CENTRE FOR APPLIED MACROECONOMIC ANALYSIS

The Australian National University

CAMA Working Paper Series **December, 2010**

REAL-TIME INFLATION FORECAST DENSITIES FROM ENSEMBLE PHILLIPS CURVES

Anthony Garratt
Birkbeck College, University of London,

James Mitchell
National Institute of Economic and Social Research, UK

Shaun P. Vahey
Centre for Applied Macroeconomic Analysis, ANU

Elizabeth C. Wakerly
Centre for Applied Macroeconomic Analysis, ANU

CAMA Working Paper 34/2010
http://cama.anu.edu.au

Real-time Inflation Forecast Densities from Ensemble Phillips Curves*

Anthony Garratt
(Birkbeck College, University of London, UK)

James Mitchell
(National Institute of Economic and Social Research, UK)

Shaun P. Vahey†
(Centre for Applied Macroeconomic Analysis, ANU, Australia)

Elizabeth C. Wakerly
(Centre for Applied Macroeconomic Analysis, ANU, Australia)

December 20, 2010

Abstract

We examine the effectiveness of recursive-weight and equal-weight combination strategies for forecasting using many time-varying models of the relationship between inflation and the output gap. The forecast densities for inflation reflect the uncertainty across models using many statistical measures of the output gap, and allow for time-variation in the ensemble Phillips curves. Using real-time data for the US, Australia, New Zealand and Norway, we find that the recursive-weight strategy performs well, consistently giving well-calibrated forecast densities. The equal-weight strategy generates poorly-calibrated forecast densities for the US and Australian samples. There is little difference between the two strategies for our New Zealand and Norwegian data. We also find that the ensemble modelling approach performs more consistently with real-time data than with revised data in all four countries.

Keywords: Density combination; Ensemble forecasting: Phillips curve
JEL classification: C32, C53, E37

*Financial support from the ESRC (grant No. RES-062-23-1753) and the ARC (grant No. LP0991098) is gratefully acknowledged. We benefitted greatly from comments by Dean Croushore, Domenico Giannone, James Hamilton, Kirdan Lees, Simon van Norden, Les Oxley, David Papell and two referees. Anne Sofie Jore and Tim Robinson kindly provided real-time data for Norway and Australia, respectively. We are also indebted to conference and seminar participants at the Reserve Bank of Australia and the CIRANO Data Revisions Workshop October 2008.

†Corresponding author. Tel: +61 2 61256220. Email address: shaun.vahey@anu.edu.au.

1 Introduction

A number of applied macro-econometric studies have found that forecast combination using recursive weights, based on historical forecast performance, is an ineffective strategy for improving point forecasts. Stock and Watson (2004), and Clark and McCracken (2010), among others, have found that an equal-weight strategy is more effective in terms of root mean squared forecast error. However, forecasters and policymakers are often interested in forecast densities rather than point forecasts. Considering the evidence for density forecasting performance, Jore, Mitchell and Vahey (JMV, 2010) report strong density forecasting performance from the recursive-weight strategy, but not from the equal-weight strategy, with vector autoregressions using US data. Garratt, Mitchell and Vahey (GMV, 2009) report similar findings for recursive weights for US inflation using a Phillips curve relationship based on the output gap, but do not consider the equal-weight strategy.

In this paper, we investigate the generality of the JMV finding by examining the recursive and the equal-weight strategies for inflation forecast densities in four countries. The model space is similar to that of GMV. That is, we consider many time-varying models of the relationship between inflation and the output gap. The forecast densities for inflation in each country reflect the uncertainty across models using many statistical measures of the output gap, and allow for time-variation in the ensemble Phillips curves.

To implement our recursive-weight strategy, we adopt the forecasting methodology proposed by JMV in which a real-time forecaster (or policymaker) recursively selects a combination of component forecasts from a set of models to produce an ensemble forecast density. Each component forecast density for inflation is produced by a single Phillips curve model for inflation based on lags of inflation and lags of the output gap. We utilise a "linear opinion pool" (LOP) to take out of sample density combinations (see Timmermann (2006, p.177)) using the logarithmic score, as a measure of the Kullback-Leibler distance, to generate component weights. The resulting ensemble approximates the unknown potentially non-linear data generating process for inflation by using time-varying weights; and, the ensemble forecast densities are not restricted to be Gaussian. To implement our equal-weight strategy, we adopt an analogous ensemble methodology, again using LOP, to take equal-weighted forecast density combinations. We evaluate the recursive and equal-weight strategies by examining the probability integral transforms (*pits*) of the ensemble densities for inflation.

We consider real-time data for the US, Australia, New Zealand and Norway. For each data set, we compare and contrast the inflation forecasts from equal-weight and recursive-weight ensembles. The recursive-weight strategy performs well across the real-time data sets, consistently giving well-calibrated forecast densities. The equal-weight strategy performs less consistently, generating poorly-calibrated forecast densities for the US and Australian samples in particular. There is little difference between the two strategies for our New Zealand and Norwegian data. We also find that the ensemble modelling approach performs more reliably with real-time data than with revised data in all four countries.

The remainder of this paper is structured as follows. In Section 2, we outline the component models. In Section 3, we describe our methods for ensemble forecasting and density evaluation. In Section 4, we apply our methodology to US, Australian, New

Zealand and Norwegian data and present the results. In the final section we conclude.

2 Component models

Following Orphanides and van Norden (2005) and GMV, we start with Phillips curve forecasting models of the Linear Gaussian form:

$$\pi_{t+h} = \alpha_1^j + \sum_{p=1}^{P} \beta_{1,p}^j \pi_{t-p+1} + \sum_{p=1}^{P} \gamma_{1,p}^j y_{t-p+1}^j + \sigma_1^j \varepsilon_{1,t+h}^j, \tag{1}$$

where inflation is defined as the log difference in the price level, and there are many output gap measures denoted y_t^j, where $j = 1, \ldots, J$; the number of lags of inflation and the output gap is denoted P, $\varepsilon_{1,t+h}^j \sim i.i.d.\ N(0,1)$ and h is the forecast horizon.[1] The predictive densities for π_{t+h} denoted $g(\pi_{\tau,h} \mid I_{i,\tau})$, allowing for small sample issues (with non-informative priors) are Student-t; see Zellner (1971, 233-236) and, for a more recent application, Garratt, Koop, Mise and Vahey (2009).

A number of Phillips curve studies have noted the scope for parameter change to improve forecasting performance; see, for example, Groen, Paap and Ravazzolo (2009). Accordingly we expand the model space to allow for a single structural break of unknown timing, assuming a coincident break in the conditional mean and variance, in each Phillips curve specification. That is, we forecast using a variety of expanding windows for parameter estimation. With the computational burden in mind, the break date is restricted to occur before the start of the evaluation period in which density combination occurs. The break models are locally linear and Gaussian so that the predictive densities from the break models are also Student-t.

To define the output gap we consider seven "flexible time trends", derived from univariate filters ($J = 7$), where the output gap is defined as the difference between observed output and unobserved potential (or trend) output for each detrending type. Let q_t denote the (logarithm of) actual output in period t reported in period $t+1$, and μ_t^j be its trend using definition j where $j = 1, 2, \ldots, J$. Then the output gap, y_t^j, is defined as the difference between actual output and its j^{th} trend measure. That is, we use the following trend-cycle decomposition:

$$q_t = \mu_t^j + y_t^j.$$

The seven methods of univariate trend extraction are: quadratic, Hodrick-Prescott, forecast-augmented Hodrick-Prescott (with forecasts generated from a recursively estimated univariate AR(8) model in output growth, using the appropriate vintage of data), Christiano and Fitzgerald, Baxter-King, Beveridge-Nelson, and Unobserved Components. An appendix describes the specification of each detrending approach.

For each country, we allow the maximum number of lags to vary between 1 and 4 ($P = 1, \ldots, 4$). With $J = 7$ detrending methods, we therefore consider 28 specifications without breaks, to be combined together with the break variants. The number of breaks considered—and therefore the total number of component models—varies by country

[1] We set $h = 1$ in our applications that follow.

according to the length of the sample in each case. We note that although the component model space defined in this paper uses single equation models, multi-equation extensions are feasible.

3 Ensemble methodology

We construct the predictive densities for the component models specified above using an ensemble methodology, following JMV and GMV. The ensemble methodology can approximate a non-Linear and/or non-Gaussian process with a combination of specifications based on an incomplete model space.

A recent paper by Bache, Mitchell, Ravazzolo and Vahey (2010) describes the embryonic ensemble forecasting literature in macro-econometrics, and provides a characterisation. Typically, ensemble modelling applications use many component specifications, with time-varying weights, to combine the evidence across components, with the aim of producing predictive densities for the variables or features of interest. Within the ensemble framework, there are a number of ways to implement the approach in practice. These include opinion pools and finite mixture models.

The opinion pool approach has a long tradition in management science, where the focus is on combining the evidence supplied by a number of experts to a decision-maker or policymaker. As emphasised by Wallis (2005), the approach is particularly useful for the combination of survey information since no information is required about the model used by each expert; see also Hall and Mitchell (2007). Common methods of pooling opinions include the linear opinion pool (LOP) and the logarithmic opinion pool. Kascha and Ravazzolo (2010) contrast the properties of linear and logarithmic opinion pools. Sometimes the application of opinion pools in economic forecasting is referred to as "density combination". JMV and GMV both utilise linear opinion pools to combine the forecast densities from several hundred vector autoregressive component models. The analysis of ensemble Phillips curves in this paper builds on this tradition; as does the study of forecasting at Norges Bank by Bjornland, Gerdrup, Jore, Smith, and Thorsrud (2010). Geweke (2009) discusses the differences between opinion pools and mixture models, providing specific empirical examples with small numbers of component models.

We note that the literature within economics on point forecast combinations (see for example, Clark and McCracken (2010); Smith and Wallis (2009); Stock and Watson (2004)) does not pertain to forecast densities. One motivation for the focus on density forecasts in our paper is that central banks such as the Bank of England and Norges Bank publish non-Gaussian forecast densities which are hard to reconcile with the (near) Linear Gaussian models predominantly used by the staff of experts. Hence, there are a number of similarities between our ensemble framework and the monetary policymaking environment.

3.1 Forecast density combinations

To construct our ensemble forecasts, we consider a policymaker who aggregates forecasts supplied by "experts". Each expert uses a unique Phillips curve specification to produce a forecast density for inflation based on Eq. (1). For simplicity, we ignore the variation in the number of component models by country, and index the N component Phillips curves $i = 1, \ldots, N$. The ensemble densities for inflation are defined by the convex combination:

$$p(\pi_{\tau,h}) = \sum_{i=1}^{N} w_{i,\tau,h}\; g(\pi_{\tau,h} \mid I_{i,\tau}), \qquad \tau = \underline{\tau}, \ldots, \overline{\tau}, \tag{2}$$

where $g(\pi_{\tau,h} \mid I_{i,\tau})$ are the h-step ahead forecast densities from model i, of inflation π_τ, conditional on the information set I_τ and $\tau = \underline{\tau}, \ldots, \overline{\tau}$ is the out of sample evaluation period. The publication delay in the production of real-time data ensures that this information set contains lagged variables, here assumed to be 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, Stock and Watson (2003). Hence, the macro variables used to produce an h-step ahead forecast density for τ are dated $\tau - h - 1$. (In applying this framework to the data from four countries we focus on $h = 1$ below.) The non-negative weights, $w_{i,\tau,h}$, in this finite mixture sum to unity.[2]

Since each component specification produces a forecast density that is Student-t, the combined density defined by the linear opinion pool described in Eq. (2) will be a mixture—accommodating the potential for multi-modality, skewness and kurtosis. That is, the policymaker seeks a more flexible distribution than each of the individual densities supplied by the experts from which it was derived. For large N, the combined density becomes very flexible, with the potential to approximate non-Linear and non-Gaussian specifications.

We consider two distinct strategies for constructing the weights, $w_{i,\tau,h}$: recursive weights and equal weights. In the former, the weights change with each recursion in the evaluation period $\tau = \underline{\tau}, \ldots, \overline{\tau}$. In the latter, we restrict the weights to be constant and equal throughout the evaluation.

3.1.1 Recursive weights (RW)

With the RW strategy, our policymaker constructs the ensemble weights based on the fit of the component forecast densities. Like Amisano and Giacomini (2007) and Hall and Mitchell (2007), we use the logarithmic score to measure density fit for each component model through the evaluation period. The logarithmic scoring rule gives a high score to a density forecast that assigns a high probability to the realised value and can be interpreted as a measure of the Kullback-Leibler distance. The logarithmic score of the i^{th} density forecast, $\ln g(\pi_{\tau,h} \mid I_{i,\tau})$, is the logarithm of the probability density function $g(. \mid I_{i,\tau})$, evaluated at the outturn $\pi_{\tau,h}$. Specifically, following JMV, the recursive weights for the h-step ahead densities take the form:

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

$$w_{i,\tau,h} = \frac{\exp\left[\sum_{\underline{\tau}-tr}^{\tau-1-h} \ln g(\pi_{\tau,h} \mid I_{i,\tau})\right]}{\sum_{i=1}^{N} \exp\left[\sum_{\underline{\tau}-tr}^{\tau-1-h} \ln g(\pi_{\tau,h} \mid I_{i,\tau})\right]}, \qquad \tau = \underline{\tau}, \ldots, \overline{\tau} \tag{3}$$

where the $\underline{\tau} - tr$ to $\tau - 1 - h$ window comprises the training period used to initialise the weights. Computation of these weights is feasible for a large N ensemble.

From a Bayesian perspective, density combination based on recursive logarithmic score weights, RW, has some similarities with an approximate predictive likelihood approach. Given our definition of density fit, the model densities are combined with equal (prior) weight on each model—which a Bayesian would term non-informative priors. Given these prior weights, we construct an aggregate forecast density for inflation (recursively, at each horizon). Nevertheless, there are important differences with predictive Bayesian model averaging. For example, since the policymaker using LOP assumes that the experts explore an incomplete model space, the conventional Bayesian interpretation of the weights as reflecting the posterior probabilities of the components is inappropriate. Accordingly, we do not consider model selection using the ensemble weights; nor do we consider strategies averaging a selection of component models.

3.1.2 Equal weights (EW)

The EW approach attaches equal (prior) weight to each model with no updating of the weights through the recursive analysis: $w_{i,\tau,h} = w_{i,h} = 1/N$. Simple combination strategies such as using equal weights have commonly been found to be effective in the point forecast combination literature within economics; see, among others, Stock and Watson (2004), Smith and Wallis (2009), and Clark and McCracken (2010).

3.2 Forecast density evaluations

A popular evaluation method for forecast densities, following (for example) Dawid (1984) and Diebold, Gunther and Tay (1998), evaluates the densities relative to the "true" but unobserved density for $\pi_{\tau,h}$ using the probability integral transforms (*pits*) of the realisation 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 calibrated. That is if the *pits*, $z_{\tau,h}$, defined as:

$$z_{\tau,h} = \int_{-\infty}^{\pi_{\tau,h}} p(u)du,$$

are uniform and, for one-step ahead forecasts, independently and identically distributed. In practice, therefore, density evaluation with the *pits* requires application of tests for both goodness-of-fit and independence at the end of the evaluation period; see Mitchell and Wallis (2010).[3]

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

The goodness-of-fit tests employed include the Likelihood Ratio (LR) test proposed by Berkowitz (2001). 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. Since the LR test has a maintained assumption of normality, 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 Nocetti, Smith and Hodges (2003)). We also 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. To test independence of the *pits*, we use a Ljung-Box (LB) test, based on autocorrelation coefficients up to four (with our quarterly data).

4 Applications

We begin our analysis by describing the sample data for each of our four countries. Then we present the results, focusing on the calibration properties of the inflation forecast densities for the two strategies EW and RW.

4.1 Data

In this section we describe the four samples used, for the US, Australia, New Zealand and Norway. Throughout our analysis, we use real-time observations for real output. We note that the availability of real-time data differs across the countries, with the US and Australian data sets covering longer periods in comparison with New Zealand and Norway.

4.1.1 United States

For the US, we use the same real-time US data set as Clark and McCracken (2010). The quarterly real-time data used refer to real GDP and the GDP price deflator. Here we use 83 vintages (seasonally adjusted data observed at a specific point in time), starting in 1987q1 and ending in 2007q3. The data, avoiding the period of the Korean War, begin in $1954q3$ and go through to $1986q4$ for the the first vintage and $2007q2$ for the last. That is, data on output and the price deflator are first released with a one quarter lag.

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 Product Accounts; each vintage reflects the information available around the middle of the respective quarter. Croushore and Stark (2001) provide a description of the database. The US evaluation period is: $\tau = \underline{\tau}, \ldots, \overline{\tau}$ where $\underline{\tau} = 1991q4$ and $\overline{\tau} = 2007q3$ (64 observations), as we drop the first 20 quarters to initialise weights (training period, $tr = 20$, in Eq. (3)), reflecting the large sample size available in the case of the US. To implement density combination through the evaluation

feasible for small N.

period requires an additional assumption about which measurement is to be forecast. Following Clark and McCracken (2010), JMV and others, we 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. We emphasise that our sequential use of vintages of real-time data is intended to replicate the approach adopted by forecasters in practice; see Cook (2008) and Corradi, Fernandez and Swanson (2010) for further discussion.

To repeat, for all four countries, we consider forecasting models based on the Phillips curve given by Eq. (1), with lag lengths of one to four ($P = 1, 2, 3, 4$) and have $J = 7$ detrending methods to define the output gap. We also allow for a single structural break of unknown timing in each component model. The break occurs in the conditional mean and the variance. This pragmatic treatment of structural breaks implies that we out of sample forecast using a variety of expanding windows for parameter estimation. The break date is restricted to occur before the start of the evaluation period to reduce the computational burden. When considering structural breaks, each regime is restricted to be at least 15 percent of the (pre and post break) sample length. Hence in the case of the US, we consider 376 component models for each measure of the output gap considered. With seven measures of the output gap derived from flexible trends, the predictive densities combine 2632 component specifications for each observation in the evaluation period.

4.1.2 Australia

Our real-time real output data for Australia were obtained from Gruen, Robinson and Stone (2002) and Stone and Wardrop (2002). There are 63 vintages of quarterly real GDP seasonally-adjusted data starting in 1991q3 and ending in 2007q4. The data for each vintage begin in $1959q3$, where the end period is $1991q2$ for the first vintage and $2007q3$ for the last, hence data on output are first released with a one quarter lag. As real-time data for prices are not available for Australia, this is also true of New Zealand and Norway, we use the consumer price index from a single vintage. The consumer price series was downloaded from the IMF's International Financial Statistics data base, dated July 2009. The Australian evaluation period is: $\tau = \underline{\tau}, \ldots, \overline{\tau}$ where $\underline{\tau} = 1996q2$ and $\overline{\tau} = 2007q4$ (47 observations), where we drop the first 20 quarters to use as a training period. As the sample sizes are comparable to those used for the US, structural breaks were handled in an identical manner, requiring a minimum of 15 percent of the sample, for each recursion, for all regressions. Hence the number of component models for each measure of the output gap is 356, making for a total of 2492 models to be combined in the evaluation period.

4.1.3 New Zealand

Our real-time real output data for New Zealand were obtained from the Reserve Bank of New Zealand (described in detail at www.rbnz.govt.nz/research/2482495.html). There are 40 vintages of quarterly real GDP seasonally-adjusted data starting in 1998q1 and ending in 2007q4. The data for each vintage begin in $1987q2$, where the end period is $1997q4$ for the first vintage and $2007q3$ for the last. Data on output are first released

with a one quarter lag. The consumer price series was downloaded from the IMF's International Financial Statistics data base in July 2009. The New Zealand evaluation period is: $\tau = \underline{\tau}, \ldots, \overline{\tau}$ where $\underline{\tau} = 1999q1$ and $\overline{\tau} = 2007q4$ (36 observations). Given the shorter sample, we drop just 5 observations to use as the training period to initialise weights. Similar considerations were also applied when dealing with structural breaks, where at each recursion a minimum of 50 percent of the sample is used for estimation. As a consequence the number of component models for each measure of the output gap is 48, making for a total of 336 models to be combined in the evaluation period.

4.1.4 Norway

The real-time real output data for Norway were obtained from Norges Bank.[4] There are 29 vintages of quarterly real GDP seasonally-adjusted data starting in 2001q2 and ending in 2008q2. The data for each vintage begin in $1978q1$, where the end period is $2001q1$ for the first vintage and $2008q1$ for the last. Data on output are first released with a one quarter lag. The consumer price series was downloaded from the IMF's International Financial Statistics data base in July 2009. The Norwegian evaluation period is: $\tau = \underline{\tau}, \ldots, \overline{\tau}$ where $\underline{\tau} = 2002q2$ and $\overline{\tau} = 2008q2$ (25 observations). For Norway we drop 5 observations for the training period and restrict the sample size when considering structural breaks to a minimum of 50 percent of the sample for each recursion. Hence the number of component models for each measure of the output gap is 208, making for a total of 1456 models to be combined in the evaluation period.

4.2 Results

In this section, we present our results on the calibration properties of the forecast densities resulting from our ensemble methodology for both the EW and RW strategies. We begin with the US results (which we treat separately on the grounds that the real-time data are of exceptional quality), and then turn to the remaining three countries.[5]

4.2.1 US

Table 1 reports the $p-$values for the *pits* tests. The figures in bold denote that the forecast density is correctly calibrated for a 95 percent confidence interval on the basis of that individual test; that is, when we cannot reject at a 5 percent significance level the null hypothesis that the densities are correctly calibrated. There are four rows to the table. The first two refer to the RW strategy, with real time data (RW-RT), and final-vintage data (RW-FV), respectively. The third and fourth rows give corresponding results for the EW strategy, with real-time data (EW-RT), and final-vintage data (EW-FV), respectively.

[4] They can be obtained from Norges Bank on request.

[5] Appendix 2 contains charts showing the probability that the output gap is less than 0 percent for each country, estimated using the RW strategy.

Looking at the real-time data results, we see that the RW strategy gives well-calibrated densities on the basis of all seven tests, row 1. But the EW strategy fails three of the seven tests in real time, row 3 in Table 1.

Turning to the revised final-vintage data, we see that for both the EW and RW strategies, the performance is somewhat weaker. The RW strategy passes four of the seven tests, and EW passes three.

Table 1: US Ensembles $p-$values for the *pits* tests

	LR2	LR_l	LR_u	LR3	AD	χ^2	LB
RW-RT	**0.13**	**0.39**	**0.27**	**0.24**	**0.27**	**0.11**	**0.21**
RW-FV	0.00	**0.14**	**0.22**	0.01	**0.06**	0.03	**0.26**
EW-RT	0.01	**0.52**	0.02	0.02	**0.08**	**0.20**	**0.11**
EW-FV	0.00	**0.08**	**0.09**	0.00	0.03	0.03	**0.06**

Notes: 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*; LR_u is the p-value for the LR test of zero mean and unit variance focusing on the 10 percent upper tail; LR_l is the p-value for the LR test of zero mean and unit variance focusing on the 10 percent lower tail; LR3 supplements LR2 with a test for zero first order autocorrelation; AD is the small-sample (simulated) p-value from the Anderson-Darling test for uniformity of the *pits* assuming independence of the *pits;* 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* based on autocorrelation coefficients up to four.

4.2.2 Australia, New Zealand and Norway

Tables 2-4 present the results for Australia, New Zealand and Norway, respectively. The Australian results, in Table 2, suggest that, like the US case, the RW strategy produces well-calibrated real-time densities (row 1). Although we note that the RW strategy does fail two of the seven tests. In contrast, the EW strategy indicates calibration failure for five of the seven tests (row 3). As with the US results, we see weaker calibration for both strategies with final-vintage data. For example, with the EW strategy (row 4), the null of correct calibration is not rejected on the basis of just one test with 95 percent confidence.

Turning to the New Zealand and Norway results, Tables 3 and 4 respectively, we see that RW and EW perform similarly in real time. The RW strategy results in only one rejection at a 95 percent confidence interval for Norway and no rejections for New Zealand (row 1 in each table). And the EW strategy betters that slightly with no tests failed (row 3 in each case).

As for the Australian and US data, both strategies fail more tests with final-vintage data. With the EW strategy, seven and two tests are failed for New Zealand and Norway, respectively; see row 4 of Tables 3 and 4. Similarly for the RW strategy, seven (New Zealand) and two (Norway) tests are failed, respectively; see row 2 of Tables 3 and 4.

Table 2: Australian Ensembles $p-$values for the *pits* tests

	LR2	LR_l	LR_u	LR3	AD	χ^2	LB
RW-RT	**0.44**	**0.49**	0.00	**0.53**	**0.08**	0.03	**0.89**
RW-FV	**0.11**	0.00	0.00	**0.14**	0.01	0.00	**0.19**
EW-RT	0.01	**0.46**	0.01	0.02	0.01	0.00	**0.94**
EW-FV	0.01	0.00	0.01	0.02	0.00	0.02	**0.21**

Notes: see notes to Table 1

Table 3: New Zealand Ensembles $p-$values for the *pits* tests

	LR2	LR_l	LR_u	LR3	AD	χ^2	LB
RW - RT	**0.10**	**0.58**	**0.70**	**0.12**	**0.06**	**0.23**	**0.44**
RW - FV	0.00	0.02	0.02	0.00	0.01	0.00	0.00
EW - RT	**0.09**	**0.41**	**0.94**	**0.09**	**0.08**	**0.29**	**0.52**
EW - FV	0.00	0.02	0.02	0.00	0.02	0.00	0.01

Notes: see notes to Table 1

Table 4: Norwegian Ensembles $p-$values for the *pits* tests

	LR2	LR_l	LR_u	LR3	AD	χ^2	LB
RW - RT	**0.25**	**0.57**	0.03	**0.36**	**0.22**	**0.61**	**0.37**
RW - FV	0.03	**0.11**	0.03	**0.05**	**0.32**	**0.68**	**0.27**
EW - RT	**0.16**	**0.14**	**0.27**	**0.17**	**0.22**	**0.61**	**0.13**
EW - FV	0.03	**0.05**	0.02	**0.05**	**0.30**	**0.83**	**0.30**

Notes: see notes to Table 1

4.2.3 Interpretation

Overall, there are two substantive findings. First, the recursive-weight strategy performs consistently across the four countries. Although for the relatively short New Zealand and Norwegian samples, there is little to separate the EW and RW strategies, with both strategies giving real-time forecast densities that are well calibrated. In contrast, for the longer US and Australian real-time samples, the EW strategy fails a number of *pits* tests. The RW strategy seems more robust on these longer real-time samples.

The second finding is that density forecasting performance is less satisfactory for the ensembles with final-vintage data. Data revisions contaminate the Phillips curve relationship in all four countries considered.

5 Conclusions

In this paper, we have examined the effectiveness of recursive-weight and equal-weight strategies for combining forecast densities using a Phillips curve relationship between inflation and the output gap. Using data for the US, Australia, New Zealand and Norway, we find that the recursive-weight strategy performs consistently well. In the two cases with longer samples of real-time data—the US and Australia—the equal-weight strategy results in forecast densities that exhibit calibration failure. This result reverses the perceived wisdom that simple averages are more reliable—a result found in a number of well-known studies of point forecasting accuracy.

Appendix 1: Output trend definitions

We summarise the seven detrending specifications below.

1. For the quadratic trend based measure of the output gap we use the residuals from a regression (estimated recursively) of output on a constant and a squared time trend.

2. For the HP trend, Hodrick and Prescott (1997), we set the smoothing parameter to be 1600 for our quarterly US data.[6] This two-sided filter relates the time-t value of the trend to future and past observations. Moving towards the end of a finite sample of data, it becomes progressively one-sided, and its properties deteriorate; see Mise, Kim and Newbold (2005).

3. To address the one-sided problem resulting from the HP trend, we use a forecast-augmented HP trend (again, with smoothing parameter 1600), with forecasts generated from an univariate AR(8) model in output growth (estimated recursively using the appropriate vintage of data). The implementation of forecast augmentation when constructing real-time output gap measures for the US is discussed at length in Garratt, Lee, Mise and Shields (2008).

4. Turning to the CF measure, Christiano and Fitzgerald (2003) propose an optimal finite-sample approximation to the band-pass filter, without explicit modeling of the data. Their approach implicitly assumes that the series is captured reasonably well by a random walk model and that, if there is drift present, this can be proxied by the average growth rate over the sample.

5. We also consider the band-pass filter suggested by Baxter and King (1999). We define the cyclical component to be fluctuations lasting no fewer than six, and no more than thirty two quarters—the business cycle frequencies indicated by Baxter and King (1999). Watson (2007) reviews band-pass filtering methods.

6. Turning to the BN trend, Beveridge and Nelson (1981), we note that this permanent trend and transitory cycle decomposition relies on a priori assumptions about the correlation between permanent and transitory innovations. The BN approach imposes the restriction that shocks to the transitory component and shocks to the stochastic permanent component have a unit correlation. We assume the ARIMA process for output growth is an AR(8), the same as that used in our forecast augmentation.

7. Finally, our UC model assumes q_t is decomposed into trend, cyclical and irregular components
$$q_t = \mu_t^7 + y_t^7 + \varepsilon_t, \;\; \varepsilon_t \sim NID(0, \sigma_\varepsilon^2), \; t = 1, \ldots, T \tag{A1}$$

[6] We could, of course, allow for uncertainty in the smoothing parameter. We reduce the computational burden in this application by fixing this parameter at 1600.

where the stochastic trend is specified as

$$\mu_t^7 = \mu_{t-1}^7 + \beta_{t-1} + \eta_t, \quad \eta_t \sim NID(0, \sigma_\eta^2) \tag{A2}$$

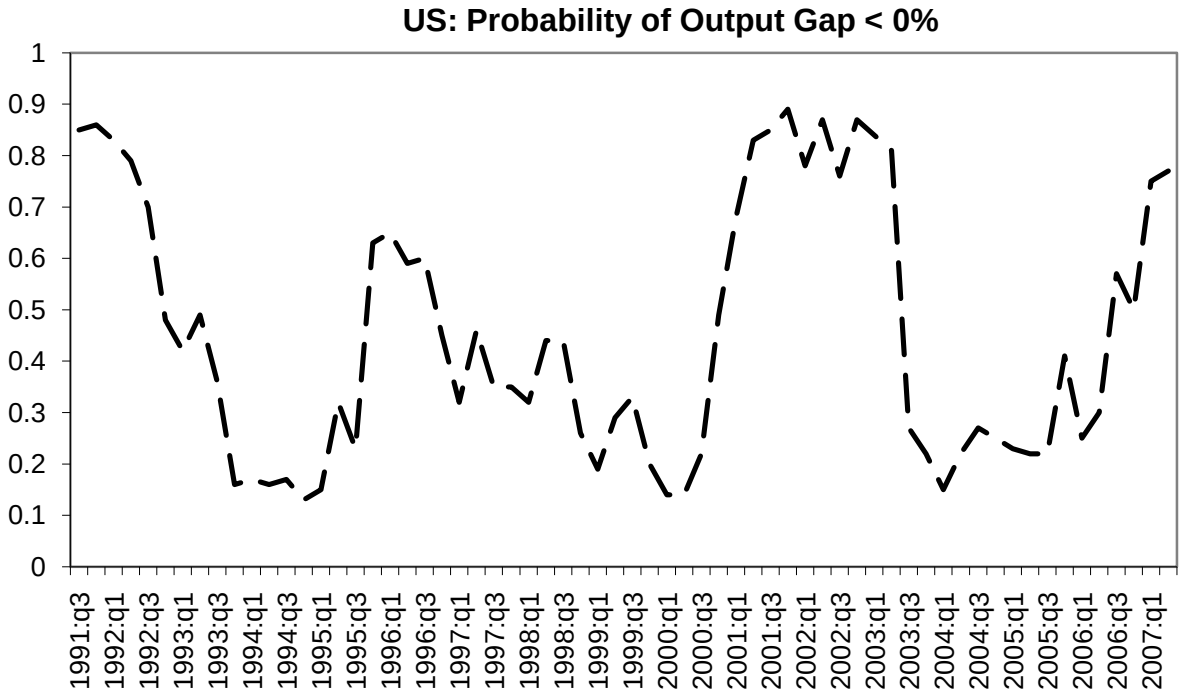
$$\beta_t = \beta_{t-1} + \zeta_t, \quad \zeta_t \sim NID(0, \sigma_\zeta^2). \tag{A3}$$

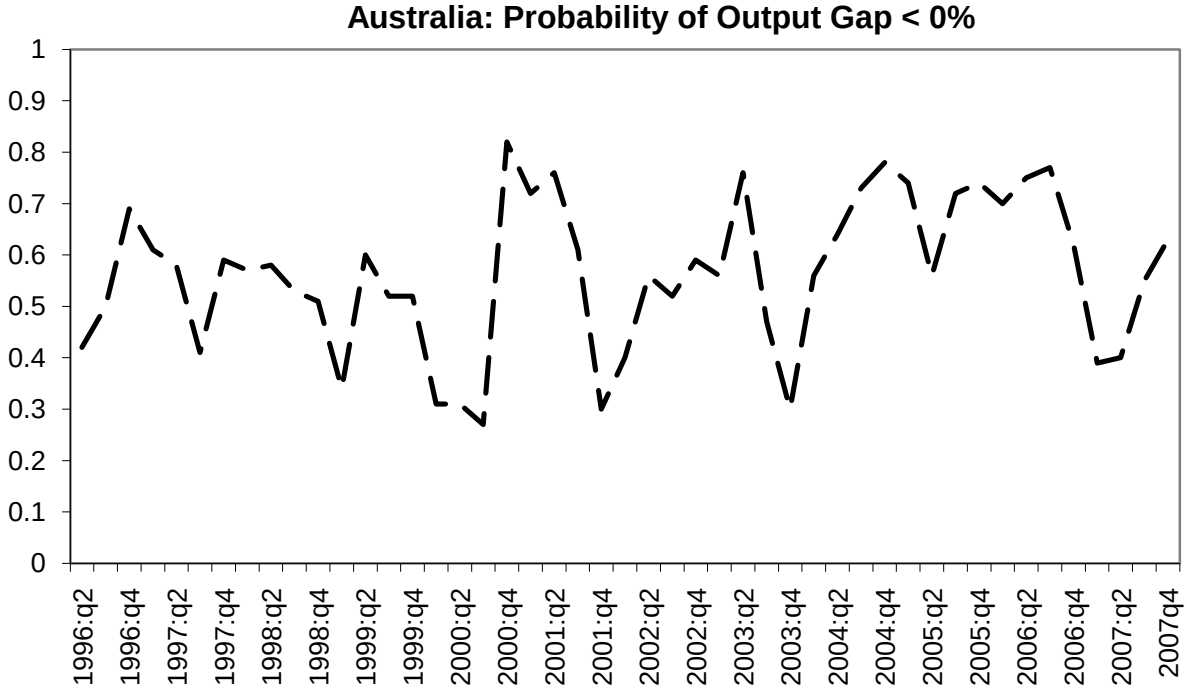
Letting $\sigma_\zeta^2 > 0$ but setting $\sigma_\eta^2 = 0$, gives an integrated random walk, which when estimated tends to be smooth. The cyclical component is assumed to follow a stochastic trigonometric process:

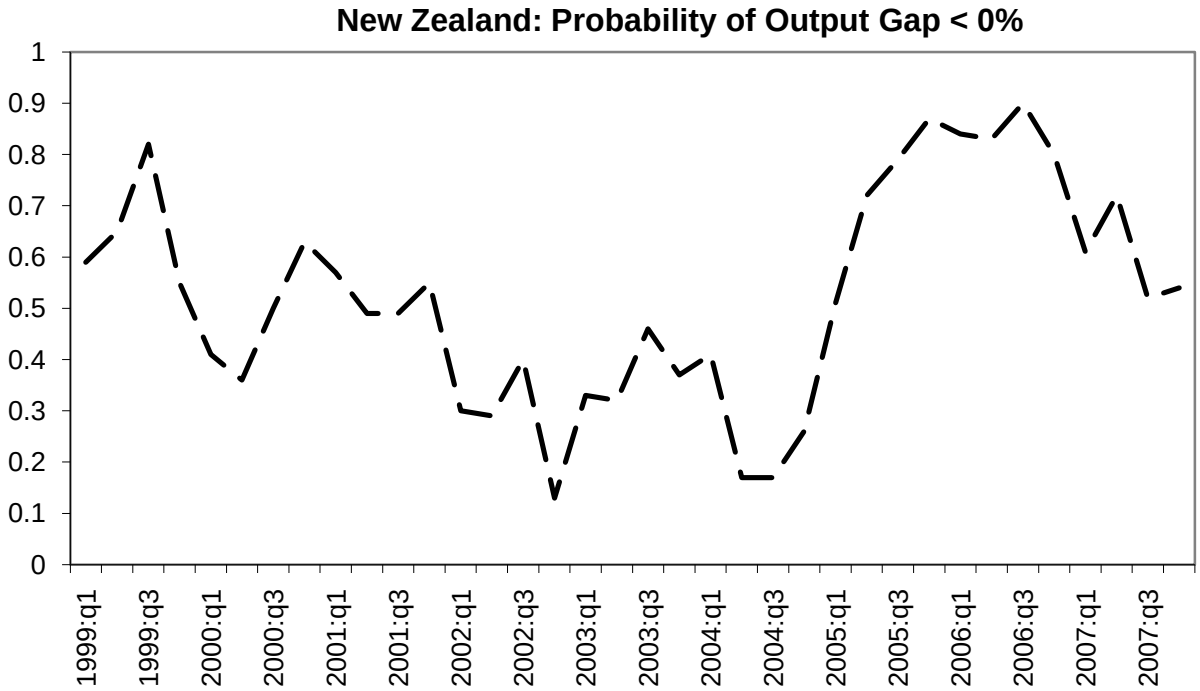
$$\begin{bmatrix} y_t^7 \\ y_t^{7*} \end{bmatrix} = \rho \begin{bmatrix} \cos\lambda & \sin\lambda \\ -\sin\lambda & \cos\lambda \end{bmatrix} \begin{bmatrix} y_{t-1}^7 \\ y_{t-1}^{7*} \end{bmatrix} + \begin{bmatrix} \kappa_t \\ \kappa_t^* \end{bmatrix} \tag{A4}$$

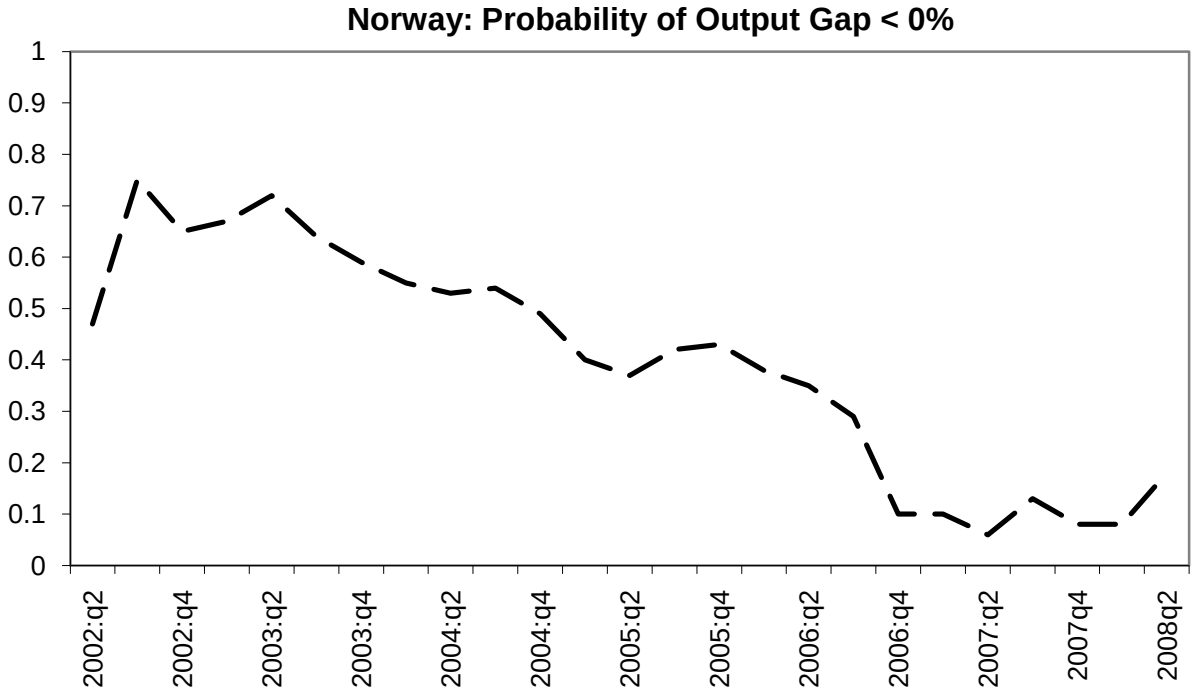
where λ is the frequency in radians, ρ is a damping factor and κ_t and κ_t^* are two independent white noise Gaussian disturbances with common variance σ_κ^2. We estimate this model by maximum likelihood, exploiting the Kalman filter, and estimates of the trend and cyclical components are obtained using the Kalman smoother.

Appendix 2


US: Probability of Output Gap < 0%
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1991:q3
1992:q1
1992:q3
1993:q1
1993:q3
1994:q1
1994:q3
1995:q1
1995:q3
1996:q1
1996:q3
1997:q1
1997:q3
1998:q1
1998:q3
1999:q1
1999:q3
2000:q1
2000:q3
2001:q1
2001:q3
2002:q1
2002:q3
2003:q1
2003:q3
2004:q1
2004:q3
2005:q1
2005:q3
2006:q1
2006:q3
2007:q1


Australia: Probability of Output Gap < 0%
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1996:q2
1996:q4
1997:q2
1997:q4
1998:q2
1998:q4
1999:q2
1999:q4
2000:q2
2000:q4
2001:q2
2001:q4
2002:q2
2002:q4
2003:q2
2003:q4
2004:q2
2004:q4
2005:q2
2005:q4
2006:q2
2006:q4
2007:q2
2007q4


New Zealand: Probability of Output Gap < 0%
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1999:q1
1999:q3
2000:q1
2000:q3
2001:q1
2001:q3
2002:q1
2002:q3
2003:q1
2003:q3
2004:q1
2004:q3
2005:q1
2005:q3
2006:q1
2006:q3
2007:q1
2007:q3


Norway: Probability of Output Gap < 0%
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
2002:q2
2002:q4
2003:q2
2003:q4
2004:q2
2004:q4
2005:q2
2005:q4
2006:q2
2006:q4
2007:q2
2007q4
2008q2

References

[1] Amisano, G. and R. Giacomini (2007), "Comparing Density Forecasts via Likelihood Ratio Tests", *Journal of Business and Economic Statistics*, 25, 2, 177-190.

[2] Bache, I.W., J. Mitchell, F. Ravazzolo and S.P. Vahey (2010), "Macro Modelling with Many Models", in D. Cobham, Ø. Eitrheim, S. Gerlach, and J. Qvigstad, eds, *Twenty Years of Inflation Targeting: Lessons Learned and Future Prospects.* Cambridge University Press, Cambridge, 398-418.

[3] Baxter, M. and R.G. King (1999), "Measuring Business Cycles: Approximate Band-Pass Filters for Economic Time Series", *Review of Economics and Statistics*, 81, 594-607.

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

[5] Beveridge, S. and C.R. Nelson (1981), "A New Approach to Decomposition of Time Series into Permanent and Transitory Components with Particular Attention to Measurements of the Buinsess Cycle", *Journal of Monetary Economics*, 7, 151-174.

[6] Bjornland, H., K. Gerdrup, A.S. Jore, C. Smith and L. A. Thorsrud (2010), "Weights and Pools for a Norwegian Density Combination", *North American Journal of Economics and Finance*, forthcoming.

[7] Christiano, L. and T.J. Fitzgerald (2003), "The Band Pass Filter", *International Economic Review*, 44, 2, 435-465.

[8] Clark, T.E. and M. W. McCracken (2010), "Averaging Forecasts from VARs with Uncertain Instabilities", *Journal of Applied Econometrics*, 25, 5-29.

[9] Cook, S. (2008), "Cross-data-vintage Encompassing", *Oxford Bulletin of Economics and Statistics*, 70, 849-865.

[10] Corradi, V., A. Fernandez and N.R. Swanson (2010), "Information in the Revision Process of Real-time Datasets", *Journal of Business and Economic Statistics*, 27, 455-467.

[11] Corradi, V. and N.R. Swanson (2006), "Predictive Density Evaluation", G. Elliot, C.W.J. Granger and A. Timmermann, eds, *Handbook of Economic Forecasting*, North-Holland, 197-284.

[12] Croushore, D. and T. Stark. (2001), "A Real-time Data Set for Macroeconomists", *Journal of Econometrics*, 105, 111-130.

[13] Dawid, A.P. (1984), "Statistical Theory: The Prequential Approach", *Journal of the Royal Statistical Society A*, 147, 278-290.

[14] Diebold, F.X., T.A. Gunther and A.S. Tay (1998), "Evaluating Density Forecasts; with applications to financial risk management", *International Economic Review*, 39, 863-83.

[15] Garratt, A, G. Koop, E. Mise, and S.P. Vahey (2009), "Real-Time Prediction with UK Monetary Aggregates in the Presence of Model Uncertainty", *Journal of Business and Economic Statistics,* 27, 480-491.

[16] Garratt, A., K. Lee, E. Mise and K. Shields (2008), "Real-Time Representations of the Output Gap", *Review of Economics and Statistics*, 90(4), 792-804.

[17] Garratt, A., J. Mitchell, and S.P. Vahey (2009), "Measuring Output Gap Uncertainty", *Birkbeck College Mimeo*, downloadable at http://www.ems.bbk.ac.uk/faculty/garratt.

[18] Genest, C. and J.V. Zidek (1986), "Combining Probability Distributions: A Critique and an Annotated Bibliography", *Statistical Science*, 1, 114-148.

[19] Geweke, J. (2009), "*Complete and Incomplete Econometric Models*", Princeton University Press.

[20] Groen, J.J.J., R. Paap and F. Ravazzolo (2009), "Real-time Inflation Forecasting in a Changing World", *Norges Bank Working Paper*, 2009-16.

[21] Gruen, D., T. Robinson and A. Stone (2002), "Output Gaps in Real Time: Are They Reliable Enough to Use for Monetary Policy?", *Reserve Bank of Australia Discussion Paper* 2002-06.

[22] Hall, S.G. and J. Mitchell (2007), "Combining Density Forecasts", *International Journal of Forecasting*, 23, 1-13.

[23] Hodrick, R. and E. Prescott (1997), "Post-War U.S. Business Cycles: An Empirical Investigation", *Journal of Money, Banking and Credit*, 29, 1-16.

[24] Jore, A. S., J. Mitchell and S.P. Vahey (2010), "Combining Forecast Densities from VARs with Uncertain Instabilities", *Journal of Applied Econometrics*, 25, 621-634.

[25] Kascha, C. and F. Ravazzolo (2010), "Combining Inflation Density Forecasts", *Journal of Forecasting*, 29, 231-250.

[26] Marcellino, M., J. Stock and M.W. Watson (2003), "A Comparison of Direct and Iterated AR Methods for Forecasting Macroeconomic Series h-steps Ahead", *Journal of Econometrics*, 135, 499–526.

[27] Mise, E., T-H. Kim and P. Newbold (2005), "On the Sub-Optimality of the Hodrick-Prescott Filter", *Journal of Macroeconomics*, 27, 1, 53-67.

[28] Mitchell, J. and K.F. Wallis (2010), "Evaluating Density Forecasts: Forecast Combinations, Model Mixtures, Calibration and Sharpness", *Journal of Applied Econometrics*, forthcoming.

[29] Nocetti, P., J. Smith and S. Hodges (2003), "An Evaluation of Tests of Distributional Forecasts" *Journal of Forecasting*, 22, 447-455.

[30] Orphanides, A. and S. van Norden (2005), "The Reliability of Inflation Forecasts Based on Output-Gap Estimates in Real Time", *Journal of Money Credit and Banking*, 37, 3, 583-601.

[31] Smith, J. and K.F. Wallis (2009), "A Simple Explanation of the Forecast Combination Puzzle", *Oxford Bulletin of Economics and Statistics*, 71, 331-355.

[32] 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.

[33] Stone, A. and S. Wardrop (2002), "Real-time National Accounts Data", *Reserve Bank of Australia Discussion Paper*, 2002-05.

[34] Timmermann, A. (2006), "Forecast Combination", in G. Elliot, C.W.J. Granger and A. Timmermann, eds, *Handbook of Economic Forecasting*, North-Holland, 197-284.

[35] 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.

[36] Wallis, K.F. (2005), "Combining Density and Interval Forecasts: a Modest Proposal", *Oxford Bulletin of Economics and Statistics*, 67, 983-994.

[37] Watson, M.W. (2007), "How Accurate are Real-time Estimates of Output Trends and Gaps?", *Federal Reserve Bank of Richmond Economic Quarterly*, Spring 2007.

[38] Zellner, A. (1971), *An Introduction to Bayesian Inference in Econometrics*, New York: John Wiley and Sons.