

# THE CALIBRATION OF PROBABILISTIC ECONOMIC FORECASTS

John W. Galbraith  
Department of Economics  
McGill University  
855 Sherbrooke St. West  
Montreal, QC, Canada  
H3A 2T7

Simon van Norden  
Affaires Internationales  
HEC Montréal  
3000 Ch. de la Côte Ste. Catherine  
Montreal, QC, Canada  
H3T 2A7

## *Abstract*

A probabilistic forecast is the estimated probability with which a future event will satisfy a specified criterion. One interesting feature of such forecasts is their calibration, or the match between predicted probabilities and actual outcome probabilities. Calibration has been evaluated in the past by grouping probability forecasts into discrete categories. Here we show that we can do so without discrete groupings; the kernel estimators that we use produce efficiency gains and smooth estimated curves relating predicted and actual probabilities. We use such estimates to evaluate the empirical evidence on calibration error in a number of economic applications including recession and inflation prediction, using both forecasts made and stored in real time and pseudo-forecasts made using the data vintage available at the forecast date. We evaluate outcomes using both first-release outcome measures as well as later, thoroughly-revised data. We find strong evidence of incorrect calibration in professional forecasts of recessions and inflation. We also present evidence of asymmetries in the performance of inflation forecasts based on real-time output gaps.

Key words: calibration, probability forecast, real-time data

\* We thank conference and seminar participants at the Federal Reserve Bank of Philadelphia, CIRANO, Dalhousie University, and the Latin American meeting of the Econometric Society, and in particular Graham Elliott, for valuable comments. The *Fonds québécois de la recherche sur la société et la culture* (FQRSC), the Social Sciences and Humanities Research Council of Canada (SSHRC) and CIRANO (*Centre Interuniversitaire de recherche en analyse des organisations*) provided support for this research.

## 1. Introduction

A probabilistic forecast is the estimated probability with which a future event will satisfy a specified criterion, such as the probability that there will be precipitation tomorrow or the probability that a particular political party will form the next government. Probabilistic forecasts are produced for economic events such as recession, failure to meet inflation targets, stock market crashes and bond defaults.<sup>1</sup> In each of these cases the event is binary (an outcome occurs in a given time interval, or not) and the forecast is an estimated probability of the event.

Calibration measures the match between forecast probabilities and actual conditional probabilities of an event occurring. Calibration is of interest in many applications because of its immediate relevance to interpretation of the forecast probabilities. As well, the bias implied by calibration error can be estimated and removed.<sup>2</sup> While probabilistic forecasts of economic events are increasingly common, studies of their calibration are still relatively rare (Diebold and Rudebusch 1989 is an important early exception.)

This paper makes several contributions to the study of the calibration of probabilistic forecasts. First, we show how kernel regression estimators can be applied to estimation of calibration. This provides improved estimation efficiency and avoids the arbitrary cell groupings used in the previous literature. It also allows for smooth graphical representations of calibration. Second, we use these methods to provide new evidence on the behavior of probabilistic forecasts of U.S. recessions and inflation. We find strong new evidence of systematic forecast errors in both forecasts from professional forecasters and simple model-based forecasts. Third, we use both first-release and highly revised economic series to evaluate forecast performance. We find several cases where this distinction has an important impact on the properties of the forecasts that we examine.

Section 2 of the paper defines the measures of interest and procedures for estimation and statistical inference. Section 3 describes the data, vintages, and forecasts or forecasting models that are subject to evaluation, and results of the analyses of these data. Section 4 concludes. In the remainder of this section, we discuss the relationship of calibration to other characterizations of forecast adequacy and we note the role of “real-time data” issues in assessing forecast performance.

Many methods have been proposed for the evaluation of probabilistic forecasts, the most general of which evaluate the full predictive density. Such methods aim at determining whether the forecast density matches the true density of future variables of interest. Both densities typically condition on some observable variables, and some methods allow for the set of conditioning variables to be mis-specified. Important contributions to this literature include Diebold, Gunther and Tay (1998), Bai (2003), and

---

<sup>1</sup>Probabilistic forecasts on financial variables are commonly traded in markets as binary or digital options.

<sup>2</sup>For example, see Hamill et al. 2003 on the usefulness of this procedure in precipitation probability forecasts.

Corradi and Swanson (2005), among many others. Diebold, Gunther and Tay (1998) suggest the use of the probability integral transform to obtain a sequence which is  $U[0,1]$  under the null of correct specification of the predictive density; Bai (2003) also obtains a statistic which is  $U[0,1]$  under this null, using a Kolmogorov-type test. Corradi and Swanson (2006) propose another Kolmogorov-type test using the probability integral transform which allows for both parameter estimation error and dynamic misspecification; since these elements enter the limiting covariance, inference can be conducted by bootstrap.

Part of the importance of the above methods is their generality, since many other features of forecasts can be seen as special cases of the properties of the predictive density. When the variable of interest is defined to be a binary indicator variable, assessing calibration can be seen as a special case of assessing the full conditional predictive density of the binary variable. However, complete forecast densities are often not available even when the simpler probabilistic forecasts which we study are. For example, the Survey of Professional Forecasters does not include a predictive density for real output growth, but it does include the forecast probability of a decline in real output. Similarly, the Survey provides only a crude and inconsistent approximation to the forecast density for inflation. Nonetheless, the inflation responses are well-suited to studies of probabilistic forecasts. We examine both the SPF contraction and inflation forecasts in the applied work which follows, below.

Moreover, even when we have the complete forecast density (as in the case of the model-based inflation forecasts which we examine), we are often most interested in a particular feature of that density rather than a global assessment. For example, negative output growth is of particular interest in macroeconomics because of evidence of non-linear dynamics (such as that of Hamilton 1989 or Teräsvirta and Anderson 1992, among many others) and the redistributive impact of recessions. For monetary policy, there is broad agreement that high inflation is qualitatively different from low and stable inflation, while at the same time there are also concerns about excessively low inflation due to the implications of the zero lower bound on nominal interest rates for monetary policy. Macroeconomists may therefore be more interested in assessments that focus on the distinction between positive and negative growth, or particular threshold levels of inflation, than a overall assessment of the forecast density. This is the motivation for the applied work on output and inflation forecasting that we present below.<sup>3</sup>

---

<sup>3</sup>Although we do not analyze them there, situations where the object of interest is the forecast probability of exceeding a fixed threshold arise frequently in finance as well as in macroeconomics. Examples include models for pricing credit derivatives (which typically require forecasts of the probability of default), portfolio managers who need to forecast the probability that losses on a portfolio of derivatives become large enough to trigger a margin call, and regulators who need to forecast the probability that a financial institution's losses become large enough to render it insolvent and pose a systemic financial risk. One result has been the development of methods that allow for

Another important task in evaluating forecasts is to ensure that the forecasts are constructed realistically, using only information that would have been available to forecasters. Croushore and Stark (2003) highlighted the potential role of data revision in distorting our analysis of historical macroeconomic events, while Orphanides and van Norden (2002) emphasized the extent of revision in estimated output gaps. To avoid such problems, two of the three applications that we consider use data from the Survey of Professional Forecasters. The Survey records the forecasts of professional forecasters, thereby avoiding the revisions problem highlighted above. We also consider the impact of revisions to our outcome variables (real output growth and GDP deflator inflation) on our estimates of forecast calibration in both of these cases. Our third application uses the same real-time output gap estimates used in Orphanides and van Norden (2005). These avoid the revision problem by using only original vintage data series published by the Philadelphia FRB.

## 2. Calibration and probability forecast evaluation

### 2.1 Methods and definitions

The calibration of a forecast can be measured with no more information than a set of point forecasts and data on outcomes. Let  $x$  be a 0/1 binary variable representing an outcome and let  $\hat{p} \in [0, 1]$  be a probability forecast of that outcome. In well calibrated forecasts, we would have  $\hat{p}$  equal to  $E(x|\hat{p})$  throughout the range of  $\hat{p}$ , so that the probability statements made by the forecaster are correct statements about probabilities of the outcome. Here we will define

$$E_f(\hat{p} - E(x|\hat{p}))^2 \tag{2.1}$$

as the *mean squared calibration error*, where  $E_f(z) = \int z f(z) dz$  and where  $f(\cdot)$  is the marginal distribution of the forecasts. Calibration error measures the deviation of the forecast probability from the true probability of the event when a given forecast is made.<sup>4</sup> The mean squared calibration error has a minimum value of zero; its maximum value is 1, where forecasts and conditional expectations are perfectly opposed. If the calibration is known to be poor in some range, we may modify our interpretation of the forecasts by making an implicit or explicit bias adjustment.

If for any forecast value  $\hat{p}_i$  the true probability that the event will occur is also  $\hat{p}_i$ , then the forecasts are correctly calibrated. If for example we forecast that the probability of a recession beginning in the next quarter is 20%, and if over all occasions

---

a local evaluation of the forecast density function; see for example Linton and Whang 2007.

<sup>4</sup>This quantity is often called simply the ‘calibration’ or ‘reliability’ of the forecasts. We prefer the term *calibration error* to emphasize that this quantity measures deviations from the ideal forecast, and we will use ‘calibration’ to refer to the general property of conformity between predicted and true conditional probabilities.

on which we would make this forecast the proportion in which a recession will begin is 20%, and if this match holds for all other possible predicted probabilities, then the forecasts are correctly calibrated. If by contrast a recession will only occur 5% of the time when  $\hat{p} = 10\%$ , the calibration error will be positive. Note that correct calibration can be achieved by setting  $\hat{p} = E(x)$ , the unconditional probability of a recession, but such forecasts are said to have no *resolution*, i.e. no ability to distinguish high- and low-probability cases.<sup>5</sup>

Calibration has been measured in both meteorological and economic literatures in the evaluation of discretely-distributed probability forecasts which may only take on only values in a finite set (e.g. precipitation-probability forecasts may take on the values  $\{0, 0.2, 0.4, \dots, 1.\}$ ), or of continuous probability forecasts in  $[0, 1]$  which are subsequently grouped into discrete cells. In the next subsection we consider methods that will allow us to estimate these quantities for continuously-distributed probability forecasts without grouping into discrete cells.

## 2.2 Estimation of the calibration error

Estimation of the quantity in (2.1) requires estimation of the conditional expectation function  $E(x|\hat{p})$  (the unconditional probability  $E(x)$  is estimated by the sample mean  $\bar{x}$  of the binary outcome). When the forecasts take on only a number of discrete values, e.g.  $\hat{p}_i = \{0, 0.1, 0.2, \dots, 1.0\}$ , the conditional expectation in (1) is estimated by a simple sub-sample mean of  $x$  for each value of  $\hat{p}$ . When the forecasts can take any value in the interval  $[0, 1]$ , the forecasts may be grouped into cells and calibration may be investigated for each cell. This is the approach taken by, for example, Diebold and Rudebusch (1989), who divide the  $N$  forecasts into  $J$  cells; the authors then compute the local squared bias measure  $N^{-1} \sum_{j=1}^J n^j (\bar{p}^j - \bar{x}^j)^2$ , where  $\bar{p}^j$  and  $\bar{x}^j$  are the mean forecast probability and the mean outcome on the  $n^j$  values contained in cell  $j$ . These authors, and others such as Casillas-Olvera and Bessler (2006), also compute the ‘global squared bias,’  $2(N^{-1} \sum_{i=1}^N \hat{p}_i - N^{-1} \sum_{i=1}^N x_i)^2$ , a measure of the match of

---

<sup>5</sup>Although the present study is concerned with calibration alone, it is worth noting that the mean squared error of a forecast can be decomposed into calibration and resolution terms: following Murphy and Winkler (1987), for example, we can condition on the forecasts to decompose the mean squared error  $E((\hat{p} - x)^2)$  of the probabilistic forecast as follows

$$E(\hat{p} - x)^2 = E(x - E(x))^2 + E_f(\hat{p} - E(x|\hat{p}))^2 - E_f(E(x|\hat{p}) - E(x))^2.$$

Note that the first right-hand side term, the variance of the binary sequence of outcomes, is a fixed feature of the problem and does not depend on the forecasts. Hence all information in the MSE that depends on the forecasts is contained in the second and third terms on the right-hand side, the mean square calibration error and the resolution.

the unconditional mean probability forecast and unconditional mean probability of the outcome. Note that the local measure is analogous to the use of a histogram to estimate a continuous density, with the corresponding loss of efficiency.

We can estimate the continuous conditional expectation function without imposing linearity or artificial grouping into cells by nonparametric (e.g. kernel) regression of  $x$  on  $\hat{p}$ . The advantage of such a continuous function is that it allows us to evaluate calibration at any point in the continuous interval  $[0, 1]$ .

There are various possible choices of estimator, including the Nadaraya- Watson (locally- constant) and locally-linear kernel regression, nearest-neighbour methods, etc. The methods estimate the true conditional mean function,

$$f(X|\hat{P} = \hat{p}) = \int_{-\infty}^{\infty} x f(x, \hat{p}) / f_{\hat{P}}(\hat{p}) dx,$$

where  $f_{\hat{P}}(\hat{p})$  is the (continuous) marginal distribution of the probability forecasts, by substituting estimates  $\hat{f}(x, \hat{p})$  and  $\hat{f}_{\hat{P}}(\hat{p})$  of the densities. See Pagan and Ullah (1989) for a general review and exposition of these methods which notes the discrete/continuous distinction, and a discussion of relative efficiency of the kernel and histogram (discrete cell) estimators. We use the Nadaraya-Watson kernel estimator in all results recorded below. Any such method requires a choice of bandwidth parameter and kernel function; while cross-validation will be our primary method for bandwidth choice, we report results for various values of the bandwidth parameter to indicate sensitivity to this choice.

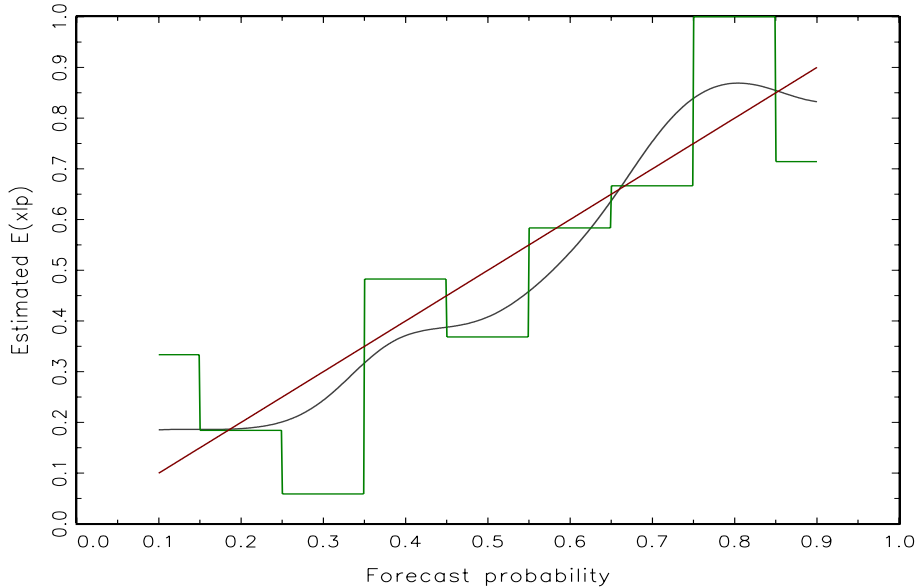
Moving from the histogram-type estimator to kernel regression produces efficiency gains. Figure 2.2.1 illustrates these gains in a single trial from a simple Monte Carlo experiment with a sample size of 200. The 45 degree line represents correct calibration, or  $E(x|\hat{p}) = \hat{p}$ , which is the true relationship in these pseudo-data; the smooth line is the kernel estimate of  $E(x|\hat{p})$ , and the step function is the cell-based estimate.

In 5000 repetitions of the simulation experiment illustrated in Figure 2.2.1, with the bandwidth chosen as in the empirical results below and with simulated forecasts centered at values between 0.2 and 0.7, the kernel estimates typically show a root integrated squared error less than two thirds that of the histogram estimates.<sup>6</sup>

---

<sup>6</sup>For example, for forecasts centered on 0.5 with a standard error of 0.2, the kernel root integrated squared error was 0.057, and the histogram 0.094; centered on 0.7 with standard error of 0.3, the corresponding values are 0.061 for kernel, 0.107 for histogram; centered on 0.2 with standard error of 0.2, the values are 0.066 and 0.119. Although simulated forecasts are differently dispersed throughout the unit interval in these different examples, each set of forecasts is correctly calibrated.

FIGURE 2.2.1  
 Illustration of kernel vs histogram estimates  
 of the conditional expectation function, N=200



### 2.3 Statistical inference on calibration

Inference on the estimated conditional expectation functions must take into account the dependence which exists in both forecasts and outcomes. Kernel regression estimates have been shown to remain consistent and asymptotically normal with dependent data under various conditions; see for example Robinson (1983, 1986).<sup>7</sup> Pointwise inference on calibration error (that is, a test of  $H_0 : \hat{p}_i = E(X|p_i)$  at any given probability  $\hat{p}_i$ ) can therefore be conducted using asymptotic confidence bands for the nonparametric regression functions. Nonetheless the sizes of sample available in macroeconomic applications yield wide confidence bands outside the central region where most observations lie, so there is little power to reject deviations from correct calibration.

We will instead consider a global test of the hypothesis of correct calibration. By a global test we mean a test of the null that the entire function  $E(X|\hat{p}_i)$  can be reduced to

---

<sup>7</sup>Nonparametric estimates are asymptotically normal at each point of estimation under standard assumptions which include smoothness of the conditional expectation function and a bandwidth parameter  $h$  that converges to zero with sample size  $n$ , having asymptotic variance  $f(x)^{(-1)}\sigma^2 \int K^2(\omega)d\omega$ ; see for example Pagan and Ullah (1999, section 3.4) in the iid case. The kernel constant  $\int K^2(\omega)d\omega$  is equal to 0.2821 for the Gaussian kernel used in our estimation.

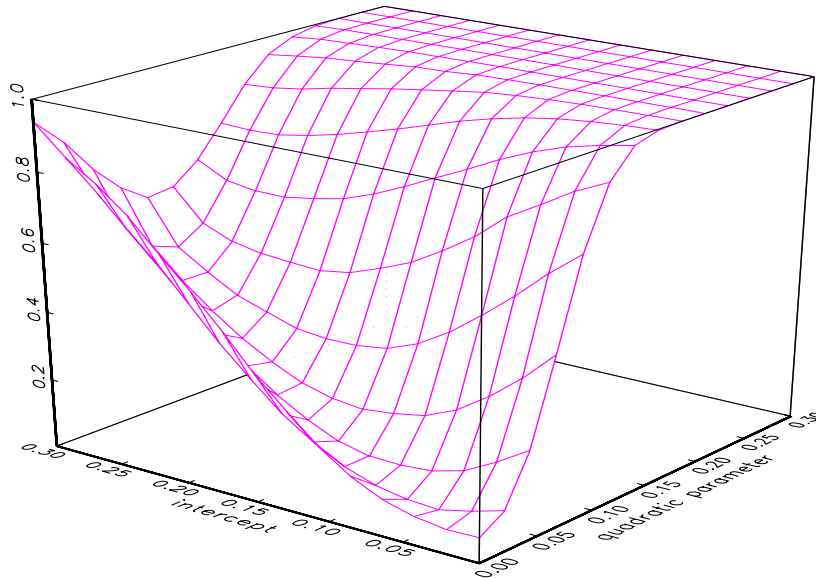
the linear form  $a + b\hat{p}_i$  with  $a = 0$ ,  $b = 1$ , as is implied by correct calibration *throughout* the interval  $[0, 1]$ . Note that this is a test purely of the property of correct calibration, not of overall forecast adequacy. It is therefore a supplement to, not an alternative to, general tests for properties such as correct specification of the forecast density.

Consider a test within a linear context which can be conducted by testing the hypotheses  $a = 0$  and  $b = 1$  jointly within a linear regression model. The model is therefore  $X = a + b\hat{p}_i + u_i$ , and  $H_0 : a = 0, b = 1$  is tested in this case with the Wald statistic  $g(\hat{\theta})'\hat{V}^{-1}g(\hat{\theta})$ , where  $\theta = (a, b)'$ ,  $g(\hat{\theta}) = (\hat{a}, \hat{b} - 1)'$  and  $\hat{V}$  is a consistent estimate of the parameter covariance matrix. Dependence in deviations from the model does imply inconsistency of the least-squares covariance matrix used in standard asymptotic tests, so a heteroskedasticity- and autocorrelation-consistent covariance matrix estimator must be used in a context such as this. We compute the Wald statistics below using the Newey-West estimator for the covariance matrix and test the joint hypothesis. This test may lose power against nonlinear deviations from the null, but is feasible under general conditions: given consistent estimates of  $\theta$  from LS regression and  $V$  from the HAC estimator, the Wald statistic has the standard asymptotic  $\chi_2^2$  distribution.

To illustrate the size conformity and power of the test, Figure 2.3.1 plots the results of a monte carlo experiment in which the true conditional expectation function  $E(X|\hat{p}_i)$  differs from  $\hat{p}_i$  via a quadratic deviation:  $E(X|\hat{p}_i) = \hat{p}_i - a\hat{p}_i^2 + b$ . The quadratic deviation parameter  $a$  and the intercept deviation  $b$  are chosen on a grid of points in the interval  $[0, 0.3]$ ; very similar results are obtained for the negatives of these values of  $a$  and  $b$ . Correct calibration holds for  $a = b = 0$ , so at these values, Figure 2.3.1 shows test size in a nominal 5% level test; at any other values of  $a$  and  $b$  it shows power of this test. The experiment recorded in the figure uses 25000 replications and, to exemplify conditions similar to those of the empirical cases below, uses sample size 250, bandwidth 0.08 and four lags in computation of the Newey-West standard errors.

We observe good but not perfect size conformity in the asymptotic test at this sample size—the rejection probability at  $a = b = 0$  is 0.07—and substantial power against deviations from correct calibration.

FIGURE 2.3.1  
 Test size and power with quadratic alternative  
 N=250, bandwidth = 0.08, 25000 replications



### 3. Data, forecasts and pseudo-forecasts

We now use the measures described in Section 2 to study three forecast data sets: the first contains recession probability forecasts while the others contain probabilistic forecasts of inflation measured by GDP deflator and CPI respectively. We investigate a variety of forecast horizons for each data set. For CPI inflation, we also evaluate a suite of forecasting models.

#### 3.1 Recession probability forecasts

The recession forecasts that we consider come from the Survey of Professional Forecasters (SPF). The survey, originally conducted by the American Statistical Association (ASA) and the National Bureau of Economic Research (NBER), began in 1968:Q4 and was taken over by the Philadelphia Fed in 1990:Q2.<sup>8</sup> Quarterly recession probability forecasts are among the many series (most of which are point forecasts) that have been recorded since the beginning of the survey. Each forecast is the median across forecasters of their estimated probabilities that the economy will experience negative real GDP growth in quarter  $t + h$ ,  $h = 0, 1, 2, 3, 4$ . Note that the definition of a ‘recession’ in the SPF is not the standard two-quarter definition, but a single quarter of contraction; note also that these are not cumulative probabilities of a recession at

---

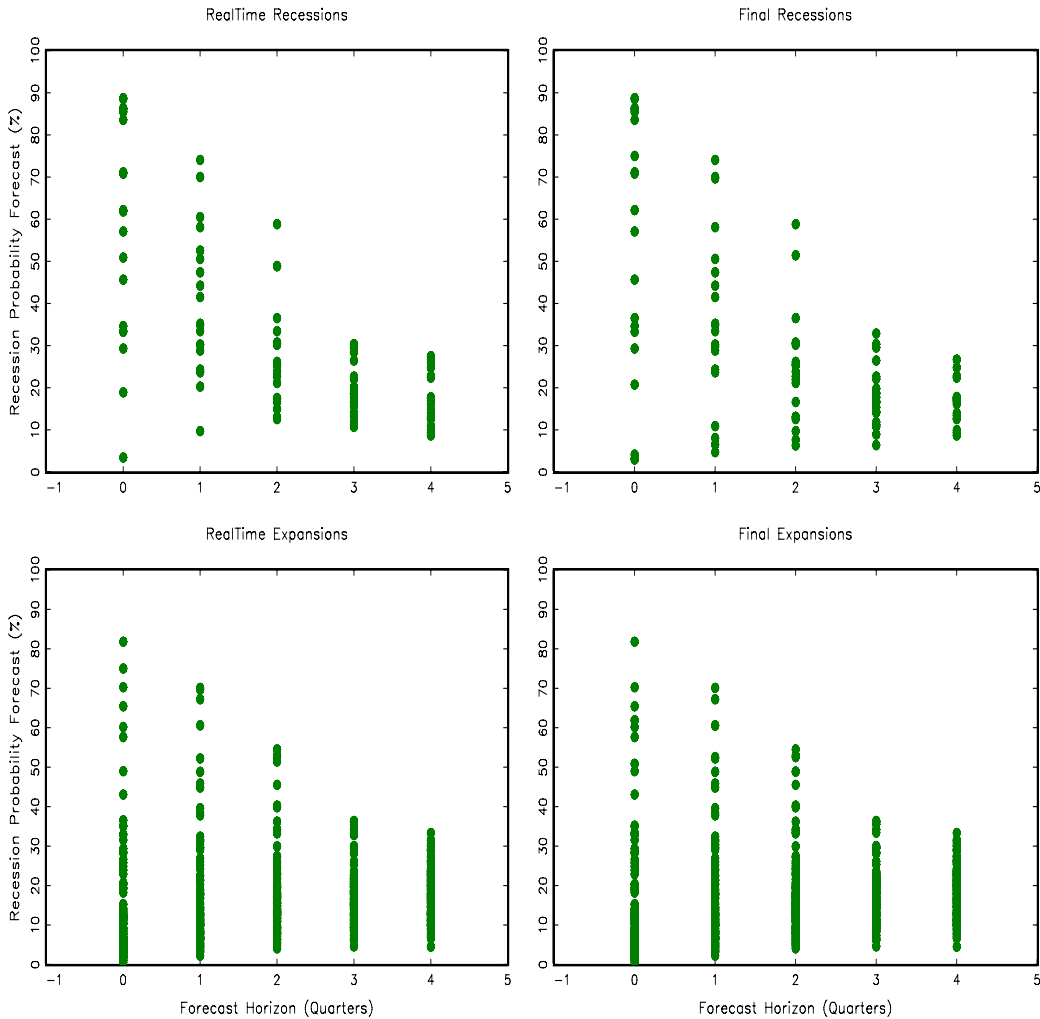
<sup>8</sup>Detailed documentation on the SPF survey is published on the FRB Philadelphia web site.

any point up to  $t + h$ , but are specific to quarter  $t + h$ . The first of these, for  $h = 0$ , is a roughly contemporaneous estimate of the current state (contraction or expansion) of the economy. We consider two measures of forecast outcomes, one based on the initial (i.e. first-release) estimate of real GDP growth and the other, which incorporates extensive data revisions, based on the 2006Q1 data vintage.

These series have also been examined by several other authors, most recently by Lahiri and Wang (2007) and Rudebusch and Williams (2007). The latter consider only the RMSFE rather than calibration. They found that while the SPF forecasts had a lower RMSFE than a naive benchmark model at all horizons, this difference was statistically significant at only the 2 and 3 quarter forecast horizons. They explain this result by noting that at shorter horizons, the SPF forecasts make a small number of large forecast errors by assigning large probabilities to a recession in the quarters just before or just after actual recessions (e.g. 1974-75, 1979-80 or 2001.) Lahiri and Wang (2007) examine calibration using a pointwise test based on 12 bins and find no evidence to reject the null of correct calibration for any forecast horizon. They also examine skill scores and other measures of forecaster ability; in contrast with Rudebusch and Williams (2007), they conclude that while the SPF appears to have important discriminatory power at shorter horizons, it appears to have little or no skill at longer horizons.

Figure 3.1.1 shows the distribution of the recession forecasts, with those forecasts for periods in which the economy subsequently contracted shown in the upper panels and those for periods in which it subsequently expanded shown in the lower panels. The revision in GDP growth figures causes the small difference between the outcomes measured with initial estimates (panels on the left) and with the recent vintage estimates (panels on the right.) Note that for shorter horizons, the forecast recession probabilities tend to be higher for contractionary periods than for expansionary ones. At longer horizons, however, no such difference is evident.

FIGURE 3.1.1  
Dispersion of probabilistic forecasts  
Survey of Professional Forecasters data



Next we compute the calibration measure for these forecasts. This requires us to choose bandwidths and kernel functions in order to estimate the continuous conditional expectation function  $E(x|\hat{p})$  needed to evaluate (2.2) and (2.3). The tables below report results from the standard Nadaraya-Watson kernel estimator with a Gaussian kernel function; our results are typically much less sensitive to kernel choice than to the choice of bandwidth parameter. Cross-validation estimates the optimal bandwidth to be close to 0.08 on most of these data sets, although results are more erratic on the longer-horizon forecast data. We therefore take the value 0.08 as a base case and also report results in which this value is varied by  $\pm 50\%$ ; Table 3.1.1 below shows that the choice

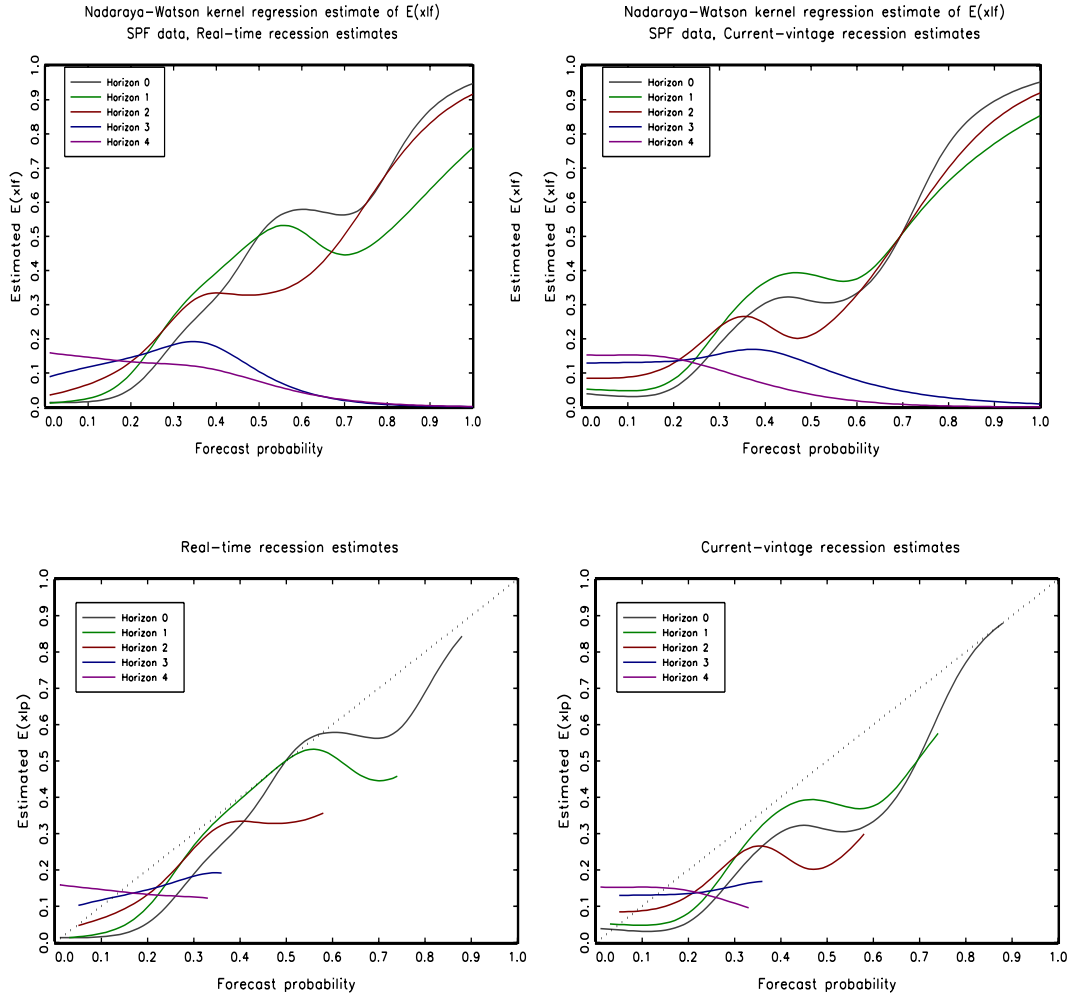
of bandwidth parameter has little effect on the RMS calibration error, nor does the use of first-release instead of recent-vintage data.

Figure 3.1.2 illustrates the estimated conditional expectations produced for the five forecast horizons and two data vintages, using the base bandwidth of 0.08. The ideal conditional expectation function would be a 45 degree line showing equality between the forecast and  $E(x|\hat{p})$ . The conditional expectation functions are only plotted over the observed range of probability forecasts; note in particular that there are forecasts near zero at all horizons whereas there are no observations of probability forecasts near one at the longer horizons. We see that at longer (3 and 4Q) horizons, there is no apparent relationship between the SPF forecast and the observed frequency of contractions; the curves in the graph are essentially horizontal at the unconditional probability of recessions. At shorter horizons we clearly see that the estimated probability of recession increases with the SPF forecast. There is some evidence that the SPF tends to overestimate the probability of recession both at low probabilities (a nominal 20 percent forecast is associated with a recession frequency of close to zero) and at the mid-range (e.g. for the 0Q forecast a forecast probability of 70 per cent is associated with an estimated recession frequency of 50 to 60 percent with initial-release data and approximately 50 percent in recent-vintage data.)

Table 3.1.1 reports the estimated root mean squared calibration errors for these SPF forecasts. The point estimates are all tightly clustered between 5 and 10 per cent, implying relatively minor overall calibration error at all horizons, regardless of the outcome measure or the bandwidth parameter used. The results of the general tests described in section 2.3 for the null of correct calibration against linear alternatives are reported in Figure 3.1.3 in the form of p-values for the null hypothesis of correct calibration.

Correct calibration is rejected at horizons 0 and 4, and at horizon 1 with real time outcomes. While evidence against correct calibration is strong at these points, the typical deviation from (0,1) slope and intercept is small: at horizon 0, the slope and intercept estimates are  $(-0.05, 0.99)$  for real-time outcomes and  $(-0.03, 0.87)$  for 2006-vintage outcomes. At horizons 3 and 4, the estimates of the linear form are imprecise, but there is nonetheless sufficient evidence at horizon 4 to reject correct calibration decisively. In general, we can conclude that while there is strong evidence of deviation from correct calibration at horizon 0 and somewhat less strong evidence at horizons 1 and 4, the magnitude of the deviation is small.

FIGURE 3.1.2  
Kernel-estimated conditional expectation of outcome given forecast  
Survey of Professional Forecasters data



### 3.2 SPF Inflation forecasts

In this and the following section we consider the calibration of inflation forecasts using two quite different types of probabilistic forecasts. The first are taken from the SPF. The survey asks forecasters to estimate the probability that inflation, as measured by the GDP deflator, will exceed various threshold levels.<sup>9</sup> While many of

<sup>9</sup>For this forecast, inflation is defined as the percentage change in the US GDP deflator over 4 consecutive quarters starting from its fourth quarter level in the current or

these thresholds varied over time as inflation varied, the 2% and 4% annual inflation threshold have been used in most surveys, thereby giving the longest consistent span of forecasts available for testing. Using all available forecasts for each forecast horizon gives us 22 to 23 observations for the 2% annual inflation threshold (using surveys from 1985Q2 to 2007Q4) and 26 to 38 for the 4% threshold (using surveys from 1968Q4 to 2007Q4.) Variation in the number of forecasts is largely due to the fact that forecasts for the 2% threshold and for the 5-7Q horizons were first recorded more than a decade after the earliest forecasts were recorded. Results for those horizons and thresholds therefore reflect a smaller sample that omits the earliest observations.

Croushore (2007) notes that revisions in inflation as measured by the GDP deflator has at times been substantial. For that reason, we again measure outcomes using both initial estimates and fixed-vintage (2008Q1) series.

Figure 3.2.1 describes the distribution of the SPF inflation forecasts. The layout of the figure is similar to that of Figure 3.1.1; each point shows a probability forecast of inflation exceeding the stated threshold, with the upper panels showing results for the 2% threshold and lower panels showing that for the 4% threshold. Within each panel, we distinguish forecasts for which measured inflation exceed the threshold (larger green dots) from those in which it did not (smaller blue dots.) Unlike the recession forecasts of the previous section, we see that these two conditional distributions look quite different at all forecast horizons for both thresholds. The measure of inflation outcomes (initial estimates in the panels on the left, recent vintage in the panels on the right) makes little difference in this respect.

The estimated calibration functions are shown in Figure 3.2.2. As was the case for Figure 3.2.1, panels on the left use the initial-release data while the upper panels show the results for the 2% threshold. In these cases we see important differences depending on whether first-release or current vintage data are used for evaluation. On first release data with 2% threshold, we see poor calibration at both low and high forecast probabilities, whereas with the most recent data forecasts appear well calibrated at most horizons; there are some apparently poor results visible at short horizons and intermediate probabilities, but relatively few actual forecasts at these points. Hence correct calibration is typically rejected (as results below describe) in initial release data, but not when the most recent data are used for evaluation. The results contrast less starkly at the 4% threshold, where correct calibration tends to be rejected in both cases.

---

previous year. This means that all the 0Q and 4Q forecasts are made in Q4, all the 1Q and 5Q forecasts are made in Q3, etc. The SPF reports the median response across its forecasters that inflation would fall in a given range. Our probabilistic forecast of inflation exceeding  $x\%$  is calculated as the sum of of the median probabilities assigned to all ranges with maximum values less than or equal to  $x\%$ .

FIGURE 3.2.1  
SPF Probabilistic Forecasts of Annual US GDP Inflation

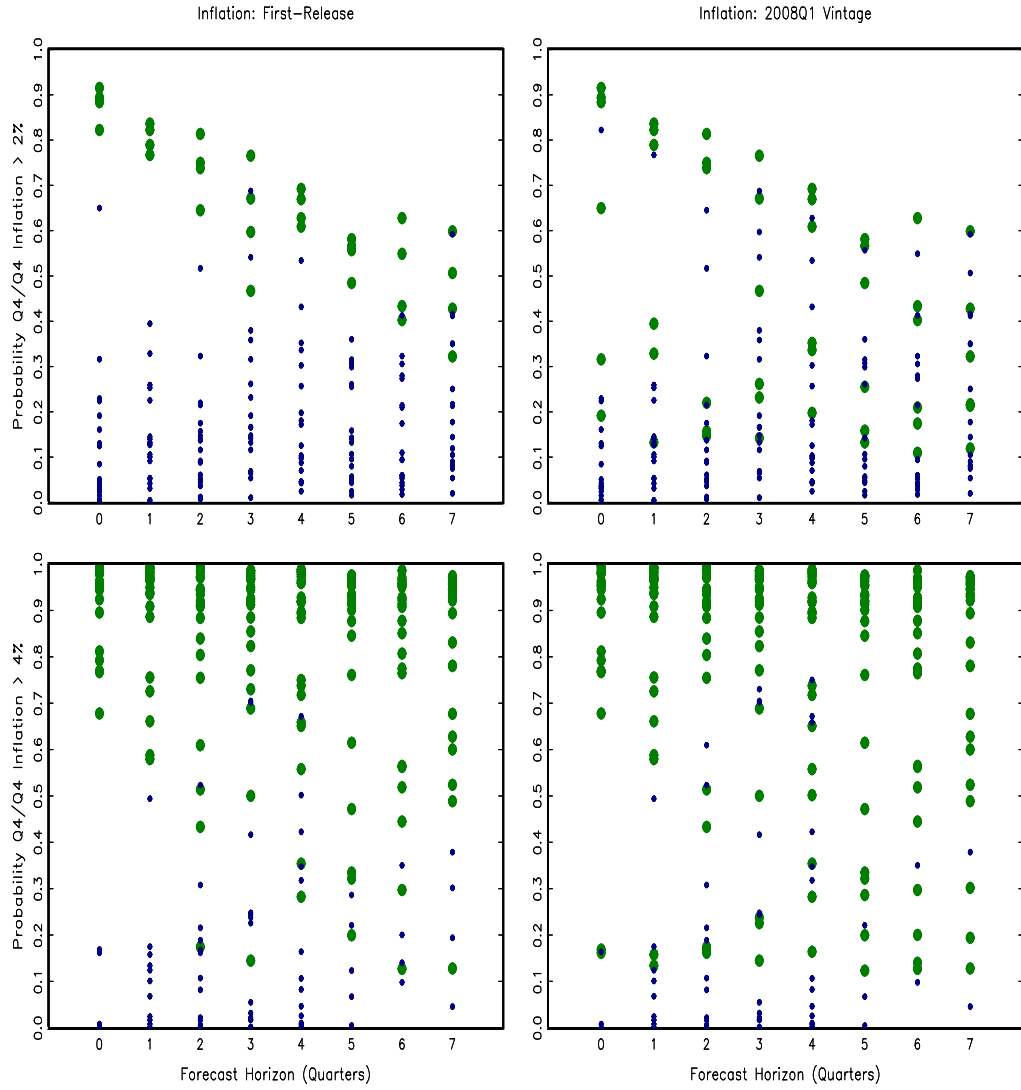
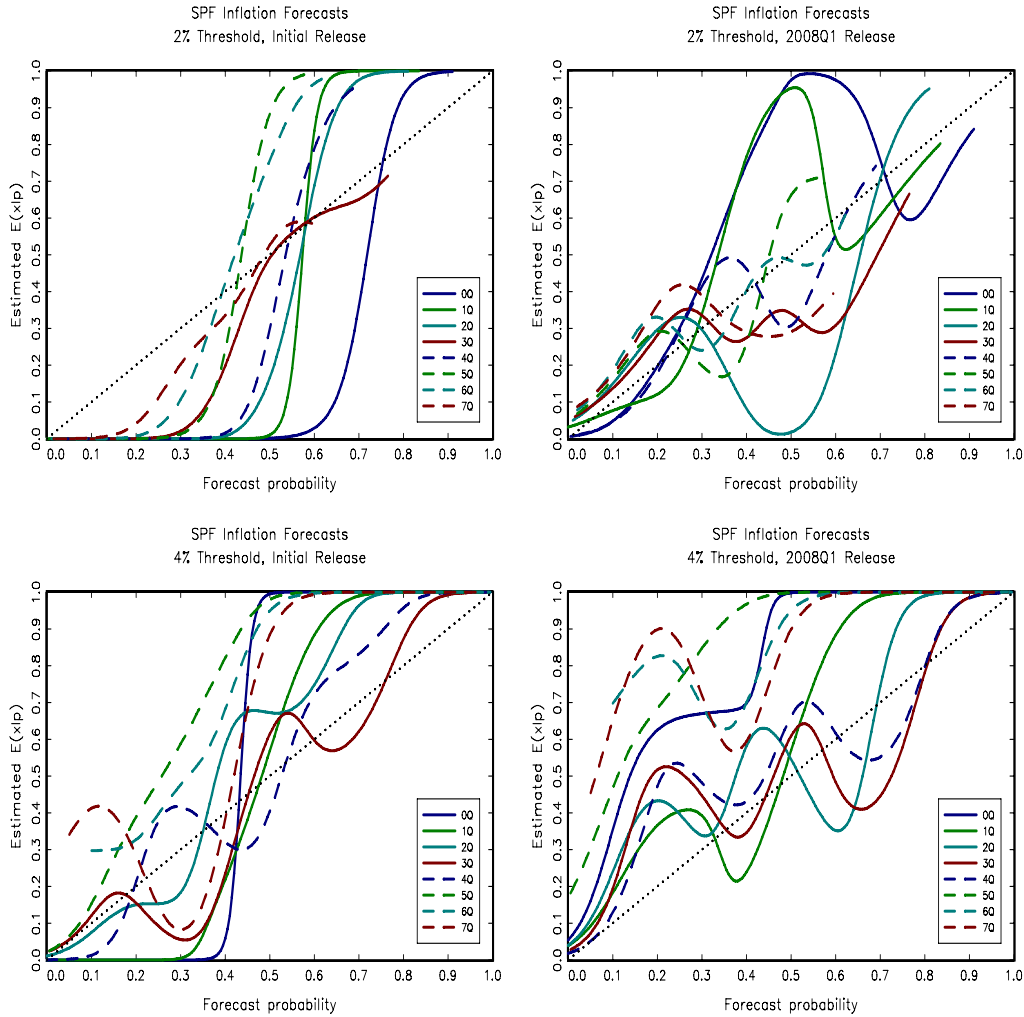


FIGURE 3.2.2  
Kernel-estimated conditional expectation of outcome given forecast  
Survey of Professional Forecasters data, two outcome measures



Numerical summaries of the behavior of the SPF inflation forecasts and their estimated calibration errors are in Table 3.2.1. The upper half of the table gives results for the 2% threshold while the lower half deals with the 4% threshold. Within each half, the upper portion uses initial estimates to measure inflation outcomes while the lower portion uses the 2008Q1 vintage. The first two lines of each portion give the variance of the inflation outcomes and the mean squared forecast error (MSFE) for comparison.<sup>10</sup> Outcomes do not vary with forecast horizon; the differences in outcome variance are

<sup>10</sup>Recall that inflation outcomes are 0 or 1, depending on whether inflation exceeded the threshold or not.

artefacts of the differing number of observations available at different horizons, as was discussed above.

Comparing the MSFE with the variance of inflation outcomes, we see that the forecasts capture most of the variance at the shortest horizons, but that this falls as the forecast horizon increases. The MSFE is always higher using the most recent data vintage rather than the initial release; this increase is often large relative to the total MSFE. With the 2% inflation threshold, this deterioration is usually much larger than the increase in the volatility of inflation outcomes.

The importance of the estimated calibration errors vary considerably, ranging from fully 100% of the MSFE to less than 5%. Calibration errors are smaller for the most recent data vintage for 7 of the 8 different forecast horizons examined, and sometimes much smaller. The tests reject the null hypothesis of correct calibration against linear alternatives at the 5% significance level in 7 of 8 forecast horizons when using first-released inflation estimated, whereas the null is never rejected using the most recent data vintage and the 2% threshold, and is rejected in only 5 of 8 cases with the 4% threshold.

Since the results for the 4% threshold at the longest horizons (5-7Q) essentially cover only the period since the Volker disinflation, we can gain some insight into how this affected the performance of the SPF inflation forecasts by comparing the properties of the longer and shorter horizon forecasts. From the lower half of Table 3.2.1, we see that the volatility of inflation outcomes dropped by roughly 50%, but with no corresponding decline in the MSFE. This suggests that the decline was largely due to a drop in predictable inflation, which would be consistent with a well-functioning inflation-targeting monetary policy. At the same time, however, calibration errors appear to be five to ten times larger during this later period, accounting for more than half of the MSFE when we use the recent data vintage to measure inflation. The presence of such apparently systematic errors is surprising; it suggests that forecasters experienced transitional difficulty in adjusting to a new inflation regime.

### *3.3 Model-based probabilistic forecasts of CPI inflation*

The other inflation forecasts that we examine are model-based pseudo-forecasts of CPI inflation: that is, in contrast with the SPF series evaluated above, these are forecasts computed now, but which could have been made using data that were available to forecasters at the time. Only historical data vintages were used in model selection, estimation and forecasting. For these U.S. inflation forecasts, we use the data, models and forecasting methodology described by Orphanides and van Norden (2005).<sup>11</sup> That study compared inflation pseudo-forecasts at various horizons from a set of fifteen simple linear models using only lagged inflation and real-time estimates of the output gap in their specification and construction.<sup>12</sup> The authors concluded that none of the

---

<sup>11</sup>CPI data are not revised, so only vintage series for real GDP from the real-time database of the FRB Philadelphia were used.

<sup>12</sup>The 15 models are Linear trend (LT) ; Quadratic trend (QT) ; Broken trend (BT);

forecasts using real-time output gap estimates seemed to perform better than models without such gaps (e.g. using output growth instead of the gap, or simply using autoregressive models of inflation) in the sense of having a consistently lower mean-squared forecast error (MSFE.) However, inflation-targeting central banks may be particularly interested in the probability that inflation stays below some upper bound that is considered consistent with explicit or implicit inflation targets (so as not to reduce the policy framework's credibility.) Alternatively, they may be particularly concerned that inflation not drop below some minimum in order to avoid problems associated with the zero bound on nominal interest rates. This suggests that evaluating the performance of probabilistic output-gap-based inflation forecasts may give additional information on whether and how such models can be of use to policy makers.

These models are used for forecasts of average US CPI inflation over 2-, 4-, 6- and 8-quarter horizons. Probabilistic forecasts are based on OLS estimates of linear forecasting equations with conventional standard error estimates and assumed Gaussian errors. While there has been no explicit inflation target in the US, we considered inflation thresholds of 2% and a 4% per annum to be of interest for policymakers for the reasons just stated. We examine quarterly forecast performance over the period 1969Q2 to 2002Q3 (the same as that used by Orphanides and van Norden 2005.) Recognizing that there have been long periods where inflation has stayed well above or well below the 4% threshold, we also examine forecasts of a positive *change* in inflation.

We find that results are often similar across the different forecasting models, but often very different depending on the threshold value used. We will therefore focus on how the results vary across the different thresholds and mention differences across the various models only briefly.

---

Hodrick-Prescott (HP); Band Pass (BP); Beveridge-Nelson (BN); SVAR-Blanchard-Quah (BQ); Watson (1986) (WT); Harvey-Clark (CL); Harvey-Jaeger (HJ); Kuttner (KT); Gerlach-Smets (GS); TOFU (TF); Nominal Output (YN); Autoregressive (AR). See Orphanides and van Norden (2005) for references and details on model specification and estimation.

FIGURE 3.3.1  
 Dispersion of probabilistic forecasts of US inflation target exceedance  
 Real-time pseudo-forecasts from fifteen models

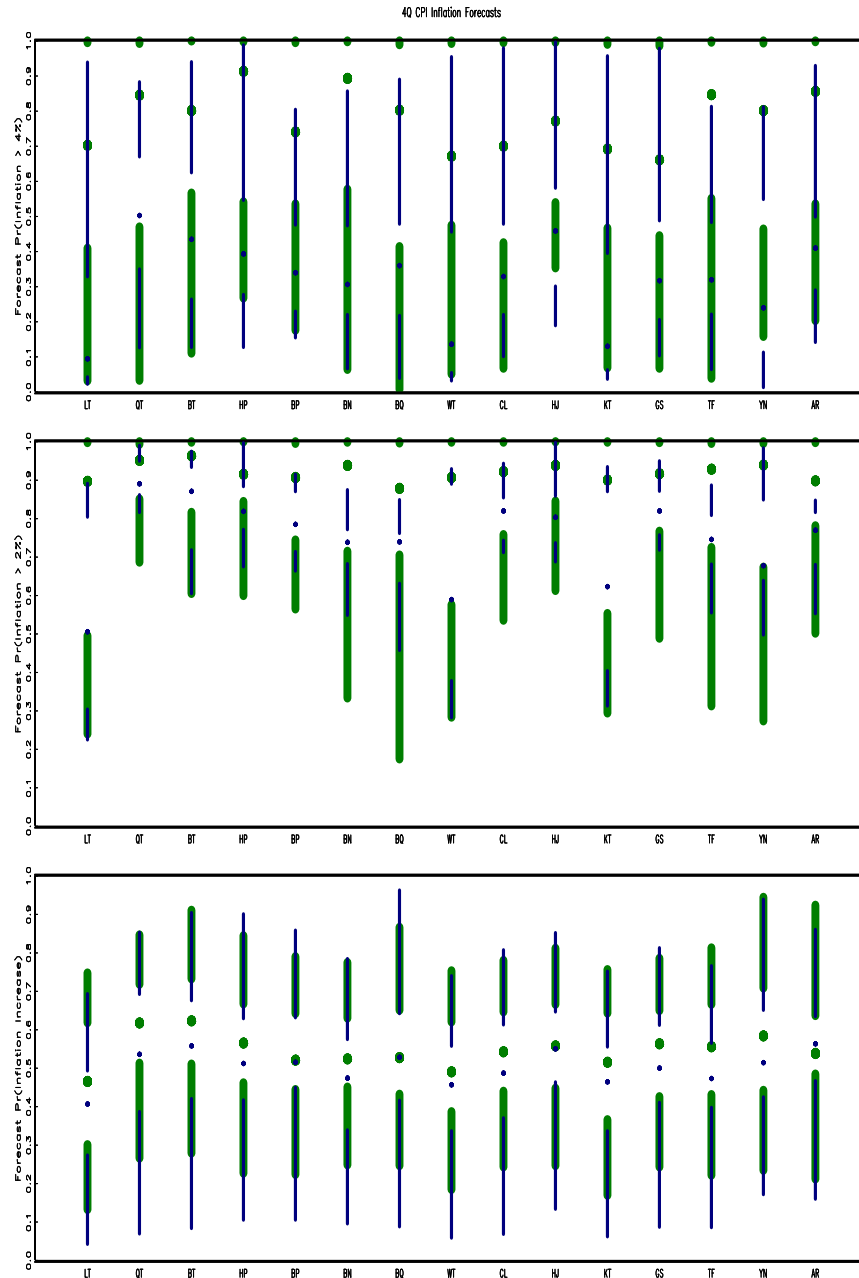


Figure 3.3.1 shows the distribution of these probabilistic forecasts. Results are similar for all forecast horizons; here we report only the results for the 4Q forecast horizon.<sup>13</sup> Because a larger number of forecasts is available, we summarize the distribution of the forecasts with a modified box plot rather than indicating each data point. The median of the distribution is indicated by a dot, which is flanked above and below by two line segments. The line segment endpoints closest to the median indicate the 25th and 75th percentiles, while their furthest endpoints indicate the 5th and 95th percentiles. (Note that in cases where the 75th and 95th percentiles are both equal to 1.0, the line segment appears as a dot.) Again, the broader green dots and lines correspond with outcomes where inflation exceeded the threshold and the narrower blue dots and lines correspond to cases where it did not.

The top panel shows the results for the 4% inflation threshold. We see that forecasts are widely distributed for every model, which contrasts to some degree with the 2% threshold case, as shown in the middle panel. In the latter case distributions are generally more compressed and the differences when inflation does vs. does not exceed the threshold are generally much smaller. There is also more variation in the results across the models; for example, the QT and BT models have forecast distributions that are very similar across the two outcomes, while the WT, KT and YN do not. The bottom panel shows the results for forecasts of the change in inflation; we see distributions that are very similar across models and across inflation outcomes.

Figures 3.3.2A and 3.3.2B show the estimated calibration functions corresponding to the above results. At the 2Q horizon and 4% threshold (Figure 3.3.2A, middle panel) we see considerable similarity in the calibration functions across models, most of which lie close to or somewhat below the diagonal (i.e. correct calibration) and are largely monotonic aside from a slight increase in the estimated frequency of threshold breaches at the very lowest forecast probabilities. With an inflation threshold of 2% (top panel), however, all models underestimate the probability of inflation breaching the 2% threshold. While slopes are generally positive for forecast probabilities above 60%, once below that conditional expectation functions are close to being flat. The lowest forecast probabilities are almost exclusively associated with breaches of the inflation threshold. This implies that such models generate alarms of very low inflation that are almost always false, even at very short horizons. For forecasts of the directional change in inflation (bottom panel), we see a tight clustering of the calibration function across models. The functions lie directly on the diagonal and show only modest calibration errors for the highest and lowest forecasts.

---

<sup>13</sup>Results for all horizons are available from the authors.

FIGURE 3.3.2A  
Kernel-estimated conditional expectation of outcome given forecast  
Two-quarter horizon, fifteen inflation forecasting models

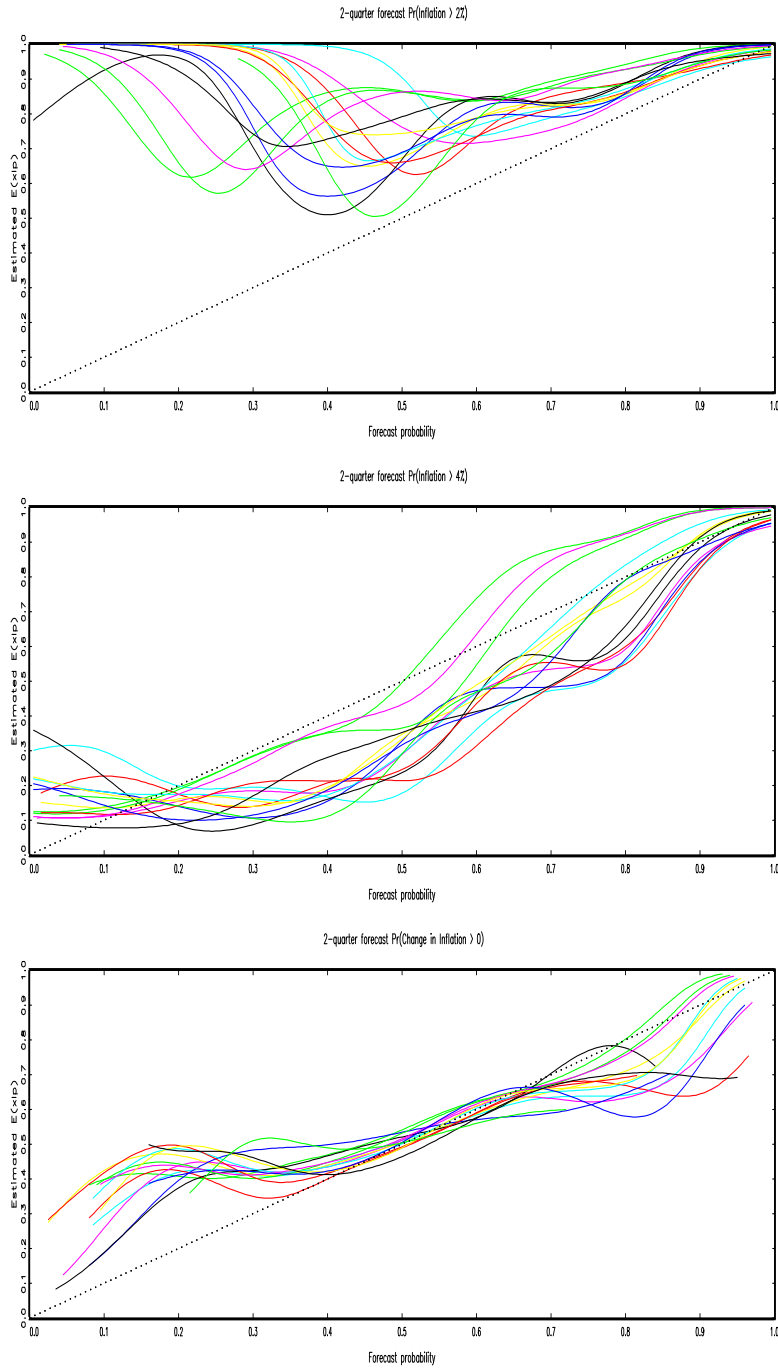
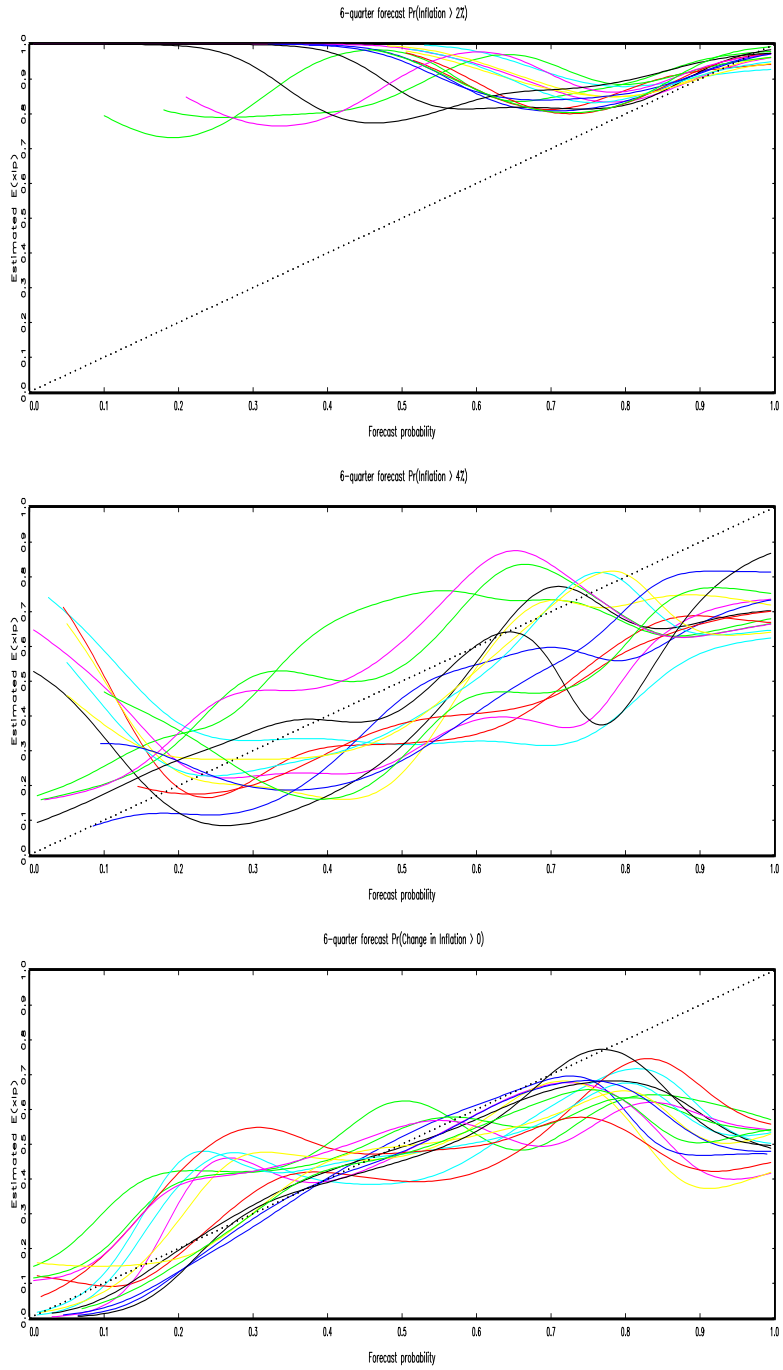


FIGURE 3.3.2B  
 Kernel-estimated conditional expectation of outcome given forecast  
 Six-quarter horizon, fifteen inflation forecasting models



As the forecast horizon increases (Figure 3.2.2.B), the estimated calibration errors become slightly more substantial. At the 2% threshold (top panel) they are increasingly tightly clustered across models. For the change in inflation (bottom panel) they remain very close to correct calibration, except for the very highest probability forecasts which tend to overstate the possibility of an increase in inflation. At the 4% threshold, we see increasing dispersion across models. While most show relatively good calibration for much of the range of forecasts, all models tend to overstate the risk of high inflation when they give very high probabilities of exceeding the threshold, and most tend to understate the risk when they give very low probabilities.

Tables 3.3.1A and B summarize forecast performance and present estimates and tests of each model's calibration. Results were generally quite similar for the 4, 6 and 8Q forecast horizons; for brevity we present the results for the 2Q and 6Q horizons.<sup>14</sup> For the short-term forecasts and the 4% threshold (Table 3.3.1A, upper panel), we see that most models have similar MSFEs and explain slightly less than 50% of the variability of inflation outcomes. Calibration errors appear to be small, never more than 0.04 and often around 10% of the MSFE. They are never significant at the 5% level. The 2% threshold (shown in the middle panel) produces quite different results, however; MFSEs are larger than the variance in inflation outcomes in every case (i.e. they perform worse than a constant forecast equal to the unconditional probability of exceeding the threshold.)<sup>15</sup> Calibration errors are much larger, typically above 0.025, and are significant at the 10% level in every case. The results for the forecast direction of the change in inflation (bottom panel) are different again; calibration errors are consistently small (never more than 0.01) and never significant at even the 10% level.

As shown in Table 3.3.1B, results are broadly similar at longer forecast horizons, although forecast performance is generally poorer. The deterioration is greatest for the 4% threshold, where few models explain any appreciable amount of the variation in inflation outcomes. Calibration errors tend to be larger, usually between 0.04 and 0.06, but are rarely important relative to MSFE and are generally insignificant. At the 2% threshold, MSFEs are larger still, while calibration errors are generally larger than before and even more significant. For changes in inflation, while MSFEs show only a modest decline in explanatory power, calibration errors are often substantially larger than before (0.03-0.05) and are now significant at the 5% level for 5 of the 15 models.

To summarize, there are very substantial differences both across these forecasting models in the quality of calibration, and across thresholds in the apparent difficulty of the problem of producing well calibrated forecasts. For the 4% threshold, most models produce results which are not significantly different from correct calibration, whereas at a 2% threshold, most do show significant deviations. This difference may provide a useful indicator of directions for development of these relatively simple forecasting

---

<sup>14</sup>Full results are available from the authors.

<sup>15</sup>This is suggestive of forecasts made beyond the content horizon; see Galbraith 2003, Galbraith and Tkacz 2006.

models.

#### 4. Concluding remarks

Estimation of the calibration of continuous probabilistic economic forecasts can be carried out without arbitrary grouping into intervals, using kernel regression estimates of the necessary conditional expectation function. In computing these estimates on a number of data sets, we find results that are quite stable across reasonable choices of smoothing parameter. The estimates provide insight into the performance and interpretation of forecasts and information that may be useful in improving these forecasts, either through direct adjustment to correct biases observed in past forecasts (again see Hamill et al. 2003, for example), or through the impetus that the analysis provides to revisit and respecify the forecasting model itself.

The three sets of probabilistic forecasts that we examined showed qualitatively different results. The Survey of Professional Forecasters recession probability forecasts show low calibration error at all horizons, although there are a number of statistically significant deviations from correct calibration; comparison of our SPF results with those of previous authors (where correct calibration is often not rejected) suggests that the global tests of correct calibration have relatively high power to detect deviations from correct calibration. The SPF forecasts of the probability of inflation exceeding a threshold, by contrast, show quite high calibration error; it is possible that these errors reflect forecasters' difficulty in learning about, or adjusting to, a new inflation regime over part of the sample period.

In the set of model-based forecasts of the probability of inflation exceeding a threshold, we also found widespread evidence of substantial calibration errors, although at some threshold values calibration error was generally low and often not significantly different from zero. Clear differences among the models were discernible, suggesting that examination of the calibration may be one useful diagnostic for builders of forecast models.

## References

- Ang, A., G. Bekaert and M. Wei (2007) "Do macro variables, asset markets, or surveys forecast inflation better?" *Journal of Monetary Economics* 54, 1163-1212.
- Azzalini, A. and A. Bowman (1993) "On the use of nonparametric regression for checking linear relationships." *Journal of the Royal Statistical Society Ser. B* 55, 549-527.
- Brier, G.W. (1950) "Verification of forecasts expressed in terms of probabilities." *Monthly Weather Review* 78, 1-3.
- Casillas-Olvera, G. and D.A. Bessler (2006) "Probability forecasting and central bank accountability." *Journal of Policy Modelling* 28, 223-234.
- Croushore, Dean and Tom Stark (2003) "A Real-Time Dataset for Macroeconomists: Does the Data Vintage Matter?" *Review of Economics and Statistics* 85(3), 605-617.
- Diebold, F.X. and G.D. Rudebusch (1989) "Scoring the leading indicators." *Journal of Business* 62, 369-391.
- Elliott, G. and R.P. Leili (2005) Predicting binary outcomes. Working paper, UCSD.
- Fan, Y. and Q. Li (1999) "Central limit theorem for degenerate u-statistics of absolutely regular processes with applications to model specification tests." *Journal of Nonparametric Statistics* 10, 245-271.
- Galbraith, J.W. (2003) "Content Horizons for Univariate Time Series Forecasts". *International Journal of Forecasting* 19, 43-55.
- Galbraith, J.W. and G. Tkacz (2006) "Forecast content and content horizons for some important macroeconomic time series." *Canadian Journal of Economics* 40, 935-953.
- Gneiting, T., F. Balabdaoui and A.E. Raftery (2007) "Probabilistic forecasts, calibration and sharpness." *Journal of the Royal Statistical Society Ser. B* 69, 243-268.
- Hamill, T.M., J.S. Whitaker and X. Wei (2003) "Ensemble reforecasting: improving medium-range forecast skill using retrospective forecasts." *Monthly Weather Review* 132, 1434-1447.
- Hamilton, J.D. (1989) "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle." *Econometrica* 57, 357-384.
- Härdle, W. and E. Mammen (1993) "Comparing nonparametric versus parametric regression fits." *Annals of Statistics* 21, 1925-1947.

- Lahiri, K. and J.G. Wang (2007) "Evaluating probability forecasts for GDP declines." Working paper, SUNY.
- Linton, O. and Y.-J. Whang (2007) "The quantilogram: With an application to evaluating directional predictability." *Journal of Econometrics* 141, 250-282.
- Murphy, A.H. and R.L. Winkler (1987) "A general framework for forecast verification." *Monthly Weather Review* 115, 1330-1338.
- Orphanides, A. and S. van Norden (2002) "The unreliability of output gap estimates in real time." *Review of Economics and Statistics* 84, 569-583.
- 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, 583-601.
- Pagan, A. and A. Ullah (1999) *Nonparametric Econometrics*. Cambridge University Press, Cambridge.
- Robinson, P.M. (1983) "Nonparametric estimators for time series." *Journal of Time Series Analysis* 4, 285-297.
- Robinson, P.M. (1986) "On the consistency and finite-sample properties of non-parametric kernel time series regression, autoregression and density estimators." *Annals of the Institute of Statistics and Mathematics* 38, 539-549.
- Rudebusch, G. and J.C. Williams (2007) "Forecasting recessions: the puzzle of the enduring power of the yield curve." Working paper, FRB San Francisco.
- Teräsvirta, T. and H.M. Anderson (1992) "Characterizing nonlinearities in business cycles using smooth transition autoregressive models." *Journal of Applied Econometrics* 7, S119-S136.
- Watson, M.W. (1986) "Univariate detrending methods with stochastic trends." *Journal of Monetary Economics* 18, 49-75.
- Yates, J.F. and S.P. Curley (1985) "Conditional distribution analyses of probabilistic forecasts." *Journal of Forecasting* 4, 61-73.

TABLE 3.1.1  
RMS calibration error and test  $p$ -values  
Recession probability forecasts, SPF data

---

	Horizon: (quarters)				
	0	1	2	3	4
Real-time outcome series					
$h=0.08$	0.08	0.08	0.07	0.06	0.09
$h=0.04$	0.08	0.09	0.08	0.06	0.09
$h=0.12$	0.09	0.08	0.07	0.07	0.08
test $p$ -value	0.002	0.014	0.277	0.502	0.021
2006 vintage outcome series					
$h=0.08$	0.10	0.08	0.09	0.07	0.09
$h=0.04$	0.10	0.09	0.10	0.07	0.10
$h=0.12$	0.10	0.09	0.09	0.08	0.08
test $p$ -value	0.048	0.147	0.134	0.092	0.001

---

TABLE 3.2.1  
RMS calibration error and test  $p$ -values  
SPF Inflation probability forecasts, 2% and 4% thresholds

Forecast Horizon (Q)	0	1	2	3	4	5	6	7
<b>2% Threshold</b>	<b>First Release Inflation Data</b>							
Outcome Variance	0.150	0.150	0.150	0.156	0.150	0.150	0.150	0.156
MSFE	0.035	0.032	0.040	0.093	0.066	0.065	0.074	0.109
Calibration	0.029	0.032	0.029	0.023	0.041	0.053	0.025	0.009
p-value	<b>0.002</b>	<b>0.000</b>	<b>0.010</b>	<b>0.001</b>	0.062	<b>0.021</b>	<b>0.049</b>	<b>0.000</b>
	<b>2008Q1 Inflation Data</b>							
Outcome Variance	0.202	0.202	0.202	0.208	0.202	0.202	0.202	0.208
MSFE	0.093	0.111	0.138	0.180	0.131	0.153	0.166	0.196
Calibration	0.007	0.008	0.017	0.015	0.005	0.009	0.005	0.016
p-value	0.997	0.972	0.930	0.493	0.423	0.869	0.837	0.538
<b>4% Threshold</b>	<b>First Release Inflation Data</b>							
Outcome Variance	0.190	0.248	0.245	0.248	0.239	0.157	0.135	0.135
MSFE	0.014	0.027	0.055	0.069	0.074	0.084	0.100	0.080
Calibration	0.014	0.014	0.009	0.005	0.007	0.033	0.043	0.042
p-value	<b>0.014</b>	<b>0.001</b>	<b>0.023</b>	0.508	<b>0.007</b>	<b>0.004</b>	<b>0.000</b>	<b>0.001</b>
	<b>2008Q1 Inflation Data</b>							
Outcome Variance	0.148	0.234	0.239	0.243	0.239	0.103	0.074	0.074
MSFE	0.060	0.065	0.096	0.108	0.113	0.128	0.150	0.119
Calibration	0.028	0.013	0.016	0.017	0.008	0.084	0.103	0.082
p-value	<b>0.001</b>	<b>0.035</b>	0.374	0.641	0.878	<b>0.000</b>	<b>0.000</b>	<b>0.000</b>

TABLE 3.3.1A  
RMS calibration error and test  $p$ -values, fifteen models, 2Q horizon

2Q CPI Inflation Forecasts															
Model	LT	QT	BT	HP	BP	BN	BQ	WT	CL	HJ	KT	GS	TF	YN	AR
Pr(Ave. Inflation > 4%)															
MSFE ( $\sigma_x^2 = 0.249$ )	0.120	0.191	0.171	0.160	0.142	0.150	0.157	0.125	0.145	0.164	0.119	0.149	0.144	0.133	0.146
Calibration	0.005	0.035	0.027	0.018	0.012	0.023	0.033	0.004	0.013	0.020	0.003	0.013	0.019	0.013	0.026
p-value	0.460	0.073	0.078	0.087	0.313	0.057	0.133	0.885	0.593	0.066	0.735	0.487	0.114	0.167	0.050
Pr(Ave. Inflation > 2%)															
MSFE ( $\sigma_x^2 = 0.082$ )	0.157	0.096	0.091	0.096	0.098	0.099	0.117	0.134	0.102	0.092	0.132	0.111	0.102	0.112	0.089
Calibration	0.088	0.023	0.019	0.026	0.029	0.033	0.053	0.066	0.033	0.021	0.062	0.038	0.035	0.038	0.020
p-value	<b>0.000</b>	<b>0.001</b>	0.087	<b>0.006</b>	<b>0.017</b>	<b>0.013</b>	<b>0.000</b>	<b>0.000</b>	<b>0.003</b>	<b>0.060</b>	<b>0.000</b>	<b>0.000</b>	<b>0.006</b>	<b>0.001</b>	0.087
Pr(Ave. Change in Q/Q Inflation > 0%)															
MSFE ( $\sigma_x^2 = 0.251$ )	0.243	0.240	0.236	0.239	0.241	0.246	0.235	0.241	0.243	0.241	0.240	0.241	0.240	0.242	0.246
Calibration	0.010	0.005	0.004	0.006	0.006	0.008	0.003	0.008	0.008	0.006	0.007	0.007	0.005	0.005	0.003
p-value	0.066	0.549	0.754	0.488	0.395	0.213	0.660	0.174	0.328	0.541	0.250	0.436	0.636	0.702	0.766

See notes at end of tables 3.3.1.

TABLE 3.3.1B  
RMS calibration error and test  $p$ -values, fifteen models, 6Q horizon

6Q CPI Inflation Forecasts															
Model	LT	QT	BT	HP	BP	BN	BQ	WT	CL	HJ	KT	GS	TF	YN	AR
	Pr(Ave. Inflation > 4%)														
MSFE ( $\sigma_x^2 = 0.251$ )	0.242	0.336	0.277	0.263	0.244	0.219	0.245	0.239	0.264	0.253	0.241	0.261	0.205	0.181	0.242
Calibration	0.048	0.105	0.060	0.057	0.051	0.033	0.057	0.040	0.048	0.042	0.041	0.045	0.028	0.011	0.044
p-value	0.092	<b>0.008</b>	0.058	0.112	0.195	0.152	0.201	0.116	0.108	0.132	0.117	0.125	0.256	0.401	0.157
	Pr(Ave. Inflation > 2%)														
MSFE ( $\sigma_x^2 = 0.251$ )	0.178	0.083	0.082	0.107	0.095	0.092	0.139	0.131	0.087	0.079	0.133	0.090	0.095	0.138	0.084
Calibration	0.105	0.006	0.008	0.033	0.019	0.020	0.070	0.058	0.012	0.007	0.059	0.015	0.022	0.065	0.013
p-value	<b>0.000</b>	0.074	0.168	<b>0.000</b>	<b>0.000</b>	<b>0.000</b>	<b>0.000</b>	<b>0.000</b>	<b>0.011</b>	0.067	<b>0.000</b>	<b>0.002</b>	<b>0.000</b>	<b>0.000</b>	<b>0.008</b>
	Pr(Ave. Change in Q/Q Inflation > 0%)														
MSFE ( $\sigma_x^2 = 0.251$ )	0.264	0.251	0.252	0.272	0.276	0.246	0.243	0.258	0.254	0.285	0.262	0.255	0.247	0.250	0.255
Calibration	0.032	0.035	0.034	0.052	0.051	0.030	0.025	0.028	0.030	0.048	0.031	0.030	0.031	0.033	0.030
p-value	<b>0.018</b>	0.113	0.118	0.056	<b>0.036</b>	0.332	0.227	<b>0.038</b>	0.078	<b>0.044</b>	<b>0.041</b>	0.097	0.254	0.218	0.172

See notes at end of tables 3.3.2.

### NOTES TO TABLES 3.3.1A AND 3.3.1B

These two tables report the results of forecasts for the indicator variable which equals one when average CPI inflation exceeds the indicated annual threshold rate over the indicated forecast horizon and zero otherwise. The values reported in the tables are:

- $\sigma_x^2$  is the sample variance of the indicator variable.
- *MSFE* is the mean squared forecast error
- *Calibration* is the estimated mean-squared calibration error.
- *p-value* gives the asymptotic marginal significance level of the test of the null hypothesis of perfect calibration against linear alternatives.

The forecasting models are those examined in Orphanides and van Norden (2005). Aside from the AR model (a univariate autoregressive model of inflation), all are recursively-estimated bivariate VAR models of CPI inflation and a real-time measure of the output gap. The models are:

- LT: Linear trend
- QT: Quadratic trend
- BT: Broken trend
- HP: Hodrick-Prescott
- BP: Band pass
- BN: Beveridge-Nelson
- BQ: SVAR-Blanchard-Quah
- WT: Watson (1986)
- CL: Harvey-Clark
- HJ: Harvey-Jaeger
- KT: Kuttner
- GS: Gerlach-Smets
- TF: TOFU
- YN: Nominal output
- AR: Autoregressive

See Orphanides and van Norden (2005) for references and details on model specification and estimation. Probabilistic forecasts are constructed under the assumption of gaussian forecast errors. Our forecast sample is the same as that used in Orphanides and van Norden (2005) and spans the period from 1969Q2 to 2002Q3.