2009s-36

Calibration and Resolution Diagnostics for Bank of England Density Forecasts

John W. Galbraith, Simon van Norden

Série Scientifique
Scientific Series

Montréal
Août 2009

CIRANO
Allier savoir et décision

Centre interuniversitaire de recherche en analyse des organisations

CIRANO

Le CIRANO est un organisme sans but lucratif constitué en vertu de la Loi des compagnies du Québec. Le financement de son infrastructure et de ses activités de recherche provient des cotisations de ses organisations-membres, d'une subvention d'infrastructure du Ministère du Développement économique et régional et de la Recherche, de même que des subventions et mandats obtenus par ses équipes de recherche.

CIRANO is a private non-profit organization incorporated under the Québec Companies Act. Its infrastructure and research activities are funded through fees paid by member organizations, an infrastructure grant from the Ministère du Développement économique et régional et de la Recherche, and grants and research mandates obtained by its research teams.

Les partenaires du CIRANO

Partenaire majeur
Ministère du Développement économique, de l'Innovation et de l'Exportation

Partenaires corporatifs
Banque de développement du Canada
Banque du Canada
Banque Laurentienne du Canada
Banque Nationale du Canada
Banque Royale du Canada
Banque Scotia
Bell Canada
BMO Groupe financier
Caisse de dépôt et placement du Québec
DMR
Fédération des caisses Desjardins du Québec
Gaz de France
Gaz Métro
Hydro-Québec
Industrie Canada
Investissements PSP
Ministère des Finances du Québec
Power Corporation du Canada
Raymond Chabot Grant Thornton
Rio Tinto
State Street Global Advisors
Transat A.T.
Ville de Montréal

Partenaires universitaires
École Polytechnique de Montréal
HEC Montréal
McGill University
Université Concordia
Université de Montréal
Université de Sherbrooke
Université du Québec
Université du Québec à Montréal
Université Laval

Le CIRANO collabore avec de nombreux centres et chaires de recherche universitaires dont on peut consulter la liste sur son site web.

ISSN 1198-8177

Partenaire financier
Développement économique, Innovation et Exportation
Québec

Calibration and Resolution Diagnostics for Bank of England Density Forecasts [*]

John W. Galbraith[†], *Simon van Norden*[‡]

Résumé / *Abstract*

Cet étude développe et applique des nouvelles techniques pour diagnostiquer les prévisions de densité de la Banque d'Angleterre (leur "fan charts"). Nous calculons leurs probabilités implicites pour des taux d'inflation et de croissance du PIB qui dépassent des seuils critiques (soit le taux d'inflation ciblé, soit 2.5%.) En contraste avec des travaux antérieurs sur ces prévisions, nous gaugeons leur calibration aussi bien que leur résolution, en donnant des tests formels et des interprétations graphiques. Les résultats renforcent des conclusions déjà existant sur les limites de ces prévisions et ils donnent de nouvelles évidences sur leurs valeurs ajoutées.

Mots clés : calibration, prévisions de densité, probabilités implicites, résolution.

This paper applies new diagnostics to the Bank of England's pioneering density forecasts (fan charts). We compute their implicit probability forecast for annual rates of inflation and output growth that exceed a given threshold (in this case, the target inflation rate and 2.5% respectively.) Unlike earlier work on these forecasts, we measure both their calibration and their resolution, providing both formal tests and graphical interpretations of the results. These results both reinforce earlier evidence on some of the limitations of these forecasts and provide new evidence on their information content.

Keywords: *calibration, density forecast, probability forecast, resolution.*

[*] We thank Ken Wallis and conference and seminar participants at the Federal Reserve Bank of Philadelphia and the Reserve Bank of New Zealand for valuable comments. We also 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.
[†] Department of Economics, McGill University, 855 Sherbrooke St. West, Montreal, Quebec, Canada, H3A 2T7.
‡ Finance, HEC Montréal, 3000 Ch. de la Côte Ste Catherine, Montreal, Quebec, Canada, H3T 2A7, email: Simon.Van_Norden@cirano.qc.ca.

1. Introduction

Since their introduction in the 1993 Inflation Report, the Bank of England's probability density forecasts ("fan charts") for inflation, and later output growth, have been studied by a number of authors. Wallis (2003) and Clements (2004) studied the inflation forecasts and concluded that while the current and next-quarter forecast seemed to fit well, the year-ahead forecasts significantly overestimated the probability of high inflation rates. Elder, Kapetanios, Taylor and Yates (2005) found similar results for the inflation forecasts, but also found significant evidence that the GDP forecasts do not accurately capture the true distribution of risks to output growth at very short horizons.[1] They also explored the role of GDP revisions in accounting for GDP forecast errors and noted that the dispersion associated with predicted GDP outcomes was increased as a result of their research. Dowd (2008) examined the GDP fan charts and found that while short-horizon forecasts appear to capture the risks to output growth poorly, results for longer horizon forecast are sensitive to the vintage of data used to evaluate the forecasts.

The Bank of England's most recent performance evaluation of the fan charts was published in their August 2009 Inflation Report. It noted the failure of the forecasts made in early 2008 to assign significant probabilities to the inflation and growth outcomes witnessed over the subsequent year as a result of the financial crisis. While not mentioning any methodological changes made to address this experience, the Report noted that the Bank increased the dispersion of GDP outcomes and added negative skewness to their distribution.

Throughout this evaluative work, the focus has been on whether the risks implied by the Bank's fan charts are well matched (in a statistical sense) by the frequency of the various inflation and output growth outcomes. Gneiting, Balabdaoui and Raftery (2007) refer to this property as 'probabilistic calibration'; asymptotic theory for tests of correct calibration was provided by Diebold, Gunther and Tay (1998). However, it is well known that different density functions may satisfy this correct calibration property yet convey quite different amounts of information.[2] Mitchell and Wallis note that this extra information has been referred to as 'sharpness', 'refinement' or 'resolution' in various contexts and note its relationship to the Kullback-Leibler Information Criterion. Although forecast sharpness or resolution is desireable, no empirical studies of the Bank's density forecasts have investigated this property.

One reason for this may be the difficulty in estimating sharpness. In general, this requires an empirical characterization of the density of future outcomes $f(x)$ conditional on some density forecast $g(x)$. Without assumptions restricting the set of density functions $\{f(x), g(x)\}$ to be considered, this will in general require much more data than is

[1] Noting the hazards of drawing firm conclusions from small samples, these authors suggested that "the fan charts gave a reasonably good guide to the probabilities and risks facing the MPC [monetary policy committee]."

[2] See Corradi and Swanson (2006) and Mitchell and Wallis (2009) for a discussion.

typically available in typical macroeconomic settings.[3] In this paper, we use an alternative approach which provides a practical and general solution to a simpler problem.

Instead of the full forecast density, we work with the implied probabilistic forecasts: we compute the forecast probabilities of failing to achieve the Bank's inflation target, or for GDP growth falling below a fixed threshold. (The methods that we use to do so can of course be applied to probability forecasts for other thresholds simply by integrating under different regions of the density forecast; different choices of threshold allow one to focus on different parts of the forecast distribution.) This is similar in spirit to Clements' (2004) examination of the fan chart's implied interval forecasts.[4] We are able to investigate the calibration of the probabilistic forecasts, that is, the degree to which predicted probabilities correspond with true probabilities of the outcome. We are also able to evaluate their resolution: their ability to discriminate among different outcomes. We also build on Galbraith and van Norden (2008) by extending their tests of forecast calibration to tests of forecast resolution.

In contrast with earlier evaluations, our results provide strong evidence of a miscalibration of the inflation forecasts at very short horizons, even though the degree of miscalibration appears to be small. Despite the much shorter sample available for the GDP forecasts, we again find significant evidence of mis-calibration and its magnitude appears to be much larger than for inflation. Results on the discriminatory power of the forecasts shows that inflation forecasts appear to have important power to distinguish high- and low- probability cases up to horizons of about one year, while that of GDP forecasts is much less and is almost negligible beyond a one-quarter horizon.

The next section of the paper introduces the Bank of England's probability forecasts and provides an informal analysis of descriptions of the data that we use. We then review the literature on methods for evaluating probability forecasts, introducing the decomposition of mean-squared forecast errors into calibration errors and forecast resolution. We briefly discuss the methods introduced in Galbraith and van Norden (2008) for tests of calibration error and their extension to tests of zero forecast resolution. The penultimate section of the paper then applies these methods to the Bank of England data and discusses the findings. The final section concludes.

2. Data and forecasts

The Bank of England's Inflation Report provides probabilistic forecasts of inflation and, more recently, output growth in the form of 'fan charts' to represent the density of

[3]For example, consider the problem of determining the effect of variations in the right tail of $g(x)$ on the right tail of $f(x)$. Only observations in which there is variation in the relevant region of $g(x)$ will be informative for this problem. However, since we only observe the outcomes x rather than $f(x)$ directly, we will need to have many of the above observations on x in order to make inferences about the tail region of $f(x)$.

[4]The Bank of England also examines such interval forecasts from time to time. For example, see Table 1 (p. 47) of the August 2008 Inflation Report.

the forecast distribution.[5] Fan charts for RPIX inflation were published from 1993Q1 to 2004Q1, when they were replaced by CPI Inflation fan charts. Both measure inflation as the percentage change in the corresponding price index over four quarters. The GDP fan chart was first published in the 1997Q3 report and also forecasts the total percentage growth over 4 quarters. 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, but we will nonetheless present most results below for both sequences of density forecasts.

For both inflation and GDP growth, we use all available forecasts up to and including that published in 2008Q2. We also measure inflation and output growth outcomes using the 2008Q2 vintage data series. While a few sets of forecast densities are available at horizons exceeding eight quarters, the sample sizes involved are small. We therefore report results for nine horizons, zero (the 'nowcast' of the eventual current-quarter release) through eight.[6] With the four cases noted above (GDP growth and inflation, assuming interest rates are constant or follow market expectations) we then have thirty-six sets of density forecasts for evaluation.

While published charts provide a visual guide to the degree of uncertainty that the Bank of England (BoE) Monetary Policy Committee (MPC) 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 (2003), among others. Forecast errors are assumed to follow a 'two-piece normal' or 'bi-normal' distribution, whose behaviour is completely characterized by three parameters: a mean μ, a measure of dispersion σ, and a parameter which controls skewness, γ.[7] These parameters therefore allow us to estimate the implied forecast probabilities that inflation or GDP growth would exceed any given threshold level.[8] For GDP forecasts, we examine the probability that annual real growth is less than 2.5%, while for inflation we examine

[5]See Wallis (1999) for a careful discussion of the interpretation of these charts; note in particular that the different bands do not correspond straightforwardly with quantiles in the general, asymmetric, case.

[6]The figures below also show some estimates based on longer horizons, although the sample sizes available are small.

[7]See Wallis (2003, Box A on p. 66) for a description of the bi-normal distribution and its alternative parameterizations. Spreadsheets containing the parameter settings for all of the published fan charts are publicly available on the BoE's web site (presently at http://www. bankofengland.co.uk/publications/inflationreport/irprobab.htm). Note that in the latter part of our sample, γ is commonly set to 0, in which case the implied forecast distribution is simply Gaussian.

[8]Like the normal distribution, the bi-normal lacks an exact closed-form formula for its cumulative distribution function (CDF). When $\gamma \neq 0$, we therefore estimated the im-

the probability that it does not exceed the Bank's announced inflation target.

2.1 Density forecasts

Figures 1 and 2 plot the results of the probability integral transforms for each of the sets of forecasts. As these are U(0,1) under the null of correct specification of the conditional density, the histograms should show roughly the same proportion of observed forecasts in each of the ten cells. In order to represent the results for nine forecast horizons (0–8 inclusive) in each figure, we have indicated the height of each histogram with a colour coding; each row of the figure represents a different horizon, and each column a particular bin with width 0.1. Uniformly distributed results would imply a frequency of 0.1 in each bin, and therefore a uniform colour in the figure. Values well below 0.1 show up as dark blue, and well above 0.1 as red.

While some sampling variation is of course inevitable, these patterns are in general far from conformity with this condition. GDP growth forecasts (Figure 1) often show an excessive number of values in the highest cell (near 1) at short horizons, an insufficient number at long horizons, and an insufficient number of values near zero at virtually all horizons. Inflation forecasts (Figure 2) show better conformity with the desired pattern of uniformity, but some of the same tendency is observable. Note that an insufficient number of values of the probability integral transform near the extremes is an indication of forecast densities that are too dispersed: actual outcomes occur near the tail of the forecast density less often than would arise with the true conditional density, and therefore observed outcomes tend to be in intermediate regions of the relevant CDF.

2.2 Threshold probability forecasts

Figures 3 and 4 show the implications of the BoE's density forecasts for the probabilities that real GDP is less than the 2.5% threshold we mentioned above and that inflation (based on RPIX or CPI) is less than the Bank's target. Each point corresponds to the probability (on the vertical axis) that inflation or output growth would be less than the chosen threshold at the given forecast horizon (given on the horizontal axis.) The larger (green) dots represent cases in which the eventual outcome was below the relevant threshold, while the smaller (blue) dots are cases in which the outcome was above threshold.

Ideal forecasts would have assigned probability one to all the large green dots and probability zero to the smaller blue dots. Instead, for GDP growth we observe several high probability blue dots and low probability green dots (see horizons 2-4 in particular) which indicate "surprises." We also find most outcomes clustered in the center of the probability range at horizons 4-8. The probabilistic outcomes for inflation show similar

---

plied forecast probabilities by numerical integration from probability density function. The resulting CDF estimates appeared to be accurate to at least 0.001. When $\gamma = 0$ the normal CDF was used. Additional details and code are available from the authors upon request.

features with respect to changes across horizons, but at the shorter horizons we see a more marked concentration of large green dots at the higher probabilities and small blue dots at the lower, suggesting that the inflation forecasts had more discriminatory power than the GDP forecasts, at least at the short horizons.

While the results in the graph presented this far are suggestive, we would like to be able to test for systematic problems in the probability forecasts and their ability to discriminate between different outcomes. In the next section, we review some of the literature on density forecast evaluation before focusing on tests of probabilistic forecasts and properties of forecast calibration and resolution.

3. Probability forecast evaluation

3.1 Predictive density evaluation

Let X be a random variable with realizations x_t and with probability density and cumulative distribution functions $f_X(x)$ and $F_X(x)$ respectively. Then for a given sample $\{x_t\}_{t=1}^T$, the corresponding sample of values of the CDF, $\{F_X(x_t)\}_{t=1}^T$, is a U(0,1) sequence. This well-known result (often termed the probability integral transform of $\{x_t\}_{t=1}^T$) is the basis of much predictive density testing, following pioneering work by Diebold, Gunther and Tay (1998).

These authors noted that if the predictive density $\hat{f}_X(x)$ is equal to the true density, then using the predictive density for the probability integral transform should produce the same result, i.e. a U(0,1) sequence. This allows us to test whether a given sequence of forecast densities could be equal to the true sequence by checking whether $\{\hat{F}_X(x_t)\}_{t=1}^T$ (i.e. the sequence of CDFs of the realized values using the forecast densities) is U(0,1). This is precisely the relationship that we looked for in Figures 1 and 2, above.

If this sequence is assumed to be independent, the U(0,1) condition is easily tested with standard tests (such as a Kolmogorov-Smirnov one-sample test.) The independence is unrealistic in many economic applications, however. In particular, violation is almost certain for multiple-horizon forecasts as the $h-1$ period overlap in horizon-h forecasts induces an MA$(h-1)$ process in the forecast errors. The inferential problem is therefore more difficult: test statistic distributions are affected by the form of dependence.

3.2 Probabilistic forecasts

Rather than analyse the entire predictive density, we instead examine the probabilistic forecasts implied by the BoE forecasts; that is, we only consider the probability that an outcome (inflation or output growth) will be below some threshold. This implies a loss of information relative to the full density forecast. Of course, if we are primarily concerned with the behaviour of our forecasts around these thresholds, this loss of efficiency may be inconsequential. This seems to be the case for the BoE forecasts and the

thresholds we have chosen; considerable attention is devoted to questions of whether or not central banks will respect their inflation targets, and whether output growth will be slightly above or below its mean.[9] As we show now, probabilistic forecasts also permit a particularly simple decomposition that is useful for interpreting forecast behaviour and the sources of forecast errors.

Following 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. Forecasts and outcomes may both be seen as random variables, and therefore as having a joint distribution; see e.g. Murphy (1973), from which much subsequent work follows.

Numerous summary measures of probabilistic forecast performance have been suggested, including loss functions such as the Brier score (Brier, 1950) which is a MSE criterion. Since the variance of the binary outcomes is fixed, it is useful to condition on the forecasts: in this case we can express the mean squared error $E((\hat{p}-x)^2)$ of the probabilistic forecast as follows:[10]

$$E(\hat{p}-x)^2 = E(x-E(x))^2 + E_f(\hat{p}-E(x|\hat{p}))^2 - E_f(E(x|\hat{p})-E(x))^2, \tag{3.1}$$

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 depends on the forecasts is contained in the second and third terms on the right-hand side of (3.1).

3.3 Calibration and resolution

We will call the first of the terms involving $\hat{p}$ in (3.1),

$$E_f(\hat{p}-E(x|\hat{p}))^2, \tag{3.2}$$

[9]Particularly in light of recent events, some might argue that central banks should be more attentive to the possibility of extreme drops in output growth. We note that the methods that we use below can in principle be applied to this question as well by simply choosing a different threshold level of output growth. Given the relative short sample over which the GDP growth forecasts are available, however, it is presumably difficult to say much about the probability of a downturn as severe as that witnessed in late 2008 with what is essentially a single realization.

[10]The MSE is of course only one of many possible loss functions, and is inappropriate in some circumstances. We focus on it here because there is no concensus on the precise form of an appropriate loss function for an inflation-targetting central and because we argue that the decomposition it presents is helpful in understanding forecast performance.

the (mean squared) *calibration error:* it measures the deviation from a perfect match between the predicted probability and the true probability of the event when a given forecast is made.[11] 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. 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.

Calibration has typically been investigated using histogram-type estimates of the conditional expectation, grouping probabilities into cells. Galbraith and van Norden (2008) show how to use smooth conditional expectation functions estimated via kernel methods to estimate calibration functions and test for miscalibration. They show that this allows one to correct for the dependence caused by overlapping forecast windows and leads to an efficiency relative to histogram methods even in the absence of dependence. We use these methods (described in detail in that paper) below to further examine the performance of the BoE inflation and growth forecasts.

The last term on the right-hand side of (3.1), $E_f(E(x|\hat{p}) - E(x))^2$, 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). In this special case, the probability forecasts are always 1 when the outcome will be below the threshold, and are 0 otherwise.

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

[11] 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.

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 random variable with mean μ. Therefore

$$\frac{E_f(E(x|\hat{p}) - \mu)^2}{\mu(1-\mu)^2 + (1-\mu)\mu^2} \in [0,1]. \tag{3.3}$$

The information in the resolution is correlated with that in the calibration; the decomposition just given is not an orthogonal one (see for example Yates and Curley 1985). However the resolution also has useful interpretive value which we will see below in considering the empirical results. The calibration and/or resolution of probabilistic economic forecasts have been investigated by a number of authors, including Diebold and Rudebusch (1989), Galbraith and van Norden (2007), and Lahiri and Wang (2007). The meteorological and statistical literatures contain many more examples; some recent contributions include Hamill et al. (2003) and Gneiting et al. (2007). We now use these methods to examine the BoE forecasts.

4. Empirical results

Figures 5 and 6 plot the estimated conditional expectation of outcome given forecast, which for correctly calibrated forecasts would lie along the line $E(x|\hat{p}) = \hat{p}$, i.e. the 45 degree line. Again, see Galbraith and van Norden (2008) for a description of the methods used to construct these estimates, including bandwidth choice; all results depicted here use bandwidth 0.08.

Figure 5 shows these conditional expectations for the GDP growth forecasts at each forecast horizon. Deviations from perfect calibration (the 45-degree line) are widespread and often large. Most strikingly, *high* predicted probabilities of growth falling below threshold correspond to *low* observed frequencies of this outcome for virtually all horizons. Calibration is much more reasonable for probabilities in the 0-0.5 range, but falls dramatically beyond predicted probabilities of around 0.8. Calibration for the small sample of very long horizon forecasts appears very poor, with a strongly negative slope everywhere. In contrast, calibration of the inflation forecasts appears more reasonable at all predicted probabilities, even for the small sample of longer-horizon forecasts available in the market interest rate case. The corresponding numerical mean squared calibration errors are presented in Table 1.

One problem in interpreting these graphs and numerical estimates is that the precision of the estimated conditional expectations may vary widely across the graph. This makes it difficult to judge whether any of the deviations from the 45 degree are statistically significant. For that, we turn to formal tests of the null hypothesis of correct calibration (that is, $E(x|\hat{p}) = \hat{p}$) given in Table 2. At short horizons, correct calibration is decisively rejected for both GDP growth and inflation forecasts. At longer

horizons the results are mixed, with differences emerging between the fixed interest rate and market interest rate forecasts. Despite the apparently dramatic mis-calibration visible in the figures for GDP forecasts, we are often unable to reject the null of correct calibration for the market-interest-rate forecasts. This may be related to the relatively small number of sample points lying at high forecast probabilities, which would give the test low power to detect such deviation. Alternatively, it could be due to the highly non-linear evidence of miscalibration shown in the graphs. Our test is based on a linear alternative hypothesis and so may have reduced power against some non-linear alternatives. In contrast, the graphs for inflation showed that the forecast calibration was much more nearly linear, which should give our tests higher power. This may therefore explain the relative high number of rejections of the null despite the relatively better fit shown in the inflation graphs.

Figures 5 and 6 also provide us with information on the resolution of the BoE forecasts. A forecast with zero resolution will have a constant $E(x|p)$, so the conditional expectation should simply be a horizontal line. While it is conceivable that some of the results for the GDP growth forecasts shown in Figure 5 could be consistent with such an outcome (such as the results for the constant interest rate forecasts at 6-8Q horizons) this appears much less likely for the inflation forecasts, which all show a persistently positive slope.

Figures 7 and 8 provide another way of understanding the resolution of the probability forecasts which does not require estimation of the conditional expectation. For each horizon (only horizons up to four quarters are shown), each figure presents a pair of empirical CDF's: that of probability forecasts in cases for which the eventual outcome was below threshold, and in cases in which the eventual outcome was above threshold. In a near-ideal world, probability forecasts should be near one in cases where the outcome turned out to be below threshold, and near zero when the outcome turned out to be above. In that case the two CDF's would lie close to the lower horizontal axis in the first case, and close to the upper horizontal axis in the second. More generally, good probability forecasts will discriminate effectively between the two possible outcomes, and the two empirical CDF's should be widely separated. At longer horizons, the value of conditioning information declines and this separation becomes more difficult to achieve; we therefore expect to see the pairs of CDF's less widely separated at longer horizons.

This pattern of reduced separation with horizon is in fact readily observable; at 3-4 quarter horizons, we observe little separation on either forecast series. However, at shorter horizons, we observe a clear distinction between the GDP growth and inflation forecasts; separation is much greater in the inflation forecast case (Figure 8, both panels; there is little observable distinction between the fixed and market interest rate cases), suggesting much higher forecast resolution. GDP growth forecasts (Figure 7) in fact show little separation of the CDF's after the shortest horizons.[12] However, for GDP forecasts there is some observable distinction between the fixed and market interest

[12]This result which mirrors the low 'content horizon' on U.S. and Canadian GDP

rate cases; the fixed cases show somewhat higher resolution in at short horizons.

These differences are confirmed by the numerical results on scaled resolution presented in Table 3. Scaled resolution (recall that this estimate is bounded to the [0,1] interval) in GDP growth forecasts is low even at short horizons, and approximately zero at moderate and long horizons; by contrast, inflation forecasts show substantial resolution for several quarters. The market-interest-rate forecast for inflation shows the highest resolution.

The test results for the null hypothesis of zero resolution (Table 4) are again compatible with these observations. These are simple $t-$ type tests of $H_0 : b = 0$ in $E(x|\hat{p}) = a + b\hat{p}$; robust (Newey-West) standard errors are used in the computation. Table 4 reports p-values from the asymptotic normal distribution applying to these test statistics. For inflation forecasts, there is strong evidence against the null in all but one (horizon 8) case, whereas for GDP forecasts we typically cannot reject zero resolution (the shortest-horizon fixed-interest forecasts provide the only exception). These results also reflect other results in the literature mentioned above, which note the difficulty of the GDP growth forecasting problem and the associated low information content of such forecasts, by various measures, relative to inflation forecasts.

6. Concluding remarks

By focusing our attention on particular probabilities derived from the fan charts, we have evaluated the performance of these density forecasts in problems of the type likely to be of interest to forecast users, as opposed to a general evaluation of whether the fan charts represent correct conditional densities; it is of course possible that a forecast density differs in some respects from the conditional density while nonetheless producing probability forecasts with some valuable features.

A number of interesting empirical results emerge. First, it is apparent even from the preliminary graphical results presented here that the predicted densities do not entirely conform with the true conditional densities. We then map these densities onto particular probabilistic forecasts for evaluation of the calibration and resolution. For inflation forecasts, deviations from correct calibration appear to be small, although nonetheless statistically significant at a number of forecast horizons. GDP growth forecasts produce much larger estimated deviations from correct calibration: that is, predicted probabilities of GDP falling below our threshold are in many cases far from

---

growth point forecasts reported by, for example, Galbraith 2003 and Galbraith and Tkacz 2007: that is, forecasts of GDP growth generally do not improve markedly on the simple unconditional mean beyond about one or two quarters into the future. Galbraith and van Norden (2008) estimate conditional expectation functions for the Survey of Professional Forecasters probabilistic forecasts for US real output contractions using methods very similar to those used here. They find forecasts are appear to be essentially unconditional forecasts at horizons of more than 2Q, which implies zero forecast resolution.

our estimates of the true conditional probabilities. However, only a subset of the observed deviations are statistically significant, perhaps because of the limited sample size.

Resolution falls rapidly with forecast horizon, is higher for inflation forecasts, and for GDP is in most cases difficult to distinguish statistically from zero.

These results, particularly at longer horizons, reflect differences in inflation and GDP growth observed in other contexts: the usefulness of conditioning information allowing us to make forecasts appears to decay much more quickly for GDP growth, and the persistence in the data is much lower. Whether sufficient information exists to produce forecast densities for GDP growth which differ from the unconditional densities, at the longer horizons used by the Bank of England, is an open question.

References

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. (2004) "Evaluating the Bank of England density forecasts of inflation." *Economic Journal* 114, 844-866.

Corradi, V. and N. Swanson (2006) "Predictive density evaluation." in Elliott, G., C. Granger and A. Timmerman, eds., *Handbook of Economic Forecasting,* North-Holland, Amsterdam.

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.

Dowd, K. (2008) "The GDP fan charts: an empirical evaluation." *National Institute Economic Review* 203, 59-67.

Elser, R., G. Kapetanios, T. Taylor and T. Yates (2005) "Assessing the MPC's fan charts." *Bank of England Quarterly Bulletin,* 326-348.

Galbraith, J.W. (2003) "Content horizons for univariate time series forecasts". *International Journal of Forecasting* 19, 43-55.

Galbraith, J.W. and S. van Norden (2008) "The calibration of probabilistic economic forecasts." Working paper, CIRANO.

Galbraith, J.W. and G. Tkacz (2007) "Forecast content and content horizons for some important macroeconomic time series." *Canadian Journal of Economics* 40, 935-953.

Gneiting, T., F. Balabdaoui and A.E. Raftery (2007) "Probabilistic forecasts, calibration and sharpness." *Journal of the Royal Statistical Society Ser. B* 69, 243-268.

Hamill, T.M., J.S. Whitaker and X. Wei (2003) "Ensemble reforecasting: improving medium- range forecast skill using retrospective forecasts." *Monthly Weather Review* 132, 1434–1447.

Lahiri, K. and J.G. Wang (2007) "Evaluating probability forecasts for GDP declines." Working paper, SUNY.

Mitchell, J. and K. F. Wallis (2009) "Evaluating Density Forecasts: Forecast Combinations, Model Mixtures, Calibration and Sharpness." Presented at the Conference in Honour of Adrian Pagan, Sydney, July 2009.

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.

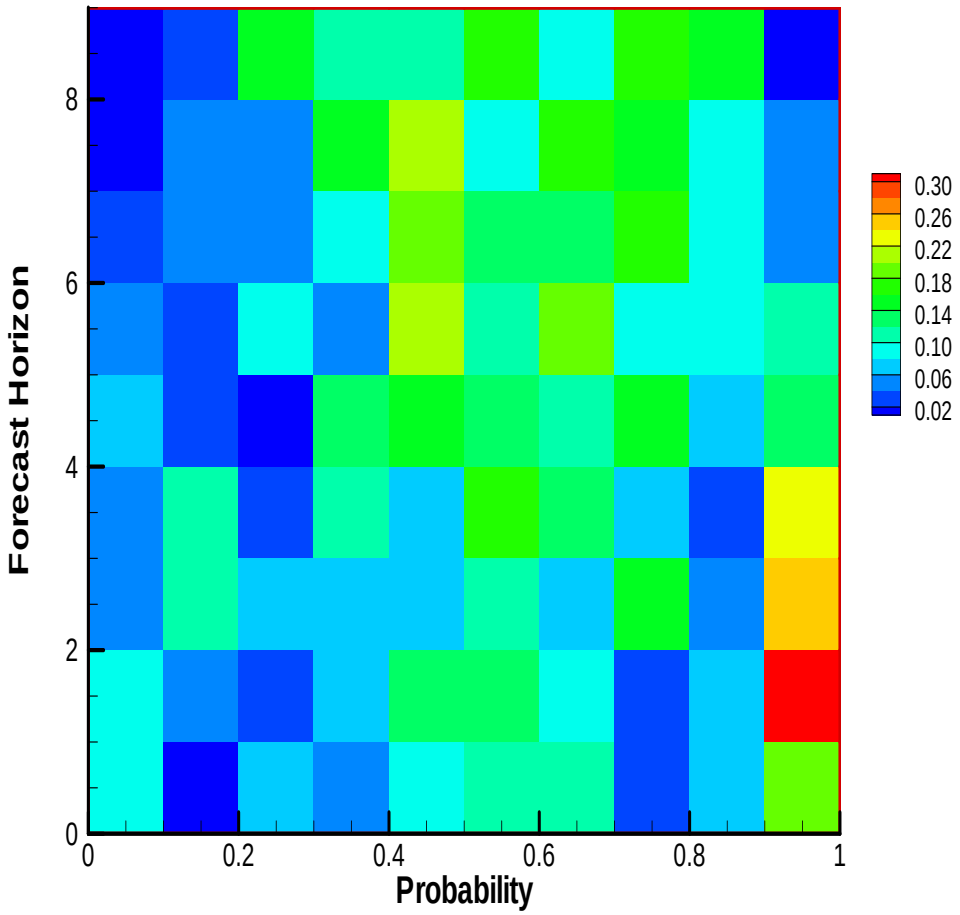
Orphanides, A. and S. van Norden (2005) "The reliability of inflation forecasts based on output gap estimates in real time." *Journal of Money, Credit and Banking* 37, 583-601.

Rudebusch, G. and J.C. Williams (2007) "Forecasting recessions: the puzzle of the enduring power of the yield curve." Working paper, FRB San Francisco.

Wallis, K.F. (1999) "Asymmetric density forecasts of inflation and the Bank of England's fan chart." *National Institute Economic Review* 167, 106-112.

Wallis, K.F. (2003) "An Assessment of Bank of England and National Institute Inflation Forecast Uncertainties." *National Institute Economic Review* 198, 64-71.

FIGURE 1(I)
Probability integral transforms[13]
GDP growth vs. threshold; fixed rates



[13]In each of the panels of figure 1 and 2, the ten columns of squares represent the bins 0–0.1, 0.1–0.2, ... 0.9–1.0 and the nine rows of squares represent the horizons 0-8. Each square represents, via colour, the height of a histogram corresponding with the horizon and bin. Ideally, the probability integral transforms would yield a U(0,1) sequence, so that each row would should uniform values equal to 0.1, and therefore uniform colour in this figure.

FIGURE 1(II)
Probability integral transforms
GDP growth vs. threshold; market rates

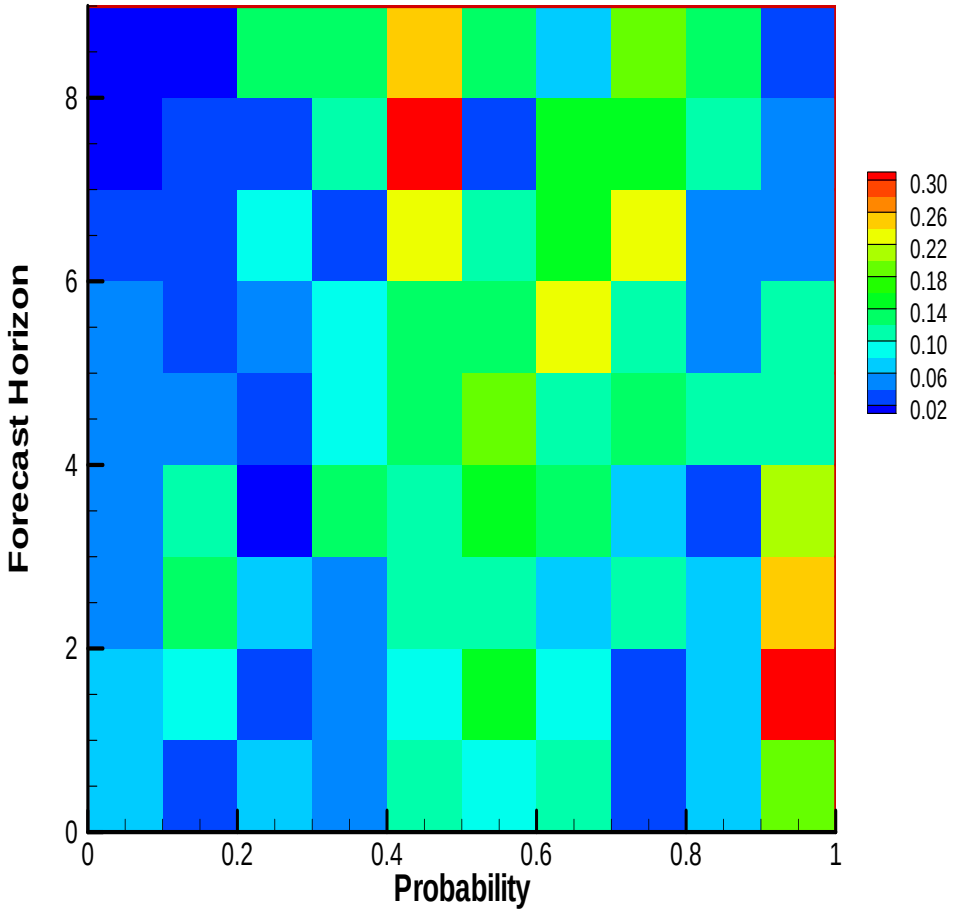


FIGURE 2(I)
Probability integral transforms
Inflation vs. threshold; fixed rates

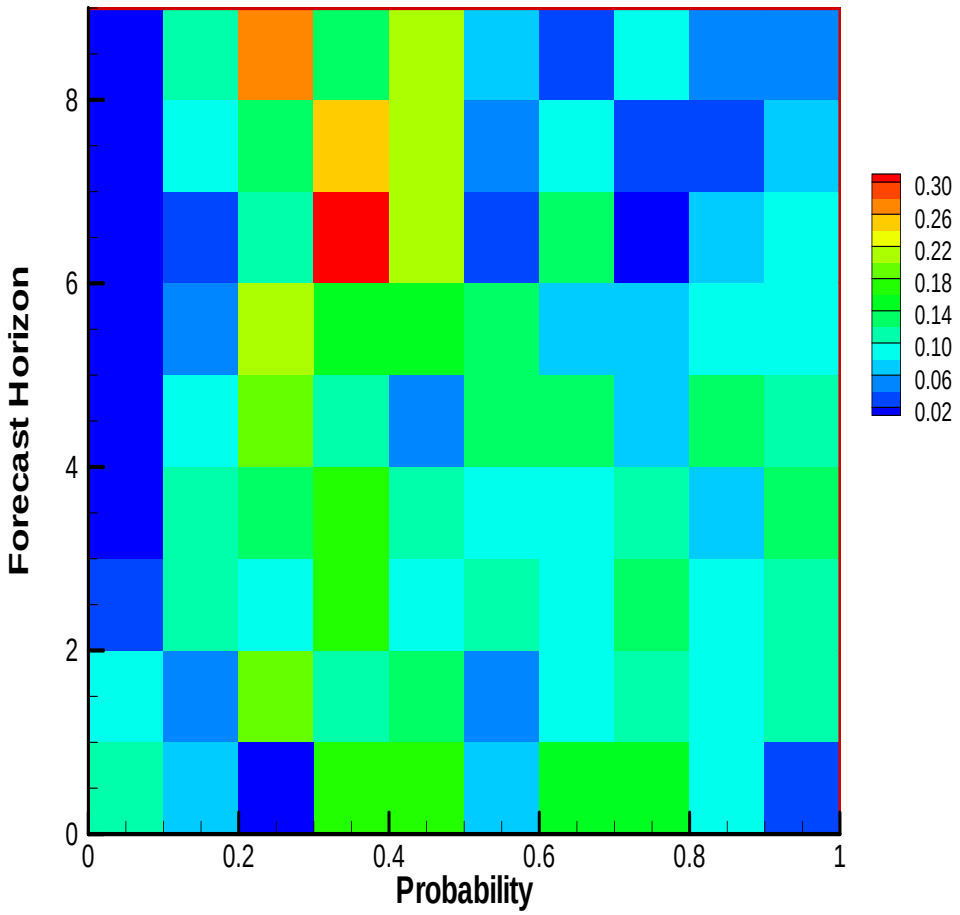
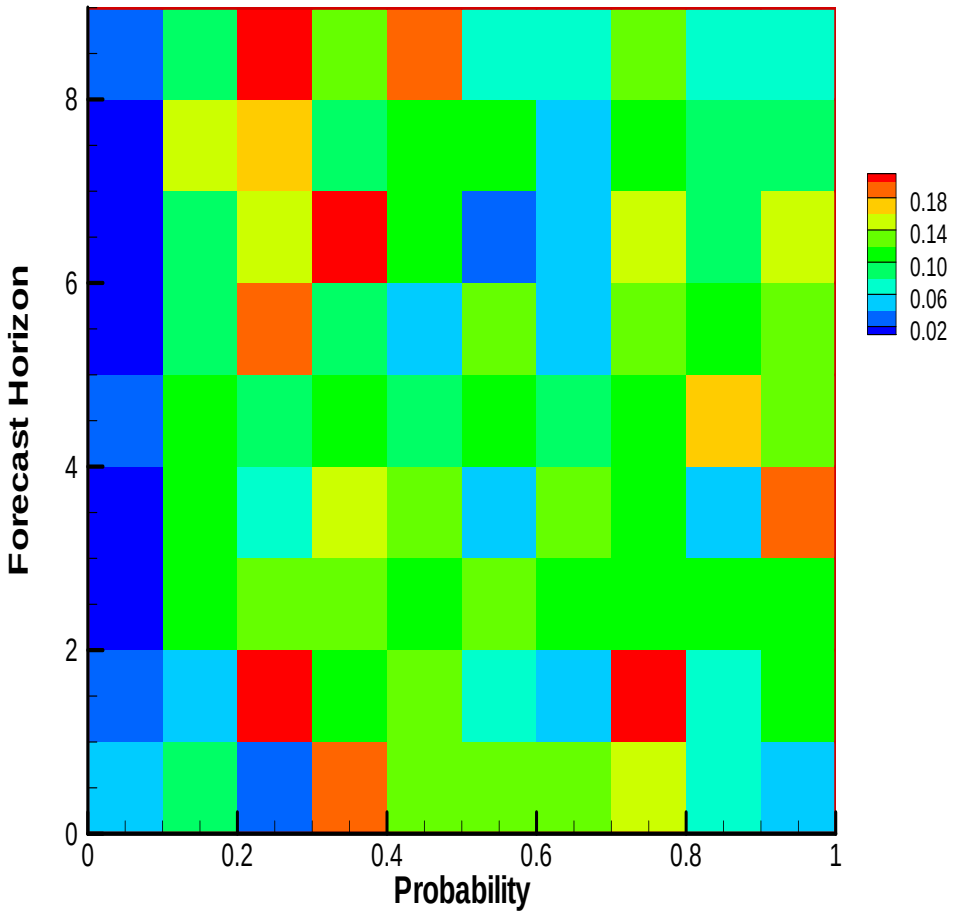


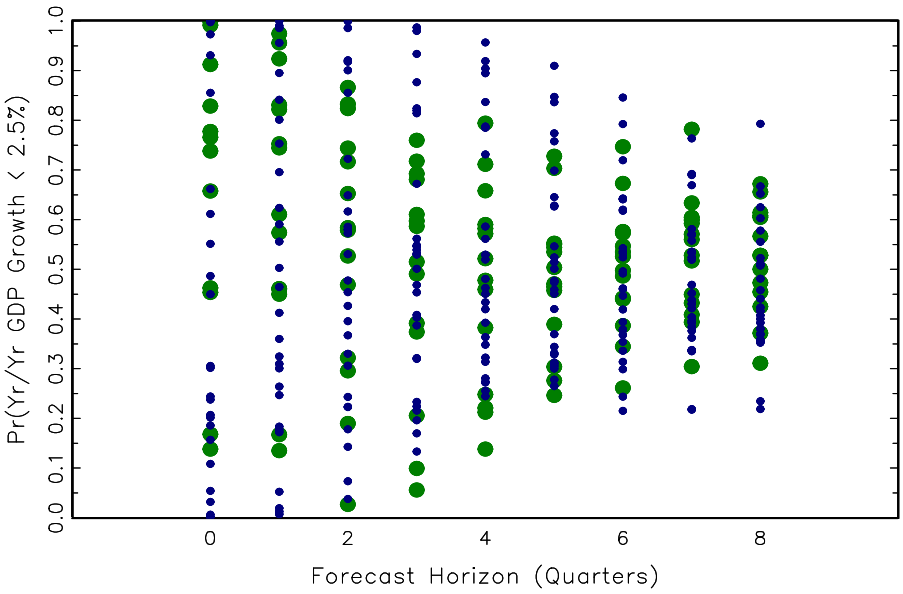
FIGURE 2(II)
Probability integral transforms[14]
Inflation vs. threshold; market rates



[14]Note that the colour scale differs substantially in Figure 2(ii); large values are less extreme than in previous figures.

FIGURE 3
Implied probability forecasts from Bank of England fan charts

(i) GDP growth vs. threshold; fixed interest rates



(ii) GDP growth vs. threshold; market interest rates

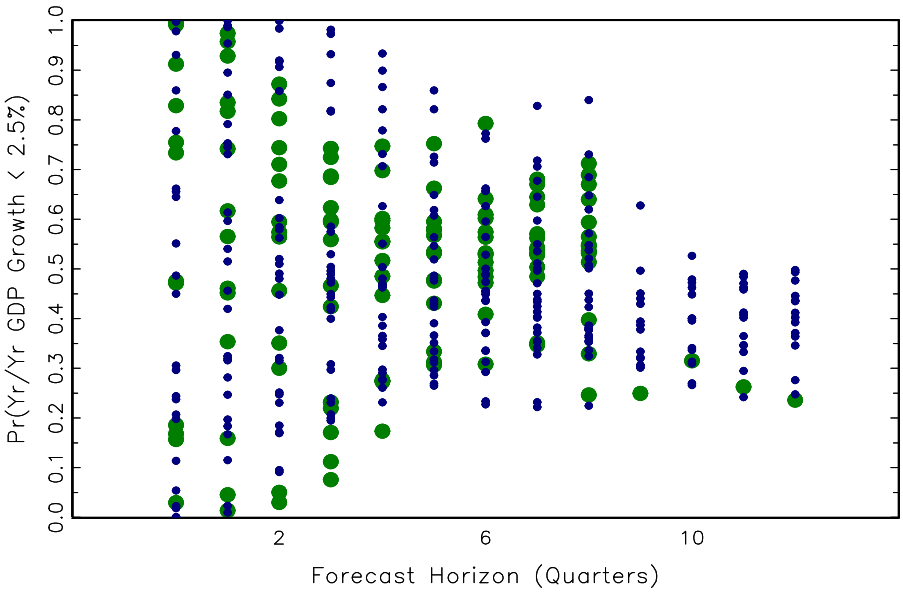
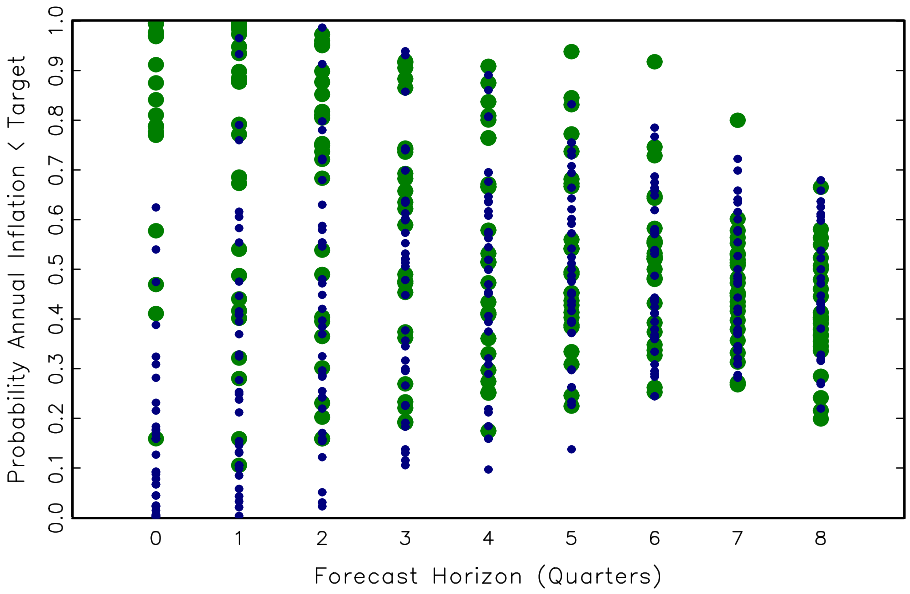


FIGURE 4
Implied probability forecasts from Bank of England fan charts

(i) Inflation vs. threshold; fixed interest rates



(ii) Inflation vs. threshold; market interest rates

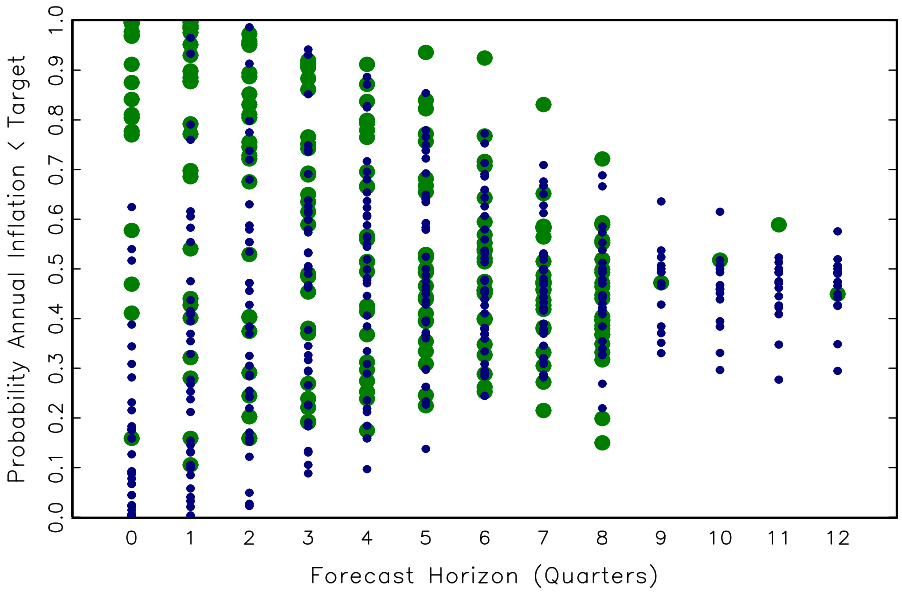
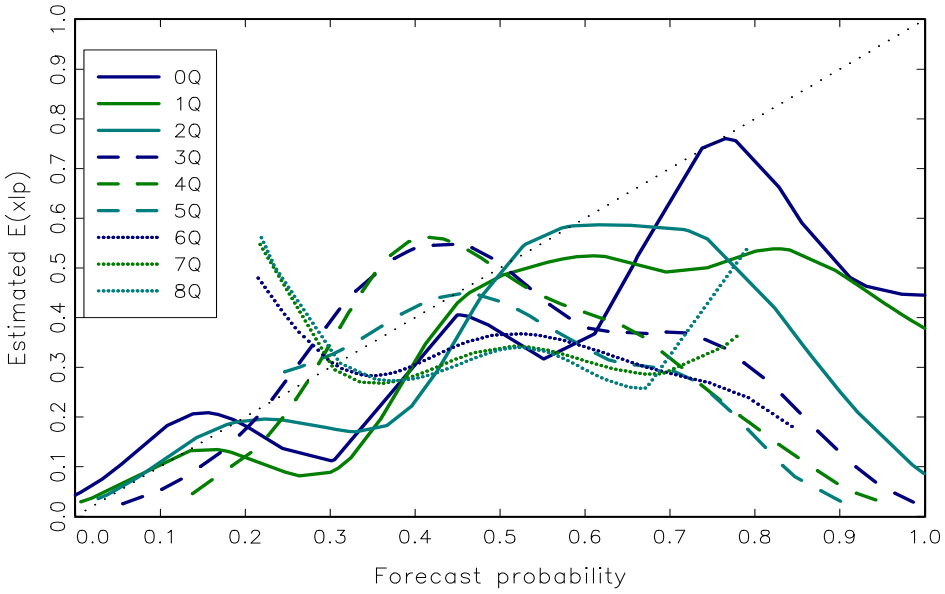


FIGURE 5
Calibration of GDP growth forecasts, 2.5% threshold

(i) 2008 Q2 vintage data, fixed interest rates



(ii) 2008 Q2 vintage data, market interest rates

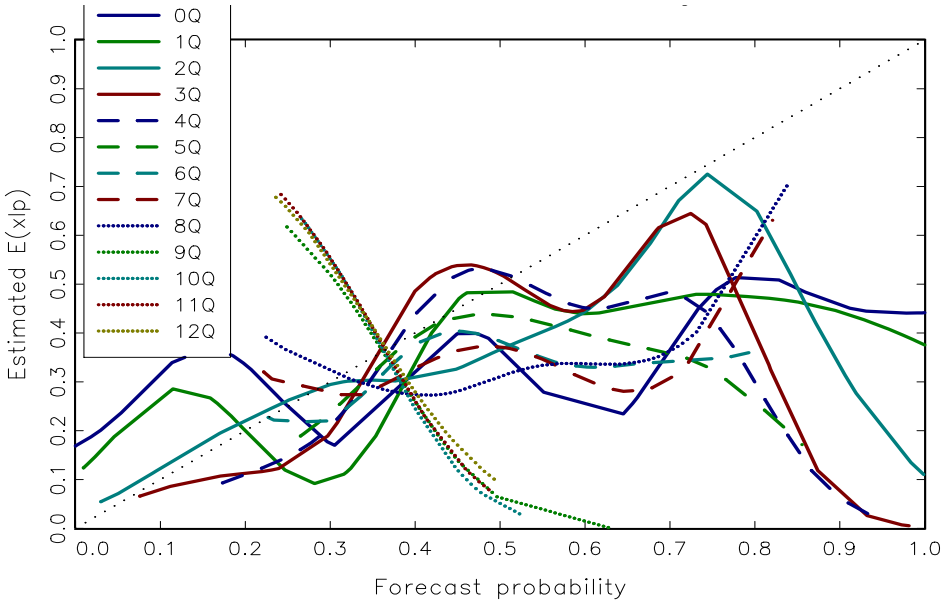
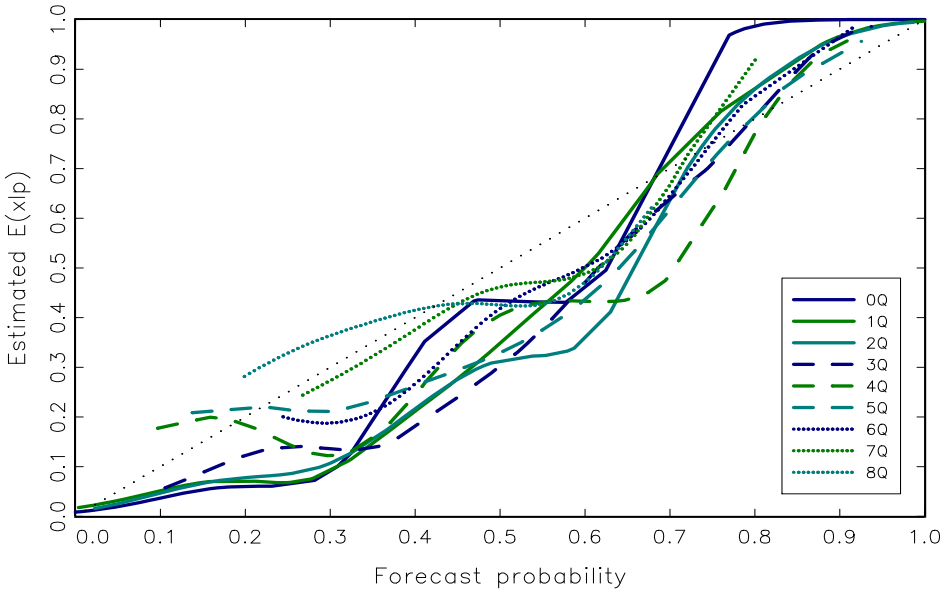


FIGURE 6
Calibration of inflation forecasts, target threshold

(i) 2008 Q2 vintage data, fixed interest rates



(ii) 2008 Q2 vintage data, market interest rates

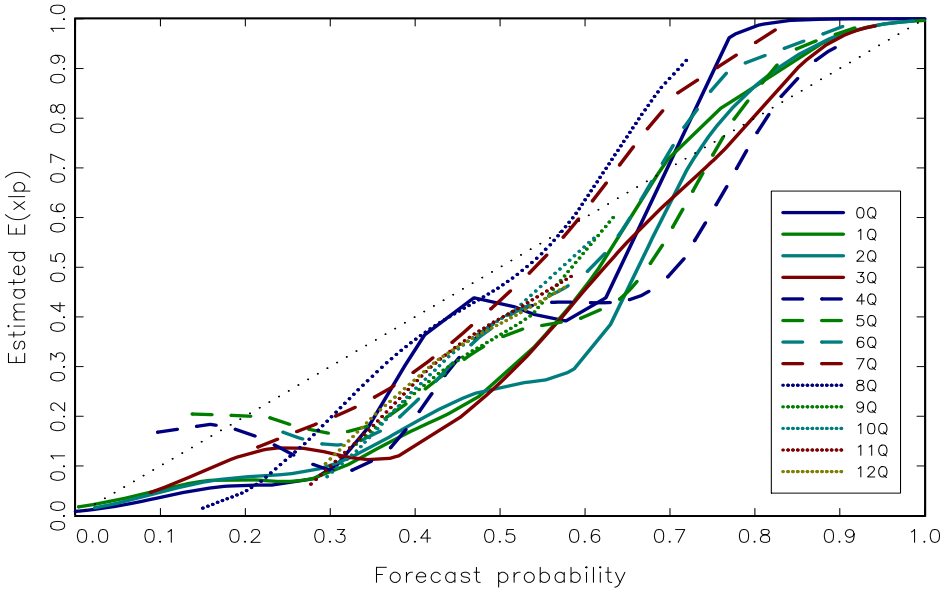
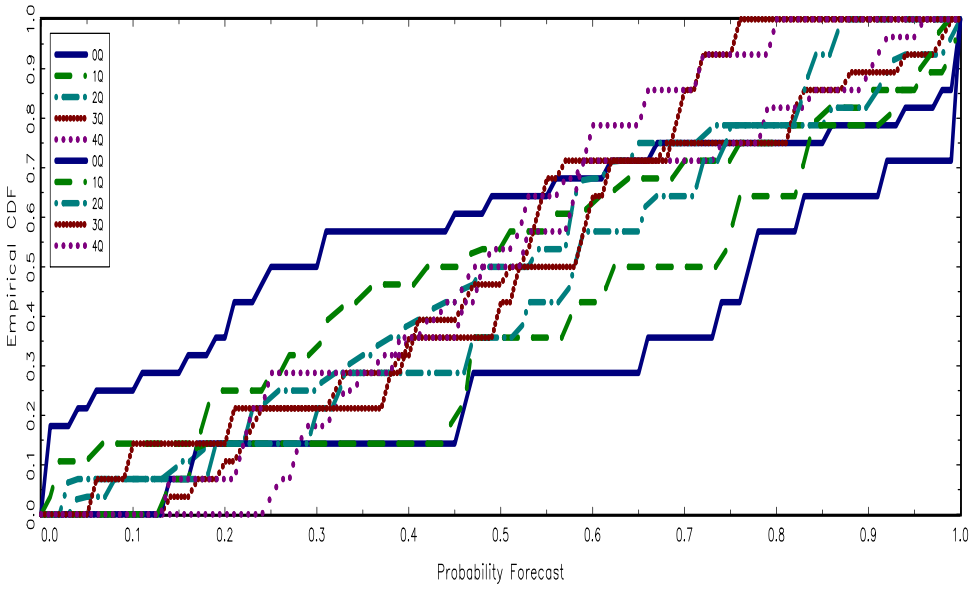


FIGURE 7
Empirical CDF's of implied probability forecasts from fan charts
Cases of GDP growth above/below threshold
(i) fixed interest rates



(ii) market interest rates

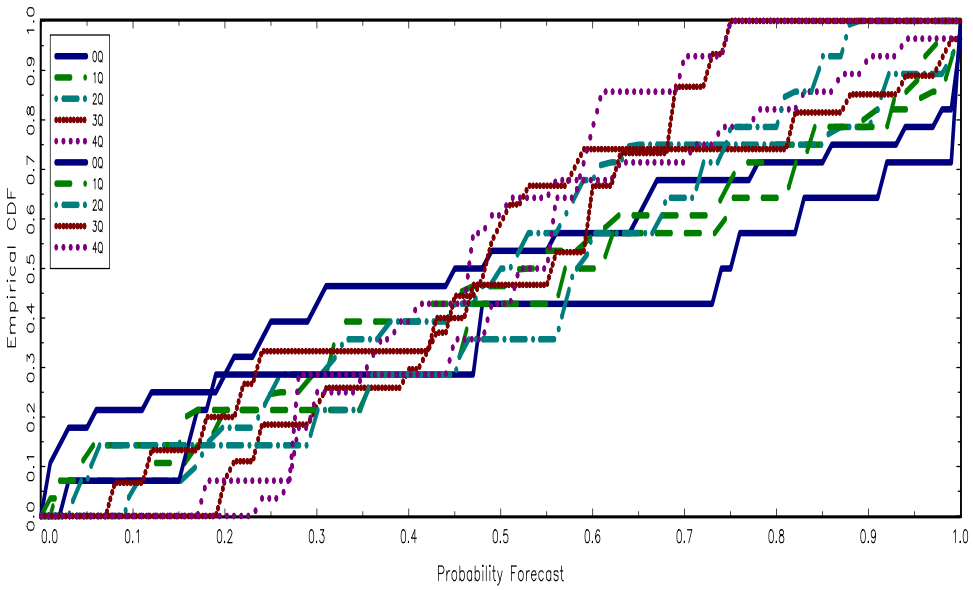
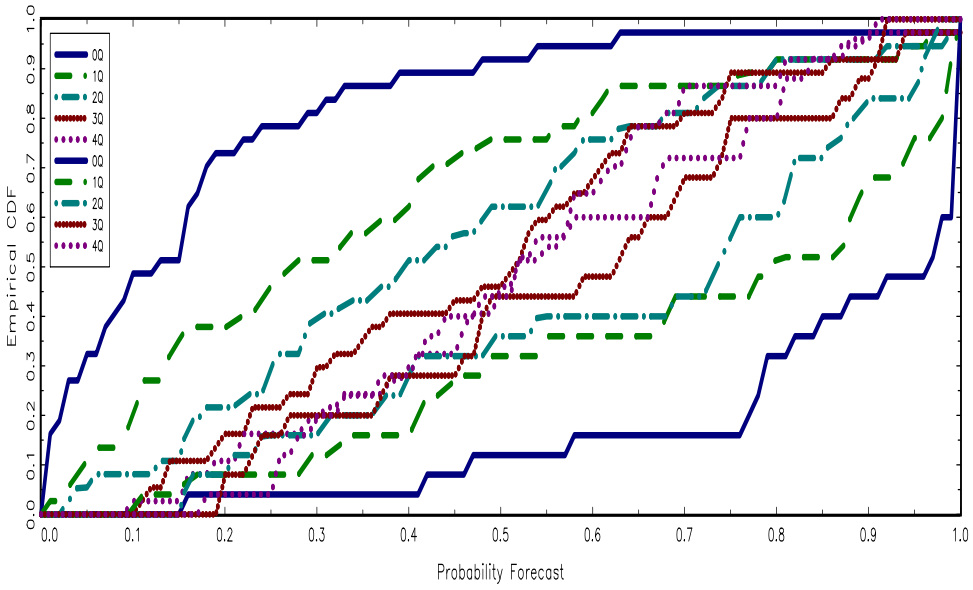


FIGURE 8
Empirical CDF's of implied probability forecasts from fan charts
Cases of inflation above/below threshold
(i) fixed interest rates



(ii) market interest rates

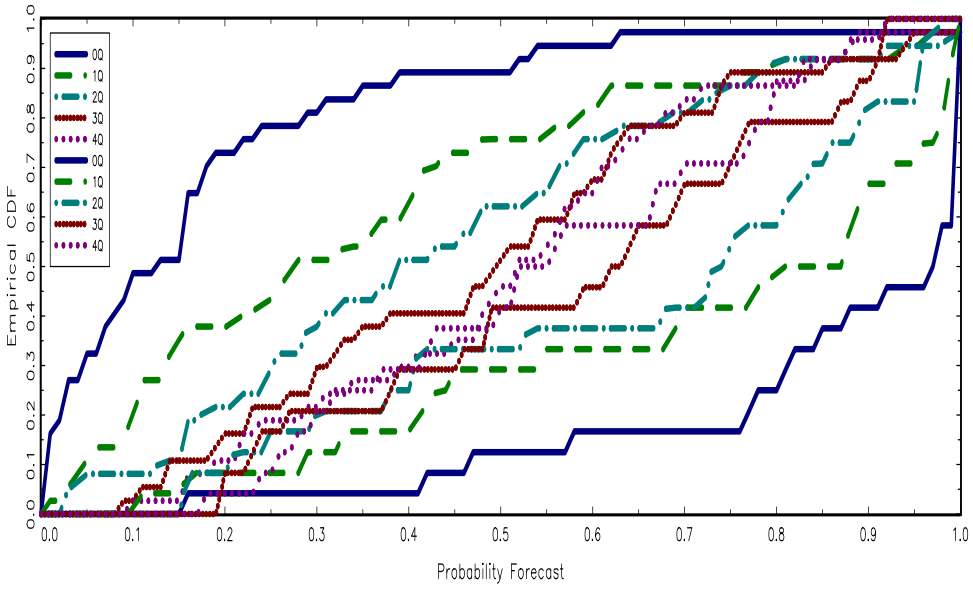


Table 1
Mean squared calibration errors

Horizon	GDP, fixed	GDP, market	Infl., fixed	Infl., market
0	0.066	0.127	0.008	0.009
1	0.071	0.088	0.015	0.012
2	0.107	0.085	0.017	0.014
3	0.134	0.127	0.017	0.014
4	0.151	0.106	0.020	0.018
5	0.124	0.059	0.018	0.017
6	0.069	0.039	0.013	0.011
7	0.052	0.033	0.005	0.006
8	0.045	0.026	0.002	0.008

Table 2
p-values in linear test of calibration[15]
$H_0 : a = 0,\ b = 1$ in $E(x|\hat{p}) = a + b\hat{p}$

Horizon	GDP, fixed	GDP, market	Infl., fixed	Infl., market
0	0.046	0.00	0.01	0.02
1	0.04	0.02	0.02	0.01
2	0.02	0.04	0.01	0.01
3	0.00	0.10	0.08	0.06
4	0.00	0.17	0.36	0.28
5	0.00	0.30	0.41	0.31
6	0.01	0.33	0.32	0.09
7	0.11	0.40	0.58	0.00
8	0.25	0.43	0.96	0.00

[15]A reported value of 0.00 in Table 2 indicates a computed p-value less than 0.005.

Table 3
scaled resolution measure ($\in [0,1]$)

Horizon	GDP, fixed	GDP, market	Infl., fixed	Infl., market
0	0.19	0.05	0.80	0.79
1	0.16	0.09	0.60	0.63
2	0.16	0.11	0.50	0.53
3	0.12	0.19	0.35	0.37
4	0.13	0.12	0.22	0.25
5	0.05	0.03	0.17	0.21
6	0.01	0.01	0.13	0.21
7	0.02	0.02	0.06	0.14
8	0.02	0.02	0.02	0.13

Table 4
p-values in linear test of zero resolution[16]
$H_0 : b = 0$ in $E(x|\hat{p}) = a + b\hat{p}$

Horizon	GDP, fixed	GDP, market	Infl., fixed	Infl., market
0	0.03	0.28	0.00	0.00
1	0.05	0.23	0.00	0.00
2	0.34	0.32	0.00	0.00
3	0.97	0.60	0.00	0.00
4	0.62	0.76	0.00	0.00
5	0.45	0.78	0.00	0.00
6	0.50	0.72	0.00	0.00
7	0.72	0.68	0.00	0.00
8	0.84	0.67	0.13	0.00

[16]A reported value of 0.00 in Table 4 indicates a computed p-value less than 0.005.