Journal of Forecasting
J. Forecast. **29**, 231–250 (2010)
Published online 6 January 2010 in Wiley InterScience
(www.interscience.wiley.com) **DOI**: 10.1002/for.1147

Combining Inflation Density Forecasts

CHRISTIAN KASCHA* AND FRANCESCO RAVAZZOLO
Research Department, Norges Bank, Oslo, Norway

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 datasets 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. Logarithmic combinations can be advantageous, in particular if symmetric densities are preferred.

KEYWORDS forecast combination; logarithmic combinations; density forecasts; inflation forecasting

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. 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 datasets and variations of the model suite.

Nowadays, most central banks want not only to control inflation but also to smooth output fluctuations over the business cycle. This strategy is often described as some sort of (flexible) inflation targeting (e.g., Svensson, 1999; Gali, 2008). Inflation itself is highly influenced by the state of the business cycle as prices are raised or lowered depending on the state of capacity utilization. As central banks face a trade-off between stabilizing output and inflation, many researchers such as Smets and Wouters (2007) have consequently tried to explain the cyclical comovements between the two variables. Good forecasts of future inflation are needed for the implementation of such a policy as the economy typically reacts with a lag to a change in the instruments of the central bank. Recognizing the fundamental importance of forecasts, modern central bank strategies have even been coined *inflation forecast targeting* (Svensson, 1997).

*Correspondence to: Christian Kascha, Research Department, Norges Bank, Bankplassen 2, 0107 Oslo, Norway. E-mail: christian.kascha@norges-bank.no

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 central banks 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. However, policy makers usually have a whole suite of forecast models at hand. In this situation, some questions 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 a 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). Timmermann (2006) provides an extensive summary of the literature 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.

Corradi and Swanson (2006b) provide a survey on the evaluation of individual density and interval forecasts. See also Clemen *et al.* (1995). Clements (2006) and Granger *et al.* (1989) have considered combinations 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. 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. 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 (1977) and Winkler (1981), Hall and Mitchell (2004) consider an approach where 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 *et al.* (1997) and Geweke and Whiteman (2006).

There are very few studies in economics that take a comparative point of view and evaluate alternative methods of density forecast combinations. Jore *et al.* (2008) provide some evidence on the performance of the weighting scheme proposed by, for example, Hall and Mitchell (2007) relative to equal weights and a pairwise averaging method. 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 datasets and focus on the combination methods' performance *throughout* these datasets 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 forecasting models and compare the results over datasets for the USA, the UK, Norway and New Zealand. The size of the model suite is relatively modest. All models are estimated using least squares or maximum likelihood estimators and a moving window of fixed size to account for structural change. We refer to the aforementioned literature for Bayesian approaches. We investigate two

DOI: 10.1002/for

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' (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 the worst models in a suite. Then we study how predictive density combinations perform relative to individual density forecasts and selecting the best-performing model at the forecasting origin.

Our results show that combining forecasts provide 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 datasets than recursive log score weights.

The rest of the paper is organized as follows. In the next section we discuss the evaluation and combination of density forecasts. In the third section we describe the data and the suite of density forecast models. The fourth section contains the results of the out-of-sample experiment. The fifth section concludes.

EVALUATING AND COMBINING DENSITY FORECASTS

Since the field of density forecasting is in a relatively infant state in economics, 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.

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 (2006b) provide an extensive survey. This question is decisive because it is central to how we design density combination schemes (Hall and Mitchell, 2007). 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 *et al.*, 1998; Bai, 2003) or conditional on a given information set (Corradi and Swanson, 2003a). 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). Alternatively, two statistical approaches have been considered in the recent literature. One is based on a distributional analog of the mean squared error norm (Corradi and Swanson, 2003b, 2006a); 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).

DOI: 10.1002/for

The measure of distributional accuracy introduced by Corradi and Swanson (2003b, 2006a) is attractive because of its analogy to the usual mean squared error norm in point forecasting. 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 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 test was 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 is the density of a real-valued, absolutely continuous random variable Y_t and we have a set of two densities f_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 and f_i is defined as

$$\begin{aligned} \mathrm{KLIC}_i &= \int f(y_t) \ln \frac{f(y_t)}{f_i(y_t)} \mathrm{d}y_t \\ &= E[\ln f(y_t) - \ln f_i(y_t)] \end{aligned} \tag{1}$$

where E denotes the expectation. We assume here and in the following that all densities are strictly positive, i.e., $f(y) > 0$, $f_i(y) > 0$ 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):

$$\mathrm{E} \ln \mathrm{S}_i = E[\ln f_i(y_t)] \tag{2}$$

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

$$\ln \mathrm{S}_i = \frac{1}{T} \sum_{t=1}^{T} \ln f_i(y_t) \tag{3}$$

Therefore, we actually do not need to know f 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.

With respect to density forecasts, we are usually interested in comparing densities conditional on different information sets instead of approximating a 'true' density. Let $f_{t+h,t,i}$ therefore denote a prediction of the density for Y_{t+h}, $h = 1, 2, \ldots$, conditional on some information set available at 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

DOI: 10.1002/for

total number of T observations. A measure of out-of-sample forecasting performance is the (out-of-sample) log score given by

$$\ln \mathrm{S}_{i,h} = \frac{1}{T-h-T^S+1} \sum_{t=T^S}^{T-h} \ln f_{t+h,t,i}(y_{t+h}) \tag{4}$$

Models or combination schemes that are associated with a high average log score give a higher probability to events which have actually occurred (Hall and Mitchell, 2007). Therefore, (4) is our preferred measure of forecast accuracy.

Combining density forecasts

There are two elementary choices in combining predictive densities. One is the method of aggregation or functional form of combining. The other is the construction of the weights attached to the individual densities. Possible methods of aggregation are described in an early review of Genest and Zidek (1986). We consider two different functional forms: linear combination and logarithmic combination. To the best of our knowledge, these are the only popular approaches in the literature for the combination of predictive densities. Some alternative approaches to combination of predictive densities are, however, given in Hall and Mitchell (2004).

As before, we consider forecasting a real-valued random variable Y_t that has a density. We consider N competitive, strictly positive, h-step-ahead density forecasts, $\{f_{t+h,t,1}, \ldots, f_{t+h,t,N}\}$, obtained using information up to time t. No other assumptions on the individual density forecasts are needed to ensure that the following combination methods actually yield densities. In particular, we are not restricted to Gaussian forecasts or linear models. The existence of $E[\ln f_{t+h,t,i}(y_t)]$ is, however, assumed such that the comparison in terms of log scores is meaningful.

The easiest combination method is linear combination (Stone, 1961):

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

where $\omega_{t+h,t,i}$ 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 $\Sigma_{i=1}^{N} \omega_{t+h,t,i} = 1$ for all $i = 1, \ldots, N$, such that the resulting combination is indeed a density function.

An alternative way of combining densities is logarithmic combination:

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

where the non-negative weights are chosen as before such that the integral in the denominator exists. Winkler (1968) points out that the logarithmic opinion pool has a natural-conjugate interpretation. For example, consider the combination of normal densities with means and variances μ_i, σ_i, $i = 1, \ldots, N$, respectively. Denote the transformed weights by $\alpha_i = \omega_i/\sigma_i^2$ accordingly. The logarithmic pool is a normal density, $N(\mu_c, \sigma_c^2)$, with mean and variance given by $\mu_c = \Sigma_{i=1}^{N} \alpha_i \mu_i / (\Sigma_{i=1}^{N} \alpha_i)$ and $\sigma_c^2 = (\Sigma_{i=1}^{N} \alpha_i)^{-1}$. Thus the logarithmic combination method retains the symmetry of the individual forecasts in this case.

DOI: 10.1002/for

An example in Figure 1 illustrates the main difference between the two aggregation schemes. Here, we combine two density functions of two normally distributed random variables with $N(-2, 1)$ and $N(2, 2)$, respectively. The weight for each individual density is 1/2. From the definition, it is immediately clear that the linear combination is typically multimodal. Logarithmic combination

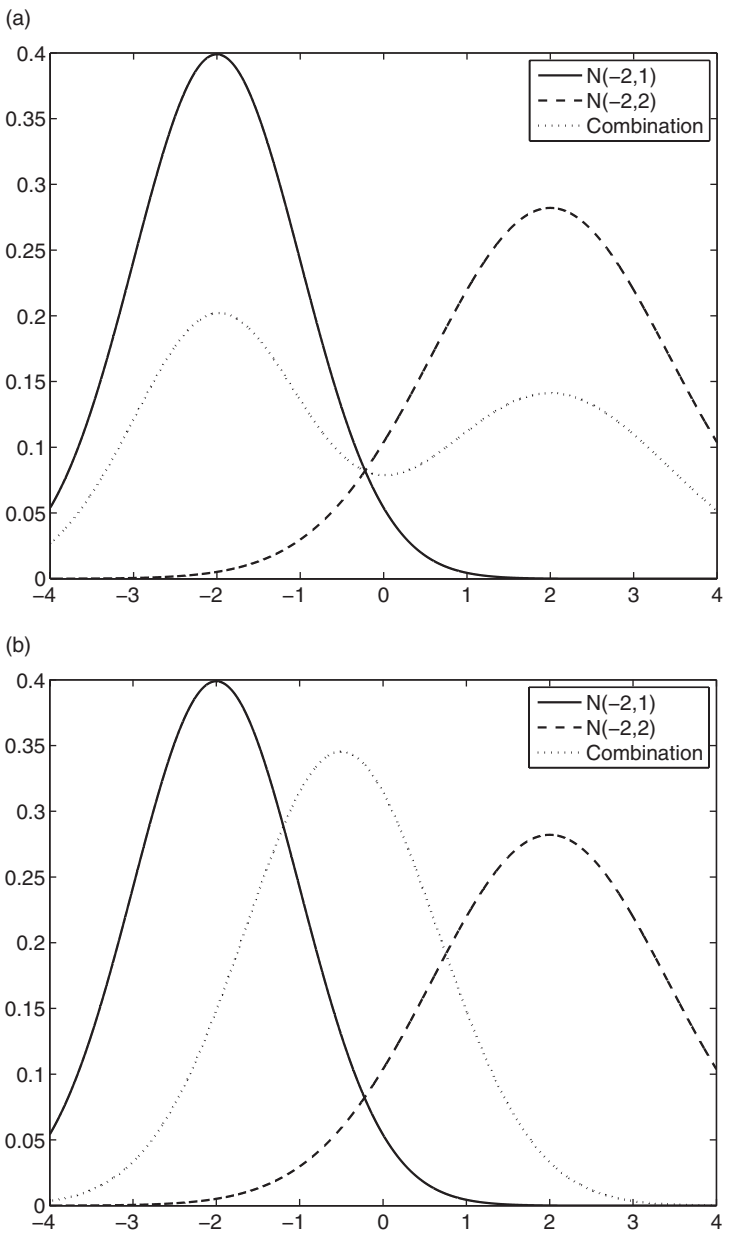


Figure 1. Individual densities and density combinations using (a) linear and (b) logarithmic combination

DOI: 10.1002/for

yields again a density of a normally distributed random variable with mean −1/2 and variance 4/3, $N(-1/2, 4/3)$. Furthermore, the linear combination 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.

Some reasons to combine density forecasts and to use the above schemes in particular have been given by Genest and Zidek (1986) in a decision-theoretic framework. Logarithmic combination has been put forward by Winkler (1968). Raftery *et al.* (1997) and Mitchell and Hall (2005) argue for linear combination 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.

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 individual density forecasts obtained in some way. A forecaster combines both with weights λ, $1 - \lambda$, $\lambda \in [0, 1]$ according to either (5) or (6). Suppose that forecast 1 is better than forecast 2 in terms of log score:

$$E[\ln f_1(y_{t+h})] \geq E[\ln f_2(y_{t+h})] \tag{7}$$

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

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

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}:

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

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:

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

DOI: 10.1002/for

Thus, combining provides insurance even in this case:

$$E\left[\ln \frac{f_1^{\lambda}(y_{t+h}) f_2^{1-\lambda}(y_{t+h})}{\int f_1^{\lambda}(y_{t+h}) f_2^{1-\lambda}(y_{t+h}) dy_{t+h}}\right] \geq \lambda E[\ln f_1(y_{t+h})] + (1-\lambda) \ln E[f_2(y_{t+h})],$$
$$\geq E[f_2(y_{t+h})]$$

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 for combining densities have 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): A weighting scheme based on the out-of-sample performance of density forecasts are recursive log score weights as proposed in, for example, Jore *et al.* (2008). The weights for the h-step-ahead density combination take the form:

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

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).

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

$$\omega_{t+h,t,i} = \frac{1/\mathrm{MSPE}_{t+h,t,i}}{\sum_{k=1}^{N} 1/\mathrm{MSPE}_{t+h,t,k}}, \tag{9}$$

$$\mathrm{MSPE}_{t+h,t,i} = \frac{1}{t-h-\underline{t}+1} \sum_{\tau=\underline{t}}^{t-h} (y_{\tau+h} - \mu_{\tau+h,\tau,i})^2$$

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. It is natural to assume that a forecaster

DOI: 10.1002/for

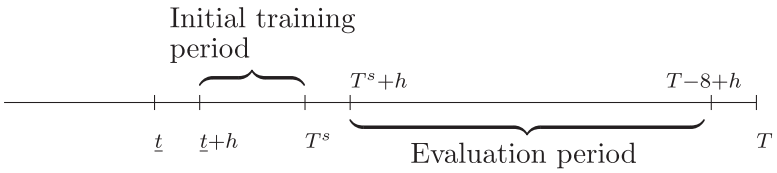
if he has to select one model chooses the model that performed best in the past. Since we are interested in predictive densities the relevant criterion is the average log score for a given horizon. We apply two selection strategies: the first is based on the evaluation of an expanding window of observations (SELEC_ew), while the second is based on the evaluation of a moving window of the last 10 observations (SELEC_mw). 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. Note that the way we select models is related to the standard AIC criterion based on the predictive likelihood.

DATA AND MODELS

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, $x_1, \ldots, x_T$ 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 USA, Consumer Price indices (CPI) for the UK and New Zealand and the Norwegian core CPI.[1] The set of potential predictors contains a quarterly M2 money measure (M1 in the case of New Zealand), M_t, a 3-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.

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 information on the out-of-sample performance of the individual models for $\underline{t} + h$ to T^s. The 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:



The used data spans from 1960Q1 to 2007Q3 for the USA, from 1978Q1 to 2007Q2 for the UK, from 1979Q2 to 2007Q3 for Norway and from 1981Q1 to 2008Q1 for New Zealand. Table I provides information on the evaluation period.

[1] We 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.

DOI: 10.1002/for

Models and forecasting

The model suite is composed of a set of univariate and multivariate specifications. The univariate models may in part be justified as simple 'forecasting devices' as in Clements and Hendry (2006). The multivariate models comprise two Philips curve-type models and different vector autoregressive (VAR) and vector autoregressive moving-average (VARMA) models that contain variables usually considered in the literature on forecasting inflation. We limit ourselves to linear models even though the focus on density forecasts gives scope to nonlinear models. The reason is that the evidence on the forecasting performance of nonlinear models is decidedly mixed (see, for example, Marcellino, 2004, 2008). We therefore leave a comparison of forecasts resulting from linear and nonlinear models for future research. 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. The complete list of models is given in Table II.

Some comments are in order here. First, we do not claim that the model suites are optimal. 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 II reveals that some choices such as the lag lengths 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 datasets 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 datasets.

All models are linear and most are estimated by least squares regressions using a moving window of the last m observations. Exceptions are the univariate moving-average (MA) and the VARMA models. The MA models are estimated using the Time Series 4.0 package in GAUSS and the VARMA models using the algorithm of Hannan and Kavalieris (1984). We apply iterative forecasting, and we compute density forecasts using a normal approximation assuming knowledge of coefficients and the entire history of the time series. Since all models are linear, we can express the inflation series π_t as a function of past errors and initial values as

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

where π summarizes the initial conditions. Assuming that the past errors and coefficients are known, the conditional expectation corresponds to the point forecast $\pi_{t+h,t} = \Sigma_{i=h}^{t-1}\phi_i\varepsilon_{t+h-i} + b\pi_0$, and the forecast error is $\pi_{t+h} - \pi_{t+h,t} = \Sigma_{i=0}^{h-1}\phi_i\varepsilon_{t+h-i}$. It follows that the forecast error variance is given by $\sigma^2(h) = E[(x_{t+h} - x_{t+h,t})^2] = \sigma^2\Sigma_{i=0}^{h-1}\phi_i^2$. 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, $N(\pi_{t+h,t}, \sigma^2(h))$. Even though parameter estimation uncertainty is not taken into account when deriving the models' predictive densities, this approximation is still useful since our focus is decidedly on the combination of a given set of forecasts. Note in particular that the actual forecast errors in the empirical applications will in general not be i.i.d. normally distributed errors, in contrast to the assumptions made above. A more elaborate approach that also takes parameter estimation uncertainty into account would have to resort to simulation techniques as, for example, in Kilian (1998). This topic, however, is left for future research.

DOI: 10.1002/for

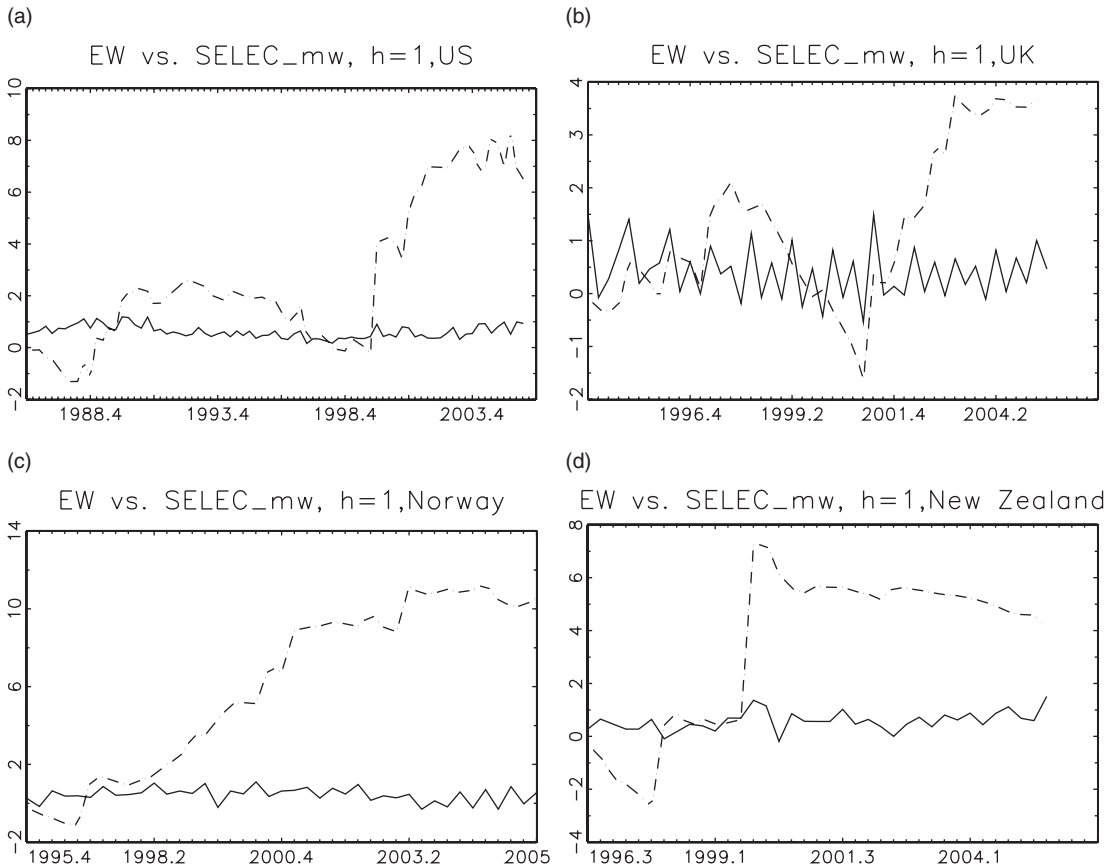


Figure 2. Cumulative log scores. The graphs show the inflation series (solid lines) and the cumulative difference between the log scores of the logarithmic combination with equal weights and selection (dashed lines): (a) USA; (b) UK; (c) Norway; (d) New Zealand

RESULTS

The results of the out-of-sample evaluation are summarized in Figures 2 and 3 and in Tables III–VI. 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 root mean square prediction error (RMSPE). We focus mainly on the lnS. Tables III–VI tabulate the out-of-sample forecasting performance of the individual models, the combination schemes and the selection strategy for each of the four countries. An explanation of the acronyms for the individual models is given in Table II.

We apply the test of equal accuracy of two density forecasts given in Mitchell and Hall (2005). Suppose there are two density forecasts, $f_{t+h,t,1}(y_{t+h})$ and $f_{t+h,t,2}(y_{t+h})$, and consider the loss differential

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

DOI: 10.1002/for

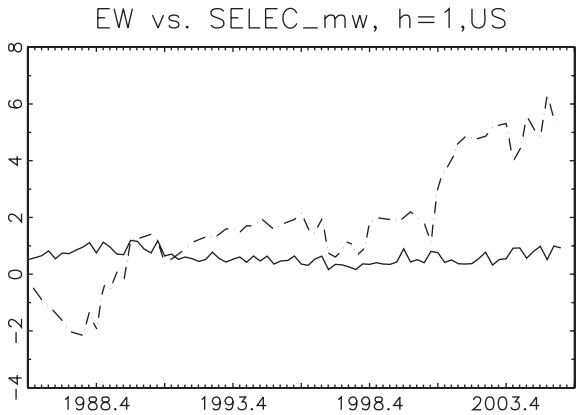


Figure 3. USA excluding AR4D1 and IMA1D1. The graph shows the US inflation series (solid line) and the cumulative difference between the log scores of the logarithmic combination with equal weights and selection (dashed line) when AR4D1 and IMA1D1 are excluded from the model suite

The null hypothesis of equal accuracy is $\mathcal{H}_0$:$E(d_{t+h}) = 0$. The sample mean, $\bar{d}_{t+h}$, has under appropriate assumptions the limiting distribution $\sqrt{T}\left(\bar{d}_{t+h} - d_{t+h}\right) \rightarrow N(0, \Omega)$. We compare the density forecasts resulting from different combinations to the one resulting from selection based on a moving window of observations. In our application, Tables III–VI show that differences are often statistically significant, excluding the UK case.

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, there are some models, both univariate and multivariate specifications, that provide good point forecasts but yield poor density forecasts. 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 datasets such as for 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 USA, the VAR2D4_pi in the case of Norway and the SVAR4D1_py in the case of the UK. In the case of New Zealand, the evidence over horizons is more in favor of trivariate VARs.

Selecting the best model at the forecast origin is generally difficult and can lead to quite inaccurate density forecasts. The moving window approach, SELEC_mw, is more accurate for the USA, the UK and New Zealand than SELEC_ew. But only in the cases of the USA and the UK, both at forecast horizons $h = 4$, does SELEC_mw perform better than the best individual model. For $h = 1$ with Norwegian data, SELEC_ew yields forecasts as accurate as the best individual model: RWD4. In all the other cases, there are several models and combination schemes that perform better. Results are qualitatively similar in terms of point forecast accuracy. Therefore our results suggest that it is quite difficult to select the best individual model at the forecast origin and findings depend on the evaluation window that is applied.

DOI: 10.1002/for

The results for the combination schemes are given in the lower part of the tables. Combined forecasts dominate the individual models' forecasts in several, but not all, cases. However, in 11 out of 12 cases there are at least two combination methods that provide higher lnSs than the selection strategy and only for the USA at forecast horizon $h = 4$ does selection give the highest average lnS. Therefore combining is a 'safe' approach to minimize density forecast errors and seems preferable to selecting a model. 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 four cases and often only marginally.

Logarithmic combination with equal and MSE weights provides high lnSs. In 11 cases the lnSs of these combinations are higher than selection. They are the highest among all the competitors in six cases. Differences with linear pooling with the same weights are, however, minor. As in Jore *et al.* (2008), recursive log score weights give marginally more accurate forecasts than other weighting schemes for the USA. For the other countries, RLSW weights yield substantially worse forecasts than alternative schemes. While RLSW weights are explicitly based on past density forecast accuracy, the estimation of these weights is apparently rather difficult in small samples, in particular when there is high instability in the relative performance of the individual models, as in the case of New Zealand.

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. 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 and we focus on $h = 1$. That is, we compute

$$\mathrm{ClnS}_t = \sum_{s=T^s}^{t} \ln f_{s+1,s,C}(y_{s+1}) - \ln f_{s+1,s,S}(y_{s+1})$$

for $t = T^s, \ldots, T - 8 + 1$, and $f_{s+1,s,C}$ and $f_{s+1,s,S}$ are the density forecasts obtained from logarithmic combining and selection, respectively. Thus ClnS_t increases when $f_{t+1,t,C}$ turns out to be more accurate than $f_{t+1,t,S}$. Ideally, we would like to see that ClnS_t increases steadily over time. This is roughly the case for Norway and New Zealand. For the USA and the UK, the pattern is less clear, even though combining is still superior on average. This is because IMA1D1 and AR4D1 perform much better than all other models in the suite for US data. Selecting the best (or the second best) model is therefore easier. When we exclude these two models from the suite, the plot is similar to the other ones (Figure 3). A similar result is found for UK data. This explains also why recursive log score weights perform so well for the USA. Assigning higher weights to dominant models is simpler in this case and improves forecast accuracy.

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 *ex post*, based on the results in Tables III–VI. We develop four exercises.[2] In the first exercise, the model suites are limited to the six best-performing models. In the second and third exercises, the model

[2]The results of the robustness exercises are available upon request.

DOI: 10.1002/for

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. Finally, only the best and the worst model of each collection are used in the model suite. All these exercises potentially favor a selection strategy. Overall, we find that combination methods, even equal-weight schemes, are still performing well compared to selection. In sum, the exercises broadly confirm the results of the main out-of-sample evaluation.

CONCLUSION

This paper extends the empirical literature on combining inflation density forecasts by evaluating several aggregation schemes over four different datasets. 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 combination schemes do not always beat the best individual models but almost always outperform a strategy which selects an individual model at the forecast origin based on past 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 nonlinear 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 of other density aggregation schemes is another interesting topic for future research.

APPENDIX: 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, and 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.

DOI: 10.1002/for

Table I. Datasets

	Sample	Evaluation period	
USA	1960 Q1–2007 Q3	(1986 Q1 + h)–(07 Q3–8 + h)	(79)
UK	1978 Q1–2007 Q2	(1994 Q1 + h)–(07 Q2–8 + h)	(46)
Norway	1979 Q2–2007 Q3	(1995 Q3 + h)–(07 Q3–8 + h)	(41)
NZ	1981 Q1–2008 Q1	(1996 Q2 + h)–(08 Q1–8 + h)	(37)

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

Table II. Definitions of forecasting models

Name	Definition	Variables	m
USA 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	$\pi_t = \mu + \alpha_1\pi_{t-1} + \ldots + \alpha_4\pi_{t-4} + \varepsilon_t$		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	$\Delta_1\mathbf{x}_t = \mu + A_1\Delta_1\mathbf{x}_{t-1} + u_t + M_1u_{t-1}$	$\mathbf{x}_t = (p_t, M_t)'$	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	$\pi_t = \mu + s_1d_{1t} + s_2d_{2t} + s_3d_{3t} + \alpha_1\pi_{t-1} + \alpha_2\pi_{t-2} + \varepsilon_t$		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	$\Delta_1\mathbf{x}_t = \mu + s_1d_{1t} + s_2d_{2t} + s_3d_{3t} + A_1\Delta_1\mathbf{x}_{t-1} + \ldots + A_4\mathbf{x}_{t-4} + u_t$	$\mathbf{x}t = (p_t, y_t)'$	50
SVAR4D1_pi		$\mathbf{x}_t = (p_t, i_t)'$	

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 and y_t is an output measure. In the PC models, $\alpha(L) = \alpha_1 + \ldots + \alpha_p L^{p-1}$ and $\beta(L) = \beta_1 + \ldots + \beta_q L^{q-1}$ are lag polynomials. Lag lengths are estimated using the BIC in the case of PC models with maximal order equal to four.

DOI: 10.1002/for

APPENDIX B: RESULTS

Table III. Out-of-sample prediction performance, USA

	$h = 1$		$h = 4$		$h = 8$	
	lnS	RMSPE	lnS	RMSPE	lnS	RMSPE
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	0.061	0.214	–0.034	0.214	–0.470	0.288
PC-U	0.070	0.207	–0.024	0.210	–0.484	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	0.039	0.209	–0.165	0.244	–0.368	0.293
Selection						
SELEC_ew	0.097	0.199	–0.009	0.226	–0.415	0.286
SELEC_mw	0.034	0.216	0.072	0.222	–0.291	0.285
Linear pooling						
EW	0.120	0.203	–0.042	0.218	–0.278	0.264
RLSW	0.155**	0.195	0.018	0.222	–0.295	0.280
MSEW	0.127	0.202	–0.020	0.213	–0.246	0.258
Log. pooling						
EW	0.117	0.202	–0.014	0.217	–0.250	0.262
RLSW	0.142**	0.195	0.014	0.221	–0.317	0.277
MSEW	0.123	0.201	0.028	0.216	–0.186**	0.259

Note: In the table lnS denotes the average log score, RMSPE denotes the root mean squared prediction error, and * and ** indicate significant differences in performance between SELEC_mw and the combination methods at the 90% and 95% level, respectively, according to the test of Mitchell and Hall (2005). See Table II for explanation of the model suite.

Table IV. 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	–0.450	0.391	–0.375	0.351	–0.550	0.454
IMA1D4	–0.468	0.314	–0.409	0.295	–0.694	0.340
PC-Y	–0.536	0.316	–0.439	0.361	–0.803	0.606
PC-U	–0.477	0.382	–0.399	0.340	–0.700	0.568
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
VAR2D4_piy	–0.457	0.349	–0.476	0.324	–0.779	0.450
VAR2D4_pmy	–0.553	0.422	–0.492	0.327	–0.770	0.426
SVAR4D1_py	–0.353	0.351	–0.376	0.358	–0.524	0.420
SVAR4D1_pi	–0.408	0.383	–0.392	0.399	–0.660	0.528

DOI: 10.1002/for

Table IV. *Continued*

	$h = 1$		$h = 4$		$h = 8$	
	lnS	RMSPE	lnS	RMSPE	lnS	RMSPE
Selection						
SELEC_ew	–0.431	0.390	–0.410	0.339	–0.611	0.412
SELEC_mw	–0.405	0.368	–0.348	0.338	–0.587	0.415
Linear pooling						
EW	–0.391	0.304	–0.374	0.303	–0.608	0.410
RLSW	–0.416	0.380	–0.386	0.322	–0.582	0.399
MSEW	–0.381	0.309	–0.370	0.299	–0.610	0.394
Log. pooling						
EW	–0.324	0.303	–0.318	0.303	–0.535	0.425
RLSW	–0.408	0.378	–0.380	0.324	–0.544	0.420
MSEW	–0.316	0.311	–0.316	0.300	–0.533	0.405

Note: See note to Table III.

Table V. 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	–0.211	0.267	–0.212	0.264	–0.418	0.297
IMA1D4	–0.216	0.263	–0.190	0.269	–0.381	0.331
PC-Y	–0.232	0.260	–0.330	0.283	–.060	0.405
PC-U	–0.241	0.264	–0.453	0.296	–0.862	0.384
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
VAR2D4_piy	–0.241	0.267	–0.200	0.248	–0.370	0.301
VAR2D4_pmy	–0.357	0.275	–0.222	0.256	–0.369	0.299
SVAR4D1_py	–0.242	0.248	–0.301	0.246	–0.509	0.300
SVAR4D1_pi	–0.213	0.248	–0.282	0.251	–0.488	0.305
Selection						
SELEC_ew	–0.165	0.268	–0.283	0.278	–0.528	0.332
SELEC_mw	–0.305	0.259	–0.419	0.280	–0.569	0.315
Linear pooling						
EW	–0.039	0.236	–0.121*	0.252	–0.325	0.299
RLSW	–0.175	0.264	–0.194	0.269	–0.376	0.311
MSEW	–0.050	0.241	–0.114*	0.251	–0.316*	0.298
Log. pooling						
EW	–0.046**	0.235	–0.131*	0.258	–0.299**	0.309
RLSW	–0.191	0.261	–0.223	0.269	–0.431	0.305
MSEW	–0.070**	0.240	–0.119**	0.255	–0.289**	0.304

Note: See note to Table III.

DOI: 10.1002/for

Table VI. 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	–0.931	0.405	–0.831	0.399	–0.801	0.399
IMA1D1	–0.731	0.384	–0.719	0.420	–0.584	0.408
PC-Y	–0.611	0.389	–0.668	0.449	–0.628	0.402
PC-U	–0.625	0.403	–0.757	0.484	–0.683	0.409
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	–0.821	0.459
VARMA11D1_pi	–0.737	0.438	–0.831	0.558	–0.847	0.528
Selection						
SELEC_ew	–0.768	0.404	–1.134	0.433	–1.156	0.434
SELEC_mw	–0.597	0.374	–1.101	0.455	–0.777	0.388
Linear pooling						
EW	–0.525	0.378	–0.597	0.413	–0.600	0.397
RLSW	–0.732	0.401	–1.098*	0.436	–1.001	0.426
MSEW	–0.476	0.378	–0.553*	0.395	–0.561	0.394
Log. pooling						
EW	–0.486	0.378	–0.554*	0.392	–0.541	0.391
RLSW	–0.755	0.398	–1.113	0.422	–0.969*	0.405
MSEW	–0.518	0.382	–0.600**	0.395	–0.556	0.392

Note: See note to Table III.

ACKNOWLEDGEMENTS

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.

REFERENCES

Amisano G, Giacomini R. 2007. Comparing density forecasts via weighted likelihood ratio tests. *Journal of Business and Economic Statistics* **25**: 177–190.

Bai J. 2003. Testing parametric conditional distributions of dynamic models. *Review of Economics and Statistics* **85**: 531–549.

Bates JM, Granger CWJ. 1969. Combination of forecasts. *Operational Research Quarterly* **20**: 451–468.

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

Bierens H. 1982. Consistent model-specification tests. *Journal of Econometrics* **20**: 105–134.

DOI: 10.1002/for

Bierens HJ, Ploberger W. 1997. Asymptotic theory of integrated conditional moments tests. *Econometrica* **65**: 1129–1151.
Clemen RT, Murphy AH, Winkler RL. 1995. Screening probability forecasts: contrasts between choosing and combining. *International Journal of Forecasting* **11**: 133–145.
Clements MP. 2004. Evaluating the Bank of England density forecasts of inflation. *Economic Journal* **114**: 844–866.
Clements MP. 2006. Evaluating the Survey of Professional Forecasters probability distributions of expected inflation based on derived event probability forecasts. *Empirical Economics* **31**: 49–64.
Clements MP, Hendry DF. 2006. Forecasting with breaks. In Handbook of *Economic Forecasting*, Elliot G, Granger CWJ, Timmermann A (eds). Elsevier: Amsterdam; 605–657.
Corradi V, Swanson NR. 2003a. Bootstrap conditional distribution tests in the presence of dynamic misspecification. *Journal of Econometrics* **133**: 779–806.
Corradi V, Swanson NR. 2003b. *A test for comparing multiple misspecified conditional distributions*. Departmental Working Papers 200314, Rutgers University.
Corradi V, Swanson NR. 2006a. Predictive density and conditional confidence interval accuracy tests. *Journal of Econometrics* **135**: 187–228.
Corradi V, Swanson NR. 2006b. Predictive density evaluation. In *Handbook of Economic Forecasting*, Elliot G, Granger CWJ, Timmermann A (eds). Elsevier: Amsterdam; 197–284.
Diebold FX, Gunther T, Tay AS. 1998. Evaluating density forecasts with applications to finance and management. *International Economic Review* **39**: 863–883.
Fernández-Villaverde J, Rubio-Ramirez JF. 2004. Comparing dynamic equilibrium models to data. *Journal of Econometrics* **123**: 153–187.
Gali J. 2008. *Introduction to Monetary Policy, Inflation, and the Business Cycle: An Introduction to the New Keynesian Framework*. Princeton University Press: Princeton, NJ.
Genest C, Zidek J. 1986. Combining probability distributions: a critique and an annotated bibliography. *Statistical Science* **1**: 114–148.
Geweke J, Whiteman C. 2006. Bayesian forecasting. In *Handbook of Economic Forecasting*, Elliot G, Granger C, Timmermann A (eds). Elsevier: Amsterdam; 3–80.
Granger CWJ, Pesaran MH. 2000. Economic and statistical measures of forecast accuracy. *Journal of Forecasting* **19**: 537–560.
Granger CWJ, Ramanathan R. 1984. Improved methods of combining forecasts. *Journal of Forecasting* **3**: 197–204.
Granger CWJ, White H, Kamstra M. 1989. Interval forecasting: an analysis based upon ARCH-quantile estimators. *Journal of Econometrics* **40**: 87–96.
Hall SG, Mitchell J. 2004. *Density forecast combination*. National Institute of Economic and Social Research Discussion Paper, No. 249.
Hall SG, Mitchell J. 2007. Combining density forecasts. *International Journal of Forecasting* **23**: 1–13.
Hannan EJ, Kavalieris L. 1984. Multivariate linear time series models. *Advances in Applied Probability* **16**: 492–561.
Hendry DF, Clements MP. 2004. Pooling of forecasts. *Econometrics Journal* **7**: 1–31.
Jore AS, Mitchell J, Vahey SP. 2008. *Combining forecast densities from VARs with uncertain instabilities*. Norges Bank Working Paper 2008/01.
Kilian L. 1998. Small-sample confidence intervals for impulse response functions. *Review of Economics and Statistics* **80**: 218–230.
Kitamura Y. 2002. *Econometric comparisons of conditional models*. Working paper, University of Pennsylvania.
Leamer E. 1978. *Specification Searches*. Wiley: New York.
Li F, Tkacz G. 2006. A consistent bootstrap test for conditional density functions with time-dependent data. *Journal of Econometrics* **127**: 863–886.
Marcellino M. 2004. Forecast pooling for European macroeconomic variables. *Oxford Bulletin of Economics and Statistics* **66**: 91–112.
Marcellino M. 2008. A linear benchmark for forecasting GDP growth and inflation? *Journal of Forecasting* **27**: 305–340.
Min CK, 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.

DOI: 10.1002/for

Mitchell J, Hall SG. 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. 1977. Combining expert judgments: a Bayesian approach. *Management Science* **23**: 679–693.
Palm FC, Zellner A. 1992. To combine or not to combine? *Journal of Forecasting* **11**: 687–701.
Raftery AE, Madigan D, Hoeting JA. 1997. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association* **92**: 179–191.
Smets F, Wouters R. 2007. Shocks and frictions in US business cycles: a Bayesian DSGE approach. *American Economic Review* **97**: 586–606.
Stock JH, Watson MW. 2007. Why has U.S. inflation become harder to forecast? *Journal of Money, Credit and Banking* **39**(02): 3–33.
Stone M. 1961. The opinion pool. *Annals of Mathematical Statistics* **32**: 1339–1342.
Svensson LEO. 1997. Inflation forecast targeting: implementing and monitoring inflation targets. *European Economic Review* **41**: 1111–1146.
Svensson LEO. 1999. Inflation targeting as a monetary policy rule. *Journal of Monetary Economics* **43**: 607–654.
Timmermann A. 2006. Forecast combinations. In *Handbook of Economic Forecasting*, Elliot G, Granger CWJ, Timmermann A (eds). Elsevier: Amsterdam; 135–196.
Vuong Q. 1989. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* **57**: 307–333.
Wallis KF. 2005. Combining density and interval forecasts: A modest proposal. *Oxford Bulletin of Economics and Statistics* **67**: 983–994.
Winkler RL. 1968. The consensus of subjective probability distributions. *Management Science* **15**: B61–B75.
Winkler RL. 1981. Combining probability distributions from dependent information sources. *Management Science* **27**: 479–488.

Authors' biographies:
Christian Kascha is an Advisor in the Research Department of Norges Bank's monetary policy wings. He received his Ph.D. in Economics from the European University Institute in 2007.

Francesco Ravazzolo is an Advisor in the Research Department of Norges Bank's monetary policy wings. He received his Ph.D. in Economics from Erasmus University Rotterdam in 2007.

Authors' addresses:
Christian Kascha and **Francesco Ravazzolo**, Norges Bank, Research Department, Bankplassen 2, 0107 Oslo, Norway.

DOI: 10.1002/for