Finance and Economics Discussion Series
Divisions of Research & Statistics and Monetary Affairs
Federal Reserve Board, Washington, D.C.

The Reliability of Inflation Forecasts Based on Output Gap Estimates in Real Time

Athanasios Orphanides and Simon van Norden

2004-68

The Reliability of Inflation Forecasts Based on Output Gap Estimates in Real Time

Athanasios Orphanides and Simon van Norden*

November 2004

Abstract

A stable predictive relationship between inflation and the output gap, often referred to as a Phillips curve, provides the basis for countercyclical monetary policy in many models. In this paper, we evaluate the usefulness of alternative univariate and multivariate estimates of the output gap for predicting inflation. Many of the ex post output gap measures we examine appear to be quite useful for predicting inflation. However, forecasts using real-time estimates of the same measures do not perform nearly as well. The relative usefulness of real-time output gap estimates diminishes further when compared to simple bivariate forecasting models which use past inflation and output growth. Forecast performance also appears to be unstable over time, with models often performing differently over periods of high and low inflation. These results call into question the practical usefulness of the output gap concept for forecasting inflation.

KEYWORDS: Phillips curve, output gap, inflation forecasts, real-time data.

JEL Classification System: E37, C53.

Athanasios Orphanides is an adviser in the Division of Monetary Affairs at the Board of Governors of the Federal Reserve System, a research fellow of the Centre for Economic Policy Research, and a fellow of the Center for Financial Studies. E-mail: Athanasios.Orphanides@frb.gov. Simon van Norden is a Professeur Agrégé at the HEC Montréal and a CIRANO fellow. E-mail: simon.van-norden@hec.ca.

* We benefited from presentations of earlier drafts at the European Central Bank, CIRANO, the Federal Reserve Bank of Philadelphia Conference on Real Time Data Analysis, the Centre for Growth and Business Cycle Research, as well as at the annual meetings of the American Economics Association, the European Economics Association and the Canadian Economics Association. We would also like to thank Sharon Kozicki, Tim Cogley, Jeremy Piger, Todd Clark, Desire Vencatachellom, Ken West and two anonymous referees for useful comments and discussions. Athanasios Orphanides wishes to thank the Sveriges Riksbank and European Central Bank for their hospitality during September 2001 when part of this work was completed. Simon van Norden wishes to thank the SSHRC and the HEC Montréal for their financial support. The opinions expressed are those of the authors and do not necessarily reflect the views of the Board of Governors of the Federal Reserve System.

1 Introduction

A stable predictive relationship between inflation and a measure of deviations of aggregate demand from the economy's potential supply—the "output gap"—provides the basis for many formulations of activist countercyclical stabilization policy. Such a relationship, referred to as a Phillips curve, is often seen as a helpful guide for policymakers aiming to maintain low inflation and stable economic growth. According to this paradigm, when aggregate demand exceeds potential output, the economy is subject to inflationary pressures and inflation should be expected to rise. Under these circumstances, policymakers aiming to contain the acceleration in prices might wish to adopt policies restricting aggregate demand. Similarly, when aggregate demand falls short of potential supply, inflation should be expected to fall, prompting policymakers to consider the adoption of expansionary policies.[1] Even assuming that the theoretical motivation for a relationship between the output gap and inflation is fundamentally correct, a number of issues may complicate its use for forecasting in practice. First, the definition of "potential output"—and the accompanying "output gap"—that might be useful in practice is far from clear. Given a definition of the output gap, its exact empirical relationship with inflation is not known a priori and would need to be determined from the data. Second, even if the proper conceptual and empirical relationships were identified, the operational usefulness of the output gap will be limited by the availability of timely and reliable estimates of the identified concept. As is well known, empirical estimates of the output gap are generally subject to significant and highly persistent revisions. (For example, see Orphanides and van Norden (2002).) The subsequent evolution of the economy leads to improved historical estimates of the gap by providing useful information about the state of the business cycle. As a result, considerable uncertainty regarding the value of the gap remains even long after it would be needed for

[1]The widespread use of models featuring estimated "Phillips curves" of various forms for monetary policy analysis at numerous central banks and other institutions is evidence of the appeal of this paradigm. See Bryant, Hooper and Mann (1993) and Taylor (1999) for collections of monetary policy evaluations that feature such estimated models.

forecasting inflation. This suggests that although the output gap may be quite useful for historical analysis, its practical usefulness for forecasting inflation in real time may be quite limited.

In this paper we assess the usefulness of alternative estimation methods of the output gap for predicting inflation, paying particular attention to the distinction between *suggested usefulness*—based on *ex post* analysis using revised output gaps, and *operational usefulness*—based on simulated real-time out-of-sample analysis.[2] First, using out-of-sample analysis based on ex post estimates of the output gap, we confirm that many concepts appear to be useful for predicting inflation. This is as would be expected since the implicit Phillips curve relationships recovered in this manner are similar to the relationships commonly found in empirical macroeconometric models. To assess their operational usefulness, we generate out-of-sample forecasts based on *real-time* output gap measures; those constructed using only data (and parameter estimates) available at the time forecasts are generated.[3] We compare the resulting forecasts to both autoregressive forecasts of inflation and bivariate forecasts that employ information from output growth as well as past inflation.

Our findings show that forecasts using *ex post* estimates of the output gap severely overstate the gap's usefulness for predicting inflation. Real-time forecasts using the output gap are often less accurate than forecasts that abstract from the output gap concept altogether. And the relative usefulness of real-time output gap estimates diminishes further when compared to simple bivariate forecasting models which use past inflation and output growth. In some cases, we find certain measures of the output gap produce superior forecasts of inflation. However, relative performance seems to vary considerably over time, with models which perform relatively well in some periods performing relatively poorly in others. Thus,

[2]Our analysis is related to investigations of the usefulness of the unemployment gap for forecasting inflation, such as Stock and Watson (1999), Atkeson and Ohanian (2001), and Fisher, Liu and Zhou (2002). In some macroeconometric models, unemployment gaps and output gaps are related through Okun's law.

[3]For this exercise, we rely on the real-time dataset for macroeconomists which was created and is maintained by the Federal Reserve Bank of Philadelphia. See Croushore and Stark (2001) for background information regarding this database.

past forecast performance may provide little guidance in selecting an operationally useful definition of the output gap going forward.

The remainder of this paper is organized as follows. In sections 2 and 3 we define the output gap concepts used and detail the methodology of our forecasting exercise. The main results are presented in section 4 and section 5 concludes.

2 Trends and Cycles Ex Post and in Real Time

One way to define the output gap is as the difference between actual output and an underlying unobserved trend towards which output would revert in the absence of business cycle fluctuations. Let q_t denote the (natural logarithm of) actual output during quarter t, and μ_t its trend. Then, the output gap, y_t can be defined as the cyclic component resulting from the decomposition of output into a trend and cycle component:

$$q_t = \mu_t + y_t$$

Since the underlying trend is unobserved, its measurement, and the resulting measurement of the output gap, very much depends on the choice of estimation method, underlying assumptions and available data that are brought to bear on the measurement problem. For any given method, simple changes in historical data and the availability of additional data can change, sometimes drastically, the resulting estimates of the cycle for a given quarter.

Evidence of the difference between historical and real-time estimates of output gaps has been presented by Orphanides and van Norden (2002). In Table 1, we present some of the summary reliability indicators they examine for twelve alternative measures of the output gap which we employ in our analysis.[4] These results mirror those of Orphanides and van Norden (2002). We find that revisions in real-time estimates are often of the same magnitude as the historical estimates themselves and that, for many of the alternative

[4]Brief descriptions of the various measures appear in Appendix A. Further details, including the output gaps used in this study, as well as the programs and data used to create them, are freely available from the authors at *http://www.hec.ca/pages/simon.van-norden.*

methods, historical and real-time estimates frequently have opposite signs.

The importance of *ex post* revisions to output gap estimates suggests that the presence of a predictive relationship between inflation and *ex post* estimated output gap measures does not guarantee that the output gap will be useful for forecasting inflation in practice. Simply, the *ex post* estimates of output gaps at a point in time may differ substantially from estimates which could be made without the benefit of hindsight. As well, these differences may hinder the real-time estimation of the presumed predictive relationship, further complicating the real-time forecasting problem.

2.1 Data Sources and Vintages

We use the term *vintage* to describe the values for data series as published at a particular point in time. Most of our data is taken from the real-time data set compiled by Croushore and Stark (2001); we use the quarterly vintages from 1965Q1 to 2003Q3 for real output. Construction of the output series and its revision over time is further described in Orphanides and van Norden (1999, 2002). We use 2003Q3 data as "final data" recognizing, of course, that "final" is very much an ephemeral concept in the measurement of output.

To measure inflation, we use the change in the log of the consumer price index (CPI). We use this both for our forecasting experiments and also to estimate measures of the output gap in multivariate models that include inflation. CPI data are revised much less than output data, with changes in seasonal factors causing most of the revisions. We therefore use the 2003Q3 vintage of CPI data for all of our analysis. This allows us to focus on the effects of revisions in the output data and the estimated output gap in our analysis. One of our models (Structural VAR) also uses data on interest rates, which are never revised.

2.2 Measuring Output Gaps

We construct output gap estimates using a variety of different models, as listed in Table 1. Each of the output gap models is used to produce gap estimates of varying vintages. Each

output gap vintage uses precisely one vintage of the output data. An estimated output gap is called a final estimate if it uses the final data vintage. Note that all the output gap estimation techniques (aside from the Hodrick-Prescott filter) require that one or more parameters be estimated to fit the data. Such estimation was repeated for every combination of technique and vintage. This means, for example, that in constructing output gap vintages from an unobserved components (UC) model spanning the period 1969Q1-2003Q3 (139 quarters), we reestimate the model's parameters 139 times, and then store 139 series of smoothed estimates.

3 A Forecasting Experiment

We are interested in quantifying the extent to which the output gap concept provides a practical means of improving forecasts of inflation. The answer will clearly depend on a large number of factors, such as the time period of interest, the way in which forecasts are constructed, the benchmark against which such forecasts are compared, and the loss function used to evaluate the quality of different forecasts. We restrict our attention to US CPI inflation since 1969 and use the mean-squared forecast error (MSFE) to compare forecast quality.

3.1 Forecasting Inflation and Benchmarks

Let $\pi_t^h = log(P_t) - log(P_{t-h})$ denote inflation over h quarters ending in quarter t. We examined forecasts of inflation at various horizons but use one year (h=4) as our baseline. Note that because of reporting lags, data for quarter t first become available in quarter $t+1$. Thus, a four-quarter ahead forecast is a forecast five quarters ahead of the last quarter for which actual data are available.[5] Our objective, therefore, is to forecast π_{t+4}^4 with data for quarter $t-1$ and earlier periods.

[5]Since the last datapoint in our sample is for the 2003Q2 quarter, this implies that 2002Q1 is the last datapoint available for forming a forecast we can use in our evaluation experiment.

We examine simple linear forecasting models of the form:

$$\pi^h_{t+h} = \alpha + \sum_{i=1}^{n} \beta_i \cdot \pi^1_{t-i} + \sum_{i=1}^{m} \gamma_i \cdot y_{t-i} + e_{t+h} \tag{1}$$

where n and m denote the number of lags of inflation and the output gap in the equation. We estimate the unknown coefficients $\{\alpha, \beta_i, \gamma_i\}$ by ordinary least squares. We set n and m using a variety of different methods; in the results presented here we use the Bayes Information Criterion (BIC). Results with other lag selection methods were found to give similar conclusions.

To provide a benchmark for comparison, we estimate a univariate forecasting model of inflation based on equation (1) but omitting the output gaps. We refer to this model as the autoregressive (AR) benchmark. Of course, the problem faced by forecasters in practice is more complex than the one we consider. One obvious and important difference is that the information set available to policymakers is much richer. It is therefore possible that output gaps might improve on simple univariate forecasts of inflation but not on forecasts using a broader range of inputs. For this reason, tests against an autoregressive forecast benchmark should be considered to be weak tests of the utility of empirical output gap models.

To provide a slightly stronger test, we also consider benchmark forecasts which replace the output gap in (1) with the first difference of the log of real output. As St-Amant and van Norden (1998) argue, using output growth in this way can be interpreted as *implicitly* defining an estimated output gap as a one-sided filter of output growth with weights based on the estimated coefficients of equation (1). van Norden (1995) refers to such estimates as TOFU gaps (Trivial Optimal Filter–Unrestricted). We refer to this as the TF benchmark forecast and interpret it as a simple reduced-form inflation forecast that uses a slightly larger information set than the AR benchmark, one which contains historical information on both prices and output growth. Comparing forecasts based on output gaps to the TF benchmark aids in isolating the usefulness (or lack thereof) of the economic structure and other restrictions embedded in the construction of the output gaps.

3.2 Forecasting and Output Gap Revisions

Several practical issues complicate the use of (1) for inflation forecasting. Since the suitable number of lags of inflation and the output gap n and m, and the coefficients of the equation are not known a priori, these need to be estimated with available data. As our sample increases and additional data become available, these estimates change. In addition, output gap estimates (like output data) are revised over time. This in turn, can influence the selected number of lags and the coefficients of equation (1) estimated in any given sample. In addition, given the parameters of the equation, revisions in the output gap will directly change the forecast value of inflation.

We therefore use (1) to construct 3 to 4 different kinds of forecasts for each output gap model. These forecasts differ in the way lag lengths are determined and in the way the output gap model is used.

Let $y_t^{i,j}$ be an estimate of the output gap at time t formed using data of vintage i, where $i > t$ and $j = t$ or $i - 1$. For non-UC models (i.e. all except the Watson, Harvey-Clark, Harvey-Jaeger, Kuttner and Gerlach-Smets models) the index j is irrelevant; $y_t^{i,t} = y_t^{i,i-1}$. For UC models, $j = t$ denotes a *filtered* output gap estimate; although the model parameters are estimated from using data up to $i - 1$, the Kalman filter recursions to estimate the gap do not use data beyond t. For these same models, $j = i - 1$ denotes a *smoothed* estimate; although $y_t^{i,t}$ and $y_t^{i,i-1}$ use the same parameter estimates to calculate the output gap, the latter also uses the data after t to recursively update its estimate of y_t. When T= 2003Q3, the terminology of Orphanides and van Norden (2002) refers to the time series $\{y_t^{T,T-1}\}$ as *Final* estimates of the gap and to $\{y_t^{T,t}\}$ as *Quasi-Final* estimates. We will commonly refer to these as FL and QF estimates.

These different kinds of output gap estimates are used to construct different kinds of forecasts. The first of these uses fixed lag lengths with final estimates of the output gap to

recursively estimate the forecasting equation

$$\pi_{t+h}^{h} = \hat{\alpha}^{t-1} + \sum_{i=1}^{\hat{n}} \hat{\beta}_{i}^{t-1} \cdot \pi_{t-i}^{1} + \sum_{i=1}^{\hat{m}} \hat{\gamma}_{i}^{t-1} \cdot y_{t-i}^{T,T-1} + e_{t+h} \tag{2}$$

where T refers to 2003Q3. This replicates the kind of recursively-estimated, out-of-sample forecasting experiments which are commonly performed but which ignore output gap revision. These forecasts are infeasible because they require information (Final estimates of output gaps) which is not available at the time the forecast is made. They also estimate the optimal lag lengths $\hat{m}, \hat{n}$ ex post. We refer to this Fixed-Lag Final-estimate forecast as FL-FL.

In the case of UC models, we can construct similar forecasts using Quasi-Final rather than Final estimates of the output gap

$$\pi_{t+h}^{h} = \hat{\alpha}^{t-1} + \sum_{i=1}^{\hat{n}} \hat{\beta}_{i}^{t-1} \cdot \pi_{t-i}^{1} + \sum_{i=1}^{\hat{m}} \hat{\gamma}_{i}^{t-1} \cdot y_{t-i}^{T,t} + e_{t+h} \tag{3}$$

Orphanides and van Norden (2002) note that the difference between the Final and Quasi-Final estimates of the output accounts for the bulk of the revisions in the output gaps they examine. The difference between the accuracy of these and the Final gap forecasts above helps us to understand the relative importance of errors in gap estimation for forecast accuracy. Like the Final gap forecasts, these forecasts are infeasible. We refer to these as FL-QF forecasts.

We also construct feasible forecasts which attempt to mirror closely the forecasts which practitioners would construct using such output gap models. Specifically, in these forecasts the lag lengths for both explanatory variables vary over time and are estimated recursively. The output gap series is also updated with its latest available vintage every time the parameters of the forecasting equation are re-estimated. The resulting Variable-Lag Real-Time

output gap (VL-RT) forecasting equation takes the form[6]

$$\pi_{t+h}^{h} = \hat{\alpha}^{t-1} + \sum_{i=1}^{\hat{n}^{t-1}} \hat{\beta}_{i}^{t-1} \cdot \pi_{t-i}^{1} + \sum_{i=1}^{\hat{m}^{t-1}} \hat{\gamma}_{i}^{t-1} \cdot y_{t-i}^{t,t-1} + e_{t+h} \tag{4}$$

where the superscripts on $(\hat{m}, \hat{n})$ indicate the information set used to estimate the lag lengths. While these are the most realistic forecasts we examine, they are also the most difficult to compute. Among other things, they require more than just the real-time gap estimates presented in Orphanides and van Norden (2002); they require *all vintages* of the complete estimated output gap series.

To summarize, we can construct two or three series of forecasts for each output gap model we analyze: (1) using recursive estimation, fixed lag lengths and final output gap estimates, (2) using recursive estimation, fixed lag lengths and quasi-final output gap estimates (which are only available for the 5 UC models we examine), and (3) using recursive estimation, variable lag lengths and all vintages of smoothed output gap estimates. We also examine one other type of forecast, one which uses variable lag lengths and final output gaps and which we refer to as VL-FL. Like the FL-QF forecast, this helps to isolate the contribution of output gap revision to forecast accuracy. As we will see below, however, these methods differ in the appropriate ways one should conduct inference.

3.3 Forecast Evaluation

We wish to evaluate the quality of the resulting forecasts by testing the null hypothesis that a given pair of models have equal MSFEs. Various tests of equal forecast accuracy have been proposed in recent years, notably by Diebold and Mariano (1995) for forecasting models without estimated parameters and by West (1996) for models with estimated parameters. While such tests have been popular, the assumptions they require are unfortunately violated

[6]Note that in equation (4) we use *smoothed* estimates of the output gap ($y_{t-i}^{t,t-1}$) rather than *filtered* estimates ($y_{t-i}^{t,t-i}$). This reflects the common practice of practitioners, which is to use the most accurate possible estimate of the gap in estimating their forecast equations. Limited experiments which replaced these smoothed estimates with filtered estimates suggest that this does not have a major impact on forecast performance. Koenig, Dolmas and Piger (2003) discuss how the use of data of varying vintage affects forecast accuracy.

for some of the hypotheses of interest here.

First, the use of Diebold-Mariano statistics with standard normal critical values for asymptotic inference is justified only if the two models being compared are not nested. However, when using suitable lag lengths, the output gap models nest the AR benchmark model. Clark and McCracken (2001) suggest alternative tests for the case of nested models, while Clark and McCracken (2002) find that the limiting distribution of these statistics is non-pivotal for forecast horizons greater than one period. To compare these models, we therefore use the MSE-F statistic proposed by McCracken (2000), which takes the form

$$\textit{MSE-F} = P \cdot \frac{(MSFE_1 - MSFE_2)}{MSFE_2} \tag{5}$$

where P is the number of forecasts, $MSFE_1$ is the MSFE of the restricted model and $MSFE_2$ is the MSFE of the unrestricted model. The distribution of the statistic under the null hypothesis of equal MSFE is estimated via a bootstrap experiment with 2000 replications, as detailed in Appendix B. Because these distributions are non-pivotal, the test statistics are bootstrapped anew for every different choice of (P, h, y, m, n). This means that every p-value we report for the AR benchmark is based on its own set of 2000 bootstrap experiments.

Second, while the available asymptotic theory underlying all such tests allows for the coefficients in an equation like (1) to be re-estimated over time, it assumes that lag lengths are fixed during the recursive estimation, that the data remain fixed during the recursive estimation, and that the data are not estimated.

All these assumptions are violated for the VL-RT forecasts we construct, so no p-values are presented for this case.

Inference in the case of the TF benchmark is more straightforward as the models of interest are no longer nested. Accordingly, we base our inference on the test statistics proposed by Diebold and Mariano (1995) and West (1996). Specifically, letting $d_t \equiv e_{it}^2 - e_{jt}^2$ be the difference in squared forecast errors between model i and model j at time t, $\overline{d} \equiv$

$T^{-1} \cdot \sum_{t=1}^{T}(d_t)$ the mean difference, and $\rho_\tau \equiv T^{-1} \cdot \sum_{t=\tau+1}^{T}(d_t - \overline{d}) \cdot (d_{t-\tau} - \overline{d})$ the estimated autocovariance of d_t at lag τ, we compute the test statictic:

$$z = \frac{\overline{d}}{\sqrt{\Omega/T}} \tag{6}$$

where $\Omega \equiv \sum_{l=-6}^{6}(1 - |l|/7) \cdot \rho_l$ is the Newey-West (1986) Heteroscedasticity and Autocorrelation (HAC) robust estimator of the long-run variance of d_t. West (1996) shows that under conventional assumptions this statistic is asymptotically normally distributed under the null hypothesis of equal forecast accuracy when the parameters of the forecast model are estimated by ordinary least squares. We therefore calculate and report 2-sided p-values for the TF benchmark using the standard normal distribution. Again, this asymptotic theory is not applicable to the VL-RT forecasts, so no p-values are reported in this case.

4 Does the Output Gap Improve Forecasts of Inflation?

4.1 Are Improvements in Forecast Accuracy Significant?

Our next step is to examine the results of the forecasting experiments described above. Table 2 shows the results of formal tests for differences in MSFE between the two benchmark models and the twelve output gap models. The upper panel of the table compares forecasts constructed using final output data, final estimates of the output gap, and constant lag lengths in the forecasting equation (FL-FL). The middle panel of the table shows the comparable results when using quasi-final rather than final (i.e. filtered rather than smoothed) estimates of the output gap (FL-QF). Since such estimates can only be constructed from UC models of the output gap, only results for the five UC models are presented. In both cases, we see the MSFE of the benchmark models, the fractional improvement in MSFE relative to the benchmark models ($(MSFE_{Benchmark} - MSFE_{Gap})/MSFE_{Gap}$) and the p-value for the test of the null hypothesis that the MSFEs of the benchmark and the gap model are equal. Differences between these two panels are entirely due to the effects of *ex post* revisions of output gaps.

The first thing apparent from the top panel of the table is that all the gap models forecast better than the autoregressive benchmark model when using final output gaps. In all but one case the differences in MSFE are greater than 10 per cent, and in four of the twelve cases they are greater than 30 per cent. The suggested improvement is statistically significant at the 5 per cent level for all but the SVAR model and at the one per cent level for nine of the twelve models. These results confirm the conventional wisdom that *ex post* output gaps appear to help forecast inflation. They also show that out-of-sample tests have sufficient power to detect relevant differences in MSFE.

The evidence supporting the usefulness of output gaps is weakened when the benchmark model is changed by adding real output growth to the forecasting equation (the TF model). As can be seen on the right side of the top panel, three of the twelve gap models now have larger MSFEs than the benchmark, and only five of the twelve show an improvement of more than 10 per cent. The differences in MSFE are significant at the 10 per cent level in only three cases and are never significant at the 5 per cent level. However, comparison of the *significance* of the differences in MSFE across the two benchmarks is complicated by differences in the tests used for nested and non-nested models, as explained in section 3.3. Note, in particular, that the reported p-values for nested models (the AR benchmark) are based on *one-sided* tests, while those for non-nested models (the TF benchmark) are based on *two-sided* tests. In addition, Clark and McCracken (2001, 2002) suggest that the MSE-F statistic, which is used for the AR benchmark, is more powerful than the z statistic used for the TF benchmark.

The apparent superiority of output-gap based forecasts is also weakened by the use of quasi-final rather than final estimates of the gap, shown in the middle panel. Improvements over the AR benchmark are now lower in every case, falling 10 to 20 per cent, and in one case output-gap-based forecasts are less accurate than the benchmark. However, improvements in forecast accuracy are still significant at or near the 5 per cent significance level in the four

remaining cases. The situation changes further if we instead use the TF benchmark. Four of the five models now forecast less accurately than the benchmark model. Ignoring the effects of output gap revisions evidently tends to overstate the importance and significance of output gaps for forecasting inflation.

The bottom panel of Table 2 shows the results of tests for differences in MSFE between the two benchmark models and the twelve output gap models when the forecasts are constructed with time-varying lag lengths and real-time output gap estimates (VL-RT). This change also increases the MSFE of the benchmark AR model by a little over 10 per cent.

The relative accuracy of these real-time forecasts is almost always lower than that of the ex post forecasts analysed in the top panel of the table. Drops in relative MSFE are substantial for many models. As noted earlier. the normal asymptotic theory results are not valid in this case so no p-values are reported. Crude simulations based on bootstrapped MSE-F statistics, however, suggested that several output gap models which appeared to forecast significantly better than the AR benchmark in the top panel no longer showed a significant difference in accuracy.

The reversal in the performance of the output gap models relative to the output growth (TF) benchmark, is even more striking. This can be seen by comparing the top and bottom panels on the right-hand side of the table. In real time, *none* of the output gap models examined forecasts better than the TF benchmark.

4.2 The Effect of Output Gap Revisions on Relative Forecast Accuracy

To better understand the causes for the changes in MSFE noted above, Table 3 compares the MSFEs of three different forecasting experiments. The first is identical to that documented in the upper panel of the previous table, using final output data and gap estimates as well as constant lag lengths in the forecasting equation (FL-FL). The second experiment uses the same output data and gap estimates, but now updates the lag lengths each time the forecast coefficients are recursively re-estimated (VL-FL). The third experiment is identical

to that documented in the bottom panel of the previous table, using time-varying lag lengths and real-time output gap estimates (VL-RT). Differences in outcomes between the first two experiments isolate the effects of variations in lag length. Differences between the second two experiments similarly isolate the effects of output gap revision.

The table shows that the introduction of time-varying lag lengths has important effects on forecast accuracy. A priori, such time-variation may improve forecasts if the underlying relationship is unstable over time. On the other hand, it may introduce another source of estimation error, which could reduce forecast accuracy. The table shows that all forecasts see a reduction in accuracy, averaging 15 per cent. The benchmarks forecasts see changes in MSFEs which are very close to the average.

Moving from Final to real-time output gap estimates has no effect on the AR benchmark forecast, but tends to make other forecasts less accurate. While the average effects of this change are smaller than those of changes in lag length, the impact varies much more across models. Four models see their accuracy improve while three see their MSFE rise by more than 20 per cent. Note that the TF benchmark sees the greatest improvement in accuracy. Evidently, revisions in output growth contain useful information about future inflation.

The net effect of the changes in lag length determination and data vintage worsens forecast accuracy in all but one case. The net effect on the AR benchmark is somewhat less than average, while the TF benchmark improves more than any other model.

The results above suggest that some output gap models forecast inflation more accurately than an autoregressive model, even when using real-time output gap estimates. However, none of the output gap models we examine forecasts inflation as well as simple models which use both past inflation and output growth. Further, the relative performance of different models is greatly affected by the use of real-time rather than ex post output gap estimates. Finally, uncertainty about the lag structure also adds considerably to MSFEs.

4.3 The Robustness of Changes in Forecast Accuracy

We now investigate the robustness of the results presented in Table 2. Table 4 examines the effects of changing the period over which forecasts are evaluated. The full 1969-2002 sample is split into two roughly equal halves, with the 1969-1983 portion characterized by relatively high and volatile inflation, whereas prices were more stable over the 1984-2002 period. The greater volatility of inflation in the former period implies that least-squares methods applied to the full sample tend to emphasize the fit of the model over the former period. Perhaps as a consequence, the full-sample results presented in Table 2 largely reflect forecast performance over the first half of the sample. Results for the low-inflation period after 1983 may be a more relevant guide for contemporary decision-making, but they differ from the full-sample results in several ways.

First, looking at forecasts with final output gaps, we see that the AR benchmark has become harder to beat. Nine of the 12 models see their relative MSFEs decline, and only five can reject the null of equal forecast accuracy at the 5 per cent level (compared to 11 in the earlier portion of the sample). This decline in the predictability of inflation has been noted previously in other studies, for example, Atkeson and Ohanian (2001), and Fisher, Liu and Zhou (2002). The picture for the TF benchmark is less clear; while the relative performance of the output gap models improves somewhat in the latter sample, there is little evidence of significantly different forecast accuracy.

Second, looking at forecasts with real-time output gaps, it appears that it has become increasingly difficult to forecast as well as the benchmarks. Out of 12 models 11 (10) have larger MSFEs than the AR (TF) benchmark in the post-83 period. The Band-Pass filter is the only model to forecast inflation better than either benchmark in the recent period, giving over a 20 per cent reduction in MSFE. It is also interesting to note that, consistent with the reported decline in the predictability of inflation, the AR benchmark now forecasts slightly better in real time than the TF benchmark.

One possible explanation for the difference in results across the two sample periods is parameter instability, a feature which has been noted by other research on inflation forecasts, in particular, Stock and Watson (1996, 1999), and Clark and McCracken (2003). Indeed, examination of changes in the period over which the forecasting model is estimated suggested some evidence of such instability for some of our output gap forecasting models. We also considered the effects of changing the forecasting horizons, forecasting changes rather than levels of inflation, using different lag selection criteria, and using nominal rather than real income growth as a benchmark. (Detailed results are available from the authors upon request.) Based on a review of these findings, it appears that the results shown in Table 2 are among the *best* that can be obtained for inflation forecasts from simple linear forecasting models using output gaps.

Having considered this evidence, one might also ask which of the output gap models examined here a practitioner should use to forecast inflation (if forced to do so.) It would appear that the deterministic trend models (Linear, Quadratic and Breaking) were often among the worst-performing in real-time, and should probably be avoided for that reason. UC models which estimated Phillips Curves (Kuttner and Gerlach-Smets) had some of the largest differences in performance when used with real-time rather than final estimates. The Band-Pass and the Beveridge-Nelson methods perform better in our simulated real-time experiments. However, their success appeared to be sensitive to the forecast horizon used. Rather than rely on any of these output gap models, our analysis suggests that a practitioner could do well by simply taking into account the information contained in real output growth without attempting to measure the level of the output gap—the TOFU model. This model was consistently among the best performers, particularly over the post-1983 forecast sample.

5 Conclusion

Forecasting inflation is a difficult but essential task for the successful implementation of monetary policy. The hypothesis that a stable predictive relationship between inflation and the output gap—a Phillips curve—is present in the data, suggests that output gap measures could be useful for forecasting inflation. This has served as the basis for empirical formulations of countercyclical monetary policy in many models. We find that many alternative measures of the output gap *appear* to be quite useful for forecasting inflation, on the basis of ex post analysis. That is, a historical Phillips curve is suggested by the data, and final (constructed ex post) estimates of the output gap are useful for understanding subsequent movements in inflation.

However, this historical usefulness does not imply a similar operational usefulness. Our simulated real-time forecasting experiment suggests, instead, that the predictive ability of many different output gap measures may be illusory. Output gaps typically can not forecast inflation as well out of sample as simple linear models of inflation and output growth (although the differences are mostly not statistically significant.) This is particularly true if we restrict our attention to the post-1983 period. These rather pessimistic findings regarding the output gap mirror earlier investigations regarding the predictive power for forecasting inflation of "unemployment gaps," that is the difference between the rate of unemployment and estimates of the NAIRU. As demonstrated by Staiger, Stock and Watson (1997a,b) and Stock and Watson (1999), estimates of the NAIRU are inherently unreliable, and simulated out-of-sample forecasting exercises do not indicate a robust improvement in inflation forecasts from using information about unemployment. Stock and Watson (1999) also show that better inflation forecasts may be obtained by indicators other than the unemployment gap. Our analysis suggests similar conclusions regarding the output gap as well. Instead of using output gaps, forecasts of inflation which simply incorporate information from the growth rate of output appear to forecast inflation as well or better.

Finally, we note that these negative findings regarding the usefulness of real-time measures of the output gap do not necessarily invalidate the potential usefulness of the theoretical Phillips curve framework per se, nor that of ex post constructed output gaps for historical analysis. That said, the dubious contribution of real-time measures of the output gap for forecasting inflation brings into question their role in the formulation of reliable real-time policy analysis.

References

Atkeson, Andrew and Lee E. Ohanian, "Are Phillips Curves Useful for Forecasting Inflation," *Federal Reserve Bank of Minneapolis Quarterly Review*, 25(1), 2-11, Winter 2001.

Baxter, Marianne; King, Robert G., "Measuring Business Cycles: Approximate Band-Pass Filters for Economic Time Series" *The Review of Economics and Statistics* 81(4) November 1999.

Beveridge, Stephen and Charles R. Nelson, "A New Approach to Decomposition of Economic Time Series into Permanent and Transitory Components with Particular Attention to Measurement of the 'Business Cycle'," *Journal of Monetary Economics,* 7, 151-174, 1981.

Blanchard. Olivier and Danny Quah, "The Dynamic Effects of Aggregate Demand and Supply Disturbances," *American Economic Review, 79(4), 655-673, September 1989.*

Bryant, Ralph C., Peter Hooper and Catherine Mann eds. *Evaluating Policy Regimes: New Research in Empirical Macroeconomics*, Brookings: Washington DC, 1993.

Cayen, Jean-Philippe and Simon van Norden "Fiabilité des estimations de l' écart de production au canada." *Bank of Canada working paper* 2002-10.

Clark, Peter K., "The Cyclical Component of U.S. Economic Activity," *Quarterly Journal of Economics* 102(4), 1987, 797-814.

Clark, Todd E. and Michael W. McCracken, "Tests of Equal Forecast Accuracy and Encompassing for Nested Models" *Journal of Econometrics*, 105, 85-110, 2001.

Clark, Todd E. and Michael W. McCracken, "Evaluating Long-Horizon Forecasts" *Federal Reserve Bank of Kansas City mimeo*, 2002.

Clark, Todd E. and Michael W. McCracken, "The predictive content of the output gap for inflation: resolving in-sample and out-of-sample evidence." *Federal Reserve Bank of Kansas City mimeo*, 2003.

Croushore, Dean and Tom Stark, "A Real-Time Data Set for Macroeconomists," *Journal of Econometrics*, 105, 111-130, November, 2001.

Diebold, Francis X. and Roberto S. Mariano, "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics*, 13, 1995, 253-265.

Fisher, Jonas D. M., Chin Te Liu, and Ruilin Zhou, "When Can we Forecast Inflation?" Federal Reserve Bank of Chicago *Economic Perspectives*, 1Q/2002, 30-42, 2002.

Gerlach, Stefan and Frank Smets, "Output Gaps and Inflation: Unobserable-Components Estimates for the G-7 Countries." Bank for International Settlements mimeo, Basel 1997.

Harvey, Andrew C., "Trends and Cycles in Macroeconomic Time Series," *Journal of Business and Economic Statistics*, 3, 216-227, 1985.

Hodrick, Robert, and Ed Prescott, "Post-war Business Cycles: An Empirical Investigation," *Journal of Money, Credit, and Banking,* 29, 1997, 1-16.

Koenig, Evan F., Sheila Dolmas and Jeremy Piger, "The Use and Abuse of 'Real-Time' Data in Economic Forecasting," *Review of Economics and Statistics*, 85(3) August 2003, 618-628.

Kuttner, Kenneth N., "Estimating Potential Output as a Latent Variable," *Journal of Business and Economic Statistics*, 12(3), 1994, 361-68.

McCracken, Michael W., "Asymptotics for Out-of-Sample Causality" *University of Missouri mimeo* 2000.

Newey, Whitney K. and Kenneth D. West, "A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica* 55(3), 703-08, May 1987.

Orphanides, Athanasios and Simon van Norden, "The Reliability of Output Gap Estimates in Real Time," Finance and Economics Discussion Series 1999-38, August 1999.

Orphanides, Athanasios and Simon van Norden, "The Unreliability of Output Gap Estimates in Real Time," *Review of Economics and Statistics*, 84(4), 569-583, November 2002.

Orphanides, Athanasios and Simon van Norden, "The Reliability of Inflation Forecasts Based on Output Gap Estimates in Real Time," *CIRANO working paper* 2003s-01.

St-Amant, Pierre and Simon van Norden, "Measurement of the Output Gap: A discussion of recent research at the Bank of Canada," Bank of Canada Technical Report No. 79, 1998.

Staiger, Douglas, James H. Stock, and Mark W. Watson, "How Precise are Estimates of the Natural Rate of Unemployment?" in Romer, Christina and David Romer, eds. *Reducing Inflation: Motivation and Strategy*, Chicago: University of Chicago Press, 1997a.

Staiger, Douglas, James H. Stock, and Mark W. Watson, "The NAIRU, Unemployment and Monetary Policy," *Journal of Economic Perspectives* 11(1), Winter 1997b, 33-49.

Stock, James H. and Mark W. Watson, "Evidence on Structural Instability in Macroeconomic Time Series Relations," *Journal of Business and Economic Statistics*, 14(1), 11-30, January, 1996.

Stock and Watson "Business Cycle Fluctuations in U.S. Macroeconomic Time Series." *NBER Working Paper* No. 6528, 1998, 83 p., prepared for *The Handbook of Macroeconomics*, edited by John B. Taylor and Michael Woodford.

Stock, James H. and Mark W. Watson, "Forecasting Inflation," *Journal of Monetary Economics,* 44, 293-335, 1999.

Taylor, John B., *Monetary Policy Rules*, Chicago: University of Chicago, 1999.

van Norden, Simon, "Why is it so hard to measure the current output gap?" Bank of Canada mimeo, 1995.

West, Kenneth D. "Asymptotic Inference About Predictive Ability." *Econometrica*, 64, 1996, 1067-84.

Table 1

Description of Alternative Output Gap Measures and Summary Reliability Statistics

Method	Data	Method Details	COR	AR	NSR	OPSIGN
Linear Trend	Univariate.		0.88	0.90	1.63	0.58
Quadratic Trend	Univariate.		0.51	0.97	1.06	0.42
Breaking Trend	Univariate.	Trend Break in 1973Q1, starting in 1977Q1.	0.77	0.87	0.81	0.28
Hodrick-Prescott	Univariate.	With $\lambda = 1600$.	0.52	0.93	1.05	0.45
Band Pass	Univariate.	6–32 quarters, series padded with AR forecasts.	0.72	0.77	0.77	0.36
Beveridge-Nelson	Univariate.	Assumes ARIMA(1,1,2).	0.84	0.09	0.63	0.30
Structural VAR	Trivariate.	Imposes long-run restrictions.	0.68	0.85	0.95	0.41
Watson	Univariate.	Local Level and AR(2).	0.88	0.87	1.50	0.55
Harvey-Clark	Univariate.	Local Linear Trend and AR(2).	0.75	0.92	0.91	0.39
Harvey-Jaeger	Univariate.	Local Linear Trend and Cycle.	0.56	0.86	0.92	0.50
Kuttner	Bivariate.	Watson model and inflation equation.	0.87	0.90	1.54	0.61
Gerlach-Smets	Bivariate.	Harvey-Clark model and inflation equation.	0.79	0.82	1.05	0.40

Notes: Univariate methods employ only real GNP/GDP data. Bivariate also employ CPI inflation. Trivariate also employs treasury bill data. The last four columns present summary measures of the reliability of real-time estimates of the output gap. All statistics are for the 1969:1–2003:1 period. **COR** denotes the correlation of the real-time and final estimates of the output gap, **AR** the first order serial correlation of the revision (the difference between the final and real-time series), **NSR** indicates the ratio of the root of the mean square revision and the standard deviation of the final estimate of the gap, and **OPSIGN** indicates the frequency with which the real-time and final gap estimates have opposite signs.

Table 2

Relative Improvement in MSFE

Method	AR	AR p-value	TF	TF p-value
Fixed Lags, Final Gaps				
Benchmark MSFE	0.494		0.436	
Linear Trend	0.302	0.009	0.148	0.164
Quadratic Trend	0.168	0.010	0.030	0.779
Breaking Trend	0.106	0.034	−0.024	0.778
Hodrick-Prescott	0.149	0.000	0.013	0.900
Band-Pass	0.134	0.000	0.000	0.997
Beveridge-Nelson	0.139	0.000	0.004	0.309
SVAR	0.047	0.121	−0.077	0.474
Watson	0.319	0.001	0.163	0.060
Harvey-Clark	0.270	0.002	0.120	0.162
Harvey-Jaeger	0.109	0.001	−0.022	0.811
Kuttner	0.336	0.008	0.178	0.079
Gerlach-Smets	0.362	0.001	0.201	0.052
Fixed Lags, Quasi-Final Gaps				
Watson	0.132	0.043	−0.002	0.979
Harvey-Clark	0.070	0.068	−0.056	0.374
Harvey-Jaeger	−0.032	0.811	−0.146	0.382
Kuttner	0.248	0.030	0.100	0.250
Gerlach-Smets	0.091	0.070	−0.038	0.414
Variable Lags, Real-time Gaps				
Benchmark MSFE	0.559		0.416	
Linear Trend	0.045		−0.219	
Quadratic Trend	0.021		−0.237	
Breaking Trend	0.043		−0.221	
Hodrick-Prescott	0.132		−0.154	
Band-Pass	0.283		−0.042	
Beveridge-Nelson	0.211		−0.095	
SVAR	−0.093		−0.323	
Watson	0.121		−0.163	
Harvey-Clark	0.147		−0.143	
Harvey-Jaeger	0.080		−0.193	
Kuttner	0.107		−0.173	
Gerlach-Smets	0.099		−0.179	

Notes: The AR benchmark is a univariate autoregressive forecast of inflation; the TF benchmark forecasts from a linear regression on lagged inflation and real output growth. Mean squared forecast errors (MSFE) for the two benchmark models are shown multiplied by 1000. The remaining figures in the AR and TF columns denote the relative improvements in MSFE for the output gap models, measured as $(A - B)/B$ where A is the MSFE of the benchmark and B is that of the output gap model. The p-values for the AR benchmark are for the null that $B \geq A$, based on the statistic in equation (5). The p-values shown for the TF benchmark are for two-sided test of the null that $A = B$, based on the statistic in equation (6). See section 3.3 and Appendix B for further discussion of the construction and interpretation of the p-values. The forecast horizon is 4 quarters and forecast performance is evaluated over the period from 1969Q1 to 2002Q1. Forecast equation estimation starts in 1955Q1. Fixed lag lengths are (1,1) while varying lag lengths are reset every quarter using BIC.

Table 3

The Effect of Lag Selection and Data Vintage

	MSFE			Change in MSFE (percent)		
Method	FL-FL	VL-FL	VL-RT	FL to VL	FL to RT	Total
AR benchmark	0.494	0.559	0.559	−13.0	0.0	−13.0
TF benchmark	0.436	0.496	0.416	−13.7	16.0	4.6
Linear Trend	0.380	0.438	0.533	−15.4	−21.7	−40.4
Quadratic Trend	0.423	0.500	0.545	−18.1	−9.0	−28.8
Breaking Trend	0.447	0.494	0.534	−10.6	−8.0	−19.5
Hodrick-Prescott	0.430	0.556	0.492	−29.2	11.5	−14.4
Band-Pass	0.436	0.502	0.434	−15.2	13.5	0.4
Beveridge-Nelson	0.434	0.482	0.460	−11.0	4.5	−6.0
SVAR	0.472	0.502	0.614	−6.4	−22.3	−30.1
Watson	0.375	0.433	0.497	−15.4	−14.9	−32.6
Harvey-Clark	0.389	0.448	0.486	−15.1	−8.4	−24.7
Harvey-Jaeger	0.446	0.577	0.516	−29.5	10.7	−15.7
Kuttner	0.370	0.402	0.503	−8.5	−25.3	−36.0
Gerlach-Smets	0.363	0.426	0.507	−17.3	−19.0	−39.6
Mean				−15.6	−5.2	−21.1
Std Dev				6.7	14.5	14.4

Notes:
MSFE denotes the mean squared forecast error (shown multiplied by 1000.)
FL-FL refers to forecasts using fixed lag lengths and final output gap estimates.
VL-FL refers to forecasts using variable lag lengths and final output gap estimates.
VL-RT refers to forecasts using variable lag lengths and real-time output gap estimates.
FL to VL refers to the change from FL-FL to VL-FL.
FL to RT refers to the change from VL-FL to VL-RT.
Total refers to the change from FL-FL to VL-RT.

Table 4

Relative Improvement in MSFE: Sub-sample Evaluation

	1969Q1–1983Q4				1984Q1–2002Q1			
Method	AR	p-value	TF	p-value	AR	p-value	TF	p-value
Fixed Lags, Final Gaps								
Benchmark MSFE	0.863		0.739		0.191		0.187	
Linear Trend	0.247	0.025	0.068	0.573	0.555	0.003	0.517	0.043
Quadratic Trend	0.194	0.014	0.023	0.838	0.079	0.139	0.054	0.859
Breaking Trend	0.120	0.038	−0.041	0.664	0.060	0.164	0.035	0.870
Hodrick-Prescott	0.178	0.000	0.009	0.942	0.051	0.063	0.025	0.868
Band-Pass	0.172	0.003	0.004	0.974	0.013	0.254	−0.011	0.938
Beveridge-Nelson	0.174	0.000	0.005	0.202	0.025	0.086	0.000	0.953
SVAR	0.013	0.359	−0.133	0.263	0.199	0.012	0.170	0.405
Watson	0.331	0.001	0.140	0.154	0.277	0.009	0.247	0.197
Harvey-Clark	0.320	0.002	0.131	0.152	0.113	0.070	0.086	0.674
Harvey-Jaeger	0.140	0.001	−0.024	0.841	0.006	0.308	−0.018	0.864
Kuttner	0.317	0.024	0.128	0.278	0.411	0.020	0.377	0.042
Gerlach-Smets	0.432	0.001	0.226	0.048	0.154	0.028	0.126	0.519
Fixed Lags, Quasi-Final Gaps								
Watson	0.091	0.117	−0.065	0.422	0.311	0.010	0.280	0.024
Harvey-Clark	0.081	0.074	−0.074	0.326	0.032	0.267	0.007	0.931
Harvey-Jaeger	0.252	0.002	0.072	0.595	−0.474	1.000	−0.487	0.198
Kuttner	0.194	0.088	0.023	0.815	0.494	0.019	0.458	0.045
Gerlach-Smets	0.115	0.076	−0.045	0.404	0.010	0.418	−0.015	0.865
Variable Lags, Real-time Gaps								
Benchmark MSFE	1.010		0.689		0.191		0.196	
Linear Trend	0.225		−0.165		−0.357		−0.341	
Quadratic Trend	0.228		−0.163		−0.405		−0.390	
Breaking Trend	0.172		−0.201		−0.289		−0.272	
Hodrick-Prescott	0.508		0.028		−0.451		−0.438	
Band-Pass	0.301		−0.113		0.215		0.244	
Beveridge-Nelson	0.288		−0.122		−0.035		−0.011	
SVAR	−0.106		−0.391		−0.018		0.006	
Watson	0.209		−0.176		−0.144		−0.123	
Harvey-Clark	0.205		−0.179		−0.046		−0.023	
Harvey-Jaeger	0.445		−0.015		−0.480		−0.468	
Kuttner	0.205		−0.179		−0.177		−0.158	
Gerlach-Smets	0.153		−0.214		−0.081		−0.059	

Notes: The AR benchmark is a univariate autoregressive forecast of inflation; the TF benchmark forecasts from a linear regression on lagged inflation and real output growth. Mean squared forecast errors (MSFE) for the two benchmark models are shown multiplied by 1000. The remaining figures in the AR and TF columns denote the relative improvements in MSFE for the output gap models, measured as $(A - B)/B$ where A is the MSFE of the benchmark and B is that of the output gap model. The p-values for the AR benchmark are for the null that $B \geq A$, based on the statistic in equation (5). The p-values shown for the TF benchmark are for two-sided test of the null that $A = B$, based on the statistic in equation (6). See section 3.3 and Appendix B for further discussion of the construction and interpretation of the p-values. The forecast horizon is 4 quarters and forecast equation estimation starts in 1955Q1. Fixed lag lengths are (1,1) while varying lag lengths are reset every quarter using BIC.

Appendix A: The Construction of Real Time Output Gaps

The output gaps used in this study, as well as the data and programs used to create them, are freely available from the authors. The estimates examined here include all those examined in Orphanides and van Norden (2002) plus the Band-Pass, Beveridge-Nelson, Harvey-Jaeger and SVAR methods described below; this is identical to the list of models considered in Orphanides and van Norden (2003). The range of available estimates were updated so that the "final" data vintage now corresponds to 2003Q3 (i.e. data available as of mid-August 2003, so data series end in 2003Q2) rather than 2000Q1 as in these two earlier papers. Data for real output were taken from the Real Time Data Archive of the Federal Reserve Bank of Philadelphia in September 2003. Observations span the period from 1947Q1 to 2003Q2. Vintages for output run from Nov. 1965 to August 2003. All CPI data are from the 2003Q3 vintage. The SVAR method also uses data for 3-month US treasury bills. Data for this rate (secondary market) from January 1934 to August 2003 were obtained from the FRED database of the Federal Reserve Bank of St Louis.

All output gap models we consider decompose the logarithm of output into trend and cycle components. The linear trend (LT) and quadratic trend (QT) models are from OLS regressions with linear and quadratic deterministic trends. The breaking trend model is identical to the LT model until 1976Q4. Starting in 1977Q1, it allows for an estimated break in the trend at the end of 1973. The Hodrick-Prescott(HP) method is based on the filter proposed by Hodrick and Prescott (1997) with their recommended smoothing parameter of 1600 for quarterly data. The band-pass method (BP) is based on the Stock and Watson (1998) adaptation of the Baxter and King (1999) approach. Following Stock and Watson (1998), we use a filter 25 observations in width and pad the available observations with forecasts from an AR(4) model. The Beveridge-Nelson follows Beveridge and Nelson (1981) in modelling output as an ARIMA(p,1,Q) series. Based on results for the full sample, we use an ARIMA(1,1,2), with parameters re-estimated by maximum likelihood methods before

each recalculation of the trend.

We examine five unobserved component (UC) models, all of which are estimated by maximum likelihood. Three of the five are univariate models. The Watson (WT) model is based on Watson (1986) and models the output trend as a random walk with drift while the cycle is assumed to follow a stationary AR(2) process. The Harvey-Clark (CL) model follows Harvey (1985) and Clark (1987), replacing the constant drift in the trend of the WT model with a random walk. The Harvey-Jaeger (HJ) model has the same trend as the CL model but replaces the AR(2) component with a stochastic cycle. All three of these univariate models require estimation of five parameters, including variances for the assumed Gaussian shocks. The Kuttner (KT) model appends a Phillips curve, as specified in Kuttner (1994), to the WT model, giving a bivariate model with eight more estimated parameters than its univariate counterpart. The Gerlach-Smets (GS) model similarly adds the Phillips curve specified in Gerlach and Smets (1997) to the CL model, yielding a bivariate model with six more estimated parameters than its univariate counterpart.

The Structural VAR measure of the output gap (BQ) is based on a VAR identified via restrictions on the long-run effects of the structural shocks, as proposed by Blanchard and Quah (1989). Our implementation is identical to that of Cayen and van Norden (2002), who use a trivariate system including output, CPI and yields on 3-month treasury bills. Lag lengths for the VAR are selected using finite-sample corrected LR tests and a general-to-specific testing approach.

Appendix B: Evaluation of Forecast Performance

As noted in section 3.3, our statistical inference for the forecast performance of the output gap models relative to the AR benchmark model is based on the MSE-F statistic proposed by McCracken (2000). This takes the form

$$MSE\text{-}F = P \cdot \frac{(MSFE_1 - MSFE_2)}{MSFE_2} \tag{B.1}$$

where P is the number of forecasts, $MSFE_1$ is the mean squared forecast error (MSFE) of the restricted model and $MSFE_2$ is the MSFE of the unrestricted model.

The distribution of the MSE-F statistic under the null hypothesis of equal MSFE is estimated via a bootstrap experiment. The bootstrap begins by estimating a constrained VAR(12) in $\pi_t^1, y_t^{T,T-1}$ in which we impose the restriction that y does not Granger-cause π. 2000 simulated realizations of this DGP are created by simulating the estimated model with shocks randomly drawn with replacement from the estimated residuals. π_{t+h}^h is then constructed as the sum of h consecutive observations of π_1^1. For each simulation, the dynamic model is initialized with historical observations starting with $\pi_{k+h}^h, \pi_{k-i}^1, y_{k-i}^{T,T-1}$ for an independently drawn value of k. MSE-F statistics are then calculated for each simulated series and their empirical distribution is used to estimate p-values for the true data's MSE-F statistics. Because these distributions are non-pivotal, the distribution of the test statistics is bootstrapped anew for every different choice of (P, h, y, m, n). The p-values for every reported MSE-F are therefore based on independent bootstrap experiments.