WBS Economic Modelling & Forecasting Group

Working Paper Series: No 1
April 2013

Measuring Output Gap Nowcast Uncertainty

Anthony Garratt
Birbeck, University of London, UK

James Mitchell
Warwick Business School, University of Warwick, UK

Shaun Vahey
ANU, Canberra, Australia

Measuring Output Gap Nowcast Uncertainty*

Anthony Garratt
(Birkbeck,
University of London)

James Mitchell
(Warwick Business School)

Shaun P. Vahey
(ANU)

April 16, 2013

Abstract

We propose a methodology to gauge the uncertainty in output gap nowcasts across a large number of commonly-deployed vector autoregressive (VAR) specifications in inflation and the output gap. Our approach utilises many output gap measures to construct ensemble nowcasts for inflation using a linear opinion pool. The predictive densities for the latent output gap utilise weights based on the ability of each specification to provide accurate probabilistic forecasts for inflation. In an application based on US real time data, nowcasting over the out of sample evaluation period from 1991q2 to 2010q1, we demonstrate that a system of bivariate VARs produces well-calibrated ensemble densities for inflation, in contrast to univariate autoregressive benchmarks. The implied nowcast densities for the output gap are multimodal and indicate considerable uncertainty. For example, we assess the probability of a negative output gap at around 45% between 2004 and 2007. Despite the Greenspan policy regime, there remained a substantial risk that the nowcast for output was below potential in real time. We extend our methodology to include distinct output gap measures, based on alternative filters. We show that, in our application, the nowcast density for the output gap displays sensitivity to the detrending method.

Keywords: Output gaps; 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, RBNZ and the RBA is gratefully acknowledged. Contact: Anthony Garratt; a.garratt@bbk.ac.uk. We benefitted greatly from comments by Knut Are Aastveit, Michael Andersson, Mike Clements, Dean Croushore, Domenico Giannone, James Hamilton, Massimiliano Marcellino, Michael McCracken, Simon van Norden, David Papell, Mike Smith, Liz Wakerly, Mark Watson, Ken West and two anonymous referees. Thanks to seminar participants at the RBA, Norges Bank, the University of Adelaide, ANU, De Nederlandsche Bank, RBNZ, CIRANO, Eurostat and the Bank of England.

1 Introduction

The sign and size of the output gap are important for formulating monetary policy; see, for example, Weidner and Williams (2011). In the aftermath of the Great Recession there has been a lively debate about the size of the current output gap and the implications for monetary policy; e.g., see Bullard (2012). This debate reflects a common perception of considerable uncertainty across contemporaneous output gap measurements. Orphanides and van Norden (2002), for example, document the large variation across output gap measures, utilising several commonly-used detrending methods.

Unfortunately, the extant literature on output gaps does not provide convenient tools for assessing nowcast uncertainty across output gap measures. This absence of appropriate probabilistic tools frustrates policymakers given that contrasting output gap point estimates often imply different paths for inflation (and other policy relevant macro aggregates). To perform probabilistic assessments about economic aggregates policymakers routinely combine the evidence across many specifications; see, for example, Greenspan (2004) and the discussions by Feinstein, King and Yellen (2004). But currently monetary policy decisions rely on judgemental probabilistic assessments with little formal analysis available for addressing specification uncertainty.

This paper proposes a methodology to gauge the uncertainty in output gap nowcasts across a large number of commonly-deployed vector autoregressive (VAR) specifications in inflation and the output gap. Our approach involves accounting for model uncertainty by varying the output gap measures across the VAR specifications. A linear mixture of experts framework, sometimes referred to as a linear opinion pool, is deployed to construct ensemble probabilistic nowcasts for inflation. Predictive densities for the latent output gap utilise weights based on the ability of each specification to provide accurate nowcasts for inflation. For each VAR specification, we measure the Kullback-Leibler “distance” between the real-time out of sample nowcast density for inflation and the “true” but unknown density using the logarithmic score. In contrast, a conventional econometric analysis based on our VAR model space would report many nowcasts of inflation and the output gap, one from each specification, or perhaps select a preferred specification. The idea of using the predictive ability for inflation to measure the latent output gap has previously been utilised by (among others) Laubach and Williams (2003) and Weidner and Williams (2011).

In our application, spurred by the analyses of output gaps and inflation in 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 in inflation and the output gap. We explore the model uncertainties in the 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 in our application are a selection of commonly-used univariate filters. But, our methodology could also be applied to multivariate detrending methods, models with optimising behaviour and survey information. Each VAR specification in our system produces one quarter ahead forecasts for both inflation and a single output gap. Given the time lag in the release of real-time macro data, these forecasts provide contemporary predictions. (Hereafter, we use the terms nowcast and one quarter ahead forecast, interchangeably.)

A key insight of this paper is that probabilistic forecasts for latent macro variables, like the output gap, accommodating a wide variety of model uncertainties, can be constructed by utilising a form of “ensemble” forecasting. Under 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 utilised to produce the ensemble forecast densities (here, the linear opinion pool), 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 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 application. Nevertheless, the ensemble nowcast densities for each macro variable need not be symmetric and can accommodate multi-modality.

We emphasise that although our approach shares the objective of accounting for specification uncertainty with Bayesian Model Averaging (BMA), our methodology offers important practical and conceptual advantages. Papers using approximate likelihood BMA methods to account for specification uncertainty include Garratt *et al.* (2003) and Garratt *et al.* (2008, 2009b). Related work by Canova and Ferroni (2011) uses Bayesian simulation. Using out-of-sample information offers greater protection against over fitting, see Eklund and Karlsson (2007), but requires a (predictive) likelihood. When the forecasts come from judgement or survey information the likelihood is typically unknown; see Wallis (2005). Furthermore, ad hoc restrictions on the distributional family for the combined density—such as, Gaussianity—can impair forecast performance with an incomplete model space. Theoretical analysis, with examples, by Geweke (2010) and Geweke and Amisano (2011) demonstrate the benefits of opinion pools when, as in our analysis, the model space is incomplete.

Turning to the results in our US application, 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 structural breaks), produces well-calibrated (unimodal and symmetric) nowcast densities for US inflation. In contrast, autoregressive benchmarks (without the output gap) produce poorly calibrated densities (with or without stochastic volatility). Our baseline ensemble 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 potential from our baseline ensemble. We find that the policymaker is almost never certain of the sign of the current output gap throughout our evaluation. For example, we calculate the probability of a negative output gap at around 45% between 2004 and 2007. Despite the Greenspan policy regime, there remained a substantial risk that the nowcast for real output was below potential in real time.

We extend our methodology to include distinct output gap measures, based on alternative filters or survey evidence. These filters could be univariate or multivariate. We

account for the uncertainty across these additional specifications by utilising a two-step combination. In effect, this is an ensemble of ensembles, which we refer to as a “Grand Ensemble”. We illustrate the extended approach in our application by considering additional measures of the output gap based on linear time trends for output (with and without breaks). We label this alternative structural perspective of the output gap a “dove” scenario since these output gap measures often imply that potential output exceeds observed output in real time. This “Grand Ensemble” produces well-calibrated nowcasts for US inflation. But consideration of our alternative scenario dramatically changes the assessment of the output gap. The nowcast density for the output gap displays sensitivity to the detrending method.

The remainder of this paper is structured as follows. In section 2 we describe our methods, including our methodology for ensemble forecasting using the linear opinion pool. In section 3 we report our results from the application forecasting US inflation and describe our predictive densities for the latent output gap. We draw some conclusions in the final section.

2 Methods

2.1 Vector autoregressive specifications

A VAR model space is a common starting point for modelling inflation. For example, Rudebusch and Svensson (2002) and Laubach and Williams (2003) start with bivariate VARs, and add additional explanatory relationships and restrictions. To illustrate our ideas, we consider a bivariate VAR model space for inflation, π_t , and the output gap, y_t , of the type explored by Orphanides and van Norden (2002, 2005):

$$\begin{bmatrix} \pi_t \\ y_t^j \end{bmatrix} = \begin{bmatrix} a_{\pi\pi}^j & a_{\pi y}^j \\ a_{y\pi}^j & a_{yy}^j \end{bmatrix} \begin{bmatrix} \pi_{t-1} \\ y_{t-1}^j \end{bmatrix} + \begin{bmatrix} \epsilon_{\pi t}^j \\ \epsilon_{y t}^j \end{bmatrix}, \quad t = 1, \dots, T, \quad (1)$$

where $[\epsilon_{\pi t}^j, \epsilon_{y t}^j]' \sim i.i.d. N(\mathbf{0}, \Sigma^j)$. 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 $j = 1, \dots, J$. For expositional ease, we ignore the intercept and restrict the lag order of all VARs to one.

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 $j = 1, \dots, J$. 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:

$$q_t = \mu_t^j + y_t^j. \quad (2)$$

In addition to the output gap measures in our VAR specifications, we consider uncertainty over two other auxiliary assumptions. First, following (among others) Stock and Watson (1999), we allow for shifts in the relationships between inflation and output gaps. To achieve this, we extend the model space to consider a class of models which allow for a break in both the conditional mean and variance. As Jore, Mitchell and Vahay (2010) show, the presence of shifting volatilities during the Great Moderation can be accommodated by allowing for breaks in the variance, as well as the conditional mean.

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 should also emphasise that nothing in our approach requires a bivariate VAR structure. We simply work with this model space for simplicity. Additional equations and even identifying conditions pose no conceptual problems for our methodology, although the computational burden would increase.¹ Our principle concern is to account for model (or specification) uncertainty; and to quantify its effect on output gap nowcasts. This model uncertainty takes the form of uncertainty, not just about which output gap measurements to use, but also how to relate the ensuing output gap estimates to inflation.²

2.2 Ensemble forecasting

Armed with our many VAR specifications, it is straightforward to produce forecast densities for both inflation and the different output gap measures. 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 Garratt *et al* (2009a). 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 utilise 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 (the VAR forecast density, in our application); see also Hall and Mitchell (2007).⁴ Sometimes the application of opinion pools in economic forecasting is referred to as “density forecast combination”. Jore, Mitchell and Vahey (2010) utilise linear opinion pools with VAR forecasts but do not consider nowcasting the output gap. Geweke (2010) discusses the differences between opinion pools and mixture models, providing specific empirical examples with small numbers of models.

¹A simple extension of our approach might consider model uncertainty regarding what macro economic indicators to be included as exogenous variables in the VAR.

²In principle, the approach could be extended to account for uncertainty about parameter estimation and data revisions.

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

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

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 utilise linear (or linearised) and Gaussian models which the policymaker believes to be false. (Currently, the aggregation of these expert opinions is conducted informally in monetary policy institutions.)

With monetary policy practice in mind, we approximate the unknown forecast densities for inflation and the output gap by taking (time-varying) combinations of a large number of 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*.⁵ In this paper, we construct the policymaker’s implied forecast densities for the output gap based on their 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 $i = 1, \dots, N$ VAR specifications, the ensemble density for inflation is defined by the LOP:

$$p(\pi_\tau) = \sum_{i=1}^N w_{i,\tau} g(\pi_\tau | I_{i,\tau}), \quad \tau = \underline{\tau}, \dots, \bar{\tau}, \quad (3)$$

where $g(\pi_\tau | I_{i,\tau})$ are the one step ahead forecast densities from model i , $i = 1, \dots, N$ 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.⁶ Furthermore, the weights may change with each recursion in the evaluation period $\tau = \underline{\tau}, \dots, \bar{\tau}$.

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

$$q(y_\tau) = \sum_{i=1}^N w_{i,\tau} h(y_\tau^j | I_{i,\tau}), \quad \tau = \underline{\tau}, \dots, \bar{\tau}, \quad (4)$$

where $h(y_\tau^j | I_{i,\tau})$ 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 construct the weights for the output gap ensemble.

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

⁵Since the forecast density refers to an unobserved variable, it would be inappropriate for the policymaker to use any single imperfectly measured output gap as the true outturn to gauge the fit of the density.

⁶The restriction that each weight is positive could be relaxed; for discussion see Genest and Zidek (1986).

$$w_{i,\tau} = \frac{\exp \left[\sum_{\underline{\tau}-\kappa}^{\tau-1} \ln g(\pi'_{\tau} | I_{i,\tau}) \right]}{\sum_{i=1}^N \exp \left[\sum_{\underline{\tau}-\kappa}^{\tau-1} \ln g(\pi'_{\tau} | I_{i,\tau}) \right]}, \quad \tau = \underline{\tau}, \dots, \bar{\tau} \quad (5)$$

where $\underline{\tau} - \kappa$ to $\underline{\tau} - 1$ comprises the training period used to initialise 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 behaviour. For example, the ensemble 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.

2.3 Grand Ensemble

Our approach so far has assumed that there are J candidate detrending methods utilised to derive the output gap measures. In practice, researchers and policymakers often disagree about detrending methods. For example, some researchers prefer (parsimonious) univariate methods, others favour multivariate (but inevitable misspecified) approaches. Furthermore, some measures of the output gap are implied by Dynamic Stochastic General Equilibrium models. And many policymakers reach probabilistic assessments of the output gap without reference to explicit models—in effect, providing survey information.

A straightforward extension of our methodology can allow distinct structural assessments, or scenarios, about the output gap utilising a two-step combination. In effect, an ensemble of ensembles, which we refer to as a “Grand Ensemble”, simply requires an alternative sequence of nowcast densities for inflation and the output gap. Regardless of the source of these alternative densities, they can be evaluated using the logarithmic score and included in a second stage linear opinion pool. Notice that, in doing so, the researcher again weights the candidate output gap measures (nowcast densities) according to the ability to provide accurate probabilistic forecasts for inflation.

3 Application: nowcast densities for inflation and the output gap in the US

We turn now to the details of our application. 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 benchmark univariate forecasting models that do not consider the output gap. Then, we describe the output gap nowcast densities produced by both the baseline ensemble and the Grand Ensemble of the baseline and an alternative dove scenario.

3.1 Data and model space details

We use the same real-time US data set as Orphanides and van Norden (2002, 2005) and 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, $\dots$, $\tau - 1$. 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 400.

Our out of sample evaluation period is: $\tau = \underline{\tau}, \dots, \bar{\tau}$ where $\underline{\tau} = 1991q2$ and $\bar{\tau} = 2010q1$ (76 observations). In order to implement the 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).

Figure 1 plots the inflation series through the evaluation period (based on a single vintage of data, dated 2010q3). The plot reveals that inflation falls from just over 4% to vary around a mean of approximately 2% through the early 2000s. The volatility of inflation increases around 2000 and remains higher thereafter. Furthermore, inflationary pressures increase from 2004 onwards, until the onset of the Great Recession from 2007. The end of the period results in low inflation outcomes and high volatility.

Our baseline VAR setup uses seven output gap measures, $J = 7$, derived from univariate filters and deployed by Orphanides and van Norden (2002, 2005). The seven methods of univariate trend extraction in our baseline VAR are: quadratic (QD), Hodrick-Prescott (HP), forecast-augmented HP (HPF), Christiano-Fitzgerald (CF), Baxter-King (BK), Beveridge-Nelson (BN) and Unobserved Components (UC). Since we are interested in real-time prediction, we recursively estimate the trends using successive vintages of the real-time data. Each recursion uses a different vintage of data as well as an additional observation. Thereby, we directly mimic real-time estimation of both the output gap and its relationship with inflation. We summarise the seven well-known univariate filters in Appendix 1. We emphasise that although our example is illustrative, the seven univariate detrending methods used are popular detrending methods, and are commonly deployed at central banks. For examples, see Edge and Rudd (2012) and Marcellino and Musso (2011), at the Federal Reserve and European Central Bank, respectively.

Figure 2 demonstrates that our baseline seven detrending methods encompass considerable disagreement about the output gap through our evaluation period. In general, there is disagreement across the seven measures about both the size and sign of the output gap. For example, during the period between 2004 and 2007, associated with the Greenspan policies immediately prior to the Great Recession, several measures indicate that output did not exceed potential - specifically, HP, HPF and CF. In contrast, the QD and UC measures imply output exceeded potential by 1 – 3% over the same period.

There is, then, considerable uncertainty about the output gap.

As we stressed, there is nothing in our proposed methodology to prevent consideration of additional univariate, multivariate or survey-based output gap estimates which may be preferred by some policymakers. We illustrate the Grand Ensemble approach with a “dove” scenario. This utilises a range of linear and segmented trend specifications, with the output gap defined as the estimated residual, $\hat{v}_t$, from the (recursively estimated using real-time data) ordinary least squares regression:

$$q_t = a + b_1t + b_2D_{73,t} + b_3D_{84,t} + v_t, \quad t = 1, \dots, T, \quad v_t \sim i.i.d. N(0, \sigma^2). \quad (6)$$

A linear time trend without breaks is defined by $D_{73,t} = D_{84,t} = 0$. The segmented trend specifications allow for one breakdate and for two breakdates. The breaks are in 1973q4 (following Orphanides and van Norden, 2002), and in both 1973q4 and 1984q1. Like the baseline case, we vary the lag length and the break date in the ensuing dove VAR ensemble, for each of these three types of linear detrending method.⁷

The output gap measures utilised in this alternative dove scenario are plotted in Figure 3, along with the UC measure for comparison. The dove measures imply that potential output is higher, and therefore the output gap lower, throughout our evaluation period. For most of the period real-time output lies below potential according to these detrending methods. Hence, the path of the output gap would be reconcilable with the view that productivity picked up, following the slowdown in the 1970s and 1980s. From the mid 1990s, there was a lively debate, including discussions at the FOMC, about whether productivity growth in the US had accelerated and whether it would continue.⁸

For each of the 10 VAR specifications - seven in the baseline, three in the dove - 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, τ , 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. For each of the VAR specifications, for a given data vintage, a single coincident break in the conditional mean and variance for both variables implies a distinct estimation window. This raises the potential number of models considerably since each candidate break date defines a new VAR.

To macro researchers used to dealing with a single specification our decision to work with a single break date might seem surprising. However, our empirical methodology is not based on the assumption that a single specification is correct. Instead, we use the linear opinion pool to approximate the true but unknown nowcast densities for inflation and the output gap. Misspecification in the type or number of breaks can, of course, impair the approximation. (In practice, increasing the maximum number of breaks to three had no impact on the density forecasting performance for inflation.)

In every VAR considered, we vary the maximum lag length between one and four. (It is, of course, feasible to select the most preferred lag length. In practice, however, we found that lag length selection with information criteria caused a minor degradation

⁷In the results section, we compare our baseline ensemble output nowcasts with some multivariate filters. They are summarized in Appendix 1. It is clear from the comparison that the divergence among the seven univariate filters is wide enough to encompass multivariate estimates.

⁸See, for example, the debate in the Minutes of the September 24 1996 meeting of the FOMC; <http://www.federalreserve.gov/fomc/minutes/19960924.htm>. And see Jorgenson *et al* (2008) about the uncertainty and changing perceptions of US productivity.

in forecast quality in our applied work.) 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 utilise $N = 2576$ 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 LOP weights based on the fit of the individual VAR forecast densities for inflation in real time. 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. With the one quarter lag in the release of real-time data measurements, the forecast densities for inflation and the output gap are nowcasts.

3.2 Inflation nowcasts

To aid evaluation of the ensemble inflation nowcasts, we deploy a range of autoregressive models (with various lag lengths) as benchmarks, together with a random walk model, and an integrated moving average (IMA) specification for inflation. The IMA model for inflation takes the form: $\Delta\pi_t = \alpha + \varepsilon_t + \theta\varepsilon_{t-1}$. An IMA model has been found, by Stock and Watson (2007) to provide competitive point forecasts for US inflation.⁹

Our analysis focuses on the baseline ensemble, plus 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 baseline and dove ensembles, which we term a “Grand Ensemble” (GE).

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 realisation of the variable with respect to the forecast densities. We gauge calibration by examining whether the *pits* z_τ , where:

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

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.¹⁰

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

⁹We experimented with a stochastic volatility version but found its density forecasting performance to be inferior to the (recursively estimated) IMA. A rolling window of 40 observations, favoured by Stock and Watson (2007) in terms of point forecasts, gave a slightly superior density forecasting performance for inflation but was poorly calibrated according to the *pits* tests considered below.

¹⁰Given 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*-based tests computationally feasible for small N .

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, utilising 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, $p(\pi_\tau | I_{1,\tau})$ and $p(\pi_\tau | I_{2,\tau})$, so that the KLIC differential between them is the expected difference in their logarithmic scores: $d_\tau = \ln p(\pi'_\tau | I_{1,\tau}) - \ln p(\pi'_\tau | I_{2,\tau})$. 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, $\bar{d}_\tau$, under appropriate assumptions, has the limiting distribution: $\sqrt{T}\bar{d}_\tau \rightarrow N(0, \Omega)$, where Ω is a consistent estimator of the asymptotic variance of d_τ .¹¹ 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*-based 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 ensemble are well calibrated at a 90% level. Instances of appropriate calibration are marked in boldface.¹² The bottom six rows, bottom panel, in Table 1 show the *pits*-based tests for the univariate time series models: the AR models, including the random walk model, and the IMA (all based on expanding samples). In contrast to the baseline ensemble, all of these benchmark univariate models fail at least one of the *pits*-based tests at a 90% level.

While we choose to present detailed results in Table 1 for the full range of univariate benchmarks, in order to demonstrate that our results are robust to which specific univariate benchmark is preferred, henceforth we confine our discussion to the AR(2) and IMA benchmarks. The IMA is the best performing benchmark, in the sense that it fails only one of the *pits*-based tests at a 90% level and it is competitive on the logarithmic score. The AR(2) provides one mid-ranking alternative. The final column of Table 1 shows the average logarithmic score over the evaluation period. The KLIC-based tests indicate that the improvement in accuracy for the baseline ensemble, relative to these univariate benchmarks, is statistically significant: the KLIC p -values are 0.00 (for AR(2)) and 0.02 (for IMA).¹³

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

¹¹When 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.

¹²To 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% confidence level, of $(100 - 90)/4 = 2.5\%$, rather than 10%.

¹³A 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 when density forecasting. And, for the longer US and Australia samples, the ensembles also outperform simple equal-weighted density combinations.

tests at a 90% level. The dove ensemble betters the univariate time series models, but not by much: the KLIC-based tests using the logarithmic scores (last column) reveal a statistically insignificant improvement relative to the AR2 (IMA) with a p -value of 0.22 (0.82).¹⁴

Even though the inflation nowcast densities from the dove ensemble are not well calibrated - see Table 1, row 2 - a policymaker might wish to make use of both the dove and baseline specifications. Members of the Federal Open Market Committee could insist, for example, on consideration of an alternative scenario consistent with high productivity. Hence, we examine the Grand Ensemble (GE) of the two, presented in the third row of the top panel of Table 1.¹⁵ We observe well-calibrated inflation forecast densities from the GE at a 90% level on the basis of each individual *pits*-based test. Furthermore, the logarithmic score is somewhat lower than for the baseline ensemble.¹⁶ 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 for inflation throughout the evaluation period; and both appear to be well-calibrated.¹⁷

These results, which show that the univariate time series benchmarks can be bettered when forecasting inflation, 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, Mitchell and Vahey (2010) explain, these benchmarks are not particularly tough to beat when 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 therefore perform quite well in terms of point forecasting too.

Although we do not report the individual model recursive weights, (5), we note that the weights on the output gap measures in the baseline ensemble, and the Grand Ensemble, vary over time and that no one single trend measure dominates. There is uncertainty over the most preferred output gap measure for forecasting inflation; and the importance of the individual specifications varies over time.

¹⁴The 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.

¹⁵The weights on each ensemble were recursively computed using equation (5). An alternative approach, described by Hall and Mitchell (2007) and Geweke and Amisano (2011), selects the weights by optimisation. We found this alternative approach yielded weights, *pits*-based tests and average logarithmic scores that were qualitatively similar.

¹⁶The KLIC-based log score tests of the GE relative to the AR2 (IMA) benchmark have a p -value of 0.00 (0.02).

¹⁷We 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.

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
AR(1)	0.00	0.08	0.14	0.03	-1.560
AR(2)	0.04	0.13	0.16	0.04	-1.473
AR(3)	0.03	0.12	0.04	0.13	-1.444
AR(4)	0.02	0.13	0.08	0.17	-1.440
Random Walk	0.00	0.42	0.06	0.03	-1.601
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).

3.3 Output gap nowcasts

Since our approach is based on the idea that none of the different 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 4, we plot the forecast densities from the GE output gap.¹⁸ The dates refer to forecast observation. The nowcasts for the GE are from 1993q3 to 2010q1 (reflecting the additional 9 observations used as a training period to initialise the weights). The shades of the forecast densities indicate probability mass, with highest mass represented by black, and lowest mass represented by white. 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 (often from the dove scenario).¹⁹

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 5, 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.

¹⁸Recall 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.

¹⁹Visual inspection of the forecast densities for the baseline (dove) case reveals that more minor multi-modality also arises in the individual ensembles.

We make two observations about the time series of probabilities displayed in Figure 5 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 – 70% for the baseline ensemble. But for the GE the probability varies between 55% and 90% and always exceeds the baseline probability. Similarly, the probability of a negative output gap fluctuates around 45% between 2004 and 2007 for the baseline case. But the GE has the risk fluctuating between approximately 60% and 80% 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 2001, the probability of a negative output gap rises very sharply, regardless of whether we look at the baseline ensemble or the GE. However, this probability does rise from a trough. While it is unclear what threshold any policymaker might have in mind if and when classifying a recession, Figure 5 shows that the probability only rises above 0.5 (a common but arbitrary threshold) well into 2001. Interestingly, the NBER dates the (classical) recession earlier, from March through November 2001.²⁰ But unlike the NBER dates, the probabilistic forecasts we present in Figure 5 are real-time nowcasts and not lagging indicators. The NBER did not date the March 2001 peak until late November 2001; and they did not date the November 2001 trough until July 2003.

Returning to Figure 5, the two ensembles are again very close from 2007 onwards, with the probability that output is 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 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 5th and 95th percentiles from our baseline ensemble with various ex post output gap measures derived from multivariate methods. We consider the Congressional Budget Office (CBO) point estimates, based on a production function approach (with unemployment data), together with the Laubach-Williams (LW) point estimates reported by Weidner and Williams (2011), based on a Philips curve view of inflation. We also plot the (pseudo) real-time point estimates from Trimbur (2009) which utilises a measure of capacity utilisation.²¹ Figure 6 shows that the (revised) CBO measure often lies within the confidence bounds of the baseline ensemble, particularly from 2001 to 2006.²² However, the CBO estimates show a very marked drop with the recent slump from 2008, giving a string of 7 consecutive measurements below the 5th 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

²⁰The NBER definition of a recession rests on more than the behaviour of output. Two consecutive periods of falling output is neither necessary nor sufficient.

²¹We 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 in CBO (2001). Appendix 1 describes these three methods in more detail.

²²The GE, which allows for the dove scenario, gives more diffuse predictive densities for the output gap, with much higher probability mass below zero.

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. The Trimbur point estimates, denoted T, for most part lie within the confidence bands. The period 2002-2005 constitutes an exception, with a deeper recession implied. The Trimbur point estimates are close to the LW prior to 2001, and close to the CBO post 2006.

Figure 6 also shows that the uncertainty around the baseline ensemble output gap estimates is often quite large. For example, while the recession in 2001 is clearly picked up in Figure 6, with the output gap probabilities indicating strong evidence of output below potential from 2000, it is only late in 2001 that the 90% level excludes zero. And while the aftermath of the Great Recession is seen in Figure 6, with the output gap becoming negative with high probability from early 2008, during the Great Moderation, from a probabilistic perspective, the economy was typically close to trend.

It is also noteworthy that the business cycle is slightly asymmetric. (What are now regarded as) long duration upswings of the economy see nowcasts of output relatively close to trend. But, slumps are of relatively short duration with strong probabilistic support output below potential. For a discussion of the asymmetric business cycle, see Morley and Piger (2012).

It is worth repeating that the ensemble forecast densities for the latent output gap are “nowcasts”: one step ahead probabilistic forecasts from macro data, arriving with a one-period lag. That is, they are more timely than the CBO and LW measures shown, which use ex post data. Furthermore, our approach yields probabilistic information.

4 Conclusions

We propose a methodology to gauge the uncertainty in output gap nowcasts across a large number of commonly-deployed VARs (vector autoregressions). Output gap uncertainty is quantified according to the ability of the many VAR specifications, using different measures of the output gap, to predict inflation, where we construct ensemble nowcast densities for the output gap and inflation using a linear opinion pool. The ensemble density forecasts for inflation yield well-calibrated forecast densities for inflation in real time, in contrast to those from univariate models which ignore the output gap. Ensemble nowcast densities for the output gap, which quantify the uncertainty, are then constructed by weighting the different output gap estimates according to their respective abilities to density forecast inflation. These 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. In our application, the illustrative “dove” scenario (which uses alternative output gap measurements) on average raises the probability of a negative output gap during our evaluation period.

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. Gneiting, P. Guttorp 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 Buiness Cycle”, *Journal of Monetary Economics*, 7, 151-174.
- [8] Bullard, J. (2012), “Global Output Gaps: Wave of the Future?”, Remarks made on March 28 at the “Monetary Policy in a Global Setting: China and the United States” conference, organised by the St. Louis Fed and Beijing’s Tsinghua University.
- [9] Canova, F. and F. Ferroni (2011), “Multiple Filtering Devices for the Estimation of Cyclical DSGE Models”, *Quantitative Economics*, 2, 1, 73-98.
- [10] Christiano, L. and T.J. Fitzgerald (2003), “The Band Pass Filter”, *International Economic Review*, 44, 2, 435-465.
- [11] Clark, T.E. and M.W. McCracken (2010), “Averaging Forecasts from VARs with Uncertain Instabilities”, *Journal of Applied Econometrics*, 25, 5-29.
- [12] Congressional Budget Office (2001), “CBO’s Method for Estimating Potential Output: An Update”, August.
- [13] 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.
- [14] Corradi, V. and N.R. Swanson (2006), “Predictive Density Evaluation”, G. Elliott, C.W.J. Granger and A. Timmermann, eds, *Handbook of Economic Forecasting*, North-Holland, 197-284.

- [15] Croushore, D. and T. Stark (2001), "A Real-time Data Set for Macroeconomists", *Journal of Econometrics*, 105, 111-130.
- [16] Dawid, A.P. (1984), "Statistical Theory: the Prequential Approach", *Journal of the Royal Statistical Society A*, 147, 278-290.
- [17] 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.
- [18] 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.
- [19] Edge, R.M and J.B. Rudd (2012), "Real-Time Properties of the Federal Reserve's Output Gap", Finance and Economics Discussion Series 2012-86. Board of Governors of the Federal Reserve System (U.S.).
- [20] Eklund, J. and S. Karlsson (2007), "Forecast Combination and Model Averaging Using Predictive Measures", *Econometric Reviews*, 26(2-4), 329-363.
- [21] 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.
- [22] Garratt, A., Lee, K., Pesaran, M.H. and Y. Shin (2003), Forecast Uncertainties in Macroeconometric Modelling: An Application to the UK Economy, *Journal of the American Statistical Association*, 98, 464, 829-838.
- [23] 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.
- [24] Garratt, A., G. Koop, E. Mise and S.P. Vahey (2009a), "Real-Time Prediction with UK Monetary Aggregates in the Presence of Model Uncertainty", *Journal of Business and Economic Statistics*, 27, 480-491.
- [25] Garratt, A., K. Lee, E. Mise and K. Shields (2009b), "Real Time Representation of the Output Gap in the UK in the Presence of Model Uncertainty", *International Journal of Forecasting*, 25, 1, 81-102.
- [26] 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.
- [27] Genest, C. and J. Zidek (1986), "Combining Probability Distributions: a Critique and an Annotated Bibliography", *Statistical Science*, 1, 114-135.
- [28] Geweke, J. (2010), *Complete and Incomplete Econometric Models*, Princeton University Press.

- [29] Geweke, J. and G. Amisano (2011), “Optimal prediction pools”, *Journal of Econometrics*, 164, 130-141.
- [30] Giacomini, R. and H. White (2006), “Tests of Conditional Predictive Ability”, *Econometrica*, 74, 1545-1578.
- [31] Gneiting, T. and T. Thorarinsdottir (2010) “Predicting inflation: Professional Experts versus No-change Forecasts”, <http://arxiv.org/abs/1010.2318>.
- [32] Greenspan, A. (2004) “Risk and uncertainty in monetary policy”, *American Economic Review*, Papers and Proceedings, May, 33-40.
- [33] Hall, S.G. and J. Mitchell (2007), “Density Forecast Combination”, *International Journal of Forecasting*, 23, 1-13.
- [34] Harvey, A.C. (2006) “Forecasting with Unobserved Components Time Series Models” G. Elliot, C.W.J. Granger and A. Timmermann, eds, *Handbook of Economic Forecasting*, North-Holland, 327-412.
- [35] Hodrick, R. and E. Prescott (1997), “Post-War U.S. Business Cycles: An Empirical Investigation”, *Journal of Money, Banking and Credit*, 29, 1-16.
- [36] 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.
- [37] Jorgenson, D.W., M.S. Ho and K.J. Stiroh (2008), “A Retrospective Look at the U.S. Productivity Growth Resurgence”, *Journal of Economic Perspectives*, 22 (1), 3-24.
- [38] Kascha, C. and F. Ravazzolo (2010), “Combining Inflation Density Forecasts”, *Journal of Forecasting*, 29, 231-250.
- [39] Laubach, T. and J.C. Williams (2003), “Measuring the Natural Rate of Interest”, *Review of Economics and Statistics*, 85(4), 1063-1070.
- [40] Marcellino, M. and A. Musso (2011), “The reliability of real-time estimates of the euro area output gap”, *Economic Modelling*, 28(4), 1842-1856. Previously available as ECB Working Paper 1157.
- [41] 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.
- [42] 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.
- [43] 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.
- [44] Mitchell, J. and K. F. Wallis (2011), “Evaluating Density Forecasts: Forecast Combinations, Model Mixtures, Calibration and Sharpness”, *Journal of Applied Econometrics*, 26, 1023-1040.

- [45] Morley, J. and J. Piger (2012), “The Asymmetric Business Cycle”, *Review of Economics and Statistics*, 94, 208-221.
- [46] Noceti, P., J. Smith and S. Hodges (2003), “An Evaluation of Tests of Distributional Forecasts”, *Journal of Forecasting*, 22, 447-455.
- [47] 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.
- [48] 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.
- [49] 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.
- [50] Rosenblatt, M. (1952), “Remarks on a Multivariate Transformation”, *The Annals of Mathematical Statistics*, 23, 470-472.
- [51] Rudebusch, G.D. and L.E.O. Svensson (2002), “Eurosystem Monetary Targeting: Lessons from US data”, *European Economic Review*, 46, 3, 417-442.
- [52] Stock, J.H. and M.W. Watson (1999), “Forecasting Inflation”, *Journal of Monetary Economics* 44, 293-335.
- [53] Stock, J.H. and M.W. Watson (2007), “Why has U.S. Inflation Become Harder to Forecast?”, *Journal of Money, Credit, and Banking*, 39, 3-34.
- [54] Timmermann, A. (2006), “Forecast Combination”, G. Elliot, C.W.J. Granger and A. Timmermann, eds, *Handbook of Economic Forecasting*, North-Holland, 197-284.
- [55] Trimbur, T.M. (2009), “Improving Real-Time Estimates of the Output Gap”, Finance and Economics Discussion Series 2009-32. Board of Governors of the Federal Reserve System.
- [56] 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.
- [57] Wallis, K.F. (2005), “Combining Density and Interval Forecasts: a Modest Proposal”, *Oxford Bulletin of Economics and Statistics*, 67, 983-994.
- [58] Watson, M.W. (2007), “How Accurate are Real-Time Estimates of Output Trends and Gaps?”, *Federal Reserve Bank of Richmond Economic Quarterly*, Spring 2007.
- [59] 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.
- [60] Zellner, A. (1971), *An Introduction to Bayesian Inference in Econometrics*, John Wiley.

Appendix 1: Output trend definitions

We summarize the seven univariate 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 1600 for our quarterly US data.²³
3. Since the HP filter is two-sided filter it relates the time- t value of the trend to future and past observations. Moving towards the end of a finite sample of data, the HP filter becomes progressively one-sided and its properties deteriorate with the Mean Squared Error (MSE) of the unobserved components increasing and the estimates ceasing to be optimal in a MSE sense. We therefore follow Mise *et al.* (2005) and seek to mitigate this loss near and at the end of the sample by extending the series with forecasts. At each recursion the HP filter is applied to a forecast-augmented output series (again with smoothing parameter 1600), with forecasts generated from an univariate AR(8) model in output growth (again 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).²⁴
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) - and set the truncation parameter (the maximum lag length) at three years. As with the HP filter we augment our sample with AR(8) forecasts to fill in the ‘lost’ output gap observations at the end of the sample due to truncation. Watson (2007) reviews band-pass filtering methods.

²³We could, of course, allow for uncertainty in the smoothing parameter. We reduce the computational burden in this application by fixing this parameter at 1600.

²⁴There is always a question empirically about what model should be used to produce the forecasts used to augment the sample. Here, in following Garratt *et al* (2008), we deliberately select a high lag order to ensure no important lags are omitted - thereby we favor unbiasedness over efficiency when estimating the AR. But, of course, this is not the only modeling option. We did experiment with use of a model selection criterion, specifically the BIC, which is known to favor more parsimonious models. In fact, inference in our application is robust to either choice with near identical output gap estimates. In principle, our methodology can avoid any selection by simply treating each variant of the HP filter, or indeed any of the other filters, as a new output gap estimator; i.e. we can increase J . But in order both to keep this application manageable and to stick to a well-known representative model space we focus on these $J = 7$ filters.

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

$$q_t = \mu_t^7 + y_t^7 + \xi_t, \quad \xi_t \sim i.i.d. N(0, \sigma_\xi^2), t = 1, \dots, T \quad (A1.1)$$

where the stochastic trend is specified as

$$\mu_t^7 = \mu_{t-1}^7 + \beta_{t-1} + \eta_t, \quad \eta_t \sim i.i.d. N(0, \sigma_\eta^2) \quad A1.2 \quad (8)$$

$$\beta_t = \beta_{t-1} + \zeta_t, \quad \zeta_t \sim i.i.d. N(0, \sigma_\zeta^2). \quad A1.3 \quad (9)$$

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:

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

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. For a detailed description of the unobserved components approach see Harvey's (2006) literature review and the references therein.

The three multivariate trends used to compute the output gap, for comparative purposes, are:

1. The Congressional Budget Office (CBO) measure of trend or potential output uses a disaggregated approach described in the CBO paper "CBO'S Method for Estimating Potential Output: An Update", August 2001. Their measure is based on the framework of the Solow growth model, with a neoclassical production function at its core. It estimates trends in the components of GDP using a variant of Okun's law. According to Okun's law actual output exceeds its potential level when the unemployment rate is below its "natural" rate of unemployment. Conversely, when the unemployment rate exceeds its natural rate, output falls short of its potential level.
2. Laubach and Williams's (2003) trend output is computed using the Kalman filter to estimate joint time-varying measures of the natural rate of interest, output and trend growth rates. The theoretical link between the natural rate of interest, unobservable trend growth plus "other" unobservable factors is combined within an unobserved components framework, with measurement equations for the (unobservable) output gap and inflation, and transition equations in the other observable factors, the natural level of output and the trend growth rate.

3. Trimbur (2009) adopts a multivariate structural time series approach to modelling trend behaviour using a bivariate unobserved components model, which combines output with capacity utilisation as an auxiliary indicator. The model employs a vector of stochastic cyclical terms in the two series of assumed order n (in our results we set $n = 1$). These cycles are *similar* (i.e. have identical autocorrelation functions and spectral densities) but are not identical in their movement. They evolve according to a specific sine-cosine specification, with a smooth trend in output (see Trimbur (2009), page 23 for details).

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.²⁵

Stack the VAR model, (1), as

$$\mathbf{Y}^j = \mathbf{Z}^j \mathbf{A}^j + \mathbf{E}^j, \quad (\text{A2.2})$$

where

$$\mathbf{Y}^j = \begin{bmatrix} \mathbf{y}_1^j \\ \mathbf{y}_2^j \\ \mathbf{y}_T^j \end{bmatrix}, \mathbf{y}_t^j = [\pi_t \ y_t^j], \mathbf{z}_t^j = \mathbf{y}_{t-1}^j, \mathbf{Z}^j = \begin{bmatrix} \mathbf{z}_1^j \\ \mathbf{z}_2^j \\ \mathbf{z}_T^j \end{bmatrix} \text{ with } \mathbf{E}^j \text{ defined from } \epsilon_t^j \text{ con-}$$

formably.

Denote the usual Ordinary Least Squares coefficient estimators as $\hat{\mathbf{A}}^j$ and $\hat{\Sigma}^j$. Integrating out the coefficients, the one step ahead posterior predictive density for $\mathbf{y}_{T+1}^j$ follows a multivariate t with mean $\mathbf{z}_{T+1}^j \hat{\mathbf{A}}^j$, covariance $[1 + \mathbf{z}_{T+1}^j (\mathbf{Z}^{j'j})^{-1} \mathbf{z}_{T+1}^{j'}] \hat{\Sigma}^j$, with T degrees of freedom; see Zellner (1971), pp. 233-236.

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).

²⁵In what follows, we use $\mathbf{Z}'$ to denote the transpose of $\mathbf{Z}$.

Figure 1: Inflation (annualized, %)

Figure 2: Output Gaps, Baseline Measures (annualized, %)

Figure 3: Output Gaps, Dove and UC Measures (annualized, %)

Figure 4: Grand Ensemble of Baseline and Dove, Output Gap Nowcast

Figure 5: Probability of Negative Output Gap, Baseline and Grand Ensemble (%)

Figure 6: Output Gap, Confidence Bands and Point Estimates (annualized, %)