Kernel-based Calibration Diagnostics for Recession and Inflation Probability Forecasts

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

Simon van Norden
Finance
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 occur. 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. Outcomes are evaluated using both first-release outcome measures and subsequent revised data. We find substantial evidence of incorrect calibration in professional forecasts of recessions and inflation from the SPF as well as in real-time inflation forecasts from a variety of output gap models.

Key words: calibration, kernel regression, probability forecast, real-time data, Survey of Professional Forecasters

JEL Classification number: C53

* 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 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 of a specific future event, such as the probability that there will be precipitation tomorrow, or that a particular party will form the next government. Probabilistic forecasts are produced for economic events such as recession, high inflation, stock market crashes and bond defaults.\(^1\) 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 occurrence.

Calibration measures the match between forecast probabilities and actual conditional probabilities of the event. Calibration is of interest in many applications because of its immediate relevance to interpretation of the forecast probabilities. 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 two contributions to the study of the calibration of probabilistic forecasts. The first is methodological: we show how kernel regression estimators can be used to estimate calibration. This avoids the arbitrary cell groupings used in the previous literature and allows for smooth graphical representations of calibration. The second is empirical; we use these methods to provide new evidence on the behavior of probabilistic forecasts of U.S. recessions and inflation, including those from the U.S. Survey of Professional Forecasters (SPF). We find evidence of systematic calibration errors both in forecasts from professional forecasters and in forecasts based on a variety of real-time output gaps. We also use both first-release and highly revised economic series to evaluate forecast performance; there are 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.

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; Brier (1950) provides a classic treatment. Let \(X\) be a 0/1 binary variable representing an outcome and let \(\hat{p} \in [0,1]\) be a forecast of the probability that \(X = 1\). A well calibrated forecast should have \(\hat{p} = E(X|\hat{p}) \forall \hat{p}\), so that the forecasts correctly reflect the conditional probability of the outcome. Following notation of Murphy and Winkler (1987), we define

\[
E_f(\hat{p} - E(X|\hat{p}))^2
\]

as the mean squared calibration error, where \(E_f(z) = \int zf(z)dz\) and where \(f(.)\) is the marginal distribution of the forecasts. The mean squared calibration error has a minimum value of zero (perfect calibration); its maximum value is 1.

\(^1\)Probabilistic forecasts on financial variables are commonly traded as binary or digital options.
If forecasts are correctly calibrated then for any forecast value $\hat{p}_i$ the true probability that the event will occur is also $\hat{p}_i$. If for example we forecast that the probability of a recession beginning in the next quarter is 20%, and if over all occasions 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.\(^2\)

Correct calibration is of interest in part because it implies that the probability forecasts are true statements: by contrast, if calibration is incorrect, the stated probability that an event will arise is not the true probability.\(^3\) As well, in detecting calibration errors and attempting to correct them, a forecaster may reduce the overall loss of a forecast (that is, calibration error is one source of forecast loss). Quite apart from using calibration information to adjust forecasts or to refine forecast methods, knowledge of the nature of deviations from correct calibration may be useful in interpreting results. For example, the function $E(X|\hat{p})$ may be approximately flat in the interval $\hat{p} \in [0.4, 0.6]$. In this case users may learn not to interpret differences between forecast probabilities in the range from 0.4 to 0.6 as meaningful, although deviations above or below this range may genuinely convey higher or lower conditional probability of the outcome.

The data we examine below will be the means of survey responses when individuals are asked to estimate the probability of a future event. If all individuals have correctly calibrated forecasts, the mean forecast will also be correctly calibrated. Conversely, it

\(^2\)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 (see also Gneiting et al. 2007), 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.

\(^3\)Of course, a statement may be true, but of limited usefulness; similarly, a probability forecast may be well calibrated but have little ability to distinguish different circumstances, as in the case just noted where all forecast probabilities are simply given as the unconditional mean.
is possible for the mean forecast to be correctly calibrated without perfect calibration of the individual forecasts provided that the individual biases are offsetting.

2.2 Evaluating predictive densities and calibration

The calibration error of probabilistic forecasts can be understood in the more general context of predictive density tests. The latter consider whether the forecast density matches the true density of a continuously or discretely valued variable of interest. For example, let \( \hat{f}_X(x) \) be a density forecast for \( X \) and let \( \hat{F}_X(x) \) be the corresponding forecast cdf. If the forecast distribution matches the true distribution, then (unconditionally) \( Pr(X \leq x) = F_X(x) = \hat{F}_X(x) \); the densities may also condition on some observable variables. In the special case where \( X \) is a binary variable, \( Pr(X = 1|\hat{p}) = E(X|\hat{p}) \), which will equal \( \hat{f}_{X|\hat{p}}(1) = \hat{p} \) if the true and forecast conditional distributions are the same.

Important contributions to this literature include those of Diebold, Gunther and Tay (1998), Bai (2003), and Corradi and Swanson (2006), 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.

Note that complete forecast densities are often not available even when the probabilistic forecasts that we study are. Even when we have the complete forecast density (as in the case of the model-based inflation forecasts considered in Section 3.3), 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 redistributional 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 on particular threshold levels of inflation, than in an overall assessment of the forecast density.\(^4\)

\(^4\)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
Calibration has been measured in both meteorological and economic literatures in the evaluation of discretely-distributed probability forecasts which may only take on a finite set of values (e.g. precipitation-probability forecasts which may take on the values 0, 0.2, 0.4,...,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.

2.3 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 $E(X)$ can of course be estimated by the sample mean $\bar{X}$ of the binary outcome). When the forecasts take on only a number of discrete values, e.g. $\hat{p}_i = \{0, 0.1, 0.2, \ldots, 1.0\}$, 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 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.

It is possible to estimate a continuous conditional expectation function without imposing linearity or artificial grouping into cells by nonparametric (e.g. kernel) regression of $X$ on $\hat{p}$. Kernel estimates are well known to have a number of advantages over histogram-type estimates. One is the fact that kernel methods produce smooth functions; here, this allows us to evaluate calibration at any point in the continuous interval $[0, 1]$. By contrast, histograms produce discontinuous estimates with a constant value within each bin. As well, kernel estimates can be shown to perform better on standard criteria such as rates of decline of the mean squared error with sample size; see for example Tapia and Thompson (1978, pp. 44-59), or see Silverman (1986) on advantages of the kernel over the histogram in general. Pagan and Ullah (1999) provide a recent general review and exposition of these methods and a discussion of relative

---

5 Note that a histogram, which uses fixed bins, is not equivalent to the use of a uniform kernel in kernel estimation, since the latter evaluates a function on an arbitrarily fine grid of points which defines a shifting set of centres for the weight function.
efficiency of the kernel and histogram (discrete cell) estimators.

There are various possible choices of estimator of the conditional expectation function, including the Nadaraya- Watson (locally- constant) and locally-linear kernel regression, nearest-neighbour methods, etc. Note that although \(X\) is binary, the conditional expectation \(E(X|\hat{p})\) given the continuous variable \(\hat{p}\) is a continuous function; because \(X\) takes only the values 0 and 1, this conditional expectation has an interpretation as the probability that the event will occur given the value of \(\hat{p}\). 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.

Given an estimate \(\hat{E}(X|\hat{p}_i)\) of the conditional expectation function \(E(X|\hat{p}_i)\) at a set of \(N\) points \(\{\hat{p}_i\}\), we compute an estimate of the quantity in (2.1) as \(N^{-1}\sum_{i=1}^{N} (\hat{p}_i - E(X|\hat{p}_i))^2\). The tables below report this measure, i.e. the mean squared calibration error.

2.4 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). Pointwise inference on calibration error (that is, a test of \(H_0: \hat{p}_i = E(X|\hat{p}_i)\) 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 the linear form \(a + b\hat{p}_i\) with \(a = 0\), \(b = 1\), throughout the interval \([0,1]\), as is implied by correct calibration.\(^7\) Note that, with histograms as well as with these kernel estimates, it is also possible to test calibration at particular points, e.g. \(H_0: E(X|\hat{p} = 0.4) = 0.4\). Pointwise tests can be based on asymptotic pointwise confidence intervals in either case.

\(^6\)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.

\(^7\)Of course \(a = 0\), \(b = 1\), is only a necessary and not a sufficient condition for correct calibration.
However, the relatively small sample sizes available here imply relatively low power in such tests, and so we compute only global tests.

To capture some nonlinearity in the estimated relation we use a quadratic specification of the conditional expectation function. The model is therefore \( x_i = a + b\hat{p}_i + c\hat{p}_i^2 + u_i \), and \( H_0 : a = 0, b = 1, c = 0 \) is tested in this case with the Wald statistic \( g(\hat{\theta})'\hat{V}^{-1}g(\hat{\theta}) \), where \( \theta = (a, b, c)' \), \( g(\hat{\theta}) = (\hat{a}, \hat{b} - 1, \hat{c})' \) and \( \hat{V} \) is a consistent estimate of the parameter covariance matrix. Unlike the numerical computations of calibration error, these tests do not use the kernel estimates of the conditional expectation functions, so that the statistical inference here does not depend on the smoothing.

Dependence in deviations from the model implies 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. Given consistent estimates of \( \theta \) from LS regression and \( V \) from the HAC estimator, the Wald statistic has the standard asymptotic \( \chi^2 \) distribution.

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 the GDP deflator and the 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 Federal Reserve Bank of Philadelphia in 1990:Q2.\(^8\) Quarterly recession probability forecasts are among the many series that have been recorded since the beginning of the survey. Each value is the mean across forecasters of their probabilities estimated at time \( t \) 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 any point up to \( t+h \), but are specific to quarter \( t+h \). 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 forecasts have also been examined by several other authors, most recently by Lahiri and Wang (2006, 2007) and Rudebusch and Williams (2007). The latter find that while the SPF forecasts have a lower root mean squared forecast error (RMSFE)

\(^8\)Detailed documentation on the SPF survey is published on the FRB Philadelphia web site.
than a naive benchmark model at all horizons, this difference is 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 as larger green circles and those for periods in which it subsequently expanded shown as black ‘+’ signs. The revision in GDP growth figures causes the small difference between the outcomes measured with initial estimates (upper panel) and with the recent vintage estimates (lower panel.) 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.
Next we choose bandwidths and kernel functions in order to estimate the continuous conditional expectation function $E(X|\hat{p})$ and evaluate the calibration. The tables below report results from the standard Nadaraya-Watson kernel estimator with a Gaussian kernel function; our results are (as is typical) 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 ±50%; Table 3.1.1 below shows that the choice of bandwidth parameter has little effect on the mean squared 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 equating the forecast \( \hat{p} \) and \( E(X|\hat{p}) \).\(^9\) We see that at longer (three and four quarter) 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 indication 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. at 0Q a forecast probability of 70 percent is associated with an estimated recession frequency of 50 to 60 percent with initial-release data and approximately 50 percent in recent-vintage data.) However, the lines are generally clustered not far from the 45\(^\circ\) line, especially for real-time estimates. It is important to note, both here and in subsequent graphics, that sampling error will account for some fluctuation of these lines; hence the need for statistical inference on deviations from the 45\(^\circ\) line.

The visible overestimation of recession probabilities for forecasts in the 60-70% range in the horizon 0 and 1 forecasts is associated with business cycle turning points. For example, based on initial releases, there were only six quarters in which the current-quarter probability of a contraction was over 50% and a contraction did not occur; 1979Q3, 1979Q4, 1980Q3, 1982Q2, 1991Q2 and 2001Q4. All corresponded to periods of limited economic growth (0 to 3% real growth at annual rates) following a contraction.\(^10\) Note that deviations from correct calibration are significant at conventional levels at these horizons, with the exception of horizon 1 in current vintage data.

\(^9\)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 probability forecasts near one at the longer horizons.

\(^10\)Real output growth in 1979Q1 was initially estimated at a quarterly rate of -0.82% but was eventually revised upwards to 0.10%.
Overestimation of recession probabilities at long horizons appears to be due to other factors, however. There were 21 quarters in which the 4-quarter ahead probability of recession exceeded 25%; this set includes the forecasts made in the aftermath of the 1987 stock market crash (all quarters from 1988Q2 to 1990Q2 inclusive except for 1989Q4) and the second OPEC oil shock (all quarters from 1979Q2 to 1980Q2.) Only one of these 13 quarters saw a contraction in real output according to recent-vintage data.

Table 3.1.1 reports the estimated 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 quadratic alternatives are also reported in Table 3.1.1 in the form of p-values.

Correct calibration is rejected at the 5% level at all horizons except 3 (which however has a p-value close to 5%) using first-release data, and at horizon 4 using current-vintage data (as well, horizon 0 is close to 5%).
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 again 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. While many of 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 (2008) notes that revisions in inflation, as measured by national accounts deflators, have 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, lower panels showing that for the 4% threshold, panels on the left measuring outcomes with initial estimates and panels on the right measuring outcomes with the 2008Q1 vintage. Within each panel, we distinguish forecasts for which measured inflation exceed the threshold (green circles) from those in which it did not (black ‘+’). 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.

\[\text{For this forecast, inflation is defined as the percentage change in the US GDP deflator over 4 consecutive quarters starting from its fourth quarter level in the current or previous year. This means that all the 0Q and 4Q forecasts are made in Q4, all the 1Q and 5Q forecasts are made in Q3, etc. The SPF reports the mean response across its forecasters that inflation would fall in a given range. Our probabilistic forecast of inflation not exceeding } a\% \text{ is calculated as the sum of the mean probabilities assigned to all ranges with maximum values less than or equal to } a\%.\]
Figure 3.2.1
SPF Probabilistic Forecasts of Annual US GDP Inflation
The estimated calibration functions are shown in Figure 3.2.2, where the panels are arranged as for Figure 3.2.1. There are several important differences between the various panels; we begin by considering the top-left panel which displays results for a 2% inflation threshold using real-time data. It is important to note that the real-time data at short horizons have perfect separation; the four years (1998, 1999, 2002 and 2003) where inflation was below 2% were also the four with the highest predicted probabilities of such an event. In Figure 3.2.2, the result is a calibration function that is effectively a vertical line. All forecasts to the left of the line resulted in inflation above the threshold and all forecasts to the right of the line resulted in inflation below the threshold.

Data revision makes these lines more complex in the upper right-hand panel. After revision, three more years (1995-97) also had inflation below 2% and 2003 did not. (All had inflation very close to the 2% threshold.) The large mis-calibration shown in the upper right panel for short-horizon forecasts in the neighbourhood of 50% should be interpreted with care. As shown in Figure 3.2.1, there are almost no observations around 50%, so the calibration function is particularly poorly estimated in this zone. This explains why we are unable to reject the null hypothesis of correct calibration below in Table 3.2.1 for this combination of inflation threshold and data vintage; in general, the statistical inference in the tables allows us to judge whether observed deviations are meaningful.

The differences between the real-time and current vintage results for the 4% threshold (lower left and right panels of Figure 3.2.2) are more muted as outcomes in only 2 of 40 years are reclassified (versus 4 of 24 for the 2% threshold.) Again, the real-time data at short horizons has perfect separation with all forecasts below 50% resulting in inflation above the threshold and all forecasts above 50% resulting in inflation below the threshold. Again, this results in an effectively vertical curve for the 0 and 1Q ahead forecasts, and the function is very imprecisely estimated around the cutoff probability, reflecting the fact that there were no forecasts between 17% and 57%. (See Figure 3.2.1.)

The results for the 4% threshold are also much less similar across forecast horizons than the results for the 2% threshold. The latter has forecasts available for each and every forecast horizon from 1985 onwards. For the 4% threshold, results start in 1968 or 1969 for the 1Q to 4Q horizons (with very few missing observations) whereas results for the other forecast horizons start between 1980 and 1982. Outcomes for the 4% threshold also fall into two distinct periods. Up to 1982, every year had (current vintage) measured inflation above the threshold. Thereafter, every year but one had inflation below the threshold. The result is that the 5Q-7Q forecasts have very few high-inflation outcomes. Furthermore, long-horizon forecasts for 1983 and 1984 assigned low (10-20%) probabilities to the possibility that inflation would fall to below 4%, which it subsequently did. This is likely the most important reason for the significant evidence of mis-calibration that we find below.
Figure 3.2.2
Kernel-estimated conditional expectation of outcome given forecast
Survey of Professional Forecasters data, two outcome measures

More analysis of the SPF inflation forecasts and their calibration is presented in Table 3.2.1. The upper half of the table gives results for the 2% threshold while the lower half refers to 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 for comparison.\textsuperscript{12} Comparing the mean squared forecast error (MSFE) with the variance of inflation outcomes, we see that the forecasts capture most of the variance of inflation outcomes at the shortest horizons, but that

\textsuperscript{12} Differences in outcome variance are artefacts of the differing number of observations available at different horizons, as was discussed above.
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.\textsuperscript{13} We also note that using revised data and the 4\% threshold, the MSFE exceeds the variance of the inflation outcomes for the three longest forecast horizons. If not due to sampling error, this would imply that these forecasts cannot be efficient predictors of inflation outcomes.\textsuperscript{14}

The importance of the estimated calibration errors vary considerably, ranging from fully 100\% of the MSFE to less than 5\%. Data vintage plays no clear role in calibration errors, with generally smaller errors using the 2008Q1 vintage with the 2\% threshold or the real-time data with the 4\% threshold. However, the significance of the calibration errors is stronger using the real-time data; the null hypothesis of no calibration error is rejected in 15 of the 16 cases examined using real-time data and only 5 of 16 using the revised data.

Since the results for the 4\% threshold at the longest horizons (5Q-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 (5Q-7Q) and shorter (3Q-4Q) 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 (and roughly equal to variance of the inflation outcomes) when we use the recent data vintage to measure inflation. The statistical evidence of mis-calibration in the latter case is very strong. The presence of such apparently systematic errors may be surprising; it suggests that forecasters experienced at least 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 not historical forecasts, but are forecasts which could have been made using data that were available at the time. Specifically, 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).\textsuperscript{15} That study compared inflation pseudo-forecasts at various horizons from a

\textsuperscript{13}Both of these results reflect the nature of the forecasting experiment; at the 1Q horizon, for example, forecasters are trying to forecast inflation over a calendar year but have official estimates of inflation over the first two quarters of that year.

\textsuperscript{14}This is suggestive of forecasts made beyond the content horizon; see Galbraith 2003, Galbraith and Tkacz 2006.

\textsuperscript{15}CPI data are not revised, so only vintage series for real GDP from the real-time database of the FRB Philadelphia were used.
set of fifteen simple linear models using only lagged inflation and real-time estimates of the output gap in their specification and construction. The authors concluded that none of the 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. 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. Evaluation of probabilistic output-gap-based inflation forecasts should therefore be of interest to policy makers.

The Orphanides-van Norden models are used to forecast average US CPI inflation over 2-, 4-, 6- and 8-quarter horizons. Probabilistic forecasts for inflation thresholds of 2% and 4% per annum are formed based on OLS estimates of linear forecasting equations with conventional standard error estimates and assumed Gaussian errors. 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 probabilistic 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.

The 15 models are Linear trend (LT); Quadratic trend (QT); Broken trend (BT); 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. 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.) The broader, lighter lines correspond with outcomes where inflation exceeded the threshold and the narrower, dark lines with 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 across inflation outcomes 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 2Q and 6Q forecast horizons. Results for the 8Q horizon were similar to those for 6Q, while those for 4Q tended to fall between the 2Q and 6Q results. These figures, which display results from all fifteen models in one graph, are intended to illustrate the typical patterns rather than to illustrate points about particular models within the set; individual lines are difficult to distinguish and so are not labeled. It is clear that the models’ performances have a good deal in common, although differences can be seen in the numerical results recorded in Tables 3.3.1A and 3.3.1B, discussed below.

Again, we note that sampling error may account for some of the deviations from the 45° line, so that these deviations must be interpreted in the light of the evidence from the statistical inference presented below. In particular, as we can see from the middle panel of Figure 3.3.1, probabilistic forecasts of inflation exceeding 2% almost always exceeded 50%, so the calibration functions to left of that level in the top panel of Figure 3.3.2 will be relatively less precise and should be interpreted with caution. Similarly, forecast probabilities of increases in inflation of over 80% are relatively rare, so the extreme right-hand portion of bottom panels of Figure 3.3.2 should also be interpreted with caution.

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 some conditional expectation functions are nearly flat; this may reflect the relative absence
of data points in this region, as noted above. 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, where there are relatively few observations.

The estimated calibration functions are fairly similar at the longer (6Q) forecast horizon of Figure 3.3.2B. At the 2% threshold (top panel) they are quite 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.

The tendency of the models to underestimate the probability of inflation > 2% but to overestimate the probability of inflation > 4% may appear surprising. One possible explanation is that the linear forecasting models used here may have overestimated the volatility of inflation forecast errors, particularly after inflation rates stabilized in the 1980’s. Using an overestimate of the variance of forecast errors would tend to exaggerate the probability that inflation will be in the tails of the distribution. In fact inflation exceeds 4% roughly half the time and 2% about 90% of the time. We would therefore expect misspecification of the forecast variance to affect the 2% inflation forecast disproportionately.

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. 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 15% of the MSFE. Despite their size, the calibration errors are statistically significant at the 5% level for just over half the models. 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.) Calibration errors are much larger, typically above 0.025, and are very strongly significant in every case. The results for the forecast direction of the change in inflation (bottom panel) are different again. The MSFEs are only marginally less than the variance of inflation outcomes, suggesting that none of models have much forecasting power. Calibration errors are consistently small (never more than 0.01) however, and are significant at the 10% level in only 1 of 15 cases.

As shown in Table 3.3.1B, forecast performance is generally poorer at longer forecast horizons. For the 4% inflation and 0% change in inflation thresholds, the MSFEs hover around (and occasionally exceed) the variance of the inflation outcomes; for the 2% threshold the MFSEs always exceed the variance of the outcomes. Calibration

\[17\text{ Again, this is suggestive of forecasts made beyond the content horizon; see Galbraith 2003, Galbraith and Tkacz 2006.}\]
errors typically account for less than 20% of the MSFE and only rarely exceed 50%. Nonetheless, this represents considerable statistical evidence of mis-calibration. For the 2% threshold, we find very strongly significant evidence of mis-calibration for every model; for the change in inflation we find evidence of mis-calibration significant at the 5% level for all but two models. Evidence of mis-calibration is relatively weaker for the 4% inflation threshold; even here, however, that evidence is statistically significant at the 10% level for 12 of the 15 models.
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
To summarize, there is considerable evidence that such output gap models give imperfectly calibrated estimates of the risk that inflation will either exceed given thresholds, or that it will simply increase. Forecasts at horizons over one year typically explain little or none of the variation in inflation outcomes and we find that most models show significant evidence of mis-calibration. At short horizons, results are more varied. Forecast performance is generally better and measured calibration error is relatively less important when using a 4% inflation threshold, although statistically significant evidence of mis-calibration is relatively common. Forecasts using a 2% inflation or for the direction of change inflation show little apparent forecasting power, although mis-calibration does not appear to be important for the latter group of forecasts. We conclude that such output gap models do not appear to be very useful for forecasting the risk of very low levels of inflation.

4. Concluding remarks

Estimation of the calibration of continuous probabilistic economic forecasts can be carried out, without arbitrary grouping, by using kernel regression estimates of the necessary conditional expectation function. Doing so allows both graphical characterization of the calibration and numerical computation of quantities such as the mean squared calibration error. In applying these methods 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. Estimates of calibration at sets of points in the [0, 1] interval may also suggest ranges of probability values for which forecast methods have performed particularly poorly.

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 was 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 statistically significant.
REFERENCES


