DP2008/18

Combining Forecast Densities from VARs with Uncertain Instabilities

Anne Sofie Jore, James Mitchell and Shaun Vahey

December 2008

JEL classification: C32, C53, E37

www.rbnz.govt.nz/research/discusspapers/

Discussion Paper Series

ISSN 1177-7567

DP2008/18

Combining Forecast Densities from VARs with Uncertain Instabilities*

Anne Sofie Jore, James Mitchell and Shaun Vahey†

Abstract

Recursive-weight forecast combination is often found to an ineffective method of improving point forecast accuracy in the presence of uncertain instabilities. We examine the effectiveness of this strategy for forecast densities using (many) VARs and ARs of output, prices and interest rates. Our proposed recursive-weights density combination strategy, based on the recursive logarithmic score of the forecast densities, produces accurate predictive densities by giving substantial weight to models that allow for structural breaks. In contrast, equal-weight combinations produce poor real-time US forecast densities for Great Moderation data.

* The views expressed in this paper are those of the author(s) and do not necessarily reflect the views of the Reserve Bank of New Zealand. We benefitted greatly from discussions with Todd Clark, John Geweke, Michael McCracken, Jim Nason, Valentyn Panchenko, Christie Smith, Simon van Norden and Ken Wallis. Particular thanks to Kirstin Hubrich and two anonymous referees for helpful comments. We are also indebted to seminar and conference participants at the Norges Bank Nowcasting Workshop 2007, the CIRANO Data Revisions Workshop 2007, the RBA Workshop 2007, Birkbeck College, the International Symposium on Forecasting 2008, the MMF Annual Conference 2008, and the Eurostat Colloquium 2008.

† Anne Sofie Jore, Norges Bank. James Mitchell (correspondence), NIESR, 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.. Shaun Vahey, Melbourne Business School, Norges Bank and Reserve Bank of New Zealand.

ISSN 1177-7567

1 Introduction

A number of studies have found that forecast combination using time-varying recursive-weights, based on historical forecast performance, is an ineffective strategy for improving point forecast accuracy. Among others, Stock and Watson (2004) and Clark and McCracken (2008) have found that an equal-weight strategy is effective in terms of Root Mean Squared forecast Error (RMSE). Forecasters and decision makers concerned with quadratic loss typically report that unknown structural changes, often referred to as "uncertain instabilities", hinder estimation and the usage of time-varying recursive weights.

The effectiveness of combination methods in the presence of uncertain instabilities for more general, but unknown, loss functions, has not been studied previously. This is surprising given the plausibility of asymmetric loss functions where the range of uncertainty about a point (conditional mean) forecast matters; see Granger and Pesaran (2000). For example, the Federal Reserve may not care equally about inflation above and below zero, but the exact loss function of the monetary policymaker is unknown.

In this paper, we examine the effectiveness of both recursive and equal-weight combination strategies for forecast densities. To facilitate comparisons with the literature on point forecast combination, we follow Clark and McCracken (2008) [CM, 2008] and consider the forecasts from Vector Autoregressive models (VARs) of output, prices and the short-term interest rate using the same real-time US data. The models include a wide range of VARs and ARs, which differ in their sensitivity to structural changes, including full sample, rolling windows, and break-date variants.

As discussed by Timmermann (2006), the literature on forecast density combination has produced a number of feasible alternatives, with no consensus on the "best" approach. In keeping with the simple convex mix of point forecasts considered by CM (2008) and others, we consider the analogous convex combinations of probability forecasts known as the "linear opinion pool"; see Timmermann (2006), p.177. We evaluate the resulting combined forecast densities using their probability integral transforms (*pits*).

The empirical analysis in this paper reveals that the recursive-weight (RW) density combinations are accurate both before, and after, the US Great Moderation. Our proposed RW scheme, based on the logarithmic score of the (component) forecast densities, gives significant weight to models that allow for the shifts in volatilities associated with the Great Moderation. In con-

trast, the equal-weight (EW) strategy produces inaccurate forecast densities after 1984; equal-weights attributes excessive weight to constant parameter models which ignore the scope for volatility breaks.

The remainder of this paper is structured as follows. In section 2 we describe our methods for forecast density combination and evaluation. In section 3, we outline briefly the data set and models considered in both this paper and CM (2008). In section 4 we present the results, and, in the final section, we conclude and 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 equal-weight average used by CM (2008) can be obtained as a special case by taking an equal-weighted average of the means of the individual densities.

2.1 Forecast density combination

We formalize density combination in a way that extends the commonly-adopted convex mix of point forecasts by utilizing the linear opinion pool approach; see 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:[1]

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

where $g(y_{\tau,h} \mid I_{i,\tau})$ are the h-step ahead forecast densities from component 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.

[1] The linear opinion pool is sometimes justified by considering an expert combination problem. See for example, Morris (1974), Morris (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 and Hall (2005) combine inflation density forecasts from different institutions.

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 $\tau - h$. The non-negative weights, $w_{i,\tau,h}$, in this finite mixture sum to unity.[2] Furthermore, the weights may change with each recursion in the evaluation period $\tau = \underline{\tau}, \ldots, \overline{\tau}$.

Since each VAR and AR considered produces a forecast density that is multivariate Student-t (see the discussion in section 3.3), the combined density defined by equation (1) will be a mixture—accommodating skewness and kurtosis. That is, the combination delivers a more flexible distribution than each of the individual densities from which it was derived. As N increases, the combined density becomes more and more flexible, with the potential to approximate non-linear specifications. Notice that our focus is on the predictive accuracy of the combination, rather than the (many) individual VAR and AR components.

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

Recursive weights

We construct the weights based on the fit of the individual model forecast densities. Following Amisano and Giacomini (2007) and Hall and 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 assigns a high probability to the realized value.[3] Specifically, the recursive weights for the h-step ahead densities take the form:

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

where the $\underline{\tau} - 10$ to $\underline{\tau}$ comprises the training period used to initialize the weights. Computation of these weights is feasible even for the large N con-

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

[3] The logarithmic score of the i-th density forecast, $\ln g(y_{\tau,h} \mid I_{i,\tau})$, is the logarithm of the probability density function $g(. \mid I_{i,\tau})$, evaluated at the outturn $y_{\tau,h}$.

sidered in our application. Given the uncertain instabilities problem, the recursive weights should be expected to vary across τ.

From a Bayesian perspective, density combination based on recursive logarithmic score weights, RW, has many similarities with an approximate predictive likelihood approach (see Raftery and Zheng 2003, and Eklund and Karlsson 2007).[4] 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 and Whiteman 2006 provide recent general discussions of Bayesian model averaging methods.) Andersson and Karlsson (2007) propose Bayesian predictive likelihood methods for forecast combination with Bayesian VARs but do not consider forecast density evaluation. Hall and Mitchell (2007) and Geweke and Amisano (2008) consider iterative algorithms to select weights that maximize the logarithmic score, suitable for small N.

Equal weights

The EW strategy attaches equal (prior) weight to each model with no updating of the weights through the recursive analysis: $w_{i,\tau,h} = w_{i,h} = 1/N$. In this version of the paper, we present results for the EW strategy without (prior) truncation of the set of models to be included.[5]

2.2 Evaluation of forecast density combinations

In constructing the RW forecast densities, we evaluate forecasts using the logarithmic score at 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. These weights provide an indication of whether the support for the component models is similar, or not, based on the score

[4] When $h > 1$ the product (over the out-of-sample window) of the h-step ahead forecast densities does not equal the marginal likelihood of the out-of-sample data.

[5] CM (2008) also consider pairwise equal weight combinations which involve weighting equally the point forecasts from just two models. The working paper version of our paper reports density evaluations for that strategy, that are comparable to the EW strategy evaluated below; see Jore *et al* (2008). We note, however, that it is difficult to justify an *a priori* truncation. Put differently, a researcher would have faced considerable uncertainty about which pair of models to select according to the PEW strategy at the start of our evaluation period.

of the individual densities. A finding of similar weights across component models would be consistent with the equal-weight strategy.

A common approach to forecast density evaluation provides statistics suitable for one-shot tests of (absolute) forecast accuracy, relative to the "true" but unobserved density. A popular method, following Rosenblatt (1952), Dawid (1984) and Diebold *et al* (1998), evaluates using the probability integral transforms (*pits*) of the realization of the variable with respect to the forecast densities. A density forecast can be considered optimal (regardless of the user's loss function) if the model for the density is correctly conditionally calibrated. That is, if the *pits* $z_{\tau,h}$, where $z_{\tau,h} = \int_{-\infty}^{y_{\tau,h}} p(u)du$, are uniform and, for one-step ahead forecasts, independently and identically distributed (see Diebold, Gunther, and Tay 1998). In practice, therefore, density evaluation with the *pits* requires application of tests for goodness-of-fit and independence at the end of the evaluation period.[6]

The goodness-of-fit tests employed include the Likelihood Ratio (LR) test proposed by Berkowitz (2001), for which results are presented at $h > 1$ using a two degrees-of-freedom variant (without a test for autocorrelation; see Clements 2004). For $h = 1$, we use a three degrees-of-freedom variant with a test for independence, where under the alternative $z_{\tau,h}$ follows an AR(1) process. 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.[7] We also consider the Anderson-Darling (AD) test for uniformity, a modification of the Kolmogorov-Smirnov test, intended to give more weight to the tails (and advocated by Noceti, Smith, and Hodges 2003). Mitchell and Wallis (2008) show that the AD test tends to over-reject in the presence of autocorrelation, which is to be expected in our specifications with $h > 1$, resulting from the overlapping nature of the forecasts. With this issue in mind, we follow Wallis (2003) and employ a Pearson chi-squared test which divides the range of the $z_{\tau,h}$ into eight equiprobable classes and tests whether the resulting histogram is uniform.

Turning to the test for independence of the *pits*, at $h = 1$ we use a Ljung-Box (LB) test, based on autocorrelation coefficients up to four.[8] For $h > 1$ we

[6] Given the large number of component densities under consideration, we do not allow for estimation uncertainty when evaluating the *pits*. Corradi and Swanson (2006) review *pits* tests computationally feasible for small N.

[7] Diks *et al* (2007) discuss the use of censored LR tests for assessing relative forecast performance; we test for absolute forecast performance.

[8] To investigate possible higher order dependence we also undertook tests in the second and third powers of the *pits*; results were similar to the first power.

test for autocorrelation at lags greater than or equal to h using a modified LB test (MLB).

3 The US data, component models, and their predictive densities

In this section, we describe the US data which span both the Great Inflation and the Great Moderation. We also list the component models used to construct forecast density combinations, and outline the construction of the forecasts from those individual components.

3.1 Data issues

To facilitate comparisons with CM (2008), we use the same real-time US data set and, like them, estimate VAR and AR models in output growth, inflation and the short-term interest rate. That is, we use the same economic variables, and their respective measures. Output growth and inflation are measured as annualized logarithmic changes from t to $(t-1)$; and interest rates are measured in annualized percentage points.

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 and 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.)

The start dates for the GDP observations vary by vintage, dictated by data

availability in the FRB Philadelphia real-time database. We use $1955Q1$ as the first observation for estimation of the component 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 by recursion for rolling sample and break models. Each individual component 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 $2005Q4$.[9] 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.

3.2 Component models

The component VAR and AR models considered are listed in table 1.[10] Our models include ARs, VARs, first differenced VARs (DVARs), de-trended VARs (using an exponential smoother) and bivariate VARs always including the variable of interest. Both full-sample and rolling sample VARs (ARs) are estimated, with the latter using the last x quarters only, with $x = 40$ 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.[11]

[9] 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 $\underline{\tau} - 10$.

[10] For computationally tractability, we focus on VARs and ARs and do not consider the BVAR and factor models analyzed by CM (2008) which typically rank poorly in their forecast evaluations.

[11] The Akaike Information Criterion ranking of the models is very similar to BIC and can be obtained from the authors on request.

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+break)	as above but estimated for each candidate break date (see Section 3.2) and 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+break)	as above but estimated for each candidate break date and with a rolling sample of 60 obs ... 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+break)	as above but estimated for each candidate break date and 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+break)	as above but estimated for each candidate break date and 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+break)	as above but estimated for each candidate break date and with a rolling sample of 60 obs
EW	equal-weight average of all AR and VAR models, including the break 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). $\pi^* = \pi^*_{-1} + .05(\pi - \pi^*_{-1})$.

Consideration of the rolling variants of our many VAR and AR specifications is a computationally convenient strategy to deal with an unknown number of structural breaks, of unknown timing; that is, as a pragmatic response to uncertain instabilities. Although rolling window models offer some protection against structural breaks, they may lead to biased forecasts if a break occurs after the beginning of the rolling window, or overstate uncertainty if the (last) break occurs before the beginning of the rolling window. Hence, we enrich the model space to consider a class of models which allow for multiple breaks, of unknown number and timing, in both the conditional mean and variance.

This is achieved following Garratt *et al* (2008b) by estimating a given component model, for a given data vintage, with each candidate break date. Thereby we accommodate break-date uncertainty in a computationally convenient manner. In our RW combination, the weights based on density fit favor break dates which produce more accurate forecast densities; the EW combination treats each break-date model as equally likely.

Specifically, consider a single equation from a given VAR model. We allow for multiple breaks in the conditional mean and conditional variance of Y_t by using the following framework

$$Y_t = \begin{cases} \alpha_1 + \beta_1 X_{t-h} + \sigma_1 \varepsilon_t & \text{if} \quad s_t = 1 \\ \alpha_2 + \beta_2 X_{t-h} + \sigma_2 \varepsilon_t & \text{if} \quad s_t = 2 \\ \vdots & \\ \alpha_R + \beta_R X_{t-h} + \sigma_R \varepsilon_t & \text{if} \quad s_t = R \end{cases} \tag{3}$$

where X_{t-h} is an appropriately defined matrix of lags of the variables concerned dated $(t-h)$, $\varepsilon_t \sim i.i.d.\ N(0,1)$ and s_t is determined as follows

$$s_t = \begin{cases} 1 & \text{if} \quad t < \tau_1 \\ 2 & \text{if} \quad \tau_1 \leq t \leq \tau_2 \\ \vdots & \\ R & \text{if} \quad t > \tau_{R-1} \end{cases} \tag{4}$$

so that structural breaks in the conditional mean and conditional variance occur at times $(\tau_1, \ldots, \tau_{R-1})'$.

For computational simplicity, we restrict the break dates to be identical across equations for each VAR model, and consider every feasible break date value with a regime containing at least 15% of the observations. Even so, with new break models included for each recursion in the evaluation period,

the computational burden is considerable. Hence, we further restrict the maximum number of regimes, R, to 3.

With these additional structural break models added to the set of rolling and full sample VARs and ARs (without breaks), we consider a maximum of 2022 models for each recursion in our evaluation period.

3.3 Component model forecast densities

We utilize a direct forecast methodology to generate the h-step ahead predictive densities from each VAR (AR). Consider a single equation from a given (constant parameter) VAR model

$$Y_t = \alpha + \beta X_{t-h} + \sigma\varepsilon_t, \tag{5}$$

where $t = 1, \ldots, \tau$ refers to the sample used to fit the model, $\varepsilon_t \sim i.i.d.$ $N(0,1)$ and the parameters are collected in α, β and σ. The predictive densities for $Y_{\tau+h}$ (with non-informative priors), allowing for small sample issues, are multivariate Student-t; see Zellner (1971), pp. 233-236 and, for a more recent application, Garratt *et al* (2008a). Recall that the break models are locally linear and Gaussian so that the predictive densities from break models are also multivariate Student-t.

4 Results

We break our results into two parts: the end of evaluation period weights on the model types (e.g., rolling/break models and full sample variants) derived from the logarithmic score of the component forecast densities; and, the evaluations of the recursive weight, RW, and equal weight, EW, strategies for combination. In each case, we present results for the evaluation period split into two subperiods, 1970-1984 and 1985-2005, for horizons 1 and 5.[12] We present results forecasting inflation, output growth and the interest-rate.

In tables 2 and 3, for each variable of interest, we present the weights on each component type implied by the RW combination at the end of each evaluation period, for horizons $h = 1$ and $h = 5$, respectively. Recall that these weights use second measurements as "final data", with density fit measured by the logarithmic score.

[12] The results for other horizons up to 8 are qualitatively similar and can be obtained from the authors on request.

Table 2
Recursive Weights at the end of the evaluation period: $h = 1$

	GDP growth		Inflation		Interest rate	
	1970-1984	1985-2005	1970-1984	1985-2005	1970-1984	1985-2005
AR	0.000	0.000	0.000	0.000	0.000	0.000
AR rolling+break	0.001	0.042	0.000	0.365	0.095	0.008
VAR	0.001	0.000	0.000	0.000	0.000	0.000
VAR rolling+break	0.009	0.030	0.002	0.042	0.861	0.144
DVAR	0.000	0.000	0.041	0.000	0.000	0.000
DVAR rolling+break	0.004	0.020	0.941	0.217	0.021	0.396
Inf detrend	0.004	0.000	0.000	0.000	0.000	0.000
Inf detrend rolling+break	0.946	0.827	0.002	0.027	0.016	0.419
BiVAR	0.002	0.000	0.003	0.000	0.000	0.000
BiVAR rolling+break	0.034	0.080	0.012	0.349	0.007	0.034

Looking first in table 2, at the $h = 1$ case, we see that although one model type never receives all the weight, many types do receive zero weight. That is, the logarithmic score metric implies there is considerable model uncertainty, and that the equal weighting of all models, EW, is not supported.

Closer examination of the rows in table 2 reveals that the highest weighted models are the rolling window and break variants. The full-sample models receive less than 5% weight in all cases, with many zeros.

The weights in table 2 also reveal considerable variation across variables and time. For example, the rolling and break variant models with inflation detrending receive a high weight for GDP growth—in excess of 80%—but this specification receives less than 5% weight for inflation. For these variables, the weights on the detrended VARs vary little by evaluation period. But for interest rates, the weight is around 40% for that component type for 1985-2005, but less than 2% for 1970-1984.

In general, the earlier evaluation period gives greater weight to full-sample variants than the subsequent evaluation period. And, in all cases the full-sample variants receive less than their counterparts with rolling samples and structural breaks.

Another striking result from table 2 is that the full sample AR models receive no weight across the two evaluation periods. Despite the well-documented competitive performance of these models for RMSE, AR models do not typically produce accurate forecast densities.

Turning to table 3, we see that the $h = 5$ results are fairly similar to the short-horizon results of table 2. There is the same basic story of considerable model uncertainty, little evidence of support for EW, strong support for rolling and break variants, variation in weights across variables and evaluation periods, and low weight on full-sample AR specifications. It is worth noting that the full-sample detrended inflation models receive a weight of 33% for inflation

Table 3
Recursive Weights at the end of the evaluation period: $h = 5$

	GDP growth		Inflation		Interest rate	
	1970-1984	1985-2005	1970-1984	1985-2005	1970-1984	1985-2005
AR	0.000	0.000	0.000	0.000	0.000	0.000
AR rolling+break	0.015	0.009	0.000	0.056	0.000	0.000
VAR	0.004	0.000	0.000	0.000	0.000	0.000
VAR rolling+break	0.025	0.025	0.000	0.002	0.000	0.005
DVAR	0.003	0.000	0.000	0.000	0.000	0.000
DVAR rolling+break	0.092	0.017	0.000	0.874	0.000	0.000
Inf detrend	0.016	0.000	0.331	0.000	0.000	0.036
Inf detrend rolling+break	0.114	0.876	0.667	0.048	1.000	0.957
BiVAR	0.113	0.000	0.000	0.000	0.000	0.000
BiVAR rolling+break	0.618	0.073	0.002	0.020	0.000	0.001

over the period 1970-1984. And the full-sample bivariate VARs receive a weight of just over 10% for output growth in the earlier evaluation period. So at this longer horizon, there is weak evidence that models without breaks do matter prior to the Great Moderation. For the later evaluation period, the RW strategy puts very little weight on components with constant parameters.

The p-values for our *pits* tests are displayed in tables 4 and 5, together with the AD statistic (which has a 95% critical value of 2.5). The GDP growth, inflation and interest rate densities are evaluated in table 4 for the 1970-1984 evaluation period and, in table 5, for the 1985-2005 period. As with the previous tables, we report results by forecast horizon: $h = 1$ in the left hand panel and $h = 5$ in the right. To facilitate easy-reading of the two tables, we place the p-values, or test statistics, in bold when the density forecast is correctly calibrated at a 95% significance level—that is, when we cannot reject the null that the densities are correctly specified. In both tables, the evidence that the combination densities are correctly specified is much stronger for output growth and inflation, regardless of the combination strategy. Hence, we discuss summary statistics with and without interest rates.[13]

Table 4 reveals that the RW combination is correctly calibrated, at a 95% level, 19/24 times for inflation and output growth over the 1970-1984 eval-

[13] The weak performance of the interest rate density forecasts from our VARs and ARs is consistent with the view that a wider set of factors explain interest rates than lagged inflation, output growth and interest rates.

Table 4
Density forecast evaluation using the *pits*: 1970-1984

		$h=1$						$h=5$					
		LR3	LR_l	LR_u	AD	χ^2	LB	LR2	LR_l	LR_u	AD	χ^2	MLB
AR	GDP	**0.75**	**0.30**	**0.82**	**0.45**	**0.40**	**0.95**	0.03	0.00	0.00	**2.31**	0.10	**0.05**
	Inf	0.02	**0.30**	**0.12**	4.60	**0.09**	**0.07**	0.00	0.00	0.00	18.7	0.00	**0.92**
	roi	0.00	0.00	0.00	11.4	0.00	0.01	0.00	0.00	0.00	27.7	0.00	**0.53**
EW	GDP	**0.86**	**0.54**	**0.63**	**0.68**	**0.93**	**0.98**	**0.93**	**0.08**	**0.45**	**0.92**	**0.55**	0.02
	Inf	**0.49**	**0.18**	**0.15**	**0.65**	**0.55**	**0.06**	0.00	**0.95**	0.00	**2.12**	**0.72**	**0.60**
	roi	0.00	0.00	**0.07**	3.01	**0.09**	**0.14**	0.00	0.00	0.00	13.6	0.00	**0.68**
RW	GDP	**0.09**	**0.14**	**0.29**	**1.69**	**0.86**	**0.96**	**0.52**	**0.08**	**0.75**	**1.08**	**0.76**	**0.05**
	Inf	**0.16**	**0.59**	**0.07**	**1.10**	**0.81**	**0.07**	0.00	0.00	0.00	16.9	0.00	**0.86**
	roi	0.00	0.00	0.00	4.17	0.02	0.02	0.00	0.00	0.00	26.1	0.00	**0.98**

Notes: AR is a full-sample AR(2); EW is the equal-weighted combination; RW the recursive weights combination; LR2 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*; LR3 supplements LR2 with a test for zero first order autocorrelation. 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. χ^2 is the p-value for the Pearson chi-squared test of uniformity of the *pits* histogram in eight equiprobable classes. LB is the p-value from a Ljung-Box test for independence of the *pits*; MLB is a modified LB test which tests for independence at lags greater than or equal to h.

uation period. When we consider interest rates, the ratio drops to 20/36.[14] The RW combination delivers densities that are reasonably well calibrated by many measures. But in this Great Inflation period, the EW combination performs very well too, with correct calibration ratios of 21/24 (without interest rates), and 26/36 (with interest rates). It seems that the density forecasts from the EW strategy perform well prior to the Great Moderation. The benchmark AR (full sample) specifications perform adequately for inflation and output growth, giving a ratio of 13/24, and with interest rates, the ratio is a less impressive 14/36.

The *pits* results over the 1985-2005 period shown in table 5 are quite distinct. RW continues to forecast well, correctly calibrated 18/24 times (without interest rates) and 23/36 times (with interest rates). We conclude that the RW adapts well to uncertain instabilities, by shifting the weight to components that allow for parameter change. Strikingly, the EW densities perform very poorly. EW delivers correct calibration ratios of only 7/24 and 9/36, with and without interest rates, respectively. The AR benchmark results in ratios of 3/24 (excluding interest rates) and 8/36 (with interest rates).

As CM (2008) discuss, simple AR benchmarks forecast well over this Great

[14] These ratios simply summarize information from tables 4 and 5 - where the null hypothesis of correct calibration for the density forecasts in each case (i.e., each forecasting strategy, variable and horizon) is tested separately by each of the *pits* tests at a 95% significance level. To control the joint size of the many (six in each case) evaluation tests requires use of a stricter p-value; e.g., a Bonferroni correction would suggest using $(100\% - 95\%)/6 = 0.8\%$. In any case, the relative performance of the different forecasting strategies is similar if we use a stricter threshold.

Table 5
Density forecast evaluation using the *pits*: 1985-2005

		$h=1$						$h=5$					
		LR3	LR_l	LR_u	AD	χ^2	LB	LR2	LR_l	LR_u	AD	χ^2	MLB
AR	GDP	0.00	0.00	0.00	8.43	0.00	**0.92**	0.00	0.00	0.00	8.47	0.00	**0.92**
	Inf	0.00	0.00	0.00	6.52	0.00	0.01	0.00	0.01	0.00	14.2	0.00	**0.76**
	roi	0.00	**0.16**	0.00	8.14	0.00	0.00	**0.08**	**0.70**	0.02	4.78	**0.08**	**0.61**
EW	GDP	0.00	0.00	0.01	6.14	0.00	**0.79**	0.00	0.03	0.00	6.11	0.00	**0.90**
	Inf	0.01	**0.09**	0.02	**2.04**	0.03	**0.08**	0.00	**0.17**	0.00	6.67	0.00	**0.64**
	roi	0.00	0.03	0.00	8.26	0.00	0.00	0.00	**0.29**	0.00	8.31	0.00	**0.71**
RW	GDP	0.01	0.02	**0.41**	2.87	0.01	**0.53**	0.01	**0.13**	**0.29**	**2.36**	0.01	**0.65**
	Inf	**0.48**	**0.13**	**0.56**	**1.25**	**0.08**	**0.51**	**0.13**	**0.19**	**0.29**	**2.15**	**0.53**	**0.89**
	roi	0.00	**0.09**	0.01	3.31	0.00	0.01	0.01	**0.34**	**0.05**	4.85	**0.07**	**0.89**

See Notes to table 4.

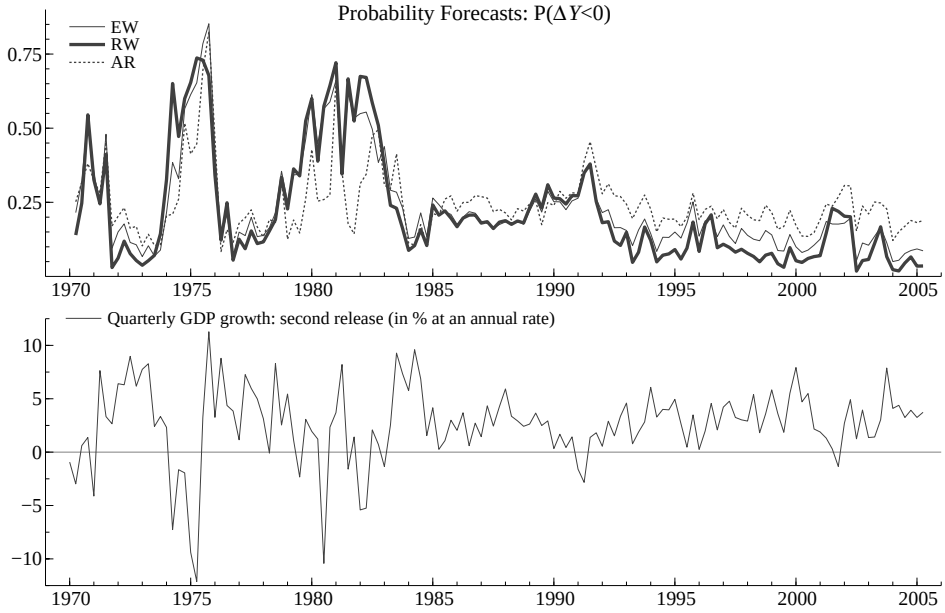
Moderation sample on the basis of RMSE. And, there is little evidence that the RW strategy dominates EW on this particular measure of (point) forecast performance. For example, over 1985-2005, at $h = 1$ ($h = 5$), EW has a RMSE for inflation and output growth, respectively, of 1.02 and 1.81 (1.22 and 1.98); while RW has a RMSE of 1.01 and 1.85 (1.33 and 2.07). In turn, a full-sample AR(2) has a RMSE of 1.03 and 1.79 (1.33 and 2.01).

We draw three main conclusions. First, although the equal-weight density combination strategy works well for Great Inflation data, density forecast performance is weaker for the Great Moderation. Second, the recursive weight strategy performs well over both evaluation periods. It seems to adapt particularly well to the shifting volatilities associated with the Great Moderation by increasing the weight on components that allow for structural change. A third finding is that simple full-sample autoregressive specifications produce particularly poor densities after the mid-1980s.

Figure 1 illustrates how EW density combinations can produce inaccurate forecast densities for output growth in the presence of the shifting volatilities exhibited in the US sample. The top panel of figure 1 plots the 1-step ahead probability that output growth is less than zero percent. The bottom panel plots the second release realization of output growth. Despite the occasional large fluctuation, the top panel of figure 1 reveals that the benchmark (full-sample) AR tracks EW and RW quite well in terms of its probabilistic forecasts over the Great Inflation period. But from the mid 1980s we observe that both AR and EW do not pick up the shifting volatilities as well as RW. Output growth very rarely drops below the zero percent threshold for Great Moderation data—only once in the last 10 years of the sample—as noted by Potter (2007). The density forecasts from an AR(2) estimated over the full-sample period give a poor indication of the probability of this particular event, forecasting a 20 to 30 percent probability of a (one-period) recession for most of the last 10 years. Broadening the model space to take an equal

weighted combination across all the models considered, including rolling and break models, does deliver some improvement. But the probability forecasts from EW are still too high, at between 10 to 20 percent over most of the last 10 years. However, using weights based on the recursive logarithmic score produces more accurate probabilities.[15]

Figure 1
Probability forecasts for GDP growth



5 Conclusions

A growing body of empirical work has demonstrated the effectiveness of taking averages of point forecasts using equal weights. In this paper, we have

[15] A similar picture emerges when we plot the 1-step ahead probability for various analogous inflation events, such as the probability that inflation is higher than five percent.

shown that neither (full sample) univariate nor equal-weight combinations produce accurate real-time forecast densities for Great Moderation data. Recursively constructed weights give greater weight to rolling and break models that allow for the shifts in volatilities. As a result, the recursive-weight density combination strategy gives accurate forecast densities in the presence of uncertain instabilities.

In future work, we plan to explore the scope for recursive density combination with alternative classes of models, including dynamic stochastic general equilibrium models and factor models.

References

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

Andersson, M and S Karlsson (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 and M W McCracken (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, A Fernandez, and N R Swanson (2007), "Information in the revision process of real-time data," Discussion Paper, Rutgers University.

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

Croushore, D and T Stark (2001), "A real-time data set for macroeconomists," *Journal of Econometrics*, 105, 111–130.

Dawid, A P (1984), "Statistical theory: the prequential approach," *Journal of the Royal Statistical Society B*, 147, 278–290.

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

Diks, C, V Panchenko, and D van Dijk (2007), "Weighted Likelihood Ratio scores for evaluating density forecasts in tails," Mimeo, University of New South Wales.

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

Garratt, A, G Koop, E Mise, and S P Vahey (2008a), "Real-time prediction with UK monetary aggregates in the presence of model uncertainty," *Journal of Business and Economic Statistics*, forthcoming. Available as Birkbeck College Discussion Paper No. 0714.

Garratt, A, G Koop, and S P Vahey (2008b), "Forecasting substantial data revisions in the presence of model uncertainty," *Economic*

Journal, 118(530), 1128–1144.

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

Geweke, J and G Amisano (2008), "Optimal prediction pools," Department of Economics, University of Iowa Working Paper.

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

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

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

Jore, A S, J Mitchell, and S P Vahey (2008), "Combining forecast densities from VARs with uncertain instabilities," Working Paper, NIESR and Norges Bank.

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

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

Marcellino, M, J Stock, and M Watson (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 and S G Hall (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.

Mitchell, J and K F Wallis (2008), "Evaluating density forecasts: is sharpness needed?" National Institute of Economic and Social Research Discussion Paper No. 320.

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, J Smith, and S Hodges (2003), "An evaluation of tests of distributional forecasts," *Journal of Forecasting*, 22, 447–455.

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

Raftery, A E and Y Zheng (2003), "Long-run performance of Bayesian model averaging," *Journal of the American Statistical Association*,

98, 931–938.

Rosenblatt, M (1952), "Remarks on a multivariate transformation," *The Annals of Mathematical Statistics*, 23, 470–472.

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

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

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

Wallis, K F (2003), "Chi-squared tests of interval and density forecasts, and the Bank of England's fan charts," *International Journal of Forecasting*, (19), 165–175.

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.

Zellner, A (1971), *An introduction to Bayesian inference in econometrics*, New York: John Wiley and Sons.