COMBINING MULTIVARIATE DENSITY FORECASTS USING PREDICTIVE CRITERIA

HUGO GERARD♠ AND KRISTOFFER NIMARK†

Abstract. This paper combines multivariate density forecasts of output growth, inflation and interest rates from a suite of models. An out-of-sample weighting scheme based on the predictive likelihood as proposed by Eklund and Karlsson (2005) and Andersson and Karlsson (2007) is used to combine the models. Three classes of models are considered: a Bayesian vector autoregression (BVAR), a factor-augmented vector autoregression (FAVAR) and a medium-scale dynamic stochastic general equilibrium (DSGE) model. Using Australian data, we find that, at short forecast horizons, the Bayesian VAR model is assigned the most weight, while at intermediate and longer horizons the factor model is preferred. The DSGE model is assigned little weight at all horizons, a result that can be attributed to the DSGE model producing density forecasts that are very wide when compared with the actual distribution of observations. While a density forecast evaluation exercise reveals little formal evidence that the optimally combined densities are superior to those from the best-performing individual model, or a simple equal-weighting scheme, this may be a result of the short sample available.

Keywords: Density forecasts, combining forecasts, predictive criteria

1. Introduction

Density forecasts, or fan charts, can help to communicate risks around a central tendency, or point forecast. Density forecasts are useful tools for inflation-targeting central banks as they can be used to quantify the probabilities of key variables being outside a given range in the future. Furthermore, multivariate or joint density forecasts can be useful in predicting the covariances across different variables of interest.

Under a Bayesian estimation framework, constructing density forecasts using a statistical model is straightforward, enabling the various types of uncertainty inherent in forecasts to be incorporated in a coherent fashion. Taking multiple draws from a model's posterior parameter distribution allows for parameter uncertainty in the forecasts. Taking many draws from a model's assumed distributions for shocks can help to characterise an inherently uncertain future. But using a single model may not result in an accurate characterisation of the true degree of uncertainty since the true data-generating process is unknown. Forecast uncertainty due to model uncertainty can also be considered by combining several models.

Date: October 2008, ♠Economic Research, Reserve Bank of Australia, 65 Martin Place, Sydney 2000 NSW, Australia. *Email:* gerardh@rba.gov.au . †CREI and Universitat Pompeu Fabra, Ramon Trias Fargas 25-27, 08005 Barcelona, Spain. *Email:* knimark@crei.cat.

The authors would like to thank Jarkko Jääskelä, Christopher Kent, Adrian Pagan and seminar participants at UNSW and the 2008 Australasian meeting of the Econometric Society for valuable discussions and comments. Responsibility for any remaining errors rests with the authors. The views expressed in this paper are those of the authors and are not necessarily those of the Reserve Bank of Australia.

There is considerable evidence that combining point forecasts from multiple models can improve forecast accuracy, see Timmermann (2006). Much less attention has been paid to combining density forecasts. Some recent work filling this gap includes Kapetanios *et al* (2005), Hall and Mitchell (2004, 2007) and Jore *et al* (2007). While point forecast combinations are usually evaluated according to root mean squared errors (RMSE), evaluating density forecasts is less straightforward. This is primarily because the true density is never observed, so computing even an in-sample measure of the accuracy of a density forecast is less straightforward.

A combined density may have characteristics quite different to those of the individual densities from which it is constructed, as noted by Hall and Mitchell (2004, 2007). For example, the weighted linear combination of two normal densities with different means and variances will be non-normal. So while density forecasts from a combination of models are more flexible than a density constructed from a single model, whether or not the combined density provides a more accurate description of the true degree of uncertainty is, in the end, an empirical question and will depend on the method used to choose weights for individual models when constructing the combined density.

This paper proposes to combine multivariate density forecasts from a suite of models consisting of a Bayesian vector autoregression (BVAR), a factor-augmented vector autoregression (FAVAR) and a dynamic stochastic general equilibrium (DSGE) model. A weighting scheme based on predictive likelihoods following Eklund and Karlsson (2005) and Andersson and Karlsson (2007) is used to combine the models. This weighting scheme also allows for different weights to be assigned to a given model at different forecast horizons. We evaluate the combination forecasts following Diebold *et al* (1998) and Diebold *et al* (1999) by assessing whether the probability integral transform of a series of observations with respect to the density forecasts are uniformly distributed and, in the case of the one-step-ahead forecasts, also independently and identically distributed.

Most of the previous literature on combining density forecasts has focused on univariate densities, that is, density forecasts for a single variable. Yet in many settings it is of interest to characterise the joint probabilities of future outcomes of several variables. For instance, a policy-maker might be interested in the joint probabilities of a target variable and a policy instrument. In this paper, we construct density forecasts of inflation, GDP growth and the cash rate using Australian data. Most central banks have a mandate to control inflation while not causing undue variation in output so these two variables can be viewed as target variables. The combined densities constructed here can thus be used to characterise the joint probability of the target variables and answer questions like "What is the probability that both inflation and GDP growth will be below average 4 quarters from now?" If this probability is deemed too high, actions can be taken to make this outcome less likely. The first step would then be to ask what path of the instrument the density forecasts of the target variables are conditioned on, which motivates the inclusion of the cash rate (the main instrument of monetary policy in Australia) in the set of variables that we construct density forecasts for.

Structural models, similar to the DSGE model included in the suite of models presented here, have become popular at central banks around the world as they can help to tell economically meaningful 'stories' around forecasts. It has also been suggested that DSGE models

are competitive with statistical models in terms of point forecast accuracy, for example by Smets and Wouters (2004) and Adolfson, Andersson *et al* (2005). This paper expands the analysis of these papers by comparing density forecasts constructed using a state-of-the-art DSGE model to density forecasts from purely statistically motivated models.

The rest of the paper is structured as follows. Section 2 outlines the suite of models and describes how they are estimated. Section 3 presents some density forecasts and discusses the motivation for using an out-of-sample-based weighting criteria to combine the models. The predictive-likelihood weighting scheme is outlined here. Section 3 also discusses the trade-offs implied by choosing the lengths of the training and hold-out samples necessary to evaluate an out-of-sample predictive criteria. Section 4 describes the data and the model weights obtained. A univariate and multivariate evaluation of the combined density forecasts is presented in Section 5. The final section concludes.

2. A Suite of Models

We consider three types of models: a BVAR with Minnesota priors, a Factor Augmented Vector Autoregression (FAVAR) and a medium-scale DSGE model. The first two are statistical models with a solid track record in forecasting (see, among others, Litterman 1986, Robertson and Tallman 1999 and Stock and Watson 2002). The structural DSGE model, on the other hand, has a shorter history as a forecasting tool, but is becoming increasingly popular in central banks. All three models restrict the dynamics of the variables of interest in order to avoid in-sample over fitting, which is a well-known cause of poor forecasting performance in unrestricted models (see, for instance, Robertson and Tallman 1999 and references therein). The BVAR shrinks the parameters on integrated variables in an unrestricted VAR towards the univariate random walk model. The FAVAR uses principal components to extract information from a large panel of data in the form of a small number of common factors. The DSGE model uses economic theory to restrict the dynamics and cross-correlations of key macroeconomic time series. Each model uses a different data set, all including a larger number of variables than the three we are interested in, that is, GDP growth, inflation and the cash rate. Each model is briefly presented below along with an overview of how the individual models are estimated.

2.1. **BVAR.** The BVAR can be represented as:

$$y_t^{bvar} = \sum_{i=1}^{p} A_i y_{t-i}^{bvar} + B + \varepsilon_t \tag{2.1}$$

where $\varepsilon_t \sim N(0, \Sigma^{bvar})$, B is a vector of constants and y_t^{bvar} is an $m \times 1$ vector that includes quarterly data on the following variables: trade-weighted measures of G7 output growth, G7 inflation and a simple average of US, euro area and Japanese interest rates; the corresponding domestic variables we are interested in forecasting (GDP growth, trimmed mean underlying inflation and the cash rate); and the level of the real exchange rate. Variables in growth rates are approximated by log differences and foreign variables are treated as exogenous to the domestic variables. We consider three specifications of Equation (2.1) denoted BVAR2, BVAR3 and BVAR4 corresponding to the number of lags $p = 2$, 3 and 4 respectively.

Minnesota-style priors, see Doan *et al* (1984) and Litterman (1986), are imposed on the dynamic coefficients A_i. The prior mean on the coefficient on the first lag of the dependent variable in each equation is set equal to zero for variables in changes and to 0.9 for the three variables specified in levels (both interest rate variables and the real exchange rate). The prior mean on coefficients for all other lags is set to zero and is tighter on longer lags than on short lags. This prior centres the non-stationary domestic and foreign price and output levels on a univariate random walk, and centres domestic and foreign interest rates and the real exchange rate (stationary variables) on AR(1) processes.[1] A diffuse prior is placed on the deterministic coefficients of the unit root processes and the constants of the stationary processes in B and we impose a diffuse prior on the variance covariance matrix of the errors, $p(\Sigma^{bvar}) \propto |\Sigma^{bvar}|^{-(m+1)/2}$.

To draw from the posterior distribution of parameters under this Normal-Diffuse prior, the Gibbs Sampler described in Kadiyala and Karlsson (1997) was used, with the number of iterations set at 10 000 and the first 500 draws used as a burn-in sample to remove any influence of the choice of starting value (the OLS estimate of Σ^{bvar}).

2.2. **FAVAR.** Factor models are based on the idea that a small number of unobserved factors $f_t = \begin{pmatrix} f_{1t} & \dots & f_{kt} \end{pmatrix}'$ can explain much of the variation between observed economic time series. Once estimated, these factors can be included as predictors in an otherwise standard VAR, forming a factor-augmented VAR, see Bernanke *et al* (2005). The FAVAR therefore takes the same form as Equation (2.1), where $y_t^{favar} = \begin{pmatrix} z_t' & f_t' \end{pmatrix}'$ and $z_t \equiv \begin{pmatrix} \Delta gdp_t & \pi_t & i_t \end{pmatrix}'$ is the vector containing the three domestic variables we are interested in forecasting.

Imposing a diffuse prior on the parameters of the model delivers the standard result that the covariance matrix of the errors Σ^{favar} is distributed as inverse Wishart (a multivariate generalisation of the inverse gamma distribution), while the regression coefficients follow a normal distribution, conditional on Σ^{favar}; see, for instance, Kadiyala and Karlsson (1997).

We use principal-components analysis following Stock and Watson (2002) to estimate f_t from a static representation of a dynamic factor model. The static representation is:

$$X_t = \Lambda f_t + e_t \tag{2.2}$$

where X_t represents a large data panel of predictor variables (demeaned and in stationary form), Λ is a matrix of factor loadings and e_t is an error term (with zero mean) which may be weakly correlated across time and series. The $k \times 1$ vector of static factors f_t can include current and lagged values of the dynamic factors. When lags are included in the dynamic factor representation, the static factors in Equation (2.2) are estimated by principal components of a matrix that augments the data panel X_t with lags of the data panel. The principal-component estimator of the factors is $\hat{f}_t = V'X_t$, where V represents the matrix of eigenvectors corresponding to the k largest eigenvalues of the variance covariance matrix of the data panel, see Stock and Watson (2002) and Boivin and Ng (2005).

The data series included in X_t are the same as in Gillitzer and Kearns (2007), except that the three foreign variables used in the BVAR are also included. Up to three static factors

[1]How strongly the overall and cross-equation elements of this prior are imposed is governed by two hyper parameters which are set at 0.5 and 0.2 respectively. Harmonic lag decay is also imposed. For a useful discussion of the Minnesota prior see Robertson and Tallman (1999).

($k = 1, 2, 3$) are included in the model, which are estimated assuming a one-lag dynamic factor model representation. The first three factors explain approximately 25 per cent of the total variation of the data panel. Finally, we allow for either 2 or 3 lags in the FAVAR itself ($p = 2, 3$), which means that in total, six different specifications of the factor model are considered, each denoted FAVARkp.

One drawback of the principal-components approach is that it gives no measure of the uncertainty surrounding the factor estimates $\hat{f}_t$, something that could be important for determining the overall uncertainty surrounding the forecasts. Bai and Ng (2006) show that when the number of series included in the data panel n grows faster than the number of observations T (that is $T/n \to 0$), then the impact from using the estimated regressors $\hat{f}_t$ on the variance of the estimated model parameters is negligible. While this condition is not met here, we expect any influence on the variance of the FAVAR's parameters to be factored into the predictive criteria used to combine the different model forecasts.

2.3. **DSGE Model.** The DSGE model is a medium-scale open economy New Keynesian model that follows the open economy extension of Christiano *et al* (2005) by Adolfson *et al* (2007) closely. It consists of a domestic economy populated with households that consume goods, supply labour and own the firms that produce the goods. Domestic households trade with the rest of the world by exporting and importing consumption and investment goods. Consumption and investment goods are also produced domestically for domestic use. The domestic economy is small compared to the rest of the world in the sense that developments in the domestic economy are assumed to have only a negligible impact on the rest of the world. The model is rich in the number of frictions and shocks, which appears to be important for matching the data.

In order to estimate the model, the structural equations are linearised and the model is then solved for the rational expectations equilibrium. This can be represented as a reduced-form VAR. Since many of the theoretical variables of the model are unobservable, the model is estimated using the Kalman filter. To do this, the solved model is first put in state space form as follows:

$$x_t^{dsge} = F_{dsge} x_{t-1}^{dsge} + u_t^{dsge} \tag{2.3}$$

$$y_t^{dsge} = \mu + H x_t + e_t^{dsge} \tag{2.4}$$

$$\begin{bmatrix} u_t^{dsge} \\ e_t^{dsge} \end{bmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} Q & \mathbf{0} \\ \mathbf{0} & R \end{bmatrix}\right) \tag{2.5}$$

where the theoretical variables are collected in the state vector x_t^{dsge} and the observable variables are collected in the vector y_t^{dsge}. The state transition Equation (2.3) governs the law of motion of the state of the model and the measurement Equation (2.4) maps the state into the observable variables. The matrices F, μ, H and Q are functions of the parameters of the model. The observable variables included in y_t^{dsge} are the real wage, real consumption growth, real investment growth, the real exchange rate, the cash rate, employment, real GDP growth, real exports and real imports growth (adjusted), trimmed mean underlying inflation, real foreign (G7) output growth, G7 inflation and the foreign interest rate. Again, variables in growth rates are approximated by log differences.

The covariance matrix R of the vector of measurement errors e_t^{dsge} in Equation (2.4) is chosen so that approximately 5 per cent of the variance of the observable time series are assumed to be due to measurement errors. The model is estimated using Bayesian methods and the posterior distributions of the 52 structural parameters are simulated using 1 000 000 draws from the Random-Walk Metropolis Algorithm, where the first 400 000 are removed as a burn-in sample. Further details of this model are available on request.

3. Combining the Model Forecasts

While each model included in the suite is estimated using different time series, the three core variables of interest - GDP growth, trimmed mean underlying inflation and the cash rate - are included in the data series for all three. To simplify notation, these three variables are collected in the vector $z_t \equiv \begin{pmatrix} \Delta gdp_t & \pi_t & i_t \end{pmatrix}'$. It is each model's forecasting performance of the joint density $p(z_{t+h}|\Omega_t)$, where Ω_t represents information available at time t, that will be used to combine the models. To simplify notation in what follows we leave out Ω_t and use $p_t(z_{t+h})$ to denote an h step-ahead conditional predictive density but the dependence on the information set available at time t should be remembered.

The three models and how they map into the observable variables that we are interested in can be represented (individually) by a state space system of the form

$$x_{k,t} = F_k x_{k,t-1} + C_k u_{k,t} \tag{3.1}$$

$$z_{k,t} = D_k x_{k,t} + e_{k,t} \tag{3.2}$$

where (3.1) is the state transition equation and (3.2) is the measurement equation. The subscript k is used to index the models and $x_{k,t}$ is the vector of model k's state variables at time t. The matrices F_k and C_k will depend on the functional forms of the models and the estimated model-specific posterior parameter distributions while the matrix D_k maps each model's state variables into the vector of interest $z_{k,t}$.

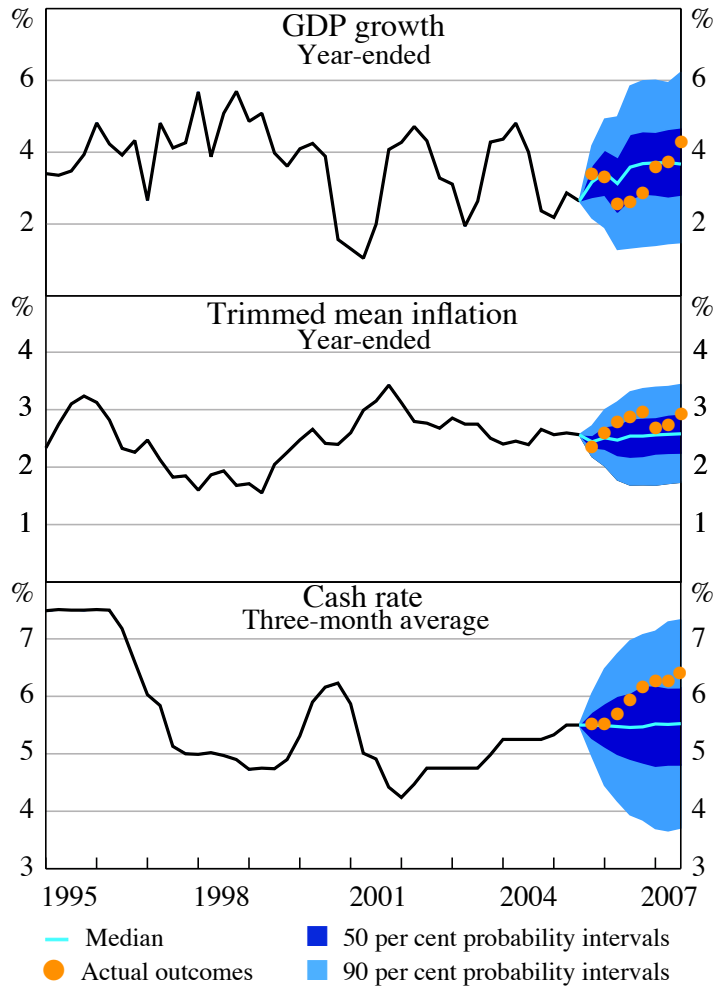
3.1. Constructing Density Forecasts.

The approach to constructing $p_t(z_{k,t+h})$ is similar for each model. Multiple draws are taken from each model's posterior parameter distribution and for each draw j, a potential multivariate realisation $z_{k,t+h}^{(j)}$ is constructed by iterating Equations (3.1) and (3.2) forward up to horizon h. At each iteration, a vector of shocks $u_{k,t+h}^{(j)}$ is drawn from a mean zero normal distribution where the variance is itself a draw from the relevant model's parameter distribution (that is, $u_{k,t+h}^{(j)} \sim N(0, \Sigma_k^{(j)})$). Repeating this procedure many times at each forecast horizon allows us to build up a complete picture of the forecast probability distribution.[2] To complete the density forecast, the potential realisations are ordered at each 'slice' in the forecast horizon. Each ordered set of realisations represents the h step-ahead conditional density forecast for $z_{k,t+h}$. The densities can be represented graphically by shading different probability interval bands in different colors, where each band represents a range in which we expect future realisations of $z_{k,t}$ to fall in with a certain probability. As an example, the density forecasts that would have been obtained using data

[2]The densities forecasts presented in this paper were all constructed using 1000 draws.

up to 2005:Q3 with the BVAR2 model are presented in Figure 1. The median projection along with 50 and 90 per cent probability intervals are shown.

FIGURE 1. BVAR2 Density Forecasts
2005:Q4-2007:Q3



A combination density forecast, denoted $p_t^c(z_{t+h})$, can be constructed as a weighted linear combination (or 'linear opinion pool') of the competing model forecasts:

$$p_t^c(z_{t+h}) = \sum_{k=1}^{K} p_t(z_{k,t+h}) w_{k,h} \tag{3.3}$$

where $w_{k,h}$ represents the weight assigned to model k when forecasting at horizon h. The remainder of this section focuses on how to go about choosing these weights.

3.2. **Equal weights.** The simplest and most straightforward weighting scheme is to put equal weight on all models in the suite. In this case, $w_{k,h} = 1/K$ at each forecast horizon. Apart from its simplicity, *a priori* this approach seems to have little going for it. For example,

an over-parameterised model that forecasts poorly would still be assigned substantial weight.[3] But Timmermann (2006) has shown that such a scheme performs well when combining point forecasts and could also prove useful in a density combination context. One reason for this unexpected success may be that an equal-weighting scheme is robust to possible small-sample problems that may arise when choosing weights 'optimally'.

3.3. **Posterior Probability Weights.** An alternative and intuitive approach to combining models can be derived in a Bayesian framework. Each model's marginal likelihood, $p(\mathbf{y}_k)$, could be used to generate posterior probability weights, a method known as Bayesian Model Averaging; see, for example, Raftery *et al* (1997). That is, the weight of model k would be given by

$$w_k = \frac{p(\mathbf{y}_k)p(M_k)}{\sum_{i=1}^{K} p(\mathbf{y}_i)p(M_i)} \tag{3.4}$$

where $p(M_k)$ represents any prior beliefs about model k being the true model.

This method is attractive as models that appear to describe the observed data better are assigned a higher weight. But a potential problem with using an in-sample measure to generate model weights is that too much weight may be placed on over-parameterised models with good in-sample fit even if they perform poorly when forecasting.[4]

A further issue is that the marginal likelihood reflects the entire fit of a model. The weights from Equation (3.4) will depend upon each model's description of all the variables making up $\mathbf{y}_k$, but $\mathbf{y}_k$ differs between models.

Another approach that can be used to help control for in-sample over-fitting, and to focus on the key variables of interest, is an out-of-sample weighting scheme based on predictive likelihoods, as demonstrated in a univariate application by Andersson and Karlsson (2007) and Eklund and Karlsson (2005). Here we extend their approach to a multivariate setting.

3.4. **Predictive-likelihood Weights.** A weighting scheme based on predictive likelihoods requires the available data to be split into two parts. A training sample is used to estimate the parameters of each model, and the remaining hold-out sample is used to evaluate each model's out-of-sample forecasting performance. Asymptotically, that is, with an infinitely long hold-out sample, predictive likelihoods would tend to put all the weight on the best model. In practice, however, there is a trade-off between the length of training and hold-out samples. With a short training sample, a model's parameters will be imprecisely estimated. But lengthening the training sample necessarily shortens the hold-out sample, which makes the evaluation of the predictive criteria less precise. Therefore, in small samples, a poor model may still be assigned substantial weight. Worse still, if there are several poor models, their combined weight can be large.

[3]Also, models that were quite similar would tend to be 'over-represented.'

[4]While it is possible to view the marginal likelihood as an out-of-sample measure, this interpretation relies on the predictive content of the prior (see, for example, Adolfson, Lindé and Villani 2005 and Eklund and Karlsson 2007). This will only be true for the DSGE model in our suite of models and, in that case, the marginal likelihood is likely to be sensitive to the choice of prior. For both the BVAR and FAVAR models, where either diffuse or relatively uninformative priors are imposed, the marginal likelihood reflects an in-sample measure of fit.

As in Andersson and Karlsson (2007), we calculate a series of small hold-out sample predictive likelihoods (PL), as shown in Equation (3.5). This involves a recursive forecasting scheme where the training sample of initial size l is expanded throughout the forecasting exercise.[5] We also restrict our attention to each model's predictive performance of the subset of variables $z_{k,t}$ as set out in Equation (3.2).

$$\mathrm{PL}_{k,h} = p(\mathbf{z}_{k,h}^{hold\text{-}out}|\mathbf{y}_k^{training}) = \prod_{t=l}^{T-h} \hat{p}(z_{k,t+h}|\mathbf{y}_{k,t}) \tag{3.5}$$

In Equation (3.5), $\mathbf{z}_{k,h}^{hold\text{-}out}$ denotes the $(T-h-l)$ hold-out observations used to evaluate model k at horizon h, $\mathbf{y}_k^{training}$ represents the (expanding) training sample and $\mathbf{y}_{k,t} = \begin{pmatrix} y_{k,1} & \dots & y_{k,t} \end{pmatrix}'$ represents each individual training sample relevant to iteration t in the recursive forecasting exercise. To calculate Equation (3.5), for each model k we use the multivariate normal distribution and take an average across multiple draws (j) from model k's predictive distribution of $z_{k,t+h}$. That is,

$$\hat{p}(z_{k,t+h}|\mathbf{y}_{k,t}) = n^{-1} \sum_{j=1}^{n} p(z_{k,t+h}^{(j)}|\mathbf{y}_{k,t}) \tag{3.6}$$

where $n = 500$ and $p(z_{k,t+h}^{(j)}|\mathbf{y}_{k,t})$ is multivariate normal. This is the same approach used by Andersson and Karlsson (2007).

The predictive-likelihood weights can be calculated by replacing the marginal likelihood in Equation (3.4) with the predictive likelihood of Equation (3.5) as follows:

$$w_{k,h} = \frac{\mathrm{PL}_{k,h} p(M_k)}{\sum_{i=1}^{K} \mathrm{PL}_{i,h} p(M_i)}. \tag{3.7}$$

In the analysis below, we assign an equal prior probability to each model being the true model, that is, $p(M_k) = 1/K$.[6]

4. Implementation and Results

In this section we describe the data and present the results of the weighting scheme. We also attempt to shed some light on the importance of the covariances of the forecasts errors for the weights assigned to each model as well as compare how a model ranking based on

[5]Theoretically, either a fixed- or rolling-window forecasting scheme would be preferred to accommodate the idea that the hold-out sample should tend towards infinity. With dynamic models, however, a fixed estimation window is not suitable as forecasts would lack information available at the time of forecasting. The rolling-window scheme is also not practical when faced with a short sample of data. We therefore prefer the recursive approach.

[6]We also generated weights numerically following Hall and Mitchell (2007) when choosing the set of weights that minimise the Kullback-Leibler divergence between the combined density forecast and the true but unknown density. When considering a small number of models, the weights obtained were similar to those of the predictive-likelihood approach, but this Kullback-Leibler information criterion weighting scheme, which involves a numerical search for the optimal set of weights, becomes impractical when considering a larger model space.

point forecast accuracy compare to the ranking implied by the weights from the predictive likelihood scheme.

4.1. **The Sample.** The data sample is from 1992:Q1 to 2007:Q3 and as discussed above, a recursive out-of-sample forecasting scheme was used to evaluate each model and generate model weights. The first $l = 36$ observations (1992:Q1-2000:Q4) were used as the initial training sample to estimate each model before constructing density forecasts up to eight quarters ahead (2001:Q1-2002:Q4). The training sample was then extended by one observation, the models re-estimated and density forecasts over the next eight quarters (2001:Q2-2003:Q1) constructed. Model weights were generated sequentially by repeating this exercise over the remaining sample. The final set of model weights were based on a hold-out sample of 27 observations at the one-step-ahead forecast horizon and 19 observations at the eight-step-ahead horizon (the final training sample was between 1992:Q1 and 2007:Q2, which allows one one-step-ahead forecast to be compared to the final observation). It should be noted that the DSGE model was only re-estimated every four quarters to save on computation time.

4.2. **Model Weights.** Table 1 shows the final set of model weights (using all observations in the hold-out sample) according to the predictive-likelihood weighting scheme when forecasting one, two, four and eight quarters ahead. Note that a simple unrestricted VAR of GDP growth, trimmed mean inflation and the cash rate (with two to four lags) was also included in the weighting scheme as a benchmark model. How the predictive-likelihood weights evolve throughout the hold-out sample (as the number of observations used to construct the weights increase) is shown in Figure 2 for the BVAR2, FAVAR12 and DSGE models.

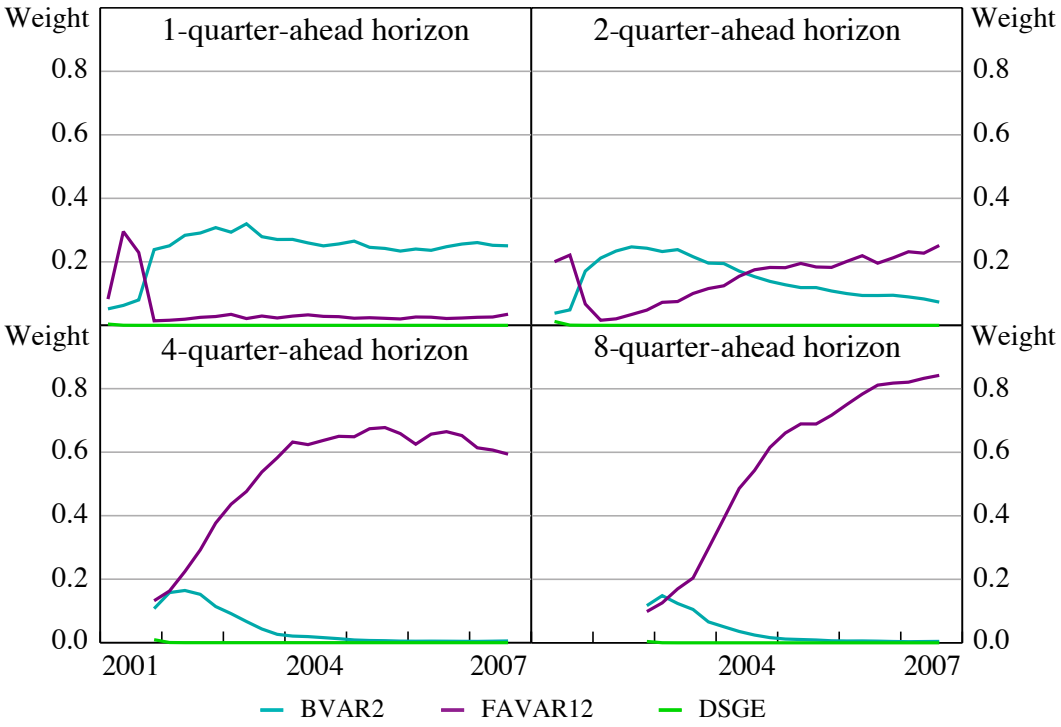
Table 1: Predictive Likelihood Weights at 2007:Q3

	Forecast horizon (h)			
	1	2	4	8
BVAR4	0.22	0.32	0.00	0.00
BVAR3	0.18	0.11	0.00	0.00
BVAR2	0.25	0.07	0.01	0.00
FAVAR33	0.00	0.00	0.00	0.00
FAVAR32	0.00	0.00	0.02	0.01
FAVAR23	0.00	0.00	0.00	0.00
FAVAR22	0.00	0.18	0.35	0.13
FAVAR13	0.00	0.00	0.02	0.00
FAVAR12	0.03	0.25	0.59	0.84
DSGE	0.00	0.00	0.00	0.00
VAR4	0.02	0.04	0.00	0.00
VAR3	0.30	0.02	0.00	0.00
VAR2	0.00	0.00	0.00	0.00

Notes: Weights based on maximum length (27 observations) of hold out sample.

Table 1 shows that, for forecasting at short horizons, the BVAR model is assigned the most weight. The BVAR specified with 2 lags is preferred at the one-quarter-ahead forecast horizon while 4 lags are preferred at the two-quarter-ahead horizon. The benchmark VAR

FIGURE 2. Predictive-likelihood Weights

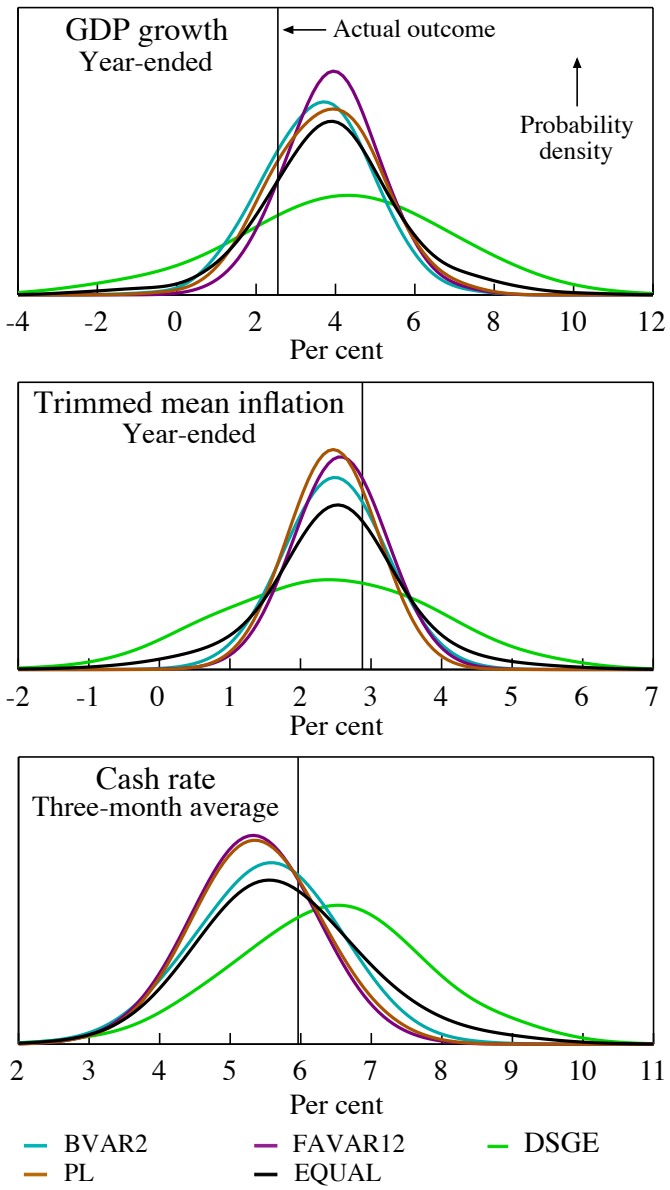


model specified with 3 lags also receives substantial weight at the one-quarter-ahead forecast horizon. When forecasting at intermediate and longer horizons, the FAVAR model is assigned the majority of the weight. The FAVAR specified with two lags and either one or two factors seems to do the best. Interestingly, the DSGE model is assigned zero weight at all forecast horizons. This result can be attributed to the large forecast error variance implied by the DSGE model. As we will see in Section 5, the DSGE model's density forecasts were typically too wide (the true degree of uncertainty was over-estimated) when compared with the distribution of actual observations in the sample.

As discussed above, combining weights 'optimally' could be troublesome when faced with a small sample of data. Therefore, we also consider an equal-weighting scheme, which assigns a one-third weight ($w_{k,h} = 1/3$) to the BVAR2, FAVAR12 and DSGE models at all horizons and in all time periods. While an equal-weighting scheme across all 13 models shown in Table 1 could have been used, that approach would tend to overweight the FAVAR models (of which there are 6 different specifications) and underweight the others, in particular the DSGE model.

To give a flavour of what the combination density forecasts may look like in a given quarter, Figure 3 shows a cross-section of the four-quarter-ahead density forecasts that would have been made in 2005:Q3 when using the predictive-likelihood and equal-weighting schemes (using weights that would have been available at the time of forecasting). The individual BVAR2, FAVAR12 and DSGE model forecasts are also shown to give an idea of how the combination forecasts differ.

FIGURE 3. Four-quarter-ahead Density Forecasts
Made in 2005:Q3



Looking at Figure 3, a couple of points are worth making. The predictive-likelihood combination density forecasts are typically similar to the FAVAR12 model's density forecasts since that model is given a large weight at the four-quarter-ahead forecast horizon. It is also clear that the DSGE model's density forecasts are characterised by a much larger degree of uncertainty than is the case with the other models or the combination density forecasts. Finally, the equally weighted combination density forecast for the cash rate has a 'fat' right-hand tail, suggesting that in 2005:Q3, according to this forecast, the risks to the central cash rate projection in four quarters time were somewhat skewed to the upside.

4.3. **Comparing multivariate and univariate weights.** The predictive likelihood scheme above rewards models that accurately describe the covariances of the forecast errors across variables. To shed some light on the importance of these covariances for the weights assigned above, we can compare the weights from the multivariate predictive likelihood to weights calculated from predictive likelihoods of density forecasts for the individual variables in z_t. This exercise suggests that when predicting single variables, a larger number of models receive more weight. For instance, there is no model that receives more weight than 20 per cent for forecasting GDP growth only at the 8 quarter horizon, but there are 5 models that receive a weight of more than 10 per cent. As can be seen in Table 1, the multivariate predictive likelihood weights are much more concentrated and the FAVAR12 is assigned a weight of 84 per cent at the 8 quarter horizon. This may at first suggest that the covariances are important for the weights assigned. However, a more careful analysis suggest that this is not the case. By computing the predictive likelihoods by stacking individual variable forecasts and imposing zero error covariances, the weights become very similar to those of the full multivariate exercise. This suggests that it is the fact that the same models tend to produce accurate density forecasts for all variables that lead to the concentration of weights in the multivariate case, not that some models are better at predicting the covariances of the forecast errors across variables. In general, these covariances turn out to be small, which explains the similarity of the results to when a zero covariance across variables is imposed.

4.4. **Point and density forecast performance.** The predictive likelihood is not an absolute measure of forecasting performance, but rather, it is a measure of forecasting accuracy relative to the variance implied by the model (see Andersson and Karlsson 2007 and Eklund and Karlsson 2007). This makes the predictive likelihood appealing when evaluating density forecasts from different models, although the ranking of models could be quite different to that obtained according to RMSEs based on point forecasts. For the model suite used here, this is indeed the case. Table 2 below reports the RMSE at different horizons for each variable for the BVAR2, FAVAR12 and DSGE models.

The BVAR2 outperforms the FAVAR12 in terms of RMSE for all variables and at all horizons. Still, the FAVAR12 is assigned a weight of 84 per cent at the 8 quarter horizon. Similarly, the DSGE has the lowest forecast RMSE of all models for GDP growth at the 8 quarter horizon. Still, it is assigned zero weight both by the multivariate predictive likelihood and the variable-by-variable predictive likelihood. For a policy maker concerned with risks or uncertainty surrounding forecasts, looking at historical RMSEs may thus be quite misleading.

Table 2: RMSE results

	Forecast horizon (h)			
Model/Variable	1	2	4	8
GDP Growth				
BVAR2	0.41	0.43	0.45	0.46
FAVAR12	0.55	0.46	0.46	0.48
DSGE	0.54	0.46	0.45	0.44
Trimmed-mean inflation				
BVAR2	0.16	0.16	0.13	0.14
FAVAR12	0.18	0.18	0.14	0.14
DSGE	0.18	0.21	0.19	0.16
Cash Rate				
BVAR2	0.23	0.39	0.54	0.51
FAVAR12	0.26	0.47	0.62	0.60
DSGE	0.39	0.55	0.73	0.77

Notes: Results are calculated over the sample 2001:Q1 to 2007:Q3. RMSE is the Root Mean Squared Error between the series of model forecasts and subsequent actual outcomes in percentage points.

5. Evaluating Density Forecasts

Accuracy is obviously a desirable feature of forecasts. For point forecasts, accuracy is usually interpreted to mean that the forecast errors are unbiased and small according to RMSEs. For density forecasts, accuracy can be interpreted in a statistical sense by comparing the distribution of observed data with the forecast distribution. Given a large enough sample of data, if a density forecast is providing an accurate characterisation of the true degree of uncertainty, that is, it provides an accurate description of reality, then we would expect observations to fall uniformly across all regions of the distribution that are forecast to contain the same probability density. As an example, if a density forecast suggests there is a 10 per cent chance of GDP growth falling between 3.5 and 3.7 per cent at a given forecast horizon, then, if economic conditions at the time of forecasting could be replicated 100 times, we would expect 10 actual observations to fall between 3.5 and 3.7 per cent. Diebold *et al* (1998) employ this result to formally evaluate density forecasts; an approach that avoids both the need to specify the unknown true density and the need to specify a loss function for the user of the forecasts.

5.1. **Probability Interval Transforms.** Diebold *et al*'s (1998) approach to evaluating univariate density forecasts is based on the probability integral transform (pit) of a sequence of n univariate observations $\{y_{t+h}\}_{t=1}^{n}$, with respect to the h-step-ahead density forecasts $\{p_t(y_{t+h})\}_{t=1}^{n}$. Each of the transformed observations or pits $\{v_{t+h}\}_{t=1}^{n}$ reflects the probability (according to the density forecast) that an outcome y_{t+h} will be less than or equal to what

was actually observed. That is,

$$v_{t+h} = \int_{-\infty}^{y_{t+h}} p_t(u)du, \ t = 1, \ldots, n \tag{5.1}$$

Equation (5.1) links where an actual observation falls relative to the percentiles of the forecasted distribution. For example, an actual observation that falls at the median of the density forecast would receive a pit value of 0.5. For an observation that falls in the upper tail, say at the 90$^{\text{th}}$ percentile, the pit value would be 0.9. If a sequence of density forecasts coincides with the true data-generating process, then the sequence of pits $\{v_{t+h}\}_{t=1}^{n}$ will be uniform $U(0,1)$ and in the case where $h = 1$, $\{v_{t+h}\}_{t=1}^{n}$ are both $U(0,1)$ and independently and identically distributed (iid). In other words, if the density forecasts are not misspecified, over a large enough sample, realisations should fall over the entire range of the forecasted density and with a probability equal to the probability specified in the density forecast.

Diebold *et al* (1999) show that the probability integral transform approach to evaluating density forecasts can be extended to the multivariate case.[7] Let $p_t(z_{t+h})$ again denote a joint density forecast of the 3×1 vector of interest $z_{t+h} = (\ z_{1,t+h} \quad z_{2,t+h} \quad z_{3,t+h}\ )'$ made at time t and suppose we have n such forecasts and n corresponding multivariate realisations. After factoring the joint density into the product of conditional densities,

$$p_t(z_{t+h}) = p_t(z_{3,t+h}|z_{2,t+h}, z_{1,t+h})p_t(z_{2,t+h}|z_{1,t+h})p_t(z_{1,t+h}) \tag{5.2}$$

the probability integral transform for each variable in the multivariate realisations can be taken with respect to the corresponding conditional distribution. This creates a set of three pit sequences, each of length n. If the joint density forecasts correspond to the true conditional multivariate density, then these three transformed sequences will each be $U(0,1)$, as will the $3n \times 1$ vector formed by stacking the individual sequences. As before, in the one-step-ahead case they will also be iid. Since the joint density in Equation (5.2) can be factored in six ways, there are six equivalent pit sequences that can be used to evaluate the multivariate density forecasts.[8]

So evaluating density forecasts can effectively be reduced to testing whether an observed series is $U(0,1)$, and in the case of the one-step-ahead forecasts, whether it is also iid. Before presenting the results, it must be highlighted that in the current context there are reasons why tests of uniformity and independence may be unreliable and it would be unwise to over-emphasise the results from these tests. Given the small sample of data on which we can evaluate the forecasts, it will always be difficult to distinguish between forecasting ability and luck. Also, as Hall and Mitchell (2007) and others have noted, the way in which dependence in the forecasts affects tests for uniformity is unknown (as is the impact of non-uniformity for tests of independence). And given that serially dependent forecasts are entirely consistent with correctly-specified density forecasts at a forecast horizon greater than one-step-ahead (see Elder *et al* 2005 for a good discussion of this point), results must be treated with some caution. In addition, formal testing of the densities presented in this paper is further

[7] See also Clements and Smith (2000) for an application of the multivariate pit approach.

[8] In the results that follow, the multivariate evaluation was based on factoring the joint density of z_t as follows: $p_t(z_{t+h}) = p_t(\Delta gdp_{t+h}|\pi_{t+h}, i_{t+h})p_t(\pi_{t+h}|i_{t+h})p_t(i_{t+h})$.

complicated by the fact that we allow for parameter uncertainty when constructing the forecasts.

5.2. **A Visual Assessment.** We present a visual assessment of the hypothesis that the pit-values corresponding to the one-quarter-ahead density forecasts are uniformly distributed in Figure 4. The results for the two- and four-quarter-ahead forecasts are provided in Appendix A (a visual assessment at longer forecast horizons is difficult due to the small number of observations available to evaluate the forecasts). This method is widely used in the literature and may also prove revealing as to how the density forecasts are misspecified. We conduct both a univariate and multivariate evaluation of the BVAR2, FAVAR12 and DSGE models, as well as the two combined density forecasts based on the predictive-likelihood and equal-weighting schemes.

Since a number of observations are used up when calculating the predictive-weighting criteria, the effective sample on which we can evaluate the combined densities is reduced. To evaluate the combined one-quarter-ahead density forecasts, 26 observations were available, while only 12 observations could be compared to the combined eight-quarter-ahead density forecasts.[9] To allow for a fair comparison with the predictive-likelihood weighting scheme, this reduced evaluation sample was also used to evaluate the equal-weighting scheme as well as the models individually.

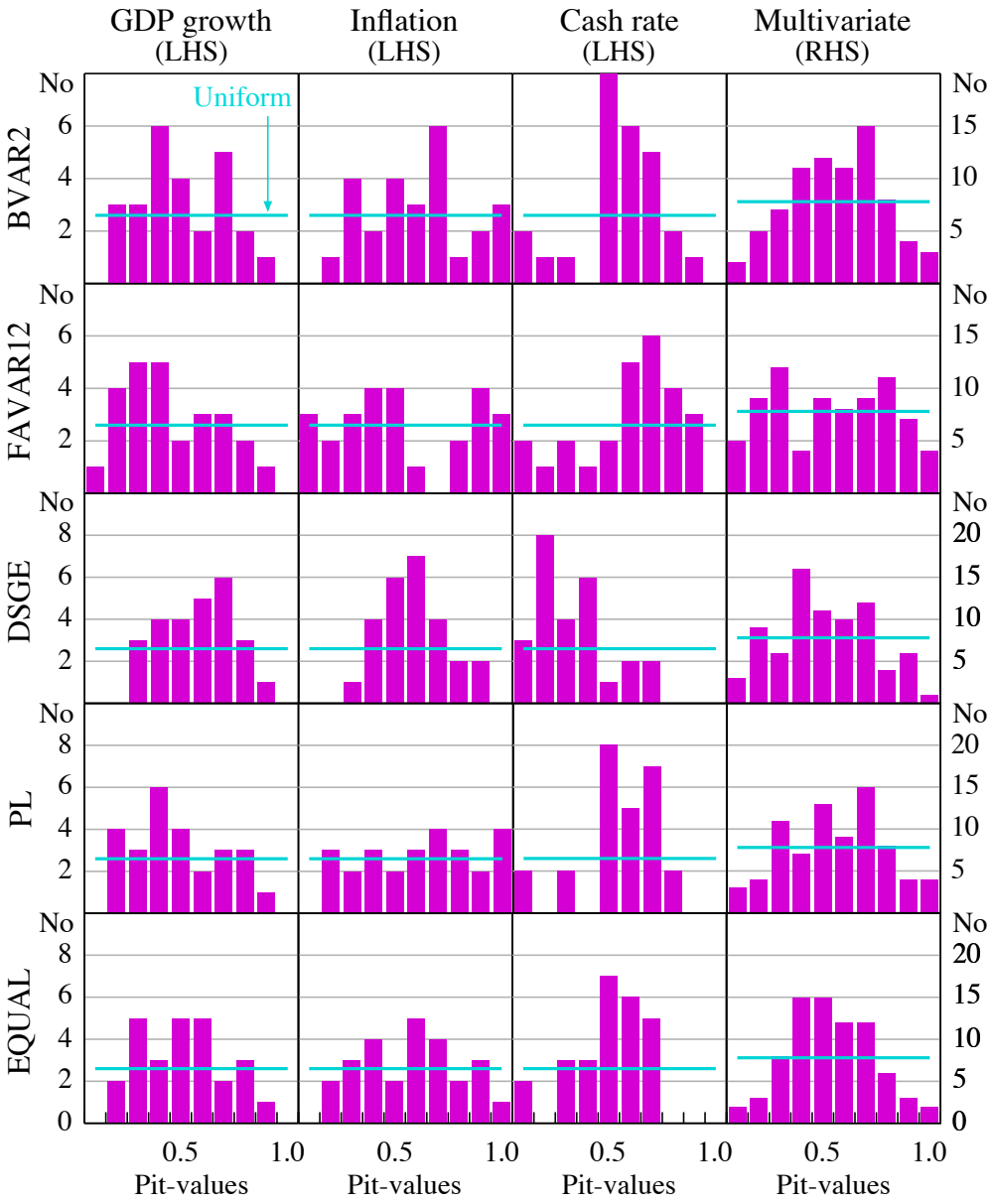
In Figures 4, 5 and 6, the horizontal line represents the theoretical distribution that pit-values would follow in the case of correctly specified density forecasts. The closer the sample histogram is to this $U(0,1)$ distribution, the better the density forecast. A hump-shaped histogram would be suggestive of density forecasts that are over-estimating the true degree of uncertainty, with too many pit-values close to 0.5 (a result of too many actual observations falling around the centre of the density forecasts over time). A histogram with peaks near 0 and 1, on the other hand, would suggest too small a probability is being assigned to outcomes in the tails of the forecasted distribution.

Some broad conclusions can be taken from the figures. It seems clear that the distributions of pit-values corresponding to the DSGE model's forecasts (the third row in each of the figures) violate the uniformity hypothesis. For both the univariate and multivariate cases, over the evaluation period, the DSGE model's density forecasts were too wide when compared to the actual distribution of observations. The hump-shaped distribution of pit-values is particularly evident at the two- and four-quarter-ahead forecast horizons (Figures 5 and 6).

Looking at the univariate cases (the first three columns in each figure) it appears that, across the different models and weighting schemes, the density forecasts for inflation perform the best. Apart from the DSGE model, the distribution of pit-values for the inflation forecasts show a reasonable coverage in the tails of the distribution, with the overall distribution typically close to the $U(0,1)$ line. The distribution of the cash rate variable seems to be

[9]To see this, consider the sequence of eight-quarter-ahead combined density forecasts. The first such forecast can only be made once the first set of eight-quarter-ahead weights are constructed (which is in 2002:Q4). And being an eight-quarter-ahead forecast, it is evaluated against the 2004:Q4 observation. A second eight-quarter-ahead forecast can be made in 2003:Q1 (using an updated set of eight-quarter-ahead weights) and evaluated in 2005:Q1. This pattern continues until the sample is exhausted, which occurs after 12 eight-quarter-ahead forecasts are made.

FIGURE 4. Pit Distributions
One-quarter-ahead horizon



Notes: Rows in the figure refer to the three individual model density forecasts and the two combination density forecasts. The first three columns refer to the pit-values corresponding to the univariate one-quarter-ahead density forecasts for GDP growth, inflation and the cash rate. The final column refers to the multivariate forecasts where the histogram is constructed using the 'stacked' sequence of pit-values as described in the main text. The height of each bin (vertical axis) reflects the number of observations that fell within different percentile bands (horizontal axis) over the evaluation period (26 observations in total in the univariate cases and 26 observations for each of the three variables in the multivariate case).

the most poorly forecast across the various methods. Turning to the multivariate cases, it seems that the FAVAR12 model provides the best description of the joint distribution of GDP growth, inflation and interest rates over the evaluation period. This seems to be true at each forecast horizon. The combination density forecasts constructed using the predictive-likelihood weights also perform well, although it is not clear that the combination density performs that much better than the individual FAVAR12 model's forecasts. There is perhaps some evidence that the optimally combined density forecasts outperform those based on an equal-weighting scheme, although this is most likely due to the poor performance of the DSGE model's density forecasts, which receive one-third weight in the equal-weighting scheme.

5.3. **Formal Tests.** Formal statistical tests of the uniformity hypothesis have also been suggested.[10] For example, Berkowitz (2001) suggests taking a further transformation using the standard normal inverse cumulative density function to convert the test for uniformity into a more powerful test for normality. In Appendix A we present a variation of the Berkowitz-type test for normality which allows serial correlation in the forecasts (see Elder *et al* 2005). While the test delivers broadly the same conclusion as the visual assessment, given the difficulties faced when assessing the uniformity (or normality) hypothesis discussed earlier, the results should still be treated with some caution.

To test the hypothesis that the pit-values corresponding to the one-quarter-ahead density forecasts are iid, Ljung-Box (LB) tests for up to fourth-order serial correlation are shown in Table 3. LB tests on the first three moments were considered to allow for the possibility of higher-order dependence. Except for the univariate density forecasts for inflation, the tests do show evidence of serial correlation. This suggests that the GDP growth, cash rate and multivariate one-quarter-ahead density forecasts are misspecified to some extent. Taking the multivariate evaluation as an example, the LB tests show dependence in the stacked sequence of pit-values in all of the first three moments when forecasting with the FAVAR12 model. The BVAR2 model seems to fare better, although there is evidence of serial correlation in the second moment. Similarly, pit-values corresponding to the predictive-likelihood and equal-weighting scheme combination density forecasts show evidence of serial correlation in the second moment, which is inconsistent with the hypothesis of correctly-specified density forecasts at the one-step-ahead forecast horizon.

[10]Corradi and Swanson (2006a) provide a detailed summary. See also Hall and Mitchell (2004) for an application of the various testing procedures to density forecasts of UK inflation.

Table 3: Ljung-box Tests for Independence
One-quarter-ahead forecast horizon

	GDP growth			Inflation			Cash rate			Multivariate		
Model/Moment	1	2	3	1	2	3	1	2	3	1	2	3
BVAR2	**0.07**	**0.36**	**0.04**	0.78	0.70	0.95	0.66	**0.04**	**0.08**	0.90	**0.00**	0.66
FAVAR12	0.19	0.62	0.39	0.20	0.56	0.20	**0.01**	0.21	**0.03**	**0.00**	**0.06**	**0.01**
DSGE	0.77	0.79	0.55	0.85	0.99	0.96	0.14	0.99	0.25	**0.00**	0.27	**0.07**
PL	**0.09**	**0.08**	**0.10**	0.94	0.76	0.95	0.94	**0.03**	0.39	0.91	**0.00**	0.87
EQUAL	0.21	0.41	0.21	0.94	0.84	0.95	0.75	0.11	0.34	0.58	**0.00**	0.56

Note: Numbers in the table are p-values corresponding to Ljung-Box tests of up to fourth-order serial correlation in the pit-values. Numbers in bold indicate that the null hypothesis is rejected at the 10 per cent significance level.

Overall, based on these results, it is hard to draw strong conclusions about the accuracy of the combined density forecasts. But one result that does seem clear is that the density forecasts constructed using the DSGE model were inconsistent with the data; the density forecasts were too wide when compared with the actual distribution of observations. One possible reason for the large forecast uncertainty implied by the DSGE model could be the many restrictions imposed on the dynamics of the model. If the data 'disagree' with these restrictions, larger shocks will be needed to explain the patterns seen in the data and, as a consequence, greater shock uncertainty will be introduced into the forecasts. So while DSGE models have been shown to produce relatively accurate point forecasts (see, for example, Adolfson, Andersson *et al* 2005), our results suggest they may be less successful at characterising the uncertainty surrounding point forecasts. However, this does not mean that density forecasts from DSGE models are not useful for policy analysis. As structural models with economically interpretable state variables, DSGE models still have the advantage of lending themselves to scenario analysis and 'story telling'; something that purely statistical models cannot do. This is equally true for density forecasts as it is for point forecasts.

6. Conclusion

In this paper, we have looked at a relatively unexplored area of the forecast combination literature, that of combining multivariate density forecasts. We have used predictive-likelihood scores to combine density forecasts produced by a suite of models consisting of a BVAR, a FAVAR and a DSGE model. The weighting scheme suggests that the DSGE model should be assigned a very low weight in the combined density forecast. Inspecting the probability integral transforms of the models' forecasts suggests that this low weight is due to the fact that over the evaluation sample the DSGE produced density forecasts that were too wide when compared with the actual distribution of observations.

We also performed both a visual and formal assessment of the performance of the density forecasts using probability integral transforms that should produce uniform distributions if the uncertainty characterised by the forecasts is correctly modelled. Overall, this exercise

returned mixed results, and it is not clear that the combined forecasts are superior to those of the best-performing individual model or an equal-weighting scheme. This may be a result of the short sample available to evaluate the forecasts. However, broadly, the evaluation exercise suggested that individual models that received a large weight in the combined density outperformed models that received a low weight.

We also find that a model that performs well in terms of point forecast accuracy does not necessarily produce the most accurate density forecasts. For instance, the BVAR2 dominates the FAVAR12 in terms of forecast RMSE at all horizons, and yet, the density forecast from the FAVAR12 is assigned a much larger weight in the optimally combined density forecast at horizons longer than 1 quarter ahead. In addition, the DSGE is competitive with the other models in terms of point forecast accuracy when forecasting output growth. Still, it is assigned essentially zero weight in the combined density forecasts. This suggests that while DSGE models may be useful for constructing point forecasts, they are not (yet) competitive with statistically motivated models in terms of characterising forecast uncertainty.

References

[1] Adolfson M, J Linde and M Villani (2005), 'Forecasting Performance of an Open Economy Dynamic Stochastic General Equilibrium Model', Sveriges Riksbank Working Paper No 190.

[2] Adolfson M, MK Andersson, J Linde, M Villani and A Vredin (2005), 'Modern Forecasting Models in Action: Improving Macroeconomic Analyses at Central Banks', Sveriges Riksbank Working Paper No 188.

[3] Adolfson M, S Lasen and M Villani (2007), 'Bayesian Estimation of an Open Economy DSGE Model with Incomplete Pass-Through', Journal of International Economics, 72(2), pp 481-511.

[4] Andersson MK and S Karlsson (2007), 'Bayesian Forecast Combination for VAR Models', Sveriges Riksbank Working Paper No 216.

[5] Bai J and S Ng (2006), 'Confidence Intervals for Diffusion Index Forecasts and Inference for Factor-Augmented Regressions', Econometrica, 74(4), pp 1133-1150.

[6] Berkowitz J (2001), 'Testing Density Forecasts, with Applications to Risk Management', Journal of Business and Economic Statistics, 19(4), pp 465-474.

[7] Bernanke BS, J Boivin and PS Eliasz (2005), 'Measuring the Effects of Monetary Policy: A Factor-Augmented Vector Autoregressive (FAVAR) Approach', Quarterly Journal of Economics, 120(1), pp 387-422.

[8] Boivin J and S Ng (2005), 'Understanding and Comparing Factor-Based Forecasts', International Journal of Central Banking, 1(3), pp 117-151.

[9] Christiano LJ, M Eichenbaum and CL Evans (2005), 'Nominal Rigidities and the Dynamic Effects of a Shock to Monetary Policy', Journal of Political Economy, 113(1), pp 1-45.

[10] Clements MP and J Smith (2000), 'Evaluating the Forecast Densities of Linear and Non-Linear Models: Applications to Output Growth and Unemployment', Journal of Forecasting, 19(4), pp 255-276.

[11] Corradi V and NR Swanson (2006), 'Chapter 5: Predictive Density Evaluation', in G Elliott, CWJ Granger and A Timmermann (eds), Handbook of Economic Forecasting, Volume 1, Elsevier, Amsterdam, pp 197-284.

[12] Diebold FX, TA Gunther and AS Tay (1998), 'Evaluating Density Forecasts with Applications to Financial Risk Management', International Economic Review, 39(4), pp 863-883.

[13] Diebold FX, J Hahn and AS Tay (1999), 'Multivariate Density Forecast Evaluation and Calibration in Financial Risk Management: High-Frequency Returns on Foreign Exchange', Review of Economics and Statistics, 81(4), pp 661-673.

[14] Doan T, RB Litterman and CA Sims (1984), 'Forecasting and Conditional Projection Using Realistic Prior Distributions', Econometric Reviews, 3(1), pp 1-100.

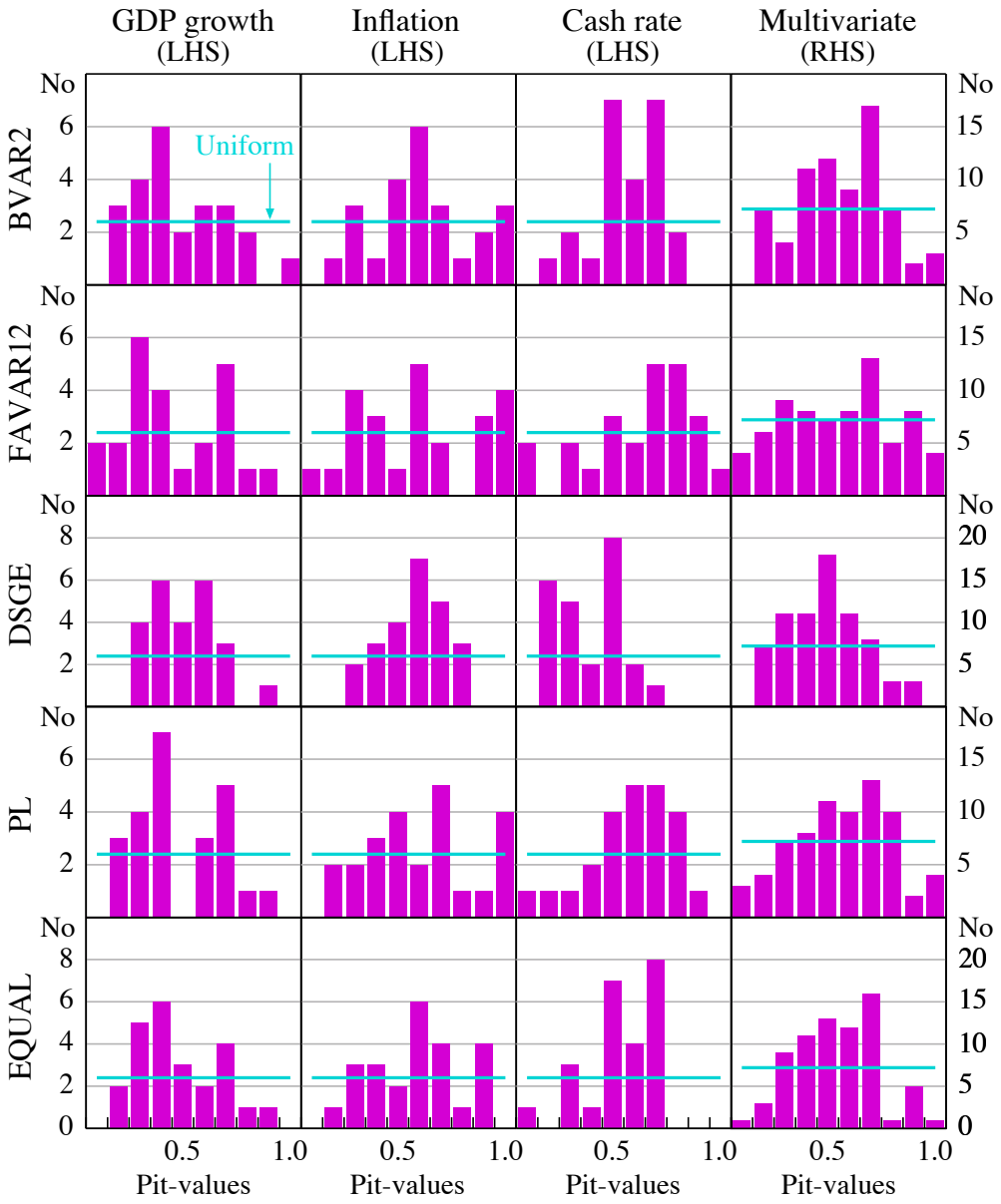
[15] Eklund J and S Karlsson (2007), 'Forecast Combination and Model Averaging Using Predictive Measures', Econometric Reviews, 26(2-4), pp 329-363.
[16] Elder R, G Kapetanios, T Taylor and T Yates (2005), 'Assessing the MPC's Fan Charts', Bank of England Quarterly Bulletin, Autumn, pp 326-345.
[17] Gillitzer C and J Kearns (2007), 'Forecasting with Factors: The Accuracy of Timeliness', RBA Research Discussion Paper No 2007-03.
[18] Hall SG and J Mitchell (2004), 'Density Forecast Combination', National Institute of Economic and Social Research Discussion Paper No 249.
[19] Hall SG and J Mitchell (2007), 'Combining Density Forecasts', International Journal of Forecasting, 23(1), pp 1-13.
[20] Jore AS, J Mitchell, J Nicolaisen and SP Vahey (2007), 'Combining Forecast Densities from VARs with Uncertain Instabilities', Paper presented to the Research Workshop on Monetary Policy in Open Economies, Reserve Bank of Australia, Sydney, 17-18 December.
[21] Kadiyala KR and S Karlsson (1997), 'Numerical Methods for Estimation and Inference in Bayesian VAR-Models', Journal of Applied Econometrics, 12(2), pp 99-132.
[22] Kapetanios G, V Labhard and S Price (2005), 'Forecasting Using Bayesianand Information Theoretic Model Averaging: An Application to UK Inflation' Bank of England Working Paper No 268.
[23] Litterman RB (1986), 'Forecasting with Bayesian Vector Autoregressions: Five Years of Experience', Journal of Business and Economic Statistics, 4(1), pp 25-38.
[24] Raftery AE, D Madigan and JA Hoeting (1997), 'Bayesian Model Averaging for Linear Regression Models', Journal of the American Statistical Association, 92(437), pp 179-191.
[25] Robertson JC and EW Tallman (1999), 'Vector Autoregressions: Forecasting and Reality', Federal Reserve Bank of Atlanta Economic Review, First Quarter, pp 4-18.
[26] Smets F and R Wouters (2004), 'Forecasting with a Bayesian DSGE Model: An Application to the Euro Area', Journal of Common Market Studies, 42(4), pp841-867.
[27] Stock JH and MW Watson (2002), 'Macroeconomic Forecasting Using Diffusion Indexes', Journal of Business and Economic Statistics, 20(2), pp 147-162.
[28] Timmermann A (2006), 'Chapter 4: Forecast Combinations', in G Elliott, CWJ Granger and A Timmermann (eds), Handbook of Economic Forecasting,Volume 1, Elsevier, Amsterdam, pp 135-196.

Appendix A. Visual and Statistical Assessment

Figures 5 and 6 present a visual assessment of the hypothesis that the pit-values corresponding to the two- and four-quarter-ahead density forecasts are uniformly distributed.

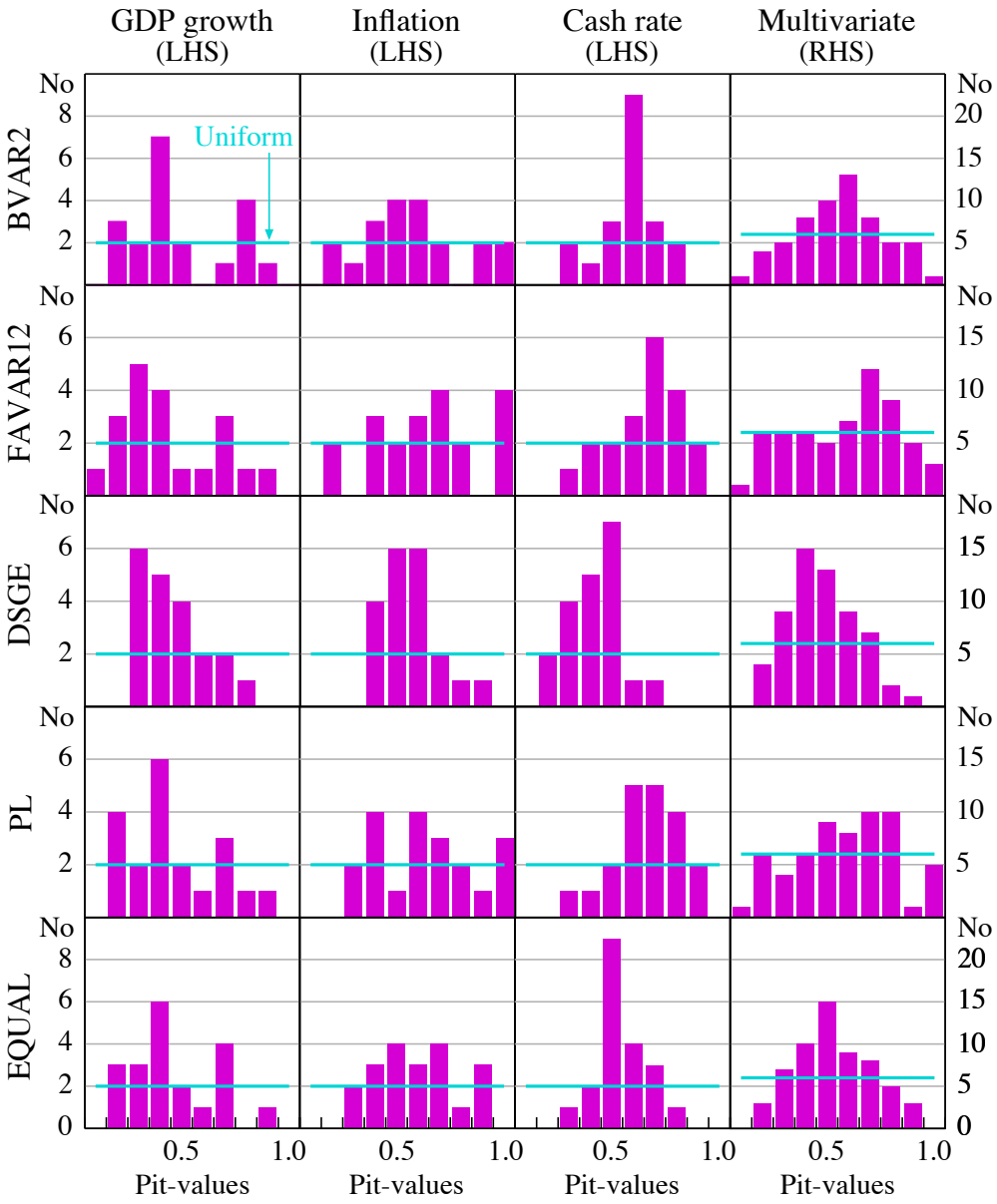
Table 4 reports p-values for likelihood ratio tests of the null hypothesis that the density forecasts are correctly specified at different forecast horizons. The test is a variant of the tests suggested by Berkowitz (2001) and is described in Elder *et al* (2005) with two degrees of freedom. The results are broadly in line with the visual assessment of the uniformity hypothesis conducted in the main text, although it is difficult to make a direct comparison. According to the tests, the univariate density forecasts of GDP growth and the cash rate were, in general, poorly characterised, while the inflation density forecasts tended to fare better. We were unable to reject the null hypothesis at the 95 per cent significance level that the BVAR12, FAVAR12, and predictive-likelihood weighted combination density forecasts for inflation were not misspecified at any forecast horizon. The test of correctly-specified multivariate density forecasts proved difficult to pass, and except for the FAVAR12 model at the one- and two-quarter-ahead forecast horizons, the null hypothesis that the multivariate density forecasts coincide with the actual joint density was rejected at the 95 per cent significance level. Again, there is little evidence to suggest that the 'optimally' combined density forecasts are superior to the best-performing individual model or the equally weighted forecasts, although the small sample makes it difficult to draw strong conclusions.

FIGURE 5. Pit Distributions
Two-quarter-ahead horizon



Notes: Rows in the figure refer to the three individual model density forecasts and the two combination density forecasts. The first three columns refer to the pit-values corresponding to the univariate two-quarter-ahead density forecasts for GDP growth, inflation and the cash rate. The final column refers to the multivariate forecasts where the histogram is constructed using the 'stacked' sequence of pit-values as described in the main text. The height of each bin (vertical axis) reflects the number of observations that fell within different percentile bands (horizontal axis) over the evaluation period (24 observations in total in the univariate cases and 24 observations for each of the three variables in the multivariate case).

FIGURE 6. Pit Distributions
Four-quarter-ahead horizon



Notes: Rows in the figure refer to the three individual model density forecasts and the two combination density forecasts. The first three columns refer to the pit-values corresponding to the univariate four-quarter-ahead density forecasts for GDP growth, inflation and the cash rate. The final column refers to the multivariate forecasts where the histogram is constructed using the 'stacked' sequence of pit-values as described in the main text. The height of each bin (vertical axis) reflects the number of observations that fell within different percentile bands (horizontal axis) over the evaluation period (20 observations in total in the univariate cases and 20 observations for each of the three variables in the multivariate case).

Table 4: Likelihood Ratio Tests

	GDP growth	Inflation	Cash rate	Multivariate
One-quarter-ahead horizon				
BVAR2	**0.00**	0.09	**0.00**	**0.00**
FAVAR12	**0.01**	0.90	0.24	0.24
DSGE	**0.00**	**0.00**	**0.01**	**0.00**
PL	**0.01**	0.25	**0.00**	**0.00**
EQUAL	**0.00**	**0.01**	**0.00**	**0.00**
Two-quarter-ahead horizon				
BVAR2	**0.01**	0.08	**0.00**	**0.00**
FAVAR12	**0.02**	0.78	0.18	0.30
DSGE	**0.00**	**0.00**	**0.00**	**0.00**
PL	**0.01**	0.23	**0.00**	**0.00**
EQUAL	**0.00**	**0.01**	**0.00**	**0.00**
Four-quarter-ahead horizon				
BVAR2	**0.01**	0.11	**0.00**	**0.00**
FAVAR12	**0.04**	0.30	**0.03**	**0.00**
DSGE	**0.00**	**0.00**	**0.00**	**0.00**
PL	**0.02**	0.22	**0.01**	**0.00**
EQUAL	**0.00**	**0.01**	**0.00**	**0.00**
Eight-quarter-ahead horizon				
BVAR2	**0.01**	0.47	**0.00**	**0.00**
FAVAR12	**0.01**	0.59	**0.00**	**0.05**
DSGE	**0.00**	**0.00**	**0.00**	**0.00**
PL	**0.01**	0.64	**0.00**	**0.02**
EQUAL	**0.00**	0.14	**0.00**	**0.00**

Note: Numbers in the table are the p-values for the likelihood ratio test of zero mean and unit variance of the inverse normal cumulative distribution function transformed pit-values, with a maintained assumption of normality. Numbers in bold indicate that the null hypothesis is rejected at the 5 per cent significance level.