrestricted VARs and two affine models, see table 16. The affine model is a structural VAR where parameters are restricted to ensure no-arbitrage between yields of different maturities, and builds on the work by Ang et al. (2006).

[8] See Aastveit and Trovik (2007) for an example of a factor model for Norwegian GDP.

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

Table 16: Term structure models for GDP in the model suite in SAM

Model	**Description of model**	**Incl. in ensemble**
TStruct1	An unrestricted VAR model for the 3-month yield, the yield spread (5-years minus 3-months) and GDP	eTstruc
TStruct1a	An unrestricted VAR model for the 3-month yield, the yield spread (5-years minus 3-months) and GDP. The Kalman-filter is used to forecast GDP in case of jagged edge.	eTstruc
TStruct2	Affine model (structural VAR) with yields on different maturities (3-month NIBOR, 12-month NIBOR, and 2, 4 and 5 year interest rate swaps (in NOK)	eTstruc
TStruct3	Affine model (structural VAR) with yields on different maturities (3-month NIBOR, 12-month NIBOR, and 2, 4 and 5 year interest rate swaps (in NOK)	eTstruc

Note: The sources of the data is Statistics Norway and Norges Bank. GDP is for Mainland-Norway (s.a.). Due to data limitations, the term structure models only produce forecasts from 2004 and onwards. Prior to 2004, we use forecasts from an AR-model in the evaluation.