The Resolution and 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

draft version of 2 April 2007

Abstract

A probabilistic forecast is the estimated probability with which a future event will satisfy a specified criterion. Economic forecasts of this type are increasingly often produced, but have been subject to relatively little formal evaluation. By contrast, meteorological probability forecasts which fall into discrete categories are routinely evaluated with respect to their calibration error (a measure of the match between forecast and actual probabilities) and resolution (measuring the ability of forecasts to distinguish cases of differing conditional probability). We show here that we can also estimate these quantities in economic contexts in which the forecast probability may take any value in the unit interval, and we evaluate the empirical evidence on calibration error and resolution for a number of important economic applications, including recession and inflation prediction. We also indicate the maximum horizons at which forecasts have useful resolution and compare calibration and resolution across alternative models. Real-time data are used as much as possible in these applications to obtain realistic estimates of performance. Although a number of the forecasts that we evaluate show low calibration error and substantial resolution at short horizons, we also observe cases in which models show poor performance.

Key words: calibration, probability forecast, resolution

* We thank 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)* for support of 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.[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 the event, on the continuous interval $[0, 1]$. While probabilistic forecasts of economic events are increasingly common, they have not been subjected to much of the verification that meteorological precipitation-probability forecasts, for example, routinely undergo; with the exception of a few studies such as that of Diebold and Rudebusch (1989), little is known about key features of sequences of economic probability forecasts.

Although it is possible to generate sequences of historical pseudo-forecasts by applying probabilistic forecasting models to historical time series, the resulting forecasts typically do not correspond to those that could have been made at a given historical date, even if parameter estimation and model selection are carried out only on data available at that date. A primary reason for this is the revision of important macroeconomic data: typically, the currently-available data corresponding to a historical date not only differ from those available at the time, but differ by having higher information content. It is well known that, as a result, the use in forecast evaluation of measures based on current vintages of data may overstate the actual usefulness of these measures in real-time forecasting; see for example Croushore and Stark (2003).

To evaluate probabilistic forecast sequences, therefore, we prefer sequences of forecasts made and recorded in real time, or alternatively pseudo-forecasts computed recursively on real-time vintages of data (that is, the forecast computed for historical date t uses only the revision of each data point which was actually available to the forecaster at t). The existence of a number of sets of vintage macroeconomic data makes these computations feasible.[2] As well, some actual sequences of historical probability forecasts have been recorded and stored and are available for evaluation.

For each of these sequences, this study will evaluate the calibration, resolution, and other features of the forecasts. Calibration measures the match between forecast probabilities and actual conditional probabilities of an event occurring in a given state; resolution is a measure of the ability of forecasts to distinguish states with relatively high conditional probabilities of the event from states with relatively low conditional

[1] Probabilistic forecasts on financial variables are commonly traded in markets as binary or digital options.

[2] The Federal Reserve Banks of St. Louis and Philadelphia make vintage data available. The Federal Reserve Bank of Philadelphia makes available a number of series of historical forecasts, both point and probability. The Bank of England also publishes historical forecasts from their *Quarterly Report on Monetary Policy*.

probabilities. These terms will be made precise below. Together, the calibration and resolution of a sequence of forecasts exhaust the model-specific information in the mean squared error. Clearly, one aim of this evaluation is to provide a direction for improvement of forecasting models: the bias implied by calibration error, for example, can be estimated and removed (see for example Hamill et al. (2003) on the usefulness of this procedure in precipitation probability forecasts).

The present paper will work with several sequences of probabilistic forecasts of recessions or of breaching inflation targets. In the former case, evaluation is complicated by the fact that the actual state of the economy never becomes perfectly observable; we use both real-time and final-revision estimates of the economy's state in the evaluation. Section 2 of the paper defines the measures of interest and procedures for pointwise and uniform statistical inference. Section 3 describes the data, vintages, and forecasts or forecasting models that are subject to evaluation. Section 4 implements these measures and tests.

2. Methods for probabilistic forecast evaluation

2.1 Definitions

The evaluation of probability forecasts begins with the observation that both forecasts and outcomes are random variables, and that the two sequences have a joint distribution; see in particular Murphy (1973), from which much subsequent work follows. Numerous decompositions of the forecast MSE are possible; in particular, we can treat either the forecast or the outcome as conditioning variables.

Following most of the notation of Murphy and Winkler (1987), 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. Since the variance of the binary outcomes is fixed, it is useful to use a decomposition that conditions on the forecasts: in this case the mean squared error $E((\hat{p} - x)^2)$ of the probabilistic forecast can be decomposed as follows:

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

where $E_f(z) = \int z f(z) dz$ with $f(.)$ the marginal distribution of the forecasts, $\hat{p}$. 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 does depend on the forecasts is contained in the second and third terms on the right-hand side of (2.1).

The first of these,

$$E_f(\hat{p} - E(x|\hat{p}))^2, \tag{2.2}$$

will be called the (mean squared) *calibration error,* and measures the deviation from a perfect match between the predicted probability and the true probability of the event

when a given forecast is made.[3] 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 perfectly calibrated. 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 perfectly 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 perfect calibration can be achieved by setting $\hat{p} = E(x)$, the unconditional probability of a recession, since the expectation is taken over the possible values or range of values that the probability forecast can take on.

The last term on the right-hand side of (2.1), $E_f(E(x|\hat{p}) - E(x))^2$, is called the forecast *resolution*, and measures the ability of forecasts to distinguish among relatively high-probability and relatively low-probability cases. Note again that the expectation is taken with respect to the marginal distribution of the forecasts. If resolution is high, then in typical cases the conditional expectation of the outcome differs substantially from its unconditional mean: the forecasts are successfully identifying cases in which probability of the event is unusually high or low. The resolution enters negatively into the MSE decomposition; high resolution lowers MSE. To return to the previous example, the simple forecast that always predicts a 5% probability of recession, where 5% is the unconditional probability, will have zero resolution. Perfect forecasts would have resolution equal to variance (and zero calibration error, so that MSE = 0).

These concepts are employed in the meteorological literature in the evaluation of discretely-distributed probability forecasts which may take on only values in a finite set (e.g. precipitation-probability forecasts may take on the values $\{0, 0.2, 0.4, \ldots, 1.\}$). By contrast, econometric probability forecasts are normally allowed to take any value in the interval $[0, 1]$. In the next subsection we consider methods that will allow us to estimate these quantities for continuously-distributed probability forecasts.

The calibration error has a minimum value of zero; its maximum value is 1, where forecasts and conditional expectations are perfectly opposed. The resolution also has a minimum value of zero, but its maximum value is equal to the variance of the binary outcome process. In order to report a more readily interpretable measure, scaled into $[0, 1]$, we divide the resolution by the variance of the binary outcome process. Let n be the number of observed forecasts and $\mu = E(x)$; then the maximum resolution achievable arises where there are $n\mu$ 1's and $n - n\mu$ 0's constituting the sequence $E(x|\hat{p})_i$. The resulting maximum total is $n\mu(1-\mu)^2 + n(1-\mu)\mu^2$. Divide by n for the mean; this quantity is then the maximum resolution and is also equal to the variance of a 0/1

[3] 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.

random variable with mean μ. Therefore

$$\frac{E_f(E(x|\hat{p}) - \mu)^2}{\mu(1-\mu)^2 + (1-\mu)\mu^2} \in [0,1]. \tag{2.3}$$

2.2 Estimation of the calibration error and resolution

Estimation of either (2.2) or (2.3) requires estimation of the conditional expectation function $E(x|\hat{p})$ (estimation of the unconditional probability $E(x)$ is of course by the sample mean $\overline{x}$ of the binary outcome). When the forecasts take on only a certain number of 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 forecasts into J cells each containing N^j forecasts; the authors then compute the local squared bias measure $N^{-1}\sum_{j=1}^{J} 2n^j(\overline{\hat{p}}^j - \overline{x}^j)$, where $\overline{\hat{p}}^j$ and $\overline{x}^j$ are the mean forecast probability and the mean outcome on the N^j values of 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.

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 $\hat{p}$. Note that this allows us in principle to evaluate calibration at any point in the continuous interval $[0,1]$, rather than only globally or within a set of discrete cells.

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. 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.

2.3 Statistical inference on calibration

Our ability to conduct inference on the estimated conditional expectation functions is limited by the dependence which exists in both forecasts and outcomes. However, kernel regression estimates have been shown to remain consistent and asymptotically normal with dependent data under various conditions; see for example Robinson (1983, 1986).[4] Pointwise inference on calibration error (that is, a test of $H_0 : p_i = E(X|p_i)$ at any given probability 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.

A 'global' test would address the question of whether or not the entire function $E(X|p_i)$ can be reduced to the linear form $a + bp_i$ with $a = 0,\ b = 1$, as is implied by the null hypothesis of correct calibration throughout the interval $[0, 1]$ (recall that for correctly-calibrated data, the conditional expectation function takes a linear form; $E(X|p_i) = p$). For iid data, tests of a linear specification vs a general nonparametric alternative have been given by, e.g., Härdle and Mammen (1993) and Azzalini and Bowman (1993), but no such procedure appears to be available for dependent data. A more limited global test can however be conducted by testing the hypotheses $a = 0$ and $b = 1$ jointly, or individually, within a linear model. Such a test loses power against nonlinear deviations from the null, but is feasible in a general context. Dependence in deviations from the model does imply inconsistency of the 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 such statistics below using the Newey-West estimator for the covariance matrix and test the joint hypothesis using a Wald statistic.

3. Data, forecasts and pseudo-forecasts

We implement the measures described in Section 2 on two classes of data: recession probability forecasts and forecasts of the probability that inflation will exceed a target (taken as 3% annually in these evaluations). In each case we evaluate several sets of forecasts or forecasting models, and a variety of forecast horizons. In examining multiple horizons we are in part interested in the point beyond which forecasts cease to be useful for each series. In the context of point forecasts, Galbraith (2003) refers to this point as the *content horizon* and gives a definition based on an arbitrary loss

[4]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.

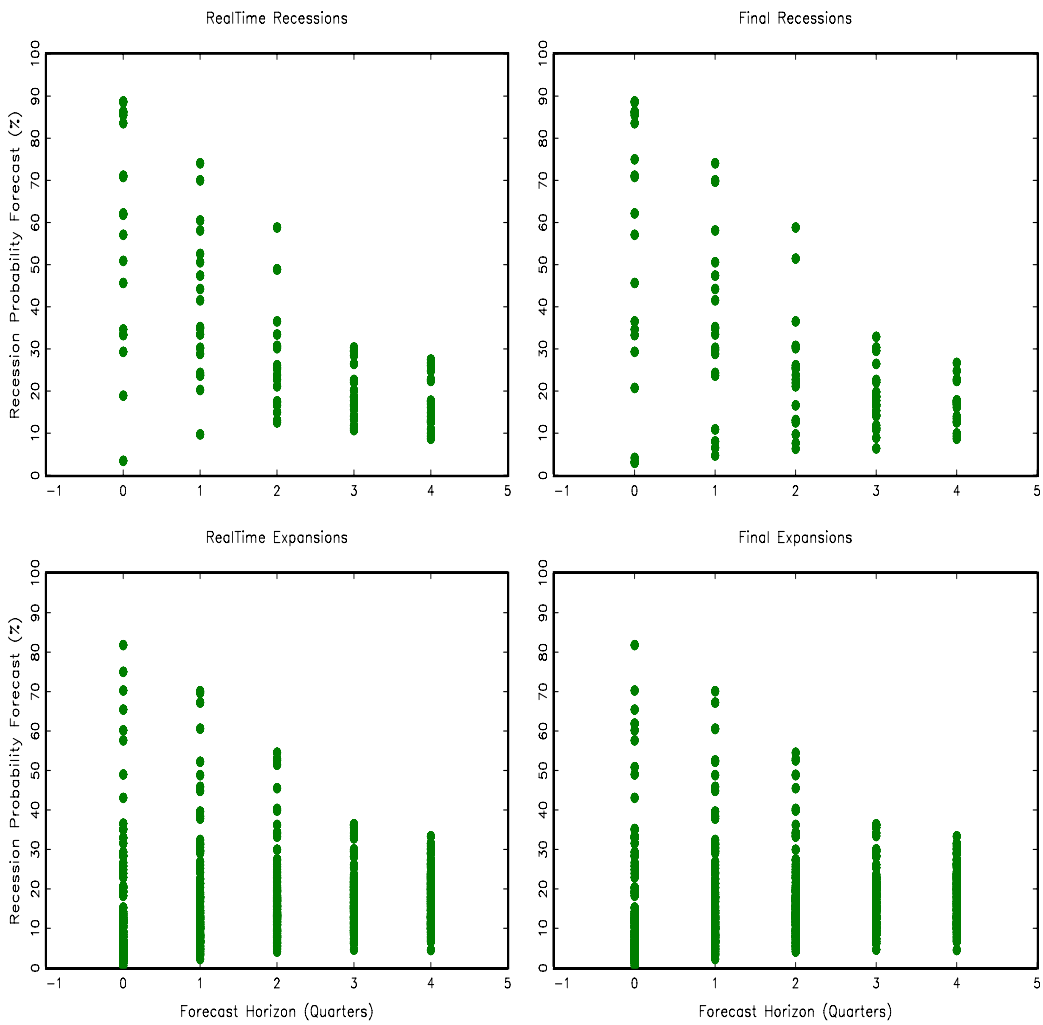
function, usually implemented using the mean squared error; in the present context, the analogue of this content horizon will be the point at which forecast resolution becomes approximately zero.

3.1 Recession probability forecasts

The recession forecasts that we consider come from the Survey of Professional Forecasters (SPF), compiled since the last quarter of 1968 first by the American Statistical Association and National Bureau of Economic Research, more recently by the Federal Reserve Bank of Philadelphia. 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. The forecasts are estimated probabilities that the economy will experience negative 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 $t+h$. The first of these, for $h=0$, is a roughly contemporaneous estimate of the current binary state (recession or expansion) of the economy. Since the true state never becomes observable, we consider two estimates of the economy's binary state, one based on the initial ('real time') estimate of GDP growth, and the other based on the final estimate, made some time later.

Figure 3.1.1 depicts the actual recession forecasts, divided into cases in which the economy did/did not subsequently enter a recession state in the relevant quarter. The revision in GDP growth figures causes the small difference between the real time (based on first-released estimates of growth) and final (2006Q1) estimates of GDP growth; both of these are reported in Figure 3.1.1. Note that, at least for shorter horizons, the panels describing expansion quarters show a concentration of forecasts at low probabilities, whereas those describing recession quarters show fewer low probability results. By horizon 4, however, there is little visible difference in the range of typical outcomes.

FIGURE 3.1.1
Dispersion of probabilistic forecasts
Survey of Professional Forecasters data



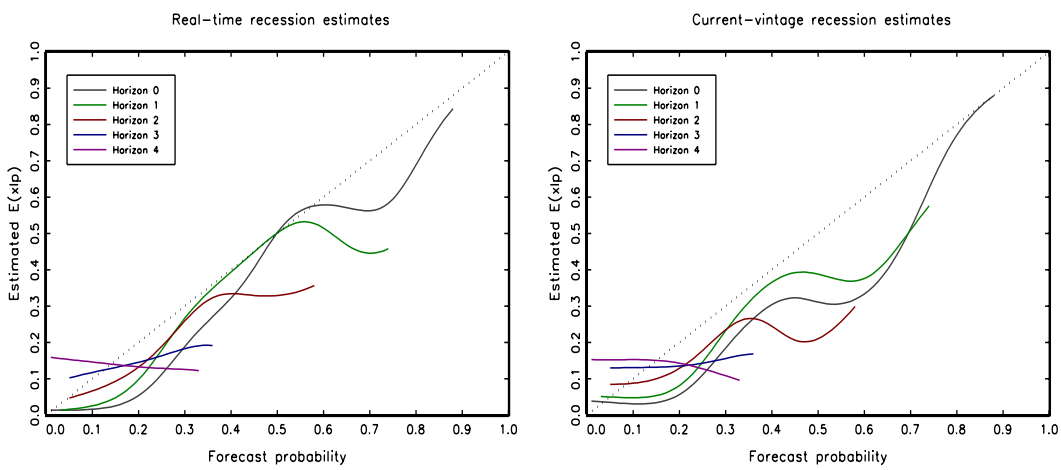
We now compute the calibration and resolution measures for these data. We have two choices of outcome data series (real time estimate and 2006 vintage estimate) by which to measure the 0/1 outcome.

Recall that we also must estimate the continuous conditional expectation function $E(x|\hat{p})$ in order to evaluate (2.2) and (2.3). We have choices of bandwidth and kernel function to make in order to do so. The tables below will report results from the standard Nadaraya-Watson kernel estimator with a Gaussian kernel function; results are typically much less sensitive to kernel choice than to the choice of bandwidth parameter. For bandwidth parameter selection we begin with cross-validation, which estimates an

optimal bandwidth 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 report results in which this value is varied by ±50%.

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 conditional expectation function for the ideal forecast would be a 45 degree line showing equality between forecast and $E(x|\hat{p})$. The conditional expectation functions are plotted over the actual range of observed probability forecasts, which can be found from Figure 3.1.1; 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.

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



The following tables record the root mean squared calibration error and scaled resolution results; point estimates of calibration error suggest *prima facie* that the forecasts are quite well calibrated.

Table 3.1.1
RMS calibration error ($\in [0,1]$)
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
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

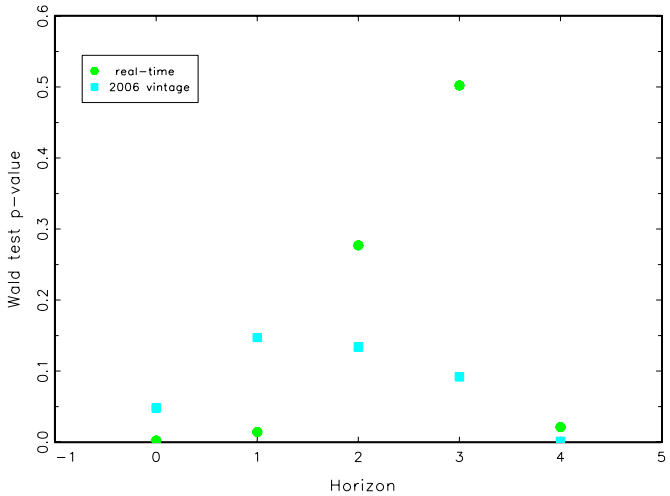
Table 3.1.2b
Scaled resolution measure ($\in [0,1]$)
Recession probability forecasts, SPF data

	Horizon: (quarters)				
	0	1	2	3	4
Real-time outcome series					
h=0.08	0.42	0.21	0.06	<0.01	<0.01
h=0.04	0.48	0.26	0.12	0.02	<0.01
h=0.12	0.37	0.16	0.03	<0.01	<0.01
2006 vintage outcome series					
h=0.08	0.35	0.15	0.03	<0.01	<0.01
h=0.04	0.39	0.19	0.07	<0.01	0.01
h=0.12	0.31	0.11	0.01	<0.01	<0.01

With respect to resolution, the measures show substantial resolution at horizons 0 and 1, but very little at horizons 3 and 4, a result which mirrors the fact that quarterly GDP forecasts have approximately zero 'skill' or 'forecast content' at these horizons (see for example Galbraith and Tkacz 2006). Roughly speaking, the horizon beyond which these forecasts have little or no remaining resolution appears to be about two quarters. Lower resolution appears in the vintage-2006 estimates of where recessions actually occurred than in the estimates made at the time.

Pointwise statistical inference on the hypothesis that calibration error is zero at each point yields very few rejections (none at horizons 0–2) and will not be reported in detail; the tests have little power in regions where there are relatively few forecasts, and confidence intervals are often very wide. A more powerful test uses the entire set of calibration results to test zero intercept and unit slope in the conditional expectation, as required by correct calibration. The p-values for Wald tests of this null hypothesis are reported in Figure 3.1.3.

FIGURE 3.1.3
p-values of Wald tests
H_0 : zero intercept and unit slope of $E(X|p_i)$
Survey of Professional Forecasters data, two outcome measures



Perfect calibration is rejected decisively at horizons 0 and 4, and at horizon 1 with real time outcomes. Note however that, while evidence against perfect calibration is strong at these points, the 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. That is, the forecasts appear to be well, but clearly not perfectly, calibrated. At horizons 3 and 4, the estimates of the linear

form are imprecise, but there is nonetheless sufficient evidence at horizon 4 to reject perfect calibration decisively.

3.2 Probability of exceeding an inflation target

Probabilistic inflation forecasts are sometimes made in the form of estimated probabilities that inflation will exceed a particular level. In this section we evaluate two types of such forecasts: model-based pseudo-forecasts and a sequence of forecasts derived from density forecasts published by the Bank of England. For the model-based pseudo-forecasts, we take care to use only information that would have been available at the time in model selection, estimation and forecasting, while the Bank of England forecasts represent true conditional forecasts published in real time by the central bank.

For US inflation, we use the models and forecasting methodology described by Orphanides and van Norden (2005). That study compared inflation pseudo-forecasts at various horizons from a set of fifteen simple linear models using only real-time data in their specification and construction.[5] 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 (MSFE.) However, in inflation-targeting countries the central banks may have an asymmetric forecast loss function. For example, they may be particularly interested in the probability that inflation stays below some upper bound that is considered consistent with explicit or implicit targets, so as not to reduce the policy framework's credibility. This suggests that evaluating the performance of probabilistic inflation forecasts may give additional information on whether such models can be of use to policy makers.

These models are used for forecasts of *cumulative* US CPI inflation over 2-, 4-, 6- and 8-quarter horizons. As there has been no explicit inflation target in the US, we considered forecasts of the probability that inflation would exceed a 3% per annum threshold.[6]

Figures 3.2.1a-c, like Figure 3.1.1, record the actual pseudo-forecasts made by these models over the sample period of 1969 Q2 through 2002 Q3. The upper panel in each figure records cases in which the eventual outcome was that inflation exceeded the 3 per cent threshold, while the lower panel shows the remaining cases. The fifteen different

[5]The 15 models are 1. Linear trend; 2. Quadratic trend; 3. Broken trend; 4. Hodrick-Prescott; 5. Band pass; 6. Beveridge-Nelson; 7. SVAR–Blanchard-Quah; 8. Watson (1986); 9. Harvey-Clark; 10. Harvey-Jaeger; 11. Kuttner; 12. Gerlach-Smets; 13. TOFU; 14. Nominal output; 15. Autoregressive. See Orphanides and van Norden (2005) for references and details on model specification and estimation. In some cases our results are numbered rather than labeled to conserve space; however we present all model results in the order shown above.

[6]Probabilistic forecasts are based on OLS estimates of linear forecasting equations with conventional standard error estimates and assumed Gaussian errors.

models are distributed along the horizontal axis. Here, in contrast with recession-forecast case, the inflation data are not subject to substantial revision and there is no real-time/final distinction. Because of the large number of models, a separate figure is used for each forecast horizon. Figures 3.2.1a-c plot every probabilistic forecast from each of the models and superimpose on them the box-whisker plot for their distribution. (The mark in the interior of the box indicates the median of the distribution, the box itself covers the interquartile range, and the whiskers extend from the 10th to the 90th percentile.)

Figure 3.2.1a
Dispersion of probabilistic forecasts of US cumulative 2Q inflation
Real-time pseudo-forecasts from fifteen models

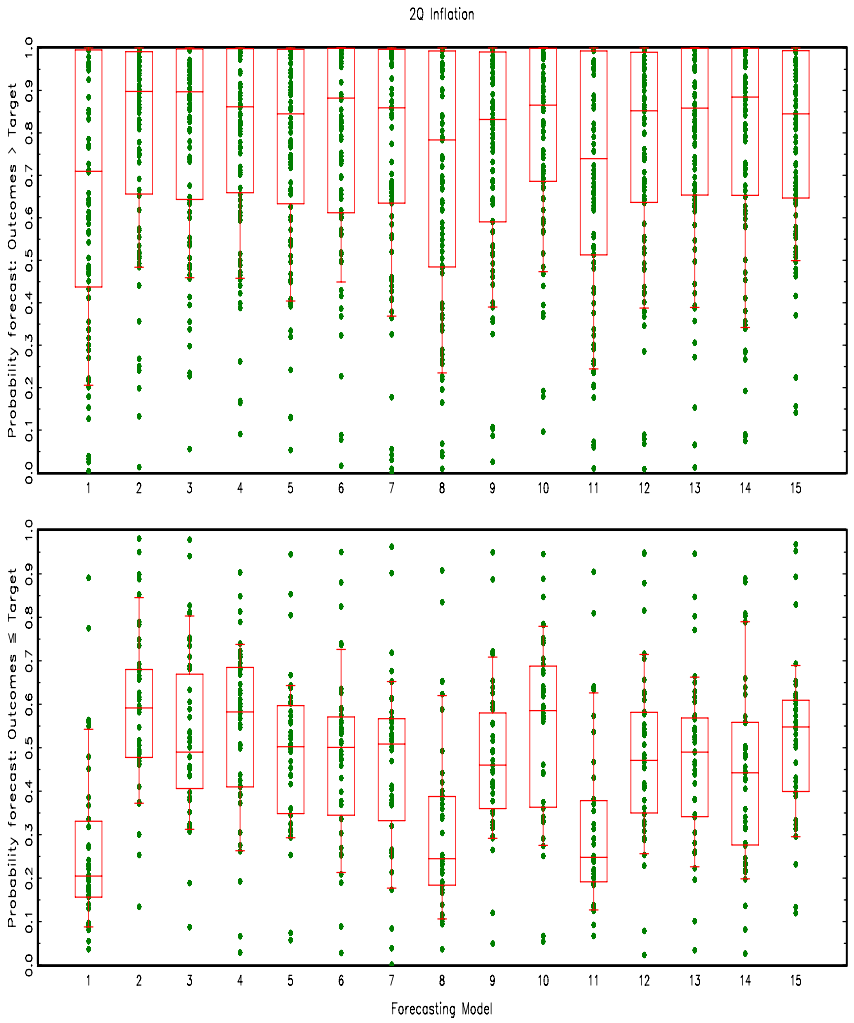


Figure 3.2.1b
Dispersion of probabilistic forecasts of US cumulative 4Q inflation
Real-time pseudo-forecasts from fifteen models

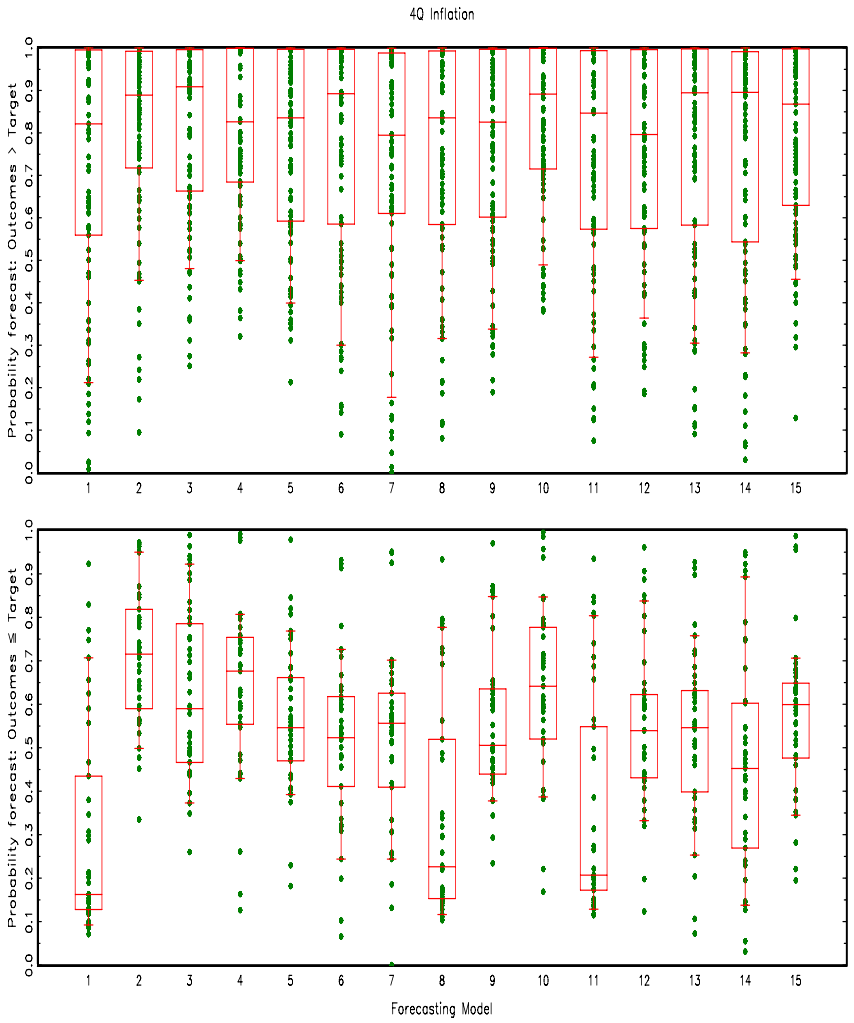
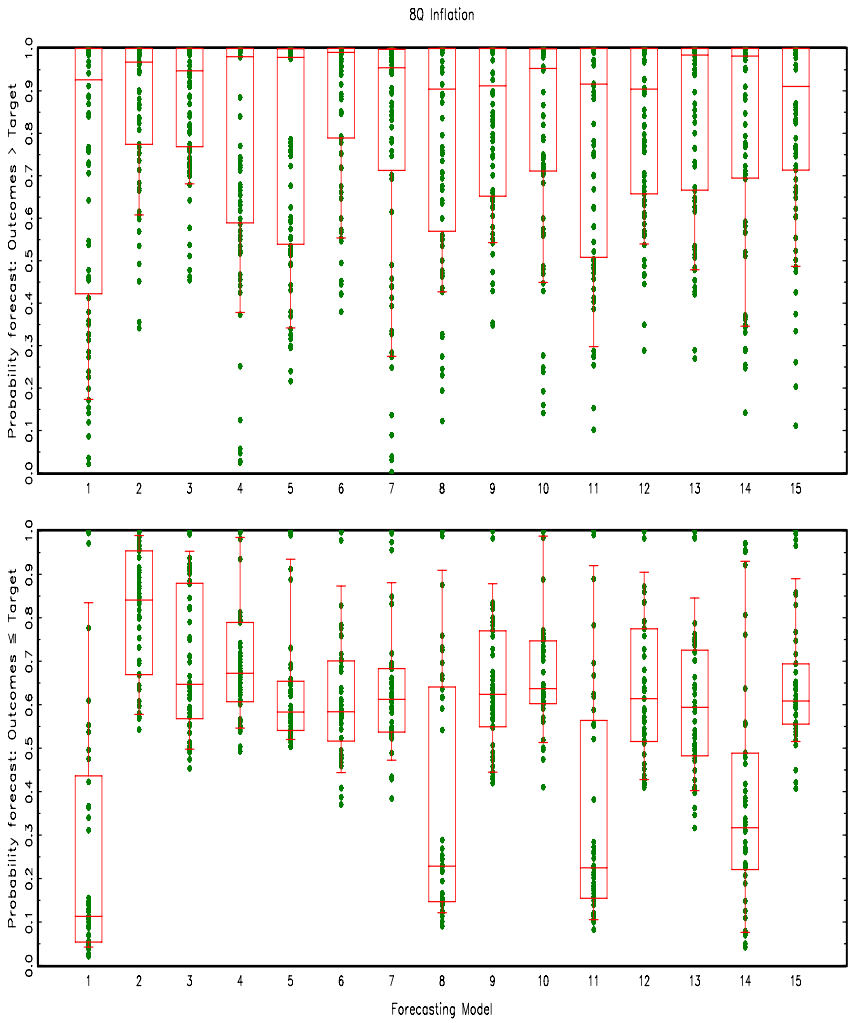


Figure 3.2.1c
Dispersion of probabilistic forecasts of US cumulative 8Q inflation
Real-time pseudo-forecasts from fifteen models



These figures allow simple *prima facie* comparisons of the various models. An excellent model will have forecasts tightly clustered around 1.0 in the upper panel and 0.0 in the lower panel. For comparison, model 15 shows the results from a univariate AR forecast of inflation, model 14 uses a bivariate VAR in nominal output growth and inflation, while model 13 uses a bivariate VAR in real output growth and inflation.

The graphs suggest that all models have at least some resolution; the distributions of the forecasts are systematically lower when inflation turns out to be below the threshold than when it turns out to be above it. We also see that while some models display a smaller downward shift (indicative of lower resolution) than the benchmark

AR and VAR models, three models display signs of higher resolution; models 1 (Linear Trend), 8 (Watson) and 11 (Kuttner.) These results appear to be robust across the three different forecast horizons (2, 4 and 8 quarters) shown here. The relatively good apparent performance of these three models is surprising given all three assume the presence of a constant trend growth rate of real output. Orphanides and van Norden (2002) note that such models tend to produce much less reliable estimates of output gaps, as evidenced by the particularly large ex post revisions in their estimates of the output gap.[7] We return to the discussion of resolution below with numerical measures.

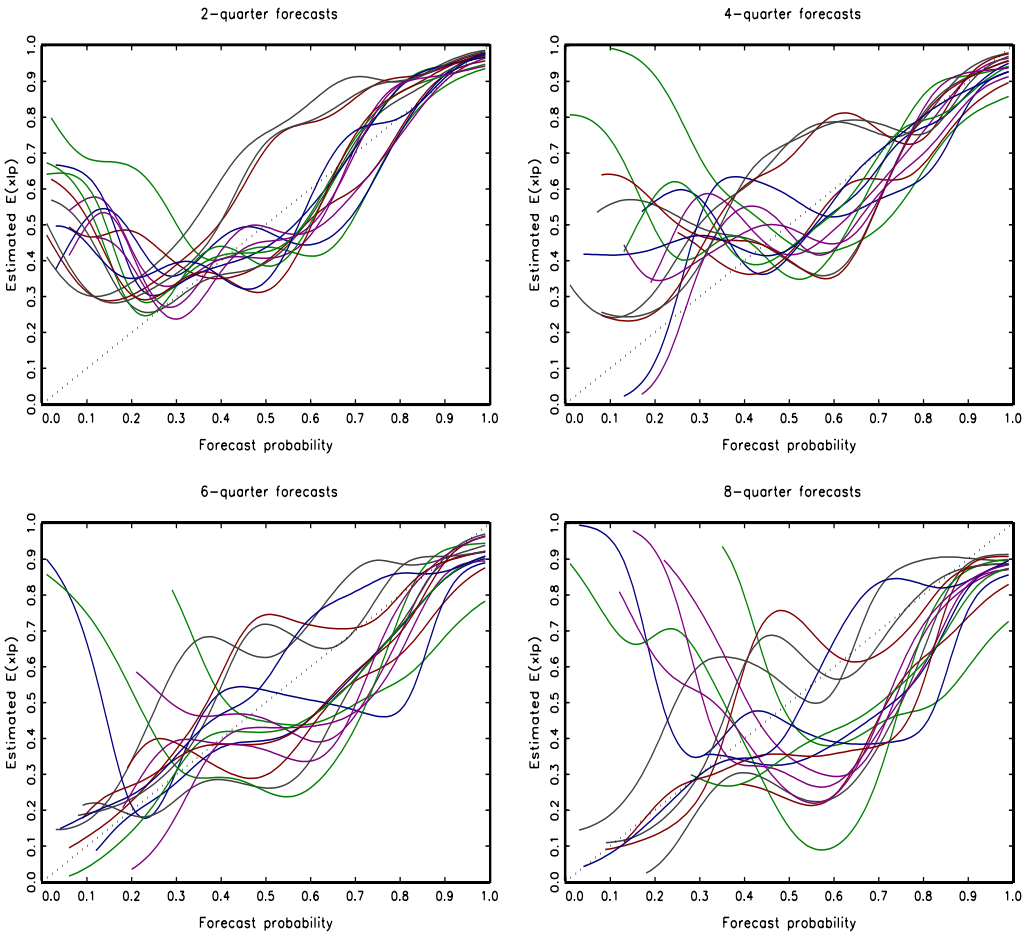
Next we apply the methods used in the previous section to these forecasts of the probability of breaching an inflation target, beginning with estimates of the conditional expectation functions. These are depicted for the base bandwidth and for each of the four cumulative horizons in Figure 3.2.2. The models are not individually labeled; the purpose of the figure is to indicate the dispersion across models in these estimated functions. These functions are again plotted only over the ranges of observed probability forecasts: a number of the models show no forecast probabilities near zero.

In considering the results for the 2Q horizon, we see that all models show large distortions for low forecast probabilities. When models estimate the probability of inflation$>$ 3% to be under 30 per cent, it is systematically higher. This distortion increases as we go lower; when they estimate the probability to be around 10%, inflation exceeds the threshold about 50 per cent of the time. This is common across all models: all output gaps, the AR model of inflation, TOFU and YN.

As we move to longer forecast horizons, the overall results change. First, performance varies more across models. Six of the 15 models show the same type of problem mentioned above but to an even stronger degree. There is increased dispersion of the results at low probabilities; at high probabilities, nearly all the models overstate the probability of exceeding the threshold.

[7]Orphanides and van Norden(2002), Table 3 shows that these three models all had noise-to-signal ratios above 1, compared with 0.69 for their best-performing model. One possible explanation for this surprising result is that the consistent over-estimation of the output gap these models tended to produce in the 1970s helped them correctly forecast an increase in inflation, and the reverse in the 1990s as output growth accelerated. We intend to investigate this hypothesis further.

FIGURE 3.2.2
Kernel-estimated conditional expectation of outcome given forecast
Fifteen inflation forecasting models



The RMS calibration error and scaled resolution estimates are in Tables 3.2.1 and 3.2.2. To avoid an excessively large number of entries, we report only the estimates using the base bandwidth, which are typical: as with the recession probability forecasts, sensitivity to the ±50% changes in bandwidth is modest.

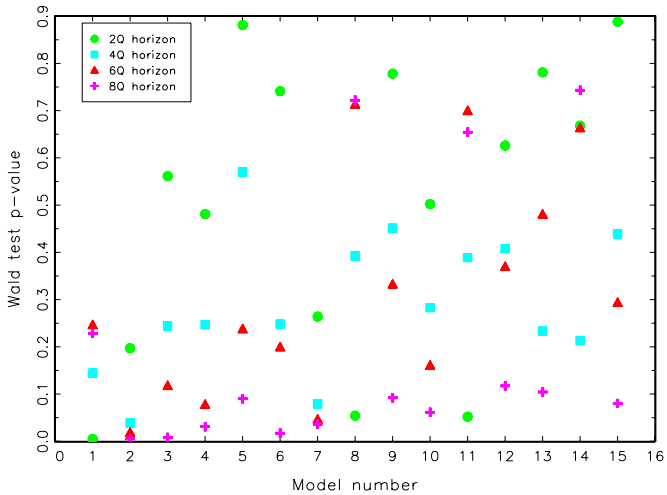
Table 3.2.1
RMS calibration error ($\in [0,1]$)
Probability of US Inflation > 3%, 15 models

	Horizon (cumulative, quarters)			
	2	4	6	8
Model 1 Linear trend	0.17	0.13	0.13	0.13
Model 2 Quadratic trend	0.15	0.20	0.21	0.28
Model 3 Broken trend	0.09	0.09	0.14	0.22
Model 4 Hodrick-Prescott	0.12	0.12	0.21	0.29
Model 5 Band pass	0.11	0.09	0.15	0.23
Model 6 Beveridge-Nelson	0.11	0.15	0.13	0.21
Model 7 SVAR–Blanchard-Quah	0.17	0.22	0.27	0.32
Model 8 Watson (1986)	0.14	0.10	0.09	0.11
Model 9 Harvey-Clark	0.13	0.11	0.10	0.17
Model 10 Harvey-Jaeger	0.09	0.10	0.14	0.25
Model 11 Kuttner	0.14	0.10	0.08	0.10
Model 12 Gerlach-Smets	0.13	0.11	0.10	0.15
Model 13 TOFU	0.13	0.17	0.09	0.17
Model 14 Nominal output	0.11	0.13	0.08	0.07
Model 15 Autoregressive	0.08	0.10	0.14	0.23

While a number of these forecasts appear to be quite well calibrated at short horizons (comparable with the SPF recession forecasts), there are a number for which calibration is markedly poorer, notably models 7 (SVAR) and 2 (Quadratic Trend). As we saw in Figure 3.2.2, as we move to larger horizons the performance of some models tends to improve while others tend to deteriorate. Model 14 (Nominal Output) improves at longer horizons, becoming the best model in terms of RMS calibration at the 8Q horizon. The constant-trend models (numbers 1, 8 and 11) also all appear better calibrated at long horizons than at short horizons. In contrast, models such as the breaking trend or the AR model perform much more poorly at long horizons than short.

As in the recession forecast data, pointwise statistical inference, on H_0 : calibration error = zero at each point yields very few rejections, and we record instead potentially more powerful global tests of zero intercept and unit slope in the linear approximation to the conditional expectation function. The Wald test p-values are recorded graphically in Figure 3.2.3.

FIGURE 3.2.3
p-values of Wald tests
H_0 : zero intercept and unit slope of $E(X|p_i)$
Fifteen inflation forecasting models, four horizons



While the point estimates of calibration are in many cases farther from zero than in the SPF data, perfect calibration is more difficult to reject in these inflation forecasting models: standard errors of estimated parameters tend to be higher. However, at longer horizons (at eight quarters, and to some extent at six quarters), models 2, 3, 4, 6 and 7 tend to show significant deviations from perfect calibration at conventional levels. While there is some evidence against the null at some horizon for models 8, 10, 11, models 8 through 15 tend overall to show a pattern not too dissimilar from the uniform distribution of p-values over $[0, 1]$ which characterizes a true null.

Table 3.2.2
Scaled resolution measure ($\in [0, 1]$)
Probability of US Inflation > 3%, 15 models

	Horizon (cumulative, quarters)			
	2	4	6	8
Model 1 Linear trend	0.35	0.37	0.42	0.40
Model 2 Quadratic trend	0.22	0.13	0.08	0.08
Model 3 Broken trend	0.28	0.17	0.20	0.24
Model 4 Hodrick-Prescott	0.25	0.18	0.18	0.22
Model 5 Band pass	0.30	0.22	0.22	0.27
Model 6 Beveridge-Nelson	0.32	0.27	0.40	0.41
Model 7 SVAR–Blanchard-Quah	0.29	0.25	0.38	0.46
Model 8 Watson (1986)	0.34	0.36	0.37	0.41
Model 9 Harvey-Clark	0.28	0.21	0.22	0.22
Model 10 Harvey-Jaeger	0.24	0.19	0.22	0.27
Model 11 Kuttner	0.35	0.36	0.36	0.38
Model 12 Gerlach-Smets	0.26	0.20	0.23	0.21
Model 13 TOFU	0.34	0.26	0.27	0.31
Model 14 Nominal output	0.32	0.23	0.34	0.41
Model 15 Autoregressive	0.26	0.23	0.33	0.31

Table 3.2.2 shows, as we saw in Figure 3.2.1, that all models have some resolution. Differences between the best and worst models appear to be smaller at shorter horizons and tend to increase somewhat with the forecast horizon; the best model has a scaled resolution of roughly four times greater than the worst model at the longest horizon. At the shorter horizon, the ratio of resolutions is approximately 1.5. We also see that the constant trend models (1, 8 and 11) identified in Figure 3.2.1 again show good resolution and are among the best performing models at all horizons. However, their performance now appears to be equaled or surpassed by several other models. At the shortest horizons, the TOFU (real output growth) models performs roughly as well, while at the longest horizons they are joined by the Nominal Output and Beveridge-Nelson models, and surpassed by the SVAR.

. . .

The second set of inflation forecasts that we evaluate is from the Bank of England (BoE) *Quarterly Monetary Policy Report.* We examine thirty-six quarterly forecasts made from 1998Q1 through 2006Q4 inclusive; because of the small sample size we report a more limited set of results.

The Bank of England's *Quarterly Monetary Policy Report* provides probabilistic forecasts of inflation in the form of 'fan charts,' indicating quantiles of the forecast distribution. Fan charts for RPIX inflation forecasts were published from 1993Q1 to 2003Q4, when they were replaced by CPI inflation forecasts. The GDP fan chart was first published in the 1997Q3 report. In addition to providing forecast distributions for roughly 0 to 8 quarters into the future, from the 1998Q1 forecast onwards these are provided conditional on the assumption of both fixed interest rates, and a "market-expectation-based" interest rate profile. The two assumptions typically provide very similar results; here, we limit ourselves to the analysis of the market-based interest rate results.

While the published charts provide only a visual guide to the degree of uncertainty that the BoE associate with their forecasts, they are based on an explicit parametric model of forecast uncertainty, as documented by Brittan, Fisher and Whitley (1998) and Wallis (2004), among others. Forecast errors are assumed to follow a two-piece normal or "binormal" distribution, whose behaviour is completely characterized by three parameters: a mean μ, a measure of dispersion σ, and a parameter γ which controls skewness.[8] These parameters therefore allow us to estimate the implied forecast probabilities that inflation or GDP growth would exceed any given threshold level; we use 2% here.[9]

Because there is only one short series, we will summarize these results in the text rather than in tabular form. Calibration is poor beyond horizon 0: estimated RMS calibration error ranges from 0.13 at horizon 0 to 0.44, 0.51, 0.49 at horizons 6–8. Tests of the null of (0,1) slope and intercept for calibration yield strong rejections for all horizons beyond 1. These results are compatible with other evaluations of the BoE density forecasts (again see Wallis 2004), which suggest that, particularly in the earlier forecasts, an excessively high estimate of the variance of future inflation was implicit in the forecast density. Scaled resolution is high at horizon 0 (0.73), but drops rapidly thereafter to estimated values in the range 0.03–0.11 for horizons 4–8.

4. Concluding remarks

Estimation of the calibration and resolution of continuous probabilistic economic forecasts estimation can be carried out without arbitrary grouping into intervals, using

[8]See Wallis (2004, p.66, Box A), for a discussion of the binormal distribution and its alternative parameterizations. Spreadsheets containing the parameter settings for all of the published fan charts are available on the Bank of England's web site (presently http://www.bankofengland.co.uk/publications/inflationreport/irprobab.htm).

[9]Like the normal distribution, the binormal lacks an exact closed-form formula for its cumulative distribution function (cdf). We therefore estimated the implied forecast probabilities by numerical integration from the probability density function. The resulting cdf estimates appear to be accurate to at least 0.001. Additional details and code are available from the authors upon request.

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 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.

We have analyzed both forecasts made and recorded in real time, and pseudo-forecasts from models estimated after the forecast dates. In the latter case, we use only the vintage of data that was available at date t in estimating, and in forecasting with, a given model. Even in the former case, different data vintages can have an effect on the results, since current and past estimates of whether or not the event took place (recession, breaching an inflation target) may differ.

The overall picture of the quality of economic forecast calibration is mixed. 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 perfect calibration. Resolution of these forecasts is substantial at horizons 0 and 1 (quarters), but is approximately zero at horizons 3 and 4, a fact which is compatible with studies of the performance of point forecasting models for quarterly real GDP growth.

With respect to the inflation forecasting models evaluated, calibration error appears to be substantial in both the US and UK data examined here. While the UK fan charts appeared to have high resolution at short horizons and little or no resolution at long horizons, the pseudo-forecasts examined for US inflation provided more mixed results. One surprise they provided was the relatively high resolution of output gap models which assume a constant growth trend. Given the existing evidence on the bias and lack of precision in such estimates, we think this finding should be investigated further.

References

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.

Brittan, E., P. Fisher and J. Whitley (1998) "The *Inflation Report* Projections: Understanding the Fan Chart." *Bank of England Quarterly Bulletin,* 30-37.

Casillas-Olvera, G. and D.A. Bessler (2006) "Probability forecasting and central bank accountability." *Journal of Policy Modelling* 28, 223-234.

Clements, M.P. "Evaluating the Bank of England density forecasts of inflation." *Economic Journal* 114, 844-866.

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.

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." Forthcoming, *Canadian Journal of Economics* .

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.

Härdle, W. and E. Mammen (1993) "Comparing nonparametric versus parametric regression fits." *Annals of Statistics* 21, 1925-1947.

Murphy, A.H. (1973) "A new vector partition of the probability score." *Journal of Applied Meteorology* 12, 595-600.

Murphy, A.H. and R.L. Winkler (1987) "A general framework for forecast verification." *Monthly Weather Review* 115, 1330-1338.

Murphy, A.H. and D.S. Wilks (1998) "A case study of the use of statistical models in forecast verification: precipitation probability forecasts." *Weather and Forecasting* 13, 795-810.

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 nonparametric kernel time series regression, autoregression and density estimators." *Annals of the Institute of Statistics and Mathematics* 38, 539-549.

Wallis, K.F. (2004) "An Assessment of Bank of England and National Institute Inflation Forecast Uncertainties." *National Institute Economic Review* 198, 64-71.

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 Finance* 4, 61-73.