

2008 | 22

Working Paper

Research Department

Combining inflation density forecasts

by Christian Kascha and Francesco Ravazzolo

Working papers fra Norges Bank kan bestilles over e-post:

tjenestetorget@norges-bank.no
eller ved henvendelse til:
Norges Bank, Abonnementservice
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 endelige 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 can be ordered by e-mail:

tjenestetorget@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.

Combining Inflation Density Forecasts*

Christian Kascha[†] and Francesco Ravazzolo[‡]

Norges Bank

November 26, 2008

Abstract

In this paper, we empirically evaluate competing approaches for combining inflation density forecasts in terms of Kullback-Leibler divergence. In particular, we apply a similar suite of models to four different data sets and aim at identifying combination methods that perform well throughout different series and variations of the model suite. We pool individual densities using linear and logarithmic combination methods. The suite consists of linear forecasting models with moving estimation windows to account for structural change. We find that combining densities is a much better strategy than selecting a particular model *ex-ante*. While combinations do not always perform better than the best individual model, combinations always yield accurate forecasts and, as we show analytically, provide insurance against selecting inappropriate models. Combining with equal weights often outperforms other weighting schemes. Also, logarithmic combinations can be advantageous, in particular if symmetric densities are preferred.

JEL classification: C53, E37

Keywords: Forecast Combination, Logarithmic Combinations, Density Forecasts, Inflation Forecasting

*We are very grateful to James Mitchell, Shaun Vahey, to participants at the 14th Conference on Computing in Economics and Finance, at the 2008 Australasian Meeting of the Econometric Society, at the 5th Eurostat Colloquium on Modern Tools for Business Cycle Analysis as well as to seminar participants at Norges Bank for comments. The views expressed in this paper are our own and do not necessarily reflect the views of Norges Bank.

[†]*Corresponding author:* Christian Kascha, Norges Bank, Research Department, Bankplassen 2, 0107 Oslo, Norway. Telephone: +47 22 31 67 19. Fax: +47 22 42 40 62. christian.kascha@norges-bank.no

[‡]Norges Bank, Research Department. francesco.ravazzolo@norges-bank.no

1 Introduction

This paper compares some common approaches for combining density forecasts from a given suite of models using the Kullback-Leibler divergence as a measure of accuracy. A similar model suite is applied to data sets from four different countries to forecast inflation. The aim of the paper is to assess the performance of these combination schemes relative to each other and to the performance of the individual models. In particular, we base our evaluation on the combinations' performance *throughout* different data sets and variations of the model suite. We find that some simple combination schemes perform well in all cases considered and often outperform the best individual model.

The value of a point forecast can be increased by supplementing it with some measure of uncertainty. Interval and density forecasts are considered an important part of the communication from policymakers to the public. For example, the Bank of England as well as the central bank of Norway, Norges Bank, publish so-called fan charts for inflation that communicate the banks' views on possible paths of future inflation. However, policymakers usually have a whole suite of forecast models at hand. In this situation, some questions naturally arise whether one is just interested in point forecasts or whether one is trying to predict densities. First, should one choose a single model or combine the individual models to form some sort of consensus forecast? Second, in which way should one possibly combine the individual forecasts?

For the combination of point forecasts, the literature has reached a relatively mature state dating back to papers such as Bates and Granger (1969). In their framework, they show that the forecast error variance resulting from the combined forecast is less or equal to the forecast error variance of the best individual model in the suite, provided that optimal weights are chosen. They also show that major improvements are most likely to happen when the individual forecasts are not highly correlated. Timmermann (2006) provides an extensive summary of the literature on combining point forecasts and the success of forecast combinations in this field motivates quite naturally the extension to density forecasts. However, the literature on density forecasting and on density combinations emerged only recently.

Tay and Wallis (2000) and Corradi and Swanson (2006a) provide surveys on the evaluation of individual density and interval forecasts. See also Clemen, Murphy and Winkler (1995). Clements (2006) and Granger, White and Kamstra (1989) have considered a combination of event and quantile forecasts. While Genest and Zidek (1986) provided a survey on density combination in a rather decision-theoretic framework, Wallis (2005) is one of the recent papers in economics on density combinations. Some combination schemes for density forecasts have been proposed probably because of their success in combining point forecasts (Hendry and Clements, 2004; Granger and Jeon, 2004; Wallis, 2005). Mitchell and Hall (2005) and Hall and Mitchell (2007) provide some justification for density combination and propose the Kullback-Leibler divergence as a unified measure for the evaluation and combination of density forecasts.

Bayesian approaches naturally lend themselves to density combination schemes. Therefore, it might not be surprising that various possible approaches have emerged in this field. For example, Min and Zellner (1993) propose simple combinations based on posterior odds ratios. Palm and Zellner (1992) propose a combination method that captures the full correlation structure between the forecast errors resulting from different models by explicitly modeling their dynamic interaction. Following Morris (1974, 1977) and Winkler (1981), Hall and Mitchell (2004) consider an approach where competing density forecasts are combined by a “decision maker” who views these forecasts as data that are used to update a prior distribution. Bayesian model averaging (BMA) methods have been proposed by Leamer (1978), Raftery, Madigan and Hoeting (1997) and Geweke and Whiteman (2006).¹

There are very few studies in economics that take a comparative point of view and evaluate alternative methods of density forecast combination. Jore, Mitchell and Vahey (2007) provide some evidence on the performance of the weighting scheme proposed by e.g. Mitchell and Hall (2005) relative to equal weights and the pairwise equal averaging method of Clark and McCracken (2007). However, our knowledge of when and why predictive density combinations work is still very limited. As Hall and Mitchell (2007) state: “It is important to try to build up both an increased understanding and an empirical consensus about the circumstances in which density forecast combination works.” Taking inflation density forecasting as a relevant example, we therefore extend the empirical literature in two ways. First, we compare different functional forms of density aggregation. Second, we apply a similar model suite to different data sets and focus on the combination methods’ performance *throughout* these data sets in order to obtain results which can be expected to be more robust to variations in the model suite and sample period. To the best of our knowledge, these features are not simultaneously shared by any of the other empirical studies focusing on the combination of predictive densities.

Specifically, we compare combinations of density forecasts for inflation using a suite of linear, univariate and multivariate forecasting models and compare the results over data sets for the US, the UK, Norway and New Zealand. The size of the model suite is relatively modest. All models are estimated using a moving window of fixed size to account for structural change. We investigate two possible ways of aggregation. The first method is the “linear opinion pool” proposed by Stone (1961). This method was used almost exclusively in empirical applications on density forecast combination. The second method is the “logarithmic opinion pool” (see e.g. Winkler, 1968). We consider three different methods to construct model weights for each of the aggregation methods: equal weights, recursive log score weights and (inverse) mean squared error weights. We show that both combination methods always provide insurance against selecting bad models. Then, we study how predictive density combinations perform relative to individual density forecasts and selecting the best performing model at the forecasting origin.

¹Some applied Bayesian approaches can be found in Garratt, Lee, Pesaran and Shin (2003); Jackson and Karlsson (2004) and Andersson and Karlsson (2007).

Our results show that combining forecasts provides much more accurate forecasts than selecting a particular model at the forecast origin in almost all cases. Furthermore, the performance obtained by combining is in several cases better than the result for the ex-post best individual model. We do not find clear support for linear or logarithmic combinations. Equal weights and mean square error weights provide more uniform results over the different data sets than recursive log score weights. The latter weighting scheme provides marginally more accurate forecasts for the US but performs worse than selection for the other three countries. We show that this result is largely due to the presence of two extremely well performing models in the case of the US. Thus, there is less “model uncertainty” in this special case. As model uncertainty increases, recursive log score weights become a less adequate weighting scheme.

The rest of the paper is organized as follows. In section two we discuss the evaluation and combination of density forecasts. In section three we describe the data and the suite of density forecast models. Section four contains the results of the out-of-sample experiment. Section five concludes.

2 Evaluating and Combining Density Forecasts

Since the field of density forecasting is in a relatively infant state in economics and, importantly, no firm standards exist as in the point forecasting literature, we give a brief overview of the areas that are most relevant to this study. One is how to evaluate predictive densities and the problem here is that the true density is never observed - not even after the random variable is drawn. Another question is how to combine predictive densities and the main choices to be made are the functional form of aggregation and the weighting scheme for the individual models.

2.1 Evaluating Density Forecasts

The question of how to measure the accuracy of density forecasts has recently received a lot of attention in the theoretical literature. Corradi and Swanson (2006a) provide an extensive survey. This question is decisive because it is central to how we design density combination schemes (Hall and Mitchell, 2007). The essential problem in evaluating density forecasts is that the true density is not observed - not even *ex-post*. Additional difficulties arise if one wants to compare multiple models that are misspecified and sometimes nested.

One branch of the literature is concerned with testing whether predictive densities are correctly specified (Bierens, 1982; Bierens and Ploberger, 1997). These tests require the assumption of correct specification of the density forecast under the null hypothesis using all the relevant information (e.g. Diebold, Gunther and Tay, 1998; Bai, 2003) or conditional on a given information set (Corradi and Swanson, 2003). Among these measures, the use of

probability integral transforms (PITs) is popular.

Another branch is concerned with the evaluation of multiple, possibly misspecified models. One possibility is to evaluate density forecasts in terms of their implied economic value (Granger and Pesaran, 2000; Clements, 2004). This strategy makes a lot of sense in areas such as financial econometrics but is less meaningful for policymakers such as central banks. Therefore we stick to statistical measures. Two approaches have been considered in the recent literature. One is based on a distributional analog of the mean squared error norm (Corradi and Swanson, 2004, 2006b), the other is based on the Kullback-Leibler divergence or Kullback-Leibler Information Criterion (KLIC) (Kitamura, 2002; Mitchell and Hall, 2005; Amisano and Giacomini, 2007).

The measure of distributional accuracy introduced by Corradi and Swanson (2004, 2006b) is attractive because of its analogy to the usual mean squared error norm in point forecasting. Given a benchmark density function, a norm over a set of possible density forecasts is defined in a straightforward manner taking the expectation of the squared, point-wise difference between a candidate density and the benchmark density over all possible outcomes of the variable to be forecasted. One problem is the dependence on a benchmark density which might be difficult to justify in our case unless one uses a nonparametric estimate as in Li and Tkacz (2006).

On the other hand, measures based on the well-known KLIC can circumvent this problem. The KLIC is a sensible measure of accuracy, since it chooses the model which on average gives higher probability to events that have actually occurred. As argued by Mitchell and Hall (2005), the KLIC is advantageous because it provides a unified framework for evaluating, comparing and combining density forecasts. Also, the KLIC can be related to other measures which have been used to evaluate density forecasts *ex-post*, such as the PITs or Berkowitz's (2001) likelihood ratio tests. Measures in terms of the KLIC have also a Bayesian interpretation as the KLIC-best model is also the model with the highest posterior probability (Fernández-Villaverde and Rubio-Ramirez, 2004). For the i.i.d. case, Vuong (1989) suggests a likelihood ratio test for choosing the conditional density model that is closest to the true density in terms of the KLIC. The tests were extended by Amisano and Giacomini (2007) to cover the case of dependent observations. Also Kitamura (2002) employs a KLIC-based approach to select between misspecified models.

Specifically, suppose f_t is the true density of a random variable Y_t with domain and range \mathbb{R} and we have a set of two densities $f_{t,i}$, $i = 1, 2$ obtained from different models. We will call this set a *suite*, its elements *individual densities* and the underlying models *individual models*. The KLIC distance between f_t and $f_{t,i}$ is defined as

$$\begin{aligned} \text{KLIC}_i &= \int f_t(y_t) \ln \frac{f_t(y_t)}{f_{t,i}(y_t)} dy_t, \\ &= E[\ln f_t(y_t) - \ln f_{t,i}(y_t)], \end{aligned} \quad (1)$$

where \int means the integral taken over \mathbb{R} and E denotes the expectation. We assume here and in the following that all densities are strictly positive, i.e. $f_t(y) > 0$, $f_{t,i}(y) > 0$ for all $y \in \mathbb{R}$. In order to compare the KLIC of f_1 , f_2 we only need to evaluate the last term of the expectation in (1). That is, the *expected logarithmic score* (ElnS):

$$\text{ElnS}_i = E[\ln f_{t,i}(y_t)]. \quad (2)$$

Thus, when $\text{ElnS}_1 > \text{ElnS}_2$ then $\text{KLIC}_1 < \text{KLIC}_2$. Under some regularity conditions, a consistent estimate of (2) can be obtained from the average of the sample information, y_1, \dots, y_T :

$$\ln S_i = \frac{1}{T} \sum_{t=1}^T \ln f_{t,i}(y_t). \quad (3)$$

Therefore, we actually do not need to know f_t to compare f_1 and f_2 and we choose the model for which the expression in (3) is maximal. The last expression will be called (*average*) *logarithmic score* or simply *log score* (lnS) in the following.

Turning to density forecasts, let $f_{t+h,t,i}$ denote a prediction of the density for Y_{t+h} , conditional on information up to date t . Density forecasts are also sometimes called *predictive densities*. Let y_{t+h} be the realization of Y_{t+h} and suppose that h -step-ahead density forecasts have been obtained starting at time T^s and given a total number of T observations. A measure of out-of-sample forecasting performance is the (out-of-sample) log score given by

$$\ln S_{i,h} = \frac{1}{T - h - T^s + 1} \sum_{t=T^s}^{T-h} \ln f_{t+h,t,i}(y_{t+h}). \quad (4)$$

Models or combination schemes that are associated with a high average log score approximate well the unknown true density in terms of the KLIC (Hall and Mitchell, 2007). Therefore, (4) is our preferred measure of forecast accuracy.

2.2 Combining Density Forecasts

There are two elementary choices in combining predictive densities. One is the way of aggregation or functional form of combining. The other is the construction of the weights attached to the individual densities. Possible ways of aggregation are described in an early review of Genest and Zidek (1986). We consider two different functional forms, the linear combination and the logarithmic combination.²

Consider N competitive h -step-ahead density forecasts obtained using information up to time t . The easiest combination method is linear combination (Stone, 1961):

$$f_{t+h,t}^c(y_{t+h}) = \sum_{i=1}^N \omega_{t+h,t,i} f_{t+h,t,i}(y_{t+h}), \quad (5)$$

where the suite of individual density forecasts is $\{f_{t+h,t,1}, \dots, f_{t+h,t,N}\}$ and $\omega_{t+h,t,i}$ are the corresponding weights. The weights have to be a convex linear combination, that is, $0 \leq \omega_{t+h,t,i} \leq 1$ and $\sum_{i=1}^N \omega_{t+h,t,i} = 1$ for all $i = 1, \dots, N$ such that the resulting combination is indeed a density function (see also Genest, 1983).

An alternative way of combining densities is logarithmic combination:

$$f_{t+h,t}^c(y_{t+h}) = \frac{\prod_{i=1}^N f_{t+h,t,i}^{\omega_{t+h,t,i}}(y_{t+h})}{\int \prod_{i=1}^N f_{t+h,t,i}^{\omega_{t+h,t,i}}(y_{t+h}) dy_{t+h}}, \quad (6)$$

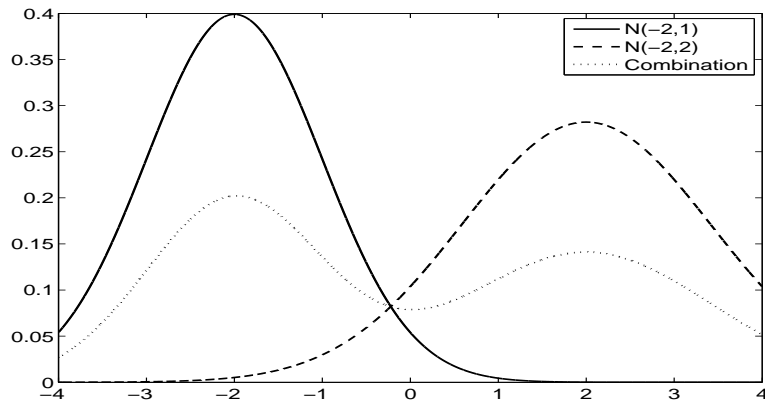
where the non-negative weights are chosen such that the integral in the denominator exists. In this paper we focus on weighting schemes that satisfy the conditions given for the linear opinion pool. Winkler (1968) points out that the logarithmic opinion pool has a natural-conjugate interpretation. The logarithmic combination method has the important property that it retains the symmetry of individual forecasts whereas linear combination does not. This is relevant for policymakers that might prefer symmetric forecast densities.

Some reasons to combine density forecasts and to use the above schemes in particular have been given in the literature. Genest and Zidek (1986) give reasons for the functional forms in (5) and (6) in a decision-theoretic framework. Logarithmic combination has been put forward by Winkler (1968) using a specific model. Raftery et al. (1997) and Mitchell and Hall (2005) argue for linear combination together with weights corresponding to the marginal likelihood of different models in a Bayesian framework. However, Mitchell and Hall (2005) point out that using only approximative Bayesian weights might lead to worse density forecasts even in-sample.

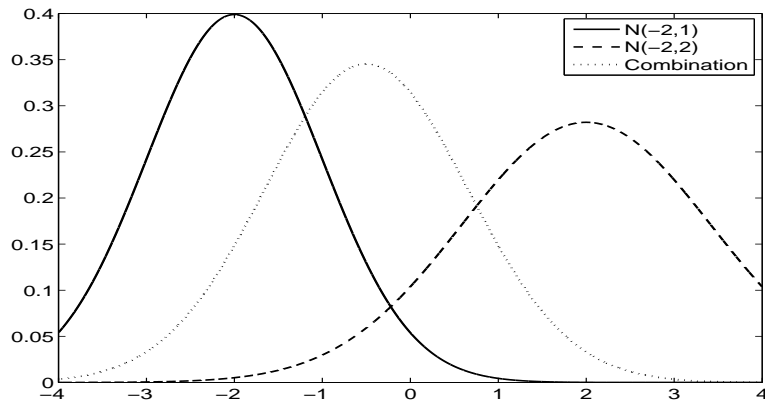
An example in Figure 1 illustrates the main difference between the two aggregation schemes. Let f_1, f_2 be the density functions of two normally distributed random variables

²Some alternative approaches are also given in Robertson, Tallman and Whiteman (2005) and Hall and Mitchell (2004).

with $N(-2, 1)$ and $N(2, 2)$, respectively. Let the weight for each individual density be $1/2$. Linear combination leads to $f^{lin} = 0.5f_1 + 0.5f_2$ with mean 0 and variance 4.5. Logarithmic combination leads to f^{log} which is the density of a normal random variable with mean $-1/2$ and variance $4/3$, $N(-1/2, 4/3)$ (See Appendix B). From the definition, it is immediately clear that f^{lin} is typically multimodal while f^{log} is unimodal. Furthermore, f^{lin} is generally more dispersed than any of the individual densities. The same weighting scheme can therefore yield substantially different densities, depending on the functional form of aggregation.



(a) Linear combination



(b) Logarithmic combination

Figure 1: Individual densities and density combinations using linear and logarithmic combination.

Similar to Hendry and Clements (2004), one can show that the combination of forecasts with deterministic weights such as equal weights provides “insurance” against selecting a bad

model. This means that a combination of density forecasts with equal weights will never be worse than the worst individual forecast. Let $f_1(y_{t+h}), f_2(y_{t+h})$ be the individual density forecasts obtained in some way. A forecaster unaware of the actual data-generating process combines both density forecasts with weights $\lambda, 1 - \lambda, \lambda \in [0, 1]$ according to either (5) or (6). Suppose that forecast 1 is better than forecast 2 in terms of the KLIC. This implies

$$E[\ln f_1(y_{t+h})|I_t] \geq E[\ln f_2(y_{t+h})|I_t], \quad (7)$$

where the expectation is taken conditional on the available information at time t, I_t . For the linear combination, we get for all $y_{t+h} \in \mathbb{R}$:

$$\ln(\lambda f_1(y_{t+h}) + (1 - \lambda)f_2(y_{t+h})) \geq \lambda \ln f_1(y_{t+h}) + (1 - \lambda) \ln f_2(y_{t+h}),$$

because of the concavity of \ln . Since the above relationship is true for all y_{t+h} , we have

$$\begin{aligned} E[\ln(\lambda f_1(y_{t+h}) + (1 - \lambda)f_2(y_{t+h}))|I_t] &\geq \lambda E[\ln f_1(y_{t+h})|I_t] + (1 - \lambda) E[\ln f_2(y_{t+h})|I_t], \\ &\geq E[\ln f_2(y_{t+h})|I_t], \end{aligned}$$

because λ is deterministic and (7). That is, the linear combination is never worse than the worst individual forecast. For the logarithmic combination, we get for all y_{t+h} :

$$\begin{aligned} \ln \frac{f_1^\lambda(y_{t+h})f_2^{1-\lambda}(y_{t+h})}{\int f_1^\lambda(y_{t+h})f_2^{1-\lambda}(y_{t+h})dy_{t+h}} &= \lambda \ln f_1(y_{t+h}) + (1 - \lambda) \ln f_2(y_{t+h}) \\ &\quad - \ln \int f_1^\lambda(y_{t+h})f_2^{1-\lambda}(y_{t+h})dy_{t+h}, \\ &\geq \lambda \ln f_1(y_{t+h}) + (1 - \lambda) \ln f_2(y_{t+h}) \\ &\quad - \ln \int \lambda f_1(y_{t+h}) + (1 - \lambda)f_2(y_{t+h})dy_{t+h}, \end{aligned}$$

where we use the fact that for all $x_1, x_2 \in \mathbb{R}, x_i > 0$ and $\lambda \in [0, 1]$ it holds that $x_1^\lambda x_2^{1-\lambda} \leq \lambda x_1 + (1 - \lambda)x_2$ (Jensen's inequality). Therefore,

$$\begin{aligned} \ln \frac{f_1^\lambda(y_{t+h})f_2^{1-\lambda}(y_{t+h})}{\int f_1^\lambda(y_{t+h})f_2^{1-\lambda}(y_{t+h})dy_{t+h}} &\geq \lambda \ln f_1(y_{t+h}) + (1 - \lambda) \ln f_2(y_{t+h}) \\ &\quad - \ln \left[\lambda \int f_1(y_{t+h})dy_{t+h} + (1 - \lambda) \int f_2(y_{t+h})dy_{t+h} \right], \\ &= \lambda \ln f_1(y_{t+h}) + (1 - \lambda) \ln f_2(y_{t+h}) \\ &\quad - \ln[\lambda + (1 - \lambda)], \\ &= \lambda \ln f_1(y_{t+h}) + (1 - \lambda) \ln f_2(y_{t+h}). \end{aligned}$$

Therefore, combining provides insurance even in this case:

$$\begin{aligned} E \left[\ln \frac{f_1^\lambda(y_{t+h}) f_2^{1-\lambda}(y_{t+h})}{\int f_1^\lambda(y_{t+h}) f_2^{1-\lambda}(y_{t+h}) dy_{t+h}} \middle| I_t \right] &\geq \lambda E[\ln f_1(y_{t+h}) | I_t] + (1 - \lambda) \ln E[f_2(y_{t+h}) | I_t], \\ &\geq E[f_2(y_{t+h}) | I_t]. \end{aligned}$$

While equal weights provide insurance in the above sense, other weighting schemes might yield even better density combinations. We therefore consider several recent proposals in the emerging literature as well as the empirical evidence on the combination of point forecasts.

Equal weights (EW): Equal weights are used in the aggregation of the forecasts in the Survey of Professional Forecasters to publish a combined density forecast for inflation. Equal weights for combining densities have also been proposed in the literature by Hendry and Clements (2004) and Wallis (2005). Formally, $\omega_{t+h,t,i} = 1/N$ for all t, h, i .

Recursive log score weights (RLSW): If we are measuring density fit by the logarithmic score, then it is only natural to base the construction of combination weights on past out-of-sample forecast performance measured in the same way. A promising candidate weighting scheme are recursive log score weights as proposed in e.g. Jore et al. (2007). The weights for the h -step ahead density combination take the form:

$$\omega_{t+h,t,i} = \frac{\exp[\sum_{\tau=\underline{t}}^{t-h} \ln f_{\tau+h,\tau,i}(y_{\tau+h})]}{\sum_{k=1}^N \exp[\sum_{\tau=\underline{t}}^{t-h} \ln f_{\tau+h,\tau,k}(y_{\tau+h})]}, \quad (8)$$

where \underline{t} is the beginning of the evaluation period and is taken as fixed. The weights can be regarded as derived in a Bayesian framework to approximate the models' posterior probabilities. Mitchell and Hall (2005) discuss the relation to the KLIC.

Mean squared error weights (MSEW): In point forecast combination, weights are often derived by the models' relative inverse mean squared prediction error (MSPE) performances computed over a window of previous observations. These are not "optimal" weights in a linear framework as MSPE weights ignore the correlation structure between forecasts (Granger and Ramanathan, 1984). However, these weights tend to outperform more sophisticated weighting schemes as the correlation matrix of the forecast errors is quite difficult to estimate. The weights for the h -step ahead density combination take the form:

$$\begin{aligned} \omega_{t+h,t,i} &= \frac{1/MSP E_{t+h,t,i}}{\sum_{k=1}^N 1/MSP E_{t+h,t,k}}, \\ MSP E_{t+h,t,i} &= \frac{1}{t-h-\underline{t}+1} \sum_{\tau=\underline{t}}^{t-h} (y_{\tau+h} - \mu_{\tau+h,\tau,i})^2, \end{aligned} \quad (9)$$

where $\mu_{\tau+h,\tau,i}$ denotes the mean (or point) forecast of model i and \underline{t} is again the beginning of the evaluation period.

Selection (SELEC): Selecting the best model is the obvious alternative to combining and therefore we also describe how we implement selection in this section. Our strategy uses only the information available at the forecast origin (abstracting from real-time issues). It is natural to assume that a forecaster, *if* he has to select one model, chooses the model that performed best in the past. Since we are interested in predictive densities, the relevant criterion is here the past performance of the models in terms of the average log score for a given horizon. Note that it is less interesting to compare the performance of the combination methods to the performance of each individual model *ex-post* because this kind of comparison ignores the model uncertainty at the time forecasts are made. For example, the integrated moving-average model to forecast US inflation proposed in Stock and Watson (2007) performs very well. However, just the date of Stock and Watson (2007) indicates that probably nobody would have chosen this particular model, say, 20 years ago. Note that the way we select models should be closely related to the standard AIC criterion based on the predictive likelihood.

3 Data and Models

3.1 Data

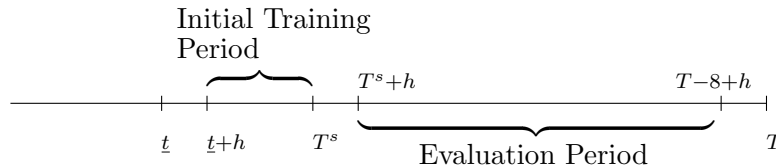
We take inflation density forecasting as a relevant example to evaluate different ways of combining predictive densities. For each country, there is a sample of available observations, x_1, \dots, x_T of size T and x_t is a vector of observations including the price level series p_t . We are interested in forecasting quarter-to-quarter inflation measured by the quarterly log change, $\pi_t = \Delta_1 \ln p_t = \ln p_t - \ln p_{t-1}$. We consider the Personal Consumption Expenditure (PCE) index for the US, Consumer Price indices (CPI) for the UK and New Zealand and the Norwegian core CPI.³ The set of potential predictors contains a quarterly M2 money measure (M1 in the case of New Zealand), M_t , a three-month quarterly interest rate, i_t , a quarterly output measure, y_t , and a quarterly unemployment rate, ur_t . We use real output as a measure of US GDP. Quarterly real GDP series are used for the other three countries. We use seasonally unadjusted series apart from the New Zealand production and unemployment series. Also, we abstract from the real time aspects and use the latest available vintage for simplicity. Data sources can be found in the Appendix.⁴

We start to compute individual forecasts 1 to h -steps ahead beginning at time \underline{t} . At time $T^s = \underline{t} + 10$ we start to compute forecasts also for the combination methods using the information on the out-of-sample performance of the individual models for $\underline{t} + h$ to T^s . The

³We focus on core CPI for Norway as energy prices have a dominant role in the Norwegian CPI. Norwegian energy prices in turn are affected largely by weather conditions.

⁴All data are available from the authors upon request.

evaluation period for all models and combinations is, depending on the horizon, $T^s + h$ to $T - 8 + h$ since 8 is the maximal forecasting horizon. The following graph illustrates our approach:



We consider sample periods that run from 1960 Q1 to 2007 Q3 for US data ($T = 191$), from 1978 Q1 to 2007 Q2 for UK data ($T = 118$), from 1979 Q2 to 2007 Q3 for Norwegian data ($T = 113$), and from 1981 Q1 to 2008Q1 for New Zealand data ($T = 109$).

The evaluation periods start at 1986 Q1, 1994Q2, 1995 Q4 and 1996 Q2. Therefore, $T^s = 105, 65, 65, 65$ and there are 79, 46, 41 and 37 evaluated forecasts for the US, the UK, Norway and New Zealand, respectively. Table 1 summarizes this information.

3.2 Models and Forecasting

The model suite is composed of a set of univariate and multivariate specifications. The univariate models may be in part justified as simple “forecasting devices” as in Clements and Hendry (2006). These simple models also present different assumptions about the orders of integration of the price level series. They can be quite serious forecasting devices as pointed out by e.g. Castle and Hendry (2007). We also use two Philips curve-type models which link inflation to output growth or the unemployment rate, respectively. The multivariate models are different VARs and VARMA that contain variables usually considered in the literature on forecasting inflation. UK and Norwegian data display seasonality, while US and New Zealand data do not. Therefore, we construct two model suites taking this difference into account. Each suite consists of 12 models: six univariate models and six multivariate models. The complete list of models is given in Table 2.

Some comments are in order here. First, we do not claim that the model suites are optimal in any sense. We do claim, however, that they represent a collection of reasonable models that might be used in a real world application. Second, a glance at Table 2 reveals that some choices such as the lag lengths or the size of the estimation windows are quite ad hoc. However, the focus is not on finding the best possible specification for each individual model. The question we ask is: How do different forms of density combinations perform *given* a set of realistic models? Third, using a similar suite of models for different data sets is disputable. On the one hand, it might be more realistic to work out a specific model suite for every country. On the other hand, this strategy lessens to some extent the dependence of the results on a particular model suite and makes the findings more comparable across data sets.

All models are estimated using a moving window of the last m observations. All our

models are linear and estimated by least squares regressions. Exceptions are the univariate MA models and the VARMA models. The MA models are estimated using the Time Series 4.0 package in GAUSS. The parameter estimates of the first iteration of a Gauss-Newton algorithm for maximizing the (conditional) likelihood are used in the VARMA case (Hannan and Kavalieris, 1984). We apply iterative forecasting, and we compute density forecasts using a normal approximation (Lütkepohl, 2005). Since all models are linear, we can express the inflation series π_t as a function of past errors and an initial value as

$$\pi_t = \sum_{i=0}^{t-1} \phi_i \varepsilon_{t-i} + b\pi_0, \quad \varepsilon_t \sim \text{i.i.d.} N(0, \sigma^2).$$

Assuming that the past errors and coefficients are known, the conditional expectation corresponds to the point forecast $\pi_{t+h,t} = \sum_{i=h}^{t-1} \phi_i \varepsilon_{t+h-i} + b\pi_0$, and the forecast error is

$$\pi_{t+h} - \pi_{t+h,t} = \sum_{i=0}^{t-1} \phi_i \varepsilon_{t+h-i} - \sum_{i=h}^{t-1} \phi_i \varepsilon_{t+h-i} = \sum_{i=0}^{h-1} \phi_i \varepsilon_{t+h-i}.$$

It follows that the forecast error variance is given by

$$\begin{aligned} \sigma^2(h) &= E[(x_{t+h} - x_{t+h,t})^2], \\ &= E \left[\left(\sum_{i=0}^{h-1} \phi_i \varepsilon_{t+h-i} \right)^2 \right] = \sigma^2 \sum_{i=0}^{h-1} \phi_i^2. \end{aligned}$$

The predictive density given by any of the models in the suite is therefore normally distributed with mean given by the usual point forecast and variance given by the above expression, $N(\pi_{t+h,t}, \sigma^2(h))$.

4 Results

The results of the out-of-sample evaluation are summarized in Figures 2 - 3 and in Tables 4 - 7. We focus here on one, four and eight-step-ahead density forecasts. Out-of-sample forecasting performance is measured both in terms of the average log score, lnS, and RMSPE. We focus mainly on the lnS as a measure of density forecast accuracy and use the RMSPE for comparison. Tables 4 - 7 tabulate the out-of-sample forecasting performance of the individual models, the combination schemes and the selection strategy for each of the four countries. The explanation of the acronyms for the individual models are given in Table 2.

We apply the test of equal accuracy of two density forecasts developed by Mitchell and Hall (2005)⁵. Suppose there are two density forecasts, $f_{t+h,t,1}(y_{t+h})$ and $f_{t+h,t,2}(y_{t+h})$, and

⁵See Mitchell and Hall (2005) page 1005 for further details. Amisano and Giacomini (2007) propose an alternative approach.

consider the loss differential

$$d_{t+h} = \ln(f_{t+h,t,1}(y_{t+h})) - \ln(f_{t+h,t,2}(y_{t+h})).$$

The null hypothesis of equal accuracy is then:

$$H_0 : E(d_{t+h}) = 0$$

The sample mean $\bar{d}_{t+h} = \frac{1}{T-h-T^s+1} \sum_{t=T^s}^{T-h} d_{t+h}$ has, under appropriate assumptions, the limiting distribution:

$$\sqrt{T}(\bar{d}_{t+h} - d_{t+h}) \rightarrow N(0, \Omega).$$

We compare the averaging strategies to selection. Tables show that differences are often statistically significant.

The results for the individual models show that there is a close relation between a model's average lnS and its RMSPE. Models with the highest lnS often have the lowest or one of the lowest RMSPE. The relationship is, however, not one-to-one. For example, for all horizons in the case of Norway and for $h = 4$ and $h = 8$ in the case of New Zealand, there are some models that provide good point forecasts but yield poor density forecasts. As expected, the forecasting performance of the individual models varies considerably over data sets. As in Stock and Watson (2007), the IMA1D1 model performs very well in terms of RMSPE and lnS for the US inflation series. However, the same model (in fourth differences) performs poorly for other data sets such as for the UK data. Some of the multivariate models generate good predictive densities. The VAR2D1_pi is the best model among the VARs in the case of the US, the VAR2D4_pi in the case of Norway and the SVAR2D1_py in the case of the UK. In the case of New Zealand, the evidence over horizons is not in favor of a particular VAR.

Selecting the best model at the forecast origin is generally difficult and can lead to quite inaccurate density forecasts. This approach, SELEC, never provides the best statistics, and for the UK and New Zealand there are several models and combination schemes that perform better. Results are qualitatively similar in terms of point forecast accuracy. Only in the case of the US is SELEC comparable to the best model, even if the difference in relation to the IMA1D1 is substantial. For $h = 1$ with Norwegian data, SELEC yields as accurate forecasts as the best individual model, RWD4. The explanation is of course that the RWD4 model provides very good forecasts *over time* such that it is always selected. Excluding this case, our results suggest that it is quite difficult to select the best individual model at the forecast origin. However, note that the forecasts obtained by combining are more accurate than selection even in this situation.

The results for the combination schemes are given in the lower part of Tables 4 - 7. Combined forecasts dominate the individual models' forecasts in several but not all cases.

However, almost all six combination methods provide higher lnSs than the selection strategy. Therefore, combining is a “safe” approach to minimize density forecast errors and seems preferable to selecting a model - even when the weights are estimated.

There are important differences between the four data sets. For US data, the IMA1D1 provides more accurate statistics than any of the combination schemes. Also, the evidence in favor of combining is weaker when we measure forecast accuracy in terms of RMSPE. Combination schemes give the lowest RMSPE only in five cases and often only marginally. This means that there is scope for combining density forecasts even in cases in which combinations do not deliver superior point forecasts.

We do not find clear support in favor of one combination method over the other. Logarithmic combination provides high lnS with UK data, but the evidence is mixed for other countries. Importantly, logarithmic combinations never yield much worse forecasts than linear combinations. Therefore, the logarithmic combination method is superior to the linear combination method if one prefers to produce symmetric density forecasts. For the different weighting schemes, our findings are generally mixed. As in Jore et al. (2007), recursive log score weights give marginally more accurate forecasts than the other weighting schemes for the US. For the other countries, RLSW weights yield substantially worse forecasts than alternative schemes. Moreover, combinations with RLSW weights occasionally provide forecasts which are less accurate than the ones obtained by SELEC. While RLSW weights are explicitly based on past density forecast accuracy, the estimation of RLSW weights is apparently rather difficult in small samples. EW and MSEW weights provide more uniform results over the different data sets and are in this sense more reliable weighting schemes. Among all the six combination methods, logarithmic combination with equal weights is one of the combinations that yielded very good results throughout data sets.

The tables give statistics over the full evaluation period but it is also interesting to investigate how different methods perform over time. In Figures 2 and 3, we compare the log scores of the logarithmic combination method with equal weights and selection over time. We choose the logarithmic combination with equal weights since this combination performs generally very well. As a performance measure over time we use cumulative log scores. That is, we compute

$$\text{ClnS}_t = \sum_{s=T^s}^t \ln f_{s+h,s,C}(\pi_{s+h}) - \ln f_{s+h,s,S}(\pi_{s+h}),$$

for $t = T^s, \dots, T - 8 + h$ and $f_{s+h,s,C}$ and $f_{s+h,s,S}$ are the density forecasts obtained from logarithmic combining and selection, respectively. Thus, ClnS_t increases when $f_{t+h,t,C}$ turns out to be more accurate than $f_{t+h,t,S}$. Ideally, we would like to see that ClnS_t increases steadily over time. Apart from the the US data set, the graphs show that this is roughly the case. For the US, the pattern is less clear even though combining is still superior on average.

This is because IMA1D1 and AR1D4 perform much better than all other models in the suite. Selecting the best (or the second best) model is therefore easier and as Hall and Mitchell (2007) and Geweke and Amisano (2008) discuss, the recursive log score measure selects the *dominant* models. When we exclude these two models from the suite, the plot is similar to the other ones (Figure 3). This explains also why recursive log score weights perform so well for the US. Assigning higher weights to dominant models is simpler in this case and improves forecast accuracy.

4.1 Robustness analysis

The presented results might still depend on the predetermined collection of individual forecasting models. Therefore, we investigate the out-of-sample performance of the combination schemes and the selection strategy using different model suites. The models were chosen based on the results in Tables 4 to 7. That is, they were essentially chosen *ex-post*. Table 3 describes the new model suites for the four robustness exercises undertaken in this section.

In the first exercise, the model suites are limited to the six best performing models. This exercise should give an advantage to the SELEC strategy as well as to RLSW. It potentially informs us about the usefulness of trimming the model space. As can be seen in Table 8, SELEC improves as the chance to choose a poor model is lower but combining is again superior. Also, RLSW weights become more accurate but EW and MSEW still outperform them to some extent. An exception is the US for reasons outlined above. The combinations' lnSs do not always improve, indicating that there are few gains from reducing the model space even more.

The second exercise and the third exercise are similar to the first one. In both cases, the model suites contain two well performing models for each country. While we choose models which are highly correlated in terms of RMSPE in the second exercise, we choose two well performing but lowly correlated models in the third one. The second exercise should give an advantage to SELEC (and RLSW), but it is comforting to see that density combinations still perform very well and provide more accurate forecasts than SELEC (Table 9). The third exercise (Table 10) basically confirms the results obtained in section 4. Comparing Tables 9 and 10, we see that it is usually advantageous to combine good models which display low correlation. This is a well known result in the point forecasting literature and it makes sense that it apparently carries over to density forecasts.

Finally, only the best and the worst model of each collection are used in the model suite. This exercise should give a disadvantage to combination methods that are based on the EW weighting scheme. It turns out, however, that combination schemes with equal weights are still performing satisfactorily. This might be because the best and the worst model also display low correlation such that there is still scope for combination. Also, SELEC yields the best forecasts in terms of lnS in the case of the US - exactly equal to those of the IMA1D1.

For the other countries, combining is still superior to selection but less advantageous than in the other exercises.

In sum, we think that these exercises confirm broadly the results of the main out-of-sample evaluation.

5 Conclusion

This paper extends the empirical literature on combining inflation density forecasts by evaluating several aggregation schemes over four different data sets. We consider both different combination methods and weighting schemes. Linear and logarithmic combinations with equal weights, recursive log score weights and mean squared error weights are used to combine density forecasts from a set of univariate and multivariate models for US, UK, Norwegian and New Zealand inflation. Results are mainly evaluated in terms of average log score.

Combinations always provide relatively accurate forecasts and, as we show, provide insurance against selecting an inappropriate model. We find that the combination schemes do not always beat the best individual models but always outperform a strategy which selects an individual model at the forecast origin based on past forecasting performance. We do not find strong evidence in favor of one combination method over the other. Equal weights and mean squared error weights were generally superior to recursive log score weights. Only in the case in which there were a small number of outstanding models in the suite did recursive log score weights yield competitive forecasts. Thus, the success of this weighting scheme crucially depends on the degree of “model uncertainty” in the overall suite of models.

Our study ignores some interesting issues which might be explored in the future. First, all models in the suite are linear. As the focus shifts from the usual MSE framework to density forecasts, there is a potential for mixtures of linear and non-linear models. Second, we combine and evaluate density forecasts for each horizon separately. A promising line of research might be the joint evaluation of sequences of forecasts or “forecasting paths”. Last but not least, we only evaluate a limited number of functional forms and weighting schemes. The development and evaluation of other density aggregation schemes is another interesting topic for future research.

References

- Amisano, G. and Giacomini, R. (2007), ‘Comparing density forecasts via weighted likelihood ratio tests.’, *Journal of Business & Economic Statistics* **25**(2), 177–190.
- Andersson, M. and Karlsson, S. (2007), Bayesian forecast combination for VAR models. Unpublished manuscript, Sveriges Riksbank.
- Bai, J. (2003), ‘Testing parametric conditional distributions of dynamic models.’, *Review of Economics and Statistics* **85**, 531–549.
- Bates, J. M. and Granger, C. W. J. (1969), ‘Combination of forecasts’, *Operational Research Quarterly* **20**(4), 451–468.
- Berkowitz, J. (2001), ‘Testing density forecasts, with applications to risk management’, *Journal of Business & Economic Statistics* **19**(4), 465–74.
- Bierens, H. (1982), ‘Consistent model-specification tests.’, *Journal of Econometrics* **20**, 105–134.
- Bierens, H. J. and Ploberger, W. (1997), ‘Asymptotic theory of integrated conditional moments tests.’, *Econometrica* **65**, 1129–1151.
- Castle, J. L. and Hendry, D. F. (2007), Forecasting uk inflation: The roles of structural breaks and time disaggregation. University of Oxford, Department of Economics, Discussion Paper.
- Clark, T. E. and McCracken, M. W. (2007), Averaging forecasts from VARs with uncertain instabilities. Revision of Federal Reserve Bank of Kansas City Working Paper 06-12.
- Clemen, R. T., Murphy, A. H. and Winkler, R. L. (1995), ‘Screening probability forecasts: Contrasts between choosing and combining.’, *International Journal of Forecasting* **11**, 133–145.
- Clements, M. P. (2004), ‘Evaluating the bank of england density forecasts of inflation.’, *Economic Journal* **114**, 844–866.
- Clements, M. P. (2006), ‘Evaluating the survey of professional forecasters probability distributions of expected inflation based on derived event probability forecasts.’, *Empirical Economics* **31**, 49–64.
- Clements, M. P. and Hendry, D. F. (2006), *Handbook of Economic Forecasting*, Elsevier, chapter 9.
- Corradi, V. and Swanson, N. (2003), ‘Bootstrap conditional distribution tests in the presence of dynamic misspecification.’, *Journal of Econometrics* **133**(2), 779–806.
- Corradi, V. and Swanson, N. (2004), A test for comparing multiple misspecified conditional distributions. Working Paper, Rutgers University.
- Corradi, V. and Swanson, N. (2006a), *Handbook of Economic Forecasting*, Elsevier, chapter 2.
- Corradi, V. and Swanson, N. (2006b), ‘Predictive density and conditional confidence interval accuracy tests.’, *Journal of Econometrics* **135**, 187–228. 1-2.

- Diebold, F. X., Gunther, T. and Tay, A. S. (1998), ‘Evaluating density forecasts with applications to finance and management.’, *International Economic Review* **39**, 863–883.
- Fernández-Villaverde, J. and Rubio-Ramírez, J. F. (2004), ‘Comparing dynamic equilibrium models to data’, *Journal of Econometrics* **123**, 153–187.
- Garratt, A., Lee, K., Pesaran, M. H. and Shin, Y. (2003), ‘Forecast uncertainties in macroeconomic modelling: An application to the UK economy.’, *Journal of the American Statistical Association* **98**, 829–838.
- Genest, C. (1983), Towards a consensus of opinion, PhD thesis, University of British Columbia.
- Genest, C. and Zidek, J. (1986), ‘Combining probability distributions: A critique and an annotated bibliography.’, *Statistical Science* **1**, 114–148.
- Geweke, J. and Amisano, G. (2008), Optimal prediction pools, Working Paper Series 22-08, Rimini Centre for Economic Analysis.
- Geweke, J. and Whiteman, C. (2006), Bayesian forecasting, in G. Elliot, C. Granger and A. Timmermann, eds, ‘Handbook of Economic Forecasting’, North-Holland.
- Granger, C. W. J. and Jeon, Y. (2004), ‘Thick modeling.’, *Economic Modelling* **21**, 323–343.
- Granger, C. W. J. and Pesaran, M. H. (2000), ‘Economic and statistical measures of forecast accuracy.’, *Journal of Forecasting* **19**, 537–560.
- Granger, C. W. J. and Ramanathan, R. (1984), ‘Improved methods of combining forecasts’, *Journal of Forecasting* **3**, 197–204.
- Granger, C. W. J., White, H. and Kamstra, M. (1989), ‘Interval forecasting: An analysis based upon ARCH-quantile estimators.’, *Journal of Econometrics* **40**, 87–96.
- Hall, S. G. and Mitchell, J. (2004), Density forecast combination. National institute of economic and social research discussion paper, No. 249.
- Hall, S. G. and Mitchell, J. (2007), ‘Combining density forecasts.’, *International Journal of Forecasting* **23**, 1–13.
- Hannan, E. J. and Kavalieris, L. (1984), ‘Multivariate linear time series models’, *Advances in Applied Probability* **16**(3), 492–561.
- Hendry, D. F. and Clements, M. P. (2004), ‘Pooling of forecasts.’, *Econometrics Journal* **7**, 1–31.
- Jackson, T. and Karlsson, S. (2004), ‘Finding good predictors for inflation: A bayesian model averaging approach.’, *Journal of Forecasting* **23**, 479–498.
- Jore, A. S., Mitchell, J. and Vahey, S. P. (2007), Combining forecast densities from VARs with uncertain instabilities. Norges Bank, NIESR and RBNZ, working paper.
- Kitamura, Y. (2002), Econometric comparisons of conditional models. Working Paper, University of Pennsylvania.

- Leamer, E. (1978), *Specification Searches.*, Wiley, Oxford.
- Li, F. and Tkacz, G. (2006), ‘A consistent bootstrap test for conditional density functions with time-dependent data.’, *Journal of Econometrics* **127**, 863–886.
- Lütkepohl, H. (2005), *New Introduction to Multiple Time Series Analysis*, Springer-Verlag, Berlin.
- Min, C. K. and Zellner, A. (1993), ‘Bayesian and non-bayesian methods for combining models and forecasts with applications to forecasting international growth rates.’, *Journal of Econometrics* **56**, 89–118.
- Mitchell, J. and Hall, S. G. (2005), ‘Evaluating, comparing and combining density forecasts using the klic with an application to the bank of england and nieser “fan” charts of inflation.’, *Oxford Bulletin of Economics and Statistics* **67**, 995–1033.
- Morris, P. (1974), ‘Decision analysis expert use.’, *Management Science* **20**, 1233–1241.
- Morris, P. (1977), ‘Combining expert judgments: A bayesian approach.’, *Management Science* **23**, 679–693.
- Palm, F. C. and Zellner, A. (1992), ‘To combine or not to combine?’, *Journal of Forecasting* **11**, 687–701.
- Raftery, A. E., Madigan, D. and Hoeting, J. A. (1997), ‘Bayesian model averaging for linear regression models.’, *Journal of the American Statistical Association* **92**, 179–191.
- Robertson, J. C., Tallman, E. W. and Whiteman, C. H. (2005), ‘Forecasting using relative entropy.’, *Journal of Money, Credit and Banking*. **37**, 383–401.
- Stock, J. H. and Watson, M. W. (2007), ‘Why has u.s. inflation become harder to forecast?’, *Journal of Money, Credit and Banking* **39**(s1), 3–33.
- Stone, M. (1961), ‘The opinion pool’, *Annals of Mathematical Statistics* **32**, 1339–1342.
- Tay, A. S. and Wallis, K. F. (2000), ‘Density forecasting: A survey.’, *Journal of Forecasting* **19**, 235–254.
- Timmermann, A. (2006), *Forecast Combinations*, Elsevier, chapter 4.
- Vuong, Q. (1989), ‘Likelihood ratio tests for model selection and non-nested hypotheses’, *Econometrica* **57**, 307–333.
- Wallis, K. F. (2005), ‘Combining density and interval forecasts: A modest proposal.’, *Oxford Bulletin of Economics and Statistics* **67**, 983–994.
- Winkler, R. (1981), ‘Combining probability distributions from dependent information sources.’, *Management Science* **27**, 479–488.
- Winkler, R. L. (1968), ‘The consensus of subjective probability distributions’, *Management Science* **15**, B61–B75.

A Logarithmic Combination of Normal Densities

Consider a collection of normal densities, $f_i(y)$, $i = 1, \dots, N$. Their logarithmic combination is given by

$$f_{log}(y) = \frac{\prod_{i=1}^N f_i^{\omega_i}(y)}{\int \prod_{i=1}^N f_i^{\omega_i}(y) dy}, \quad (\text{A-1})$$

where ω_i is the combination weight for the density $f_i(y)$. Denote the means and variances of the individual forecasts by μ_i and σ_i^2 , $i = 1, \dots, N$, respectively. In this special case, it follows that

$$\prod_{i=1}^N f_i^{\omega_i}(y) = c \prod_{i=1}^N \left(\exp \left[-\frac{1}{2} \frac{(y - \mu_i)^2}{\sigma_i^2} \right] \right)^{\omega_i}, \quad (\text{A-2})$$

where $c = (2\pi)^{-\frac{1}{2}} \prod_{i=1}^N \sigma_i^{-\omega_i}$. Define $\alpha_i = \frac{\omega_i}{\sigma_i^2}$ and write

$$\begin{aligned} \prod_{i=1}^N f_i^{\omega_i}(y) &= c \prod_{i=1}^N \exp \left[-\frac{\omega_i}{2\sigma_i^2} (y - \mu_i)^2 \right] = c \prod_{i=1}^N \exp \left[-\frac{\alpha_i}{2} (y - \mu_i)^2 \right], \\ &= c \exp \left[-\frac{1}{2} \sum_{i=1}^N \alpha_i (y^2 + \mu_i^2 - 2y\mu_i) \right], \\ &= c \exp \left[-\frac{1}{2} \left(y^2 \sum_{i=1}^N \alpha_i + \sum_{i=1}^N \alpha_i \mu_i^2 - 2y \sum_{i=1}^N \alpha_i \mu_i \right) \right]. \end{aligned}$$

The term between square brackets can be decomposed as

$$\begin{aligned} \left(y^2 \sum_{i=1}^N \alpha_i + \sum_{i=1}^N \alpha_i \mu_i^2 - 2y \sum_{i=1}^N \alpha_i \mu_i \right) &= \sum_{i=1}^N \alpha_i \left(y^2 + \frac{\sum_{i=1}^N \alpha_i \mu_i^2}{\sum_{i=1}^N \alpha_i} - 2y \frac{\sum_{i=1}^N \alpha_i \mu_i}{\sum_{i=1}^N \alpha_i} \right), \\ &= \sum_{i=1}^N \alpha_i \left(\left(y - \frac{\sum_{i=1}^N \alpha_i \mu_i}{\sum_{i=1}^N \alpha_i} \right)^2 + \frac{\sum_{i=1}^N \alpha_i \mu_i^2}{\sum_{i=1}^N \alpha_i} - \frac{(\sum_{i=1}^N \alpha_i \mu_i)^2}{(\sum_{i=1}^N \alpha_i)^2} \right), \\ &= \sum_{i=1}^N \alpha_i (y - \mu_c)^2 + d, \end{aligned}$$

where $\mu_c = \sum_{i=1}^N \alpha_i \mu_i / \sum_{i=1}^N \alpha_i$ and $d = \sum_{i=1}^N \alpha_i \mu_i^2 - (\sum_{i=1}^N \alpha_i \mu_i)^2 / \sum_{i=1}^N \alpha_i$. We can express the weighted product of the individual densities compactly as

$$\begin{aligned} \prod_{i=1}^N f_i^{\omega_i}(y) &= c \exp \left[-\frac{1}{2} \left(\sum_{i=1}^N \alpha_i (y - \mu_c)^2 + d \right) \right], \\ &= c \exp \left[-\frac{1}{2} \left(\sum_{i=1}^N \alpha_i \right) (y - \mu_c)^2 \right] \exp[-1/2d], \\ &= c \exp[-1/2d] \exp \left[-\frac{1}{2} \frac{1}{\sigma_c^2} (y - \mu_c)^2 \right]. \end{aligned}$$

Since c and $\exp[-1/2d]$ do not depend on y , we have

$$\int \prod_{i=1}^N f_i^{\omega_i}(y) dy = c \exp[-1/2d] \int \exp \left[-\frac{1}{2} \frac{1}{\sigma_c^2} (y - \mu_c)^2 \right] dy = c \exp[-1/2d] (2\pi)^{1/2} \sigma_c.$$

Thus, the logarithmic combination in (A-1) is given by

$$\begin{aligned} \frac{\prod_{i=1}^N f_i^{\omega_i}(y)}{\int \prod_{i=1}^N f_i^{\omega_i}(y) dy} &= \frac{c \exp[-1/2d]}{(c \exp[-1/2d] (2\pi)^{-1/2} \sigma_c)} \exp \left[-\frac{1}{2} \frac{1}{\sigma_c^2} (y - \mu_c)^2 \right] \\ &= (2\pi)^{-1/2} \sigma_c^{-1} \exp \left[-\frac{1}{2} \frac{1}{\sigma_c^2} (y - \mu_c)^2 \right] \end{aligned}$$

which is a normal density, $N(\mu_c, \sigma_c^2)$, with mean and variance μ_c and σ_c^2 , respectively, given by

$$\begin{aligned} \mu_c &= \frac{\sum_{i=1}^N \alpha_i \mu_i}{\sum_{i=1}^N \alpha_i}, \\ \sigma_c^2 &= \left(\sum_{i=1}^N \alpha_i \right)^{-1} = \left(\sum_{i=1}^N \frac{\omega_i}{\sigma_i^2} \right)^{-1}, \\ &= \frac{\prod_{i=1}^N \sigma_i^2}{\sum_{i=1}^N \omega_i \prod_{j=1, j \neq i}^N \sigma_j^2}. \end{aligned}$$

B Data and Models

We collect US PCE from the NIPA accounts available from the Bureau of Economic Analysis, US GDP, M2 and the unemployment rate from the Federal Reserve Bank of Philadelphia's Real time Data Set for Macroeconomists, US interest rates from the Federal Reserve Economic Data (FRED database). UK CPI, interest rates, money and unemployment rate are obtained from the OECD database, and UK GDP from EUROSTAT. Norwegian data are collected from Norges Bank's database and New Zealand data from the Reserve Bank of New Zealand database.

Table 1: Data Sets

	sample	evaluation period	
US	1960 Q1 - 07 Q3	(1986 Q1+h) - (07 Q3-8+h)	(79)
UK	1978 Q1 - 07 Q2	(1994 Q1+h) - (07 Q2-8+h)	(46)
Norway	1979 Q2 - 07 Q3	(1995 Q3+h) - (07 Q3-8+h)	(41)
NZ	1981 Q1 - 08 Q1	(1996 Q2+h) - (08 Q1-8+h)	(37)

Note: The table reports the sample period, the forecasting evaluation period and between parenthesis the number of evaluated forecasts for different countries.

Table 2: Definitions of Forecasting Models

Name	Definition	Variables	m
US and New Zealand			
RWD1	$\pi_t = \pi_{t-1} + \varepsilon_t$		20
AR1D1	$\pi_t = \mu + \alpha_1 \pi_{t-1} + \varepsilon_t$		20
AR4D1	$\pi_t = \mu + \alpha_1 \pi_{t-1} + \dots + \alpha_4 \pi_{t-4} + \varepsilon_t$		40
IMA1D1	$\pi_t = \pi_{t-1} + \varepsilon_t + \theta \varepsilon_{t-1}$		40
PC-Y	$\pi_{t+h} = \mu + \alpha(L)\pi_t + \beta(L)\Delta_1 y_t + \varepsilon_{t+h}$		50
PC-U	$\pi_{t+h} = \mu + \alpha(L)\pi_t + \beta(L)\Delta_1 ur_t + \varepsilon_{t+h}$		50
VAR2D1_pm	$\Delta_1 \mathbf{x}_t = \mu + A_1 \Delta_1 \mathbf{x}_{t-1} + A_2 \mathbf{x}_{t-2} + u_t$	$\mathbf{x}_t = (p_t, M_t)'$	50
VAR2D1_pi		$\mathbf{x}_t = (p_t, i_t)'$	
VAR2D1_piy		$\mathbf{x}_t = (p_t, i_t, y_t)'$	
VAR2D1_pmy		$\mathbf{x}_t = (p_t, M_t, y_t)'$	
VARMA11D1_pm	$\Delta_1 \mathbf{x}_t = \mu + A_1 \Delta_1 \mathbf{x}_{t-1} + u_t + M_1 u_{t-1}$	$\mathbf{x}_t = (p_t, M_t)'$	50
VARMA11D1_pi		$\mathbf{x}_t = (p_t, i_t)'$	
UK and Norway			
RWD4	$\pi_t^a = \pi_{t-1}^a + \varepsilon_t$		20
AR1D4	$\pi_t^a = \mu + \alpha_1 \pi_{t-1}^a + \varepsilon_t$		20
SAR2D1	$\pi_t = \mu + s_1 d_{1t} + s_2 d_{2t} + s_3 d_{3t} + \alpha_1 \pi_{t-1} + \alpha_2 \pi_{t-2} + \varepsilon_t$		40
IMA1D4	$\pi_t^a = \pi_{t-1}^a + \varepsilon_t + \theta \varepsilon_{t-1}$		40
PC-Y	$\pi_{t+h}^a = \mu + \alpha(L)\pi_t^a + \beta(L)\Delta_4 y_t + \varepsilon_{t+h}$		40
PC-U	$\pi_{t+h}^a = \mu + \alpha(L)\pi_t^a + \beta(L)ur_t + \varepsilon_{t+h}$		40
VAR2D4_pm	$\Delta_4 \mathbf{x}_t = \mu + A_1 \Delta_4 \mathbf{x}_{t-1} + A_2 \Delta_4 \mathbf{x}_{t-2} + u_t$	$\mathbf{x}_t = (p_t, M_t)'$	50
VAR2D4_pi		$\mathbf{x}_t = (p_t, i_t)'$	
VAR2D4_piy		$\mathbf{x}_t = (p_t, i_t, y_t)'$	
VAR2D4_pmy		$\mathbf{x}_t = (p_t, M_t, y_t)'$	
SVAR4D1_py	$\Delta_1 \mathbf{x}_t = \mu + s_1 d_{1t} + s_2 d_{2t} + s_3 d_{3t}$	$\mathbf{x}_t = (p_t, y_t)'$	50
SVAR4D1_pi	$+ A_1 \Delta_1 \mathbf{x}_{t-1} + \dots + A_4 \mathbf{x}_{t-4} + u_t$	$\mathbf{x}_t = (p_t, i_t)'$	

Note: The table shows the definitions of the forecasting models used in the suite together with the moving window of observations, m , that is used for estimation. In the table, π_t and π_t^a are quarter-to-quarter and year-to-year inflation, respectively; p_t is the price level, i_t is a short-term interest rate, M_t is a money measure, ur_t is the unemployment rate, y_t is an output measure. In the PC models, $\alpha(L) = \alpha_1 + \dots + \alpha_p L^{p-1}$ and $\beta(L) = \beta_1 + \dots + \beta_q L^{q-1}$ are lag polynomials. Lag lengths are estimated using the BIC in the case of the PC models with maximal order equal to four.

Table 3: Robustness checks

US	
Six best models	AR4D1, IMA1D1, PC-U, VAR2D1_pi VAR2D1_piy, VARMA11D1_pi
Best and high corr	AR4D1, IMA1D1
Best and low corr	IMA1D1, PC-U
Best and worst	RWD1, IMA1D1_pi
UK	
Six best models	RWD4, AR1D4, IMA1D4, PC-U SVAR2D1_py, VAR2D4_pmy
Best and high corr	RWD4, AR1D4
Best and low corr	RWD4, VAR2D4_pmy
Best and worst	RWD4, VAR2D4_pm
Norway	
Six best models	RWD4, IMA1D4, VAR2D4_pm, VAR2D4_pi SVAR2D1_pi, VAR2D4_piy
Best and high corr	RWD4, VAR2D4_pm
Best and low corr	RWD4, VAR2D4_piy
Best and worst	RWD4, AR1D4
New Zealand	
Six best models	AR2D1, IMA1D1, PC-Y, PC-U VAR2D1_piy, VAR2D1_pmy
Best and high corr	PC-Y, PC-U
Best and low corr	IMA1D1, PC-Y
Best and worst	RWD1, PC-Y

Note: The table indicates the model suites used for the robustness checks in section 4.1 for each country.

C Results

Table 4: Out-of-sample prediction performance, US

	h=1		h=4		h=8	
	lnS	RMSPE	lnS	RMSPE	lnS	RMSPE
Individual Models						
RWD1	-0.005	0.231	-0.259	0.222	-0.586	0.278
AR1D1	-0.035	0.216	-0.123	0.239	-0.346	0.265
AR4D1	0.093	0.200	0.008	0.218	-0.336	0.285
IMA1D1	0.167	0.196	0.052	0.210	-0.201	0.265
PC-Y	0.022	0.218	-0.034	0.214	-0.562	0.297
PC-U	0.065	0.208	-0.025	0.211	-0.483	0.284
VAR2D1_pm	-0.015	0.223	-0.187	0.241	-0.392	0.281
VAR2D1_pi	0.057	0.212	-0.165	0.238	-0.391	0.278
VAR2D1_piy	0.033	0.217	-0.178	0.249	-0.388	0.283
VAR2D1_pmy	-0.051	0.230	-0.218	0.259	-0.409	0.295
VARMA11D1_pm	-0.000	0.220	-0.266	0.270	-0.466	0.328
VARMA11D1_pi	0.035	0.210	-0.159	0.245	-0.366	0.299
Selection						
SELEC	0.097	0.199	-0.009	0.226	-0.415	0.286
Linear Pooling						
EW	0.122	0.203	-0.042	0.218	-0.280	0.265
RLSW	0.155*	0.195	0.018	0.222	-0.295*	0.280
MSEW	0.129	0.202	-0.019	0.213	-0.247	0.258
Log. Pooling						
EW	0.114	0.203	-0.014	0.218	-0.254	0.263
RLSW	0.141*	0.195	0.014	0.221	-0.317**	0.277
MSEW	0.120	0.202	0.028	0.216	-0.187**	0.259

Note: In the table, lnS denotes the average log score evaluated out-of-sample, RMSPE denotes the root mean squared prediction error. For the Mitchell and Hall (2005) test, * and ** indicate rejection at the 90% and 95% significance level, respectively, of the null hypothesis of equal forecasting accuracy of the selection strategy and a particular combination method. See Table 2 for definitions of the individual models.

Table 5: Out-of-sample prediction performance, UK

	h=1		h=4		h=8	
	lnS	RMSPE	lnS	RMSPE	lnS	RMSPE
Individual Models						
RWD4	-0.347	0.298	-0.357	0.295	-0.539	0.340
AR1D4	-0.464	0.325	-0.361	0.302	-0.486	0.343
SAR2D1	-0.534	0.405	-0.368	0.333	-0.535	0.409
IMA1D4	-0.468	0.314	-0.409	0.295	-0.694	0.340
PC-Y	-0.451	0.314	-0.473	0.441	-0.785	0.607
PC-U	-0.496	0.397	-0.340	0.325	-0.698	0.545
VAR2D4_pm	-0.542	0.402	-0.487	0.319	-0.764	0.414
VAR2D4_pi	-0.459	0.342	-0.483	0.329	-0.778	0.447
SVAR2D1_py	-0.353	0.351	-0.376	0.358	-0.524	0.420
SVAR2D1_pi	-0.408	0.383	-0.392	0.399	-0.660	0.528
VAR2D4_piy	-0.450	0.395	-0.522	0.409	-0.686	0.504
VAR2D4_pmy	-0.499	0.423	-0.532	0.420	-0.695	0.511
Selection						
SELEC	-0.483	0.419	-0.395	0.331	-0.701	0.429
Linear Pooling						
EW	-0.372	0.307	-0.363	0.310	-0.577	0.410
RLSW	-0.430	0.389	-0.351	0.319	-0.578	0.404
MSEW	-0.370	0.320	-0.347	0.298	-0.574	0.389
Log. Pooling						
EW	-0.297*	0.308	-0.295*	0.313	-0.502*	0.428
RLSW	-0.419	0.389	-0.349	0.321	-0.577*	0.435
MSEW	-0.303**	0.323	-0.287*	0.300	-0.489*	0.409

Note: See footnote in Table 4.

Table 6: Out-of-sample prediction performance, Norway

	h=1		h=4		h=8	
	lnS	RMSPE	lnS	RMSPE	lnS	RMSPE
Individual Models						
RWD4	-0.165	0.268	-0.163	0.269	-0.347	0.331
AR1D4	-0.464	0.304	-0.472	0.321	-0.720	0.448
SAR2D1	-0.435	0.266	-0.346	0.272	-0.578	0.304
IMA1D4	-0.216	0.263	-0.190	0.269	-0.381	0.331
PC-Y	-0.315	0.268	-0.245	0.269	-0.580	0.349
PC-U	-0.269	0.258	-0.211	0.267	-0.572	0.328
VAR2D4_pm	-0.243	0.256	-0.227	0.259	-0.354	0.301
VAR2D4_pi	-0.189	0.250	-0.193	0.251	-0.361	0.303
SVAR2D1_py	-0.242	0.248	-0.301	0.246	-0.509	0.300
SVAR2D1_pi	-0.213	0.248	-0.282	0.251	-0.488	0.305
VAR2D4_piy	-0.182	0.259	-0.391	0.273	-0.589	0.307
VAR2D4_pmy	-0.235	0.270	-0.362	0.269	-0.581	0.307
Selection						
SELEC	-0.165	0.268	-0.216	0.271	-0.767	0.324
Linear Pooling						
EW	-0.019	0.230	-0.126	0.250	-0.323	0.293
RLSW	-0.175	0.264	-0.182*	0.272	-0.435	0.319
MSEW	-0.027	0.235	-0.124	0.252	-0.321	0.297
Log. Pooling						
EW	-0.023	0.230	-0.127*	0.255	-0.296*	0.300
RLSW	-0.191	0.261	-0.256*	0.274	-0.635**	0.317
MSEW	-0.048	0.236	-0.130*	0.256	-0.308*	0.303

Note: See footnote in Table 4.

Table 7: Out-of-sample prediction performance, New Zealand

	h=1		h=4		h=8	
	lnS	RMSPE	lnS	RMSPE	lnS	RMSPE
Individual Models						
RWD1	-0.782	0.458	-0.894	0.480	-1.131	0.496
AR1D1	-0.732	0.382	-0.931	0.387	-0.775	0.385
AR2D1	-0.835	0.404	-0.738	0.410	-0.558	0.393
IMA1D1	-0.731	0.384	-0.719	0.420	-0.584	0.408
PC-Y	-0.612	0.385	-0.641	0.410	-0.634	0.410
PC-U	-0.657	0.409	-0.725	0.475	-0.685	0.413
VAR2D1_pm	-0.573	0.391	-0.710	0.447	-0.758	0.450
VAR2D1_pi	-0.623	0.420	-0.705	0.423	-0.734	0.414
VAR2D1_piy	-0.644	0.479	-0.688	0.392	-0.732	0.410
VAR2D1_pmy	-0.546	0.378	-0.694	0.399	-0.751	0.426
VARMA11D1_pm	-0.664	0.421	-0.781	0.474	-0.822	0.459
VARMA11D1_pi	-0.737	0.438	-0.831	0.558	-0.847	0.528
Selection						
SELEC	-0.768	0.404	-1.134	0.433	-1.052	0.428
Linear Pooling						
EW	-0.531	0.379	-0.598	0.411	-0.593*	0.398
RLSW	-0.732	0.401	-1.098	0.436	-0.959**	0.422
MSEW	-0.480	0.379	-0.552**	0.395	-0.556**	0.394
Log. Pooling						
EW	-0.490	0.378	-0.549*	0.391	-0.526**	0.392
RLSW	-0.755	0.398	-1.113	0.422	-0.917**	0.400
MSEW	-0.508*	0.381	-0.586**	0.395	-0.543**	0.392

Note: See footnote in Table 4.

Table 8: Out-of-sample prediction performance, best six models

	h=1		h=4		h=8	
	lnS	RMSPE	lnS	RMSPE	lnS	RMSPE
US						
SELEC	0.127	0.196	0.032	0.208	-0.300	0.280
Linear Pooling						
EW	0.146	0.198	-0.015	0.217	-0.270	0.264
RLSW	0.159	0.194	0.049	0.211	-0.243	0.275
MSEW	0.150	0.197	0.011	0.215	-0.235	0.263
Log. Pooling						
EW	0.139	0.197	-0.010	0.215	-0.273	0.261
RLSW	0.148	0.193	0.048	0.211	-0.248	0.271
MSEW	0.141	0.197	0.015	0.213	-0.237	0.260
UK						
SELEC	-0.421	0.393	-0.395	0.331	-0.615	0.435
Linear Pooling						
EW	-0.359	0.301	-0.338	0.297	-0.551	0.370
RLSW	-0.422	0.389	-0.350	0.319	-0.557	0.380
MSEW	-0.361	0.312	-0.326	0.292	-0.547	0.352
Log. Pooling						
EW	-0.300	0.302	-0.289	0.300	-0.482	0.403
RLSW	-0.416	0.389	-0.349	0.321	-0.523	0.417
MSEW	-0.303	0.313	-0.287	0.296	-0.476	0.386
Norway						
SELEC	-0.165	0.268	-0.163	0.269	-0.449	0.332
Linear Pooling						
EW	-0.045	0.234	-0.139	0.251	-0.316	0.301
RLSW	-0.165	0.266	-0.161	0.269	-0.336	0.322
MSEW	-0.059	0.241	-0.135	0.252	-0.314	0.304
Log. Pooling						
EW	-0.056	0.234	-0.134	0.253	-0.299	0.302
RLSW	-0.160	0.265	-0.169	0.269	-0.363	0.315
MSEW	-0.083	0.242	-0.131	0.254	-0.300	0.305
New Zealand						
SELEC	-0.816	0.396	-0.764	0.405	-0.562	0.376
Linear Pooling						
EW	-0.546	0.377	-0.628	0.404	-0.583	0.397
RLSW	-0.620	0.377	-0.738	0.410	-0.561	0.393
MSEW	-0.518	0.376	-0.619	0.399	-0.587	0.399
Log. Pooling						
EW	-0.588	0.379	-0.626	0.401	-0.557	0.398
RLSW	-0.690	0.382	-0.738	0.403	-0.549	0.393
MSEW	-0.628	0.382	-0.651	0.400	-0.572	0.399

Note: The table reports lnS and RMSPE using only the best six models in the suite.

Table 9: Out-of-sample prediction performance, two good, highly correlated models

	h=1		h=4		h=8	
	lnS	RMSPE	lnS	RMSPE	lnS	RMSPE
US						
SELEC	0.127	0.196	0.032	0.208	-0.300	0.280
Linear Pooling						
EW	0.175	0.194	0.062	0.211	-0.214	0.271
RLSW	0.162	0.194	0.049	0.212	-0.243	0.276
MSEW	0.174	0.194	0.060	0.211	-0.211	0.268
Log. Pooling						
EW	0.157	0.194	0.059	0.211	-0.227	0.270
RLSW	0.150	0.194	0.048	0.211	-0.249	0.272
MSEW	0.158	0.194	0.060	0.209	-0.215	0.265
UK						
SELEC	-0.347	0.298	-0.406	0.301	-0.537	0.344
Linear Pooling						
EW	-0.363	0.306	-0.350	0.298	-0.508	0.339
RLSW	-0.349	0.298	-0.364	0.298	-0.510	0.340
MSEW	-0.361	0.305	-0.351	0.297	-0.512	0.339
Log. Pooling						
EW	-0.377	0.308	-0.352	0.297	-0.504	0.339
RLSW	-0.344	0.298	-0.366	0.298	-0.506	0.339
MSEW	-0.372	0.307	-0.352	0.297	-0.507	0.338
Norway						
SELEC	-0.165	0.268	-0.163	0.269	-0.347	0.331
Linear Pooling						
EW	-0.124	0.256	-0.125	0.262	-0.304	0.313
RLSW	-0.165	0.268	-0.161	0.269	-0.327	0.330
MSEW	-0.127	0.256	-0.119	0.263	-0.303	0.314
Log. Pooling						
EW	-0.136	0.261	-0.126	0.264	-0.283	0.313
RLSW	-0.166	0.268	-0.162	0.269	-0.340	0.331
MSEW	-0.140	0.261	-0.129	0.265	-0.288	0.314
New Zealand						
SELEC	-0.619	0.386	-0.642	0.417	-0.692	0.414
Linear Pooling						
EW	-0.624	0.379	-0.663	0.436	-0.644	0.411
RLSW	-0.624	0.377	-0.653	0.431	-0.648	0.411
MSEW	-0.626	0.381	-0.664	0.437	-0.646	0.411
Log. Pooling						
EW	-0.626	0.377	-0.669	0.436	-0.657	0.411
RLSW	-0.626	0.377	-0.657	0.432	-0.661	0.411
MSEW	-0.628	0.379	-0.669	0.438	-0.658	0.411

Note: The table reports lnS and RMSPE using two good, highly correlated models.

Table 10: Out-of-sample prediction performance, two good, lowly correlated models

	h=1		h=4		h=8	
	lnS	RMSPE	lnS	RMSPE	lnS	RMSPE
US						
SELEC	0.167	0.196	0.052	0.210	-0.201	0.265
Linear Pooling						
EW	0.156	0.195	0.051	0.204	-0.277	0.260
RLSW	0.167	0.195	0.052	0.210	-0.201	0.265
MSEW	0.161	0.194	0.052	0.209	-0.217	0.264
Log. Pooling						
EW	0.158	0.195	0.042	0.206	-0.287	0.267
RLSW	0.168	0.195	0.052	0.210	-0.201	0.265
MSEW	0.164	0.195	0.051	0.209	-0.214	0.266
UK						
SELEC	-0.499	0.423	-0.357	0.295	-0.554	0.368
Linear Pooling						
EW	-0.338	0.316	-0.379	0.316	-0.570	0.379
RLSW	-0.499	0.422	-0.354	0.292	-0.563	0.375
MSEW	-0.358	0.344	-0.351	0.294	-0.545	0.350
Log. Pooling						
EW	-0.290	0.312	-0.319	0.321	-0.500	0.407
RLSW	-0.502	0.421	-0.343	0.293	-0.526	0.397
MSEW	-0.314	0.335	-0.315	0.301	-0.483	0.378
Norway						
SELEC	-0.165	0.268	-0.163	0.269	-0.347	0.331
Linear Pooling						
EW	0.017	0.232	-0.119	0.256	-0.317	0.302
RLSW	-0.165	0.267	-0.162	0.269	-0.331	0.329
MSEW	0.001	0.238	-0.112	0.258	-0.320	0.309
Log. Pooling						
EW	0.005	0.231	-0.148	0.267	-0.337	0.315
RLSW	-0.159	0.265	-0.163	0.269	-0.345	0.330
MSEW	-0.037	0.239	-0.157	0.269	-0.358	0.318
New Zealand						
SELEC	-0.771	0.388	-0.725	0.422	-0.715	0.403
Linear Pooling						
EW	-0.539	0.374	-0.631	0.418	-0.562	0.400
RLSW	-0.649	0.384	-0.726	0.419	-0.620	0.401
MSEW	-0.535	0.379	-0.674	0.414	-0.588	0.406
Log. Pooling						
EW	-0.618	0.379	-0.643	0.405	-0.561	0.400
RLSW	-0.752	0.387	-0.723	0.416	-0.620	0.400
MSEW	-0.671	0.383	-0.696	0.413	-0.582	0.405

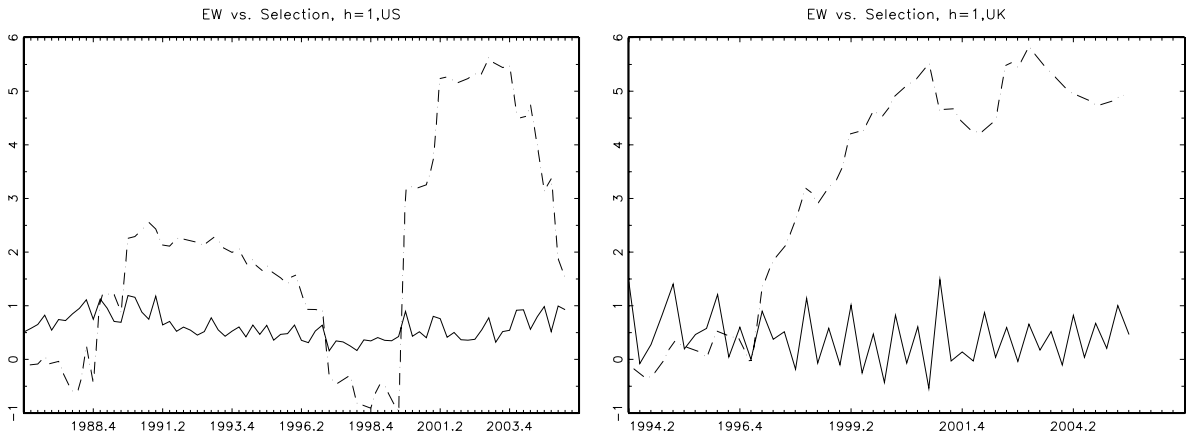
Note: The table reports lnS and RMSPE using two good models with low correlation.

Table 11: Out-of-sample prediction performance, best and worst model

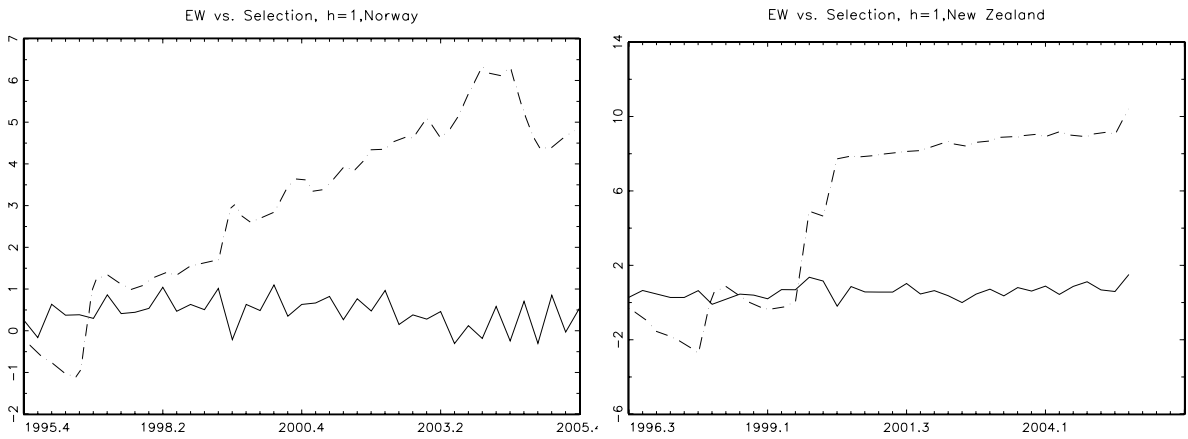
	h=1		h=4		h=8	
	lnS	RMSPE	lnS	RMSPE	lnS	RMSPE
US						
SELEC	0.167	0.196	0.052	0.210	-0.201	0.265
Linear Pooling						
EW	0.125	0.207	-0.061	0.209	-0.330	0.266
RLSW	0.165	0.197	0.047	0.210	-0.208	0.265
MSEW	0.135	0.204	-0.057	0.208	-0.322	0.265
Log. Pooling						
EW	0.128	0.205	0.005	0.207	-0.237	0.264
RLSW	0.165	0.197	0.048	0.210	-0.206	0.265
MSEW	0.138	0.203	0.008	0.207	-0.233	0.264
UK						
SELEC	-0.491	0.393	-0.392	0.313	-0.641	0.384
Linear Pooling						
EW	-0.411	0.326	-0.388	0.299	-0.631	0.368
RLSW	-0.478	0.391	-0.385	0.301	-0.598	0.372
MSEW	-0.404	0.325	-0.380	0.297	-0.610	0.369
Log. Pooling						
EW	-0.368	0.327	-0.357	0.298	-0.590	0.367
RLSW	-0.463	0.388	-0.369	0.299	-0.575	0.369
MSEW	-0.363	0.326	-0.352	0.296	-0.574	0.369
Norway						
SELEC	-0.165	0.268	-0.216	0.271	-0.445	0.335
Linear Pooling						
EW	-0.245	0.281	-0.253	0.290	-0.452	0.366
RLSW	-0.174	0.268	-0.192	0.271	-0.415	0.334
MSEW	-0.235	0.280	-0.249	0.289	-0.453	0.364
Log. Pooling						
EW	-0.286	0.283	-0.290	0.292	-0.468	0.366
RLSW	-0.185	0.268	-0.207	0.272	-0.432	0.335
MSEW	-0.273	0.281	-0.283	0.290	-0.469	0.364
New Zealand						
SELEC	-0.782	0.458	-0.893	0.473	-1.013	0.476
Linear Pooling						
EW	-0.557	0.391	-0.646	0.415	-0.689	0.426
RLSW	-0.782	0.458	-0.872	0.476	-0.994	0.480
MSEW	-0.553	0.408	-0.695	0.443	-0.863	0.460
Log. Pooling						
EW	-0.517	0.401	-0.520	0.393	-0.536	0.407
RLSW	-0.778	0.457	-0.857	0.482	-0.935	0.473
MSEW	-0.574	0.412	-0.553	0.419	-0.633	0.414

Note: The table reports lnS and RMSPE using only the best and the worst model in the suite.

Figure 2: Cumulative Log scores



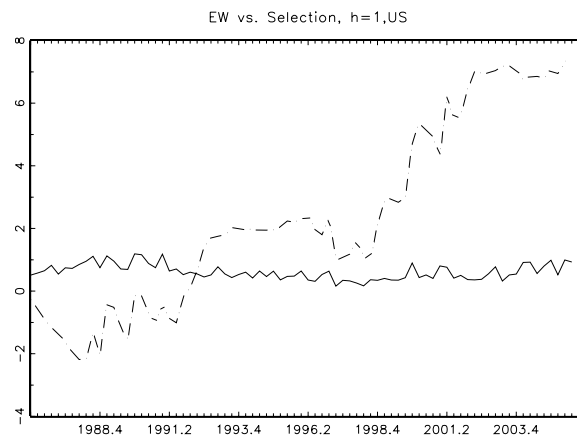
(a) US – UK



(b) Norway – New Zealand

Note: The graphs show the inflation series (solid lines) and the cumulative difference between the realized log scores of the log pooling combination with equal weights and selection (dashed lines).

Figure 3: US excluding AR4D1 and IMA1D1



Note: The graph shows the US inflation series (solid line) and the cumulative difference between the realized log scores of the log pooling combination with equal weights and selection (dashed line) when AR4D1 and IMA1D1 are excluded from the model suite.

WORKING PAPERS (ANO) FROM NORGES BANK 2005-2008

Working Papers were previously issued as Arbeidsnotater from Norges Bank, see Norges Bank's website <http://www.norges-bank.no>

- 2005/1 Q. Farooq Akram
*Efficient consumption of revenues from natural resources –
An application to Norwegian petroleum revenues* Research Department, 33 p
- 2005/2 Q. Farooq Akram, Øyvind Eitrheim and Lucio Sarno
*Non-linear dynamics in output, real exchange rates and real money balances: Norway, 1830-
2003* Research Department, 53 p
- 2005/3 Carl Andreas Claussen and Øistein Røisland
Collective economic decisions and the discursive dilemma Monetary Policy Department, 21 p
- 2005/4 Øistein Røisland
Inflation inertia and the optimal hybrid inflation/price level target
Monetary Policy Department, 8 p
- 2005/5 Ragna Alstadheim
Is the price level in Norway determined by fiscal policy? Research Department, 21 p
- 2005/6 Tommy Sveen and Lutz Weinke
Is lumpy investment really irrelevant for the business cycle? Research Department, 26 p
- 2005/7 Bjørn-Roger Wilhelmsen and Andrea Zaghini
Monetary policy predictability in the euro area: An international comparison
Economics Department, 28 p
- 2005/8 Moshe Kim, Eirik Gaard Kristiansen and Bent Vale
What determines banks' market power? Akerlof versus Herfindahl Research Department, 38 p
- 2005/9 Q. Farooq Akram, Gunnar Bårdsen and Øyvind Eitrheim
Monetary policy and asset prices: To respond or not? Research Department, 28 p
- 2005/10 Eirik Gard Kristiansen
Strategic bank monitoring and firms' debt structure Research Department, 35 p
- 2005/11 Hilde C. Bjørnland
Monetary policy and the illusionary exchange rate puzzle Research Department, 30 p
- 2005/12 Q. Farooq Akram, Dagfinn Rime and Lucio Sarno
Arbitrage in the foreign exchange market: Turning on the microscope
Research Department, 43 p
- 2005/13 Geir H. Bjønnes, Steinar Holden, Dagfinn Rime and Haakon O.Aa. Solheim
"Large" vs. "small" players: A closer look at the dynamics of speculative attacks
Research Department, 31 p
- 2005/14 Julien Garnier and Bjørn-Roger Wilhelmsen
The natural real interest rate and the output gap in the euro area: A joint estimation
Economics Department, 27 p
- 2005/15 Egil Matsen
Portfolio choice when managers control returns Research Department, 31 p
- 2005/16 Hilde C. Bjørnland
Monetary policy and exchange rate interactions in a small open economy
Research Department, 28 p
- 2006/1 Gunnar Bårdsen, Kjersti-Gro Lindquist and Dimitrios P. Tsomocos
Evaluation of macroeconomic models for financial stability analysis
Financial Markets Department, 45 p

- 2006/2 Hilde C. Bjørnland, Leif Brubakk and Anne Sofie Jore
Forecasting inflation with an uncertain output gap Economics Department, 37 p
- 2006/3 Ragna Alstadheim and Dale Henderson
Price-level determinacy, lower bounds on the nominal interest rate, and liquidity traps
Research Department, 34 p
- 2006/4 Tommy Sveen and Lutz Weinke
Firm-specific capital and welfare Research Department, 34 p
- 2006/5 Jan F. Qvigstad
When does an interest rate path „look good“? Criteria for an appropriate future interest rate path Norges Bank Monetary Policy, 20 p
- 2006/6 Tommy Sveen and Lutz Weinke
Firm-specific capital, nominal rigidities, and the Taylor principle Research Department, 23 p
- 2006/7 Q. Farooq Akram and Øyvind Eitrheim
Flexible inflation targeting and financial stability: Is it enough to stabilise inflation and output? Research Department, 29 p
- 2006/8 Q. Farooq Akram, Gunnar Bårdsen and Kjersti-Gro Lindquist
Pursuing financial stability under an inflation-targeting regime Research Department, 29 p
- 2006/9 Yuliya Demyanyk, Charlotte Ostergaard and Bent E. Sørensen
U.S. banking deregulation, small businesses, and interstate insurance of personal income
Research Department, 57 p
- 2006/10 Q. Farooq Akram, Yakov Ben-Haim and Øyvind Eitrheim
Managing uncertainty through robust-satisficing monetary policy Research Department, 33 p
- 2006/11 Gisle James Natvik:
Government spending and the Taylor principle Research Department, 41 p
- 2006/12 Kjell Bjørn Nordal:
Banks' optimal implementation strategies for a risk sensitive regulatory capital rule: a real options and signalling approach Research Department, 36 p
- 2006/13 Q. Farooq Akram and Ragnar Nymoen
Model selection for monetary policy analysis – importance of empirical validity
Research Department, 37 p
- 2007/1 Steinar Holden and Fredrik Wulfsberg
Are real wages rigid downwards? Research Department, 44 p
- 2007/2 Dagfinn Rime, Lucio Sarno and Elvira Sojli
Exchange rate forecasting, order flow and macroeconomic information
Research Department, 43 p
- 2007/3 Lorán Chollete, Randi Næs and Johannes A. Skjeltorp
What captures liquidity risk? A comparison of trade and order based liquidity factors
Research Department, 45 p
- 2007/4 Moshe Kim, Eirik Gaard Kristiansen and Bent Vale
Life-cycle patterns of interest rate markups in small firm finance Research Department, 42 p
- 2007/5 Francesco Furlanetto and Martin Seneca
Rule-of-thumb consumers, productivity and hours Research Department, 41 p
- 2007/6 Yakov Ben-Haim, Q. Farooq Akram and Øyvind Eitrheim
Monetary policy under uncertainty: Min-max vs robust-satisficing strategies
Research Department, 28 p

- 2007/7 Carl Andreas Claussen and Øistein Røisland
Aggregating judgments on dependent variables: an (im)possibility result
Research Department, 17 p
- 2007/8 Randi Næs, Johannes Skjeltorp og Bernt Arne Ødegaard
Hvilke faktorer driver kursutviklingen på Oslo Børs?
Forskningsavdelingen, 68 s
- 2007/9 Knut Are Astveit and Tørres G. Trovik
Nowcasting Norwegian GDP: The role of asset prices in a small open economy
Research Department, 29 p
- 2007/10 Hilde C. Bjørnland, Kai Leitemo and Junior Maih
Estimating the natural rates in a simple new Keynesian framework
Economics Department, 33 p
- 2007/11 Randi Næs and Bernt Arne Ødegaard
Liquidity and asset pricing: Evidence on the role of investor holding period
Research Department, 31 p
- 2007/12 Ida Wolden Bache
Assessing estimates of the exchange rate pass-through
Research Department, 60 p
- 2007/13 Q. Farooq Akram
What horizon for targeting inflation?
Research Department, 45 p
- 2007/14 Q. Farooq Akram, Yakov Ben-Haim and Øyvind Eitrheim
Robust-satisficing monetary policy under parameter uncertainty
Research Department, 33 p
- 2007/15 Ida Wolden Bache and Bjørn E. Naug
Estimating New Keynesian import price models
Research Department, 40 p
- 2008/1 Anne Sofie Jore, James Mitchell and Shaun P. Vahey
Combining forecast densities from VARs with uncertain instabilities
Economics Department, 26 p
- 2008/2 Henrik Andersen
Failure prediction of Norwegian banks: A logit approach
Financial Markets Department, 49 p
- 2008/3 Lorán Chollete, Randi Næs and Johannes A. Skjeltorp
The risk components of liquidity
Research Department, 28 p
- 2008/4 Hilde C. Bjørnland and Kai Leitemo
Identifying the interdependence between US monetary policy and the stock market
Economics Department, 28 p
- 2008/5 Christian Kascha and Karel Mertens
Business cycle analysis and VARMA models
Research Department, 40 p
- 2008/6 Alan S. Blinder
On the design of Monetary policy committees
Norges Bank Monetary Policy, 22 p
- 2008/7 Francesco Furlanetto
Does monetary policy react to asset prices? Some international evidence
Research Department, 44 p
- 2008/8 Christian Huerman, Francesco Ravazzolo and Chen Zhou
The power of weather. Some empirical evidence on predicting day-ahead power prices through weather forecasts
Research Department, 28 p
- 2008/9 Randi Næs, Johannes A. Skjeltorp and Bernt Arne Ødegaard
Liquidity at the Oslo Stock Exchange
Research Department, 49 p
- 2008/10 Francesco Furlanetto and Martin Seneca
Fiscal shocks and real rigidities
Research Department, 41 p

- 2008/11 Randi Næs, Johannes A. Skjeltop and Bernt Arne Ødegaard
Liquidity and the business cycle Research Department, 45 p
- 2008/12 Q. Farooq Akram
Commodity prices, interest rates and the dollar Research Department, 33 p
- 2008/13 Henrik Andersen, Sigbjørn Atle Berg and Eilev S. Jansen
The dynamics of operating income in the Norwegian banking sector
Financial Markets Department, 27 p
- 2008/14 Lars Fredrik Øksendal
Monetary policy under the gold standard – examining the case of Norway, 1893-1914
Norges Bank's bicentenary project, 103 p
- 2008/15 Hilde C. Bjørnland and Jørn I. Halvorsen
How does monetary policy respond to exchange rate movements? New international evidence
Research Department, 45 p
- 2008/16 Hilde C. Bjørnland
Oil price shocks and stock market booms in an oil exporting country
Research Department, 37 p
- 2008/17 Özer Karagedikli, Troy Matheson, Christie Smith and Shaun Vahey
RBCs and DSGEs: The computational approach to business cycle theory and evidence
Research Department, 36 p
- 2008/18 Ida Wolden Bache and Kai Leitemo
The price puzzle: Mixing the temporary and permanent monetary policy shocks
Monetary Policy Department, 12 p
- 2008/19 Q. Farooq Akram, Dagfinn Rime and Lucio Sarno
Does the law of one price hold in international financial markets? Research Department, 43 p
- 2008/20 Amund Holmsen, Jan F. Qvigstad, Øistein Røisland and Kristin Solberg-Johansen
Communicating monetary policy intentions: The case of Norges Bank
Monetary Policy Department, 33 p
- 2008/21 Pelin Ilbas
Revealing the preferences of the US Federal Reserve Research Department, 54 p
- 2008/22 Christian Kascha and Francesco Ravazzolo
Combining inflation density forecasts Research Department, 39 p

KEYWORDS:

Forecast combination

Logarithmic combinations

Density forecasts

Inflation forecasting