Subjective Probability Forecasts for Recessions

EVALUATION AND GUIDELINES FOR USE

By Kajal Lahiri and J. George Wang

Kajal Lahiri is distinguished professor of economics and director of the Econometric Research Institute at the University at Albany, SUNY. He has been a visiting scholar at the Social Security Administration, U.S. Department of Transportation, and the International Monetary Fund. He has received grants and contracts from the National Science Foundation, World Bank, and many federal and New York state agencies. He earned a B.A. from Calcutta University, and a Ph.D. from the University of Rochester.

J. George Wang is an assistant professor of finance at the business department of the College of Staten Island (CSI) of the City University of New York. Prior to joining CSI, he was a lead analyst of AT&T Bell Labs. He holds a Ph.D. in economics from the State University of New York at Albany and a M.A. and a B.A. in economics from Peking University in China.

Probabilistic forecasts are often more useful in business than point forecasts. In this paper, the joint subjective probabilities for negative GDP growth during the next two quarters obtained from the Survey of Professional Forecasters (SPF) are evaluated using various decompositions of the Quadratic Probability Score (QPS). Using the odds ratio and other forecasting accuracy scores appropriate for rare event forecasting, we find that the forecasts have statistically significant accuracy. However, compared to their discriminatory power, these forecasts have excess variability that is caused by relatively low assigned probabilities to forthcoming recessions. We suggest simple guidelines for the use of probability forecasts in practice.

Forecasting relatively rare business events like recessions or major stock market corrections is inherently risky, resulting in frequent misses and false signals. However, when uncertainty about future events is expressed in terms of probabilities (e.g., the probability of a recession next year is 30 percent), these forecasts are more informative and useful than purely categorical forecasts (e.g., recession or no recession) in that the probabilities can be used in the calculation of various measures of interest such as expected payoffs and downside risks. Also, because more information is imbedded in probability forecasts, there may be more scope to improve prediction.

The failure of point forecasts from large scale structural macro and VAR models or from professional surveys (e.g., Blue Chip, OECD, Survey of Professional Forecasters, National Association for Business Economics, etc.,) in predicting—or even timely recognition of—postwar recessions is well documented.[1] Admittedly, recessions that are caused by external shocks cannot, by definition, be predicted. However, the trans-

[1]See, for instance, Filardo (1999, 2004), Fintzen and Stekler (1999), and Juhn and Loungani (2002).

mission of the exogenous shocks through the economy can take some time to generate a full-fledged recession. Moreover, anti-inflationary monetary policies, that have often caused recessions in the U.S. economy, take quarters to take effect. Thus, the basic challenge is whether one can identify, at least probabilistically, an impending recession by understanding the structure of the transmission mechanism. Not surprisingly, in recent years, economists have developed advanced macroeconomic models to generate probability forecasts for business cycle turning points.[2]

However, one such model—the dynamic single index model developed by Stock and Watson (1993) —could not identify its first two out-of-sample recessions, viz., those of 1990 and 2001. Since the Stock-Watson model is built on one of the strongest scientific foundations found in the literature and on extensive use of time series data, the failures of their recession indexes represent a significant challenge for today's business cycle researchers. In explaining forecast failures, Stock and Watson (2003) painfully fall back on Leo Tolstoy in *Anna Karenina*, "Happy families are all alike; every unhappy family is unhappy in its own way." That is, econometric models typically fail to predict recessions because each recession is special in its own novel way. For example, while the decline of the stock market gave some advance warning of the 2001 recession, it was not otherwise reliable during the 1980s and the 1990s. In short, the structure of the economy changes—sometimes abruptly—and no single model specification or a set of variables can do justice to all forthcoming recessions.

Yet, recessions inflict enormous costs to society, the exact extent of which we have just begun to explore. For instance, Bangia, et al. (2002) showed that the economic capital required to capitalize a bank during a recession year is about 25-30 percent higher than that during an expansion year. Carey (2002) found that losses of a typical bank portfolio during a recession are about the same as losses in the 0.5 percent tail during an expansion. Human costs due to lay-offs and stock market declines are well known, and need no elaboration.

In this paper, we will study the usefulness of subjective probability forecasts obtained from the Survey of Professional Forecasters (SPF) as predictors of cyclical downturns. Since these probability forecasts are generated from no specific models or variables but are based on subjective probability heuristics of professional economists, there may be certain advantages in using them over models based macro forecasts (Kahnemann and Tversky, 1973). Even though the probability forecasts are available since 1968, and have drawn some media attention,[3] very little systematic analysis has been conducted to look into their usefulness as possible business cycle indicators.[4] Fortunately, there is a rich history of probability forecasts of rare events in meteorology, psychology, and geophysics, see for example, Murphy (1991), Doswell, et al. (1990), and Ogata, et al. (1994). We will utilize verification methodologies developed in these disciplines to see if the SPF probability forecasts have any value and then explore ways of reading these forecasts for monitoring cyclical downturns.

The plan of this paper is as follows: In the next sections, we will introduce the data, explain the set up, and evaluate the probability forecasts using procedures developed in other disciplines. We will also suggest simple ways to monitor and interpret time series movements in the data in terms of odds ratios and other accuracy score measures appropriate for rare-event forecasting. Finally, concluding remarks will be summarized.

The SPF Forecasts and the Joint Probability Predictor

Thanks to the ingenuity of Victor Zarnowitz, one of the world's leading scholars on business cycles, indicators, and forecast evaluation, the Survey of Professional Forecasters (SPF)[5] has been collecting subjective probability forecasts of real GDP/GNP declines during the current and four subsequent quarters since its inception in 1968.[6] At the end of the first month of each quarter, the individual forecasters in SPF form their forecasts. The survey collects probability assessments for a decline in real GDP in the current quarter, in the next quarter conditional on the growth in the current period, and so on. The number of respondents has varied between 15 and 60 over the quarters. In this study, we use probabilities averaged over individuals. The joint probability of GDP

[2]For examples of models generating probability forecasts, see Diebold and Rudebusch (1991), Hamilton (1989), Stock and Watson (1991, 1993), and Zellner, et al. (1991).

[3]*The New York Times* columnist David Leonhardt (September 1, 2002) calls the one-quarter-ahead GDP decline probability the "Anxious Index".

[4]Notable exceptions include Braun and Yaniv (1992), Wang (1993), Graham (1996), and Stock and Watson (2003). However, these studies emphasized different aspects of the probability forecasts. Baghestani (2005) suggests a way of improving interest rate forecasts available in SPF.

[5]Formerly the surveys were done under the auspices of American Statistical Association and National Bureau of Economic Research (ASA-NBER). Since 1990, the Federal Reserve Bank of Philadelphia has conducted the survey. See Croushore (1993) for an introduction to SPF.

[6]The definition of real output in the survey has changed from real GNP to real GDP since 1992:1.

declining in both the current and in the next quarter can be obtained by multiplying the two probabilities using Bayes' rule of conditional probability. Note that the product of the current and first quarter forecasts is effectively a two-quarter ahead forecast of real GDP.

Even though "two consecutive quarters of negative GDP growth" is a popular definition of a recession, the NBER defines recession using a number of monthly indicators, and thus the two do not match exactly. Using the July revisions of the real-time GDP growth, during our sample period there were six episodes of negative GDP growth in two or more consecutive quarters – those beginning 1969:4, 1974:1, 1981:4, 1982:3, 1990:4 and 2001:1. These six separate episodes of two or more consecutive quarters of real GDP declines match with five of the six NBER-defined U.S. recessions over this period. The 1980 NBER recession exhibited negative real GDP growth only in one quarter, and hence did not match with our definition of two consecutive quarters of negative real GDP growth as a cyclical downturn. Otherwise, the two definitions are very close.

It is now well accepted that the currently available (benchmark) revised data should not be used in evaluating forecasts. (Diebold and Rudebusch, 1991). We used the annualized real-time, real GDP growth issued every July as the forecasting target in this analysis, against which the forecasting performance of the proposed predictor will be evaluated. We also considered the 30-day preliminary announcements as the target variable. Except for the substantial revisions during the last recession, these two data vintages do not make much difference so far as SPF probabilities are concerned. The SPF recession probability and the real time real GDP growth are depicted in Figure 1. The shaded bars represent the NBER-defined recessions. We find that the joint probabilities rise sharply and contemporaneously during quarters with negative real GDP growths.

Validity Tests for Recession Probability Forecasts

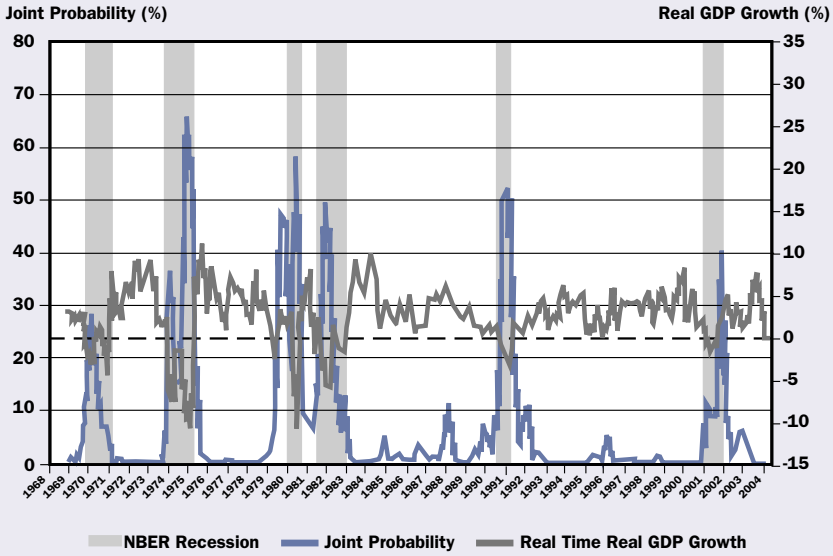
Consider the joint distribution of the probability forecasts and observations $p(f, x)$, where f is the probability forecast, and x the negative GDP growth indicator. A conventional way to evaluate probability forecasts is to calculate the mean square error (MSE), or half Brier's $QPS = \frac{1}{T}\sum_{t=1}^{T}(f_t - x_t)^2$ where f and x are the forecasts and the (0, 1) event variable, respectively. The QPS varies between 0 and 1, with 0 implying perfect forecasts. In our case, the QPS was calculated to be 0.071 suggesting impressive accuracy. However, a more meaningful measure of performance is the skill score (SS), which measures the QPS accuracy relative to a chosen benchmark. We calculated

$$(1)\quad SS(f,x) = 1 - [MSE(f,x)/MSE(\mu_x,x)]$$

to be 0.295, where MSE (μ_x, x) is the accuracy associated with the constant base rate prediction at our sample average value $\mu_x = 0.11$. The base rate is defined as the proportion of quarters with two or more consecutive quarters of negative real GDP growth in our sample (i.e., 16/143). It may be noted that using the historical average as the base rate presumes substantial knowledge on the part of the forecaster and is more demanding than the use of no-change forecast in our sample. It is interesting that the average forecast probability of a recession in our sample (μ_f) was 0.072 which is considerably less than μ_x = 0.11, suggesting under-confidence. In contrast, the rare events like earthquakes or snowstorms, are typically over-predicted (Murphy, 1991). Thus, the cost/loss structure in recession forecasting must be quite different from that of weather forecasting. Note that MSE (μ_x, x) = σ_x^2 = 0.10 and σ_f^2 = 0.12 in our sample.

FIGURE 1

JOINT PROBABILITY OF TWO CONSECUTIVE (Q0-Q1) GDP DECLINES: Q4 1968-Q2 2004



Murphy Decomposition

It is widely recognized that a single measure like QPS is grossly inadequate for evaluating the goodness of probability forecasts, particularly of rare events (Lahiri and Wang, 1994).[7] There are several features that characterize good probability forecasts. Murphy (1972) decomposed the MSE or the Brier score into three components

$$(2)\quad MSE(f,x) = \sigma_x^2 + E_f(\mu_{x|f} - f)^2 - E_f(\mu_{x|f} - \mu_x)^2$$

The first term of the right-hand-side (RHS) of equation (2) is the variance of the observations and can be interpreted as the MSE of constant forecasts equal to the base rate. It represents forecast difficulty. The second term on the RHS of equation (2) is the calibration or reliability of the forecasts, which measures the difference between the conditional mean of the occurrence in a probability group and the forecast probability. The third term on the RHS of equation (2) is a measure of the resolution or discrimination that requires significant subtleties in interpretation (Yates, 1994). In general, resolution implies that it is desirable for the relative frequency of occurrence of an event to be larger (smaller) than the unconditional relative frequency of occurrence when f is larger (smaller). Even though calibration is a natural feature to have, it is resolution that makes the forecasts useful in practice. Therefore, equation (2) can be written as:

(3) Accuracy of the forecasts = Uncertainty + Calibration - Resolution

The calibration and resolution refer to two distinct attributes of the forecast. A sequence of probability forecasts is said to be perfectly calibrated if, for all forecast values, the relative frequency of occurrence of the event for these observations associated with a particular forecast probability f, $p(x = 1 \mid f)$, is equal to that probability value f. The magnitude of any difference between the forecast probability and the frequency of the occurrence would indicate the degree of miscalibration. For perfectly calibrated forecasts, $\mu_{x|f} = f$ and $\mu_x = \mu_f$, and the resolution term equals the variance of the forecasts, σ_f^2. Resolution or discrimination (or sharpness) refers to the marginal or predictive distribution of the forecasts $p(f)$. A sample of probability forecasts is said to be completely resolved if the probability only takes values zero and one. Thus, completely refined forecasts would be miscalibrated due to the inability of the forecasters to predict the future with certainty. Conversely, well-calibrated probability forecasts generally exhibit only a moderate degree of refinement. Thus, a possible trade-off between the calibration and resolution exists to minimize MSE. Forecasts possess positive skill when resolution reward exceeds the miscalibration penalty.

The distributions of $p(x \mid f)$ and $p(f)$ are depicted in Figures 2 and 3. In Figure 2, $\mu_{x|f}$ is plotted against f, and this is referred to as the attributes diagram. The calculations are explained in Table 1. In contrast, Figure 3 depicts the marginal or predictive distribution of the forecasts $p(f)$, in which $p(f)$ is plotted against f. Figure 2 indicates the relationship between $\mu_{x|f}$ and f for the relevant sample of the forecasts and observations and also contains several reference or benchmark lines. The straight 45^0 line for which $\mu_{x|f} = f$ represents perfectly calibrated forecasts. The horizontal line represents completely unresolved forecasts, for which $\mu_{x|f} = \mu_x$ for all $f \in F$. The dotted line equidistant between the 45^0 line and the horizontal lines represents forecasts of zero skill in terms of SS. To the right (left) of the vertical auxiliary line at $f = \mu_x$ and above (below) the zero-skill line, skill is positive; and skill is negative below (above) it. Therefore, Figure 2 permits qualitative evaluation of resolution and skill as well as calibration for individual values of $f \in F$.

TABLE 1

CALIBRATION CALCULATIONS

f = probability	N = number of the observations	χ = number of realizations	$\mu_{x\|f}$ = relative frequency
0.025	100	0	0.00
0.1	23	6	0.26
0.2	7	2	0.29
0.3	1	1	1.00
0.4	4	2	0.50
0.5	6	4	0.67
0.6	1	0	0.00
0.7	1	1	1.00
0.8	0	0	0.00
0.9	0	0	0.00
0.975	0	0	0.00
Total	*143*	*16*	

An examination of the empirical curves in Figure 2 indicates that the joint probability forecast is generally well calibrated. Most points on the empirical curves fall in regions of positive skill. In Figure 3, the distribution of $p(f)$ indicates that low probability values or probability values near the historical base rate value (0.11) are used

[7]Diebold and Rudebusch (1989) introduced this measure and the Murphy decomposition in economics.

FIGURE 2

ATTRIBUTES DIAGRAM

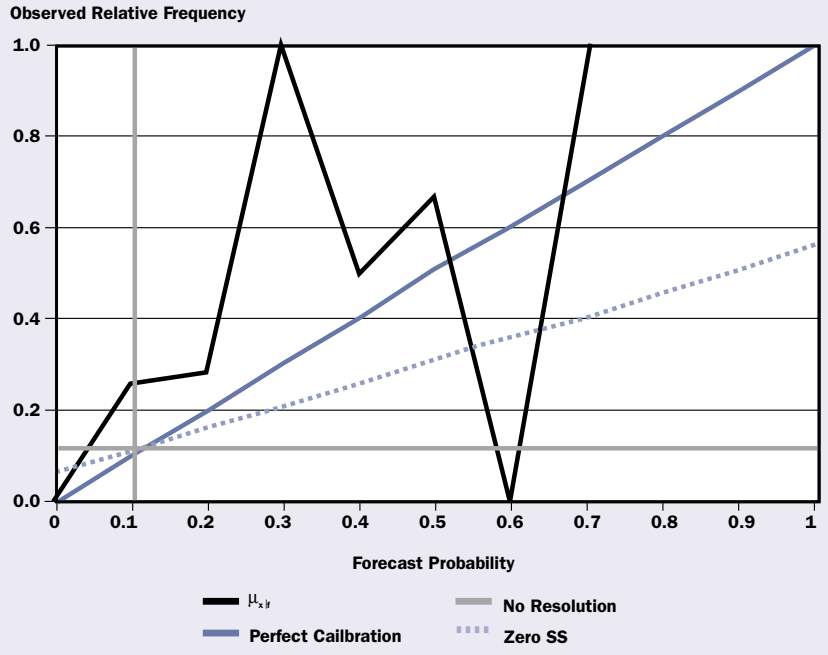
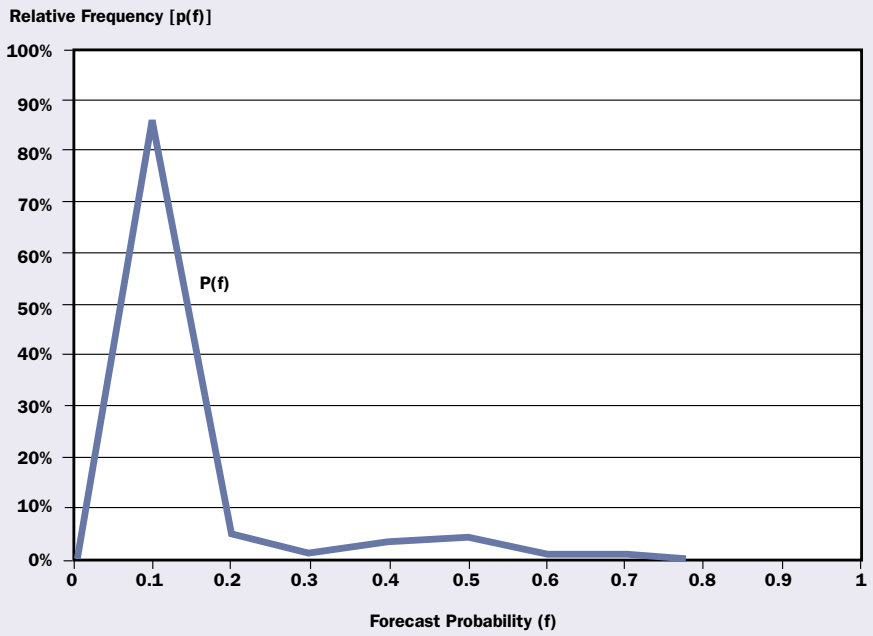


FIGURE 3

PREDICTIVE DISTRIBUTION

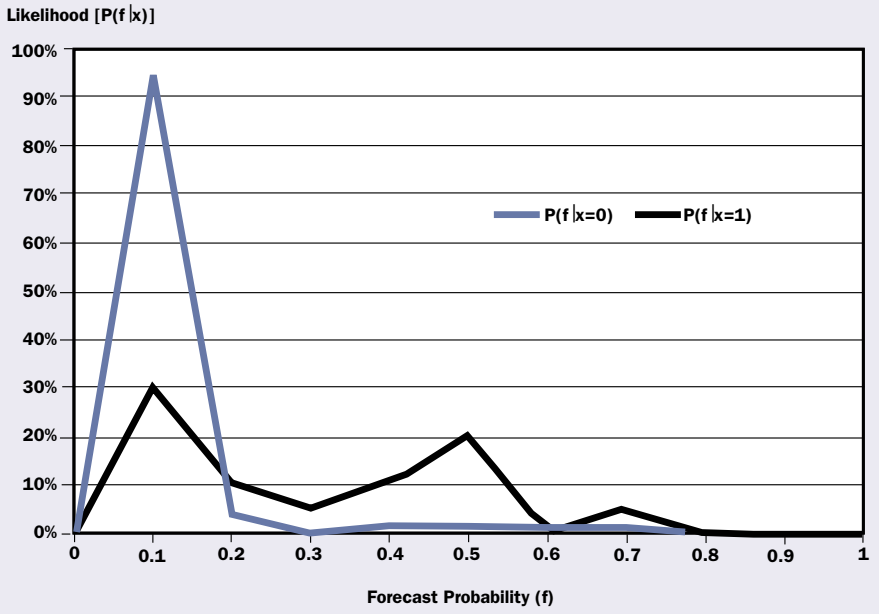


much often than high probability values. There is some accumulation of probability mass at point 0.5 that is associated with negative real GDP growth. In Figure 4 this graph is split into two conditional likelihood distributions given $x = 1$ (recession) and $x = 0$ (no recession). For these two conditional distributions, the means and variances were calculated to be (0.04, 0.01) for $x = 0$ and (0.29, 0.04) for $x = 1$, respectively. Good discriminatory forecasts will give two largely non-overlapping marginal distributions; and, in general, their vertical differences will be large. Due to the overuse of low probabilities during regime $x = 1$, the two lines overlap. However, the vertical difference at each probability value and the difference in their means (0.04 vs. 0.29) are suggestive of reasonable sharpness. Note that the difference between two conditional means, the so-called forecast slope, reflects the forecaster's ability to respond properly to cues that are predictive of the target event and to ignore cues that are not predictive of that event. It approaches one in a strong predictive situation. Numerical values of the three components in the Murphy decomposition of QPS in equation (3) were found to be 0.100, 0.013 and 0.042 respectively.

We find that the overall forecast performance, as measured by MSE, is improved by about 30 percent over the constant relative frequency forecast (CRFF) (from 0.100 to 0.071). The major contributor for the improvement in MSE is resolution, which helps reduce the baseline MSE (CRFF) by about 42 percent. On the other hand, the miscalibration increases CRFF by 12 percent. As indicated by the attributes figure (Figure 2) and the overlapping of the $p(f|x = 1)$ distribution over $p(f|x = 0)$ (Figure 4), the SPF probabilities are conservative in assigning high probability to the quarters when recession occurs. This also suggests

FIGURE 4

LIKELIHOOD DIAGRAM



that distinguishing between occurrences and non-occurrences, and assigning higher probabilities to the quarters when recession occurs, can possibly improve the resolution of the forecasts. It may be noted that the assignment of lower probability for rare events is not unusual. It is quite common in weather forecasting. When the diagnostic information or "cue" is not adequate enough to make the forecast, the tendency for the forecaster is to assign a low base-rate probability.

Yates Decomposition

In a series of influential papers, Yates (1982) and Yates and Curley (1985) showed that calibration and resolution components in the Murphy decomposition are not independent of each other and suggested a covariance decomposition of MSE that can shed additional light on the characteristics of probability forecasts. It is written as:

$$(4)\quad MSE(f,x) = \sigma_x^2 + \sigma_f^2 + (\mu_f - \mu_x)^2 - 2\sigma_{f,x}^2$$

Since the variance of x is $\mu_x(1-\mu_x)$, (4) can be transformed into a more revealing decomposition:

$$(5)\quad MSE(f,x) = \mu_x(1-\mu_x) + \Delta\sigma_f^2 + \sigma_{f,\min}^2 + (\mu_f - \mu_x)^2 - 2\sigma_{f,x}^2$$

where, $\sigma_{f,\min}^2 = (\mu_{f|x=1} - \mu_{f|x=0})^2 \mu_x(1-\mu_x)$, $\Delta\sigma_f^2 = \sigma_f^2 - \sigma_{f,\min}^2$

The outcome index variance σ_x^2 provides a benchmark reference for the interpretation of MSE. It can be shown that the Brier's score is fully determined by this term when an unskilled forecaster makes a constant forecast by setting the constant probability to the relative frequency of the outcome (i.e., the base rate). It is also an important reference point when comparing different forecasters' performance, because it indicates the degree of the difficulties of the target being forecasted.

The conditional minimum forecast variance $\sigma_{f,\min}^2$ reflects the double role that the variance of the forecast plays in forecasting performance. On the one hand, minimized σ_f^2 will help reduce the MSE; on the other hand, minimized forecast variance can be achieved only when the constant forecast is offered. The constant forecast leads to zero covariance of the forecast and event, which would increase the MSE. So the solution is to minimize the forecast variance given the covariance. This strategy demonstrates the fundamental forecast ability of the forecasters. The conditional minimum value of forecast variance is achieved when the forecaster has perfect foresight such that he or she could exhibit perfect discrimination of the instances in which the event does and does not occur.

Since $\Delta\sigma_f^2 = \sigma_f^2 - \sigma_{f,\min}^2$, the term may be considered as the excess variability in the forecasts. If the covariance indicates how responsive the forecaster is to information related to event's occurrence, $\Delta\sigma_f^2$ might reasonably be taken as a reflection of how responsive the forecaster is to information that is not related to event's occurrence. Another variation of Yates decomposition is as follows:

$$(6)\quad MSE(f,x) = \mu_x(1-\mu_x) + Scatf + \sigma_{f,\min}^2 + (\mu_f - \mu_x)^2 - 2\sigma_{f,x}^2$$

where $Scatf = (N_1\sigma_{f|x=1}^2 + N_0\sigma_{f|x=0}^2)/N$, N_1, i = 0,1 is the number of the periods associated with the occurrence (i =1) and non-occurrence (i = 0), $N_1 + N_0 = N$. So the term is the weighted mean of the conditional forecast variances.

Using the SPF probability forecast data, the components in equation (5) were computed and are presented below in parentheses as:

$$(7)\quad MSE(f,x) = \mu_x(1-\mu_x) + \Delta\sigma_f^2 + \sigma_{f,\min}^2 + (\mu_f - \mu_x)^2 - 2\sigma_{f,x}^2$$
$$(0.071) = (0.10) + (0.012) + (0.006) + (0.002) - (0.049)$$
$$Bias = \mu_f - \mu_x = -0.040 \qquad \mu_{f|x=1} - \mu_{f|x=0} = 0.248$$

The overall MSE value 0.071, which is less than constant relative frequency forecast variance 0.10, demonstrates the skillfulness of the SPF joint probability forecasts. The primary contributor to the performance is the covariance term that helps reduce the forecast variance by almost 50 percent. The covariance reflects the forecaster's ability to make distinctions between individual occasions in which the event might or might not occur. It assesses the sensitivity of the forecaster to specific cues that are indicative of what will happen in the future. It also shows whether that cue responsiveness is oriented in the proper direction.

As noted before, the forecasts exhibit some degree of overall bias (under confidence) as evidenced by $\mu_f - \mu_x = -0.04$. The conditional minimum forecast variance $\sigma^2_{f,\min}$ is 0.006. Compared to the overall forecast variance 0.018, the observed variability of SPF probability forecasts is about three times the variability that is necessary, given the difference in conditional means $\mu_{f|x=1} - \mu_{f|x=0} = 0.248$. This means the subjective probabilities are scattered unnecessarily around $\mu_{f|x=1}$ and $\mu_{f|x=0}$. The Yates decomposition shows that the primary reason for an excess MSE over the CRFF variance is the variance of the forecasts. As seen above, the forecast variance 0.018 is about three times greater than what is necessary as measured by the minimum forecast variance $\sigma^2_{f,\min}$ = 0.006. The excess variation, as measured by $\Delta Var(f)$ = 0.012 increases the CRFF variance by about 12 percent. The choice of relatively low probabilities when the event actually occurred seems to be the root cause of the inflated MSE. During 50 percent of the quarters (8 out of 16) when GDP growth was negative, the probabilities assigned were below 20 percent. In contrast, 92 percent of the quarters (117 out of 127) when GDP growth did not go down, the probabilities assigned were correctly below 15 percent. That explains why $Var(f|x = 1)$ is much bigger, about four times, than $Var(f|x = 0)$.

Overall, both the Murphy and Yates decompositions support the usefulness of the SPF probability as a predictor of the two consecutive quarters of negative real GDP growth and suggest ways of improving the forecasts. The probabilities embody effective information related to the occurrence of the event, and the overall average forecast probabilities are close to the relative frequency of the occurrence of the event. However, improvement can be made by further distinguishing factors related to the occurrence of recessions, while keeping the sensitivity of the forecasts to information that is actually related to the occurrence of GDP declines. This would imply a reduction of unnecessary variance of forecasts, particularly during GDP declines, thereby increasing resolution further.

Ogata's AIC Difference as Skill Score

Similar to the concept of skill score that measures the improvement of QPS associated with a particular set of forecasts over average base rate forecasts, Ogata (1995) developed a forecast improvement measure using the Akaike Information Criterion (AIC). For the constant forecast ($f_0 = \mu_x$),

$$(8)\quad AIC_0 = (-2)\sum_{i=1}^{n}\{x_i \log f_0 + (1-x_i)\log(1-f_0)\}$$

and for the SPF forecast (f_i),

$$(9)\quad AIC_1 = (-2)\sum_{i=1}^{n}\{x_i \log f_i + (1-x_i)\log(1-f_i)\}.$$

AIC measures how close the forecast is to the occurrence of the event, so the forecast with smaller AIC is considered to be a better fit. The difference $\Delta AIC = AIC_1 - AIC_0$ measures the quality of the SPF forecast performance over the base rate forecasts with

$$(10)\quad \Delta AIC = (-2)\sum_{i=1}^{n} Q_i$$

where

$$(11)\quad Q_i = x_i \log(f_i / f_0) + (1-x_i)\log\{(1-f_i)/(1-f_0)\}$$

indicates the size of gain or loss of the SPF probability forecast against the constant relative frequency forecast for each i. Over our sample, it was calculated to be

$$\Delta AIC = AIC_1 - AIC_0 = 60.99 - 100.23 = -39.24$$

showing that the SPF probability forecasts improve the forecast quality over the constant relative frequency forecast in a significant way as evidenced by reducing the AIC_0 by over 39 percent. This is very similar to what we found earlier using the conventional QPS-based skill score and has not been used in economics before.

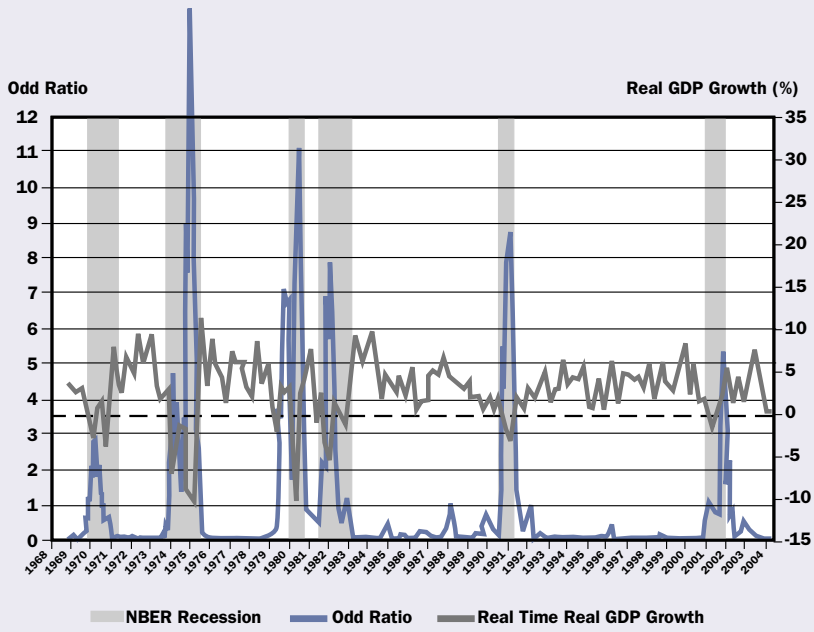
How to Use and Interpret the Probability Forecasts

One important issue for the users of the probability forecasts of rare business events, such as recessions or stock market crashes, is how to use and interpret the probabilities. Given the infrequent nature of the event and the lack of strong diagnostic information or cues, the probabilities assigned to these events are seldom very high. Psychologists have shown that individuals have a propensity to bias their estimated probabilities towards an anchor, the base rate in this case, particularly when they face difficult forecast situations. That is, in difficult fore-

cast situations, individuals do not adjust enough to new information, making the value of the anchor very critical. This is the theory of anchoring and adjustment due to Kahnemann and Tversky (1973). Whatever may be the reason, some users may tend to ignore such relatively modest probability values. Murphy (1991) suggested using odds in addition to the probabilities to communicate the probability forecasts to the users. For example, if the relative frequency of a recession during the observed sample is 0.13, and the probability forecast for the occurrence is 0.39, the occurrence of the event is less likely than its nonoccurrence for this occasion. However, it is considerably more likely this period than its relative frequency. Specifically, it is three times more likely than the average base rate.

FIGURE 5

ODD RATIO OF JOINT PROBABILITY Q4 1968 - Q2 2004



The odds or "risk" of an event is the ratio of the probability that the event occurs to the probability that the event does not occur. Thus, a recession in this case has an odds of 0.39/(1-0.39) = 0.64 (or 0.6 to 1 "on/for" in bookmaker's language). One interesting property of odds is that the odds for the complement of the event is the reciprocal of the odds for the event. Forecast information imbedded in subjective forecasts can be judged by comparing the odds of a recession using SPF probabilities to the odds of a recession corresponding to the base rate. Base rate odds (BO) in favor of the event in this case is $BO = \mu_x/(1- \mu_x) = 0.13/(1-0.13) = 0.15$. The forecast odds (FO) in favor of the event is $FO = f/(1-f) = 0.39/(1-0.39) = 0.64$. The odds ratio (OR) in favor of the event is $OR = FO/BO = 0.64/0.15 = 4.28$.

Thus, the forecast odds in favor of the event is more than four times greater than the base rate odds. According to Murphy (1991), looking at these odd ratios may overcome the problem of relatively modest forecast probabilities, and enable users to identify those occasions where the risk of the rare event is sufficiently high to warrant taking appropriate precautionary measures.

We plotted the odds ratios (OR) against real GDP growth over 1968-2004 in Figure 5. As expected, during the cyclical downturns, OR sharply increases and gives useful signals. By studying its historical behavior before two-consecutive GDP declines, we found that a threshold value in excess of 1.5 suggests a cyclical downturn until it comes down below 1.4. Using these thresholds, we found that OR provided timely and dependable warnings for negative GDP growth periods over the sample. Using real time GDP figures (July revisions), OR identified five of the six episodes of two or more consecutive quarters of negative real GDP growth in our sample period—those beginning: 1969:4, 1974:1, 1981:4, 1982:3, and 1990:4; of these, on two occasions (1981:4 and 1982:3) OR had a lead of one quarter, on one occasion (1969:4) it lagged the onset by one quarter and on the other two occasions (1974:1 and 1990:4), OR gave signals coincidentally. It should be pointed out that one-quarter lag for the two-quarter-recession beginning 1969:4 means the signal came in the middle of the episode in January 1970.

Apparently, SPF could not foresee the 2001 recession; the signal came with a lag in 2001:4 when the negative growth period had already passed (Stock and Watson, 2003). Thus, using the July-revised real time GDP data, SPF not only missed the 2001 recession, but the lagged signal has to be considered as a false signal. There are two additional false signals, one beginning 1980:1 and 1975:2. The 1980:1 false signal can be explained by the fact that 1980:1-1980:3 is a NBER defined recession even though it was not characterized by a two-consecutive quarterly fall in real GDP. We will see that if we use the real time GDP figures based on the initial 30-day announcements as the true SPF target, then the false

TABLE 2

PROBABILITIES OF A QUARTERLY DECLINE IN REAL GDP FROM THE SURVEY OF PROFESSIONAL FORECASTERS

Quarter	Target Date Actual Growth		Forecasts Made In 2000		2001			
	30-day Preliminary	Revised	Q3	Q4	Q1	Q2	Q3	Q4
2000 Q4	1.37	1.10	**7%**	4%				
2001 Q1	1.97	-0.60	13%	**11%**	37%			
2001 Q2	0.73	-1.60	16%	17%	**32%**	32%		
2001 Q3	-0.36	-0.30	17%	19%	23%	**29%**	35%	
2001 Q4	0.22	2.70		19%	18%	23%	**26%**	82%
2002 Q1	5.70	5.00			13%	18%	20%	**49%**
2002 Q2	1.06	1.30				13%	16%	27%
2002 Q3	3.10	4.00					15%	18%

Note: Forecast entries are the probability that real GDP growth will be negative, averaged across SPF forecasters.
The forecasted probability that growth will be negative in the quarter after the forecast is made (that is, the one-quarter-ahead forecast) appears in bold. Table is adapted from Stock and Watson (2003).

alarm of 1975:2 and the missed signal of the latest recession will disappear. During the five-quarter recession beginning 1974:1, SPF continued to give a high recession probability even in 1975:2 when real GDP was no longer negative according to the July-revised real time data. But according to the 30-day preliminary data, the growth was negative in 1975:2.[8]

If the SPF forecasters were predicting the 30-day revised GDP growth data, the probabilities and the outcomes would match better with these than those with July-revised growth data. There is direct evidence that SPF respondents target the growth rates based on the initial 30-day GDP announcements. In terms of QPS scores, the SPF probabilities of negative GDP growth during next six months explains the real GDP declines better when the target is defined in terms of the 30-day real time data ($QPS = 0.056$) compared to those based on July revisions ($QPS = 0.071$). This result is not entirely unexpected in view of the fact that when SPF respondents form their forecasts, they have the 30-day announcements as the most recent available information on GDP. Also, in real time, the 30-day GDP announcements are possibly more important to actual market analysts than the revised data or the NBER recession chronologies that are more academic in nature.

In Table 2, we have reproduced Table 2 of Stock and Watson (2003) where they argued that SPF recession probabilities missed the last recession. Their GDP growth figures were the revised data as of February 28, 2003. These were, however, very similar to our July revisions. We have augmented their table with real GDP growth data that were available in the real time one month after the end of the quarter (i.e., the 30-day announcements). As is well known, and can be seen from Table 2, these two actual quarterly growth rate series are remarkably different during the last recession in that there was only one quarter (2001:3) during 2000:4-2002:3 in which the GDP growth was negative if we follow the 30-day preliminary data. However, the subsequent revisions showed that during the three-quarter period 2001:1-2001:3. GDP growth was negative. Thus, if we use the 30-day preliminary real time data, we would conclude that SPF forecasters were correct in not issuing high probability assessments for a recession. However, the anemic growth during the period was signaled as the probability jumped from around ten percent in 2000:4 to 37 percent in 2001:1 and stayed that way throughout 2001. The high current quarter probability of 82 percent in 2001:4 (and hence a false signal) can be explained by the fact that the survey forecasters were

TABLE 3

CONTINGENCY TABLE

Forecast/ Event	Occur	Not Occur	Total
Yes	X = 12	Z = 8	X+Z = 20
No	Y = 4	W = 119	Y+W = 123
Total	X+Y = 16	Z+W = 127	X+Y+Z+W = 143

[8]Graham (1996) noted that momentum following, that is, repeating the same forecast in a number of quarters consecutively, was not a problem with SPF data over shorter horizons.

responding within a month of the 9/11 attack, and the high extraordinary GDP growth during the 2002:1 can be explained by the effect of 9/11 during the last quarter of 2001. Admittedly, catastrophic events like 9/11 can adversely affect judgmental forecasts, and model-based forecasts could do better in such situations.

Based on the OR values and the thresholds, we constructed a 2x2 contingency table to study the predictive content of the OR. Counts of the correct classifications, the misses, and the false alarms are presented in Table 3 with notations. The most common summary measure of verification skill using contingency tables is the so-called Kuipers' Performance Index (Granger and Pesaran, 2000). This is a frequently used measure of skill that is obtained by taking the difference between the hit rate and the false alarm rate, and can be computed from the contingency table as:

$$(12)\quad Kuipers'\ score = (xw - yz)/[(x+y)(z+w)]$$

However, in evaluating rare events where one expects the dominance of the occurrence of non-events, this index is subject to hedging behavior, (i.e., deviating from the forecaster's true beliefs in order to increase the verification score). Doswell, et al. (1990) argued that Kuiper's Index is an improper scoring rule for rare event forecasting. It should be emphasized that for rare events like recessions, the forecast value comes from correctly forecasting the rare events and not the nonevents. Following Doswell, et al. (1990), we use the Heidke skill score that is defined as the proportion of correct classifications compared to that obtained under no-skill random forecasts and can be easily calculated using numbers from the contingency table as

$$(13)\quad Heidke\ Skill\ Score = 2(xw - yz)/[y^2 + z^2 + 2xw + (y+z)(x+w)]$$

It also ranges from –1 to +1. Stephenson (2000) has suggested a skill score based on a comparison of odds of making a good forecast (a hit) to the odds of making a bad forecast (a false alarm). In other words, he suggests using the odds ratio $\theta = (H/1-H)/(F/1-F)$ where $H = x/(x+y)$, the hit rate, and $F = z/(z+w)$, the false alarm rate. This statistic has a long history in medical diagnosis. From the contingency table, this was easily calculated as $\theta = xw/yz = 44.625$. A simple skill score ranging from –1 to +1 can be obtained from the odds ratio θ by the transformation

$$(14)\quad Odds\ Ratio\ Skill\ Score\ (ORSS) = (\theta - 1)/(\theta + 1)$$

Based on data in Table 3, the Kuipers', Heidke's and Odds Ratio skill scores were calculated to be 0.687, 0.619, and 0.956 respectively. All these values suggest impressive forecast skill. Kuipers' score seems to be slightly inflated compared to Heidke score. ORSS value shows that the OR-based predictions have excellent skill score. Stephenson (2000) also suggested the use of the odds ratio to test the statistical significance of the skill. It is also well known that the log *(odds ratio) = (log x + log w – log z – log y)* is approximately normally distributed with standard error $1/(n_j)^{1/2}$ where $1/n_j = (1/x)+(1/z)+(1/y)+(1/w)$. Based on Table 3, the log odds ratio was calculated as 1.648 with standard error 0.683. Thus the value is more than 1.96 standard errors away from zero implying that there is less than five percent chance the positive skill found using the odds ratio approach could be due to pure chance.

Conclusions

In this paper we have evaluated the subjective probability forecasts for real GDP declines during 1968-2004. The Survey of Professional Forecasters record probability forecasts for real GDP declines during the current and next four quarters. Using the current and the one-quarter-ahead forecasts we generated forecasts for GDP declines during the next two quarters. By using forecast evaluation methodologies developed in meteorology, psychology, and other disciplines, we studied the quality of these probability forecasts in terms of calibration, resolution, and alternative variance decompositions. We found conclusive evidence that these forecasts possess significant skill and are acceptably calibrated and resolved. These results are similar in sprit to those found in Graham (1996), even though Graham's study was based on a small subset of the complete data set. We also found evidence that the SPF targets the initial 30-day preliminary GDP growth figures, and not the subsequently revised figures.

The variance of the forecasts, particularly during cyclical downturns, was found to be three times more than necessary. This result implies that forecasters are responsive to cues or predictors that are not related to the occurrence of negative GDP growth. Thus, forecast improvement is possible by further distinguishing factors related to the event from those that are not, while keeping the sensitivity of the forecasts to correct information. However, this may not be an easy task in practice.

Similar to the record of subjective forecasts of rare events in other disciplines, the recession probabilities seldom rose very high and were muted. Following recent climatological literature, we used odds ratios to identify signals in these forecasts compared to base line forecasts. Based on historical data, we found that an odds ratio in

excess of 1.50 signals two consecutive GDP declines quite successfully. By using skill score measures that are appropriate for rare event forecast evaluation, we found that the odd ratios have statistically significant forecasting power.

It may be noted that "over-confidence" in judgment is frequently reported in the meteorology and psychology literature of probability forecasts of rare events. This reminds one of a dictum of the 17th century French moralist La Rochefoucauld, "Everybody complains about their memory but no-one complains about their judgment." Our results, however, indicate that the recession probability forecasts are under-forecasted and lack confidence. The mean forecast probability for two consecutive declines in real GDP was only 0.0723, which is about 35 percent less than the relative frequency of the occurrence of the event, which was 0.1119 in our sample.

This under-forecasting may be a result of a different loss function in the business world compared with that in meteorology and psychology. In meteorology, rare events such as tsunamis or earthquakes occur suddenly and pass very quickly, but the damage they leave behind can be tremendous. Therefore, the cost of the missing target would be much higher than that of false signals. By contrast, a rare business event such as a recession occurs more gradually and lasts longer. Also, the effects or the course of a recession can be negated or changed by government policies. Thus, the cost of missing the target (and not reducing shipment orders, work force, etc. immediately) could be much less than the cost of a false signal (dislocations due to mistakenly laying off workers, reducing orders when it is unnecessary, etc.). So the basic costs/incentives patterns seem to be different in business situations from predicting natural disasters.

The decomposition methodologies introduced in this paper have much broader application in evaluating model fit in Logit, Probit and other limited dependent variable models. These models generate probabilities of discrete events as model predictions. Again, often in economics, we try to identify events that are relatively rare. Usually the model fit criteria based on maximized likelihood look excellent, but the estimated model hardly identifies the small but special population of interest. By using the evaluation methodology of probability forecasts analyzed in this paper, one can study the true value of many estimated limited dependent variable models for out-of-sample predictive purposes.

The application of the concept of the odds ratio to recession probabilities provides a simple but powerful monitoring scheme for impending recessions. Considering the fact that the chronologies of NBER recessions are usually determined long after the recession is over, negative GDP growth in two consecutive quarters is probably a more realistic and practical metric for tracking business cycles in real time. We have found conclusive evidence that SPF subjective probability forecasts are useful in this regard. ■

ACKNOWLEDGEMENT

We are grateful to Richard Cohen, Nigel Harvey, Thad Mirer, Herman Stekler, Kenneth Wallis, Arnold Zellner, two anonymous referees, and the editor for many helpful comments and suggestions. However, we are solely responsible for any remaining errors and omissions.

REFERENCES

Baghestani, H. 2005 "Improving the Accuracy of Recent Survey Forecasts of the T-bill Rate." *Business Economics.* 40: 2, pp. 36-40.

Bangia, A., F. X. Diebold, A. Kronimus, C. Schagen, and T. Schuermann. 2002. "Ratings Migration and the Business Cycle, with Application to Credit Portfolio Stress Testing." *Journal of Banking and Finance.* 26, pp. 445- 474.

Braun, P. and I. Yaniv. 1992. "A Case Study of Expert Judgment: Economists' Probabilities versus Base Rate Model Forecasts." *Journal of Behavioral Decision Making.* 5, pp. 217-231.

Carey, M. 2002. "A Guide to Choosing Absolute Bank Capital Requirements." *Journal of Banking and Finance.* 26, pp. 929-951.

Croushore, D. 1993. "Introducing: The Survey of Professional Forecasters." Federal Reserve Bank of Philadelphia Business Review. November/December, pp.3-13.

Diebold, Francis X. and G. D. Rudebusch. 1989. "Scoring the Leading Indicators." *Journal of Business.* 64, pp.369-391.

———. 1991. "Turning Point Prediction with the Composite Leading Index: An Ex Ante Analysis," in K. Lahiri and G.H. Moore (eds.), *Leading Economic Indicators: New Approaches and Forecasting Records.* Cambridge: Cambridge University Press, pp. 231-256.

Doswell, C. A., R. Davies-Jones, and D. L. Keller. 1990. "On Summary Measures of Skill in Rare Event Forecasting Based on Contingency Tables." *Weather and Forecasting.* 5, pp. 576-585.

Filardo, A. J. 1999. "How Reliable Are Recession Prediction Models?" *Federal Reserve Bank of Kansas City Economic Review,* pp. 35-55.

———. 2004. "The 2001 US Recession: What Did the Recession Prediction Models Tell Us?" *Bank of International Settlements, BIS Working Paper.* No 148, March.

Fintzen, D. and H. O. Stekler. 1999. "Why Did the Forecasting Fail to Predict the 1990 Recession?" *International Journal of Forecasting.* 15, pp. 309-323.

Graham, H. R. 1996. "Is a Group of Forecasters Better Than One? Than None?" *Journal of Business.* 69: 2, pp. 193-232.

Granger, C. W. and M. H. Pesaran. 2000. "Economic and Statistical Measures of Forecast Accuracy." *Journal of Forecasting.* 19, pp. 537-560.

Hamilton, J. D. 1989. "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle." *Econometrica.* 57, pp. 375-384.

Juhn, G. and P. Loungani. 2002. "Further Cross-Country Evidence on the Accuracy of the Private Sector Output Forecasts." *IMF Staff Papers.* 49, pp. 49-64.

Kahnemann, D. and A. Tversky. 1973. "On the Psychology of

Prediction." *Psychological Review.* 80, pp. 237-251.

Lahiri, K and J. G. Wang. 1994. "Predicting Cyclical Turning Points with Leading Index in a Markov Switching Model." *Journal of Forecasting.* pp. 245-263.

Murphy, A. 1972. "Scalar and Vector Partitions of the Probability Score: Part I. Two-state Situation." *Journal of Applied Meteorology.* 11, pp. 273-282.

Murphy, A. H. 1991. "Probabilities, Odds, and Forecasters of Rare Events." *Weather and Forecasting.* 6, pp. 302-306.

Ogata, Y. 1995. "Evaluation of Probability Forecasts of Events." *International Journal of Forecasting.* 11, pp. 539-541.

Ogata, Y., T. Utsu, and K. Katsura. 1994. "Statistical Features of Foreshocks in Comparison with Other Earthquake Clusters." *Geophysical Journal International.* 121, pp. 233-254.

Stephenson, D.B. 2000. "Use of the 'Odds Ratio' for Diagnosing Forecast Skill." *Weather and Forecasting.* 15, pp. 221-232.

Stock, J. H. and M. W. Watson. 1991. "A Probability Model of the Coincident Economic Indicators" in K. Lahiri and G.H. Moore (eds.), *Leading Economic Indicators: New Approaches and Forecasting Records.* Cambridge: Cambridge University Press, pp. 63-85.

———. 1993. "A Procedure for Predicting Recessions with Leading Indicators: Econometric Issues and Recent Experience." in J.H. Stock and M.W. Watson (eds.), *New Research on Business Cycles, Indicators, and Forecasting,* Chicago: University of Chicago Press, pp. 95-153.

———. 2003. "How Did Leading Indicator Forecasts Perform During the 2001 Recession?" *Federal Reserve Bank of Richmond Economic Quarterly.* 89: 3, pp. 71-90.

Wang, J. G. 1993. "On the Use of Markov Regime-Switching Model in Forecasting Business Cycles." Unpublished PhD dissertation, University at Albany-SUNY.

Yates, J. F. 1982. "External Correspondence: Decompositions of the Mean Probability Score." *Organizational Behavior and Human Performance.* 30, pp. 132-156.

———. 1994. "Subjective Probability Accuracy Analysis," in G. Wright and P. Ayton (eds.), *Subjective Probability.* Chichester, UK: John Wiley, pp. 381- 410.

Yates, J. F. and S. P. Curley. 1985. "Conditional Distribution Analysis of Probabilistic Forecasts." *Journal of Forecasting.* 4, pp. 61-73.

Zellner, A., C. Hong, and C-K. Min. 1991. "Forecasting Turning Points in International Growth Rates Using Bayesian Exponentially Weighted Autoregression, Time Varying Parameter, and Pooling Techniques." *Journal of Econometrics.* 49, pp. 275-304.