Combining Forecast Densities from VARs with Uncertain Instabilities[*]

Anne Sofie Jore
(Norges Bank)

James Mitchell
(NIESR)

Jon Nicolaisen
(Norges Bank)

Shaun P. Vahey
(Norges Bank and RBNZ)

December 3, 2007

Abstract

Clark and McCracken (2008) argue that combining real-time point forecasts from VARs of output, prices and interest rates improves point forecast accuracy in the presence of uncertain model instabilities. In this paper, we generalize their approach to consider forecast density combinations and evaluations. Whereas Clark and McCracken (2008) show that the point forecast errors from particular equal-weight pairwise averages are typically comparable or better than benchmark univariate time series models, we show that neither approach produces accurate real-time forecast densities for recent US data. If greater weight is given to models that allow for the shifts in volatilities associated with the Great Moderation, predictive density accuracy improves substantially.

[*] Address for correspondence: James Mitchell, National Institute of Economic and Social Research, 2 Dean Trench Street, Smith Square, London, SW1P 3HE, U.K. Tel: +44 (0) 207 654 1926. Fax: +44 (0) 207 654 1900. E-Mail: j.mitchell@niesr.ac.uk. We thank Todd Clark and Michael McCracken for providing the data used in this study. We benefitted greatly from comments by Todd Clark, Jim Nason, Christie Smith and Simon van Norden. We are also indebted to seminar participants at Norges Bank Nowcasting Workshop 2007, and the CIRANO data revisions Workshop 2007. The views in this paper represent those of the authors and are not those of the Reserve Bank of New Zealand or Norges Bank.

1 Introduction

Clark & McCracken (2008) [CM, 2008] find that forecast combination improves the Root Mean Squared forecast Error (RMSE) of Vector Autoregressions (VARs) in the presence of unknown structural changes, referred to as "uncertain instabilities". They endorse two forecast selection strategies. In the first, the forecasts from two models are weighted equally, which we refer to as pairwise equal weights (PEW). The second strategy involves averaging the forecasts from all the VARs considered, which we term equal weights (EW). The finding that these combinations help is reassuring to forecasters and decision makers concerned with quadratic loss. In this special case, uncertain instabilities in VARs can (in general) be circumvented by simply attaching an equal weight to each forecast.

For more general but unknown loss functions, the effectiveness of simple averages for VARs in the presence of unknown structural changes has not been studied previously. This is surprising given the plausibility of asymmetric loss functions where the range of uncertainty matters; see Granger & Pesaran (2000). For example, the Federal Reserve may not care equally about inflation above and below the zero bound, but of course the exact loss function of the monetary policymaker is unknown.

In this paper, we generalize the analysis of CM (2008) to study the forecast densities produced by averaging. As discussed by Timmermann (2006), the literature on averaging densities has produced a number of feasible alternatives, with no consensus on the "best" approach. In keeping with the simple pairwise convex mix of point forecasts considered by CM (2008), we consider the analogous convex combinations of probability forecasts known as the "linear opinion pool"; see Timmermann (2006) (p.177). We evaluate the resulting forecast densities by three methods. First, by testing whether the probability integral transforms of the forecast density with respect to the realization of the variable are uniform and, via a transformation, normal (e.g., Diebold et al. (1998)). Second, by using the logarithmic score (e.g., Gneiting & Raftery (2007), Hall & Mitchell (2007) and Amisano & Giacomini (2007)). And finally by considering the probabilities of tail events or economic events of interest, such as a (one period) recession. Corradi & Swanson (2006) review methods for predictive density evaluation.

To facilitate comparisons with CM (2008), we use the same real-time data and estimate their preferred VAR models (and combinations) in output, prices and the short-term interest rate. The models include a wide range of VARs and ARs, including the full sample and their respective rolling window variants considered by CM (2008).

In contrast to CM (2008), we find a substantial difference between the accuracy of simple averaging before, and after, the US Great Moderation. In the 1985-2005 period, the PEW forecast density combinations favored by CM (2008) rank very badly. The univariate benchmark models are roughly as accurate as the simple pairwise combinations for all evaluation periods. The strategy of equally weighting all models considered, EW, is more successful. But it is often dominated by the strategy of time varying recursive weights, RW, proportional to the logarithmic score.

The following figure illustrates how PEW density combinations can produce inaccurate forecast densities in the presence of the shifting volatilities exhibited in the US sample data. The top left-hand panel of Figure 1 plots the 1-step ahead out-of-sample forecast mean for output growth from two models favored by CM (2008). The first is an order 4 VAR (using detrended inflation), the second is an autoregressive model with two lags. Neither model performs particularly well, exhibiting substantial forecast failure at times. The top right-hand panel illustrates the modest benefits of the PEW strategy for point forecasts.

The bottom panel of the figure shows the corresponding plots for the 1-step ahead probability that output growth is less than zero percent. The bottom left-hand panel shows that both models suggest implausibly high probabilities of this event. Output growth very rarely drops below this threshold in our sample—only once in the last 10 years of the sample—as noted by Potter (2007). The bottom right-hand panel reveals that the pairwise equally-weighted (PEW) forecast density gives a poor indication of the probability of this particular event, indicating a 10 to 20 percent probability of a (one period) recession for most of the last 10 years. Broadening the model space to take a convex combination across all the models considered, with weights based on the recursive logarithmic score, produces more accurate probabilities. The RW forecasts achieve this improved performance by giving greater weight to models that allow for the shifts in volatilities associated with the Great Moderation.

The rest of this paper provides more formal evidence that equally weighted combinations provide poor density forecasts for US real-time data spanning the Great Moderation. We consider a wide range of models, together with many combinations of their forecast densities. Furthermore, we evaluate the densities, rather than the conditional means (as in CM, 2008), and the probability of large losses (along the lines of the above output growth example).

The remainder of this paper is structured as follows. In Section 2 we describe our methods for forecast density combination. In Section 3, we outline briefly the data set used in both this paper and CM (2008). In Section 4, we present the results, and in the final section, we discuss the scope for future research in this area.

2 Methods for density combination and evaluation

We begin by describing the density combination methods used in this study. The high-ranking PEW and EW combinations used by CM (2008) can be obtained as special cases by taking the means of the individual densities. Table 1 summarizes the complete set of model types.

2.1 Forecast density combination

To formalize density combination in a way that extends the pairwise convex mix of point forecasts utilized by CM (2008), we adopt the linear opinion pool approach described by Timmermann (2006), (p.177), and the references described therein. Given $i = 1, \ldots, N$ VAR and AR models, the combined densities are defined by the convex combination (linear opinion pool):[1]

$$p_{\tau}(y_{\tau,h}) = \sum_{i=1}^{N} w_{i,\tau,h}\ g(y_{\tau,h} \mid I_{i,\tau}), \qquad \tau = \underline{\tau}, \ldots, \overline{\tau}, \tag{1}$$

where $g(y_{\tau,h} \mid I_{i,\tau})$ are the h-step ahead forecast densities from individual model i, $i = 1, \ldots, N$ of a variable y_{τ}, conditional on the information set I_{τ}. The publication delay in the production of real-time data ensures that this information set contains macroeconomic variables dated $\tau - 1$ and earlier. Each individual model is used to produce h-step ahead forecasts via the direct approach; see the discussion by Marcellino et al. (2003). Hence, the macro variables used to produce an h-step ahead forecast density for τ are dated $\tau - 1 - h$. The set of non-negative weights, $w_{i,\tau,h}$, in this finite mixture sum to unity.[2] Furthermore, the weights may change through the evaluation period $\tau = \underline{\tau}, \ldots, \overline{\tau}$.

Since the VAR and AR models considered produce forecast densities that are normal, but with different means and variances, the combined density by Equation (1) will be mixture normal—accommodating skewness and kurtosis. The method delivers a more

[1] The linear opinion pool is sometimes justified by considering an expert combination problem. See for example, Morris (1974, 1977) and Winkler (1981), Lindley (1983) and McConway (1990). Wallis (2005) proposes the linear opinion pool as a tool to aggregate forecast densities from survey participants. Mitchell & Hall (2005) combine inflation density forecasts from different institutions.

[2] The restriction that each weight is positive might be relaxed; for discussion see Genest & Zidek (1986).

flexible distribution than each of the individual densities from which it was derived.

We consider a number of different methods for constructing the weights, $w_{i,\tau,h}$.

2.1.1 Recursive weights

We construct the weights based on the fit of the individual model forecast densities. Following Amisano & Giacomini (2007) and Hall & Mitchell (2007), we use the logarithmic score to measure density fit for each model through the evaluation period. The logarithmic scoring rule is intuitively appealing as it gives a high score to a density forecast that provides a high probability to the value y_τ that materializes. Specifically the weights for the h-step ahead densities take the form:

$$w_{i,\tau,h} = \frac{\exp\left[\sum_{\underline{\tau}-10}^{\tau-1-h} \ln g(y_{\tau,h} \mid I_{i,\tau})\right]}{\sum_{i=1}^{N} \exp\left[\sum_{\underline{\tau}-10}^{\tau-1-h} \ln g(y_{\tau,h} \mid I_{i,\tau})\right]}, \qquad \tau = \underline{\tau}, \dots, \overline{\tau} \tag{2}$$

where the $\underline{\tau} - 10$ to $\underline{\tau}$ comprises the training period used to initialize the weights. Hall & Mitchell (2007) demonstrate that these weights minimize the Kullback-Leibler Information Criterion (KLIC) distance between the combined density forecast and the true but unknown density.

From a Bayesian perspective, the forecast density averaging based on recursive logarithmic score weights, RW, has many similarities with an approximate predictive likelihood approach (see Raftery & Zheng (2003), and Eklund & Karlsson (2007)). Given our definition of density fit, the model densities are combined using Bayes rule with equal (prior) weight on each model—which a Bayesian would term non-informative priors. (Koop (2003) (chapter 11) and Geweke & Whiteman (2006) provide recent general discussions of Bayesian Model Averaging methods.) Andersson & Karlsson (2007) propose Bayesian predictive likelihood methods for forecast combination with Bayesian VARs but do not consider forecast density evaluation.

2.1.2 Equal weights

The EW strategy suggested by CM (2008) attaches equal (prior) weight to each model with no updating of the weights through the recursive analysis: $w_{i,\tau,h} = w_{i,h} = 1/N$. Pairwise equal weight combinations, PEW, can be thought of as a truncation of the (prior) model space. We note, however, that it is difficult to justify *a priori* truncation. Put differently, a researcher would have faced considerable uncertainty about which pairs of models to select according to the PEW strategy at the start of our evaluation period.

2.2 Evaluation of forecast density combinations

In constructing the RW forecast densities, the model forecasts are evaluated using the logarithmic score for each recursion. We emphasize that in deriving the weights based on this measure of density fit, the many models are repeatedly evaluated using real-time data.[3] By construction, the PEW and EW strategies are dominated by the RW approach using the logarithmic score. The forecast density evaluation carried out for each evaluation period in the recursive analysis provides the weights on each model, and by summing the weights, the weights attached to particular PEW combinations.

It is common in the density forecast evaluation literature to provide classical statistics suitable for a single model evaluation in a one shot test. We report *pits* statistics at the end of the evaluation period by using a number of methods. Diebold et al. (1998) have shown that the *pits* are uniform when the density forecast is optimal (well calibrated) and also i.i.d. for one step-ahead forecasts. When the *pits* are (i.i.d.) uniform the combined density forecast will be preferred irrespective of the loss function. In practice, forecast density evaluation by this method requires application of a test for uniformity (or normality, via the inverse normal cumulative density function transformation).

We evaluate the forecast densities from various strategies, such as RW, PEW and EW, using a number of *pits* statistics. These include the Likelihood Ratio (LR) test proposed by Berkowitz (2001) for which results are presented using a two degrees-of-freedom variant (without a test for autocorrelation); see Clements (2004).[4] We also follow Berkowitz (2001) and report a censored LR test which focuses on the 10% top and bottom tails. This is designed to detect forecast failure in the tails of the forecast density (see figure 1). Finally, we consider the Anderson-Darling (AD) test, a modification of the Kolmogorv-Smirnov test, intended to give more weight to the tails (and advocated by Noceti et al. (2003)). Given our limited number of macroeconomic observations, the LR test is probably a better guide than using nonparametric tests.

[3] Clark & McCracken (2007) discuss tests of mean squared forecast error performance with real-time data.

[4] Results are similar when the three degrees-of-freedom test is employed.

3 The US data

To facilitate comparisons with CM (2008), we use the same real-time US data set and, like them, estimate VAR models in output, inflation and the short-term interest rate. That is, we use the same economic variables, and their respective measures.

The raw data for GDP (in practice, GNP for some vintages) are taken from the Federal Reserve Bank of Philadelphia's Real-Time Data Set for Macroeconomists. This is a collection of vintages of National Income and Production Accounts; each vintage reflects the information available around the middle of the respective quarter. Croushore & Stark (2001) provide a description of the database. Although CM (2008) consider two measures of the output gap and output growth, to save space we simply report results for the output growth case. (Results for the other two definitions, which are qualitatively very similar to those reported here, can be obtained from the authors on request.)

The short term interest rate is a T-bill rate taken from the Board of Governor's FAME database. For inflation, we report results using a GDP deflator series, which unlike the T-bill commonly suffers from revisions. (Additional results using an alternative measure of inflation derived from the CPI (Bureau of Labor Statistics, 1967 base year), seasonally adjusted using an X-11 filter by CM (2008), can be provided on request.)

For completeness, the set of models comprising the 68 VARs and ARs are listed in Table 1.[5] The models include ARs, VARs, first differenced VARs (DVARs), de-trended VARs (using an exponential smoother) and bivariate VARs always including the variable of interest (i.e. restricted trivariate VARs). Both full-sample and rolling sample ARs and VARs are estimated. The rolling models are estimated over the last x quarters only. Following CM (2008) we set $x = 40$ for the ARs and $x = 60$ for the VARs. We consider lag lengths of one to four, as well as recursively selecting the lag length using the Schwarz Bayesian Information Criterion (BIC); see Schwarz (1978) for details.[6] The density forecasts from each model are assumed Gaussian, with conditional mean centered on the recursively computed point forecast and the conditional variance equal to the approximate mean squared error of the forecast with parameter uncertainty (see Lutkepohl (1991), p. 87, eq. 3.5.9).

The start dates for the GDP observations vary by vintage, dictated by data availability in the FRB Philadelphia Real-time database. Like CM (2008, p4), we use $1955Q1$ as the

[5] We focus on the VARs and ARs preferred by CM (2008) but do not consider their BVAR and factor models which typically rank poorly in their forecasts evaluations.

[6] CM (2008) also report results for the Akaike Information Criterion. The ranking of the models is very similar to BIC and can be obtained from the authors on request.

first observation for parameter estimation of the models, or failing that, the first quarter available (allowing for five quarters for differencing and lags). So each model forecast is based on a sample, $t = t_{0,i_\tau}, \ldots, \tau - 1$ to estimate the parameters of interest, where the start date, t_{0,i_τ}, can vary by the model, and for rolling sample models, by recursion. Each individual model produces forecast densities based on the sequence of data vintages starting in $1965Q4$ and ending in $2005Q4$.

To match the approach taken by CM (2008), we break our evaluation period, $\tau = \underline{\tau}, \ldots, \overline{\tau}$ into two subperiods: with τ from $1970Q1$ to $1984Q4$, and from $1985Q1$ to $2005Q4$.[7] To implement density combination through the evaluation period requires an additional assumption about which measurement is to be forecast. CM (2008) use the second estimate as the "final" data to be forecast. For consistency, we report results for the same definition of "final" data for all forecast density combinations and evaluations. CM (2008) discuss the robustness of their results to other definitions of realized outturns; see also the discussion in Corradi et al. (2007). For our reported results, the delay in observing the outturn introduces an additional one period lag in the construction of the recursive density combination weights.

4 Results

We break our results into three components: the ranking of the model types (e.g. rolling and full sample variants); forecast density evaluations of the pairwise equal weight combinations, PEW, and benchmark AR strategies found to work well for point forecasts by CM (2008); and summaries of forecast density accuracy for the recursively selected weights, RW, and equal weights, EW, strategies. In each case, we present results for the evaluation period split into two subperiods, 1970-1984 and 1985-2005, for horizons 1 (denoted $h = 0$ since this is a current quarter forecast) and 5 (denoted $h = 4$).[8]

For each variable of interest, we present the weights on each model type at the end of each evaluation period in Table 2 for horizons $h = 0$ and $h = 4$.[9] Recall that these weights use second measurements as "final data" with density fit measured by the logarithmic

[7] CM (2008) present results for point forecast combinations based on a static "training period" window from $1965Q4$ to $1969Q4$. All our density combinations with recursive weights use an expanding window, starting in $\underline{\tau} - 10$.

[8] The results for horizons $h = 1, \ldots, 8$ are qualitatively similar and can be obtained from the authors on request.

[9] Note that like CM (2008), we analyze the forecast for each variable individually, not jointly. We plan to explore variable dependence in subsequent work.

score. Looking first at the 1985-2005 period for GDP growth, the highest weighted models are rolling window variants. For every model type, the rolling version receives greater weight than the full sample equivalent. For example, for the VAR models at horizon $h = 1$, the rolling weight is approximately 25 percent and the full sample weight is zero. For the DVARs the respective numbers are six percent and zero. However, in the earlier 1970-1984 period, rolling window variants fare less well for GDP growth. The longer horizon, $h = 4$, results for GDP growth display a similar pattern across the two evaluation periods, with the rolling models receiving greater weight in the second evaluation period, but not the first.

A second striking feature of the GDP growth results is that the benchmark full sample AR class of models rarely receive substantial weight. Regardless of the evaluation period, this class typically receives a zero weight. The only exception is the 7.2 percent weight for the ARs in the 1970-1984 evaluation.

For inflation and interest rates, the patterns are similar to GDP growth for the later evaluation period, but the rolling models also receive large weights in the earlier 1970-1984 period. In general, the weights at the longer horizon are more concentrated than in the GDP growth case, with the predominant models typically rolling.

To shed further light on the generally poor performance of the full sample AR benchmarks, Figure 2 plots the recursively selected weights through the 1985-2005 evaluation period for the AR(2) benchmark, found by CM (2008) to produce competitive point forecasts. We also plot the detrended VAR(4) model recursive weights. Recall that these two models make up CM's preferred PEW combination. The recursively computed forecast density weights typically exhibit very little time variation (a notable exception is interest rates before 1995). The general picture is one of declining weights as the evaluation period progresses: all the weights are negligible by 2005. The PEW combination of these two models also receives an approximately zero weight (the sum of the individual weights), as does the rolling equivalent (not shown). We note, however, that the detrended VAR(4) does better for both GDP growth and inflation in the first evaluation period.

To summarize, the analysis of our recursively generated weights based on the logarithmic score suggests little evidence that the benchmark AR models, or the PEW combination favored by CM (2008), produce accurate forecast densities, particularly over the 1985-2005 period.

Turning to the *pits* test p-values displayed in Tables 3-5, we see a similar story. The GDP growth densities are evaluated in Table 3 for the 1970-1984 evaluation period, top panel, and for 1985 to 2005, bottom panel. As with the previous table, we also report results by forecast horizon: $h = 0$ in the left hand panel and $h = 4$ in the right. Looking at the later evaluation period, bottom panel, the benchmark AR(2) results indicate that the densities are rejected at the one percent level for the LR and censored LR statistics; and the AD test statistic exceeds the 2.5 critical value. The rolling equivalent of the benchmark performs much better on all the tests. The PEW combination favored by CM (2008) which averages the AR(2) with the detrended VAR (4) gives a similar density performance to the benchmark. However the rolling variant of that combination does not produce as much improvement as in the AR benchmark case.

Moving on to the longer horizon, those same patterns emerge again. But the evidence based on the 1970-1984 period is more supportive of the benchmark and AR densities. The 1985-2005 period is when these models and PEW combinations perform worst.

Looking at Table 4 and 5, which show comparable results for the other variables, we see that the forecast densities for inflation give fairly poor LR and AD statistics for all time periods and horizons. The PEW combination favored by CM (2008) does poorly too. But the rolling variant of it does better for the $h = 0$ 1985-2005 case. The interest rate forecast densities are inadequate in all cases.

Since the evidence does not offer much support for the benchmark densities or those from the PEW strategy, we now turn to forecast densities produced by our RW strategy. Tables 3-5 show that the RW densities perform relatively well, with some differences by test statistic. Typically the LR p-values are greater than 5 percent, often much greater, and the AD statistics often less than 2.5. For example, the bottom left quadrant of Table 3 shows that the censored LR statistic p-values are around 30 percent and 9 percent for the lower (10 percent) and upper (10 percent) tail LR tests, respectively. The AD test is borderline. But, in contrast, the overall LR test is not, with a significance of less than one percent.

For the other variables shown in Tables 4 and 5, the RW forecast densities give some favorable statistics, but these are mostly restricted to the 1985-2005 evaluation period for inflation. The results for the 1970-1984 evaluation period and for interest rates in both evaluation periods are weaker. Evaluations of the density forecasts produced each quarter in real-time by the Survey of Professional Forecasters have also found the GDP growth forecasts to be better calibrated than those for inflation; see Diebold et al. (1999) for an evaluation of the inflation densities and Mitchell (2007) for an evaluation of both the inflation and real GDP growth densities.

Finally, we turn to the evaluation of the EW strategy, which attaches equal weight to all the models considered (see Table 1). Over the earlier evaluation period, 1970-1984, the results are similar to the RW case. For example, for GDP growth both combinations deliver densities that appear well calibrated with at $h = 0$ a p-value for the (overall) LR test of 0.479 for EW and 0.231 for RW. For inflation, again the RW performance is matched by EW, although the forecast densities do not perform as well as for GDP growth. For interest rates, as the top panel of Table 4 indicates, there is even stronger rejection of the forecast densities over the 1970-1984 period.

The more interesting story is over the later evaluation period, 1985-2005, where the EW strategy performs poorly. For example, the bottom panel of Table 3 shows that for GDP growth at both $h = 0$ and $h = 4$, for 1985-2005, we reject the adequacy of the EW combination, with p-values generally less than five percent and AD statistics greater than 2.5 (the 95% critical value). A similar pattern emerges for inflation as for GDP growth in Table 4. For interest rates, Table 5, there is little difference between the RW and EW strategies.

Overall, we draw out three main findings. First, typically the RW forecast densities are the most satisfactory: the RW combinations appear particularly well-calibrated for GDP growth. Consistent with the impression given by Figure 1, the RW combination appears to perform particularly well relative to PEW in the tails of the distribution. Second, simple full sample PEW and the full sample AR benchmarks perform equally badly. Third, the EW strategy fares better than the PEW strategy and is comparable to the RW approach for the 1970-1984 evaluation period but is worse for the 1985-2005 period.

5 Conclusions

CM (2008) argue that combining real-time point forecasts from VARs of output, prices and interest rates improves point forecast accuracy in the presence of uncertain model instabilities. They show that the (root mean squared) forecast errors from simple pairwise averages are typically comparable with, or better, simple benchmark univariate time series models. In contrast, in this paper, we have shown that neither approach produces accurate real-time forecast densities for recent US data. Recursively constructed weights give greater weight to rolling models that allow for the shifts in volatilities associated with the Great Moderation. This improves substantially predictive density accuracy.

Although VAR combinations are widely used for forecasting macroeconomic variables, particularly by central banks, there have been very few studies that systematically evaluate forecast performance in the presence of uncertain instabilities. The framework developed by CM (2008) takes an important first step towards formal evaluation by focusing on the point forecasts. Their results suggest little scope for improvement beyond taking simple pairwise averages. When generalizing their framework to consider forecast densities, we have found that recursively constructed weights give better forecasts, by bringing together evidence for a wide range of VAR models. The recursive weight strategy frequently attaches high weight to models that allow for structural change through rolling windows. In future work, we plan to explore two further avenues for improving density forecast accuracy: the scope for formal modeling of structural breaks (emphasized by Pesaran & Timmermann (2007)); and the potential of restrictions derived from dynamic stochastic general equilibrium models.

References

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

Andersson, M. & Karlsson, S. (2007), Bayesian forecast combination for VAR models. Unpublished manuscript, Sveriges Riksbank.

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

Clark, T. E. & McCracken, M. W. (2007), Tests of equal predictive ability with real-time data. Research Working Paper 07-06, FRB Kansas City.

Clark, T. E. & McCracken, M. W. (2008), 'Averaging forecasts from VARs with uncertain instabilities', *Journal of Applied Econometrics* . Forthcoming. Revision of Federal Reserve Bank of Kansas City Working Paper 06-12.

Clements, M. P. (2004), 'Evaluating the Bank of England density forecasts of inflation', *Economic Journal* **114**, 844–866.

Corradi, V., Fernandez, A. & Swanson, N. R. (2007), Information in the revision process of real-time data. Discussion Paper, Rutgers University.

Corradi, V. & Swanson, N. R. (2006), Predictive density evaluation, *in* G. Elliott, C. W. J. Granger & A. Timmermann, eds, 'Handbook of Economic Forecasting', North-Holland, North Holland, pp. 197–284.

Croushore, D. & Stark, T. (2001), 'A real-time data set for macroeconomists', *Journal of Econometrics* **105**, 111–130.

Diebold, F. X., Gunther, A. & Tay, K. (1998), 'Evaluating density forecasts with application to financial risk management', *International Economic Review* **39**, 863–883.

Diebold, F. X., Tay, A. S. & Wallis, K. F. (1999), Evaluating density forecasts of inflation: the Survey of Professional Forecasters, *in* R. Engle & H. White, eds, 'Cointegration, causality and forecasting: a festschrift in honour of Clive W. J. Granger', Oxford University Press.

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

Genest, C. & Zidek, J. (1986), ‘Combining probability distributions: a critique and an annotated bibliography’, *Statistical Science* **1**, 114–135.

Geweke, J. & Whiteman, C. (2006), Bayesian forecasting, *in* G. Elliott, C. W. J. Granger & A. Timmermann, eds, ‘Handbook of Economic Forecasting Volume 1’, North-Holland, pp. 3–80.

Gneiting, T. & Raftery, A. E. (2007), ‘Strictly proper scoring rules, prediction, and estimation’, *Journal of the American Statistical Association* **102**, 359–378.

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

Hall, S. G. & Mitchell, J. (2007), ‘Combining density forecasts’, *International Journal of Forecasting* **23**, 1–13.

Koop, G. (2003), *Bayesian Econometrics*, Wiley.

Lindley, D. (1983), ‘Reconciliation of probability distributions’, *Operations Research* **31**, 866–880.

Lutkepohl, H. (1991), *Introduction to Multiple Time Series Analysis*, Berlin, Spinger-Verlag.

Marcellino, M., Stock, J. & Watson, M. (2003), ‘A comparison of direct and iterated AR methods for forecasting macroeconomic series h-steps ahead’, *Journal of Econometrics* **135**, 499–526.

McConway, C. G. . K. J. (1990), ‘Allocating the weights in the linear opinion pool’, *Journal of Forecasting* **9**, 53–73.

Mitchell, J. (2007), Constructing bivariate density forecasts of inflation and output growth using copulae: modelling dependence using the Survey of Professional Forecasters. Discussion Paper No. 297, NIESR.

Mitchell, J. & Hall, S. G. (2005), ‘Evaluating, comparing and combining density forecasts using the KLIC with an application to the Bank of England and NIESR “fan” charts of inflation’, *Oxford Bulletin of Economics and Statistics* **67**, 995–1033.

Morris, P. (1974), ‘Decision analysis expert use’, *Management Science* **20**, 1233–1241.

Morris, P. (1977), ‘Combining expert judgments: A Bayesian approach’, *Management Science* **23**, 679–693.

Noceti, P., Smith, J. & Hodges, S. (2003), ‘An evaluation of tests of distributional forecasts’, *Journal of Forecasting* **22**, 447–455.

Pesaran, M. H. & Timmermann, A. (2007), ‘Selection of estimation window in the presence of breaks’, *Journal of Econometrics* **137**(1), 134–161.

Potter, S. (2007), Forecasting the frequency of recessions. Unpublished manuscript, FRB New York.

Raftery, A. E. & Zheng, Y. (2003), ‘Long-run performance of Bayesian model averaging’, *Journal of the American Statistical Association* **98**, 931–938.

Schwarz, G. (1978), ‘Estimating the dimension of a model’, *Annals of Statistics* **6**, 461–464.

Timmermann, A. (2006), Forecast combinations, *in* G. Elliott, C. W. J. Granger & A. Timmermann, eds, ‘Handbook of Economic Forecasting Volume 1’, North-Holland, pp. 135–196.

Wallis, K. F. (2005), ‘Combining density and interval forecasts: a modest proposal’, *Oxford Bulletin of Economics and Statistics* **67**, 983–994.

Winkler, R. (1981), ‘Combining probability distributions from dependent information sources’, *Management Science* **27**, 479–488.

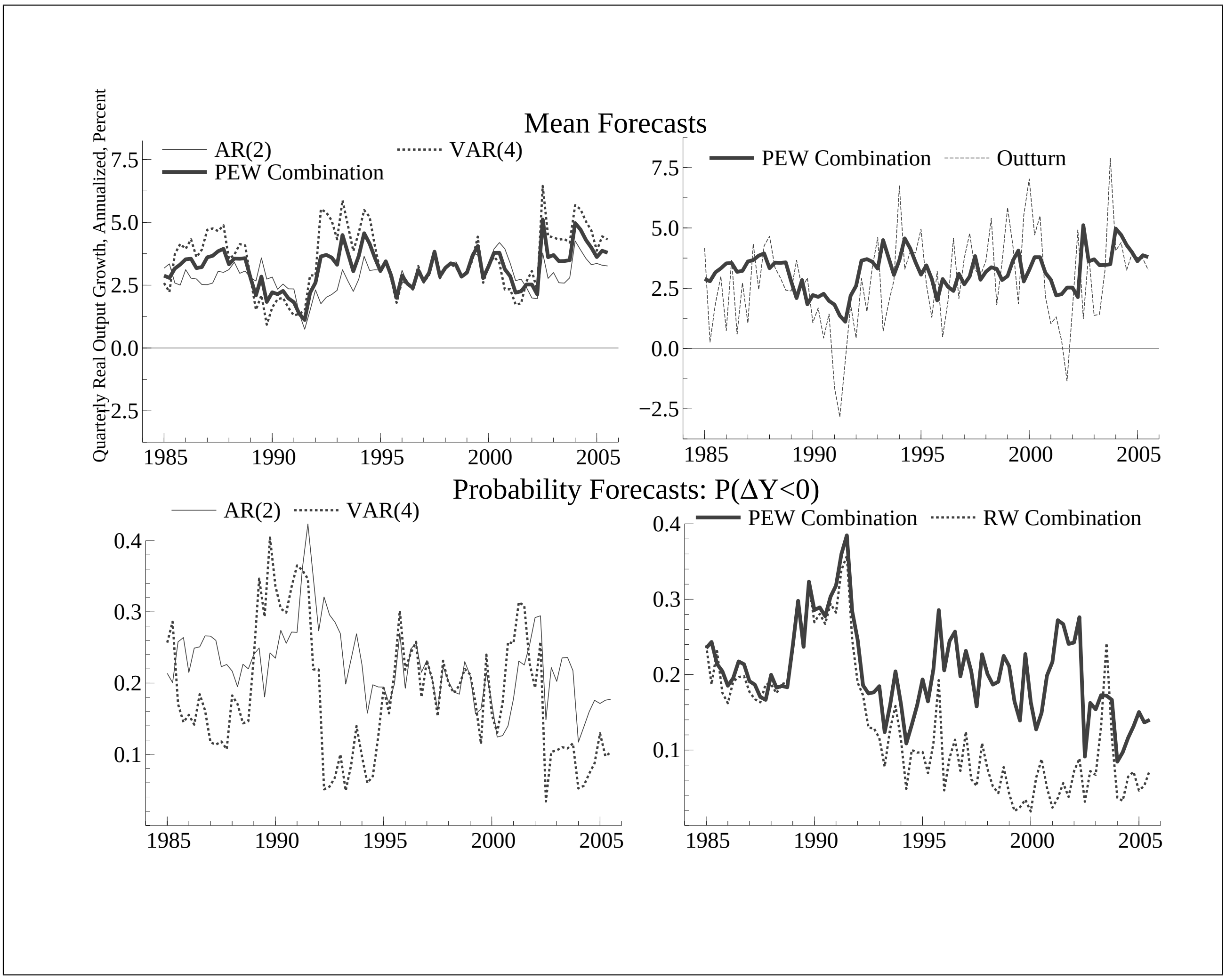

Figure 1: Forecasting performance of AR(2) and de-trended VAR(4) models for GDP growth

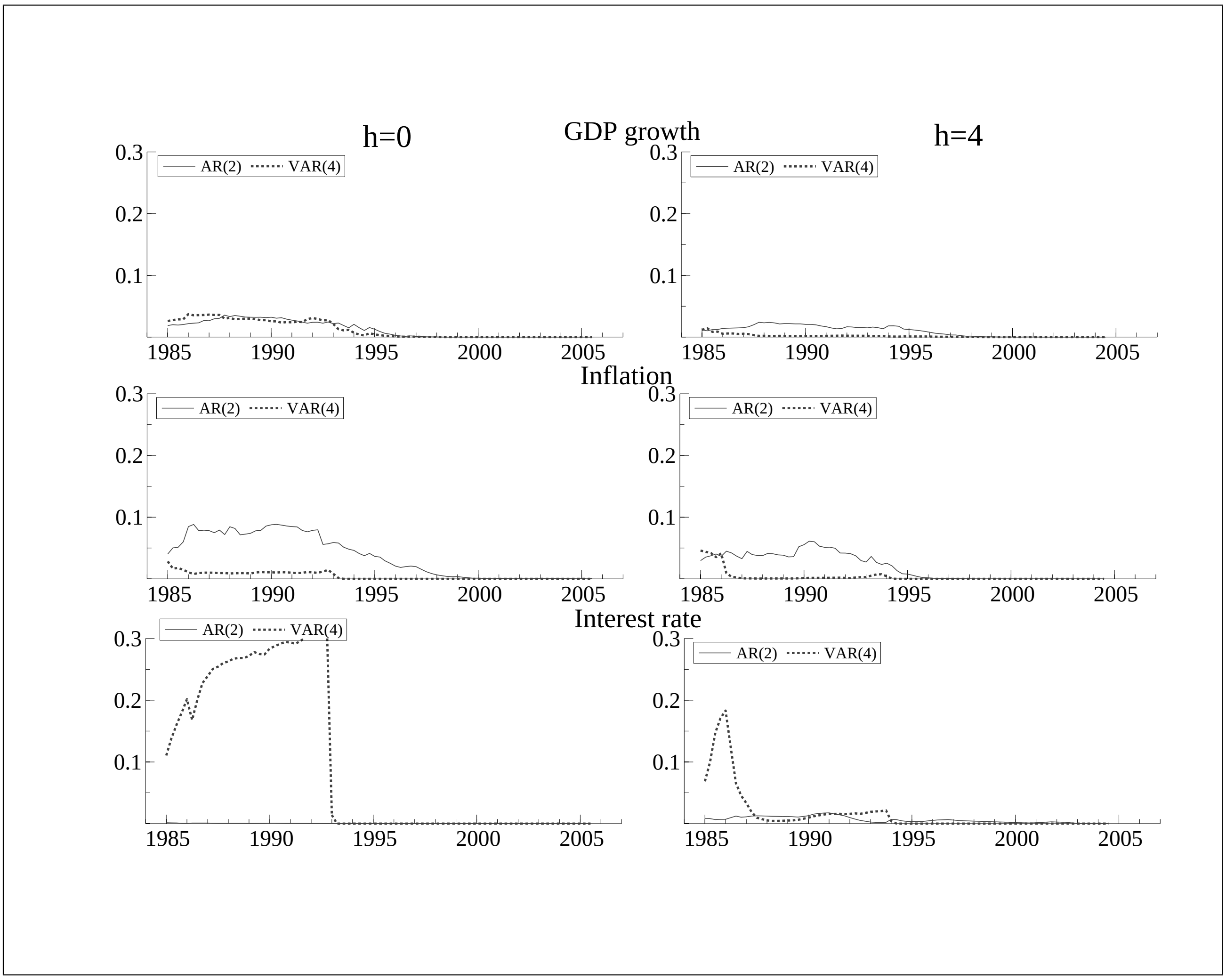

Figure 2: Recursively selected weights on the AR(2) and the detrended VAR(4)

Table 1: Model type: a summary of the complete set of VAR models and combination methods

method	details
AR	ARs with fixed lags of $1-4$ and determined at each t by BIC
AR (rolling)	same as above but estimated with a rolling sample of 40 obs
VAR	VARs in $\Delta y, \pi$ and i with fixed lags of $1-4$ and determined at each t by BIC
VAR (rolling)	same as above but estimated with a rolling sample of 60 obs and for
	... VAR(1) and VAR(BIC) also with rolling samples of 30, 40, 50, 70 and 80 obs
DVAR	VARs in $\Delta y, \Delta\pi$ and Δi with fixed lags of $1-4$ and determined at each t by BIC
DVAR (rolling)	same as above but estimated with a rolling sample of 60 obs
Inf. detrend	VARs in $\Delta y, \pi - \pi^*_{-1}$ and $i - \pi^*_{-1}$ with fixed lags of $1-4$
	... and determined at each t by BIC
Inf. detrend (rolling)	same as above but estimated with a rolling sample of 60 obs
BiVAR	Bivariate VARs in $\Delta y, \pi$ and $\Delta y, i$ for Δy; in $\pi, \Delta y$ and π, i for π;
	... in $i, \Delta y$ and i, π for i with fixed lags of $1-4$ and determined at each t by BIC
BiVAR (rolling)	same as above but estimated with a rolling sample of 60 obs
PEW	equal-weight pairwise average of the AR(2) and the Inf. detrended VAR(4)
PEW (rolling)	same as above but AR (VAR) estimated with a rolling sample of 40 (60) obs
EW	equal-weight average of all AR and VAR models
RW	recursive-weight average of all models determined at each t by the log-score

Notes: The variables $\Delta y, \pi$ and i refer to, respectively, GDP growth, inflation and the interest rate. The BIC lag orders range from 0 (minimum) to 4 (the maximum allowed). $\pi^* = \pi^*_{-1} + .05(\pi - \pi^*_{-1})$.

Table 2: Recursive weights at the end of 2005

h=0	GDP growth		Inflation		Interest rates	
	1970-1984	1985-2005	1970-1984	1985-2005	1970-1984	1985-2005
AR	0.000	0.000	0.000	0.009	0.000	0.000
AR(rolling)	0.000	0.115	0.000	0.924	0.229	0.105
VAR	0.001	0.000	0.016	0.003	0.000	0.241
VAR (rolling)	0.004	0.246	0.661	0.001	0.616	0.361
DVAR	0.001	0.000	0.000	0.006	0.000	0.064
DVAR(rolling)	0.000	0.060	0.225	0.025	0.108	0.205
Inf. Detrend	0.986	0.000	0.002	0.000	0.000	0.000
Inf. Detrend (rolling)	0.001	0.522	0.004	0.000	0.047	0.000
BiVAR	0.006	0.000	0.005	0.018	0.000	0.018
BiVAR rolling	0.000	0.056	0.087	0.014	0.000	0.006

h=4	GDP growth		Inflation		Interest rates	
	1970-1984	1985-2005	1970-1984	1985-2005	1970-1984	1985-2005
AR	0.072	0.000	0.000	0.000	0.000	0.001
AR(rolling)	0.000	0.001	0.812	0.000	0.000	0.000
VAR	0.002	0.000	0.000	0.000	0.000	0.965
VAR (rolling)	0.000	0.021	0.000	0.000	0.000	0.000
DVAR	0.574	0.000	0.000	0.311	0.000	0.000
DVAR(rolling)	0.000	0.068	0.000	0.689	0.000	0.000
Inf. Detrend	0.014	0.001	0.000	0.000	0.002	0.000
Inf. Detrend (rolling)	0.000	0.760	0.000	0.000	0.998	0.000
BiVAR	0.337	0.001	0.081	0.000	0.000	0.034
BiVAR rolling	0.000	0.149	0.108	0.000	0.000	0.000

Notes: Weights on selected subsets of models; see Table 1 for a description of the models.

Table 3: GDP growth density forecasts

	1970-1984				1970-1984			
	h=0				h=4			
	LR	LR_{lower}	LR_{upper}	AD	LR	LR_{lower}	LR_{upper}	AD
AR(2)	0.272	0.138	0.881	0.532	0.072	0.006	0.177	1.348
AR(2) roll	0.013	0.007	0.135	1.262	0.031	0.002	0.836	1.474
avg. AR(2),VAR(4)	0.982	0.918	0.594	0.239	0.443	0.106	0.337	0.944
avg. AR(2),VAR(4) roll	0.242	0.185	0.397	0.781	0.335	0.003	0.133	0.709
EW	0.479	0.814	0.208	1.352	0.850	0.096	0.457	1.287
RW	0.231	0.407	0.438	1.290	0.096	0.002	0.764	0.830
	1985-2005				1985-2005			
	h=0				h=4			
	LR	LR_{lower}	LR_{upper}	AD	LR	LR_{lower}	LR_{upper}	AD
AR(2)	0.000	0.000	0.002	7.884	0.000	0.004	0.000	7.078
AR(2) roll	0.134	0.481	0.600	1.438	0.975	0.610	0.903	0.813
avg. AR(2),VAR(4)	0.000	0.000	0.000	8.480	0.000	0.005	0.000	8.160
avg. AR(2),VAR(4) roll	0.001	0.001	0.191	2.994	0.019	0.057	0.595	1.658
EW	0.000	0.001	0.003	5.847	0.000	0.039	0.000	5.436
RW	0.004	0.302	0.087	2.507	0.000	0.033	0.222	2.863

Notes: LR is the p-value for the Likelihood Ratio test of zero mean and unit variance of the inverse normal cumulative distribution function transformed *pits*, with a maintained assumption of normality for the transformed *pits*; LR_{upper} is the p-value for the LR test of zero mean and unit variance focusing on the 10 percent upper tail; LR_{lower} is the p-value for the LR test of zero mean and unit variance focusing on the 10 percent lower tail; AD is the Anderson-Darling test statistic for uniformity of the *pits* which assuming independence of the *pits* has an associated 95 percent asymptotic critical value of 2.5. roll denotes models estimated using a rolling window of length 40 quarters for ARs and 60 quarters for VARs

Table 4: Inflation density forecasts

	1970-1984				1970-1984			
	h=0				h=4			
	LR	LR_{lower}	LR_{upper}	AD	LR	LR_{lower}	LR_{upper}	AD
AR(2)	0.000	0.013	0.000	4.356	0.000	0.005	0.000	8.970
AR(2) roll	0.000	0.019	0.003	2.803	0.000	0.000	0.000	9.833
avg. AR(2),VAR(4)	0.000	0.009	0.000	2.829	0.000	0.000	0.000	10.137
avg. AR(2),VAR(4) roll	0.001	0.070	0.005	2.001	0.000	0.000	0.000	10.476
EW	0.009	0.609	0.003	1.518	0.000	0.034	0.000	6.077
RW	0.000	0.004	0.000	5.338	0.000	0.000	0.000	18.032
	1985-2005				1985-2005			
	h=0				h=4			
	LR	LR_{lower}	LR_{upper}	AD	LR	LR_{lower}	LR_{upper}	AD
AR(2)	0.001	0.241	0.008	2.624	0.000	0.008	0.000	8.213
AR(2) roll	0.762	0.591	0.903	1.225	0.028	0.341	0.340	4.098
avg. AR(2),VAR(4)	0.000	0.068	0.013	3.423	0.000	0.050	0.000	8.037
avg. AR(2),VAR(4) roll	0.247	0.984	0.331	1.801	0.013	0.011	0.750	4.776
EW	0.005	0.086	0.019	2.080	0.000	0.157	0.005	5.612
RW	0.210	0.010	0.844	1.321	0.420	0.284	0.419	0.766

See Notes to Table 3

Table 5: Interest rate density forecasts

	1970-1984				1970-1984			
	h=0				h=4			
	LR	LR_{lower}	LR_{upper}	AD	LR	LR_{lower}	LR_{upper}	AD
AR(2)	0.000	0.001	0.002	6.511	0.000	0.000	0.000	19.819
AR(2) roll	0.000	0.009	0.018	3.493	0.000	0.000	0.000	24.500
avg. AR(2),VAR(4)	0.000	0.000	0.000	3.910	0.000	0.001	0.000	13.657
avg. AR(2),VAR(4) roll	0.001	0.000	0.103	1.930	0.000	0.000	0.000	10.660
EW	0.000	0.000	0.045	2.565	0.000	0.001	0.000	13.761
RW	0.000	0.000	0.005	3.959	0.000	0.000	0.000	18.136
	1985-2005				1985-2005			
	h=0				h=4			
	LR	LR_{lower}	LR_{upper}	AD	LR	LR_{lower}	LR_{upper}	AD
AR(2)	0.000	0.054	0.000	10.370	0.000	0.568	0.001	8.153
AR(2) roll	0.006	0.578	0.052	4.672	0.000	0.000	0.242	30.974
avg. AR(2),VAR(4)	0.000	0.010	0.000	7.851	0.000	0.094	0.001	11.071
avg. AR(2),VAR(4) roll	0.000	0.229	0.013	6.379	0.000	0.000	0.032	18.923
EW	0.000	0.079	0.000	7.661	0.000	0.018	0.000	8.812
RW	0.002	0.025	0.000	4.469	0.005	0.263	0.014	4.390

See Notes to Table 3