

An evaluation of the Survey of
Professional Forecasters probability distributions
of expected inflation and output growth.

Michael P. Clements*

Department of Economics,

University of Warwick

November 22, 2002

Abstract

Regression-based tests of forecast probabilities of particular events of interest are constructed. The event forecast probabilities are derived from the SPF density forecasts of expected inflation and output growth. Tests of the event probabilities supplement statistically-based assessments of the forecast densities using the probability integral transform approach. The regression-based tests assess whether the forecast probabilities of particular events are equal to the true probabilities, and whether any systematic divergences between the two are related to variables in the agents' information set at the time the forecasts were made. Forecast encompassing tests are also used to assess the quality of the event probability forecasts.

Journal of Economic Literature classification: C53.

Keywords: Density forecasts, event probabilities, encompassing, SPF inflation forecasts.

First version: April 2002.

*Financial support from the U.K. Economic and Social Research Council under grant L138251009 is gratefully acknowledged. Helpful comments were received from Ken Wallis and seminar participants at Southampton University. Computations were performed using code written in the the Gauss programming language (Aptech Systems) and Givewin 2 and PcGive 10: see Doornik and Hendry (2001).

1 Introduction

In recent years a number of papers have gone beyond the traditional concern with the production and evaluation of point forecasts to consider interval forecasts and density forecasts,¹ in recognition that some measure of the degree of uncertainty surrounding a ‘central tendency’ will enhance the usefulness of the forecast. In this paper we consider the usefulness of extant techniques for the evaluation of short series of probability distributions of macroeconomic variables, such as the distributions of expected inflation and output growth generated by the Survey of Professional Forecasters (SPF). In some instances particular events may be of special interest, such as the event that the rate of inflation will fall within a target range. Having accurate forecasts of events of this type may be more important than the forecast density being correctly calibrated throughout its range. To this end, we propose a framework for testing various aspects of the rationality of forecasts of the probabilities of particular events derived from the SPF density forecasts. We also propose forecast encompassing tests of these event probabilities relative to rival sets of probabilities – one rival being based on a ‘no change’ forecast density. No change forecasts have recently claimed some success in the inflation point-forecasting literature (see Atkeson and Ohanian (2001)).

The literature on interval and density forecast evaluation is large and varied, and covers a range of situations. A number of the concerns addressed in the literature are not relevant to the evaluation of probability distributions that are derived from surveys, as opposed to being model based, and some techniques are not applicable in this instance. We provide a brief overview of the field, organized around a number of key distinctions, to highlight the issues that are relevant to our evaluation of survey-based probability distributions and event probability forecasts.

The first distinction we draw is whether reference is made to the method of construction of the forecast, which is equally germane to the evaluation of forecasts of any type (point, interval or den-

¹On the former, see Granger, White and Kamstra (1989), Chatfield (1993), Christoffersen (1998) and Clements and Taylor (2002), and on density forecasts, Diebold, Gunther and Tay (1998), Diebold, Tay and Wallis (1999), Diebold, Hahn and Tay (1999), Clements and Smith (2000), and the review by Tay and Wallis (2000). Wallis (2002) evaluates both types of forecast in a single framework using chi-squared goodness of fit tests.

sity). For point forecasts, for example, Chong and Hendry (1986) proposed a method of comparing rival forecasts that requires only the sequences of forecasts and the actual values of the variable being forecast, as a viable way of evaluating forecasts from large-scale macroeconomic models. Diebold and Mariano (1995) discussed extant methods of testing for statistically significant differences in rival forecasts which also made no recourse to their method of construction, and proposed a new, more general test for this purpose. As noted by Diebold, Tay and Wallis (1999) in evaluating the survey-based SPF probability distributions, where the forecasts are formed by averaging the probability distributions of the respondents, it is difficult to see how the method of construction could be part of the evaluation. To clarify this aspect, contrast the approach of Li and Tkacz (2001), where the conditional density function of a particular parametric model is compared to a non-parametric estimate of the conditional density function. The test statistic is based on the integrated squared distance between the two densities. The density function is derived from a parametric model which *defines* the conditioning variables in the non-parametric estimate of the ‘true’ (given those conditioning variables) conditional density.

Secondly, whilst the point forecast literature often assesses accuracy relative to some benchmark or rival forecast (see, e.g., Clements and Hendry (1998, ch. 3.3)) this is less common in the interval and density forecast evaluation literature. Granger *et al.* (1989) is an important exception. Evaluation is based on interval combination using the quantile regression techniques of Koenker and Bassett (1978, 1982), including combination of the interval forecasts (strictly, estimated quantiles) with a ‘constant quartile’ to assess the scope for bias-correction, again using quantile regression. More commonly, forecast intervals and densities are tested for correct conditional calibration (as defined below). Relative forecast performance is important in the point forecast literature, in part because of ignorance of what the lowest achievable mean squared forecast error (say) of a sequence of forecasts may be, and supplements tests of the ‘rationality’ of the forecasts, and tests that compare in-sample model fit to out-of-sample performance. In the case of interval and density forecasts, the emphasis probably reflects the difficulties in determining which is the best of two rival forecasts, for example, whether differences in coverage

rates of rival interval forecasts are statistically significantly different – in short, the absence of Diebold-Mariano type tests for interval and density forecasts. We construct two competitors to the SPF density forecasts. The first is motivated by the commonly-used point forecast no-change predictor, and the second is based on the notion of the unconditional average. These two rival sets of forecasts are used in formal forecast encompassing tests of the expected probabilities of events derived from the SPF forecasts.

Thirdly, there is an important distinction to be drawn between whether the forecasts are intended for a specific use, for example, as an input in a specific decision-making problem, or are general purpose. Granger and Pesaran (2000) consider the evaluation of density forecasts from a decision-theoretic approach, and Pesaran and Skouras (2002) review this approach and contrast it with evaluation based on ‘purely statistical measures’. The decision-theoretic approach evaluates forecasts in terms of their implied economic value. Economic value is the average loss that results from making decisions based on the likelihood of different states of nature eventuating, where the probabilities that underscore the calculations are given by the specific forecast density. Forecast densities can then be ranked on the basis of lowest expected loss, providing a metric for determining which is ‘best’. Granger and Pesaran (2000, p.243) give a number of reasons for the low take-up of this approach in macroeconomics. These include that the complete specification of the decision problem is often lacking. We follow the mainstream of the literature by focusing on general-purpose statistical approaches, but complement this with an empirical analysis of the performance of the forecasts at predicting two ‘events of interest’.

Fourthly, because our forecasts are not model-based, the recent concerns over the impact of parameter estimation uncertainty on comparisons of predictive accuracy of point forecasts (West (1996), West and McCracken (1998); West and McCracken (2002) provide an exposition) do not arise. But there is an interesting parallel. When forecasts are model-based, replacing the model’s parameters by their estimates and then effectively assuming these are the population values in calculating the forecasts ignores a source of uncertainty (we assume away other sources, such as the uncertainty in the model

specification – alternatively, the forecasts are conditional on the model specification). In the case of the survey-based forecasts, we may ask how confident the respondent is in their probability assessments (if a 30% probability is attached to the ‘bin’ that inflation will be between 2 and 3% next year, is this assessment made with 50% confidence or 95% confidence?). It is difficult to imagine how a formal treatment could be brought to bear on this issue.

Fifthly, there is a long tradition of using past performance to improve point forecasts (e.g., Mincer and Zarnowitz (1969), and the review by Clements and Hendry (1998, ch. 3)). In the density forecast literature, this is commonly known as ‘calibration’, and dates back at least to Dawid (1984): Kling and Bessler (1989) applies calibration techniques to the forecast densities of a vector autoregression of money, prices, interest rates and output. Our sample is too short to allow a ‘holdout period’ for which probability assessments can be adjusted based on past performance, but we suggest a way in which calibration could be used to obtain a joint distribution of inflation and output growth based on the two ‘marginals’.

The plan of the paper is as follows. Section 2 discusses evaluation methods based on the probability integral transform approach, and an approach that looks at the properties of forecasts of event probabilities derived from the density forecasts. Section 3 describes the nature of the SPF forecasts. Section 4 presents an evaluation of the SPF forecasts using the probability integral transform approach, as well as analyzing the properties of single and joint event forecasts. Section 5 concludes.

2 Density forecast evaluation techniques

2.1 The probability integral transform approach

Diebold, Tay and Wallis (1999) evaluate SPF inflation forecasts for the period 1969 to 1995 using the probability integral transform as described by Diebold *et al.* (1998). The key idea can be traced back at least to Rosenblatt (1952). Suppose there are a series of 1-step forecast densities for the value of a variable Y_t made at $t-1$, denoted by $p_{Y,t-1}(y_t)$, where $t = 1, \dots, n$. The probability integral transforms

(pits) of the realizations of the variable with respect to the forecast densities are given by:

$$z_t = \int_{-\infty}^{y_t} p_{Y,t-1}(u) du \equiv P_{Y,t-1}(y_t) \quad (1)$$

for $t = 1, \dots, n$, where $P_{Y,t-1}(y_t)$ is the forecast probability of Y_t not exceeding the realized value y_t . Then z_t has support on the unit interval by construction. Let $q(z_t)$ denote the density of z_t . If $f_{Y,t-1}(y_t)$ denotes the true density, then using the change of variable formula for $z_t = P_{Y,t-1}(y_t)$, we obtain q_t as:

$$\begin{aligned} q_t(z_t) &= f_{Y,t-1}\left(P_{Y,t-1}^{-1}(z_t)\right) \left| \frac{\partial P_{Y,t-1}^{-1}(z_t)}{\partial z_t} \right| \\ &= \frac{f_{Y,t-1}\left(P_{Y,t-1}^{-1}(z_t)\right)}{p_{Y,t-1}\left(P_{Y,t-1}^{-1}(z_t)\right)}. \end{aligned} \quad (2)$$

Consequently, when the forecast density equals the true density $q(z_t) = 1$ for $z_t \in [0, 1]$, i.e., $z_t \sim U(0, 1)$. Even though the actual conditional densities may be changing over time (as indicated by the t subscript) provided the forecast densities match the actual densities at each t , then $z_t \sim U(0, 1)$ for each t , so that the time subscript on q_t is redundant. Moreover the z_t are independently distributed, such that the time series $\{z_t\}_{t=1}^n$ is independently identically uniform distributed, i.e., iid $U[0, 1]$.

Evaluating the forecast densities by assessing whether $\{z_t\}_{t=1}^n$ is iid $U[0, 1]$ thus requires testing the joint hypothesis of independence and uniformity. Independence can be assessed by examining correlograms of $\{z_t\}_{t=1}^n$, and of powers of this series (as a check for dependence in higher moments, which would be incompatible with the independence claim), and formal tests of autocorrelation can be performed. Uniformity can also be assessed in a number of ways: whether the empirical cdf of the $\{z_t\}$ is significantly different from the theoretical uniform cdf (a 45° line) using the Kolmogorov Smirnov (KS) test of whether the maximum difference between the two cdfs exceeds some critical value. Alternatively, chi-squared goodness-of-fit statistics can also be used. However, in either case the effects of a failure of independence on the distribution of the test for uniformity is unknown, and could be exacerbated

by a small sample size. Moreover tests of autocorrelation will be affected by failure of the uniformity assumption. Graphical analyses are often reported as an adjunct to formal tests of the two parts of the joint hypothesis.

These tests evaluate the whole densities. As shown by Diebold *et al.* (1998) and Granger and Pesaran (2000), a density forecast that coincides with the data generating process will be optimal in terms of minimizing expected loss whatever users' loss functions, whereas in general rankings between rival forecasts will not be invariant to the loss function. This provides a strong case for evaluating the whole forecast density, when available, and for using the probability integral transform approach, which tests whether the forecast density closely corresponds to the actual density. However, allied to the concern of Diebold *et al.* (1998) that the outcomes of the tests may not be informative about the reasons for the rejections, that is, which aspect of the densities are deficient, it may be that a set of density forecasts are still valuable for the purpose at hand. For example, probabilities of events of primary interest to the user may nevertheless be well calibrated. An advantage of having the complete densities is that it is possible to read off the probabilities of particular events of interests. As well as being of interest in their own right, an assessment of the implied probabilities of certain events may also be informative about the reasons for the 'whole-density' rejection, thereby operating in a constructive manner. Below we adopt an approach to event probability forecast evaluation that is closely related to that of Christoffersen (1998) for interval forecasts. This is because interval forecasts can be viewed as event forecasts. Interpreted in this way, the nominal coverage rate of the interval is the forecast probability, and the event occurs when the forecast interval contains the realized value. The density forecast is comprised of all such intervals generated by allowing the nominal coverage rate to take on any value in the unit interval. Thus the evaluation of a specific sequence of interval forecasts can also be thought of as assessing just one aspect of an underlying sequence of forecast densities, such as the Value-at-Risk (VaR) in risk management exercises².

²See Lopez (1996) for a discussion of the relationship between VaR analysis and interval forecasting.

2.2 An event-based evaluation framework

Interest often focuses on the probabilities of certain events, which may be composites. For example, Garratt, Lee, Pesaran and Shin (2001) generate out-of-sample forecasts of “the probability that inflation will be in the range $1\frac{1}{2} - 3\frac{1}{2}\%$ next year, and growth will be ‘reasonable’ ” from a VAR model of the UK economy. Attaching probabilities to events of this sort is obviously of paramount importance in guiding policy.³

The SPF provides ‘marginal’ forecasts of inflation and output growth. Joint densities of inflation and output growth can only be constructed under the restrictive assumption that the two variables are assumed to be independently distributed, so that the joint is the product of the marginals. In section 4.3 we will consider events such as that just described which require the joint density of the two variables, and describe how event probabilities may be recalibrated to remove distortions induced by ignoring dependence between the two variables.

2.2.1 Regression-based tests of event probabilities

Suppose events are defined as occurring when output and inflation fall within certain ranges. Consider an event E defined in this way. The ranges might be constant over the forecast period, or might depend on time, as would be the case, for example, if E were the event that inflation (and/or output growth) were lower (higher) than in the previous year. The probabilities p_t attached to the event in each period are calculated by integrating over the relevant ranges of the joint density of inflation and output growth. For the SPF forecasts this amounts to calculating the product of the probabilities of the two marginal probabilities.

By way of contrast to interval evaluation, or the Pearson goodness-of-fit approach to density forecast evaluation,⁴ the evaluation of a sequence of event probability forecasts requires an assessment of forecast

³Since the Spring of 1997, the Monetary Policy Committee in the UK has been charged with delivering an inflation rate in a certain range, but whether formalised or not, considerations of this type are an integral part of any macroeconomic stabilisation policy.

⁴The Pearson goodness-of-fit approach to density forecast evaluation begins by dividing the range of the probability

probabilities which typically vary over time, with the ranges defining the events being fixed (or varying ‘exogenously’ over t). Whereas for a sequence of interval forecasts the nominal coverage level is fixed and the length and location of the intervals varies over t . Nevertheless, we shall show that tests from the interval evaluation literature can be simply adapted to evaluate the event probability forecasts.

Two sets of considerations arise in the definition of the events. Firstly, some events may be of particular interest, such as the ability to forecast declines in output growth, or recessions. Secondly, a choice of events such that some events occur only rarely will reduce the powers of tests of the adequacy of the probability forecasts.

Christoffersen (1998, p. 849–50) and Engle and Manganelli (1999) present regression-based tests of interval forecasts based on:

$$I_t = \alpha + \beta W_{t-1} + \epsilon_t, \quad t = 1, \dots, n \quad (3)$$

as a way of testing ‘conditional forecast efficiency’. Conditional efficiency is the requirement that a sequence of forecasts has correct conditional coverage, rather than simply that the ex post coverage of the set of forecasts equals the nominal coverage rate. In (3), I_t is an indicator variable that equals unity when the t^{th} interval (with nominal coverage p) contains the actual – a ‘hit’ – and W_{t-1} is a subset of variables from the information set Ω_{t-1} , which typically includes lagged values of the indicator function $\{I_{t-1}, I_{t-2}, \dots\}$, and in principle can include any variables known at period $t - 1$ and earlier. The joint null that $\alpha = p$ and $\beta = 0$ gives $E[I_t | \Omega_{t-1}] = p$, with the interpretation that for each t , the probability that the actual lies within the interval is equal to the nominal coverage rate, and is independent of whether or not previous intervals contained the actual values, and of any other factors. A rejection of $\beta = 0$ would signify that the likelihood of a hit varies systematically with information known at the time the interval forecast was made. Clements and Taylor (2002) consider the evaluation of interval forecasts

integral transform into k fixed, *equiprobable* classes, with boundaries j/k , $j = 0, 1, \dots, k$. The proportion of the actuals that falls into each is used to construct the statistic: see Wallis (2002). In the multivariate case, equiprobable classes would require classes of different lengths, because the joint probability integral transform (assuming independence) is not uniform but has as cdf $F_Z = Z - Z \ln(Z)$: see Clements and Smith (2000).

of intraday hourly returns, where the latter are marked by clear movements in volatility which repeat each day, and so include hourly dummies in Ω_t to test whether the intervals capture these patterns. Typically variables will suggest themselves depending on the context. Testing for independence in interval forecasts with very high (or low) nominal coverage is akin to testing the independence of ‘rare events’: misses occur only rarely so that power to detect patterns in these events will be low.

A simple way of testing the adequacy of the probability forecasts of the event E_t at periods $t = 1, \dots, n$ based on the above discussion is as follows. The null is that the forecast probability of E_t , denoted p_t , equals the true probability of the event, which we denote by p_t^0 . Let $I_t = 1(E_t)$ be an indicator variable that takes the value 1 if E_t occurs in period t , and the value 0 otherwise. In place of (3), we estimate the following equation where p_t is an explanatory variable and W_{t-1} is assumed to include an intercept:

$$I_t = \gamma p_t + \beta W_{t-1} + \epsilon_t, \quad t = 1, \dots, n \quad (4)$$

so that $p_t^0 = E[I_t | \Omega_{t-1}] = p_t$ if $\gamma = 1, \beta = 0$, where $\beta = [\beta_0 \ \beta_1]$ and $W_{t-1} = [1 \ W_{t-1}^*]'$. Strictly this is only a necessary condition, because $p_t^0 = p_t = \gamma p_t + \beta W_{t-1}$ may have solutions other than $\{\gamma = 1, \beta = 0\}$.⁵

Clements and Taylor (2002) note that the binary nature of the dependent variable in equation (3) means that the linear probability model is not an efficient way of estimating γ and β or testing hypotheses. Models with discrete dependent variables can be estimated by fitting a regression model to a logistic transformation of the dependent variable. In terms of the event forecast evaluation regression (4), the logit model is:

$$\Pr(I_t = 1) = \Lambda(\gamma, \beta; \alpha_t, W_{t-1}), \quad t = 1, \dots, n \quad (5)$$

where

$$\Lambda(\gamma, \beta; \alpha_t, W_{t-1}) = e^{\gamma \alpha_t + \beta W_{t-1}} / (1 + e^{\gamma \alpha_t + \beta W_{t-1}})$$

⁵This parallels the ‘standard’ condition for unbiasedness in the forecast-realization regression in the point forecast evaluation literature being only sufficient, see Holden and Peel (1990).

and $\alpha_t = \ln(p_t/(1 - p_t))$. The transformation of p_t given by α_t gives $\Pr(I_t = 1) = p_t$ under the null, $\gamma = 1, \beta = 0$.

The test of $\gamma = 1$ and $\beta_0 = 0$ in (4) with the W_{t-1} variables excluded ($\beta_1 = 0$) is akin to the Mincer and Zarnowitz (1969) test of weak efficiency, while testing in addition $\beta_1 = 0$ as part of the null is a test of ‘orthogonality’ or ‘strong-form’ efficiency.

When the events cover all possible outcomes, as for example if one were to consider the set of directional events defined by – output growth and inflation higher than in the previous period, both lower, and the two combinations of one higher and the other lower – then the adequacy of the set of event probability forecasts can be assessed using the multinomial logit model for unordered multi-responses (see, e.g., Greene (2000, sect. 19.7)), as described by Patton (2002) working within a similar framework. Alternatively, the specific events can be considered individually.

When a rival set of event probability forecasts is available, tests of forecast encompassing based on Chong and Hendry (1986) (see Newbold and Harvey (2002) for a recent review) allow for a formal assessment of relative performance. Denoting the SPF probability forecasts of a particular event by $\{p_t, t = 1, \dots, n\}$, and a rival set by $\{\tilde{p}_t, t = 1, \dots, n\}$, the SPF expected probabilities forecast encompass the rival set if we fail to reject the null that $\delta = 0$ in:

$$I_t = (1 - \delta)p_t + \delta\tilde{p}_t + \varepsilon_t \quad (6)$$

where I_t is the binary event indicator as above. Under the null, $E[I_t | p_t, \tilde{p}_t] = p_t$; the rival forecasts contain no additional information about I_t over and above that contained in the SPF forecasts. A failure to forecast encompass, i.e., $\delta \neq 0$, and e.g., $0 < \delta < 1$, would imply that a convex combination of both sets of forecasts is preferable to either alone, suggesting pooling of the two forecasts. A test of forecast encompassing in the opposite direction, that the rival forecast encompasses the SPF probabilities, is simply that $\delta = 1$. The form of (6) restricts the weights on the two rival forecasts to sum to unity, and supposes the two forecast are unbiased, as otherwise an intercept would be warranted, but we do not

restrict the weights to be positive by requiring $0 \leq \delta \leq 1$.

The form of the regression equation (6) is superficially similar to (4), with $W_{t-1} = \tilde{p}_t$ and $\beta = 1 - \gamma$, say, suggesting use of the logit approach as in (5). But this is misleading. Defining $\alpha_t = \ln(p_t/(1 - p_t))$ and $\tilde{\alpha}_t = \ln(\tilde{p}_t/(1 - \tilde{p}_t))$, and specifying

$$\Pr(I_t = 1) = \Lambda(\delta; \alpha_t, \tilde{\alpha}_t) = e^{\delta\alpha_t + (1-\delta)\tilde{\alpha}_t} / (1 + e^{\delta\alpha_t + (1-\delta)\tilde{\alpha}_t})$$

delivers $\Pr(I_t = 1) = p_t$ when $\gamma = 1$, and $\Pr(I_t = 1) = \tilde{p}_t$ when $\gamma = 0$, as required, under the null hypotheses, but under the alternative or maintained hypothesis:

$$\Pr(I_t = 1) = \frac{p_{r,t}}{1 + p_{r,t}}$$

where:

$$p_{r,t} = \left(\frac{\frac{p_t}{1-p_t}}{\frac{\tilde{p}_t}{1-\tilde{p}_t}} \right)^\delta \frac{\tilde{p}_t}{1-\tilde{p}_t},$$

which is hard to interpret.

One solution is to test for forecast encompassing based on OLS estimation of (6) transformed to:

$$I_t - p_t = \delta (\tilde{p}_t - p_t) + \varepsilon_t. \quad (7)$$

Because the dependent variable is bounded between plus and minus one, we report the results of a small simulation study of the standard t -tests that $\delta = 1$ and $\delta = 0$ in a setup designed to mimic the inflation forecasting application.

3 The SPF probability distributions of expected future inflation and output growth

The SPF⁶ is a quarterly survey of macroeconomic forecasters of the US economy that began in 1968 as the ASA-NBER survey, administered by the American Statistical Association and the National Bureau of Economic Research, and since June 1990 has been run by the Philadelphia Fed, under its current name. The majority of the survey questions ask respondents to report their point forecasts for a number of variables at various forecast horizons, from which median forecasts are calculated, but respondents are also asked to report discrete probability forecasts, or histograms, for output growth and inflation for the current and next year, which are then averaged to produce the density forecasts.

Diebold, Tay and Wallis (1999) discuss the survey and the complications that arise in using the inflation forecasts. In order to obtain a non-overlapping series of forecasts – in the sense that the realization of inflation in period t is known before making the next forecast of $t + 1$ at period t – they take the density forecasts made in the first quarter of each year of the annual change in that year on the preceding year. This avoids the counterpart of the well-known problem in the point forecast evaluation literature that a sequence of optimal h -step ahead forecasts, where the forecasting interval is one period, will follow a moving-average process of $h - 1$. Further complications are that both the base years of the price indices and the indices themselves have changed over time. The change in base years is likely to have had a minor effect on the inflation rate, and we construct a series of realizations of annual inflation that matches the indices for which probability assessments were requested. Thus, for 1969 to 1991 we use the implicit GNP deflator, for 1992 to 1995 the implicit GDP deflator, and for 1996 to 2001 the chain-weighted deflator, correcting for the changes in the definition of the index but not for base-year changes. Moreover, we use the latest available estimates of the realized values.⁷

⁶Detailed information on the survey as well as the survey results are available at the URL <http://www.phil.frb.org/econ/spf>. An academic bibliography of articles that either discuss or use data generated by the SPF is also maintained online.

⁷The series were taken from the Federal Reserve Bank of St Louis database (FRED), available at the URL <http://www.stls.frb.org/fred/data/> and have the codes GNPDEF, GDPDEF and GDPCTPI.

To obtain a non-overlapping series of output growth forecasts we adopt the same timing convention. There were similar changes in base years (which we again ignore) but a more important change in the definition of the series being forecast. Between 1969 and 1981 probability forecasts were of nominal GNP, between 1982 and 1991 real GNP, 1992 to 1995 real GDP, and 1996 to 2001 real chain-weighted GDP.⁸ We create a series of realizations that allow for the changes in definition. The change from nominal to real output growth in 1982 suggests caution in the analysis of this sequence of forecasts.

Finally, as documented by the Philadelphia Fed, the form in which the respondents report their probability assessments has changed over time, with changes in the number of bins and/or their locations and lengths as the perceived likely ranges of the target variables has changed. However, this complication is minor because for the most part we will want to read off probabilities of certain values, and the values that define given probabilities, and both can be achieved by piecewise linear approximation – this approximation ‘undoes’ the discretization in the histogram.⁹

4 Results

4.1 Graphical analyses of densities and probability integral transforms

Figure 1 portrays the inflation density forecasts as Box-Whisker plots along with the realizations. Because we closely follow the approach of Diebold, Tay and Wallis (1999), the observations for 1969 to 1996 match those recorded in their figure, and are discussed in detail by those authors. We note that the forecasts and realizations for 1997 and 1998 indicate a continuation of the tendency from the early

⁸Data source is FRED, as for inflation (see previous footnote). The series codes are (in order) GNP, GNP divided by GNPDEF, GDP96, and GDPC1.

⁹As an example, suppose we wish to calculate the forecast probability of observing a value less than $Y = 3.5$. Suppose $\Pr(Y < 2)$ is 0.5, and the bin defined by $[2, 4)$ has a probability of 0.2. Then:

$$\Pr(Y < 3.5) = \Pr(Y < 2) + \frac{1.5}{2} \Pr(Y \in [2, 4)) = 0.5 + \frac{1.5}{2} 0.2 = 0.65.$$

Linear interpolation follows the assumption implicit in the histogram – that probability mass is uniform within a bin. If a bin is bordered by a high probability bin and a relatively low probability bin, it might be desirable to attach higher probabilities to points near the boundary with the high probability bin.

part of the decade to both over-estimate the uncertainty (the densities seem too dispersed) and level (the actual is in or close to the lower quartile) at low and falling rates of inflation. The ‘errors’ (median less realized value) are small, at least compared to the turbulent 1970’s. This trend appears to have come to an end in 1999 – 2001 as inflation rose a little.

Care is required in interpreting the equivalent figure for output growth (figure 2). The first thirteen observations relate to nominal GNP (1969 to 1981), and there is the same pattern of realizations in the upper quartiles of the forecast densities in the early 1970’s as observed for the inflation forecasts at the time of rapid increases in the price level. Of the twenty forecasts of real output growth, only around a quarter contain the realized value within the interquartile range. The forecasts and realizations for 1982 – 86 are worthy of comment. The fall in output in 1982 was below the 25th percentile of the forecast distribution, but negative growth was not unexpected at the beginning of the year (recall the forecasts are made in the first quarter) with the median being below zero. The strength of the recovery in 1983, and especially in 1984, was somewhat of a surprise, but forecasters completely missed the dip in growth in 1985. In 1985 the ‘whiskers’ were 5.3 and 7.7, indicating relatively little uncertainty, but the actual was 3.5. In 1986 the actual was very similar at 3.2, but there was perceived to be a relatively small risk of a large decline, that is, a large left tail to the probability distribution – the 25th percentile was 2.5, but the lower whisker was -2.9!¹⁰

The other distinctive feature of the real output growth forecasts is the sequence in the second half of the 1990’s where the economy grows more rapidly than is expected, at the same time that inflation is generally lower than expected.

Table 1 records the results of formal testing of the pits’ for uniformity and independence. The KS test statistic for inflation is 0.14, compared to a 5% critical value of 0.24, so there is no evidence against the null of uniformity, but independence is rejected because the null of no first-order serial correlation in the $\{z_t\}$ is rejected at the 1% level. In addition, Berkowitz (2001) suggests taking the

¹⁰The Philadelphia Fed note that it is unclear whether the surveys for 1985 and 1986 relate to the current year or the following year, so there are doubts concerning the reliability of these observations (see <http://www.phil.frb.org/files/spf/spfprob.txt>).

Table 1 Tests of SPF density forecasts of inflation based on probability integral transforms.

Test	
KS test of uniformity	0.14
Bowman-Shenton asy normality	0.13
Doornik-Hansen	0.02
Berkowitz Independence (assuming $N(0, 1)$. 1 d.o.f.)	0.00
Berkowitz $iid(0, 1)$ (assuming normality. 3 d.o.f.)	0.00

All the test outcomes are recorded as p -values, except for the KS test, which is the test statistic value. The 5% critical value is 0.24.

inverse normal CDF transformation of the $\{z_t\}_{t=1}^n$ series, to give, say, $\{z_t^*\}_{t=1}^n$. Thus, the z_t 's, which are $iidU[0, 1]$ under the null, become iid standard normal variates. Berkowitz argues that more powerful tools can be applied to testing a null of $iidN(0, 1)$ compared to one of iid uniformity. He proposes a one-degree of freedom test of independence against a first-order autoregressive structure, as well as a three-degree of freedom test of zero-mean, unit variance and independence. In each case the maintained assumption is that of normality, so that standard likelihood ratio tests are constructed using the gaussian likelihoods. In our case, both the one and three-degree of freedom tests rejected at the 1% level. Testing for a zero mean and unit variance, assuming independence, yielded a p -value of 0.10. The assumption of normality of $\{z_t^*\}_{t=1}^n$ is also amenable to testing – the Shenton and Bowman (1977) two-degree of freedom asymptotic chi-squared test returned a p -value of 0.13, whilst the test recommended by Doornik and Hansen (1994) rejects at the 5% level. Notice though that these tests of distribution assume a random sample.

For output growth over the whole period 1969 – 2001, the KS test statistic value is significant at the 5% level, and there is clear evidence of that the $\{z_t^*\}_{t=1}^n$ series is non-normal. If we confine attention to the real output growth forecasts, 1982 – 2001, the KS statistic is on the borderline of significance at the 5% level, but there is no evidence of correlation in the $\{z_t\}_{t=1}^n$, or of non-normality in the $\{z_t^*\}_{t=1}^n$ series, or of rejection on the Berkowitz tests, although the sample now consists of only 20 observations.

4.2 Inflation event forecasts

In section 2.1 the SPF inflation forecasts were rejected using tests based on the probability integral transform. The formal tests are not especially informative about the reasons for the rejection, although the time series Box-Whisker plots of the densities suggest that the deficiencies include a tendency over the recent period to both over-estimate the uncertainty or variability and level of the rate of inflation. In the absence of a fully-formulated decision/cost based approach to the evaluation of the forecasts (see e.g., the references in the Introduction), it is unclear to what extent these deficiencies detract from the value of the forecasts from a user's perspective. We provide two complementary assessments of the quality of the density forecasts, in terms of how accurately two events are forecast. The first is the event that inflation will be in some target range, say $1\frac{1}{2}$ to $3\frac{1}{2}$ %. In a number of countries the monetary authorities target a range for the inflation rate, which either formally or informally is an important determinant of monetary policy. The second event is the direction-of-change of inflation. Accurately forecasting increases versus decreases in rates of growth has obvious appeal in the context of real activity variables, because of the correlation with business cycle phases of contraction and expansion. Canova (2002) recommends evaluating inflation forecasts in this way.

Table 2 records the p -values of the tests of weak and strong efficiency described in section 2.2 for assessing the quality of the SPF forecast probabilities of the two events. For the strong efficiency test we include as explanatory variables the lagged rate of import price inflation and the (linearly detrended) unemployment rate.¹¹ Lagged values of the indicator variables were rarely at all significant. Import price inflation captures the commodity (oil) price shocks of the seventies, and the unemployment rate is suggested as an important predictor of inflation by textbook Phillips Curve models. The results indicate that we are unable to reject the nulls that the SPF forecasts of both events are both weakly and

¹¹The civilian unemployment rate u was taken directly from Ray Fair's Fairmodel web site <http://fairmodel.econ.yale.edu> (model release July 31, 2002). The import price variable was constructed as described in Fair (2000), using data series taken from the Bureau of Economic Analysis National Income and Product Accounts Tables web site <http://www.bea.gov/bea/dn/nipaweb/index.asp>. The import price deflator is current price Imports (Billions of \$s, Table 1.1, line 17) divided by chained constant price imports (Billions of chained 1996 \$s, Table 1.2, line 17). Import price inflation is the percentage rate of change in this variable.

strongly efficient. Because the failure to reject might simply reflect a lack of power due in part to the small number of observations (only 33), we consider in addition two benchmarks. The first is a density forecast analog to the widely-reported ‘no change’ point forecast of Theil (1966). This is commonly viewed as a naive predictor, but has recently been shown to be a good forecasting device in certain states of nature (see, e.g., Clements and Hendry (1999)). Moreover, Atkeson and Ohanian (2001) show that this ‘naive’ predictor outperforms modern Phillips Curve-based model predictions of inflation in the U.S. over the last fifteen years. The density forecast analog assumes a gaussian density with means equal to the realised rates of inflation in the previous periods, and the variances the estimated sample variances for the $t - 4$ to $t - 1$ observations (for a forecast of period t). Given the preceding comments it is perhaps unsurprising that no change probability forecasts are not rejected for either event. Note that by construction the directional event probabilities are constant at 0.5 for all periods,¹² so $\alpha_t = 0 \forall t$, and is excluded, so the tests are one and three-degree of freedom tests respectively. Thus we are unable to reject a set of forecasts that states that there is an evens chance that inflation will be lower (higher) than in the previous period.

An alternative approach is to use an ‘unconditional’ probability distribution to derive the event forecast probabilities. We assume a gaussian density with constant mean and variance set to forecast period estimates. In contrast to the no change forecasts, which suppose that the future will be like the recent past, the average approach requires that the observations are generated by a constant, stationary mechanism. (Note that the averages are calculated of the forecast period observations.) These forecasts are rejected by both tests for both events (but only at the 10% level for the target-range event weak-efficiency test). Now α_t is a constant for the target-range event, so the constant (β) is omitted from the logit regressions.

The results of the forecast encompassing test described in section 2.2 are recorded in table 3 for the SPF probabilities against each rival.¹³ The SPF forecasts of the direction of inflation clearly forecast

¹²Note that the ‘no change’ probability forecasts of the ‘target range’ event are not the same for all periods. ‘No change’ relates to the motivation behind the form of the forecast density, not the event probabilities.

¹³We consider the SPF against each rival one at a time, although forecast encompassing tests could be used: see Harvey and

Table 2 Inflation Events Forecasts.

	Weak efficiency $\beta_0 = 0, \gamma = 1,$ (W_{t-1} excluded)	Strong efficiency $\beta_0 = 0, \gamma = 1, \beta_1 = 0$
Target range event		
SPF	0.168	0.870
No change	0.467	0.576
Unconditional	0.085	0.047
Directional event		
SPF	0.702	0.174
No change	0.386	0.143
Unconditional	0.000	0.002

Event probability forecasts are calculated for the SPF density forecasts, and for ‘no change’ and ‘unconditional’ density forecasts.

The table records the p -values of LR tests of the linear restrictions implied by the null hypotheses in logit regressions. These are two-degree of freedom tests in the second column, except for the ‘unconditional’ forecasts in the first panel, and the ‘no change’ in the second panel (one-degree of freedom), and four-degree of freedom tests in the third column, except for the ‘unconditional’ in the first panel, and the ‘no change’ in the second panel (three-degree of freedom tests).

The explanatory variables in W_{t-1} in the tests in the third column are the linearly-detrended civilian unemployment rate and import price inflation.

encompass both rivals, while the two reverse direction hypotheses are rejected at the 1% level. For the target range event the SPF forecast encompasses the no change forecasts, and encompassing in the reverse direction is rejected at the 5% level, while the hypothesis that the SPF forecast encompass the Unconditional is borderline at the 5% level, but the hypothesis is emphatically rejected when run in the reverse direction. From table 3 we see that the weight on the SPF forecasts ($1 - \hat{\delta}$) exceeds one ($\hat{\delta} < 1$) in three of the four cases, though (as mentioned) not significantly so. These results support the SPF event probability forecasts by indicating that the two rival sets of forecasts do not contain useful additional information.

The forecasting encompassing tests reported in table 3 are based on the assumption that it is reasonable to suppose that the ‘ t -statistics’ are approximately Student t in samples of only 33 observations. Table 4 reports the results of a small simulation study of the adequacy of this assumption. We looked at two sample sizes, $T = 30$, and $T = 100$. The data generating process for inflation is an AR(1)-

ARCH(1), approximately calibrated on the US annual data:

$$\Delta p_t = 0.5 + 0.8\Delta p_{t-1} + \varepsilon_t$$

$$\varepsilon_t = z_t \sqrt{h_t}$$

$$h_t = 0.4 + 0.85\varepsilon_{t-1}^2$$

where z_t is a standard normal random variable. h_0 and Δp_0 are set to their model-mean values, but in any case on each of the 50000 replications the first 50 values are discarded. We consider only the event that inflation falls within the target range $1\frac{1}{2}$ to $3\frac{1}{2}\%$. The ‘true’ probability event forecasts are derived from the 1-step ahead Gaussian densities implied by the data generating process with the true values of the coefficients in the conditional mean and variance functions assumed known. There is no suggestion that this process is a good representation of the way in which the SPF forecast densities are constructed. All we require is that it captures some of the salient features, such as changing conditional means and variances, so that we can explore the properties of testing procedures against alternatives in which these aspects are absent or otherwise incorrectly specified. The no change and unconditional event probabilities are calculated as in the empirical work.

The first and third rows in table 4 show that the sizes are close to the nominal 5% even at $T = 30$. Given the large number of replications, the simulation error is small, and the divergences between the Monte Carlo estimates of the actual size, and the nominal 5% size, though small are significant. The second and fourth rows report the rejection frequencies of the tests that the No change and Unconditional forecast encompass the true probabilities, and show reasonable power even at $T = 30$, especially for the Unconditional. These simulation results suggest we can be reasonably confident about the forecast encompassing results.

Figure 3 shows the occurrences of the two events and the forecast probabilities. The no-change forecasts are uninformative about the directional event by construction, but the SPF forecasts are also

Table 3 Forecast Encompassing Tests of Inflation Events Forecasts.

$1 - \delta$	δ	$\hat{\delta}$	p -value $\delta = 0$	p -value $\delta = 1$
Target range event				
SPF	No change	0.378	0.123	0.013
SPF	Unconditional	-0.338	0.053	0.000
Directional event				
SPF	No change	-0.018	0.950	0.001
SPF	Unconditional	-0.231	0.306	0.000

Forecast encompassing tests are calculated for the SPF event probability forecasts versus the ‘no change’ and ‘unconditional’ probabilities.

The table records the p -values of the null that one forecast encompasses the other, as described in the text based on equation 6.

Table 4 Forecast Encompassing Tests Simulated Sizes and Powers, for a nominal size of 5%, and the target range event.

	$T = 30$	$T = 100$
True vs. No change	0.070	0.066
No change vs. True	0.432	0.944
True vs. Unconditional	0.065	0.039
Unconditional vs. True	0.721	0.996

Simulated sizes and powers of forecast encompassing tests of event probability forecasts, as described in text.

The Monte Carlo standard error is $\sqrt{p(1-p)/M}$, where p is the nominal size and M the number of replications, so for $p = 0.05$ and $M = 50,000$ an approximate 95% confidence interval is (0.048, 0.052).

of little value over the last decade. But during the 1990’s inflation variability was low and directional changes are often small changes in percentage point terms, making this event particularly difficult to predict correctly.

4.3 Output growth and inflation joint events

Of obvious interest to policy makers will be the likelihood of inflation remaining relatively low and output growth being reasonably buoyant, so we evaluate the reliability of the SPF forecasts along this dimension. We suppose the event of interest is that inflation falls in the target range $1\frac{1}{2}$ to $3\frac{1}{2}$ % (as above) at the same time that the rate of output growth exceeds $2\frac{1}{2}$ % per annum. Figure 4 plots the time series and associated SPF probabilities of this event. Table 5 records the proportion of times that the event occurred for probability assessments in particular ranges. Because there are only 20 observations following the change in the output series in 1982 we adopt a coarse grid, assigning forecast probabilities

Table 5 Reliability of Joint Event Probability Forecasts.

Range	Proportion	Actual No.	Forecast no.
0,0.25	0.3	3	10
0.25,0.5	0.5	3	6
0.5,0.75	0.75	3	4
0.75,1	—	0	0

to the ranges $[0, 0.25)$, $[0.25, 0.5)$, $[0.5, 0.75)$ and $[0.75, 1]$.¹⁴

From table 5 it is apparent that the probability assessments understate the likelihood of the event occurring, because in each of the first three ranges the observed proportion of times the event occurs is at the upper end of the nominal range. This is clearly evident from figure (4). Testing the null that $\gamma = 1$ and $\beta_0 = 0$ in a logit regression as described in section 2.2 yielded a p -value of 0.166. The estimated value of γ was close to 1, and the constant, β_0 , though not significant, was positive, consistent with the general under-estimation of the probabilities of the event. The graphical evidence suggests the failure to reject the null is due to the inadequate sample size rather than reflecting positively on the quality of the event forecasts.

4.3.1 Recalibration

One reason why the joint event probability forecasts may not be well calibrated is that we have ignored any dependence between the inflation and output growth forecasts. The rationale behind (re)-calibration is most easily seen for univariate density forecasts, although the logic applies equally to multivariate densities and to probability event forecasts. Suppose a series of density forecasts are dynamically well-specified, in the sense that the independence of the $\{z\}_{t=1}^n$ series approximately holds, but the uniformity assumption is clearly invalid. Then it may be possible to obtain better future forecast densities by calibration: see e.g., Dawid (1984), and Kling and Bessler (1989) for an illustration. This can be most

¹⁴A typical calibration plot would be a cross-plot of the first two columns of the table.

easily seen by rearranging (2) to give:

$$f_{Y,t-1}(y_t) = p_{Y,t-1}(y_t) q(z_t)$$

where the independence (or correct dynamic specification) allows us to drop the time subscript on $q(\cdot)$.

Under the iid assumption, we can estimate $q(\cdot)$ by the empirical pdf of $\{z_t\}_{t=1}^n$, which we denote by $\hat{q}(\cdot)$, and use this in place of $q(\cdot)$.¹⁵ The re-calibrated forecast density is then given by:

$$\hat{p}_{Y,t-1}(y_t) = p_{Y,t-1}(y_t) \hat{q}(z_t)$$

which should provide a better estimate of $f(\cdot)$. Note that the $t - 1$ forecast origin indicates that the forecasts underlying the calculation of $\hat{q}(\cdot)$ are being calibrated, though typically future forecasts would be adjusted, as noted above.

Thus, if the joint event forecasts are dynamically well specified, but poorly calibrated because they fail to allow for dependence, then in principle the calibration function of table 5 could be used to re-calibrate the joint event forecasts to obtain a set of forecasts that better reflects the interdependencies between inflation and output growth. In our case original probability assessments would be increased. This is simple to do, but would require a ‘hold-out’ sample of observations for a proper evaluation.

5 Conclusions

We have applied regression-based tests to probability event forecasts derived from the SPF density forecasts to complement a purely statistically-based assessment of the forecast densities using the probability integral transform approach. The regression-based tests assess whether the forecast probabilities of particular events are equal to the true probabilities, and whether any systematic divergences between the

¹⁵Fackler and Kling (1990) fit a beta distribution to the empirical distribution of the z 's, but see also Kling and Bessler (1989).

two are related to variables in the agents' information set at the time the forecasts were made. These tests are analogous to the tests of weak and strong efficiency that are a mainstay of the point forecast evaluation literature. The SPF event probability forecasts are not found wanting in this regard, but neither are forecasts derived from 'no change' inflation density forecasts. The literature suggests that, at least in the point forecasting arena, 'no change' inflation forecasts are competitive with economic model-based forecasts. Nevertheless, the SPF event probabilities forecast encompass those derived from the no-change forecast densities for predicting the direction of change of inflation. In terms of predicting the event that inflation will be in the range $1\frac{1}{2}$ to $3\frac{1}{2}\%$ a year ahead, we are able to reject the null that the no-change probabilities forecast encompass the SPF at the 5% level, but not the 1% level, although there is no evidence against the hypothesis in the reverse direction.

Notwithstanding the relatively small number of forecasts at our disposal, that formal tests of the event probabilities based on the notion of forecast encompassing suggest that the SPF forecasts contain all the information in two rival sets of forecasts for the purpose of predicting two events of interest. In principle formal analyses of the sort developed in this paper may provide useful additional information to the popular probability integral transform approach to assessing the quality of density forecasts.

Figure 1 Inflation forecast probability distributions shown as Box-Whisker plots and realizations. The boxes represent the inter-quartile, the outer 'whiskers' the 10 and 90th percentiles, and the inner line the median. The realizations are circles with dots at the centres.

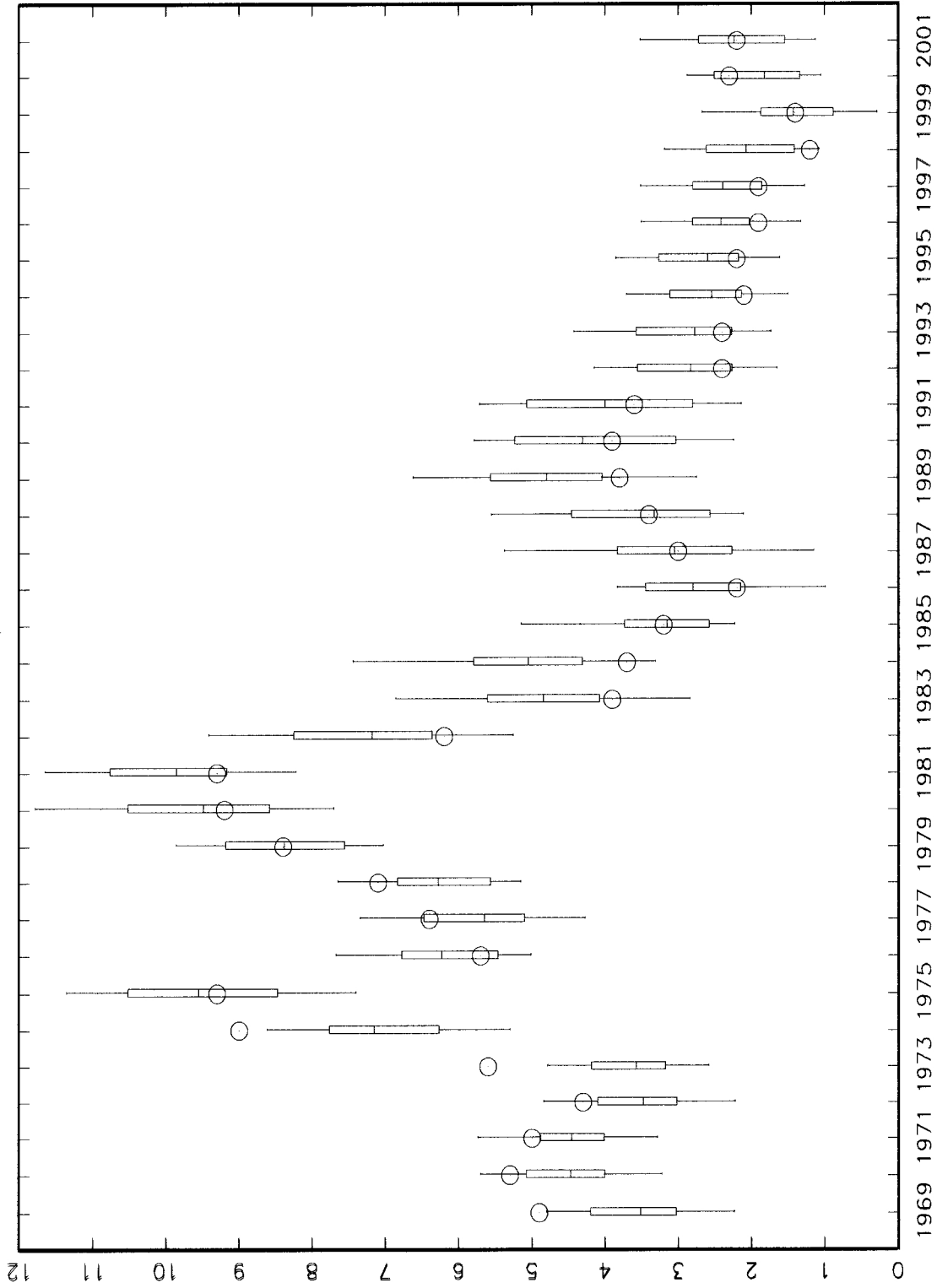
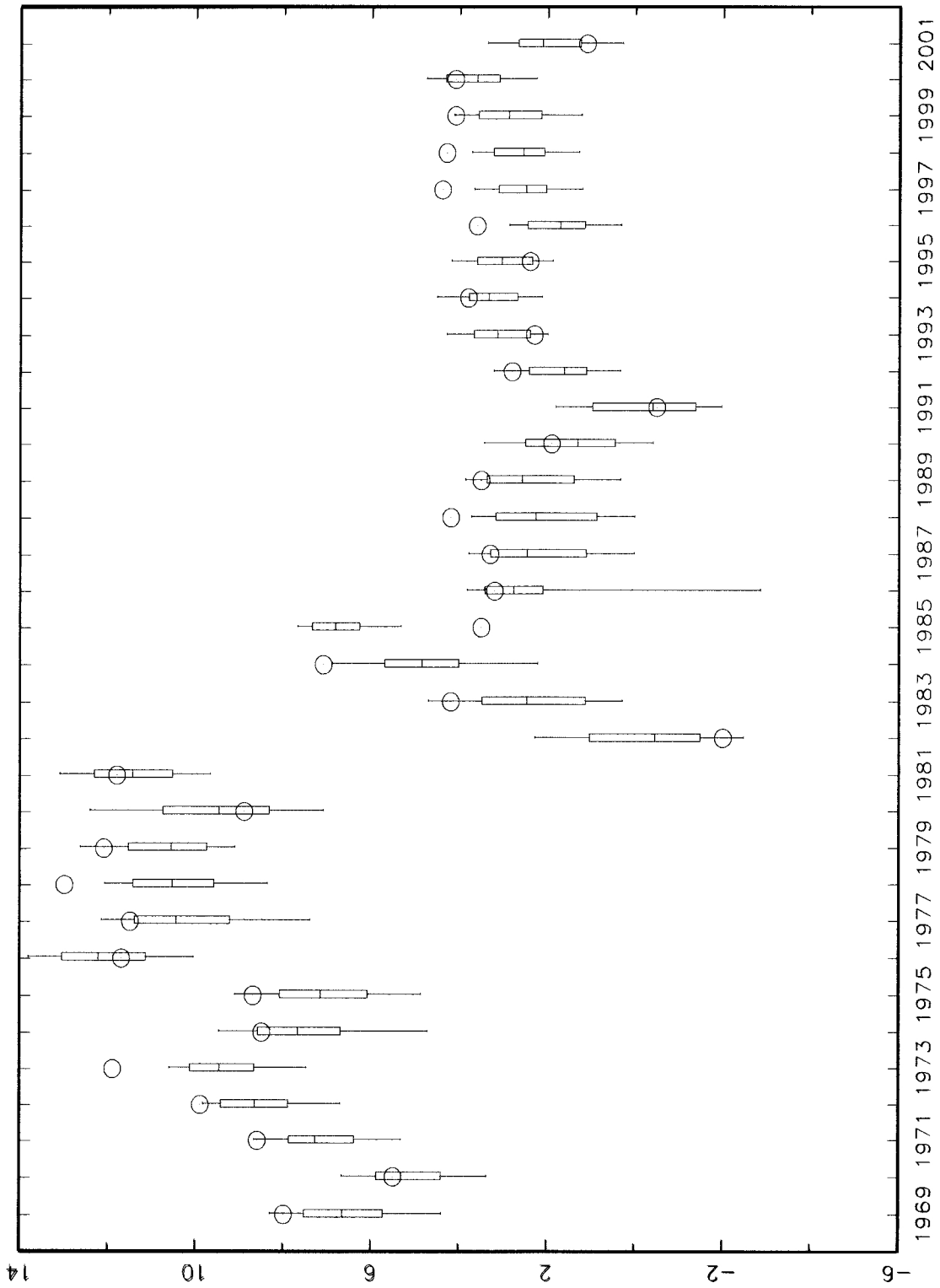


Figure 2 Output growth forecast probability distributions shown as Box-Whisker plots and realizations. The boxes represent the inter-quartile, the outer 'whiskers' the 10 and 90th percentiles, and the inner line the median. The realizations are circles with dots at the centres.



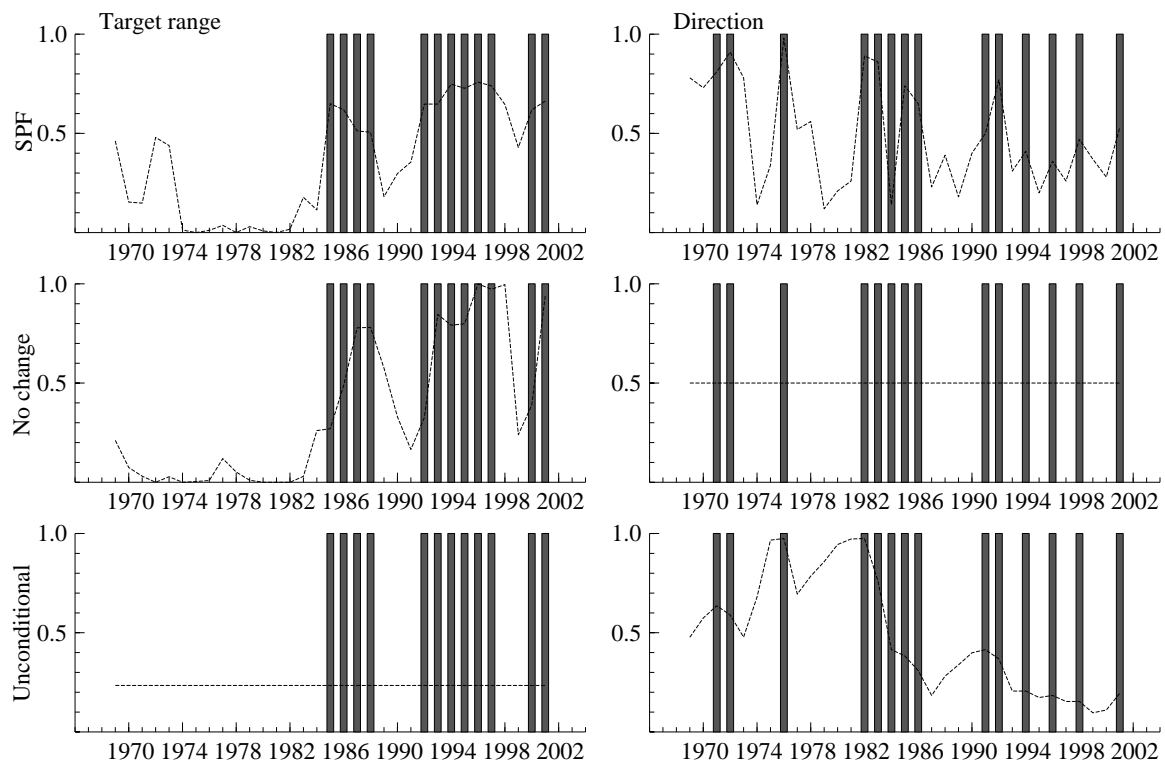


Figure 3 Time series of events ‘inflation in the range 1.5 to 3.5%’ and ‘lower inflation than last year’ and forecast probabilities of these event. Each column refers to one of the two events. The rows relate to the SPF, ‘No change’ and ‘Unconditional’ event forecast probabilities respectively. In each panel, the bars are one-zero event indicators, and the lines the forecast probabilities.

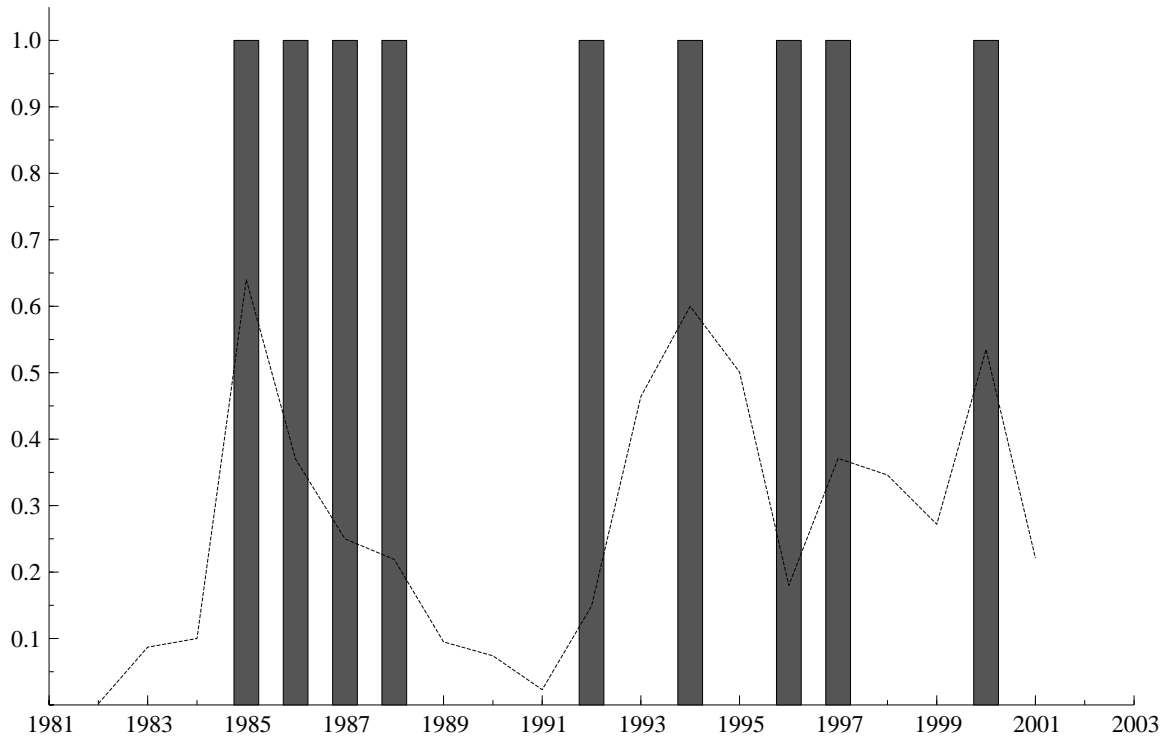


Figure 4 Time series of occurrences of event ‘inflation in the range 1.5 to 3.5% and output growth in excess of 2.5% per annum’ and associated forecast probabilities. The bars are the zero-one event indicators, and the lines the SPF forecast probabilities.

References

- Atkeson, A., and Ohanian, L. (2001). Are Phillips curves useful for forecasting inflation?. *Federal Reserve Bank of Minneapolis, Quarterly Review*, **25**, 2–11. (1).
- Berkowitz, J. (2001). Testing density forecasts, with applications to risk management. *Journal of Business and Economic Statistics*, **19**, 465–474.
- Canova, F. (2002). G-7 Inflation forecasts. mimeo, Universitat Pompeu Fabra.
- Chatfield, C. (1993). Calculating interval forecasts. *Journal of Business and Economic Statistics*, **11**, 121–135.
- Chong, Y. Y., and Hendry, D. F. (1986). Econometric evaluation of linear macro-economic models. *Review of Economic Studies*, **53**, 671–690. Reprinted in Granger, C. W. J. (ed.) (1990), *Modelling Economic Series*. Oxford: Clarendon Press.
- Christoffersen, P. F. (1998). Evaluating interval forecasts. *International Economic Review*, **39**, 841–862.
- Clements, M. P., and Hendry, D. F. (1998). *Forecasting Economic Time Series: The Marshall Lectures on Economic Forecasting*. Cambridge: Cambridge University Press.
- Clements, M. P., and Hendry, D. F. (1999). *Forecasting Non-Stationary Economic Time Series*. Cambridge, Mass.: MIT Press. The Zeuthen Lectures on Economic Forecasting.
- Clements, M. P., and Smith, J. (2000). Evaluating the forecast densities of linear and non-linear models: Applications to output growth and unemployment. *Journal of Forecasting*, **19**, 255–276.
- Clements, M. P., and Taylor, N. (2002). Evaluating prediction intervals for high-frequency data. *Journal of Applied Econometrics*. Forthcoming.
- Dawid, A. P. (1984). Statistical theory: The prequential approach. *Journal of The Royal Statistical Society, ser. A*, **147**, 278–292.
- Diebold, F. X., Gunther, T. A., and Tay, A. S. (1998). Evaluating density forecasts: With applications to financial risk management. *International Economic Review*, **39**, 863–883.

- Diebold, F. X., Hahn, J. Y., and Tay, A. S. (1999). Multivariate density forecast evaluation and calibration in financial risk management: High frequency returns on foreign exchange. *Review of Economics and Statistics*, **81**, 661–673.
- Diebold, F. X., and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, **13**, 253–263.
- Diebold, F. X., Tay, A. S., and Wallis, K. F. (1999). Evaluating density forecasts of inflation: The Survey of Professional Forecasters. In Engle, R. F., and White, H. (eds.) , *Cointegration, Causality and Forecasting: A Festschrift in Honour of Clive Granger*, pp. 76–90. Oxford: Oxford University Press.
- Doornik, J. A., and Hansen, H. (1994). A practical test for univariate and multivariate normality. Discussion paper, Nuffield College.
- Doornik, J. A., and Hendry, D. F. (2001). *GiveWin: An Interface to Empirical Modelling*. London: Timberlake Consultants Press.
- Engle, R. F., and Manganelli, S. (1999). CAViaR: Conditional autoregressive Value-at-Risk by regression quantiles. Ucsd discussion paper 99-20, Department of Economics, UCSD.
- Fackler, P. L., and Kling, R. P. (1990). Calibration of options-based probability assessments in agricultural commodity markets. *American Journal of Agricultural Economics*, **72**, 73–83.
- Fair, R. C. (2000). Testing the NAIRU model for the United States. *Review of Economics and Statistics*, **82**, 64–71.
- Garratt, A., Lee, K., Pesaran, M. H., and Shin, Y. (2001). Forecast uncertainties in macroeconomic modelling: An application to the UK economy. mimeo, Department of Applied Economics, University of Cambridge.
- Granger, C. W. J., and Pesaran, M. H. (2000). A decision-based approach to forecast evaluation. In Chan, W. S., Li, W. K., and Tong, H. (eds.) , *Statistics and Finance: An Interface*: London: Imperial College Press.

- Granger, C. W. J., White, H., and Kamstra, M. (1989). Interval forecasting: An analysis based upon ARCH-quantile estimators. *Journal of Econometrics*, **40**, 87–96.
- Greene, W. H. (2000). *Econometric Analysis (ed. 4)*: Prentice Hall.
- Harvey, D. I., and Newbold, P. (2000). Tests for multiple forecast encompassing. *Journal of Applied Econometrics*, **15**, 471–482.
- Holden, K., and Peel, D. A. (1990). On testing for unbiasedness and efficiency of forecasts. *Manchester School*, **58**, 120–127.
- Kling, J. L., and Bessler, D. A. (1989). Calibration-based predictive distributions: An application of prequential analysis to interest rates, money, prices and output. *Journal of Business*, **62**, 477–499.
- Koenker, R., and Bassett, G. (1978). Regression quantiles. *Econometrica*, **46**, 33–55.
- Koenker, R., and Bassett, G. (1982). Robust tests for heteroscedasticity based on regression quantiles. *Econometrica*, **50**, 43–62.
- Li, F., and Tkacz, G. (2001). A consistent bootstrap test for conditional density functions with time dependent data. Discussion paper, Dept. of Monetary and Financial Analysis, Bank of Canada.
- Lopez, J. (1996). Regulatory evaluation of Value-at-Risk models. Discussion paper 95-6, Federal Reserve Bank of New York.
- Mincer, J., and Zarnowitz, V. (1969). The evaluation of economic forecasts. In Mincer, J. (ed.) , *Economic Forecasts and Expectations*. New York: National Bureau of Economic Research.
- Newbold, P., and Harvey, D. I. (2002). Forecasting combination and encompassing. In Clements, M. P., and Hendry, D. F. (eds.) , *A Companion to Economic Forecasting*, pp. 268–283: Oxford: Blackwells.
- Patton, A. J. (2002). Modelling time-varying exchange rate dependence using the conditional copula. Mimeo, University of California, San Diego.
- Pesaran, M. H., and Skouras, S. (2002). Decision-based methods for forecast evaluation. In Clements,

- M. P., and Hendry, D. F. (eds.) , *A Companion to Economic Forecasting*, pp. 241–267. Oxford: Blackwells.
- Rosenblatt, M. (1952). Remarks on a multivariate transformation. *Annals of Mathematical Statistics*, **23**, 470–472.
- Shenton, L. R., and Bowman, K. O. (1977). A bivariate model for the distribution of $\sqrt{b_1}$ and b_2 . *Journal of the American Statistical Association*, **72**, 206–211.
- Tay, A. S., and Wallis, K. F. (2000). Density forecasting: A survey. *Journal of Forecasting*, **19**, 235–254.
Reprinted in Clements, M. P. and Hendry, D. F. (eds.) *A Companion to Economic Forecasting*, pp.45 – 68, Oxford: Blackwells (2002).
- Theil, H. (1966). *Applied Economic Forecasting*. Amsterdam: North-Holland.
- Wallis, K. F. (2002). Chi-squared tests of interval and density forecasts, and the Bank of England’s fan charts. *International Journal of Forecasting*. Forthcoming.
- West, K. D. (1996). Asymptotic inference about predictive ability. *Econometrica*, **64**, 1067–1084.
- West, K. D., and McCracken, M. W. (1998). Regression-based tests of predictive ability. *International Economic Review*, **39**, 817–840.
- West, K. D., and McCracken, M. W. (2002). Inference about predictive ability. In Clements, M. P., and Hendry, D. F. (eds.) , *A Companion to Economic Forecasting*, pp. 299–321: Oxford: Blackwells.