Finance and Economics Discussion Series
Divisions of Research & Statistics and Monetary Affairs
Federal Reserve Board, Washington, D.C.

Has Output Become More Predictable? Changes in Greenbook Forecast Accuracy

Peter Tulip

2005-31

Has Output Become More Predictable?
Changes in Greenbook Forecast Accuracy

Peter Tulip *

Abstract

Several researchers have recently documented a large reduction in output volatility. In contrast, this paper examines whether output has become more predictable. Using forecasts from the Federal Reserve Greenbooks, I find the evidence is somewhat mixed. Output seems to have become more predictable at short horizons, but not necessarily at longer horizons. The reduction in unpredictability is much less than the reduction in volatility. Associated with this, recent forecasts had little predictive power.

JEL classification: E37

Keywords: Predictability, Variability, Forecast Errors, Greenbook

---

* Division of Research and Statistics, Federal Reserve Board, Washington, D.C. 20551. Email: ptulip@frb.gov. I would like to acknowledge helpful comments from Dave Reifschneider, David Wilcox, Neil Ericsson, Charles Goodhart, Michael Kiley, seminar participants at the Federal Reserve Board, and, especially, John Roberts. The views presented in this paper are solely those of the author and do not necessarily represent those of the Federal Reserve Board or its staff.

Introduction

The volatility of the US economy has declined dramatically. The standard deviation of annualized changes in quarterly real seasonally-adjusted GDP declined from 1.2 percentage points in the period 1947-1983 to 0.5 percentage points in 1984-2004. This "Great Moderation" has been described as one of the most striking changes in the business cycle in recent decades (Ben Bernanke, 2004a; James Stock and Mark Watson, 2003). It is the subject of a large and growing literature, of which Margaret McConnell and Gabriel Perez-Quiros (2000), Chang Jin Kim and Charles Nelson (1999), and Olivier Blanchard and John Simon (2001) are prominent examples.

However, what matters to most people is not *volatility* but *uncertainty*. Because resources can generally be transferred from known periods of high income to those of low income, predictable variations are not a serious concern. Presumably, it is unpredictable changes that cause large welfare losses. When people cannot accurately predict the future, they make decisions that, with hindsight, turn out to be mistakes. Firms build factories when they shouldn't. Central banks raise interest rates when they should have lowered them. Resources are wasted taking precautions against events that do not occur. And so on.

The clearest evidence of the importance of uncertainty relative to volatility is the lack of interest in seasonal economic variations. Seasonal variations are huge, accounting for about 85 percent of the variability of output (J. Joseph Beaulieu and Jeffrey A. Miron, 1992, table 1). But because they are predictable, almost no-one pays attention to them (at a macro-economic level). Even the studies of so-called "volatility" use seasonally-adjusted data. They do not measure the total variation in the data; only the variation not accounted for by one specific influence. But there is no obvious reason for singling out seasonality. Just as predictable seasonal variations are appropriately removed from the data, so should other predictable influences.

If one is interested in unpredictability, one can measure it directly, as the difference between actual outcomes and what people were expecting. There are many available measures of expectations. I use the forecasts of the staff of the Federal Reserve

Board of Governors, as published in a document called the Greenbook. Differences between these forecasts and actual outcomes are the Greenbook errors.

The Greenbook errors provide a good measure of uncertainty for several reasons. Previous researchers have found that the Greenbook forecasts are more accurate than other forecasts (Christina Romer and David Romer, 2000; Christopher Sims, 2002). So they can be taken as representing the state-of-the-art or the envelope of predictability. Furthermore, the data on Greenbook forecasts is richer than for many private sector forecasts. The forecast horizon is longer and the data extend further back in time.[1]

Trends in the Greenbook errors are also interesting because of their relevance to monetary policy. As Chairman Alan Greenspan (2004, p. 8) has noted, "the success of monetary policy depends importantly on the quality of forecasting". So, from a historical perspective, changes in the quality of the forecasts might help explain changes in policy performance, to the extent that policy was guided by the staff forecasts. From a normative perspective, the accuracy of forecasts and its stability help determine the extent to which monetary policy should be "forward-looking". Lastly, if the forecast errors are stable over time then the monetary policy environment can be described as one of "risk" rather than "Knightian uncertainty". That is, we can quantify what we do not know. In particular, the distribution of outcomes about previous forecasts would provide a reliable guide to the distribution of possible outcomes about the current forecast. This is relevant both to the FOMC's assessment of risks, and (more so in other countries than in the US) the public presentation of policy.

Although the paper is indirectly motivated by these monetary policy issues, its primary focus is whether uncertainty has declined. I find that there has been a clear and large reduction in uncertainty at short horizons, but not necessarily at longer horizons. I also find that the reduction in uncertainty is much less than the reduction in volatility. Closely associated with this, recent forecasts have had remarkably little predictive power.

[1] For example, whereas the Greenbook forecasts for real GDP began in 1965, the Survey of Professional Forecasters began in 1968, DRI forecasts began in 1970, Blue Chip forecasts began in 1977, and The Wall Street Journal survey began in 1986.

Whereas the Fed predicted a large share of the fluctuations in output in the 1970s and 1980s, more recent fluctuations have been surprises.

II. Related literature

The view that unpredictability is of greater interest than volatility is not new. As noted above, almost all of the studies of volatility remove predictable seasonal influences from the data. Many others remove the predictions of a vector autoregression. Several papers in this literature – for example, Stock and Watson (2003) – explicitly discuss unpredictability.

However, insofar as measures of uncertainty are presented, it is typically in the form of the errors of an econometric model. After-the-event regression residuals are easier to compile than real-time forecast errors, and they facilitate decomposition and analysis. But otherwise, they provide an unsatisfactory measure of the uncertainty that faced decision makers in real time. On the one hand, they understate real-time uncertainty because regressions are estimated after the event and so benefit from hindsight. For example, they "know" the sample mean (unless estimated recursively) and data revisions (unless real time data is used). Unavoidably, their specifications reflect information that was unavailable to forecasters. On the other hand, they tend to overstate uncertainty because they are simple. Even the most complicated econometric models incorporate much less information than the Greenbook forecast, which reflects the pooling of many variables, models, and statistical methods by a large team of economists.

Previous comparisons suggest that the second of these biases has usually been more important. The Greenbook and private sector forecasts have been much more accurate (over a limited range of measures) than autoregressions, and slightly more accurate than large econometric models, such as MPS.[2] That is, autoregressions have tended to overstate uncertainty.

[2] Examples of forecast comparisons include Romer and Romer (2000), Sims (2002), Campbell (2004) and unpublished studies conducted by the Federal Reserve staff.

Several recent papers have analyzed real-time forecast errors, including Scott Schuh (2001), Charles Goodhart (2004), and Sean Campbell (2004). Schuh and Goodhart find some similar results to mine, using different data sets, which I note below. However, neither of these papers is directly focused on changes in the errors over time.

Campbell's work, circulated while this paper was in preparation, overlaps to a greater extent. We both find that short-horizon forecast errors have narrowed by less than the decline in output volatility. However, Campbell's focus is on differences between private sector forecasts and autoregressions, rather than assessing whether uncertainty has changed. Also, his data comes from the Survey of Professional Forecasters (SPF), whereas I use the Federal Reserve Greenbooks. Accordingly, Campbell's analysis is more relevant to private sector decision-making while mine is more relevant to monetary policy. Furthermore, the horizon of the Greenbooks is longer than that of the SPF. Partly because of this, my conclusions are slightly different. Whereas Campbell (p2) finds that "macroeconomic uncertainty … (has) exhibited a substantial decline since 1984," I find that the evidence of a reduction is mixed. At longer horizons, point estimates of uncertainty have not substantially declined.

III. Data

Before scheduled meetings of the Federal Reserve's Federal Open Market Committee (FOMC), the staff of the Board of Governors prepares a detailed forecast. This is published in a document universally, though unofficially, called the Greenbook. The purpose of the Greenbook is to facilitate the deliberations of the FOMC. The forecasts reflect the views of the staff, not the Committee members, who may hold quite different views about the evolution of the economy.[3]

The Greenbook forecasts are available at the website of the Federal Reserve Bank of Philadelphia, except for those from the last five years, which are confidential. The

[3] The Committee members report their own forecasts for GDP growth, unemployment, and inflation to Congress twice a year.

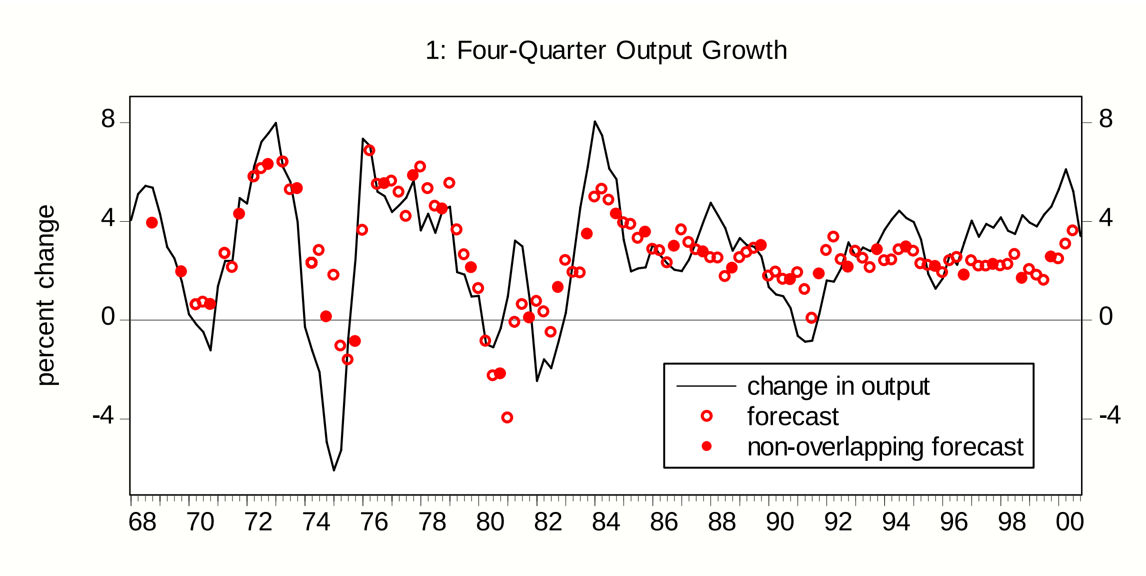
first current-quarter forecast for real GNP was published in November 1965. The forecast horizon has been extended since then, reaching four quarters (including the current quarter) in 1968, eight quarters in 1979 and ten quarters in 1990. The horizon typically rolls forward to cover a new calendar year every twelve months (currently in September). Because of this, the data are discontinuous, particularly at longer horizons. I use forecasts through November 1999, which have a horizon extending to 2001q4.

I use one Greenbook per quarter, although the actual frequency of publication is higher. I assume that the potential loss of information is outweighed by the convenience of measuring forecasts and outcomes at the same frequency. I choose the Greenbook closest to the middle of the quarter, for comparability with the spreadsheets maintained by the Philadelphia Federal Reserve. I focus on the forecast for real output, defined as GNP prior to 1991, then GDP. This series uses prices from fixed base years until 1996, then is chain-weighted.

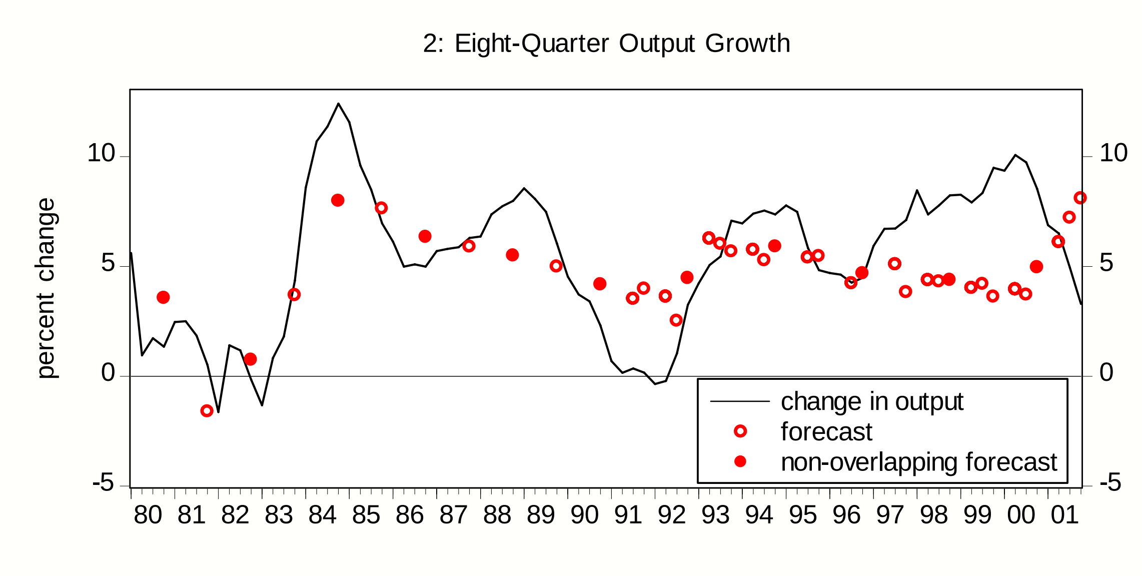
To calculate forecast errors, I compare these predictions with real-time data. Specifically, I use the GDP/GNP estimate as of the middle of the quarter two quarters after the relevant event, also available from the Federal Reserve Bank of Philadelphia's real-time data set. Hence, "truth" for, say, the change in output in the four quarters to 2000q1 is the estimate as of mid-August 2000. Typically, these estimates represent the "first final" estimate (also called the "second revision") of the BEA. These data reflect a more comprehensive analysis of source data than earlier estimates, while usually adhering to the same data definitions as at the time of the forecast.

The use of real-time data differs from the approach of Sims (2002, p7), Campbell (2004), and many others, who use latest available estimates. Using recent estimates is easier but involves treating changes in data definitions as forecast errors. There are several problems with this approach, of which two are important for my purposes. First, use of recent data would bias results toward showing that predictability has increased over time, because recent forecasts would use data definitions that were closer to the "truth" than earlier forecasts. Second, using later data definitions would make forecast errors correlated, lowering the information content of individual errors. Other reasons for preferring real-time to current data are noted in Romer and Romer (2000), Robertson and Tallman (1998) and several references cited by Schuh (2001, n.14).[4]

Charts 1 and 2 show some illustrative data. The lines show real-time measures of changes in output. The dots show corresponding forecasts, dated by time of the event, not the time of the forecast. That is, a dot that is close to the line represents an accurate forecast. (The filled-in dots represent a set of non-overlapping forecasts I use in the Appendix). Chart 1 shows four-quarter changes, with forecasts for the current quarter and three following quarters. Chart 2 shows eight-quarter changes (the current quarter and seven following quarters). Elsewhere, these forecasts are sometimes called "three-quarter ahead" and "seven-quarter ahead" forecasts.



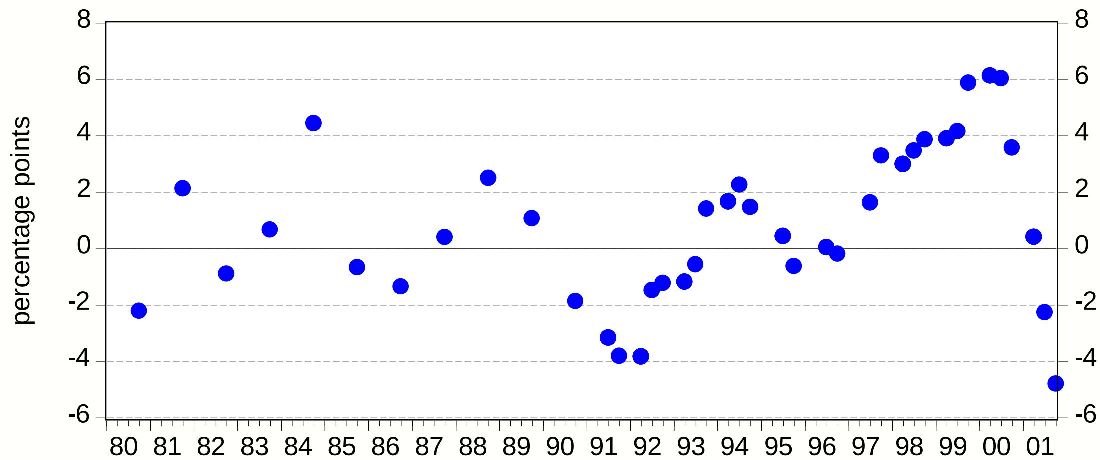
[4] Other approaches would also be possible. For example, one could use even earlier estimates, such as the advance or preliminary NIPA. These are based on incomplete data augmented by the BEA assumptions. Hence "forecast errors" measured on this basis reflect the extent to which the forecaster shares the BEA's assumptions, rather than consistency with actual economic conditions. Another possibility would be to only use data defined exactly the same way as the forecast, excluding observations at the time of benchmark revisions. However, this would substantially reduce the number of long-horizon errors in my sample. In practice, one-off changes arising from redefinitions to GDP are small relative to overall forecast errors. For example, the root mean squared difference between the current measures of four-quarter changes in GNP and GDP between 1991q3 and 1993q4, the period affected by this change in definition, is 0.13 percent, tiny relative to the 4-quarter RMSE, 1.6 percent. Of course, were this discrepancy to be applied to all previous forecasts (as in the use of latest available data) its effect would cease to be trivial.



As the charts show, there were large swings in activity in the 1970s and 1980s. Interestingly, the Fed staff anticipated a substantial share of these. But more recently, the staff missed the boom of the 1990s and subsequent downturn. Schuh (2001, Figure 1) shows a similar deterioration in the performance of the Survey of Professional Forecasters.

Forecast errors are simply the difference between the forecast and outcomes. For illustration, chart 3 shows eight-quarter errors. Note that they have tended to grow larger over time, with the errors made over the most recent business cycle being especially bad. I discuss this point further below.

3: Eight-Quarter Output Errors

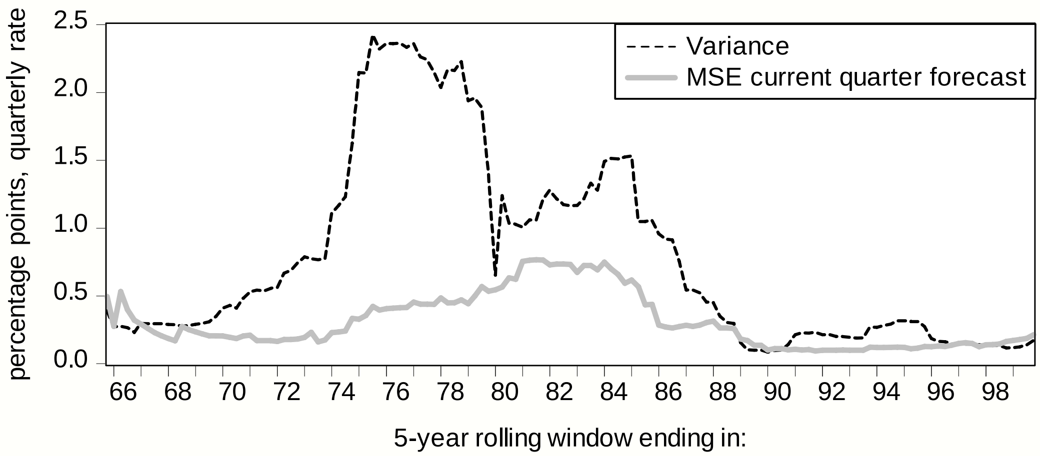


IV. Changes in Predictability

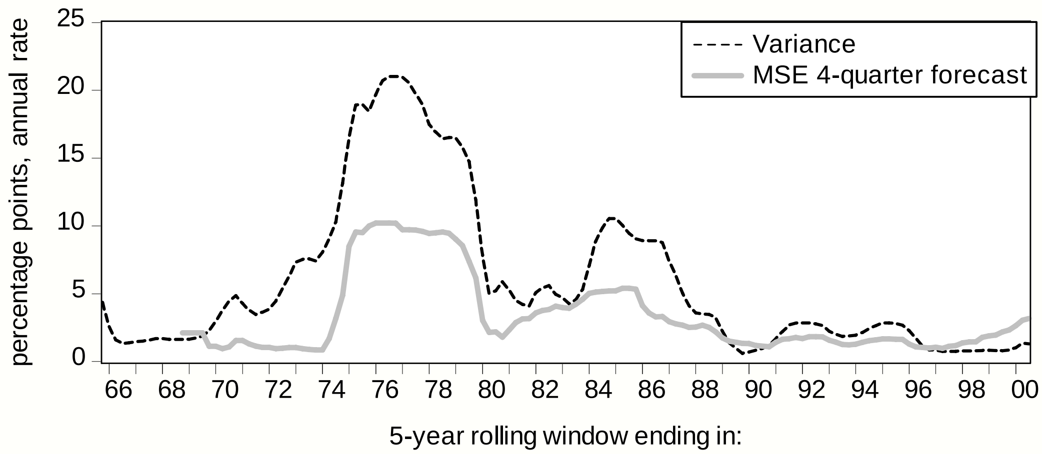
The main results in this paper are presented in charts 4, 5, and 6. Each chart shows two series. The dashed black lines show the variance of output growth. The solid grey lines show unpredictability, measured as the Mean Squared Error (MSE) of forecasts of output growth. Both series are measured using 5-year rolling windows, following the approach of Olivier Blanchard and John Simon (2001). Note that the sample of forecast errors is incomplete; the MSEs are calculated using whatever observations are available within the window.

The charts differ by forecast horizon (and, accordingly, by the frequency with which changes in actual output are measured). Chart 4 shows current-quarter errors, chart 5 shows four-quarter errors, and chart 6 shows eight-quarter errors. These three charts are representative of other horizons. The forecast errors can be interpreted as applying to both the change and the level of GDP, relative to its level in the previous quarter.

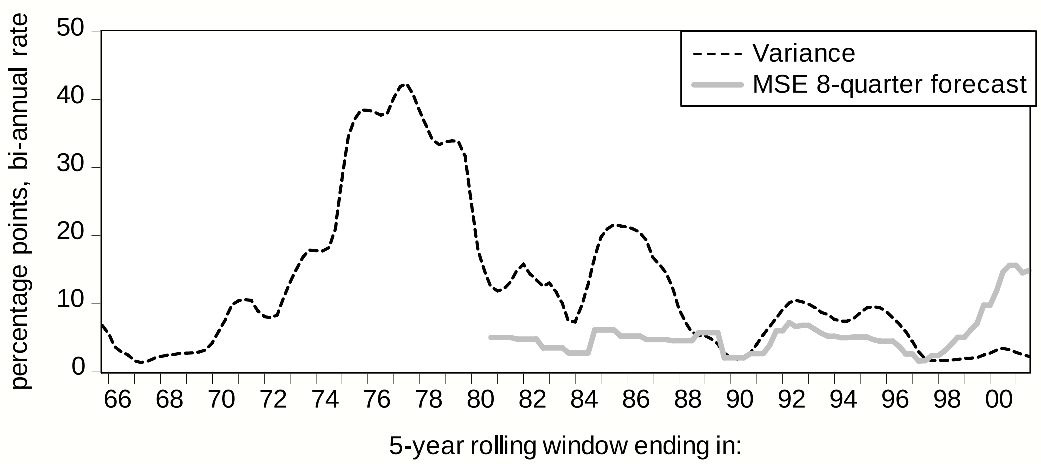
4: Variance and Unpredictability of Quarterly GDP Growth



5: Variance and Unpredictability of 4-Quarter GDP Growth



6: Variance and Unpredictability of 8-Quarter GDP Growth



The variances and MSEs shown in the charts are algebraically related. Let y_t represent actual output growth in quarter t and f_t its forecast. The forecast error is then $e_t = y_t - f_t$. I use the same real-time measure of y_t in both the MSE and the variance. Rearranging, subtracting the sample mean $\bar{y}$ from each side, squaring and averaging over n quarters (n = 20 for a five-year window), gives:

$$\frac{1}{n}\sum_{t=1}^{n}\left(y_t-\bar{y}\right)^2 = \frac{1}{n}\sum_{t=1}^{n}e_t^2 + \frac{1}{n}\sum_{t=1}^{n}\left(f_t-\bar{y}\right)^2 + \frac{2}{n}\sum_{t=1}^{n}\left(f_t-\bar{y}\right)e_t$$

$$\text{Variance} = \text{MSE} + \text{predicted variation} + \text{covariance}$$

The distance between the two lines in each chart equals the sum of the last two terms in the equation. Loosely speaking, this can be called the predictable component of output growth. Strictly speaking, this requires that the covariance be small, which is not always the case. This is in contrast to after-the-event econometric analysis, where the event y is known before its prediction f. Then error-minimization means the covariance of predictions and errors is zero (otherwise, errors could be reduced by changing the prediction). But when f is determined before y, as in forecasting, the forecast does not minimize errors (though it tries to) and the covariance need not equal zero.

There are four key points evident in charts 4, 5, and 6. These are examined in more detail in subsequent sections:

1) As the literature on the Great Moderation has documented, the variance of output growth declines substantially, in the sense that it has been much smaller in the last two decades than it was in the previous two decades.

2) In contrast, the trend in unpredictability is less clear. Although mean squared prediction errors of short-horizon forecasts tend to be larger before the early 1980s than after, the change is not as large or obvious as for the variance, and is more sensitive to timing. Moreover, the eight-quarter forecast errors seem to trend up, albeit over a shorter sample period.

3) The predictable component of output growth has virtually disappeared. Although output was highly variable in the 1970s and early 1980s, most of this variation was predicted. In contrast, variations since the late 1980s have been surprises.

4) Indeed, recent Mean Squared Errors have been larger than variances, particularly at longer horizons.

Much of the literature on volatility (with several exceptions) has focused on changes at a quarterly frequency. Previous analyses of Greenbook forecasts (for example, Romer and Romer, 2000, or Sims, 2002) have also tended to emphasize quarterly changes at different horizons (the three-quarter ahead forecast of quarterly GDP growth, the four-quarter ahead forecast of quarterly GDP growth … and so on). Chart 4, which shows quarterly changes and errors, is included to permit comparisons with this research and because of its slightly longer span of data.

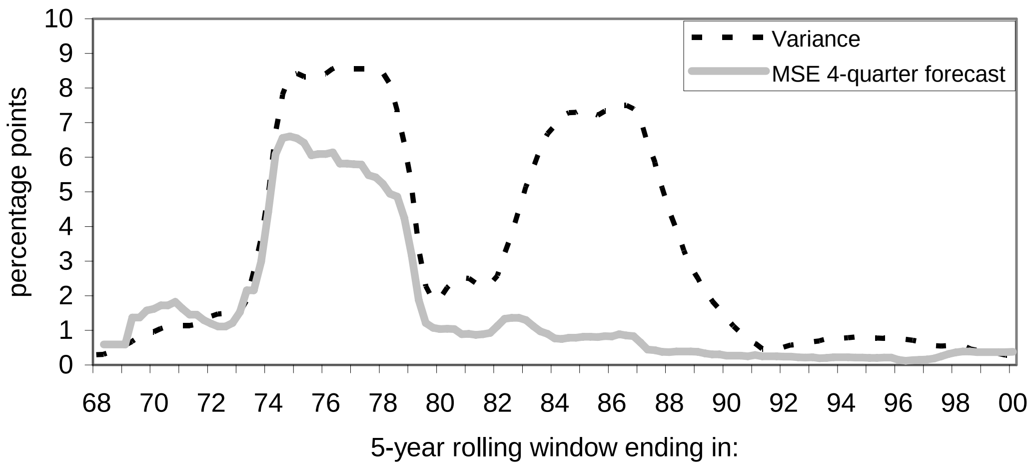
However, measurement at this frequency places equal weight on transient and persistent errors. But an error that is reversed the following quarter is less important than one that is sustained. Accordingly, the cumulative sum of errors over multiple quarters, shown in charts 5 and 6, is more interesting for most purposes. For example, it is more closely related to intermediate objectives of monetary policy, such as the level of the output gap. Moreover, the much greater magnitude of longer horizon errors can be seen from the scaling of the vertical axes of the charts. Focusing on errors beyond the current quarter also reflects the perspective of central bankers. Bernanke (2004b, p4) and Goodhart (2004, p5) suggest that many monetary policy decisions are based on the inflation outlook 7 to 8 quarters ahead.[5] However, one limitation of the longer-horizon forecasts is that they are available for a shorter sample, as noted above. Another limitation is that they overlap more, so provide fewer independent observations for testing hypotheses.

---

[5] In personal communication, Charles Goodhart suggests that monetary policy should focus on inflation two years ahead and output one year ahead, reflecting corresponding lags in the transmission mechanism.

A detailed analysis of inflation uncertainty is outside the scope of this paper. Nevertheless, chart 7 shows 4-quarter changes and errors for the GNP/GDP deflator, for comparison. The inflation story is somewhat different from that for output in several respects. A reduction in uncertainty is clearer. And the outlier of 1974 (discussed in section VIII) has a greater effect. Notwithstanding these differences, points 1, 3, and 4 above apply to inflation as well as to output.

7: Variance and Unpredictability of Inflation
(4-Quarter Change in GDP deflator)



V. Changes in Unpredictability

One possibly surprising feature of charts 4, 5 and 6 is how little the MSEs have changed over a fairly long period of time. The errors made over the last decade appear similar to those made in the late 1960s and early 1970s. It may seem as though the Fed has not learned anything about forecasting output over the last three decades.

But a simple comparison of errors at the beginning and end of the sample may be misleading. These observations may not be representative of overall performance. For a more comprehensive assessment of longer term changes, I examine whether uncertainty declined after 1984. Several papers have concluded that a discrete break in output volatility occurred about then, so a natural question is whether there was a similar break in uncertainty. Of course, this is just one of many possible ways that instability in the errors could be measured. One could also look for deterministic or stochastic trends or a discrete change at other, possibly unknown, breakpoints. At first glance, a deterministic trend would seem to provide a poor description of the MSE series in charts 4, 5 and 6. Examination of other alternatives would require more detailed analysis.

Table 1 shows MSEs before and after 1984q1, and their ratio. These measures are presented for each horizon of errors through eight quarters. As the column of ratios indicates, there was a sizeable reduction in uncertainty after 1984. In particular, short-horizon Mean Squared Errors were approximately two-fifths as large after 1984 as before. However, at horizons of six or seven quarters, the change in the size of errors is relatively small. Eight-quarter errors *increased* substantially, being three times as large after 1984 as before.

Table 1: Forecast Mean Squared Errors (MSEs)

Before and after 1984

Horizon	Earliest to 1984 (a)		1984 to 2001 (a)		Ratio	*p*-value
	MSE (b)	Observations	MSE (b)	Observations	(c)	(d)
1 (current) quarter	0.4	74	0.2	64	.39	0.008%
2 quarters	1.6	68	0.6	65	.37	0.02%
3 quarters	3.3	61	1.2	66	.35	0.16%
4 quarters	5.4	54	2.0	67	.37	0.68%
5 quarters	7.9	42	3.1	68	.40	2.3%
6 quarters	7.5	22	4.7	63	.64	18.0%
7 quarters	6.1	8	6.4	53	1.04	44.6%
8 quarters	2.7	4	8.5	36	3.19	84.6%

(a) Errors are dated by the time of the event, not the time of the forecast. The last forecast used was made in November 1999, for the period 1999 to 2001, depending on the horizon.
(b) Mean Squared Error, in percentage points
(c) Ratio of later MSE (column 4) to early MSE (column 2)
(d) probability of a ratio smaller than that observed, under the null hypothesis that variances are equal. Constructed by Monte Carlo, as described in the text.

Do these differences mean that the distribution of errors changed? Or could they be due to chance? The final column helps answer this question, but needs some explanation.

If forecast errors are unbiased and normally, identically, and independently distributed, then the ratio of their mean squared errors, before and after a given breakpoint will have an exact F-distribution, with degrees of freedom equal to the number of errors in each sub-sample. One can confidently reject the hypothesis that MSEs are equal if the F-statistic is very different from one.

The assumption that errors are unbiased seems a reasonable approximation. Over different horizons, the smallest p-value for a two-tailed t-test of the hypothesis that the mean error is zero is 13 percent (for seven-quarter errors). And even if bias were to become evident in a small sample, the staff would presumably react to remove it from the population.

A more serious concern is that forecast errors are not normally distributed. For example, the four-quarter errors fail a Jarque-Bera test for normality with a p-value of less than 0.0001 percent. This matters because, with a fat-tailed distribution, a reduction in measured uncertainty might simply reflect some large outliers fortuitously falling before the breakpoint, rather than after. However, the non-normality of the forecast errors can largely be attributed to their overlapping nature, which means that unusually large errors, such as 1974, tend to be repeated more often than would normally occur. The residuals from modeling the 4-quarter errors as an MA(3) process (discussed below) have a Jarque-Bera p-value of 3 percent. This suggests that, once serial correlation is removed, the errors are approximately normal. Nevertheless, as I discuss in Section VIII, removing the outliers of 1974 from my sample weakens the evidence of a reduction in uncertainty.

The main difficulty with conducting F-tests is that they require that the errors be independent. For a forecast with a horizon of h quarters, the outcome becomes known (by assumption) in $h + 2$ quarters. If forecasts are efficient, that error will be uncorrelated with the errors in following forecasts. However, because the forecast horizon exceeds the frequency of observation, the forecasts overlap, and these overlapping forecasts will be correlated: a surprise that causes an error for one forecast will also contribute to errors for all forecasts already made but for which the data are yet to be realized. Accordingly the Greenbook errors could, in principle, have a MA($h+1$) structure.

In practice, the errors can reasonably be described as only MA(h-1). Although the previous quarter's error is not exactly known when the current quarter is being forecast, it seems that that information would not actually be helpful. The h'th coefficient in moving average regressions is typically near zero. For example, the correlation of current-quarter forecast errors with their one-quarter lag is only 0.02. So for current-quarter forecasts, serial correlation can be ignored. For multi-quarter forecasts, it is only the overlap of the events that needs to be controlled for, not the overlap of the forecasts.[6]

Serial correlation can be dealt with in different ways. In the Appendix, I use a sample of non-overlapping errors. Non-overlapping errors are simple to construct and interpret, and provide a test statistic with a distribution that is both standard and exact. Accordingly, they provide a natural "first cut" at the data. However, this approach involves disregarding a lot of relevant information. Surprisingly perhaps, much more powerful tests can be constructed by Monte Carlo.

Specifically, I estimate an MA(h-1) model for each horizon of errors over the longest continuous sample available. I then use these estimated coefficients, together with random draws from a standard normal distribution, to construct a sequence of moving averages. I then impose the same frequency of missing observations as my sample and draw a sub-sequence that has the same timing and size as my forecast errors.

[6] Correlation among variances is another potential concern. However, the correlation between squared current-quarter forecast errors and their one-quarter lag is only 0.09. There are signs of higher order autoregressive heteroskedasticity, but I prefer to view that as a possible explanation of any rejections of constant variance.

I then take the ratio of the mean square of these artificial errors, before and after 1984q1. I store that ratio and repeat 100,000 times.[7]

The final column of table 1 shows how much of the distribution of these artificial ratios lies to the left of the actual ratio of mean squared errors. The *p*-values for short horizons are very low, indicating that the reduction in MSEs after 1984 is difficult to attribute to sampling variability. More likely, the distribution has changed. In contrast, the increase in 8-quarter errors, with a *p*-value of 15 percent (≈1-.846) on a right-tailed test, could easily be a fluke, reflecting a few lucky forecasts in the early 1980s.[8]

A third approach that could be used to assess statistical significance of a change in uncertainty is to regress the squared errors on a constant and post-1984 dummy. The coefficient on the dummy equals the change in the variance of the errors. A conventional *t*-test on this dummy would imply a *p*-value of 0.6 percent for current-quarter errors. This is about 80 times larger than the *p*-value estimated by the variance-ratio tests above. Similarly, Newey-West *t*-statistics on dummies in regressions for multi-quarter forecasts also have much higher *p*-values than those in table 1. Using absolute errors, instead of squared errors, results in even larger differences. Because, given their assumptions, the variance-ratio tests are exact, this implies that the regression approach lacks power. The reason is that the residuals, being *squared* errors, have a highly skewed distribution, approximately $\chi^2(1)$. Whereas variance-ratio tests reflect this, conventional least squares inference assumes incorrectly that the residuals are normally distributed. This turns out

---

[7] One complication in this approach is that beyond a horizon of six quarters, long continuous series of errors are not available, which makes estimation of moving average models difficult. However, regression coefficients tend to increase approximately linearly as the horizon of errors and the order of the moving average expand. Accordingly, it seems reasonable to extrapolate coefficients for the seven and eight quarter errors, based on the coefficients from other regressions. The results are not sensitive to how this is done, partly because significance levels at these horizons are far from marginal.

Indeed, overall the results are not sensitive to most specification choices. For example, varying the estimation period, using higher-order moving average models, or not using backcasting can change individual coefficients, but leaves overall results little changed. When extreme parameter values are chosen, for example setting the moving average coefficients to zero (assuming independence) or to 2, the quantitative results change, but qualitative results are fairly similar.

[8] This interpretation is reinforced by the anomalous reduction in pre-84 MSEs when the horizon extends from seven to eight quarters. However, even if the eight-quarter MSE were slightly larger than the seven-quarter MSE, it would remain smaller than the post-84 eight-quarter MSE.

to be a very poor approximation. One implication of this is that simple modifications of regressions, such as Andrews-Ploberger tests for instability at unknown dates, would also give distorted results.[9]

Table 2 compares the reduction in uncertainty after 1984 with the reduction in volatility. The change in MSE from table 1 is reproduced in the second column. The third column shows the corresponding change in the variance of output growth. The variance measures differ from the MSEs solely in that they substitute the mean for the Greenbook forecast. The number and timing of observations are the same as those in the MSE calculations in the second column. The reduction in volatility after 1984 is large and relatively uniform across changes of different frequencies. This contrasts with the reduction in uncertainty, which is much weaker. The ratio of MSEs is typically more than twice as large as the ratio of variances.

To assess the statistical significance of these differences, I repeat the Monte Carlo exercise of table 1. However, in place of a null hypothesis that the variance is stable, I now test the hypothesis that it falls by as much as the variance of output growth. That is, after allocating initial draws of random numbers to a quarter, I multiply those after 1984 by the square root of the variance ratios in table 2. Then, as before, I compare actual MSE ratios with the distribution of artificial ratios of mean squared moving averages, but this time using a right-tailed test. As the final column of table 2 shows, *p*-values from this experiment tend to be significant. One can reject the hypothesis that uncertainty has fallen by as much as the reduction in volatility.

---

[9] Perhaps for this reason, McConnell and Perez-Quiros bootstrap their Andrews-Ploberger tests. However the bootstrap is harder to apply to my data set, because it is serially correlated and the frequency of observation increases over time. If one simply took draws at regular intervals, the results would be similar to those for the non-overlapping sample, discussed in the Appendix.

Table 2: Uncertainty versus Volatility

Before and after 1984

	Change in Uncertainty post-84 MSE / pre-84 MSE	Change in Volatility post-84 variance / pre-84 variance	*p*-value of difference
1 (current) quarter	.39	.18	0.10%
2 quarters	.37	.18	0.76%
3 quarters	.35	.18	2.9%
4 quarters	.37	.20	7.3%
5 quarters	.40	.26	19.8%
6 quarters	.64	.33	14.4%
7 quarters	1.04	.28	8.9%
8 quarters	3.19	.31	2.1%

The greater reduction in volatility than in uncertainty corresponds to the dramatic decline in the predictable component of output variations, shown in the earlier charts. The reasons for this decline are not clear but are an interesting subject for speculation. One possibility is that early fluctuations were heavily influenced by changes in monetary policy, the effects of which were relatively predictable.

VII. Recent forecasts have been bad

One surprising feature of Charts 3 and 6 is that eight-quarter errors have tended to increase over the last two decades. Although not shown, a similar trend is evident, though not as marked, in six-quarter and seven-quarter errors.

This may be relevant to historical assessments of monetary policy. For example, Bernanke (2004b n.11) suggests that “technical improvements in modeling and forecasting" may help to explain the favorable performance of the economy lately. This has some basis if one is comparing macroeconomic performance before and after the early 1980s. However, the failure of longer-horizon errors to narrow since then suggests that it would not account for more recent successes of monetary policy. For example, the mildness of the 2001 recession does not represent a success of economic forecasting. The forecasts made in the lead-up to this recession, at horizons of over five quarters, produced the largest errors since 1975.

A greater concern than the trends in unpredictability is the observation that, since the late 1980s, mean squared prediction errors have been similar to, and sometimes greater than, the variance. Put another way, the sample mean has provided a more accurate guide to GDP growth than the actual forecasts. So, given the mean, the forecast can be characterized as uninformative or even misleading. It can also be characterized as having a zero or negative R^2.

To illustrate the point slightly differently, consider a regression of the two-year change in GDP (the solid line in chart 2) on a constant and the corresponding forecast (the dots in chart 2). This regression generates a *negative* coefficient on the forecast when estimated over the last decade of my sample (from 1992 through 2001). So when the staff predicted that output growth was likely to be high, it actually tended to be low; when the prediction was low, actual growth tended to be high.

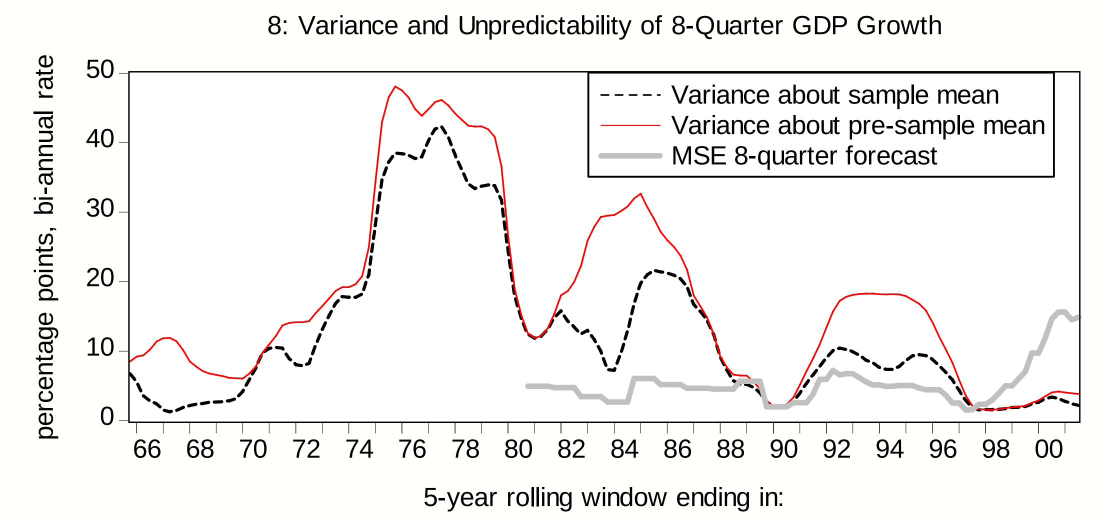
This experience is not confined to the performance of the Federal Reserve staff. Campbell (2004, p.9) finds that short-term output forecasts of the Survey of Professional Forecasters also have a negative R^2 over the period 1984-2003. Goodhart (2004, table 5) reports negative coefficients on most of the Bank of England's longer-horizon output forecasts for 1998 to 2003.

The disappearance of the predictable component of real GDP growth seems to contradict the finding by Blanchard and Simon (2001), Stephen Cecchetti et. al. (2005), and others, that there has not been a change in the dynamics of US output. Reasons for this difference are not clear. Perhaps it is the use of hindsight or revised data. Or perhaps simple autoregressions, such as those used in these studies, lack the power to detect important structural changes.

The disappearance of the predictable component of real GDP growth also has implications for the debate over whether monetary policy should be guided by simple feedback rules or whether it should be based on economic forecasts. For a survey of this debate, see Bernanke (2004). Advocates of feedback rules (sometimes called "backward-looking" policy) suggest that forecasts are unreliable and give rise to over-confidence and hence policy errors. In contrast, proponents of "forward-looking" policy argue that we have some ability to forecast the economy, and that this information is useful. The evidence from the 1990s offers more support to the advocates of feedback rules than to those of forecast-based policies.

The poor recent performance highlights a danger in "forward-looking" policies. However, it is not typical. When estimated over the full sample for which two-year forecast errors are available (that is, 1980-2001), the Mean Squared Error of eight-quarter forecasts (7.9) is appreciably smaller than the variance of eight-quarter output growth (9.7). Calculations for shorter-horizon forecasts, extending further back in time, show larger differences. Similarly, when the regression noted four paragraphs above is estimated over the 1980-2001 period, the forecast has a coefficient of 0.95 (with Newey-West standard error 0.25), which is correctly signed and not statistically or economically different from one. So, over larger samples, the forecast seems to have been a useful guide, on average. The poor recent performance may simply represent a run of bad luck.

Recent forecasts have been bad both relative to history and in the sense that they have had a negative correlation with actual outcomes. However, it does not necessarily follow from this that the Fed staff should have done better. In comparing MSEs with variances, one is implicitly using the sample mean as a benchmark. This approach is appropriate for many purposes, but it does not represent a forecasting rule that was available when the forecasts were made. To address this, chart 8 reproduces the information from chart 6, together with a thin solid line labeled "variance about pre-sample mean". This line shows the squared deviation of output growth over the preceding five years from its pre-sample mean, where that mean is taken from 1947 through to the beginning of the five-year window, measured with real-time data. It represents errors from the simplest possible forecasting rule: projecting forward the historical mean. The picture is not very different from chart 6. Since the early 1990s, the variance about the pre-sample mean is still much smaller than the forecast MSE,[10] implying that the naïve alternative would have generated more accurate forecasts than recent Greenbooks. In contrast, the Greenbooks still tend to do better before the early-1990s. Hence, on the question of whether the Fed staff could have done better at the time, this evidence is mixed.



VIII. The Outliers of 1974

In May 1974, the staff forecast that GNP would grow 1.8 percent in the four quarters to 1975q1. In the event, it fell 6.1 percent. This 7.9 percentage point error is 4.2 times as large as the standard deviation of four-quarter errors. In a large sample of normally distributed errors, such an extreme event would happen once every 34,000 observations.

---

[10] This is evident in the chart, given that 1997 observations represent the five-year period since 1992.

This outlier, and the forecast errors it overlaps, account for some unusual features of my results. For example, in chart 5, they give rise to the jump in the 5-year Mean Squared Error of four-quarter forecasts in 1974 and collapse 5 years later. This effect is even more pronounced in the inflation errors shown in chart 7. In table 1, they explain the surprising reduction in MSE as the forecast horizon moves from 5 quarters to 6 quarters. Without the 1974 outliers, MSEs increase approximately proportionally to the forecast horizon (as one might expect), up to a horizon of seven quarters.

The 1974 errors account for a substantial fraction of the reduction in output uncertainty. This can be seen in table 3. Column 2, reproduced from table 1, shows MSE ratios using all observations. Column 3 shows MSE ratios when the errors for 1974q3-1975q2 are excluded from the pre-1984 sample. This exclusion lowers the pre-1984 MSEs by up to 40 percent; nevertheless, they remain higher than post-1984 MSEs. So the reduction in uncertainty does not just reflect the accident that the “perfect storm” of 1974 happened to fall before the breakpoint.

Table 3: MSE ratios excluding the outliers of 1974-75

Horizon	Ratio of MSE before and after 1984q1	
	All observations	Excluding 1974q3:1975q2
1 (current) quarter	.39	.40
2 quarters	.37	.44
3 quarters	.35	.46
4 quarters	.37	.61
5 quarters	.40	.64
6 quarters	.64	.84
7 quarters	1.04	1.04
8 quarters	3.19	3.19

IX. Conclusion

The Greenbook short-horizon forecast errors are smaller after 1984. In that sense, uncertainty has diminished. However, this reduction was significantly smaller than the reduction in output volatility. Moreover, longer-horizon errors, which are more important for some purposes, do not seem to have narrowed.

This paper has attempted to document how the unpredictability of output has changed over recent decades. It has not directly addressed the questions of why the Fed staff made the errors it did, how decision-makers should react to that unpredictability, nor how that unpredictability might be reduced. These are topics for future research. Hopefully, the results in this paper may contribute to their analysis.

Appendix: Non-overlapping errors

A simple method of removing serial correlation from forecast errors is to construct a sample of non-overlapping errors – or, more precisely, forecasts of non-overlapping events. This sample comprises all current-quarter forecasts, every second 2-quarter forecast, every third three-quarter forecast, and so on. This approach also has the advantage that it provides an equal weighting to periods with frequent repetitive forecasts relative to periods when forecasts were less frequent.

Table 4 shows forecast MSEs before and after 1984q1 using this sample. These are broadly similar to those shown in table 1. The main difference is that six-quarter and seven-quarter MSEs are noticeably larger after 1984 than before. This partly reflects the re-weighting, but mainly seems attributable to sampling variability. Assuming that the non-overlapping errors are independent, the ratio of these MSEs has an exact *F* distribution, with degrees of freedom parameters equal to the number of errors in each sub-sample.

P-values from *F*-tests are shown in the final column. Note that the *p*-value for current quarter forecasts, which use all observations and assume no serial correlation, is about the same as that in table 1 estimated by Monte Carlo, as would be expected. Other *p*-values are much higher than those in table 1. These tests do not clearly reject the hypothesis that the variance of errors is stable. However, because these tests involve disregarding a great deal of relevant information, it is arguable that this failure to reject is uninformative.

Table 4: Forecast Mean Squared Errors (MSEs)

Before and after 1984

Non-Overlapping Forecast Errors

Horizon	Earliest to 1984		1984 to 2001 (a)		Ratio	*p*-value
	MSE (b)	Observations	MSE (b)	Observations		(c)
1 (current) quarter	0.4	74	0.2	63	.39	0.008%
2 quarters	1.3	36	0.7	32	.55	4.5%
3 quarters	2.1	21	1.4	22	.64	15.5%
4 quarters	4.0	16	1.8	16	.45	6.2%
5 quarters	4.2	10	2.8	14	.66	23.6%
6 quarters	3.1	7	3.7	11	1.18	57.0%
7 quarters	3.8	2	5.2	10	1.36	50.3%
8 quarters	2.9	2	7.0	9	2.42	67.3%

(a) Errors are dated by the time of the event, not the time of the forecast. The last forecast used was made in November 1999, for the period 1999 to 2001.

(b) Mean Squared Error, in percentage points

(c) probability of a ratio smaller than that observed, under the null hypothesis that variances are equal. Specifically, the tail of an F-distribution to the left of the given ratio, with degrees of freedom given by the number of non-overlapping observations shown.

References

Atkeson, Andrew and Lee E. Ohanian, (2001), "Are Phillips Curves Useful for Forecasting Inflation?" *Federal Reserve Bank of Minneapolis Quarterly Review,* Vol 25, No 1, pp. 2-11.

Beaulieu; J. Joseph and Jeffrey A. Miron (1992) "A Cross Country Comparison of Seasonal Cycles and Business Cycles" *The Economic Journal* Vol. 102, No. 413 (Jul., 1992), pp. 772-788

Bernanke, Ben S. (2004a) "*The Great Moderation",* Speech to the Eastern Economic Association, Washington, DC, February 20. http://www.federalreserve.gov/boarddocs/speeches/2004/20040220/default.htm

Bernanke, Ben S. (2004b) "*The Logic of Monetary Policy",* Speech to the National Economists Club, Washington, DC, December 2. http://www.federalreserve.gov/boarddocs/speeches/2004/20041202/default.htm

Blanchard, Olivier, and John Simon (2001). "The Long and Large Decline in U.S. Output Volatility," *Brookings Papers on Economic Activity*, 1, pp. 135-64.

Campbell, Sean D. (2004) "Macroeconomic Volatility, Predictability and Uncertainty in the Great Moderation: Evidence from the Survey of Professional Forecasters" *Finance and Economics Discussion Series* 2004-52 Federal Reserve Board, Washington DC

Cecchetti, Stephen G., Alfonso Flores-Lagunes and Stefan Krause, (2005) "Assessing the Sources of Changes in the Volatility of Real Growth," prepared for Reserve Bank of Australia conference on "The Changing Nature of the Business Cycle," Sydney Australia, July, 2005. http://www.rba.gov.au/PublicationsAndResearch/Conferences/2005/index.html

Goodhart, C.A.E., (2004) "The Interaction between the Bank of England's Forecasts and Policy, and the Outturn," presented at Federal Reserve Board of Governors, May, 2004

Greenspan, Alan (2004) "Risk and Uncertainty in Monetary Policy," Speech to the American Economic Association, San Diego, California, January 3, 2004. http://www.federalreserve.gov/boarddocs/speeches/2004/20040103/default.htm

Kim, Chang-Jin, and Charles Nelson (1999). "Has the U.S. Economy Become More Stable? A Bayesian Approach Based on a Markov-Switching Model of the Business Cycle," *Review of Economics and Statistics*, 81, pp. 608-16.

McConnell, Margaret, and Gabriel Perez-Quiros (2000). "Output Fluctuations in the United States: What Has Changed since the Early 1980s?" *American Economic Review*, 90, pp. 1464-76.

Robertson John C. and Ellis W. Tallman, (1998) "Data Vintages and Measuring Forecast Model Performance" *Economic Review,* Federal Reserve Bank of Atlanta, Fourth Quarter, Volume 83, Number 4

Romer, Christina D. and David H. Romer (2000) “Federal Reserve Information and the Behavior of Interest Rates”, *American Economic Review*, Vol.90, No. 3, 429-457

Schuh, Scott (2001) “An Evaluation of Recent Macroeconomic Forecast Errors”, *New England Economic Review,* Federal Reserve Bank of Boston, Jan.-Feb., pp. 35-56

Sims, Christopher (2002) “The Role of Models and Probabilities in the Monetary Policy Process” *Brookings Papers on Economic Activity*, 2, 1-62

Stock, James H., and Mark W. Watson (2003). "Has the Business Cycle Changed? Evidence and Explanations," in *Monetary Policy and Uncertainty: Adapting to a Changing Economy,* Federal Reserve Bank of Kansas City symposium, Jackson Hole, Wyoming.

Tetlow, Robert J. and Brian Ironside (2005) "Real-Time Model Uncertainty in the United States: the Fed from 1996-2003," unpublished manuscript, http://www.roberttetlow.com