CENTRE FOR APPLIED MACROECONOMIC ANALYSIS

The Australian National University

CAMA Working Paper Series **June, 2011**

MEASURING OUTPUT GAP NOWCAST UNCERTAINTY

Anthony Garratt
Birkbeck College

James Mitchell
NIESR

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

CAMA Working Paper 16/2011
http://cama.anu.edu.au

Measuring Output Gap Nowcast Uncertainty*

Anthony Garratt (Birkbeck College)
James Mitchell (NIESR)
Shaun P. Vahey (CAMA, ANU)

March 29, 2011

Abstract

We propose a methodology to gauge the uncertainty in output gap nowcasts across a large number of commonly-deployed vector autoregressions in US inflation and various measures of the output gap. Our approach constructs ensemble nowcast densities using a linear opinion pool. This yields well-calibrated nowcasts for US inflation in real time from 1991q2 to 2010q1, in contrast to those from a univariate autoregressive benchmark. The ensemble nowcast densities for the output gap are considerably more complex than for a single VAR specification. They cannot be described adequately by the first two moments of the forecast densities. To illustrate the usefulness of our approach, we calculate the probability of a negative output gap at around 45 percent between 2004 and 2007. Despite the Greenspan policy regime, and some large point estimates of the output gap, there remained a substantial risk that output was below potential in real time. Our ensemble approach also facilitates probabilistic assessments of "alternative scenarios". A "dove" scenario (based on distinct output gap measurements) typically raises substantially the probability of a negative output gap (including 2004 through 2007) but has little impact in slumps, in our illustrative example.

Keywords: Output gap uncertainty; Ensemble Forecasting; Linear Opinion Pools; VAR models

JEL codes: C32; C53; E37

*Financial support from the ESRC (grant No. RES-062-23-1753), the ARC (grant No. LP0991098), Norges Bank, the Reserve Bank of New Zealand, and the Reserve Bank of Australia is gratefully acknowledged. Contact: Anthony Garratt, Birkbeck, University of London, Malet Street, Bloomsbury, London, WC1E 7HX, U.K. Tel: +44 (0) 207 631 6410. Fax: +44 (0) 631 6416. E-Mail: a.garratt@bbk.ac.uk. We benefitted greatly from comments by Knut Are Aastveit, Michael Andersson, Dean Croushore, Domenico Giannone, James Hamilton, Massimiliano Marcellino, Michael McCracken, Simon van Norden, David Papell, Mike Smith, Liz Wakerly, Mark Watson and Ken West. We are also indebted to conference and seminar participants at the Reserve Bank of Australia, Norges Bank, the University of Adelaide, ANU, De Nederlandsche Bank, the Reserve Bank of New Zealand, CIRANO, Eurostat, and the Bank of England. An earlier version of this paper was entitled "Measuring output gap uncertainty".

1 Introduction

Although the formulation of monetary policy in many central banks gives prominence to the output gap, conventional wisdom has it that there is considerable uncertainty about contemporaneous gap measurements. For example, Orphanides and van Norden (2002) document the large variation and the unreliability of real-time point estimates of the output gap for several commonly-used filters or detrending methods.

Unfortunately, the extant literature on output gaps does not permit a policymaker, or a researcher, to quantify the uncertainty across different detrending methods. This presents a considerable impediment to monetary policy in practice. Policymakers routinely combine the evidence across many specifications to perform probabilistic assessments of aggregate behavior; see, for example, Greenspan (2004), and the discussions by Feinstein, King, and Yellen (2004). In the absence of formal methods accounting for specification uncertainty, monetary policy based on probabilistic assessments must be based on informal judgements.

This paper proposes a methodology to gauge the uncertainty in output gap nowcasts across a large number of empirical specifications. Spurred by the analyses of output gaps and inflation by Orphanides and van Norden (2002, 2005), we use their US real-time data (supplied by the Federal Reserve Bank of Philadelphia; see Croushore and Stark, 2001) and a model space comprising many linear Gaussian vector autoregressive (VAR) models. We explore the model uncertainties in a baseline VAR by varying the "auxiliary assumptions"regarding the detrending method used to construct the output gap, the lag lengths in the VAR, and the timing of a single structural break. For simplicity, the detrending methods considered are a selection of commonly-used univariate filters. (In principle, our methodology could also be applied to multivariate detrending methods, and models with optimizing behavior.) Each VAR specification in our system produces one quarter ahead forecasts for inflation and a single output gap. Given the time lag in the release of real-time macro data, these forecasts indicate the current state of the macro economy. (Hereafter, we use the terms nowcast and one quarter ahead forecast, interchangeably.)

Whereas a conventional econometric analysis based on our VAR model space would report many nowcasts of inflation and the output gap, or perhaps use regression diagnostics to select a single forecast vector, our aim is to build forecast densities that reflect the uncertainty across the many specifications considered. To achieve this, we utilize a linear mixture of experts framework, also known as the "linear opinion pool" (LOP), to construct ensemble forecast densities for the variables of interest. For each VAR specification we measure the Kullback-Leibler "distance" between the real-time out of sample inflation forecast density and the "true" but unknown density using the logarithmic score. We use this information to construct the weights on the various specifications to build the ensemble forecast density vector. In this way, our output gap predictive densities reflect the ability of the many VAR specifications to predict inflation. The idea of using the Phillips curve to inform analysis of real-time output gaps is used (in larger systems) by Laubach and Williams (2003) and Weidner and Williams (2011).

A key insight of this paper is that probability forecasts for latent macro variables, accommodating a wide variety of model uncertainties, can be constructed by utilizing a form of "ensemble" forecasting. By this approach, the researcher considers a model space comprising a large number of forecasting models, generated by varying measurements and/or model features. Gneiting and Thorarinsdottir (2010) provide an example based on forecasting inflation with survey information and discuss the related literature in meteorology. In our analysis, we vary the auxiliary assumptions in our baseline VAR specification. Regardless of the technology utilized to produce the ensemble forecast densities (here LOP), the aim of ensemble forecasting is to approximate the unknown process with a large number of misspecified forecasting models. Although the focus is on model uncertainty and using misspecified models to construct predictive densities, concepts which are often associated with Bayesian analysis, the individual models can be estimated by either frequentist or Bayesian techniques. The linear and Gaussian VARs are estimated by frequentist methods in our empirical work. Nevertheless, the ensemble nowcast densities for each macro variable need not be symmetric and can accommodate multi-modality.

Turning to our results, we find that a baseline ensemble, based on seven commonly-used output gap measures (derived from univariate filters), and allowing for VAR model uncertainty (regarding lag length and the breakdate), produces well-calibrated (unimodal and symmetric) nowcast densities for US inflation. In contrast, a simple autoregressive benchmark (without the output gap) produces poorly calibrated densities, as does an integrated moving average model for inflation. Both of these commonly-utilized univariate time series models also have a weaker point forecasting performance (in terms of root mean squared forecast error) than our baseline ensemble. Our baseline ensemble typically produces a multi-modal nowcast density for the (latent) output gap. Hence, the nowcast density for the current output gap cannot be described adequately by a point (conditional mean) estimate, together with a conventional measure of dispersion, such as the forecast variance.

To illustrate the usefulness of our output gap nowcasts to monetary policymakers, we calculate the one step ahead probability that output is below trend from our baseline ensemble. We find that the policymaker is almost never certain of the sign of the output gap throughout our evaluation. For example, we calculate the probability of a negative output gap at around 45 percent between 2004 and 2007. Therefore despite the Greenspan policy regime, and some large positive point estimates of the output gap, there remained a substantial risk that real output was below potential in real time.

Since monetary policy is often concerned with the probabilistic impacts of "alternative scenarios", we repeat our analysis assuming that potential output can be captured by a linear time trend (allowing for breaks in the trend). We label our alternative scenario a "dove" to reflect that the output gap measures often imply that output is less than potential. (The linear trends approach fails to match the productivity slowdown apparent in revised US data.) We consider an ensemble of the dove scenario ensemble and our original baseline ensemble. This "Grand Ensemble" produces well-calibrated nowcasts for inflation. But consideration of our alternative scenario dramatically changes the assessment of the output gap. The inclusion of our dove scenario raises the probability of a negative

output gap considerably over the 2004-2007 period in our illustrative example. But it has very modest impacts on the probability of interest during the recent slump—when all our output gap measurements assess real output to be less than potential.

The remainder of this paper is structured as follows. In section 2, we describe our model space, based on Orphanides and van Norden (2002, 2005). In section 3, we present and discuss our methodology for ensemble forecasting using the LOP. In section 4, we report our results for forecasting inflation and describe our predictive densities for the latent output gap. We draw some conclusions in the final section.

2 Model space

A bivariate VAR model space is common to many analyses of inflation determination. For example, Rudebusch and Svensson (2002) and Laubach and Williams (2003) start with bivariate VARs, and add additional explanatory relationships and restrictions.[1]

To illustrate our ideas, we begin with a VAR model space for inflation, π_t, and the output gap, y_t, of the type explored by Orphanides and van Norden (2002, 2005):

$$\begin{bmatrix} \pi_t \\ y_t^j \end{bmatrix} = \mathbf{A}^j \begin{bmatrix} \pi_{t-1} \\ y_{t-1}^j \end{bmatrix} + \boldsymbol{\epsilon}_t^j, \qquad t = 1, \ldots, T, \quad \boldsymbol{\epsilon}_t^j \sim i.i.d.\ N(\mathbf{0}, \boldsymbol{\Sigma}^j). \tag{1}$$

That is, we consider a baseline VAR specification in which the measure of interest for the output gap has been varied to give J linear and Gaussian VAR models, indexed $j = 1, \ldots, J$. For expositional ease, we ignore the intercept and restrict the lag order of the J VARs to one. The conformable coefficient matrix is denoted $\mathbf{A}^j$.

Our baseline VAR setup uses seven output gap measures, $J = 7$, derived from univariate filters and deployed by Orphanides and van Norden (2002, 2005).[2] We define the output gap as the difference between observed output and unobserved potential (or trend) output. Let q_t denote the (logarithm of) actual output in period t, and μ_t^j be its trend using definition j, where $j = 1, \ldots, J$. Then the output gap, y_t^j, is defined as the difference between actual output and its j^{th} trend measure. We assume the following trend-cycle decomposition:

$$q_t = \mu_t^j + y_t^j. \tag{2}$$

The seven methods of univariate trend extraction in our baseline VAR are: quadratic, Hodrick-Prescott (HP), forecast-augmented HP, Christiano and Fitzgerald, Baxter-King, Beveridge-Nelson, and Unobserved Components. Since we are interested in real-time prediction, we estimate the trends at every vintage date. That is, estimation is recursive for all VAR specifications; each recursion uses a different vintage of data as well as an

[1] Although additional equations and identifying conditions pose no conceptual problems for our methodology, the computational burden would increase. We prefer to restrict our attention to a two-equation model space which, as Sims (2008) notes, lies at the heart of many explanations of inflation determination.

[2] Marcellino and Musso (2009) provide a recent analysis of univariate and multivariate real-time output gap measures using Euro-area data.

additional observation. We summarize the seven well-known univariate filters in Appendix 1.

We vary two other assumptions in our baseline VAR specification. First, following (among others) Stock and Watson (1999), we allow for shifts in relationships between inflation and output gaps during the transition from (what is often referred to as) the US Great Inflation to the Great Moderation. We consider every feasible single break date, assuming a single coincident break in the conditional mean and variance for both variables. This raises the potential number of models considerably since, in effect, each candidate break date defines a new VAR.

The second auxiliary assumption we vary is the lag length in the VAR (which for ease of exposition we fixed at one in equation (1)). If we have J output gap measures, and for any given y_t^j, we have K different variants defined over different values of the maximum lag length and the location of the break date, then in total we have $N = J \times K$ models, and N associated forecasts of inflation and the output gap from the entire system.

We emphasize that even though univariate filters have received considerable attention in the literature, there is no particular reason to restrict policymakers to this approach. Since policymakers often disagree strongly over their assessments of the output gap, we repeat our ensemble analysis with an alternative scenario. We construct our alternative measures of the output gap using a linear time trend (allowing for breaks in the trend) to approximate potential output. Many applications of "Taylor rules" use similar time-trend based measures; see Taylor (1993). Orphanides and van Norden (2002) discuss the real-time properties of linear time trend based measures of the output gap, noting that typically output is less than potential by this approach in recent years. The linear time trend is not sufficiently flexible to capture the productivity downturn (even when, following common practice, trend breaks are admitted in the 1970s and 1980s). We label our alternative scenario a "dove" to reflect this feature.

More precisely, we explore alternative specifications with the output gap defined as the estimated residual, $\widehat{\upsilon}_t$, from the ordinary least squares regression of:

$$q_t \;=\; a + b_1 t \;+\; b_2 D_{73,t} t \;+\; b_3 D_{84,t} t \;+\; \upsilon_t, \qquad t = 1, \ldots, T, \qquad \upsilon_t \sim i.i.d.\ N(0, \sigma^2). \quad (3)$$

A linear time trend without breaks is defined by $D_{73,t} = D_{84,t} = 0$. We also consider two further specifications with one breakdate, and two breakdates. The breaks are 1973q4 (following Orphanides and van Norden, 2002), and 1973q4 and 1984q1, respectively. Like the baseline case, we vary the lag length and the break date in the dove VAR ensemble, for each of the three types of linear detrending considered.

3 Ensemble forecasting

Armed with our many VAR specifications, it is straightforward to produce forecast densities for both variables of interest. Given non-informative priors, the predictive densities for both inflation and the output gap, for a given VAR specification, (1), are multivariate Student-t; see Zellner (1971), pp. 233-236 and, for a more recent application, Garratt et

al. (2009). Appendix 2 provides details of how the predictive densities are constructed from our VARs.

A key insight of this paper is that by interpreting the N VAR specifications as a "perturbed" ensemble model space, probabilistic forecasts can be constructed using the entire set of specifications. In ensemble forecasting applications, researchers consider perturbations to a single basic misspecified model.[3] The perturbations might be to the measurements and/or the model space; see (among others) Raftery et al. (2005), Doblas-Reyes *et al.* (2009) Bao *et al.* (2010) and Gneiting and Thorarinsdottir (2010). A common feature of the ensemble methodology is the aim to approximate a non-linear and/or non-Gaussian process with a combination of specifications based on an incomplete model space.

We utilize a linear opinion pool (LOP) to construct the ensemble forecast densities. 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 a policymaker. As emphasised by Wallis (2005), the approach is particularly useful for the combination of survey information since the only information required is the out of sample forecast provided by each expert (VAR); see also Hall and Mitchell (2007).[4] Sometimes the application of opinion pools in economic forecasting is referred to as "density combination". Jore *et al.*(2010) utilize linear opinion pools with VAR forecasts but do not consider the issues of measuring, or forecasting, the output gap. Geweke (2010) discusses the differences between opinion pools and mixture models, providing specific empirical examples with small numbers of models.

Our methodology is motivated by the situation faced by policymakers in central banks. The policymakers concerned discuss the forecast densities provided by experts on their staff. Typically, these many experts utilize linear (or linearized) and Gaussian models which the policymaker believes to be false.[5]

With monetary policy practice in mind, we approximate the unknown forecast densities for inflation and the output gap by using (time-varying) aggregates of the individual VAR forecasts, with each VAR being both linear and Gaussian. We assume that a policymaker would believe that the various output gap measures differ from the "true" output gap by more than conventional (Gaussian) white noise measurement error and that the policymaker believes also that the true output gap is never observed, even ex post.[6] In this paper, we construct the policy maker's implied forecast densities for the output gap using the out of sample density forecasting performance for inflation.

More formally, we consider a monetary policymaker seeking to aggregate forecasts from the VARs in the model space. Given $i = 1, \ldots, N$ VAR specifications, the ensemble

[3]Bache et al (2010) note that the technologies for ensemble density construction differ across applied statistics fields.

[4]Kascha and Ravazzolo (2010) contrast the properties of linear and logarithmic opinion pools.

[5]Currently, the aggregation of these expert opinions is conducted informally in monetary policy institutions.

[6]Since the forecast density refers to an unobserved variable, y, it would be inconsistent for the policymaker to use any single imperfectly measured output gap, $y^{j\prime}$, as the true outturn to gauge the fit of the density.

density for inflation is defined by the LOP:

$$p(\pi_\tau) = \sum_{i=1}^{N} w_{i,\tau}\ g(\pi_\tau \mid I_{i,\tau}), \qquad \tau = \underline{\tau}, \ldots, \overline{\tau}, \tag{4}$$

where $g(\pi_\tau \mid I_{i,\tau})$ are the one step ahead forecast densities from model i, $i = 1, \ldots, N$ for inflation π_τ, conditional on the information set $I_{i,\tau}$. 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. The non-negative weights, $w_{i,\tau}$, in this finite mixture sum to unity.[7] Furthermore, the weights may change with each recursion in the evaluation period $\tau = \underline{\tau}, \ldots, \overline{\tau}$.

We then construct the ensemble nowcast density for the output gap using:

$$q(y_\tau) = \sum_{i=1}^{N} w_{i,\tau}\ h(y_\tau^j \mid I_{i,\tau}), \qquad \tau = \underline{\tau}, \ldots, \overline{\tau}, \tag{5}$$

where $h(y_\tau^j \mid I_{i,\tau})$ is a forecast density from model i for the output gap, and the weights are exactly the same as in the previous equation—that is, we use the inflation densities and outturns to derive the weights for the output gap ensemble.

We utilize weights based on the fit of the individual component forecast densities. Following Amisano and Giacomini (2007), Hall and Mitchell (2007) and Jore *et al.*(2010), the logarithmic score measures density fit for each component 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. The logarithmic score of the i^{th} density forecast, $\ln g(\pi'_\tau \mid I_{i,\tau})$, is the logarithm of the probability density function $g(. \mid I_{i,\tau})$, evaluated at the outturn, π'_τ. Specifically, the recursive weights for the one step ahead densities take the form:

$$w_{i,\tau} = \frac{\exp\left[\sum_{\underline{\tau}-\kappa}^{\tau-1} \ln g(\pi'_\tau \mid I_{i,\tau})\right]}{\sum_{i=1}^{N} \exp\left[\sum_{\underline{\tau}-\kappa}^{\tau-1} \ln g(\pi'_\tau \mid I_{i,\tau})\right]}, \qquad \tau = \underline{\tau}, \ldots, \overline{\tau} \tag{6}$$

where $\underline{\tau} - \kappa$ to $\underline{\tau} - 1$ comprises the training period used to initialize the weights. It is important to note that the weight on the various specifications varies through time. Hence, the ensemble exhibits greater flexibility than any single linear VAR specification (in which the individual model parameters are recursively updated). In this sense the ensemble approach approximates a non-linear data generating process. We also reiterate that the ensemble forecast densities can accommodate non-Gaussian behavior. For example, the predictive densities for the output gap are not restricted to be symmetric. This feature reflects the properties of the LOP approach; see the discussion in Timmermann (2006). In this sense, our approach is applicable even if the policymaker believes that the true model is non-Gaussian and non-linear.

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

4 Nowcast densities for inflation and the output gap in the US

We turn now to the details of our empirical exercise. We begin this section by describing the sample data and the specifics of our VAR model space. We present the density forecasting results for US inflation using our baseline ensemble and contrast the performance with two conventional forecasting models: a second-order autoregressive benchmark (AR2) and an integrated moving average (IMA) specification for inflation. We repeat our analysis with our alternative (dove) scenario (using linear time trend measures of potential output but allowing for breaks in the trend). Then, we consider an ensemble of the two ensembles, which we term a "Grand Ensemble" (GE), and analyze the behavior of the forecast densities for inflation and the output gap. To illustrate the implications of our analysis for policymaking, we plot and discuss the time series of probabilities the event that output is below trend in the current period—an event of interest to central bankers.

4.1 Data and model space details

Our US sample spans the Great Inflation, the Great Moderation and the Great Recession. We use the same real-time US data set as Clark and McCracken (2010), extended to include more recent data. The quarterly real-time data comprise real GDP and the GDP price deflator, with 180 vintages (data observed at a specific point in time, known as the vintage date), starting in 1965q4, and ending in 2010q3. The data for each vintage, avoiding the Korean War, are for $1954q3, \ldots, \tau - 1$. Data for output and the price deflator are 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. We define inflation as the first difference in the logarithm of the price deflator, multiplied by 100.

Our out of sample evaluation period is: $\tau = \underline{\tau}, \ldots, \overline{\tau}$ where $\underline{\tau} = 1991q2$ and $\overline{\tau} = 2010q1$ (76 observations). In order to implement ensemble methodology through the evaluation period requires an additional assumption about which measurement is to be forecast. Following Clark and McCracken (2010) 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 evaluations; see the discussion in Corradi *et al.*(2009).

For each VAR, we allow for a single structural break of unknown timing. In order to reduce the computational burden, the break date is restricted to occur before the start of the evaluation period, $\underline{\tau}$, with at least 15 percent of the sample used for post-break in-sample estimation of each VAR. The break occurs in the conditional mean and the variance for both equations in the bivariate VARs.

In every VAR considered, we vary the maximum lag length between one and four. In

total, accounting for our variations in lag length and break dates we consider $K = 368$ VAR models for each measure of the output gap. With seven measures of the output gap based on univariate filters, the predictive densities for the baseline ensemble utilize $N = 2576$ specifications for each observation in the evaluation period. For the dove case, with the linear time trend model space, there are 1104 models.

Recall that we construct the weights based on the fit of the individual VAR forecast densities for inflation. We use the logarithmic score to measure density fit for inflation through the evaluation period. Given the relatively large number of quarterly observations available in our data set, we set $\kappa = 20$, allowing a training period of five years. Given the one quarter lag in the release of real-time data measurements, the forecast density for inflation and the output gap are density nowcasts.

4.2 Inflation nowcasts

We begin our results with an assessment of the calibration of the ensemble predictive densities for inflation. 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. Following Rosenblatt (1952), Dawid (1984) and Diebold *et al.*(1998), evaluation can use the probability integral transforms (*pits*) of the realization of the variable with respect to the forecast densities. We gauge calibration by examining whether the *pits* z_τ, where:

$$z_\tau = \int_{-\infty}^{\pi'_\tau} p(u)du, \tag{7}$$

are uniform and, for our one step ahead forecasts, independently and identically distributed; see Diebold *et al.* (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.[8]

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_τ follows an AR(1) process. In addition, we 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 *et al*, 2003). Finally, following Wallis (2003), we employ a Pearson chi-squared test (χ^2) which divides the range of the z_τ into eight equiprobable classes and tests for uniformity in the histogram. Turning to the test for independence of the *pits*, we use a Ljung-Box (LB) test, based on (up to) fourth-order autocorrelation.

We also investigate relative predictive accuracy by considering a Kullback-Leibler information criterion (KLIC)-based test, utilizing the expected difference in the log scores of candidate densities; see Bao *et al.* (2007), Mitchell and Hall (2005) and Amisano and Giacomini (2007). Suppose there are two forecast densities, $p(\pi_\tau \mid I_{1,\tau})$ and $p(\pi_\tau \mid I_{2,\tau})$, so

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

that the KLIC differential between them is the expected difference in their log scores: $d_\tau = \ln p(\pi'_\tau \mid I_{1,\tau}) - \ln p(\pi'_\tau \mid I_{2,\tau})$. The null hypothesis of equal forecast performance is $\mathcal{H}_0 : E(d_\tau) = 0$. A test can then be constructed since the mean of d_τ over the evaluation period, $\overline{d}_\tau$, under appropriate assumptions, has the limiting distribution: $\sqrt{T}\overline{d}_\tau \rightarrow N(0, \Omega)$, where Ω is a consistent estimator of the asymptotic variance of d_τ.[9] Mitchell and Wallis (2011) discuss the value of information-based methods for evaluating forecast densities that look well-calibrated from the perspective of the *pits*.

Examining the goodness of fit and independence *pits* tests presented in the first row of the top panel of Table 1, we see that the real-time inflation ensemble forecast densities from the baseline VAR specifications are well calibrated at a 90 percent confidence level. (Instances of appropriate calibration are marked in boldface.[10]) The two rows in the bottom panel of Table 1 show the *pits*-based tests for our two univariate time series models: the AR2 and the IMA. The AR2 fails the LR and LB tests at a 90 percent confidence level; and the IMA fails the LR.[11] The final column of Table 1 shows the average log score over the evaluation period. The KLIC-based tests for the baseline VAR ensemble relative to the AR2 (IMA) give a p-value of 0.00 (0.02). The baseline ensemble outperforms the univariate benchmarks.[12]

The densities constructed using our dove scenario (based on linear time trend measures of the output gap) are shown in the second row of the top panel of Table 1. This ensemble does not produce well-calibrated inflation forecast densities, failing both the LR and AD tests at a 90 percent confidence level. The dove ensemble betters the univariate time series models, but not by much: the KLIC-based tests using the log scores (last column) reveal an insignificant improvement relative to the AR2 (IMA) with a p-value of 0.22 (0.82).[13]

Even though the inflation nowcast densities from the dove ensemble are not particularly well calibrated, the policymaker might wish to make use of both the dove and baseline specifications. Prominent members of the Federal Open Market Committee could insist, for example, on consideration of an alternative scenario. Hence, we examine the Grand Ensemble (GE) of the two, presented in the third row of the top panel of Table 1.[14] We

[9] When evaluating the forecast densities we abstract from the method used to produce them. Amisano and Giacomini (2007) and Giacomini and White (2006) discuss more generally the limiting distribution of related test statistics.

[10] To control the joint size of the four evaluation tests applied would require the use of a stricter p-value. For example, the Bonferroni correction indicates a p-value threshold, for a 90 percent confidence level, of $(100 - 90)/4 = 2.5$ percent, rather than 10 percent.

[11] We also examined (but do not report) AR benchmarks with lag order 1,3 and 4, with no qualitative differences in the results. The IMA model for inflation takes the form: $\Delta\pi_t = \alpha + \varepsilon_t + \theta\varepsilon_{t-1}$.

[12] A companion paper, Garratt et al (2011), describes the forecasting performance for inflation of this VAR ensemble using US, Australian, Norwegian and New Zealand data. In all cases, the ensemble outperforms simple autoregressive benchmarks for density forecasting. And, for the longer US and Australia samples, the ensembles outperform simple equal-weighted density averages.

[13] The KLIC-based test of forecast density performance of the baseline ensemble relative to the linear time trend ensemble has a p-value of 0.14.

[14] The weights on each ensemble were recursively computed using equation (6). An alternative approach, described by Hall and Mitchell (2007) and Geweke (2010), selects the weights by optimization. We found this alternative approach yields weights, *pits*-based tests and average log scores that were qualitatively

observe well-calibrated inflation forecast densities from the GE at a 90 percent confidence level on the basis of each individual *pits* test. Furthermore, the log score is somewhat lower than for the baseline ensemble.[15] The similarity between the performance of the baseline and GE inflation density forecasts reflects the high degree of linear dependence between them. Both yield approximately symmetric and unimodal densities throughout the evaluation; and both appear to be well-calibrated.[16]

These results, which show that the univariate time series benchmarks can be bettered in forecasting performance, are noteworthy given the well-known result in the macro forecasting literature that parsimonious autoregressive specifications are "hard to beat"; e.g., see Stock and Watson (2007). This view is based on measures of relative point forecasting performance in general, and root mean square forecast error (RMSFE) in particular. As Jore *et al.*(2010) explain, these benchmarks are not particularly tough to beat at density forecasting. It is interesting to note, however, that the point forecast from the baseline ensemble (the conditional mean of the densities) has a RMSFE ratio of 0.9106 (0.9606), relative to the AR2 (IMA) benchmark. The corresponding ratio for the GE ensemble is 0.9155 (0.9658). Our VAR ensembles perform quite well in terms of point forecasting.

Table 1: Forecast Density Evaluation for Inflation, 1991q2-2010q1

	LR	AD	χ^2	LB	Log Score
Baseline Ensemble	**0.16**	**0.54**	**0.68**	**0.39**	-1.391
Dove Ensemble	0.04	0.03	**0.31**	**0.36**	-1.434
Grand Ensemble	**0.27**	**0.61**	**0.80**	**0.52**	-1.375
AR2	0.04	**0.13**	**0.16**	0.04	-1.473
IMA	0.08	**0.22**	**0.11**	**0.18**	-1.441

Notes: LR is the p-value for the Likelihood Ratio test of zero mean, unit variance and zero first order autocorrelation of the inverse normal cumulative distribution function transformed *pits*, with a maintained assumption of normality for the transformed *pits*; 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. Log Score is the average logarithmic score over the evaluation period. The Grand Ensemble statistics are computed over a shorter evaluation period, reflecting the need for an extra training period (here set to 9 quarters).

similar.

[15] The KLIC-based log score tests of the GE relative to the AR2 (IMA) benchmark have a p-value of 0.00 (0.02).

[16] We experimented with modelling the dependence using a Gaussian copula. The copula opinion pool yielded very similar calibration properties and a similar average log score to the GE.

4.3 Output gap nowcasts

Since our approach is based on the idea that none of the output gap measures correspond to the true output gap, and density evaluation requires an "outturn", we do not report *pits*-based tests for the output gap. Instead, in Figure 1, we focus on characterizing the forecast densities from the GE output gap.[17] The dates to which the nowcast refers are provided along the x-axis in Figure 1, and the size of the output gap is along the y-axis. The nowcasts for the GE are from 1993q3 to 2010q1 (reflecting the additional 9 observations used as a training period to initialize the weights). The shades of the forecast densities indicate probability mass, with highest mass represented by white, and lowest mass represented by black. For many observations "twin peaks" (or more) are discernible, often with one peak close to zero (from the baseline case) and another almost always substantially below zero (from the dove scenario). It is also noteworthy that the business cycle is slightly asymmetric in so far as the upswings of the economy are of long duration but contractionary spells are relatively short; see the discussion of the asymmetric business cycle in Morley and Piger (2011).

Reporting a central measure of the output gap, even augmented with a single measure of dispersion (such as the forecast error variance), would not give an accurate representation of the nowcast uncertainty in the presence of this multi-modality. Perhaps with this in mind, central banks often focus on the probability of particular events of interest, such as the sign of the output gap. In Figure 2, we plot the probability of a negative output gap $Pr(y_t < 0)$, using both our baseline ensemble (solid line), and the GE (broken line); the difference between the two reflects the impact of the dove alternative scenario.

We make two observations about the time series of probabilities displayed in Figure 2 and the impact of the alternative scenario. First, the dove scenario tends to raise the probability of a negative output gap. For example, for the 1993 to 1999 period, the probability of a negative output gap generally fluctuates in the range of 30 to 70 percent for the baseline ensemble. But for the GE the probability varies between 55 and 90 percent, and always exceeds the baseline probability. Similarly, the probability of a negative output gap fluctuates around 45 percent between 2004 and 2007 for the baseline case. But the GE has the risk fluctuating between approximately 60 and 80 percent over the same period. Although the baseline ensemble implies that through this sub-period of the Greenspan monetary policy regime there was a substantial risk that real output was below potential, the risk is assessed to be much greater with the addition of the dove scenario. Second, during the periods in which the ex post data suggest that the US economy contracted during 2000 and 2001, the probability of a negative output gap rises very sharply, regardless of whether we look at the baseline ensemble or the GE. The two ensembles are again very close from 2007 onwards, with the probability of output being below potential rising progressively until late 2009. Overall we conclude that, in general, consideration of the alternative scenario (in the GE) raises the probability of a negative output gap considerably. But it has very modest impacts on the probability of interest

[17] Recall that the GE nowcasts for the output gap are based on equation (5) with the weights given by equation (6), and that the GE inflation nowcasts based on equation (4) do not exhibit calibration failure.

during a slump—when there is less disagreement between our various specifications about the output gap.

Although, as we have noted, point estimates of the output gap can be misleading with non-Gaussian predictive densities, it is interesting to compare the 5^{th} and 95^{th} percentiles from our baseline ensemble with the (ex post) Congressional Budget Office (CBO) point estimates, and the Laubach-Williams (LW) point estimates reported by Weidner and Williams (2011).[18] Figure 3 shows that the (revised) CBO measure often lies within the confidence bounds of the baseline ensemble, particularly from 2001 to 2006.[19] However, the CBO estimates show a very marked drop with the recent slump from 2008, giving a number of observations below the 5^{th} percentile. The CBO estimates are also outside our confidence bands during the late 1990s and suggest a stronger boom. In contrast to both the CBO approach and our baseline ensemble, the LW estimates indicate a considerable boom between 2003 and 2007, followed by a rapid decline during 2008, and then sit comfortably within the confidence bounds until 2010. It is worth repeating that the ensemble forecast densities for the latent output gap are "nowcasts": one step ahead forecasts from macro data arriving with a one-period lag. That is, they are more timely than the CBO and LW measure shown; and our approach is based on real-time data.

5 Conclusions

We propose a methodology to gauge the uncertainty in output gap nowcsats across a large number of commonly-deployed VARs (vector autoregessions) in US inflation and output gap measures. We construct ensemble nowcast densities for the output gap and inflation with a linear opinion pool. Our approach yields well-calibrated forecast densities for inflation in real time, in contrast to those from simple univariate models which ignore the output gap. The ensemble forecast densities for the output gap indicate considerable uncertainty and admit multi-modality, allowing the researcher to distinguish between the impacts of baseline and alternative scenarios on nowcast probabilities. Our "dove" scenario (which uses alternative output gap measurements) on average raises the probability of a negative output gap during our evaluation period. But the alternative scenario has little impact in slumps, when there is more agreement about the sign of the output gap nowcast across our VAR specifications.

[18] We are grateful to Justin Weidner and John Williams for supplying the point estimates used in this plot. Details of the CBO approach can be found CBO (2001).

[19] The GE, which allows for the dove scenario, gives more diffuse predictive densities for the output gap, with much higher probability mass below zero; see Figure 1.

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 *Twenty Years of Inflation Targeting: Lessons Learned and Future Prospects* (eds. D. Cobham, Ø. Eitrheim, S. Gerlach and J. Qvigstad), Cambridge University Press, Cambridge, pp. 398-418.

[3] Bao, L., T. Gneiting, E.P. Grimit, P. Guttop, and A.E. Raftery (2010), "Bias Correction and Bayesian Model Averaging for Ensemble Forecasts of Surface Wind Direction", *Monthly Weather Review*, 138, 1811-1821.

[4] Bao, Y., T-H. Lee and B. Saltoglu (2007), "Comparing Density Forecast Models", *Journal of Forecasting*, 26, 203-225.

[5] 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.

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

[7] 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.

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

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

[10] Congressional Budget Office (2001), "CBO's Method for Estimating Potential Output: An Update", August.

[11] Corradi, V., A. Fernandez and N.R. Swanson (2009), "Information in the Revision Process of Real-Time Data Sets", *Journal of Business and Economic Statistics*, 27, 455-467.

[12] 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.

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

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

[15] 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.

[16] Doblas-Reyes, F.J., A. Weisheimer, M. Déqué, N. Keenlyside, M. McVean, J. M. Murphy, P. Rogel, D. Smith, and T. N. Palmer (2009) "Addressing Model Uncertainty in Seasonal and Annual Dynamical Ensemble Forecasts", *Quarterly Journal of the Royal Meteorological Society*, 135, 1538-1559.

[17] Feinstein, M., M.A. King, and J. Yellen (2004) "Innovations and Issues in Monetary Policy: Panel Discussion", *American Economic Review*, Papers and Proceedings, May, 41-48.

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

[19] 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.

[20] Garratt, A., J. Mitchell, S.P. Vahey and E. Wakerly (2011), "Real-time Inflation Forecast Densities from Ensemble Phillips Curves", *North American Journal of Economics and Finance*, 22, 77-87.

[21] Genest, C. and J. Zidek (1986), "Combining Probability Distributions: a Critique and an Annotated Bibliography", *Statistical Science*, 1, 114-135.

[22] Geweke, J. (2010), *Complete and Incomplete Econometric Models*, Princeton University Press.

[23] Giacomini, R. and H. White (2006), "Tests of Conditional Predictive Ability", *Econometrica*, 74, 1545-1578.

[24] Gneiting, T., and T. Thorarinsdottir (2010) "Predicting inflation: Professional Experts versus No-change Forecasts", http://arxiv.org/abs/1010.2318.

[25] Greenspan, A. (2004) "Risk and uncertainty in monetary policy", *American Economic Review*, Papers and Proceedings, May, 33-40.

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

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

[28] 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.

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

[30] Laubach, T. and J.C. Williams (2003), "Measuring the Natural Rate of Interest", *Review of Economics and Statistics*, 85(4), 1063-1070.

[31] Marcellino, M. and A. Musso (2009), "Real Time Estimates of the Euro Area Output Gap: Reliability and Forecasting Performance", European University Institute, unpublished manuscript.

[32] Marcellino, M., J. Stock and M.W. Watson (2006), "A Comparison of Direct and Iterated AR Methods for Forecasting Macroeconomic Series h-steps Ahead", *Journal of Econometrics*, 135, 499-526.

[33] 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.

[34] 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.

[35] Mitchell, J. and K. F. Wallis (2011), "Evaluating Density Forecasts: Forecast Combinations, Model Mixtures, Calibration and Sharpness", *Journal of Applied Econometrics*, forthcoming.

[36] Morley, J. and J. Piger (2011), "The Asymmetric Business Cycle", *Review of Economics and Statistics*, forthcoming.

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

[38] Orphanides, A. and S. van Norden (2002), "The Unreliability of Output-Gap Estimates in Real Time", *Review of Economics and Statistics*, 84, 4, 569-583.

[39] 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.

[40] Raftery, A.E., T. Gneiting, F. Balabdaoui and M. Polakowski (2005), "Using Bayesian Model Averaging to Calibrate Forecast Ensembles", *Monthly Weather Review*, 133, 1155–1174.

[41] Rosenblatt, M. (1952), "Remarks on a Multivariate Transformation", *The Annals of Mathematical Statistics*, 23, 470-472.

[42] Rudebusch, G.D. and L.E.O. Svensson (2002), "Eurosystem Monetary Targeting: Lessons from US data", *European Economic Review*, 46, 3, 417-442.

[43] Sims, C. (2008), "Inflation Expectations, Uncertainty, and Monetary Policy", Princeton University, unpublished manuscript, http://sims.princeton.edu/yftp/BIS608/

[44] Stock, J.H. and Watson, M.W. (1999). "Forecasting Inflation", *Journal of Monetary Economics* 44, pp. 293-335.

[45] Stock, J.H. and M.W. Watson (2007), "Has Inflation Become Harder to Forecast?", *Journal of Money, Credit, and Banking*, 39, 3-34.

[46] Taylor, J.B. (1993), "Discretion versus Policy Rules in Practice", *Carnegie-Rochester Conference Series on Public Policy*, 39, 195-214.

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

[48] 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.

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

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

[51] Weidner, J. and J.C. Williams (2011), "How Big is the Output Gap?" , FRBSF, *Economics Letter*, 2009-19, June 12, 2009, updated January 28, 2011.

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

Appendix 1: Output trend definitions

We summarize 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. Following Hodrick and Prescott (1997, HP), we set the smoothing parameter to be 1600 for our quarterly US data.[20] 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. Christiano and Fitzgerald (2003) propose an optimal finite-sample approximation to the band-pass filter, without explicit modelling 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. The Beveridge and Nelson (1981) decomposition relies on a priori assumptions about the correlation between permanent and transitory innovations. The 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 Unobserved Components model assumes q_t is decomposed into trend, cyclical and irregular components

$$q_t = \mu_t^7 + y_t^7 + \xi_t, \;\; \xi_t \sim i.i.d.\; N(0, \sigma_\xi^2),\; t = 1, \ldots, T \tag{A1.1}$$

[20] 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

$$\mu_t^7 = \mu_{t-1}^7 + \beta_{t-1} + \eta_t, \quad \eta_t \sim i.i.d.\ N(0, \sigma_\eta^2) \tag{A1.2}$$

$$\beta_t = \beta_{t-1} + \zeta_t, \quad \zeta_t \sim i.i.d.\ N(0, \sigma_\zeta^2). \tag{A1.3}$$

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:

$$\begin{bmatrix} y_t^7 \\ y_t^{7*} \end{bmatrix} = \rho \begin{bmatrix} \cos\lambda & \sin\lambda \\ -\sin\lambda & \cos\lambda \end{bmatrix} \begin{bmatrix} y_{t-1}^7 \\ y_{t-1}^{7*} \end{bmatrix} + \begin{bmatrix} \kappa_t \\ \kappa_t^* \end{bmatrix} \tag{A1.4}$$

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: Predictive densities from the VAR model

This appendix provides the formula for the predictive density from the VAR models with non-informative priors. For related discussion see Zellner (1971), pp. 233-236.[21]

Stack the VAR model, (1), as

$$\mathbf{Y}^j = \mathbf{Z}^j \mathbf{A}^j + \mathbf{E}^j, \tag{A2.2}$$

where $\mathbf{Y}^j = \begin{bmatrix} \mathbf{y}_1^j \\ \mathbf{y}_2^j \\ \\ \mathbf{y}_T^j \end{bmatrix}$, $\mathbf{y}_t^j = \left[\pi_t \; y_t^j\right]$, $\mathbf{z}_t^j = \mathbf{y}_{t-1}^j$, $\mathbf{Z}^j = \begin{bmatrix} \mathbf{z}_1^j \\ \mathbf{z}_2^j \\ \\ \mathbf{z}_T^j \end{bmatrix}$ with $\mathbf{E}^j$ defined from $\boldsymbol{\epsilon}_t^j$ conformably.

Denote the usual Ordinary Least Squares coefficient estimators as $\widehat{\mathbf{A}}^j$ and $\widehat{\boldsymbol{\Sigma}}^j$. Integrating out the coefficients, as in Zellner, the one step ahead posterior predictive density for $\mathbf{y}_{T+1}^j$ follows a multivariate t with mean $\mathbf{z}_{T+1}^j \widehat{\mathbf{A}}^j$, covariance $\left[1 + \mathbf{z}_{T+1}^j (\mathbf{Z}^{j\prime}\mathbf{Z}^j)^{-1} \mathbf{z}_{T+1}^{j\prime}\right] \widehat{\boldsymbol{\Sigma}}^j$, with T degrees of freedom.

We note that h-step ahead densities ($h > 1$) could also be produced in a similar manner utilizing the direct forecast methodology; see Marcellino, Stock and Watson (2006).

[21] In what follows, we use $\mathbf{Z}'$ to denote the transpose of $\mathbf{Z}$.

Figure 1: Grand Ensemble of Baseline and Dove, Output Gap Nowcast

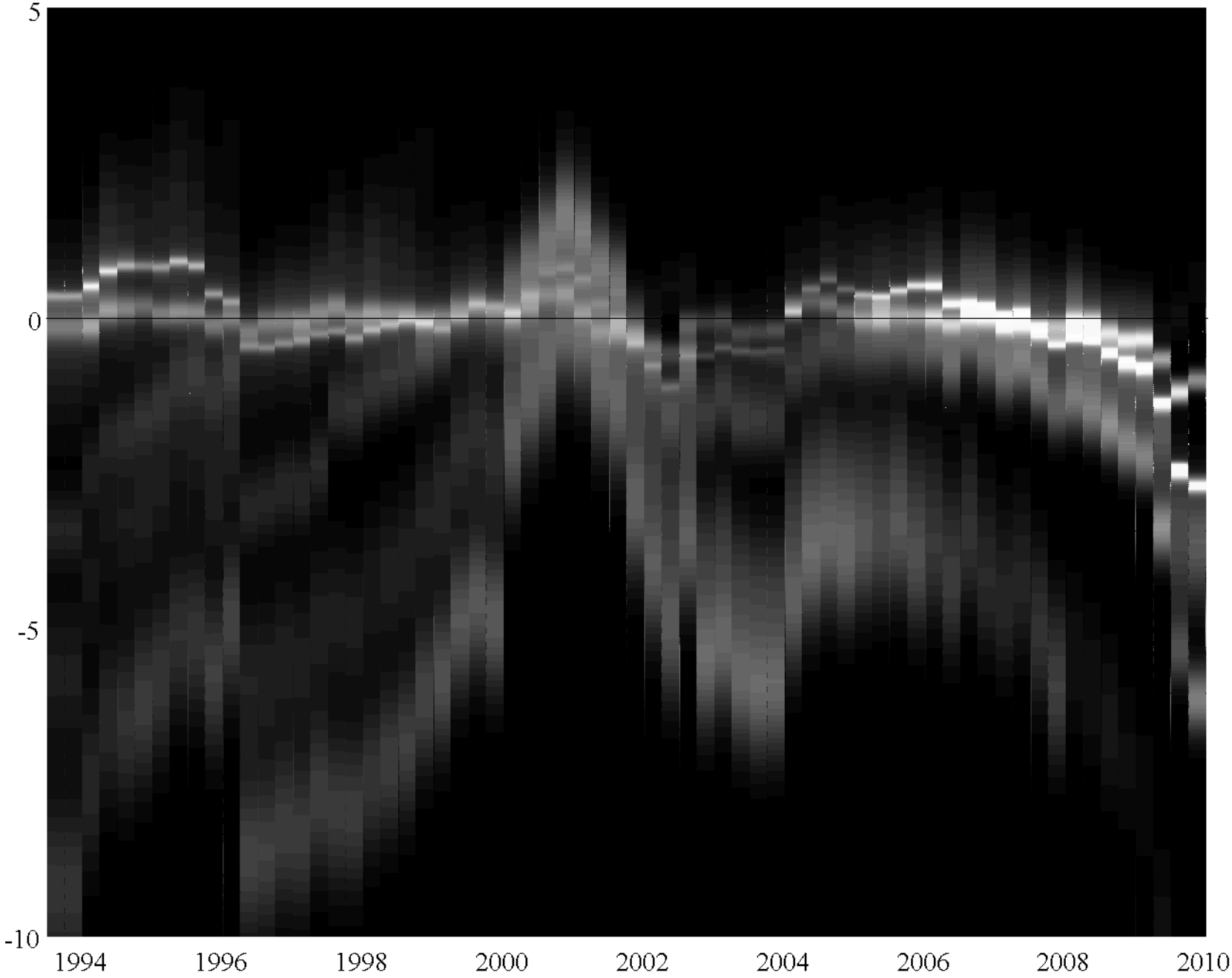

Figure 2: Probability of a Negative Output Gap, Baseline and Grand Ensemble

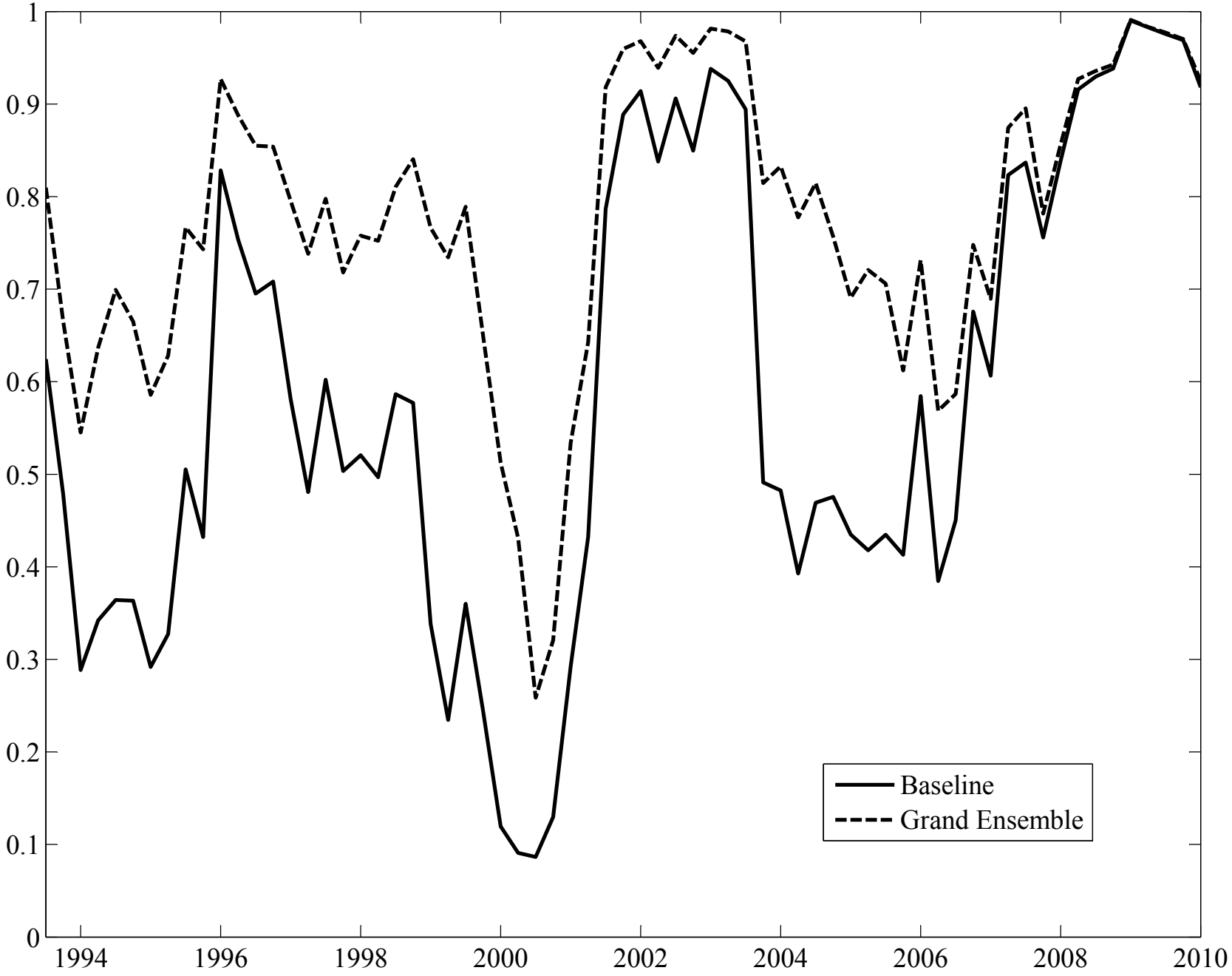

Figure 3: Baseline Ensemble Confidence Bands for the Output Gap, with Selected Point Estimates