Diagnostics for Probability Forecasts Derived from the Bank of England Fan Charts

John W. Galbraith
Department of Economics, McGill University

Simon van Norden [*]
Finance, HEC Montréal

April 21, 2010

Abstract

Density forecasts, including the pioneering Bank of England 'fan charts', are often used to produce forecast probabilities of a particular event such as exceedance of a threshold. Here we compute probability forecasts implicit in the Bank's forecast densities, where the events are annual rates of inflation and output growth that exceed a given threshold (in this case, the target inflation rate and 2.5% respectively). We subject these implicit probability forecasts to a number of graphical and numerical diagnostic checks. Unlike earlier work on these forecasts, we measure both their calibration and their sharpness or resolution, providing both numerical and graphical interpretations of the results. The results reinforce earlier evidence on limitations of these forecasts and provide new evidence on their information content and on the relative performance of inflation and GDP growth forecasts.

Key words: calibration, density forecast, probability forecast, resolution, sharpness

[*] 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.

1. Introduction

Probability forecasts have attracted increasing recent interest; see Gneiting (2008) for a discussion and overview. While many series of such forecasts have been subject to careful evaluation (particularly in meteorological contexts), important economic series of probability forecasts have received less precise attention. The present paper addresses one well known source of such forecasts, arising from the Bank of England's fan charts, and subjects them to a number of diagnostic checks.

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 (gross domestic product) growth 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]

[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.

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 desirable, 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 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 (2009) by extending their tests of forecast calibration to tests of forecast resolution.

In contrast with earlier evaluations, our results provide strong evidence of a mis-calibration 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

[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.

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 (2009) 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 the forecast distribution. Fan charts for RPIX inflation were published from 1993Q1 to 2004Q1, when they were replaced by CPI Inflation fan charts. [5] 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, and below we will present results for the market interest rate case only.)

For both inflation and GDP growth, we use all available forecasts up to and including that published in 2010Q1. We also measure inflation and output growth outcomes using the 2010Q1 vintage data series. We report results for nine horizons, zero (the 'nowcast' of the eventual current-quarter release) through eight. 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 these charts provide only a visual guide to the degree of uncertainty that the Bank of England (BoE) associate with their forecasts, they are based on an explicit parametric model. Future outcomes 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

[5] RPIX: retail prices index excluding mortgage payments; CPI: consumer price index. 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.

controls skewness, γ.[6] These parameters therefore allow us to estimate the implied forecast probabilities that inflation or GDP growth would exceed any given threshold or fall within any given range. [7]

The bi-normal distribution has the convenient property that its cumulative distribution function (CDF) can be rewritten as a function of the standard normal CDF. Specifically, if x has a bi-normal distribution and ℓ_1, ℓ_2 are two points of evaluation, then $P(\ell_1 < x < \ell_2) =$

$$\begin{array}{ll} \frac{2\sigma_1}{\sigma_1+\sigma_2}\left[\Phi\left(\frac{\ell_2-\mu}{\sigma_1}\right)-\Phi\left(\frac{\ell_1-\mu}{\sigma_1}\right)\right] & \text{if } \ell_1 < \ell_2 < \mu; \\ \frac{2\sigma_2}{\sigma_1+\sigma_2}\left[\Phi\left(\frac{\ell_2-\mu}{\sigma_2}\right)-\Phi\left(\frac{\ell_1-\mu}{\sigma_2}\right)\right] & \text{if } \mu < \ell_1 < \ell_2; \\ \frac{2}{\sigma_1+\sigma_2}\left[\sigma_2\Phi\left(\frac{\ell_2-\mu}{\sigma_2}\right)-\sigma_1\Phi\left(\frac{\ell_1-\mu}{\sigma_1}\right)+\left(\frac{\sigma_1-\sigma_2}{2}\right)\right] & \text{if } \ell_1 < \mu < \ell_2, \end{array} \tag{1}$$

where $\Phi(.)$ is the CDF of the standard normal distribution. Note that in the special case where $\sigma_1 = \sigma_2$, this simply reduces to $\Phi\left(\frac{\ell_2-\mu}{\sigma}\right)-\Phi\left(\frac{\ell_1-\mu}{\sigma}\right)$, which is the usual expression arising for the normal.

Calculation of any probabilistic forecast implied by the Bank's density forecasts therefore simply requires the standard normal cumulative density function, the points of evaluation ℓ_1 and ℓ_2, and the parameters $\{\sigma_1, \sigma_2, \mu\}$.

The Bank of England publishes five values (mode, median, mean, uncertainty and skew) for each of its density forecasts. The mode corresponds to μ, but σ_1 and σ_2 are only indirectly related to the remaining published values. Specifically, let σ be the published uncertainty and S be the published skew. Then define

$$\gamma = sgn(S)\left[1-4\left(\frac{\sqrt{1+\pi S^2}-1}{\pi S^2}\right)^2\right]^{\frac{1}{2}}, \tag{2}$$

with $sgn(S)$ denoting the sign of the skewness S; the parameters σ_1 and σ_2 can then be obtained as $\sigma_1^2 = \sigma^2/(1+\gamma)$ and $\sigma_2^2 = \sigma^2/(1-\gamma)$.

[6] See Brittan, Fisher and Whitley (1998) and Wallis (2003, particularly 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).

[7] The authors thank Rashmi Harimohan of the Bank of England for his technical assistance in reproducing the Bank of England's own probability forecasts.

2.1 Density forecasts

Figure 1 plots the results of the probability integral transforms for the Bank's GDP growth and inflation 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 part of the 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 (green) 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 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 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

Figure 2 shows the implications of the BoE's density forecasts for the probabilities that real GDP is less than the 2.5% threshold mentioned above, and that inflation (based on RPIX or CPI) is less than the Bank's target value. Note that the Bank has a scalar target value as well as a band within which inflation outcomes are considered acceptable; we present results based on each of these indicators below.

Each point in Figure 2 corresponds with the implied forecast probability (on the vertical axis) that inflation or output growth will be less than the chosen threshold, at forecast horizon given on the horizontal axis. The smaller (green) dots represent cases in which the eventual outcome was below the relevant threshold, while the larger (blue) dots are cases in which the outcome was above threshold.

Ideal forecasts would have assigned probability one to all the small green dots and probability zero to the larger 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 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 or sharpness.

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 Figure 1, 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. [8] 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:[9]

$$E(\hat{p} - x)^2 = E(x - E(x))^2 + E_f(\hat{p} - E(x|\hat{p}))^2 - E_f(E(x|\hat{p}) - E(x))^2, \quad (3)$$

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).

3.3 Calibration and resolution

We will call the first of the terms involving $\hat{p}$ in (3),

$$E_f(\hat{p} - E(x|\hat{p}))^2, \quad (4)$$

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

[8] 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.

[9] 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 consensus 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.

a given forecast is made. [10] 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 (2009) 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), $E_f(E(x|\hat{p}) - E(x))^2$, 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

[10] 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.

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 random variable with mean μ. Therefore

$$\frac{E_f(E(x|\hat{p}) - \mu)^2}{\mu(1-\mu)^2 + (1-\mu)\mu^2} \in [0,1]. \tag{5}$$

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 (2009), and Lahiri and Wang (2007). The meteorological and statistical literatures contain many more examples; some recent contributions include Hamill et al. (2003), Gneiting et al. (2007) and Thorarinsdottir and Gneiting (2010). We now use these methods to examine the BoE forecasts.

4. Empirical results

We begin by interpreting the graphical diagnostics provided in Figures 3–5.

Figure 3 provides another way of understanding the resolution of the probability forecasts that does not require estimation of the conditional expectation. For each horizon (only horizons up to four quarters are shown), each of the panels presents a pair of empirical CDF's of the forecast probabilities of an outcome below a threshold. One of the CDF's applies to cases for which the eventual outcome was below threshold, and the other, in the same colour, applies to cases for which the eventual outcome was above threshold. In a near-ideal world, these forecast probabilities should be near one in cases where the outcome did turn 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 of the same colour should be widely separated in each panel. 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 (particularly of the longest-horizon yellow lines) 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, suggesting much higher forecast resolution. GDP growth forecasts compared with current-vintage data in fact show little separation of the CDF's after the shortest horizons, although evaluated relative to preliminary-release data, there is more separation of cases. [11] Galbraith and van Norden (2009) 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, and find that forecasts appear to be essentially equivalent to unconditional forecasts at horizons of more than two quarters, which implies zero forecast resolution. However, for GDP forecasts there is some observable distinction between the fixed and market interest rate cases; the fixed cases show somewhat higher resolution at short horizons.

Figure 4 plots the estimated (kernel-smoothed) 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. It also provides some information on forecast resolution; a forecast with a constant $E(x|\hat{p})$, i.e. a horizontal conditional expectation, will have zero resolution. Again, see Galbraith and van Norden (2009) for a description of the methods used to construct these estimates, including bandwidth choice; all results depicted here use bandwidth 0.08.

The upper panels of the figure show these conditional expectations for the GDP growth forecasts at each forecast horizon. Deviations from perfect calibration (the 45-degree line) are widespread and sometimes large. Departures from the 45-degree line are more pronounced when using revised GDP figures than when using preliminary estimates of GDP growth. Indeed, although the conditional expectations for the preliminary data broadly follow the 45-degree line with some deterioration in fit as the forecast probabilities approach 1.0, when using revised GDP estimates it is unclear whether the curves have a more pronounced positive or negative slope. This implies that evaluating the forecasts using preliminary estimates of outcomes may falsely suggest that they have substantial resolution for the final data, whereas it is not obvious from the revised data that this is true.

The lower right panel shows results for forecasts of whether inflation will exceed the target rate. While the forecasts at horizons of less than 4Q are generally upward sloping and not far from the 45-degree line, as we saw for preliminary estimates of GDP, the longer horizon inflation rate forecasts have much more variable slope and show larger deviations from correct calibration, resembling more the results for revised estimates of GDP. This reflects the fact that shorter horizon forecasts will

[11] The former result mirrors the low 'content horizon' on U.S. and Canadian GDP 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.

generally have better resolution than longer horizon forecasts.

Finally, the lower left panel shows the results for interval forecasts of inflation, i.e. forecasts of the probability with which inflation will remain in the target band. For shorter horizons, there appear to be gross deviations from correct calibration: for example, when the forecast probability of inflation remaining within the target band is between 0.3 and 0.4, inflation outcomes are within the band essentially 100% of the time. For longer horizons, the results imply that inflation is very likely to remain within the target band regardless of the forecast probability of this.

However, a better perspective on the inflation interval forecasts may be gained by considering the corresponding panel in Figure 2. This shows that there were only four times in our sample when inflation fell outside the target band. These four points define much of the shape of the curve in Figure 5. There are also very few observations when the probabilistic forecast was below 0.7. Taken together, this suggests that the sampling error for the curves shown in Figure 4 will be quite high for points to the left of 0.7. To the right of that threshold, forecast calibration appears to be much more reasonable, but forecast resolution is uncertain.

As the above discussion suggests, sampling error makes it difficult to judge whether any of the deviations from the 45 degree line (calibration errors) or from a horizontal line (zero resolution) are statistically significant. For that, we require formal tests of the null hypotheses of correct calibration (that is, $E(x|\hat{p}) = \hat{p}$) and zero resolution (that is, $E(x|\hat{p}) = E(x)$) The results of these tests are given in Table 2; however we turn our attention to the remaining figure before discussing the results of the formal tests.

Figure 5 gives a different perspective on forecast calibration and resolution by showing their importance relative to MSFE. Recalling that (3) decomposes MSFE into mean-squared calibration error, resolution and the variance of outcomes, we can divide by the unconditional variance of the outcome process, $V_x = E(x - E(x))^2$, and re-arrange to obtain

$$E(\hat{p} - x)^2/V_x - E_f(\hat{p} - E(x|\hat{p}))^2/V_x + E_f(E(x|\hat{p}) - E(x))^2/V_x = 1. \quad (6)$$

The first term is the scaled MSFE, the second is the scaled calibration error and the third is the scaled resolution. Figure 5 shows how each of these three terms vary across forecast horizons. (Since calibration error enters negatively, the figure shows negative scaled calibration error.)

Using preliminary data for GDP outcomes, we see that while calibration error is roughly constant across horizons, resolution decreases steadily as the horizon increases, causing a roughly equivalent increase in MSFE. Resolution at horizons beyond 4Q are close to zero. However, using revised data for GDP outcomes, we see much larger calibration error at horizons up to one year and resolution that is close to zero at all horizons. The result is an MSFE that *exceeds* the variance of GDP outcomes at all horizons. (This is a commonly observed feature of GDP growth forecasts; see for example Galbraith and Tkacz 2007.) This implies that while the Bank forecasts appear to have some information content at shorter horizons, this content is an artefact of using unrevised data. From the perspective of forecasting the best final estimates of true outcomes, the bank's implicit probabilistic forecasts contain little information. The latter point reflects the relatively low information content of the initial-release GDP information available to economic forecasters.

For inflation, shown in the lower panels, the results are broadly similar regardless of whether we examine the probability that inflation will be within the target band or whether it will exceed the target level. In both cases, there is substantial resolution at the very shortest horizons which declines to roughly zero for forecast horizons of four quarters or more. (Recall that the measure of inflation being forecast is the four-quarter change in the price level. At forecast horizons of less than four quarters, therefore, some fraction of this four-quarter change has already been observed.) While there is some variation in calibration error, MSFE is dominated by the drop in forecast resolution.

Tables 1 and 2 provide more precise numerical descriptions of the probability forecasts and their performance. Table 1 provides a set of descriptive statistics on the forecast probabilities and outcomes. Note that, because annualized GDP growth is typically near 2.5%, mean forecast probabilities of being on either side of that value do not differ too much from 0.5. In contrast, the forecast probabilities of inflation being within the target band tend to be higher and inflation outcomes are rarely outside the band. Table 2 contains the decomposition (3) of the variance of outcomes x into MSFE, squared calibration error and resolution, as well as the results of formal tests of the null hypotheses that (a) calibration error is zero, and (b) resolution is zero.

As the decomposition of variance has been described above, consider now the tests of hypotheses (a) and (b). The tests are based on the facts that $E(x|\hat{p}) = \hat{p}$ implies correct calibration, and that $E(x|\hat{p}) = a$, a constant, implies zero resolution. To test calibration, again as in Galbraith and van Norden (2009), we estimate the model $x_i = a + b\hat{p}_i + c\hat{p}_i^2 + \epsilon_i$, and jointly test $H_0 : a = 0, b = 1, c = 0$ with a χ^2_3- distributed Wald statistic, using robust (Newey-West) standard errors in the computation. The $p-$values from these Wald tests are the values reported in Table 2. The test of zero resolution is of $H_0 : b = 0$ alone, and is computed as the simple $t-$ type test in the same regression, again using a robust standard error computation. The statistic is compared with the asymptotic normal critical values and again Table 2 reports two-sided p-values applying to these test statistics.

The results for the GDP forecasts reinforce the graphical analysis presented above. Results using preliminary GDP data are benign, with little or no significance evidence of miscalibration at any forecast horizon, and with statistically significant forecast resolution at all horizons up to and including four quarters. Results using revised GDP data are much less encouraging; there is strongly significant evidence of miscalibration at both longer and shorter forecast horizons, and there is no significant evidence of positive forecast resolution for anything beyond the current quarter. These observations are consistent with the conditional empirical CDFs shown in Figure 3 and the decompositions in Figure 5. A natural interpretation is that Bank of England forecasts of final revision data are mis-calibrated at least in part because of systematic features of the data revision process.

For the inflation level forecasts, we see that while there is strongly significant evidence of calibration errors at the longest and shortest horizons, there is no evidence of such errors for one to four quarter horizons. The resolution tests confirm the analysis presented in Figure 3 and Figure 5; while there is significant resolution at horizons up to three quarters, there is no detectable resolution beyond that point. Results for the inflation interval forecasts are quite similar with respect to forecast resolution. However, there is strong and widespread evidence of forecast miscalibration.

To summarize these empirical results: for inflation forecasts, deviations from correct calibration appear to be small, although nonetheless statistically significant at a number of forecast horizons. GDP growth forecasts, particularly of revised outcomes, produce much larger estimated deviations from correct calibration; that is, predicted probabilities of GDP falling below our threshold are in many cases far from 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 forecasts 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.

5. Discussion

By focusing our attention on the probabilistic forecasts implied by the Bank of England's fan charts, we have evaluated their performance on criteria that are relevant to policymakers: do they correctly capture the risk of high or low growth? the risk of high or low inflation? the risk that inflation will be outside the Bank's target band? Focusing on these probabilistic forecasts uses a subset of the information contained in the complete density. However, it also allows us to examine the extent to which these forecasts contain useful information about future growth and inflation. As in earlier studies, we find some statistically significant evidence of differences between the predicted probabilities and the actual probabilities of the subsequent outcomes. Both the inflation and GDP forecasts display some evidence of this calibration error. In the case of the GDP forecasts, our results emphasize the role that data revision plays in such errors.

Additionally, we are able to relate the degree of calibration error to overall forecast performance. For most of the cases that we examine, calibration errors play only a minor role in explaining changes in MSFE across different forecast horizons. The dominant factor is the fall in forecast resolution with increasing horizon. The only exception to this result lies in the evaluation of GDP forecasts with revised data; there we find low forecast resolution at *all* forecast horizons. Furthermore, the evidence of resolution in the inflation forecasts may be due to the fact that the Bank is forecasting the four-quarter change in prices; there is no significant evidence of forecast resolution for inflation at longer horizons. The Bank's forecasts on the eve of the most recent economic downturn are also consistent with a lack of resolution.[12]

A lack of resolution is important for both policymakers and forecast users to understand. For example, in the past the Bank of England has reacted to its own analysis of forecast performance by adjusting the dispersion used in its density forecasts; for example, see Elder, Kapetanios, Taylor and Yates (2005) or the Bank's August 2009 Inflation Report. This does not address the problem of the lack of forecast resolution. Whether sufficient information exists to produce forecast densities for GDP growth which have useful resolution at the longer horizons used by the Bank of England is an open question: low or zero resolution in GDP growth forecasts appears to be common. The lack of resolution in GDP growth forecasts does not imply that Bank of England forecasts use information less efficiently than competitors, but does imply that these forecasts contain little information beyond the unconditional distribution of GDP growth.

One of the key reasons for examining diagnostics such as those given above is to suggest ways in which forecast methods may be refined. When there is sufficient sample information to estimate calibration error precisely at various points,

[12] The Bank's August 2008 forecasts, which were made more than half a year after the US economy entered a recession, expected positive real GDP growth in every quarter for the subsequent three years. The November 2008 forecast, made after the collapse of major US and British financial institutions, expected negative growth through to the end of 2009, but implied that risks were symmetric around this forecast. Based on the most recently available GDP data vintages, both forecasts were much too optimistic; the cumulative probability that they assigned to outcomes as bad as or worse than what was observed in 2009 was less than 0.1%.

one option is to use a second-stage modification to produce corrected probability forecasts; see for example Hamill et al. (2003), Thorarinsdottir and Gneiting (2010). In the present context, however, both the small sample sizes and the fact that forecast methods are evolving relatively rapidly suggest that this is unlikely to be practical. The evidence of calibration error for relatively small and large probabilities suggests that the error distribution used for forecasting may assign inappropriate weight to relatively extreme events. The bi-normal distribution used in these forecasts provides much more flexibility than the normal in this respect, but this flexibility can certainly be extended; one avenue for investigation lies in the possibility that forecasts might show better performance for large deviations if a distribution with more flexible skewness and tail decay properties such as the skewed Student-t or generalized asymmetric Student-t (Fernandez and Steel 1998, Zhu and Galbraith 2010) were used in fitting the processes of interest. Another possibility is that the data revision process may contain systematic biases which make the forecaster's problem more difficult in dealing with final-release data, but to which forecasts could in principle be adapted. In any event, regular tracking of calibration and resolution may be a useful element in the evolution of the Bank of England's forecasts.

References

[1] Brier, G.W. (1950) Verification of forecasts expressed in terms of probabilities. *Monthly Weather Review* 78, 1-3.

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

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

[4] Clements, M.P. (2004) Evaluating the Bank of England density forecasts of inflation. *Economic Journal* 114, 844-866.

[5] 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.

[6] 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.

[7] Diebold, F.X. and G.D. Rudebusch (1989) Scoring the leading indicators. *Journal of Business* 62, 369-391.

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

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

[10] Fernandez, C. and Steel, M.F.J. (1998). On Bayesian modeling of fat tails and skewness, *Journal of the American Statistical Association* 93, 359-371.

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

[12] 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.

[13] Galbraith, J.W. and S. van Norden (2009) Kernel-based verification of the calibration of probabilistic economic forecasts. Working paper.

[14] Gneiting, T. (2008) Probabilistic forecasting. *Journal of the Royal Statistical Society Ser. A* 171, 319-321.

[15] 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.

[16] 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.

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

[18] 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.

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

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

[21] 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.

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

[23] Thorarinsdottir, T.L. and T. Gneiting (2010) Probabilistic forecasts of wind speed: ensemble model output statistics by using heteroscedastic censored regression. *Journal of the Royal Statistical Society Ser. A* 173, 371-388.

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

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

[26] Zhu, D. and J.W. Galbraith (2010). A generalized asymmetric Student-t distribution with application to financial econometrics. *Journal of Econometrics*.

Appendix: Tables [13]

Forecasts conditional on market interest rates

Table 1: Descriptive statistics of probabilistic forecasts

1(i): Inflation forecasts

Horizon	Mean	Std Dev	Variance	Minimum	Maximum	Obs		Outcomes
				Pr(Inflation	< Target)		> Target	< Target
0	0.515	0.386	0.149	0.000	1.000	48	25	23
1	0.549	0.346	0.120	0.000	1.000	47	24	23
2	0.556	0.310	0.096	0.000	0.986	46	23	23
3	0.567	0.259	0.067	0.002	0.942	45	22	23
4	0.573	0.212	0.045	0.035	0.905	44	21	23
5	0.568	0.179	0.032	0.118	0.927	43	20	23
6	0.553	0.128	0.017	0.301	0.904	42	20	22
7	0.517	0.088	0.008	0.314	0.774	41	19	22
8	0.484	0.074	0.006	0.255	0.637	40	18	22
			Pr(Inflation	within	Target Band)		# Outside	# Within
0	0.907	0.217	0.047	0.000	1.000	48	4	44
1	0.861	0.247	0.061	0.000	1.000	47	4	43
2	0.839	0.234	0.055	0.002	0.996	46	4	42
3	0.843	0.197	0.039	0.059	0.992	45	4	41
4	0.868	0.117	0.014	0.430	0.985	44	4	40
5	0.865	0.092	0.009	0.579	0.977	43	4	39
6	0.854	0.075	0.006	0.589	0.964	42	4	38
7	0.834	0.076	0.006	0.674	0.952	41	4	37
8	0.813	0.088	0.008	0.577	0.939	40	4	36

[13] The tables report results pertaining to forecasts for the indicator variable which equals one when either average CPI inflation or GDP growth exceeds the indicated annual threshold rate over the indicated forecast horizon, and zero otherwise. MSFE denotes the mean squared forecast error, and 'MS calibration' denotes the mean squared calibration error.

1(ii): GDP growth forecasts

Horizon	Mean	Std Dev	Variance	Minimum	Maximum	Obs	Outcomes	
		Pr(GDP Growth < 2.5%): 1st estimate					> 2.5%	< 2.5%
0	0.580	0.372	0.138	0.000	1.000	49	21	28
1	0.591	0.321	0.103	0.009	1.000	48	20	28
2	0.567	0.292	0.085	0.030	1.000	47	19	28
3	0.541	0.267	0.072	0.076	1.000	46	18	28
4	0.527	0.222	0.049	0.173	0.992	45	18	27
5	0.504	0.165	0.027	0.260	0.927	44	18	26
6	0.488	0.126	0.016	0.226	0.725	43	18	25
7	0.482	0.119	0.014	0.222	0.750	42	18	24
8	0.475	0.124	0.015	0.225	0.764	41	17	24
			Pr(GDP Growth < 2.5%): 2010Q1				> 2.5%	< 2.5%
0	0.580	0.372	0.138	0.000	1.000	49	27	22
1	0.591	0.321	0.103	0.009	1.000	48	26	22
2	0.567	0.292	0.085	0.030	1.000	47	25	22
3	0.541	0.267	0.072	0.076	1.000	46	24	22
4	0.527	0.222	0.049	0.173	0.992	45	23	22
5	0.504	0.165	0.027	0.260	0.927	44	22	22
6	0.488	0.126	0.016	0.226	0.725	43	21	22
7	0.482	0.119	0.014	0.222	0.750	42	20	22
8	0.475	0.124	0.015	0.225	0.764	41	19	22

Table 2: Calibration and resolution of forecasts

GDP growth: preliminary release data

Horizon:	0	1	2	3	4	5	6	7	8
Outcome Variance:	0.250	0.248	0.246	0.243	0.245	0.247	0.249	0.251	0.249
MSFE:	0.122	0.158	0.172	0.195	0.217	0.240	0.243	0.263	0.266
Resolution:	0.122	0.084	0.065	0.061	0.049	0.020	0.006	0.007	0.014
MS calibration:	0.002	0.006	0.008	0.019	0.027	0.027	0.015	0.034	0.052
Scaled resolution [0,1]:	0.500	0.347	0.268	0.255	0.205	0.081	0.026	0.028	0.056
Calibration p-value	0.747	0.755	0.679	0.561	0.535	0.434	0.554	0.090	0.103
Resolution p-value	0.000	0.000	0.000	0.001	0.033	0.095	0.331	0.904	0.940

GDP growth: 2010Q1 vintage data

Horizon:	0	1	2	3	4	5	6	7	8
Outcome Variance:	0.253	0.254	0.254	0.255	0.256	0.256	0.256	0.256	0.255
MSFE:	0.275	0.292	0.298	0.310	0.281	0.270	0.262	0.264	0.278
Resolution:	0.031	0.021	0.018	0.011	0.016	0.003	0.001	0.006	0.011
MS calibration:	0.065	0.067	0.072	0.070	0.048	0.025	0.015	0.024	0.044
Scaled resolution [0,1]:	0.124	0.084	0.072	0.045	0.062	0.011	0.004	0.026	0.045
Calibration p-value	0.005	0.006	0.030	0.074	0.342	0.478	0.487	0.001	0.000
Resolution p-value	0.005	0.256	0.617	0.871	0.775	0.880	0.893	0.643	0.169

Inflation: target threshold

Horizon:	0	1	2	3	4	5	6	7	8
Outcome Variance:	0.255	0.255	0.256	0.256	0.255	0.255	0.256	0.255	0.254
MSFE:	0.064	0.131	0.171	0.228	0.285	0.291	0.279	0.260	0.267
Resolution:	0.195	0.121	0.080	0.032	0.008	0.008	0.011	0.001	0.003
MS calibration:	0.011	0.006	0.008	0.015	0.047	0.053	0.042	0.010	0.014
Scaled resolution [0,1]:	0.783	0.485	0.320	0.128	0.032	0.031	0.043	0.004	0.012
Calibration p-value	0.000	0.635	0.847	0.649	0.115	0.024	0.081	0.102	0.000
Resolution p-value	0.000	0.000	0.000	0.010	0.791	0.564	0.275	0.387	0.788

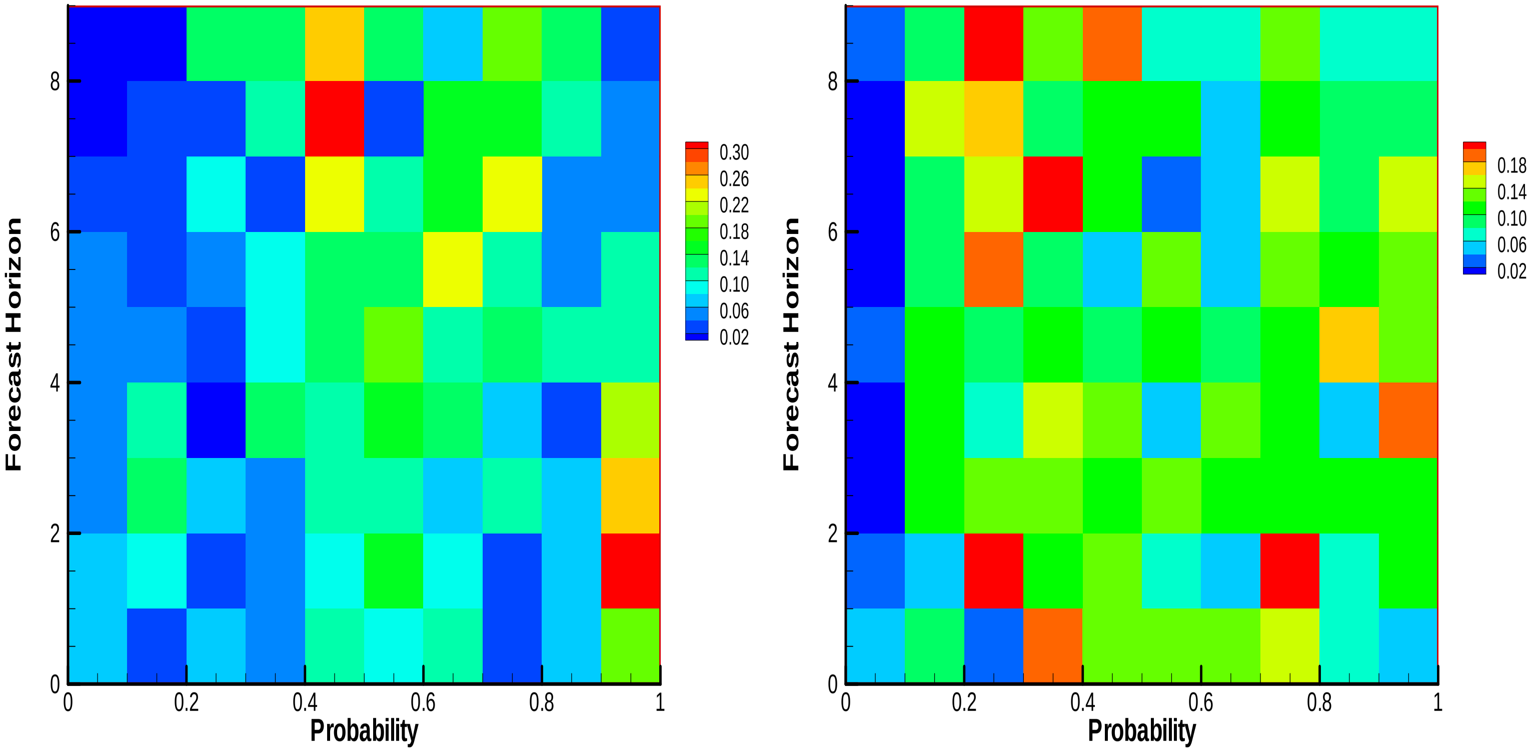
Inflation: target band

Horizon:	0	1	2	3	4	5	6	7	8
Outcome Variance:	0.078	0.080	0.081	0.083	0.085	0.086	0.088	0.090	0.092
MSFE:	0.024	0.023	0.058	0.086	0.095	0.093	0.096	0.103	0.116
Resolution:	0.057	0.062	0.042	0.011	0.001	0.001	0.000	0.001	0.002
MS calibration:	0.005	0.008	0.022	0.021	0.016	0.012	0.009	0.013	0.024
Scaled resolution [0,1]:	0.744	0.792	0.526	0.134	0.018	0.016	0.006	0.010	0.024
Calibration p-value	0.070	0.000	0.164	0.013	0.000	0.000	0.000	0.000	0.000
Resolution p-value	0.000	0.000	0.000	0.104	0.145	0.197	0.324	0.248	0.570

Figure 1

Probability integral transforms [a]

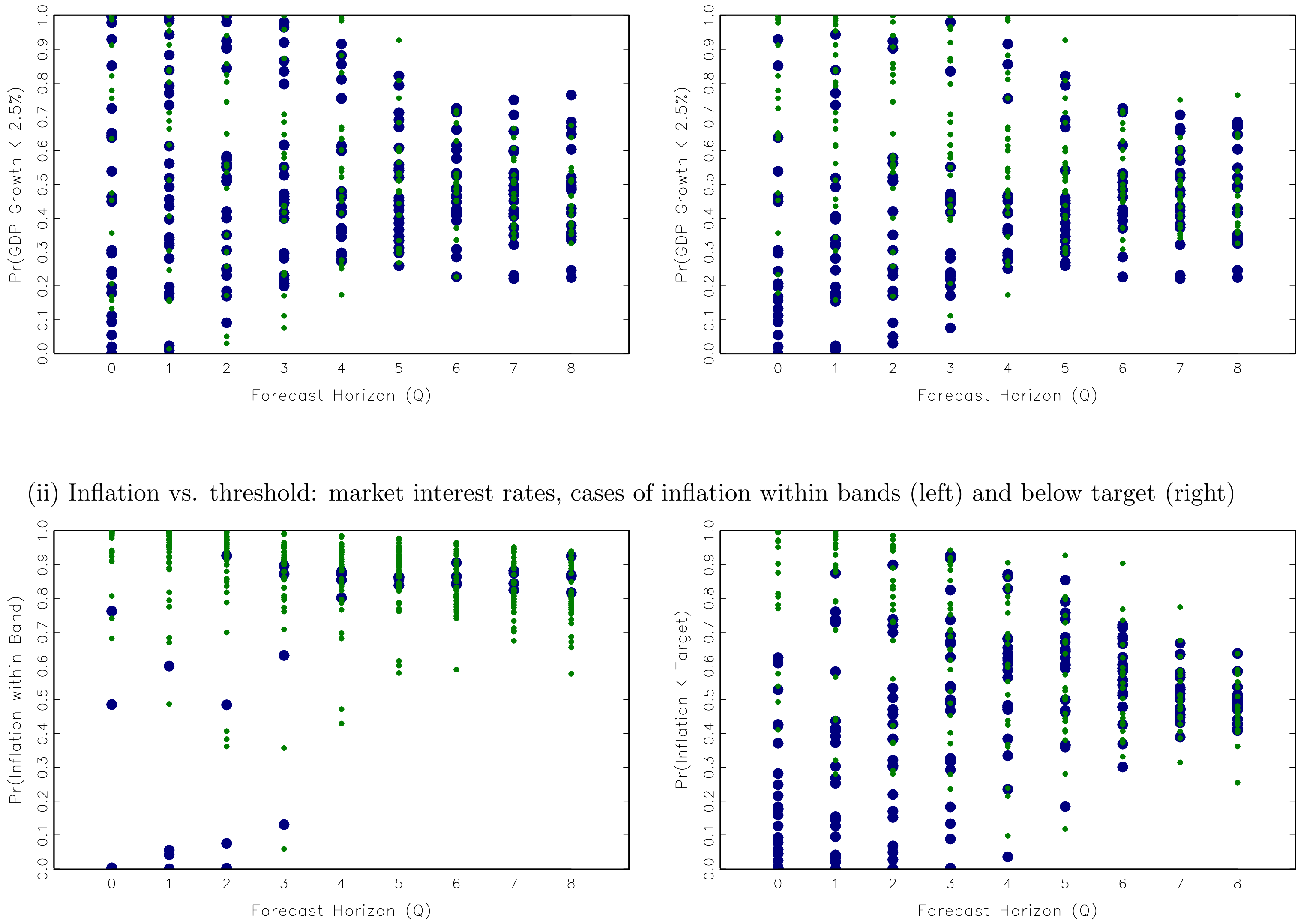
(i) GDP growth vs. threshold; (ii) Inflation vs. threshold; both at market rates



[a] In each of the panels of Figure 1, 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 2

Implied probability forecasts from Bank of England fan charts

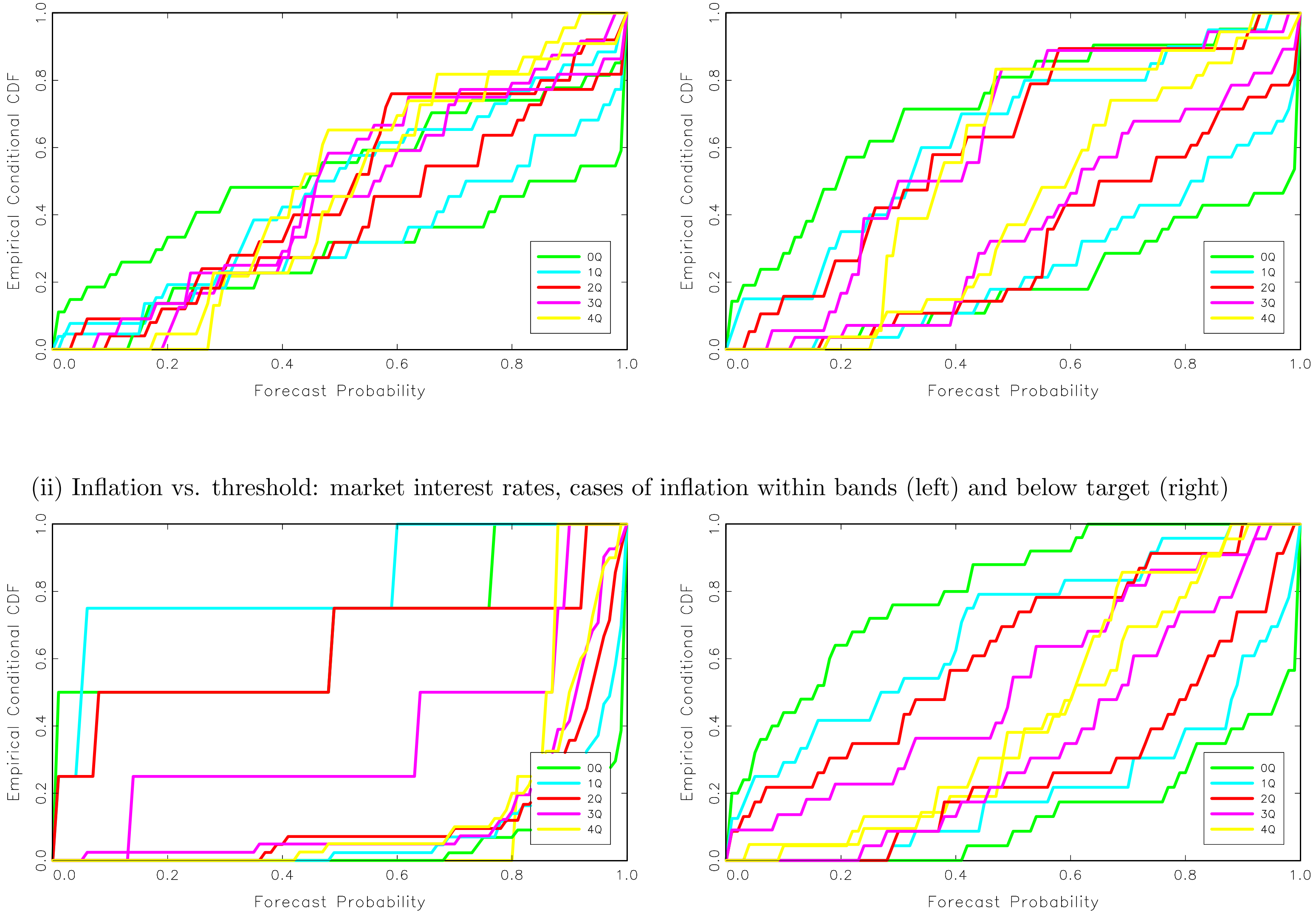


(i) GDP growth vs. threshold; market interest rates, 2010 Q1 (left) and preliminary (right) data

(ii) Inflation vs. threshold: market interest rates, cases of inflation within bands (left) and below target (right)

Figure 3

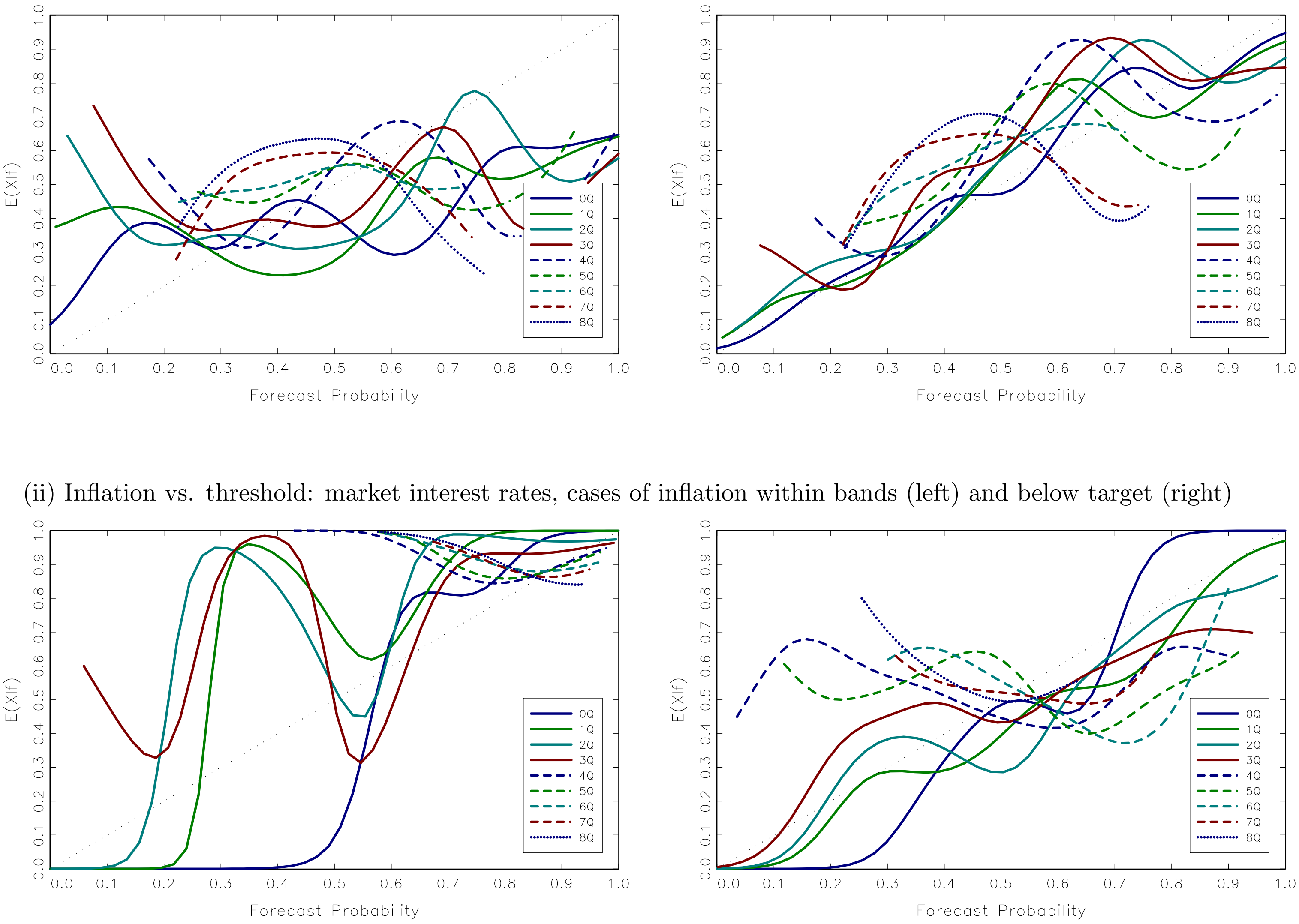
Empirical CDF's of implied probability forecasts from fan charts



(i) GDP growth vs. threshold; market interest rates, 2010 Q1 (left) and preliminary (right) data

(ii) Inflation vs. threshold: market interest rates, cases of inflation within bands (left) and below target (right)

Figure 4
Calibration of probability forecasts



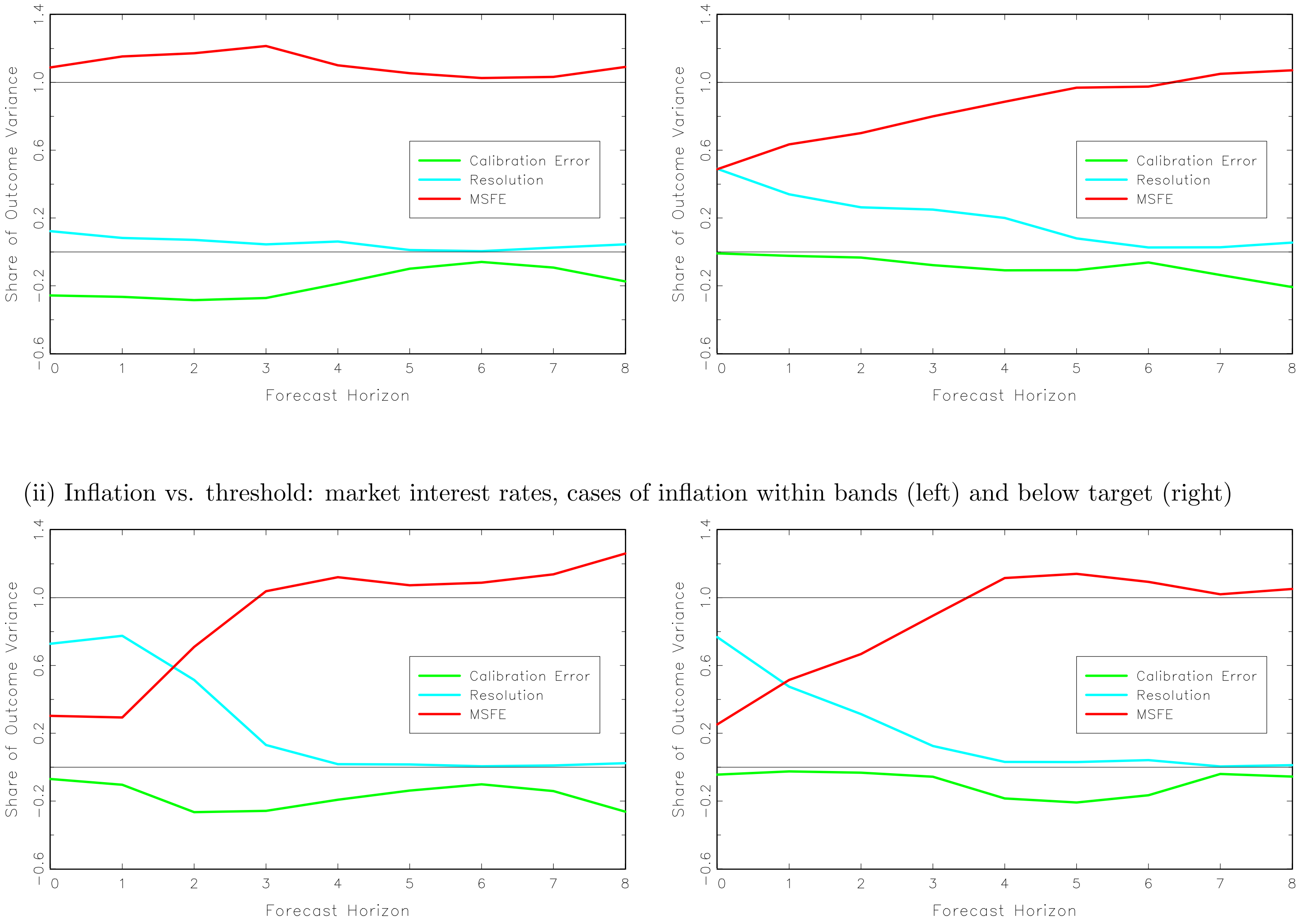
(i) GDP growth vs. threshold; market interest rates, 2010 Q1 (left) and preliminary (right) data

(ii) Inflation vs. threshold: market interest rates, cases of inflation within bands (left) and below target (right)

FIGURE 5

Decomposition of probability forecast MSE

(i) GDP growth vs. threshold; market interest rates, 2010 Q1 (left) and preliminary (right) data



(ii) Inflation vs. threshold: market interest rates, cases of inflation within bands (left) and below target (right)