2010 | 06

Working Paper

Economics Department

Weights and pools for a Norwegian density combination

By Hilde Bjørnland, Karsten Gerdrup, Christie Smith, Anne Sofie Jore and Leif Anders Thorsrud

NB NORGES BANK

Working papers fra Norges Bank, fra 1992/1 til 2009/2 kan bestilles over e-post:
servicesenter@norges-bank.no
eller ved henvendelse til: Norges Bank, Abonnementsservice
Postboks 1179 Sentrum
0107 Oslo
Telefon 22 31 63 83, Telefaks 22 41 31 05

Fra 1999 og senere er publikasjonene tilgjengelige på www.norges-bank.no

Working papers inneholder forskningsarbeider og utredninger som vanligvis ikke har fått sin e4ndelige form. Hensikten er blant annet at forfatteren kan motta kommentarer fra kolleger og andre interesserte. Synspunkter og konklusjoner i arbeidene står for forfatternes regning.

Working papers from Norges Bank, from1992/1 to 2009/2 can be ordered by e-mail:
servicesenter@norges-bank.no
or from Norges Bank, Subscription service,
P.O.Box. 1179 Sentrum
N-0107Oslo, Norway.
Tel. +47 22 31 63 83, Fax. +47 22 41 31 05

Working papers from 1999 onwards are available on www.norges-bank.no

Norges Bank's working papers present research projects and reports (not usually in their final form) and are intended inter alia to enable the author to benefit from the comments of colleagues and other interested parties. Views and conclusions expressed in working papers are the responsibility of the authors alone.

ISSN 1502-8143 (online)
ISBN 978-82-7553-552-6 (online)

Weights and Pools for a Norwegian Density Combination*

Hilde Bjørnland
Norwegian School of Management (BI) and Norges Bank

Karsten Gerdrup
Norges Bank

Anne Sofie Jore
Norges Bank

Christie Smith†
Reserve Bank of New Zealand

Leif Anders Thorsrud
Norges Bank

May 19, 2010

Abstract

We apply a suite of models to produce quasi-real-time density forecasts of Norwegian GDP and inflation, and evaluate different combination and selection methods using the Kullback-Leibler information criterion (KLIC). We use linear and logarithmic opinion pools in conjunction with various weighting schemes, and we compare these combinations to two different selection methods. In our application, logarithmic opinion pools were better than linear opinion pools, and score-based weights were generally superior to other weighting schemes. Model selection generally yielded poor density forecasts, as evaluated by KLIC.
Keywords: Model combination; evaluation; density forecasting; KLIC
JEL Codes: C32, C52, C53, E52

"My own feeling is that different combining rules are suitable for different situations, and any search for a single, all purpose, "objective" combining procedure is futile." Winkler (1986)

1 Introduction

Monetary policy-makers make policy decisions about their instruments in the context of a fundamentally uncertain world. To ensure appropriate monetary

*The views expressed in this paper are those of the authors and should not be attributed to Norges Bank or the Reserve Bank of New Zealand.

†Corresponding author: Christie.Smith@rbnz.govt.nz; Ph +64 4 471 3740; Fax +64 4 473 1209. Economics Department, Reserve Bank of New Zealand, PO Box 2498, Wellington 6140, New Zealand.

policy decisions, central bankers must provide suitable characterisations of that uncertainty. In this article we use density combination methods to characterise uncertainty and provide short term macroeconomic forecasts. In the context of a broad suite of models, we explore the performance of different density combination methods (log and linear opinion pools) and weighting schemes (weights compiled from log scores and continuous ranked probability scores). We evaluate these combinations using logarithmic scores, which can be interpreted as evaluating densities on the basis of the Kullback-Leibler information criterion.

Boiled down to its essence, our paper is a horse race to identify which combination schemes yield the best density forecasts. The imperatives of policy-making mean that forecasts are always required, irrespective of whether the forecasts are consistent with some notionally ideal model. By identifying the best forecast densities it is hoped that the losses that entail from the density misspecification will be as small as possible. The quote from Winkler (1986) above suggests that there may not be a universally 'best' combination method, a view with which we have considerable sympathy. It also implies that different combination methods should be assessed on a case-by-case basis. This paper represents one such case, and illustrates a method that can be used to discriminate between different competing densities.

Although the importance of uncertainty on decision-making has long been realised by monetary policy-makers, the vast bulk of analysis has taken refuge in a certainty equivalence framework that enables policy-makers to disregard the properties of uncertainty. If a policy-maker's loss function is quadratic and the dynamics of the economy can be adequately represented with linear equations, certainty equivalence implies it is only necessary to focus on the first moments of future outcomes, appropriately discounted, to derive optimal policy (see Simon 1956 and Theil 1957). However, if the policy-maker's loss function is more complicated or if the world is nonlinear then it no longer suffices to focus solely on the first moments of possible outcomes entering the loss function, rather one may need to characterise *all* moments or, equivalently, the entire distribution of possible outcomes.[1] Consequently, forecasters should provide *density* forecasts rather than simply point forecasts reflecting expected values.[2]

Models should be specified and estimated taking into account the objective function of the end-user of the model, but typically the objective function is unknown. Ideally, different methods for forecasting densities would have a consistent ranking irrespective of the decision-maker's loss function; the highest ranked density forecast method would then be robust to the preferences of the ultimate end-user. However, as Diebold et al. (1998) and Granger and Pesaran (2000) discuss, there is no consistent ranking over competing, misspecified

[1] Karagedikli and Lees (2007), Surico (2007), Cukierman and Muscatelli (2008) and Aguiar and Martins (2008) all find evidence of significant asymmetries in monetary policy preferences. Dolado et al. (2004) find asymmetries in the reaction function of the Volcker-Greenspan regime and infer that policy preferences are asymmetric.

[2] Timmerman (2006, sn 2.1) makes a similar point, noting that, when the loss function depends solely on forecasts, the optimal combination weights will typically depend on the entire distribution of forecast errors.

density forecasts: different decision-makers with different loss functions could favour different density forecasting methods. On the other hand, if one of the density forecasting methods miraculously coincides with the true data generating process then this true density function will be preferred above all others since it enables the optimal action to be identified, which ultimately minimizes the expected loss of the policy-maker. The goal for forecasters is thus to provide density forecasts that are as close as possible to the truth, to facilitate a good approximation to the optimal action.

Policy-makers are confronted with a wide array of candidate models and hence candidate forecast densities. We use combination methods (cf. selection) to reconcile competing forecasts. Timmerman (2006) surveys combination methods and provides theoretical rationales in favour of combination – including unknown instabilities, portfolio diversification of models, and idiosyncratic biases. Empirical evidence also supports the use of combination methods, see for example Clemen (1989), Makridakis et al. (1982), Makridakis et al. (1993), Makridakis and Hibon (2000)), Stock and Watson (2004), and Clark and McCracken (2010). For point forecasting, a number of these papers find that simple, equal-weighted combination methods out-perform more sophisticated 'adaptive' methods where the weights are based on past performance. However, Jore et al. (2010) examine *density* forecasts and conclude that adaptive weights improve upon simple weights.

Our paper is similar to Gerard and Nimark (2008) and Eklund and Karlsson (2007) in that we use out-of-sample predictive performance to weight models together and investigate whether improved forecasts can be derived from model combination. However, Eklund and Karlsson and Gerard and Nimark use a single formal method for combination – Bayesian model averaging – while we explore both linear and logarithmic opinion pools, and we have a larger and more diverse model space. Unlike Kascha and Ravazzolo (2010) and Garratt et al. (2009), who compute densities using either a normal or multivariate t-distribution approximation, some of our densities are computed using simulation techniques. The use of simulation methods affects how we represent the densities and compute our linear and logarithmic opinion pools, which we elaborate on later in the paper.

We forecast year-on-year inflation in the Norwegian consumer price index adjusted for taxes and energy (CPIATE) and the year-on-year growth in gross domestic product (GDP) for 'Mainland Norway'. CPIATE is the main reference series used to evaluate the implementation of Norwegian monetary policy, while the Mainland Norway GDP series endeavours to exclude the impact of North Sea oil production.[3] Oil production is discounted for monetary policy purposes since it is capital intensive and much of the returns are invested abroad by the 'The Government Pension Fund - Global', colloquially known as the Norwegian Petroleum Fund. Consequently, off-shore oil production has little impact on domestic demand, and hence inflationary pressure.

[3]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 implementing monetary policy.

The paper is organised as follows. Section 2 covers three issues: i) the evaluation of the predictive forecast densities; ii) the mechanisms used to combine densities; and iii) the derivation of weights used to aggregate densities. In section 3 we describe the suite of models employed in our analysis and discuss how the forecast densities from individual models are characterised. We also describe how we conduct our quasi real-time analysis. Section 4 then reports the log scores of the forecast combinations, which enables us to rank the combination methods given our suite of models. These combination forecasts are also compared to those obtained from model selection. Section 5 concludes.

2 Density evaluation, model combinations and weighting schemes

2.1 Assessment of densities

'Scoring rules' are functions of predictive distributions and realised out-turns, and are used to evaluate the predictive densities (Gneiting and Raftery, 2007). When forecasters are assessed and rewarded relative to 'proper' scoring rules, the forecasters have an incentive to report their true beliefs about the density rather than trying to game the assessment by reporting some other forecast density, see Matheson and Winkler (1976) and Good (1952).

We use logarithmic scores to assess individual and combination densities, which are proper scoring rules in the sense described above. Log scores are linked to both Shannon entropy and the Kullback-Leibler information criterion (KLIC) (Gneiting and Raftery, 2007). The KLIC for the i^{th} model is:

$$E\left(\log(f(y_t)/P_i(y_t))\right) \tag{1}$$

where this expectation is taken with respect to the true unknown density $f(y_t)$. For a continuous distribution, this expectation is:

$$\begin{aligned} KLIC_i &= \int_{-\infty}^{+\infty} \log(f(y_t)/P_i(y_t)))f(y_t)dy \\ &= \int_{-\infty}^{+\infty} \log\left(f(y_t)\right) f(y_t)dy - \int_{-\infty}^{+\infty} \log\left(P_i(y_t)\right) f(y_t)dy \end{aligned} \tag{2}$$

The KLIC of a model's density represents the expected divergence of the model density relative to the true unknown density across the entire domain of the true density. The KLIC is non-negative, and only attains its lower bound of zero when $P_i(y_t) = f(y_t)$. The first integral on the right hand side of (2) is an unknown but fixed constant. Therefore the KLIC can be minimized by making the second term as large as possible, ie by maximizing $\int \log\left(P_i(y_t)\right) f(y_t)dy$. Assuming ergodicity this expectation can be approximated using a sample mean $\frac{1}{T}\sum_{t=1}^{T} \log\left(P_i(y_t)\right)$. Maximizing this quantity given a vector of data $(y_1, \ldots, y_T)'$ is simply maximum likelihood. Thus, a density with a higher (log) likelihood

value than another – *a higher score* – has a lower Kullback-Leibler information divergence and is closer to the true but unknown density.

Scores are affected by both the location and dispersion of a density. A density that is under-dispersed will have observations fall in its tails more often than predicted resulting in low average scores. Conversely, a distribution that is over-dispersed will have actual observations fall mainly in its centre, but the scores will be lower than they should because probability mass is spread over ranges that are only infrequently visited by the random variable of interest.

The log scores of the models and combinations are derived from the out-of-sample performance of the predictive densities. Performance is also assessed on a horizon-specific basis, since models are often found to have contrasting performance at different horizons (Makridakis and Hibon, 2000).

2.2 Opinion pools

Linear and logarithmic opinion pools are the two density combination schemes that have been most prominent in the literature.[4] The linear opinion pool is the most intuitive combination density, simply being:

$$P(y_t) = \sum_{i=1}^{n} w_i P_i(y_t) \tag{3}$$

where $P(y_t)$ is the combination density, $P_i(y_t)$ is an individual density obtained from the i^{th} model (suppressing parameter vectors for notational convenience) 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. Bayesian model averaging (BMA), which is often used to combine densities, is an example of a linear opinion pool.[5] Linear opinion pools can also be generalised to allow for a constant and indeed negative weights; see Genest and Zidek (1986) and Timmerman (2006).

The log opinion pool, on the other hand, is a geometric weighted average of the individual densities:

$$P(y_t) = K \cdot \Pi_{i=1}^{n} P_i(y_t)^{w_i} \tag{4}$$

where K is a constant to ensure that the log opinion pool integrates to 1.

It is worth noting several characteristics in which linear and logarithmic opinion pools differ. 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) – Genest and Zidek (1986) refer to this as a 'veto' property. Consequently a single, poorly specified density could have a material impact on a combination density from a log opinion pool. A log opinion pool will also be normally

[4] Garratt et al. (2010), this issue, adopt terminology from the meteorological literature and refer to density combinations as ensembles.

[5] See Hoeting et al. (1999) for an introduction to Bayesian model averaging.

distributed if the individual densities underpinning it are normally distributed; see Kascha and Ravazzolo (2010). Conversely, combining normally distributed probability distributions in a linear opinion pool results in a normal mixture distribution.

In general, log opinion pools are less dispersed than linear opinion pools, and are typically uni-modal (Genest and Zidek, 1986). Linear opinion pools, on the other hand, may be multi-modal. Kascha and Ravazzolo (2010) illustrate that when combining two densities with two distinct modes, the linear and logarithmic pools can result in quite different combination densities: in particular, the log opinion pool can place high probability mass between the modes of the individual densities even though each individual density places low probability on that intermediate region. Nevertheless, a priori one cannot say that such a density will be worse in terms of KLIC, and therefore both pooling operators need to be assessed empirically.

As we have no strong preference for particular density combination properties we investigate both linear and logarithmic opinion pools, evaluating their forecasting performance against each other. The choice of pool still leaves open the question of the choice of weights.

2.3 Combination weights

The simplest combination scheme uses equal weights for all of the densities that enter the combination. Equal weights imply that the data cannot reliably inform us about the performance of individual models. Nonetheless, as discussed earlier, equally-weighted combinations have been found to be surprisingly effective, at least for point forecasting.

For density combination, one of the most natural statistics to use to derive weights is the probability that a model could have generated the evaluation data. Loosely speaking, the log-likelihood is the log of the probability that the data were generated by a particular model. One way of deriving weights is thus by taking the exponent of the log-likelihood. The weights can then be specified as:

$$w_i = \frac{\exp\left(\log(P_i(\underline{y}))\right)}{\sum_{j=1}^{n} \exp\left(\log(P_j(\underline{y}))\right)} \tag{5}$$

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

Another candidate statistic that can be used to derive model weights is the continuous ranked probability score (CRPS). In an ideal world perfect forecasters would correctly anticipate all future events, and probability mass of one would be centred on the soon-to-be realised outcomes. The corresponding cumulative density functions would be step functions, with the steps also located at the realised outcomes. The CRPS can be conceptualised as a measure of the deviation from this step-function ideal. Let X_t be the variable being forecast and denote the actual out-turn $X_t = x_t$. Then the ideal cumulative forecast distribution, the 'heaviside function' centred on the actual observation, is a step

function that we denote:

$$\mathbb{I}(x \geq x_t) = \begin{cases} 1 \text{ if } x \geq x_t \\ 0 \text{ if } x < x_t \end{cases} \tag{6}$$

Denote the predictive forecast distribution of model M_i as $F(x; M_i)$. Then the CRPS for model M_i for the single observation at time t is:

$$CRPS_{it} = \int_{-\infty}^{+\infty} (\mathbb{I}(x \geq x_t) - F(x; M_i))^2 dx \tag{7}$$

and over a sample of random variables $X_1, ..., X_T$ one can simply take the average of $CRPS_{it}$ to assess the calibration of the densities from model M_i. Figure 1 illustrates the continuous ranked probability score for a single realisation (though for simplicity the shaded regions depict the *absolute* differences between the heaviside function and the forecast density, rather than the quadratic differences of the CRPS). Two forecast densities may have the same score – the same slope of the CDF evaluated at the realisation – but one density may be considered better than another because it has more probability mass 'near' the realisation. The CRPS provides a metric for discriminating between two such densities. The CRPS is also more forgiving of outliers than scores: if a realised value x is below (less than) the support of a model density then the slope of the cumulative distribution function, the score, would be zero but the CRPS would be $\int_x^{+\infty}(1 - F(x; M_i))^2 dx$, which would be a finite number. The score of zero for a single observation would result in no probability ever being attached to that model, even if the performance of the model density was very good for all other realisations. In contrast, the CRPS might still place positive weight on this model density, despite its egregious miss for one realisation. The more forgiving CRPS weights might therefore yield a better combination density.[6] The appendix provides more detail on the computation of weights from log scores and CRPS statistics.

3 Norges Bank models and estimation

3.1 SAM and the models

In 2006 Norges Bank instituted a 'Nowcasting' project to improve its short-term forecasting. One of the objectives of the project was to provide a formal, model-based characterisation of uncertainty. This objective has been framed in terms of obtaining good density forecasts through forecast combination. We use Norges Bank's System for Averaging Models (SAM) to evaluate the models and the combinations employed in this paper. Since it is entirely possible that the best models or combinations at time t will be superseded by some other models

[6] Any non-linear transformation of a statistic will affect the weights assigned to different models. Transforming the CRPS non-linearly to optimize the combination is a possibility, but is left for future research.

or combinations at a later date, it is important for the models and combinations to be evaluated on a recurrent basis, and this ongoing evaluation also helps to forestall mistakes in the implementation of the models.

Norges Bank's System for Averaging Models is used recurrently as part of the policy process to provide model-based forecasts for GDP and inflation. The suite of models in Sam includes univariate and vector autoregressions (ARs and VARs), factor models based on quarterly and monthly data, Bayesian VARs, term structure models, and a dynamic stochastic general equilibrium (DSGE) model. For GDP many of the models are simply bivariate autoregressions with four lags of both GDP and an indicator as regressors.[7] For GDP there are 144 models, and for CPIATE there are 84. There are ongoing efforts to expand the suite of models, and to widen the set of variables being forecast.

The models are estimated using a variety of different methods, from classical estimation to Markov chain Monte Carlo (MCMC) techniques. Consequently, the density forecasts are also obtained in a variety of different ways. Many of the VAR models provide density forecasts by assuming asymptotic normality (ignoring parameter uncertainty) and thus require just a point forecast and a standard deviation to compute the density, whilst densities for some of the other models are obtained by simulation, e.g. bootstrapping the estimation errors with replacement or using simulations from MCMC methods. These 'simulation samples' are smoothed using kernel density methods.

Since some predictive densities are computed *parametrically* and others are derived from *simulations*, we need to establish a common basis that can be used to weight the densities together into a combined density. One approach to construct a linear opinion pool would be to represent each density numerically by simulating, say, 1000 realisations from the density. One could then compute a weighted combination by sampling from each density's simulation sample with the desired weight. Suppose for example that there are two models, M_1 and M_2 and model M_1 had probability $2/6$ and model M_2 had probability $4/6$. Like a mixture distribution, one could roll a die and draw with replacement from M_1's simulation sample if $\{1, 2\}$ occurred on the die, and draw from M_2's simulation sample if the die return $\{3, 4, 5, 6\}$. One could draw N such variates resulting in a sample from the weighted density. The probabilities for this weighted density could be computed using a histogram or non-parametric methods; this histogram or non-parametric density is required to enable one to assess the score of an observation that might fall anywhere in the domain of the density.

While this simulation approach could be employed to approximate a linear opinion pool, it is not immediately obvious how it could be used to approximate a logarithmic opinion pool. Kascha and Ravazzolo (2010) compute logarithmic opinion pools, but they do so parametrically, making use of the fact that when the individual densities are normal (as they assume) then so to is the combined logarithmic opinion pool.

Given the number of models in our model suite, the number of recursive

[7] A brief description of most of the models in Sam is provided in Bjørnland et al. (2008), and a complete description of the models is available from the authors upon request.

samples, and the fact that we are also interested in logarithmic opinion pools, the simulation approach described above was impractical for our application. Instead we represent cumulative distributions using piecewise linear functions, with knots at a common grid of points $g \in \{g_1, g_2, \ldots, g_N\}$, where $g_{j-1} < g_j$.

If model M_1 has a simulation sample $\{X^{(1)}, X^{(2)}, \ldots, X^{(N)}\}$ to represent a density for a variable X being forecast, then the empirical cumulative distribution for model M_1 at point g is:

$$F_N(X \leq g; M_1) = \frac{1}{N} \sum_{k=1}^{N} \mathbb{I}(X^{(k)} \leq g) \tag{8}$$

where

$$\mathbb{I}(X^{(k)} \leq g) = \begin{cases} 1 \text{ if } X^{(k)} \leq g \\ 0 \text{ if } X^{(k)} > g \end{cases} \tag{9}$$

This empirical distribution function converges uniformly as the sample size N increases (Mood et al., 1974). For model M_1, $P(g_{j-1} < X \leq g_j; M_1) = F_N(X \leq g_j; M_1) - F_N(X \leq g_{j-1}; M_1)$; given the assumption of piecewise linearity, this probability is assumed to be evenly distributed across the interval $(g_{j-1}, g_j]$.[8]

Suppose that model M_2 has a parametric cumulative distribution $F(X \leq g; M_2)$. Then, as for models that use empirical distribution functions, $P(g_{j-1} < X \leq g_j) = F(g_j; M_2) - F(g_{j-1}; M_2)$. Although it is possible to evaluate the distribution function exactly at any intermediate point using the parametric distribution function, we approximate the distribution by using a piecewise linear function, with knots at the same grid as before, $g \in \{g_1, g_2, \ldots, g_N\}$. Placing both parametric and simulation densities on a common footing is done for computational convenience to ease the combination of the individual densities. The linear and log opinion pools are computed at $\{g_1, g_2, \ldots, g_N\}$ using (3) and (4).

The grid of knots used for both GDP and CPIATE lies in the interval $[-5, +10]$ and increments in steps of 0.05. This domain is broader than the unconditional domains of historical Norwegian GDP and CPIATE. In the context of year-on-year inflation, the grid encompasses deflation of 5 percent year on year, and moderately high inflation of up to 10 percent per year. Likewise, the analysis can incorporate declines in output of up to 5 percent per year, and expansions of up to 10 percent per year.

3.2 Estimation

To evaluate the models and the different combination schemes we perform an out-of-sample evaluation. This is a quasi real-time analysis, based on the most recent vintage of data. Given that we wish to mimic the forecasting problems faced by a central bank, it would be preferable to perform a complete real-time analysis, but Norwegian real-time data were not available when this analysis was undertaken.

[8] In practice we use non-parametric methods to compute the densities at the knots and assume the densities are flat across intervals such as $(g_{j-1}, g_j]$.

The models are first estimated up to 1998Q4, and then the estimation window is recursively expanded. We report results and statistics for our entire quasi out-of-sample period, 1999Q1-2008Q4. However, not all models are estimated on the same in-sample period, since different researchers have different views as to which data should be used to forecast forward (e.g. because of structural breaks, data availability, and/or changes in policy regime). Thus, it is not possible to provide a fair comparison of the models based on their in-sample fit, which motivates our choice of out-of-sample evaluation criteria (see section 2). Out-of-sample evaluation also acts as a natural penalty on over-parameterized models.

It is comparatively uncommon for a data set to be balanced.[9] When the data set is unbalanced the analysis proceeds as if the same unbalancedness prevails each time the model is estimated. The results reported in this paper are based on an unbalanced panel, with GDP in 2008Q4 being unknown. This unbalancedness is artificially maintained in earlier forecasting periods. The unbalancedness is artificial in the sense that publication schedules are not perfectly consistent through time. Suppose there are two quarterly series $\{X_t\}_{t=1}^{\infty}$ and $\{Y_t\}_{t=1}^{\infty}$ and their publication lags in days after the end of the t^{th} quarter are respectively δ_{xt} and δ_{yt}, with $\delta_{xt} < \delta_{yt}$. Suppose at an earlier time period s that $\delta_{xs} > \delta_{ys}$. Then at time t x_t is published before series y_t, but at time s the order of the publication dates is reversed. Thus a particular unbalancedness (such as x_s is known but y_s is not) would never have existed in practice.

The quasi real-time analysis proceeds as we would have in real-time, with the exception that we use the final vintage of data. Since there are delays before any evidence of forecast performance is available to assess the models, there are also delays in our ability to compute the adaptive weights used to form combination forecasts. For example, for the one-step-ahead forecast horizon we must wait for a quarter before we can weight the models together; for the two-step-ahead forecast horizon we must wait two quarters before we can assess the individual models, and so on for higher-step forecasts. Publication lags exacerbate the delay in being able to adaptively assess model performance. In early periods the forecasts from models are equally weighted, which can be thought of as an uninformative prior for the model weights.

4 Results – model and combination performance

In this section we analyse the performance of the models in the suites employed by Sam and the performance of various combination forecasts. We compare these combinations to densities obtained through quasi real-time selection of the best densities. The next subsection illustrates some statistics for the individual models. Then in section 4.2 we illustrate how the choice of weights and the type

[9] For example, National accounts data in Norway are published with lags of up to two quarters.

of opinion pool affects the performance of the resulting forecasting combinations. Our central research questions are thus:

1. Are densities from model combination better (in KLIC/score terms) than densities from model selection?
2. Do logarithmic or linear opinions provide better predictive density forecasts (evaluated by KLIC/score)?
3. Which weights – equal weights or weights derived from scores or CRPSs – yield the best (quasi) out-of-sample density forecasting performance, again evaluated by KLIC?

In our analysis there are two pools and three weighting schemes, resulting in six different combinations. We also compare our combination results to a selection strategy, where a weight of 1 is placed on a specific 'best' model, given the performance to date. The choice of best model evolves through time, and depends on the statistic used to determine which model is best: log score or CRPS.

4.1 Model evaluation

To set the scene, figures 2 and 3 plot the last round of forecast densities for mainland GDP growth and CPIATE inflation from the *individual models*.[10] The forecast densities are for growth rates which are calculated in year-on-year terms, and many of the models are estimated on these year-on-year growth rates. For GDP there are 144 models, and for CPIATE there are 84 individual models. The aim is thus to pool the individual densities into a single density, or to select one, for each variable at each horizon in order to help monetary policy decision-makers.

As can be seen from these figures, there is substantial variation in the densities provided by the various individual models. The variation is particularly marked for GDP forecasts, which is not surprising given that Norwegian GDP has been quite variable, whereas inflation has generally been low and fairly stable.

4.2 Evaluating combinations and selections

Figures 4 and 5 illustrate the performance of the individual densities, evaluated by their average log scores. (See also figure 1 for more insight into the computation of the scores.). In light of the previous figures it will come as no surprise that there is substantial variation in the weights attached to the individual forecast densities.

Tables 1 and 2 represent the crux of the empirical analysis in this paper. In these tables we report the densities from the six *combinations* and the densities *selected* by average log score and CRPS.

[10] We report average log scores, but of course the ranking is the same as if the log scores were cumulated and exponentiated.

For GDP the results are fairly clear: conditional on a given weighting scheme, logarithmic pools are preferred at all horizons according to the Kullback-Leibler information criterion. The weights that yield the best combination density are sometimes derived from scores and sometimes derived from CRPSs. However, even when weights from the CRPSs are better, the differences relative to the performance of the score-weighted linear opinion pools are only slight. The densities obtained through selection are almost universally the worst performers. At the one step forecast horizon for GDP, logarithmic and linear opinion pools with score-based weights yield the best densities, though the linear opinion pool is not much better than logarithmic opinion pools with either equal or CRPS-based weights. At longer horizons the choice of weighting scheme (score, CRPS, or equal) is not very material, conditional on a given pool type.

For CPIATE logarithmic opinion pools have larger log scores than linear pools for all weighting schemes. And within logarithmic pools score-based weights consistently yield the best combinations (as evaluated by score/KLIC). However, using scores as the selection criterion or using linear opinion pools with score based weights also yield fairly good density forecasts for CPIATE. In contrast to the GDP results, the choice of weighting scheme does matter for CPIATE: CRPS- and equal-weighted combinations are noticeably inferior to those that use score-based weights.

Given our model suites, equal weighted combination densities are (with one exception) inferior to combinations with weights derived from either scores or CRPSs. Combinations with equal weights are of course contaminated by the poorest individual densities. However, as we saw in figures 4 and 5, some of the models that are initially worst improve substantially over time, and ruling them out entirely could be premature as more data arrives.

Figures 6 and 7 illustrate the evolution of weights as more and more forecast rounds have been performed so that there is a more extensive sample of (quasi) out-of-sample forecasts that can be used to assess the models, and hence derive their weights. Each colour (or monochrome shade) corresponds to a model, and the height of each colour/shade corresponds to the weight attached to the model after a given number of forecast rounds. Naturally, each column sums to one. The early forecast rounds, where equal weights were attached to each model, are not depicted. The most obvious feature of these plots is that the identification of which models are 'best' changes through time. Initially some models have quite high weights, but then they fail to forecast some observations and their weights drop away reflecting this poor performance. Naturally the reverse happens too, some weights increase. The upshot is that we cannot be assured that the accumulation of data will enable us to select a single best model, or a combination with a subset of best performing models.

5 Conclusion

In this paper we investigated the forecasting properties of various combined densities for Norwegian GDP growth and Norwegian inflation. We evaluated

these forecast densities by considering their log scores, which is analogous to assessing them by the Kullback-Leibler information divergence. Our results imply that the performance of the combinations depends upon the interaction between the choice of pool and the choice of weights, and the underlying model space.

For our data and model suites, and evaluating densities by their scores, we found that the logarithmic opinion pools generally provided better densities (as measured by KLIC) than linear opinion pools. Furthermore, densities obtained through *combination* were generally better than densities obtained by *selecting* a best model, where selection was based on a statistic from the out-of-sample forecast performance. Though not universally best, weights obtained from scores were generally better than simple (equal) weights, and better than weights obtained from the CRPSs. It should be noted that the score based weights did exhibit substantial variation as the forecast rounds were conducted. Generally, a relatively small number of models received substantial weight, but the exact models receiving the weight evolved considerably through time.

The analysis illustrated in this paper helps to answer a 'live' policy question about which combination method to use. We do not formally test whether one combination is better than another since ultimately we need to choose a method to provide policy-makers with forecasts, and we do not care whether one method is only slightly or significantly better than another.

It has long been realised that specification choices affect model properties, and that one should take into account the possibility that there are other models that may be better than the single model that one selects. Specification choices are also important for model combinations: choices about model weights or the type of pool can have an important effect on the properties of a combination density forecast. Decisions about the choice of weights also have an impact; the precise consequences will depend on the suite of individual models underlying the combinations. Without investigating the different possible combinations, one cannot know if one's combination choices have been inspired or unknowingly mediocre.

The introduction to this paper emphasized that density forecasting is needed in monetary policy analysis to facilitate policy decision-making, as per the simple decision environment outlined in Diebold et al. (1998, sn. 2). For monetary policy that simple decision environment is somewhat unsatisfactory because future macroeconomic outcomes, and hence also macroeconomic forecasts, should be affected by the policy actions undertaken. If, however, policy can be well-represented by a rule, so that the effects of monetary decisions are already embedded in the reduced form dynamics of the data, or if such policy actions take a while to affect the economy, then near-term forecasts may be relative immune to such considerations. Nevertheless, explicitly incorporating policy conditionality into the forecasting process – making forecasts conditional on interest rates – would be a highly desirable direction in which to proceed, particularly when forecast horizons are a year or more.

6 Acknowledgements

This paper, originally titled 'There is more than one weight to skin a cat: Combining densities at Norges Bank', developed from a presentation at the 11-12 December 2008 *Nowcasting with model combination* workshop at the Reserve Bank of New Zealand. A later version was presented at the Deutsche Bundesbank's *Forecasting and Monetary Policy* conference in Berlin 23-24 March 2009. The current version of the paper was presented at the New Zealand Econometric Study Group Meeting in Auckland 26-27 March 2010. We thank Leni Hunter, Dorian Owen, Francesco Ravazzolo, Shaun Vahey, Ken Wallis, two anonymous referees, and the participants at the conferences for comments which helped the paper evolve considerably.

References

Aguiar, A., Martins, M. M. F., 2008. Testing for asymmetries in the preferences of the euro-area monetary policymaker. Applied Economics 40 (13), 1651–1667.
URL http://search.ebscohost.com/login.aspx?direct=true&db=bch&AN=32990522&loginpage=login.asp&site=ehost-live

Bjørnland, H. C., Jore, A. S., Smith, C., Thorsrud, L. A., 2008. Improving and evaluating short term forecasts at Norges Bank. Staff Memo 4, Norges Bank.

Clark, T. E., McCracken, M. W., 2010. Averaging forecasts from VARs with uncertain instabilities. Journal of Applied Econometrics 25 (1), 5–29.
URL http://ideas.repec.org/a/jae/japmet/v25y2010i1p5-29.html

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

Cukierman, A., Muscatelli, A., 2008. Nonlinear Taylor rules and asymmetric preferences in central banking: Evidence from the United Kingdom and the United States. The B.E. Journal of Macroeconomics 8 (1).

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

Dolado, J., Pedrero, R. M.-D., Ruge-Murcia, F. J., 2004. Nonlinear monetary policy rules: Some new evidence for the U.S. Studies in Nonlinear Dynamics & Econometrics 8 (3), 1–32.
URL http://search.ebscohost.com/login.aspx?direct=true&db=bch&AN=14600374&site=ehost-live

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

Garratt, A., Mitchell, J., Vahey, S. P., Wakerly, E. C., 2009. Real-time inflation forecast densities from ensemble Phillips curves. Working Paper BWPEF 0910, Birbeck, University of London.

Garratt, A., Mitchell, J., Vahey, S. P., Wakerly, E. C., 2010. Real-time inflation forecast densities from ensemble Phillips curves. North American Journal of Economics and FinanceForthcoming.

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

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

Gneiting, T., Balabdaoui, F., Raftery, A. E., 2007. Probabilistic forecasts, calibration and sharpness. Journal of the Royal Statistical Society B 69 (2), 243–68.

Gneiting, T., Raftery, A. E., 2007. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association 102, 359–376.

Good, I., 1952. Rational decisions. Journal of the Royal Statistical Society. Series B (Methodological) 14 (1), 107–14.

Granger, C. W. J., Pesaran, M. H., 2000. Economic and statistical measures of forecast accuracy. Journal of Forecasting 19 (7), 537–560.

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

Jore, A.-S., Mitchell, J., Vahey, S. P., 2010. Combining forecast densities from VARs with uncertain instabilities. Journal of Accounting and Economics 25 (4), 621–34, forthcoming.

Karagedikli, O., Lees, K., 2007. Do the central banks of Australia and New Zealand behave asymmetrically? Evidence from monetary policy reaction functions. Economic Record 83 (261), 131–42.

Kascha, C., Ravazzolo, F., 2010. Combining inflation density forecasts. Journal of Forecasting 29 (1-2), 231–250.

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.
URL http://search.ebscohost.com/login.aspx?direct=true&db=bch&AN=6142968&loginpage=Login.asp&site=ehost-live

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, J. E., Winkler, R. L., 1976. Scoring rules for continuous probability distributions. Management Science 22 (10), 1087–96.

Mood, A. M., Graybill, F. A., Boes, D. C., 1974. Introduction to the theory of statistics. McGraw-Hill, Inc., New York.

Simon, H. A., 1956. Dynamic programming under uncertainty with a quadratic criterion function. Econometrica 24 (1), 74–81.

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

Surico, P., 2007. The fed's monetary policy rule and u.s. inflation: The case of asymmetric preferences. Journal of Economic Dynamics and Control 31 (1), 305 – 324.
URL http://www.sciencedirect.com/science/article/B6V85-4JGJJ0B-2/2/68143adcb1fb2cbc3db26b7c35c32061

Theil, H., 1957. A note on certainty equivalence in dynamic programming. Econometrica 25 (2), 346–349.

Timmerman, A., 2006. Forecast combinations. In: Elliott, G., Granger, C. W. J., Timmerman, A. (Eds.), Handbook of Economic Forecasting. Vol. 1. Elsevier, Amsterdam, pp. 135–96.

Winkler, R. L., 1986. Comment on combining probability distributions: A critique and an annotated bibliography. Statistical Science 1 (1), 138–140.

Table 1: GDP: Average log scores

	H=1	H=2	H=3	H=4
Selection-score	-1.5569	-1.701	-1.8441	-1.8781
Selection-CRPS	-1.578	-1.6899	-1.7582	-2.2781
Linear pool-score	-1.2845	-1.5537	-1.7197	-1.8176
Linear pool-CRPS	-1.4671	-1.5897	-1.7316	-1.8332
Linear pool-equal	-1.4762	-1.591	-1.7378	-1.8403
Log pool-score	**-1.2189**	-1.5014	-1.6417	**-1.7553**
Log pool-CRPS	-1.3758	**-1.4977**	**-1.6331**	-1.7653
Log pool-equal	-1.3903	-1.5001	-1.6346	-1.7654

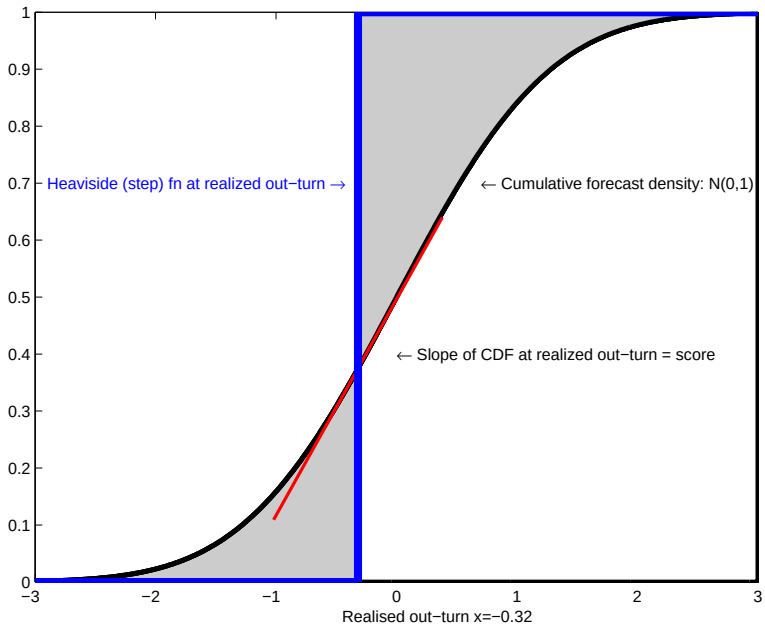
Each column corresponds to a different forecast horizon. The best performing combinations are highlighted in red/bold.

Table 2: CPIATE: Average log scores

	H=1	H=2	H=3	H=4
Selection-score	0.25833	-0.0999	-0.23992	-0.44781
Selection-CRPS	-0.16116	-0.25499	-0.64458	-0.96924
Linear pool-score	0.25526	-0.1346	-0.28698	-0.46247
Linear pool-CRPS	-0.032881	-0.45715	-0.68689	-0.8583
Linear pool-equal	-0.087209	-0.48884	-0.71791	-0.89089
Log pool-score	**0.37547**	**-0.088455**	**-0.18822**	**-0.38615**
Log pool-CRPS	0.29961	-0.1914	-0.41939	-0.59501
Log pool-equal	0.2567	-0.2202	-0.44782	-0.61869

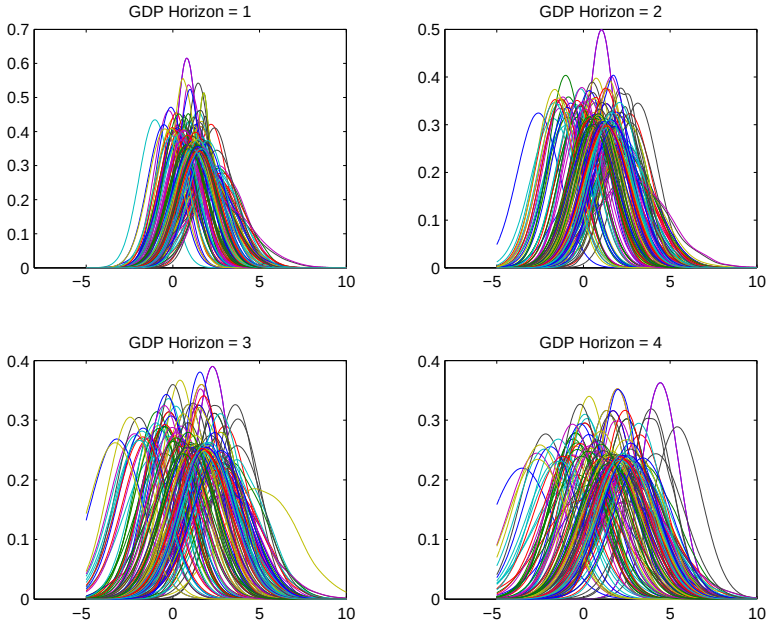
Each column corresponds to a different forecast horizon. The best performing combinations are highlighted in red/bold.

Figure 1: Illustration of the Continuous Ranked Probability Score



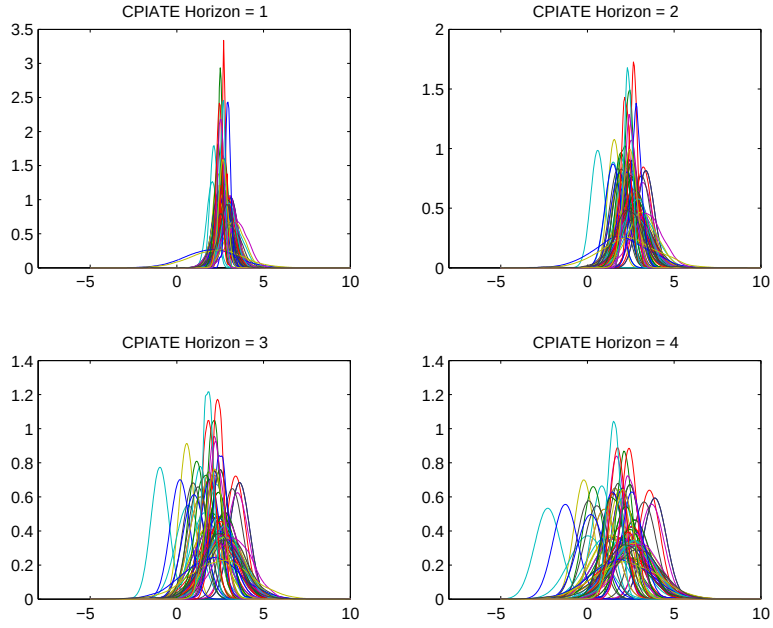
The CRPS is the integral of the squared heights in the shaded region (for convenience, the figure shades the absolute values not the squared heights of the CRPS). Thus, the CRPS is a measure of the divergence from an ideal forecast with probability mass centred on the observation that actually occurred. The *score* is the slope of the cumulative forecast density evaluated at the realisation.

Figure 2: GDP densities for all individual models



Forecast densities for GDP from the individual models from the last forecast round.

Figure 3: CPIATE densities for individual models



Forecast densities for CPIATE from the individual models from the last forecast round.

Figure 4: GDP average log scores

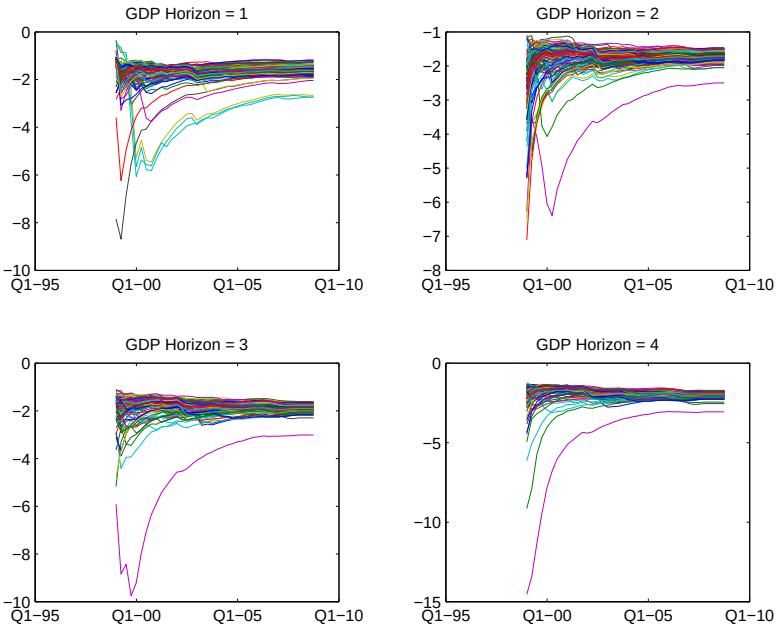


Figure 5: CPIATE average log scores

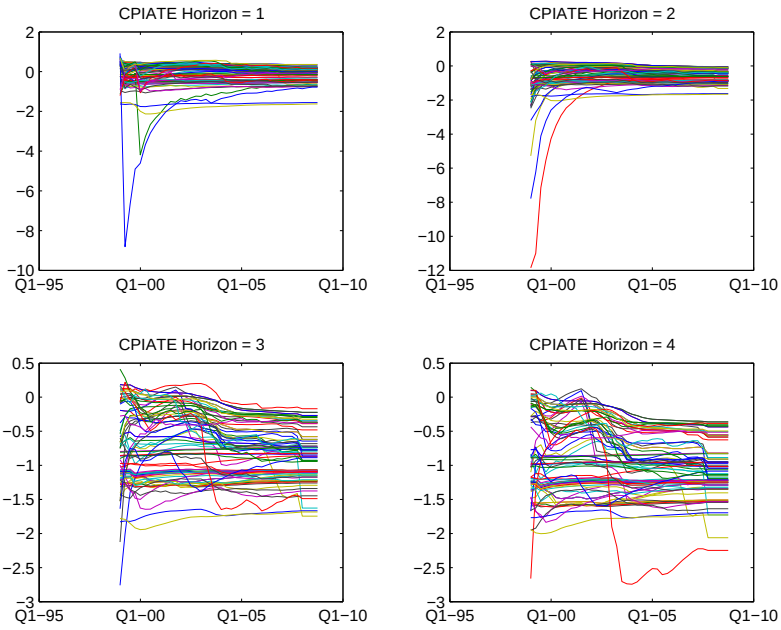
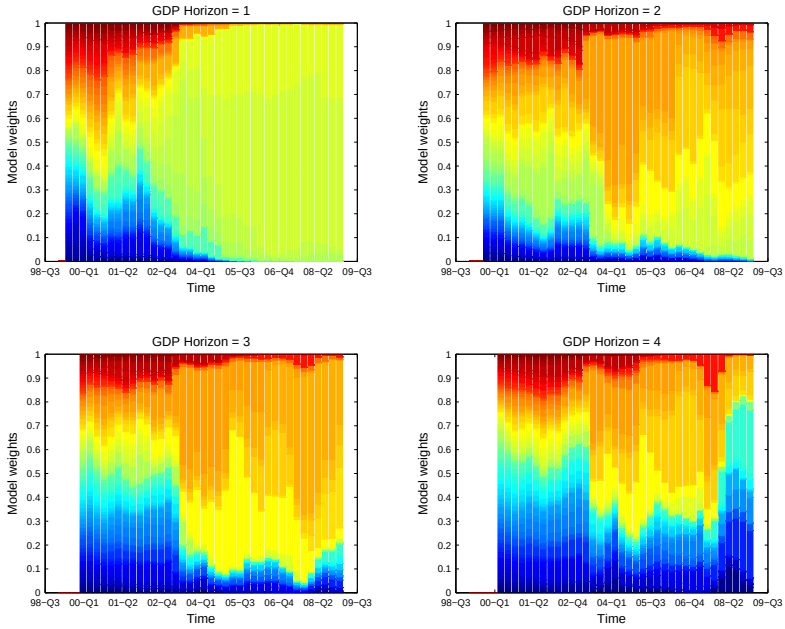
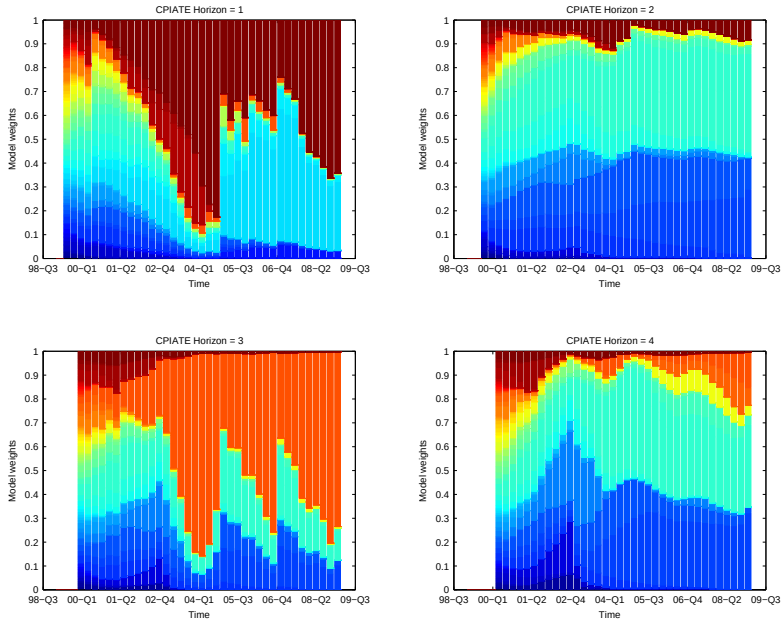


Figure 6: Evolution of GDP score-based weights



There are 144 models for GDP, but many of these models are poor with weights below 1 percent. Very low weights are not discernible in the figure. Across the recursive samples for horizon 1, at most 39 models receive weight greater than 1 percent, and for the last (full) sample only 7 models have weight greater than 1 percent.

Figure 7: Evolution of CPIATE score-based weights



There are 84 models for CPIATE, but many of these models are poor with weights below 1 percent. Very low weights are not discernible in the figure. Across the recursive samples for horizon 1, at most 41 models receive weight greater than 1 percent and only 8 models have weight greater than 5 percent. For the last (full) sample, for horizon 1, only 4 models have weight greater than 1 percent.

Appendix: Deriving weights

In principle, weights between zero and one that sum to one could be derived from any model statistic indicating performance, e.g.,

$$w_{it} = \frac{h(s_{it})}{\sum_{j=1}^{n} h(s_{jt})} \tag{10}$$

where $s_{it} \in S$ is a statistic from the i^{th} model at time t, and $h(\cdot)$ is a monotonic function, $h : S \to \mathbb{R}^+$. While such a function $h(\cdot)$ will preserve the ranking of models according to the statistic s, the function $h(\cdot)$ may have a material effect on the weight w_i that is subsequently assigned to the i^{th} model.

Given that we focus on the performance of density forecasts, our particular interest is in computing weights from statistics that account for both location and calibration, such as the score and the CRPS. Calibration in the context of density forecasting refers to the consistency of the density with the actual outcomes that are observed (Gneiting et al., 2007). In particular, if a forecast density is well-calibrated then outcomes should occur in given ranges with the frequencies predicted by the forecast density, and should therefore be neither under- nor over-dispersed relative to the data that actually arises. We include equal-weighted combinations to connect to the earlier point forecasting literature, which has found equal-weighted combinations to be difficult to surpass for point forecasting.

We compute weights using equation (10) where s is an out-of-sample log score or an average CRPS statistics. For log scores the function $h(s) = \exp(s)$ and for CRPS the function $h(s) = s$. The log scores are computed as per equation (4) in Kascha and Ravazzolo (2010). The CRPS for a single realised out-turn is computed using equation (7) (see also Gneiting and Raftery 2007, eq. 20), with $F(y; M_i)$, the cumulative distribution function of the i^{th} model, being represented as a piecewise linear function, as discussed in section 3. For computational convenience, the limits of the integral are -5 and $+10$, which encompasses the unconditional distributions of Norwegian GDP and CPIATE over the sample period used to estimate our models. The CRPS for the i^{th} model is then computed as the average CRPS from realised out-turns; using the sum of CRPSs rather than the average would result in the same weights as constants cancel out in the numerator and denominator of equation (10).