Journal of Economic Literature
Vol. XLI (September 2003) pp. 788–829

Forecasting Output and Inflation: The Role of Asset Prices

JAMES H. STOCK *and* Mark W. Watson[1]

1. *Introduction*

Because asset prices are forward-looking, they constitute a class of potentially useful predictors of inflation and output growth. The premise that interest rates and asset prices contain useful information about future economic developments embodies foundational concepts of macroeconomics: Irving Fisher's theory that the nominal interest rate is the real rate plus expected inflation; the notion that a monetary contraction produces temporarily high interest rates—an inverted yield curve—and leads to an economic slowdown; and the hypothesis that stock prices reflect the expected present discounted value of future earnings. Indeed, Wesley Mitchell and Arthur Burns (1938) included the Dow Jones composite index of stock prices in their initial list of leading indicators of expansions and contractions in the U.S. economy.

The past fifteen years have seen considerable research on forecasting economic activity and inflation using asset prices, where we interpret asset prices broadly as including interest rates, differences between interest rates (spreads), returns, and other measures related to the value of financial or tangible assets (bonds, stocks, housing, gold, etc.). This research on asset prices as leading indicators arose, at least in part, from the instability in the 1970s and early 1980s of forecasts of output and inflation based on monetary aggregates and of forecasts of inflation based on the (non-expectational) Phillips curve. One problem with using monetary aggregates for forecasting is that they require ongoing redefinition as new financial instruments are introduced. In contrast, asset prices and returns typically are observed in real time with negligible measurement error.

The now-large literature on forecasting using asset prices has identified a number of asset prices as leading indicators of either economic activity or inflation; these include interest rates, term spreads, stock returns, dividend yields, and exchange rates. This literature is of interest from several perspectives. First and most obviously, those whose daily task it is to produce forecasts—notably, economists at central banks, and business

[1] Stock: Department of Economics, Harvard University, and NBER. Watson: Woodrow Wilson School and Department of Economics, Princeton University, and NBER. We thank Charles Goodhart and Boris Hofmann for sharing their housing price data and John Campbell for sharing his dividend yield data. Helpful comments were provided by John Campbell, Fabio Canova, Steve Cecchetti, Frank Diebold, Stefan Gerlach, Charles Goodhart, Clive Granger, Mike McCracken, Marianne Nessen, Tony Rodrigues, Chris Sims, two referees, and participants at the June 2000 Sveriges Riksbank Conference on Asset Markets and Monetary Policy. We thank Jean-Philippe Laforte for research assistance. This research was funded in part by NSF grants SBR-9730489 and SBR-0214131.

economists—need to know which, if any, asset prices provide reliable and potent forecasts of output growth and inflation. Second, knowledge of which asset prices are useful for forecasting, and which are not, constitutes a set of stylized facts to guide those macroeconomists mainly interested in understanding the workings of modern economies. Third, the empirical failure of the 1960s-vintage Phillips curve was one of the crucial developments that led to rational expectations macroeconomics, and understanding if and how forecasts based on asset prices break down could lead to further changes or refinements in macroeconomic models.

This article begins in section 2 with a summary of the econometric methods used in this literature to evaluate predictive content. We then review the large literature on asset prices as predictors of real economic activity and inflation. This review, contained in section 3, covers 93 articles and working papers and emphasizes developments during the past fifteen years. We focus exclusively on forecasts of output and inflation; forecasts of volatility, which are used mainly in finance, have been reviewed recently in Ser-Huang Poon and Clive Granger (2003). Next, we undertake our own empirical assessment of the practical value of asset prices for short- to medium-term economic forecasting; the methods, data, and results are presented in sections 4–7. This analysis uses quarterly data on as many as 43 variables from each of seven developed economies (Canada, France, Germany, Italy, Japan, the United Kingdom, and the United States) over 1959–99 (some series are available only for a shorter period). Most of these predictors are asset prices, but for comparison purposes we also consider selected measures of real economic activity, wages, prices, and the money supply.

Our analysis of the literature and the data leads to four main conclusions.

First, some asset prices have substantial and statistically significant marginal predictive content for output growth at some times in some countries. Whether this predictive content can be exploited reliably is less clear, for this requires knowing *a priori* what asset price works when in which country. The evidence that asset prices are useful for forecasting output growth is stronger than for inflation.

Second, forecasts based on individual indicators are unstable. Finding an indicator that predicts well in one period is no guarantee that it will predict well in later periods. It appears that instability of predictive relations based on asset prices (like many other candidate leading indicators) is the norm.

Third, although the most common econometric method of identifying a potentially useful predictor is to rely on in-sample significance tests such as Granger causality tests, doing so provides little assurance that the identified predictive relation is stable. Indeed, the empirical results indicate that a significant Granger causality statistic contains little or no information about whether the indicator has been a reliable (potent and stable) predictor.

Fourth, simple methods for combining the information in the various predictors, such as computing the median of a panel of forecasts based on individual asset prices, seem to circumvent the worst of these instability problems.

Some of these conclusions could be interpreted negatively by those (ourselves included) who have worked in this area. But in this review we argue instead that they reflect limitations of conventional models and econometric procedures, not a fundamental absence of predictive relations in the economy: the challenge is to develop methods better geared to the intermittent and evolving nature of these predictive relations. We expand on these ideas in section 8.

2. *Methods for Evaluating Forecasts and Predictive Content*

Econometric methods for measuring predictive content can be divided into two

groups: in-sample and out-of-sample methods.

2.1 *In-Sample Measures of Predictive Content*

Suppose we want to know whether a candidate variable, X, is useful for forecasting a variable of interest, Y. For example, X_t could be the value of the term spread in quarter t and Y_{t+1} could be the growth rate of real GDP in the next quarter, so $Y_{t+1} = 400\ln(\text{GDP}_{t+1}/\text{GDP}_t) = 400\Delta\ln(\text{GDP}_{t+1})$ (the factor of 400 standardizes the units to annual percentage growth rates). A simple framework for assessing predictive content is the linear regression model relating the future value of Y to the current value of X:

$$Y_{t+1} = \beta_0 + \beta_1 X_t + u_{t+1}, \qquad (1)$$

where β_0 and β_1 are unknown parameters and u_{t+1} is an error term. If $\beta_1 \neq 0$, then today's value of X can be used to forecast the value of Y in the next period. The null hypothesis that X_t has no predictive content can be tested by computing the t-statistic on β_1. The economic significance of X_t as a predictor can be assessed using the regression R^2 and the standard error of the regression (*SER*), the estimate of the standard deviation of u_{t+1}. Because the error term can be heteroskedastic (that is, the variance of u_{t+1} can depend on X_t) and/or autocorrelated (u_{t+1} can be correlated with its previous values), the t-statistic should be computed using heteroskedasticity- and autocorrelation-consistent (HAC) standard errors.

This simple framework has an important limitation: if Y_{t+1} is serially correlated, as is typically the case for time series variables, its own past values are themselves useful predictors. Thus a more discerning question than that studied using (1) is whether X_t has predictive content for Y_{t+1}, above and beyond that contained in its past values. Moreover, additional lagged values of X_t also might be useful predictors. This leads to the extension of (1), in which multiple lagged values of X_t and Y_t appear. This multiple regression model is conventionally expressed using lag polynomials. Let $\beta_1(\text{L})$ and $\beta_2(\text{L})$ denote lag polynomials, so that $\beta_1(\text{L})X_t = \beta_{11}X_t + \beta_{12}X_{t-1} + \ldots + \beta_{1p}X_{t-p+1}$, where p is the number of lagged values of X included (we refer to X_t as a lagged value because it is lagged relative to the variable to be forecasted, which is dated $t+1$ in (1)). Then the extended regression model is the autoregressive distributed lag (ADL) model,

$$Y_{t+1} = \beta_0 + \beta_1(\text{L})X_t + \beta_2(\text{L})Y_t + u_{t+1}, \qquad (2)$$

In the context of the ADL model (2), the hypothesis that X_t has no predictive content for Y_{t+1}, above and beyond that in lags of Y, corresponds to the hypothesis that $\beta_1(\text{L}) = 0$, that is, that each of the lag polynomial coefficients equals zero. This hypothesis can be tested using the (heteroskedasticity-robust) F-statistic. This F-statistic is commonly called the Granger causality test statistic. The economic value of the additional forecasting content of X_t can be assessed by computing the partial R^2 of the regression or by computing the ratio of the *SER* (or its square) of the regression (2) to that of a univariate autoregression (AR), which is (2) in which X_t and its lags are excluded.

Equation (2) applies to forecasts one period ahead, but it is readily modified for multistep-ahead forecasts by replacing Y_{t+1} with the suitable h-period ahead value. For example, if the variable being forecast is the percentage growth of real GDP over the next eight quarters, then the dependent variable in (2) becomes $Y^8_{t+8} = 50\ln(\text{GDP}_{t+8}/\text{GDP}_t)$, where the factor of 50 standardizes the units to be annual percentage growth rates. In general, the h-step ahead forecasting regression can be written

$$Y^h_{t+h} = \beta_0 + \beta_1(\text{L})X_t + \beta_2(\text{L})Y_t + u^h_{t+h}. \qquad (3)$$

Because the data are overlapping, the error term u^h_{t+h} in (3) is serially correlated, so the test of predictive content based on (3)

(the test of $\beta_1(L) = 0$) should be computed using HAC standard errors.[2]

The stability of the coefficients in the forecasting relation (3) can be assessed by a variety of methods, including testing for breaks in coefficients and estimation of models with time-varying parameters. These methods are used infrequently in this literature, so we do not discuss them here. We return to in-sample tests for parameter stability when we describe our empirical methods in section 4.

The tools discussed so far examine how useful X would have been for predicting Y, had you been able to use the regression coefficients estimated by the full-sample regression. If coefficients change over time, this full-sample analysis can be misleading for out-of-sample forecasting. Therefore, evaluations of predictive content also should rely on statistics that are designed to simulate more closely actual real-time forecasting, which we refer to generally as pseudo out-of-sample forecast evaluation.

2.2 *Pseudo Out-of-Sample Measures of Predictive Content*

Pseudo out-of-sample measures of predictive content entail simulating real-time forecasting. Suppose the researcher has quarterly data; to make the pseudo-forecast for 1990:I she estimates the model using data available through 1989:IV, then uses this estimated model to produce the 1990:I forecast, just as she would were it truly 1989:IV. This is repeated throughout the sample, moving ahead one quarter at a time, thereby producing a sequence of pseudo out-of-sample forecasts. The model estimation stage can be complex, possibly entailing the estimation of a large number of models and selecting among them based on some criterion; for example, the lag length of an autoregression could be selected using an information criterion such as the Akaike or Bayes information criteria (AIC or BIC). Critically, however, all model selection and estimation must be done using data available prior to making the forecast—in the example, using only the data available through 1989:IV. Pseudo out-of-sample measures of forecast accuracy have several desirable characteristics, most notably from the perspective of this survey being their ability to detect changes in parameters towards the end of the sample.[3]

A common way to quantify pseudo out-of-sample forecast performance is to compute the mean squared forecast error of a candidate forecast (forecast i), relative to a benchmark (forecast 0). For example, the candidate forecast could be based on an asset price and the benchmark could come from a univariate autoregression. Let $\hat{Y}^h_{0,t+h|t}$ and $\hat{Y}^h_{i,t+h|t}$ be the benchmark and i^{th} candidate pseudo out-of-sample forecasts of Y^h_{t+h}, made using data through time t. Then, the h-step ahead mean squared forecast error (MSFE) of forecast i, relative to that of the benchmark forecast, is:

[2] An alternative to the "h-step ahead projection" approach in (3) is to estimate a vector autoregression (VAR) or some other joint one-step ahead model for X_t and Y_t, then iterate this model forward for h periods. Almost all the papers in the asset price-as-predictor literature use the h-step ahead projection method. If the VAR is correctly specified, then the VAR iterated forecasts are more efficient asymptotically, but the h-step ahead projection forecast reduces the potential impact of specification error in the one-step ahead model.

[3] Many macroeconomic time series are subject to data revisions. An additional step towards forecasting reality is to construct pseudo out-of-sample forecasts using real-time data–in the example, the vintage of the data currently available as of 1989:IV. Implementing this in practice requires using large data sets with many vintages. Such data are now available (Dean Croushore and Tom Stark 2003). The data revision issue is, however, less important in the literature of concern in this review than elsewhere: a virtue of asset price data is that they are measured with negligible error in real time and are not revised, nor is the CPI revised. One question is whether these asset prices should predict early or late ("final") vintages of GDP. The implicit view in this literature is that the best way to evaluate a true predictive relation is to use the best (final) estimate of GDP, and we adopt that view here.

$$\frac{\frac{1}{T_2 - T_1 - h + 1}\sum_{t=T_1}^{T_2-h}\left(Y_{t+h}^h - \hat{Y}_{i,t+h|t}^h\right)^2}{\frac{1}{T_2 - T_1 - h + 1}\sum_{t=T_1}^{T_2-h}\left(Y_{t+h}^h - \hat{Y}_{0,t+h|t}^h\right)^2}, \quad (4)$$

where T_1 and $T_2 - h$ are respectively the first and last dates over which the pseudo out-of-sample forecast is computed (so that forecasts are made for dates $t = T_1 + h, \ldots, T_2$).

If its relative MSFE is less than one, the candidate forecast is estimated to have performed better than the benchmark. Of course, this could happen simply because of sampling variability, so to determine whether a relative MSFE less than one is statistically significant requires testing the hypothesis that the population relative MSFE = 1, against the alternative that it is less than one. When neither model has any estimated parameters, this is done using the methods of Francis Diebold and Robert Mariano (1995). Kenneth West (1996) treats the case that at least one model has estimated parameters and the models are not nested (that is, the benchmark model is not a special case of model *i*). If, as is the case in much of the literature we review, the benchmark is nested within model *i*, then the methods developed by Michael McCracken (1999) and Todd Clark and McCracken (2001) apply. These econometric methods have been developed only recently so are almost entirely absent from the literature reviewed in section 3.

In the empirical analysis in sections 6 and 7, we use the relative mean squared forecast error criterion in (4) because of its familiarity and ease of interpretation. Many variations on this method are available, however. An alternative to the recursive estimation scheme outlined above is to use a fixed number of observations (a rolling window) for estimating the forecasting model. Also, squared error loss can be replaced with a different loss function; for example, one could compute relative mean absolute error. Other statistics, such as statistics that test for forecast encompassing or ones that assess the accuracy of the forecasted direction of change, can be used in addition to a relative error measure (see M. Hashem Pesaran and Spyros Skouras 2002, and McCracken and West 2002 for a discussion of alternative forecast evaluation statistics). For a textbook introduction and worked empirical examples of pseudo out-of-sample forecast comparison, see Stock and Watson (2003a, section 12.6). For an introduction to the recent literature on forecast comparisons, see McCracken and West (2002).

3. *Literature Survey*

This survey first reviews papers that use asset prices as predictors of inflation and/or output growth, then provides a brief, selective summary of recent developments using nonfinancial indicators. Although we mention some historical precedents, our review focuses on developments within the past fifteen years. The section concludes with an attempt to draw some general conclusions from this literature.

3.1 *Forecasts Using Asset Prices*

Interest rates. Short-term interest rates have a long history of use as predictors of output and inflation. Notably, using data for the United States, Christopher Sims (1980) found that including the commercial paper rate in vector autoregressions (VARs) with output, inflation, and money eliminated the marginal predictive content of money for real output. This result has been confirmed in numerous studies, e.g. Ben Bernanke and Alan Blinder (1992) for the United States, who suggested that the federal funds rate is the appropriate short-run measure of monetary policy, rather than the growth of monetary aggregates. Most of the research involving interest rate spreads, however, has found that the level (or change) of a short rate has little marginal predictive content for output once spreads are included.

The term spread and output growth. The term spread is the difference between interest rates on long and short maturity debt, usually government debt. The literature on term spreads uses different measures of this spread, the most common being a long government bond rate minus a three-month government bill rate or, instead, the long bond rate minus an overnight rate (in the United States, the federal funds rate).

The adage that an inverted yield curve signals a recession was formalized empirically, apparently independently, by a number of researchers in the late 1980s, including Robert Laurent (1988, 1989), Campbell Harvey (1988, 1989), Stock and Watson (1989), Nai-Fu Chen (1991), and Arturo Estrella and Gikas Hardouvelis (1991). These studies mainly focused on using the term spread to predict output growth (or in the case of Harvey 1988, consumption growth) using U.S. data. Of these studies, Estrella and Hardouvelis (1991) provided the most comprehensive documentation of the strong (in-sample) predictive content of the spread for output, including its ability to predict a binary recession indicator in probit regressions. This early work focused on bivariate relations, with the exception of Stock and Watson (1989), who used in-sample statistics for bivariate and multivariate regressions to identify the term spread and a default spread (the paper-bill spread, discussed below) as two historically potent leading indicators for output. The work of Eugene Fama (1990) and Frederic Mishkin (1990a,b) is also notable, for they found that the term spread has (in-sample, bivariate) predictive content for real rates, especially at shorter horizons.

Subsequent work focused on developing economic explanations for this relation, determining whether it is stable over time within the United States, and ascertaining whether it holds up in international evidence. The standard economic explanation for why the term spread has predictive content for output is that the spread is an indicator of an effective monetary policy: monetary tightening results in short-term interest rates that are high, relative to long-term interest rates, and these high short rates in turn produce an economic slowdown (Bernanke and Blinder 1992). Notably, when placed within a multivariate model, the predictive content of the term spread can change if monetary policy changes or the composition of economic shocks changes (Frank Smets and Kostas Tsatsaronis 1997). Movements in expected future interest rates might not account for all the predictive power of the term spread, however: James Hamilton and Dong Heon Kim (2002) suggested that the term premium (the term spread minus its predicted component under the expectations hypothesis of the term structure of interest rates) has important predictive content for output as well.

A closer examination of the U.S. evidence has led to the conclusion that the predictive content of the term spread for economic activity has diminished since 1985, a point made using both pseudo out-of-sample and rolling in-sample statistics by Joseph Haubrich and Ann Dombrosky (1996) and by Michael Dotsey (1998). Similarly, Andrew Ang, Monika Piazzesi, and Min Wei (2003) found that, during the 1990s, U.S. GDP growth is better predicted by the short rate than the term spread. In contrast, research that focuses on predicting binary recession events (instead of output growth itself) suggests that the term spread might have had some link to the 1990 recession. In particular, the *ex post* analyses of Estrella and Mishkin (1998), Kajal Lahiri and Jiazhou Wang (1996), and Michael Dueker (1997) respectively provided probit and Markov switching models that produce in-sample recession probabilities consistent with the term spread providing advance warning of the 1990 U.S. recession. These estimated probabilities, however, were based on estimated parameters that include this recession, so these are not real time or pseudo out-of-sample recession probabilities.

The real-time evidence about the value of the spread as an indicator in the 1990 recession is more mixed. Laurent (1989), using the

term spread, predicted an imminent recession in the United States; Harvey (1989) published a forecast based on the yield curve that suggested "a slowing of economic growth, but not zero or negative growth" from the third quarter of 1989 through the third quarter of 1990; and the Stock-Watson (1989) experimental recession index increased sharply when the yield curve flattened in late 1988 and early 1989. However, the business cycle peak of July 1990 considerably postdates the predicted period of these slowdowns: as Laurent (1989) wrote, "recent spread data suggest that the slowdown is likely to extend through the rest of 1989 and be quite significant." Moreover, Laurent's (1989) forecast was based in part on a judgmental interpretation that the then-current inversion of the yield curve had special (nonlinear) significance, signaling a downturn more severe than would be suggested by a linear model. Indeed, even the largest predicted recession probabilities from the in-sample models are modest: 25 percent in Estrella and Mishkin's (1998) probit model and 20 percent in Dueker's (1997) Markov switching model, for example. Harvey (1993) interpreted this evidence more favorably, arguing that because the yield curve inverted moderately beginning in 1989:II it correctly predicted a moderate recession six quarters later. Our interpretation of this episode is that the term spread is an indicator of monetary policy, that monetary policy was tight during late 1988, and that yield-curve based models correctly predicted a slowdown in 1989. This slowdown was not a recession, however, and the proximate cause of the recession of 1990 was not monetary tightening but rather special nonmonetary circumstances, in particular the invasion of Kuwait by Iraq and the subsequent response by U.S. consumers (Olivier Blanchard 1993). This interpretation is broadly similar to Benjamin Friedman and Kenneth Kuttner's (1998) explanation of the failure of the paper-bill spread to predict the 1990 recession (discussed below).

Stock and Watson (2003b) examine the behavior of various leading indicators before and during the U.S. recession that began in March 2001. The term spread did turn negative in advance of this recession: the Fed funds rate exceeded the long bond rate from June 2000 through March 2001. This inversion, however, was small by historical standards. While regressions of the form (3) predicted a slower rate of economic growth in early 2001, the predicted slowdown was modest: four quarter growth forecasts based on (3) fell 1.4 percentage points, from 3.3 percent in 2000:I to a minimum of 1.9 percent in 2000:IV, still far from the negative growth of a recession.

One way to get additional evidence on the reliability of the term spread as a predictor of output growth is to examine evidence for other countries. Harvey (1991), Zuliu Hu (1993), E. Philip Davis and S. G. B. Henry (1994), Charles Plosser and K. Geert Rouwenhorst (1994), Catherine Bonser-Neal and Timothy Morley (1997), Sharon Kozicki (1997), John Campbell (1999), Estrella and Mishkin (1997), Estrella, Anthony Rodrigues, and Sebastian Schich (2003), and Joseph Atta-Mensah and Greg Tkacz (2001) generally conclude that the term spread has predictive content for real output growth in major non-U.S. OECD economies. Estrella, Rodrigues, and Schich (2003) use in-sample break tests to assess coefficient stability of the forecasting relations and typically fail to reject the null hypothesis of stability in the cases in which the term spread has the greatest estimated predictive content (mainly long-horizon regressions). Additionally, Henri Bernard and Stefan Gerlach (1998) and Estrella, Rodrigues, and Schich (2003) provide cross-country evidence on term spreads as predictors of a binary recession indicator for seven non-U.S. OECD countries. Unlike most of these papers, Plosser and Rouwenhorst (1994) considered multiple regressions that include the level and change of interest rates and concluded that, given the spread, the short rate has little predictive content for output in almost all the economies

they consider. These studies typically used in-sample statistics and data sets that start in 1970 or later. Three exceptions to this generally sanguine view are Davis and Gabriel Fagan (1997), Smets and Tsatsaronis (1997), and Fabio Canova and Gianni De Nicolo (2000). Using a pseudo out-of-sample forecasting design, Davis and Fagan (1997) find evidence of subsample instability and report disappointing pseudo out-of-sample forecasting performance across nine EU economies. Smets and Tsatsaronis (1997) find instability in the yield curve–output relation in the 1990s in the United States and Germany. Canova and De Nicolo (2000), using in-sample VAR statistics, find only a limited forecasting role for innovations to the term premium in Germany, Japan, and the United Kingdom.

Term spreads and inflation. Many studies, including some of those already cited, also consider the predictive content of the term spread for inflation. According to the expectations hypothesis of the term structure of interest rates, the forward rate (and the term spread) should embody market expectations of future inflation and the future real rate. With some notable exceptions, the papers in this literature generally find that there is little or no marginal information content in the nominal interest rate term structure for future inflation. Much of the early work, which typically claims to find predictive content, did not control for lagged inflation. In U.S. data, Mishkin (1990a) found no predictive content of term spreads for inflation at the short end of the yield curve, although Mishkin (1990b) found predictive content using spreads that involve long bond rates. Philippe Jorion and Mishkin (1991) and Mishkin (1991) reached similar conclusions using data on ten OECD countries, results confirmed by Gerlach (1997) for Germany using Mishkin's methodology. Drawing on Jeffrey Frankel's (1982) early work in this area, Frankel and Cara Lown (1994) suggested a modification of the term spread based on a weighted average of different maturities that outperformed the simple term spread in Mishkin-style regressions.

Mishkin's regressions have a single stochastic regressor, the term spread (no lags), and in particular do not include lagged inflation. Inflation is highly persistent, however, and Bernanke and Mishkin (1992), Estrella and Mishkin (1997), and Kozicki (1997) examined the in-sample marginal predictive content of the term spread, given lagged inflation. Bernanke and Mishkin (1992) found little or no marginal predictive content of the term spread for one-month-ahead inflation in a data set with six large economies, once lags of inflation are included. Kozicki (1997) and Estrella and Mishkin (1997) included only a single lag of inflation, but even so they found that doing so substantially reduced the marginal predictive content of the term spread for future inflation over one to two years. For example, once lagged inflation is added, Kozicki (1997) found that the spread remained significant for one-year inflation in only two of the ten OECD countries she studied; in Estrella and Mishkin's (1997) study, the term spread was no longer a significant predictor at the one-year horizon in any of their four countries, although they provided evidence for predictive content at longer horizons.

Default spreads. Another strand of research has focused on the predictive content of default spreads, primarily for real economic activity. A default spread is the difference between the interest rates on matched maturity private debt with different degrees of default risk. Different authors measure this spread differently, and these differences are potentially important. Because markets for private debt differ substantially across countries and are most developed for the United States, this work has focused on the United States.

In his study of the credit channel during the Great Depression, Bernanke (1983) showed that, during the interwar period the Baa-Treasury bond spread was a useful predictor of industrial production growth. Stock

and Watson (1989) and Friedman and Kuttner (1992) studied the default spread as a predictor of real growth in the postwar period; they found that the spread between commercial paper and U.S. Treasury bills of the same maturity (three or six months; the "paper-bill" spread) was a potent predictor of output growth (monthly data, 1959–88 for Stock and Watson 1989, quarterly data, 1960–90 for Friedman and Kuttner 1992). Using in-sample statistics, Friedman and Kuttner (1992) concluded that, upon controlling for the paper-bill spread, monetary aggregates and interest rates have little predictive content for real output, a finding confirmed by Bernanke and Blinder (1992) and Martin Feldstein and Stock (1994).

Subsequent literature focused on whether this predictive relationship is stable over time. Bernanke (1990) used in-sample statistics to confirm the strong performance of paper-bill spread as predictor of output, but by splitting up the sample he also suggested that this relation weakened during the 1980s. This view was affirmed and asserted more strongly by Mark Thoma and Jo Anna Gray (1994), Rik Hafer and Ali Kutan (1992), and Kenneth Emery (1996). Thoma and Gray (1994), for example, found that the paper-bill spread has strong in-sample explanatory power in recursive or rolling regressions, but little predictive power in pseudo out-of-sample forecasting exercises over the 1980s. Emery (1996) finds little in-sample explanatory power of the paper-bill spread in samples that postdate 1980. These authors interpreted this as a consequence of special events, especially in 1973–74, which contribute to a good in-sample fit but not necessarily good forecasting performance. Drawing on institutional considerations, John Duca (1999) also took this view: Duca's (1999) concerns echo Timothy Cook's (1981) warnings about how the changing institutional environment and financial innovations could substantially change markets for short-term debt and thereby alter the relationship between default spreads and real activity.

One obvious true out-of-sample predictive failure of the paper-bill spread is its failure to rise sharply in advance of the 1990–91 U.S. recession. In their post-mortem, Friedman and Kuttner (1998) suggested that this predictive failure arose because the 1990–91 recession was caused in large part by nonmonetary events that would not have been detected by the paper-bill spread. They further argued that there were changes in the commercial paper market unrelated to the recession that also led to this predictive failure. Similarly, the paper-bill spread failed to forecast the 2001 recession: the paper-bill spread had brief moderate spikes in October 1998, October 1999, and June 2000, but the spread was small and declining from August 2000 through the end of 2001 (Stock and Watson 2003b).

We know of little work examining the predictive content of default spreads in economies other than the United States. Bernanke and Mishkin (1992) report a preliminary investigation, but they questioned the adequacy of their private debt interest rate data (the counterpart of the commercial paper rate in the United States) for several countries. Finding sufficiently long time series data on reliable market prices of suitable private debt instruments has been a barrier to international comparisons on the role of the default spread.

Some studies examined the predictive content of the default spread for inflation. Friedman and Kuttner (1992) found little predictive content of the paper-bill spread for inflation using Granger causality tests. Consistent with this, Feldstein and Stock (1994) found that although the paper-bill spread was a significant in-sample predictor of real GDP, it did not significantly enter equations predicting nominal GDP.

Four nonexclusive arguments have been put forth on why the paper–bill spread had predictive content for output growth during the 1960s and 1970s. Stock and Watson (1989) suggested the predictive content arises from expectations of default risk,

which are in turn based on private expectations of sales and profits. Bernanke (1990) and Bernanke and Blinder (1992) argued instead that the paper-bill spread is a sensitive measure of monetary policy, and this is the main source of its predictive content. Friedman and Kuttner (1993a,b) suggested that the spread is detecting influences of supply and demand (i.e. liquidity) in the market for private debt; this emphasis is similar to Cook's (1981) attribution of movements in such spreads to supply and demand considerations. Finally, Thoma and Gray (1994) and Emery (1996) argued that the predictive content is largely coincidental, the consequence of one-time events.

There has been some examination of other spreads in this literature. Mark Gertler and Lown (2000) take the view that, because of the credit channel theory of monetary policy transmission, the premise of using a default spread to predict future output is sound, but that the paper-bill spread is a flawed choice for institutional reasons. Instead, they suggest using the high-yield bond ("junk bond")–Aaa spread instead. The junk bond market was only developed in the 1980s in the United States, so this spread has a short time series. Still, Gertler and Lown (2000) present in-sample evidence that its explanatory power was strong throughout this period. This is notable because the paper-bill spread (and, as was noted above, the term spread) have substantially reduced or no predictive content for output growth in the United States during this period. However, Duca's (1999) concerns about default spreads in general extend to the junk bond-Aaa spread as well: he suggests the spike in the junk bond spread in the late 1980s and early 1990s (which is key to this spread's signal of the 1990 recession) was a coincidental consequence of the aftermath of the thrift crisis, in which thrifts were forced to sell their junk bond holdings in an illiquid market.

Stock prices and dividend yields. If the price of a stock equals the expected discounted value of future earnings, then stock returns should be useful in forecasting earnings growth or, more broadly, output growth. The empirical link between stock prices and economic activity has been noted at least since Mitchell and Burns (1938); see Stanley Fischer and Robert Merton (1984) and Robert Barro (1990). Upon closer inspection, however, this link is murky. Stock returns generally do not have substantial in-sample predictive content for future output, even in bivariate regressions with no lagged dependent variables (Fama 1981 and Harvey 1989), and any predictive content is reduced by including lagged output growth. This minimal marginal predictive content is found both in linear regressions predicting output growth (Stock and Watson 1989, 1999a) and in probit regressions of binary recession events (Estrella and Mishkin 1998).

In his review article, Campbell (1999) shows that in a simple loglinear representative agent model, the log price-dividend ratio embodies rational discounted forecasts of dividend growth rates and stock returns, making it an appropriate state variable to use for forecasting. But in his international dataset (fifteen countries, sample periods mainly 1970s to 1996, Campbell (1999) found that the log dividend price ratio has little predictive content for output. This is consistent with the generally negative conclusions in the larger literature that examines the predictive content of stock returns directly. These generally negative findings provide a precise reprise of the witticism that the stock market has predicted nine of the last four recessions.

Campbell et al. (2001) proposed an interesting variant in which the variance of stock returns, rather than the returns themselves, could have predictive content for output growth. Using in-sample statistics, they found evidence that high volatility in one quarter signals low growth in the next quarter, as it might if high volatility was associated with increased doubts about short-term economic prospects. When Hui Guo (2002) used out-of-sample statistics, however, the

evidence for predictive content was substantially weaker. These findings are consistent with the predictive content of stock market volatility being stronger during some episodes than during others.

Few studies have examined the predictive content of stock prices for inflation. One is Charles Goodhart and Boris Hofmann (2000a), who found that stock returns do not have marginal predictive content for inflation in their international data set (twelve developed economies, quarterly data, mainly 1970–98 or shorter).

Other financial indicators. Exchange rates are a channel through which inflation can be imported in open economies. In the United States, exchange rates (or a measure of the terms of trade) have long entered conventional Phillips curves. Robert Gordon (1982, 1998) finds these exchange rates statistically significant based on in-sample tests. In their international dataset, however, Goodhart and Hofmann (2000b) find that pseudo out-of-sample forecasts of inflation using exchange rates and lagged inflation outperformed autoregressive forecasts in only one or two of their seventeen countries, depending on the horizon. At least in the U.S. data, there is also little evidence that exchange rates predict output growth (e.g. Stock and Watson 1999a).

One problem with the nominal term structure as a predictor of inflation is that, under the expectations hypothesis, the forward rate embodies forecasts of both inflation and future real rates. In theory, one can eliminate the expected future real rates by using spreads between forward rates in the term structures of nominal and real debt of matched maturity and matched bearer risk. In practice, one of the few cases for which this is possible with time series of a reasonable length is for British index-linked bonds. David Barr and Campbell (1997) investigated the (bivariate, in-sample) predictive content of these implicit inflation expectations and found that they had better predictive content for inflation than forward rates obtained solely from the nominal term structure. They provided no evidence on Granger causality or marginal predictive content of these implicit inflation expectations in multivariate regressions.

Lettau and Sydney Ludvigson (2001) proposed a novel indicator, the log of the consumption-wealth ratio. They argue that in a representative consumer model with no stickiness in consumption, the log ratio of consumption to total wealth (human and nonhuman) should predict the return on the market portfolio. They find empirically that their version of the consumption–wealth ratio (a cointegrating residual between consumption of nondurables, financial wealth, and labor income, all in logarithms) has predictive content for multiyear stock returns (both real returns and excess returns). If consumption is sticky, it could also have predictive content for consumption growth. However, Ludvigson and Charles Steindel (1999) and Lettau and Ludvigson (2000) find that this indicator does not predict consumption growth or income growth in the United States one quarter ahead.

Housing constitutes a large component of aggregate wealth and gets significant weight in the CPI in many countries. More generally, housing is a volatile and cyclically sensitive sector, and measures of real activity in the housing sector are known to be useful leading indicators of economic activity, at least in the United States (Stock and Watson 1989; 1999a), suggesting a broader channel by which housing prices might forecast real activity, inflation, or both. In the United States, housing starts (a real quantity measure) have some predictive content for inflation (Stock 1998; Stock and Watson 1999b). Studies of the predictive content of housing prices confront difficult data problems, however. Goodhart and Hofmann (2000a) constructed a housing price data set for twelve OECD countries (extended to seventeen countries in Goodhart and Hofmann (2000b). They found that residential housing inflation has significant in-sample marginal

predictive content for overall inflation in a few of the several countries they study, although in several countries they used interpolated annual data which makes those results difficult to assess.

Nonlinear models. The foregoing discussion has focused on statistical models in which the forecasts are linear functions of the predictors; even the recession prediction models estimated using probit regressions are essentially linear in the sense that the predictive index (the argument of the probit function) is a linear function of the predictors. Might the problems of instability and episodically poor predictive content stem from inherent nonlinearities in the forecasting relation? The evidence on this proposition is limited and mixed. Ted Jaditz, Leigh Riddick, and Chera Sayers (1998) examine linear and nonlinear models of U.S. industrial production using asset price predictors and conclude that combined nonlinear forecasts improve upon simple linear models; additionally, Greg Tkacz (2001) reports improvements of nonlinear models over linear models for forecasting Canadian GDP. On the other hand, John Galbraith and Tkacz (2000) find limited international evidence of nonlinearity in the output–term spread relation. Similarly, in their pseudo out-of-sample comparison of VAR to multivariate neural network forecasts, Norman Swanson and Halbert White (1997) concluded that the linear forecasts generally performed better for various measures of U.S. economic activity and inflation; Swanson and White (1995) reached similar conclusions when forecasting interest rates. Given this limited evidence, we cannot rule out the possibility that the right nonlinear model will produce stable and reliable forecasts of output and inflation using interest rates, but this "right" nonlinear model has yet to be found.

3.2 *Forecasts Using Nonfinancial Variables*

The literature on forecasting output and inflation with nonfinancial variables is massive; see Stock and Watson (1999a) for an extensive review of the U.S. evidence. This section highlights a few recent studies on this topic.

The use of nonfinancial variables to forecast inflation has, to a large extent, focused on identifying suitable measures of output gaps, that is, estimating generalized Phillips curves. In the United States, the unemployment-based Phillips curve with a constant nonaccelerating inflation rate of unemployment (NAIRU) has recently been unstable, predicting accelerating inflation during a time that inflation was, in fact, low and steady or falling. This instability has been widely documented; see, for example, Gordon (1997, 1998) and Douglas Staiger, Stock, and Watson (1997a,b; 2001). One reaction to this instability has been to suggest that the NAIRU was falling in the United States during the 1990s. Mechanically, this keeps the unemployment-based Phillips curve on track, and it makes sense in the context of changes in the U.S. labor market and in the economy generally; cf. Lawrence Katz and Alan Krueger (1999). However, an imprecisely estimated time-varying NAIRU makes forecasting using the unemployment-based Phillips curve problematic.

A different reaction to this time variation in the NAIRU has been to see if there are alternative predictive relations that have been more stable. Staiger, Stock, and Watson (1997a) considered 71 candidate leading indicators of inflation, both financial and nonfinancial (quarterly, U.S.), and in a similar but more thorough exercise, Stock and Watson (1999b) considered 167 candidate leading indicators (monthly, U.S.). They found a few indicators that have been stable predictors of inflation, the prime example being the capacity utilization rate. Gordon (1998) and Stock (1998) confirmed the accuracy of recent U.S. inflation forecasts based on the capacity utilization rate. Stock and Watson (1999b) also suggested an alternative Phillips-curve-type forecast, based on a single aggregate activity index computed using 85 individual measures of real aggregate activity.

Recently, Andrew Atkeson and Lee Ohanian (2001) challenged the usefulness of all inflation forecasts based on the Phillips curve and its variants. They showed that, for the United States from 1984 to 2001, published inflation forecasts and pseudo out-of-sample Phillips curve forecasts did not beat a seasonal random walk forecast of annual inflation. Their finding poses a significant challenge to all attempts to forecast inflation, and we return to it in our empirical analysis.

The international evidence on the suitability of output gaps and the Phillips Curve for forecasting inflation is mixed. Simple unemployment-based models with a constant NAIRU fail in Europe, which is one way to state the phenomenon of so-called hysteresis in the unemployment rate. More sophisticated and flexible statistical tools for estimating the NAIRU can improve in-sample fits for the European data (Thomas Laubach 2001), but their value for forecasting is questionable because of imprecision in the estimated NAIRU at the end of the sample. Similarly, inflation forecasts based on an output gap rather than the unemployment rate face the practical problem of estimating the gap at the end of the sample, which necessarily introduces a one-sided estimate and associated imprecision. Evidence in Massimiliano Marcellino, Stock, and Watson (2000) suggests that the ability of output gap models to forecast inflation in Europe is more limited than in the United States.

Finally, there is evidence (from U.S. data) that the inflation process itself, as well as predictive relations based on it, is time-varying. William Brainard and George Perry (2000) and Timothy Cogley and Thomas Sargent (2001, 2002) suggested that the persistence of U.S. inflation was high in the 1970s and 1980s but subsequently declined, although this conclusion appears to be sensitive to the method used to measure persistence (Frederic Pivetta and Ricardo Reis 2002). George Akerlof, William Dickens, and Perry (2000) provided a model, based on near-rational behavior, that motivates a nonlinear Phillips curve, which they interpreted as consistent with the Brainard and Perry (2000) evidence.

In a similar vein, Stephen Cecchetti, Rita Chu, and Charles Steindel (2000) performed a pseudo out-of-sample forecasting experiment on various candidate leading indicators of inflation, from 1985 to 1998 in the United States, including interest rates, term and default spreads, and several nonfinancial indicators. They concluded that none of these indicators, financial or nonfinancial, reliably predicts inflation in bivariate forecasting models, and that there are very few years in which financial variables outperform a simple autoregression. Because they assessed performance on a year-by-year basis, these findings have great sampling variability and it is difficult to know how much of this is due to true instability. Still, their findings are consistent with Stock and Watson's (1996) results, based on formal stability tests, that time variation in these reduced form bivariate predictive relations is widespread in the U.S. data.

3.3 *Discussion*

An econometrician might quibble with some aspects of this literature. Many of these studies fail to include lagged endogenous variables and thus do not assess marginal predictive content. Results often change when marginal predictive content is considered (the predictive content of the term spread for inflation is an example). Many of the regressions involve overlapping returns, and when the overlap period is large relative to the sample size, the distribution of in-sample t-statistics and R^2s becomes nonstandard. Some regressors, such as the dividend yield and the term spread, are highly persistent, and even if they do not have a unit root this persistence causes conventional inference methods to break down. These latter two problems combined make it even more difficult to do reliable inference, and very few of these studies mention, far less tackle, either of these difficulties with their

in-sample regressions. Instability is a major focus of some of these papers, yet formal tests of stability are the exception. Finally, although some of the papers pay close attention to simulated forecasting performance, predictive content usually is assessed through in-sample fits that require constant parameters (stationarity) for external validity.

Despite these shortcomings, the literature does suggest four general conclusions. First, the variables with the clearest theoretical justification for use as predictors often have scant empirical predictive content. The expectations hypothesis of the term structure of interest rates suggests that the term spread should forecast inflation, but it generally does not once lagged inflation is included. Stock prices and log dividend yields should reflect expectations of future real earnings, but empirically they provide poor forecasts of real economic activity. Default spreads have the potential to provide useful forecasts of real activity, and at times they have, but the obvious default risk channel appears not to be the relevant channel by which these spreads have their predictive content. Moreover, the particulars of forecasting with these spreads seem to hinge on the current institutional environment.

Second, there is evidence that the term spread is a serious candidate as a predictor of output growth and recessions. The stability of this proposition in the United States is questionable, however, and its universality is unresolved.

Third, although only a limited amount of international evidence on the performance of generalized Phillips curve models was reviewed above, generalized Phillips curves and output gaps appear to be one of the few ways to forecast inflation that have been reliable. But this too seems to depend on the time and country.

Fourth, our reading of this literature suggests that many of these forecasting relations are ephemeral. The work on using asset prices as forecasting tools over the past fifteen years was in part a response to disappointment over the failure of monetary aggregates to provide reliable and stable forecasts or to be useful indicators of monetary policy. The evidence of the 1990s on the term spread, the paper–bill spread, and on some of the other theoretically suggested asset price predictors recalls the difficulties that arose when monetary aggregates were used to predict the turbulent late 1970s and 1980s: the literature on forecasting using asset prices apparently has encountered the very pitfalls that its participants hoped to escape. One complaint about forecasts based on monetary aggregates was the inability to measure money properly in practice. The results reviewed here suggest a more profound set of problems; after all, these asset prices are measured well, and in many cases the underlying financial instrument (a broad-based stock portfolio, short-term government debt) does not vary appreciably over time or even across countries. These observations point to a deeper problem than measurement: that the underlying relations themselves depend on economic policies, macroeconomic shocks, and specific institutions and thus evolve in ways that are sufficiently complex that real-time forecasting confronts considerable model uncertainty.

4. *Forecasting Models and Statistics*

We now turn to an empirical analysis of the predictive content of asset prices and other leading indicators for output growth and inflation using quarterly data from 1959 to 1999 (as available) for seven OECD countries. Our purpose is to provide a systematic replication, extension, and reappraisal of the findings in the literature reviewed in section 3. Real output is measured by real GDP and by the index of industrial production (IP). Inflation is measured by the percentage change of the consumer price index (CPI), or its counterpart, and of the implicit GDP deflator (PGDP). This section extends section 2 and discusses additional empirical methods. The data are discussed in section 5, and results are presented in sections 6 and 7.

4.1. *Forecasting Models*

The analysis uses h-step ahead linear regression models of the form (3). In addition, we examine the predictive content of X_t for Y^h_{t+h} after controlling for past values of Y_t *and* past values of a third predictor, Z_t; for example, we examine whether the predictive performance of a backward-looking Phillips curve is improved by adding an asset price. This is done by augmenting (3) to include lags of Z_t:

$$Y^h_{t+h} = \beta_0 + \beta_1(\mathrm{L})X_t + \beta_2(\mathrm{L})Y_t + \beta_3(\mathrm{L})Z_t + u^h_{t+h}. \quad (5)$$

The dependent variables are transformed to eliminate stochastic and deterministic trends. The logarithm of output is always treated as integrated of order one (I(1)), so that Y_t is the quarterly rate of growth of output at an annual rate. Because there is ambiguity about whether the logarithm of prices is best modeled as being I(1) or I(2), the empirical analysis was conducted using both transformations. The out-of-sample forecasts proved to be more accurate for the I(2) transformation, so to save space we present only those results. Thus for the aggregate price indices, Y_t is the first difference of the quarterly rate of inflation, at an annual rate. Transformations of the predictors are discussed in the next section.

Definitions of Y^h_{t+h}. The multistep forecasts examine the predictability of the logarithm of the level of the variable of interest, after imposing the I(1) or I(2) constraint. For output, we consider cumulative growth, at an annual percentage rate, of output over the h periods, so $Y^h_{t+h} = (400/h)\ln(Q_{t+h}/Q_t)$, where Q_t denotes the level of the real output series. For prices, we consider the h-period rate of inflation $(400/h)\ln(P_{t+h}/P_t)$, where P_t is the price level; upon imposing the I(2) constraint, this yields the dependent variable, $Y^h_{t+h} = (400/h)\ln(P_{t+h}/P_t) - 400\ln(P_t/P_{t-1})$.

Lag lengths and estimation. To make the results comparable across series and country, for the in-sample analysis we use a fixed lag length of four (so that the regressors in (3) are $X_t,\ldots, X_{t-3}, Y_t,\ldots, Y_{t-3}$). For the pseudo out-of-sample analysis, the lag length is data-dependent—specifically, chosen using the AIC—so that the model could adapt to potentially different dynamics across countries and over time. For the univariate forecasts, the AIC-determined lag length was restricted to be between zero and four. For the bivariate forecasts, between zero and four lags of Y_t were considered, and between one and four lags of X_t were considered. For the trivariate forecasts, between zero and four lags of Y_t were considered, and between one and four lags each of X_t and Z_t were considered.

4.2. *Model Comparison Statistics*

For each forecasting model, we computed both in-sample and pseudo out-of-sample statistics.

In-sample statistics. The in-sample statistics are the heteroskedasticity-robust Granger-causality test statistic, computed in a 1-step ahead regression ($h = 1$ in (3) and (5)), and the Richard Quandt (1960) likelihood ratio (QLR) test for coefficient stability, computed over all possible break dates in the central 70 percent of the sample.

The QLR statistic tests the null hypothesis of constant regression coefficients against the alternative that the regression coefficients change over time. Our version of the QLR statistic, also known as the sup-Wald statistic, entails computing the heteroskedasticity-robust Wald statistic testing for a break in the coefficients at a known date, then taking the maximum of those statistics over all possible break dates in the central 70 percent of the sample. Although this statistic is designed to detect a break at a single date, it has good power against other forms of parameter instability, including slowly drifting parameters (Stock and Watson 1998). The asymptotic null distribution of this statistic was derived by Donald Andrews (1993) (a corrected table of critical

values is provided in Stock and Watson 2003a, table 12.5).

Two versions of the QLR statistic were computed. The first tests only for changes in the constant term and the coefficients on X_t and its lags, that is, for a break in β_0 and $\beta_1(L)$ in (3) and (5) under the maintained hypothesis that the remaining coefficients are constant. The second tests for changes in all of the coefficients. The qualitative results were the same for both statistics. To save space, we report results for the first test only.

Pseudo out-of-sample statistics. The pseudo out-of-sample statistics are based on forecasts computed for each model, horizon, and series being forecasted. The model estimation and selection is recursive (uses all available prior data) as the forecasting exercise proceeds through time. We computed the sample relative MSFE defined in (4), relative to the AR benchmark, where both models have recursive AIC lag length selection. For most series, the out-of-sample forecasting exercise begins in the first quarter of 1971 and continues through the end of the sample period. For variables available from 1959 onward, this means that the first forecast is based on approximately ten years of data, after accounting for differencing and initial conditions. For variables with later start dates, the out of sample forecast period begins after accumulating ten years of data. The out of sample period is divided into two sub-periods, 1971–84 and 1985–99. These periods are of equal length for the four-quarter ahead forecasts. Because the models are nested, tests of the hypothesis that the population relative MSFE is one, against the alternative that it is less than one, are conducted using the asymptotic null distribution derived by Clark and McCracken (2001).

5. *Data*

We collected data on up to 26 series for each country from 1959 to 1999, although for some countries certain series were either unavailable or were available only for a shorter period. Data were obtained from four main sources: the International Monetary Fund IFS database, the OECD database, the DRI Basic Economics Database, and the DRI International Database. Additional series, including spreads, real asset prices, and *ex-ante* real interest rates, were constructed from these original 26 series, bringing the total number of series to 43. These 43 series are listed in table 1. The dates over which each series is available are listed in the appendix. The data were subject to five possible transformations, done in the following order.

First, a few of the series contained large outliers, such as spikes associated with strikes, variable re-definitions etc. (those series and outlier dates are listed in the Appendix). Those outliers were replaced with an interpolated value constructed as the median of the values within three periods on either side of the outlier.

Second, many of the data showed significant seasonal variation, and these series were seasonally adjusted. Seasonal variation was determined by a pre-test (regressing an appropriately differenced version of the series on a set of seasonal dummies) carried out at the 10-percent level. Seasonal adjustment was carried out using a linear approximation to X11 (Kenneth Wallis's 1974 for monthly series and Guy Larocque's 1977 for quarterly series) with endpoints calculated using autoregressive forecasts and backcasts.

Third, when the data were available on a monthly basis, the data were aggregated to quarterly observations. For the index of industrial production and the CPI (the variables being forecast) quarterly aggregates were formed as averages of the monthly values. For all other series, the last monthly value of the quarter was used as the quarterly value.

Fourth, in some cases the data were transformed by taking logarithms.

Fifth, the highly persistent or trending variables were differenced, second differenced, or computed as a "gap," that is, a deviation from a stochastic trend. Because the variables are being used for forecasting,

TABLE 1
SERIES DESCRIPTIONS

Series Label	Sampling Frequency	Description
		Asset Prices
rovnght	M	Interest rate: overnight
rtbill	M	Interest rate: short-term gov't. bills
rbnds	M	Interest rate: short-term gov't. bonds
rbndm	M	Interest rate: medium-term gov't. bonds
rbndl	M	Interest rate: long-term gov't. bonds
rrovnght	Q	Real overnight rate: rovnght–CPI inflation
rrtbill	Q	Real short-term bill rate: rtbill–CPI inflation
rrbnds	Q	Real short-term bond rate: rtbnds–CPI inflation
rrbndm	Q	Real medium-term bond rate: rtbndm–CPI inflation
rrbndl	Q	Real long-term bond rate: rtbndl–CPI inflation
rspread	M	Term spread: rbndl–rovnght
exrate	M	Nominal exchange rate
rexrate	M	Real exchange rate (exrate × relative CPIs)
stockp	M	Stock price index
rstockp	M	Real stock price index: stockp
divpr	Q	Dividend price index
house	Q	Housing price index
rhouse	Q	Real housing price index
gold	M	Gold price
rgold	M	Real gold price
silver	M	Silver price
rsilver	M	Real silver price
		Activity
rgdp	Q	Real GDP
ip	M	Index of industrial production
capu	M&Q	Capacity utilization rate
emp	M&Q	Employment
unemp	M&Q	Unemployment rate
		Wages, Goods and Commodity Prices
pgdp	Q	GDP deflator
cpi	M	Consumer price index
ppi	M	Producer price index
earn	M	Wages
commod	M	Commodity price index
oil	M	Oil price
roil	M	Real oil prices
rcommod	M	Real commodity price index
		Money
m0	M	Money: M0 or monetary base
m1	M	Money: M1
m2	M	Money: M2
m3	M	Money: M3
rm0	M	Real money: M0
rm1	M	Real money: M1
rm2	M	Real money: M2
rm3	M	Real money: M3

Notes: M indicates that the original data are monthly, Q indicates that they are quarterly, M&Q indicates that monthly data were available for some countries but quarterly data were available for others. All forecasts and regressions use quarterly data, which were aggregated from monthly data by averaging (for CPI and IP) or by using the last monthly value (all other series). Additional details are given in the appendix.

the gaps were computed in a way that preserved the temporal ordering. Specifically, the gaps here were estimated using a one-sided version of the filter proposed by Robert Hodrick and Edward Prescott (1981), details of which are given in the appendix.

For some variables, such as interest rates, it is unclear whether they should be included in levels or after first differencing, so for these variables we consider both versions. In all, this results in a maximum 73 candidate predictors per country for each of the inflation and output growth forecasts.

6. *Results for Models with Individual Indicators*

This section summarizes the empirical results for forecasts of inflation and output growth using individual predictors. Forecasts were made for two-, four- and eight-step ahead inflation and output growth (h = 2, 4, and 8 in (3) and (5)). Among the bivariate models (for which there is no Z variable in (5)), there are a total of 6,123 potential pairs of predictor and dependent variable over the three horizons and seven countries; of these, we have at least some empirical results for 5,080 cases.[4] To save space, we focus on four-quarter ahead forecasts of CPI inflation and GDP growth. A full set of results for all horizons and dependent variables is available in the results supplement, which is available on the web.[5]

[4] Because real interest rates are formed by subtracting CPI inflation for the current quarter, and because inflation is modeled in second differences, the CPI inflation forecast based on the regression (3), where X_t is the first difference of nominal interest rates, is the same as the CPI inflation forecast where X_t is the first difference of real interest rates, as long as the number of lags is the same in the two regressions and the number of interest rate lags is not more than the number of inflation lags. But because these two lag lengths are selected recursively by BIC, these conditions on the lags need not hold, so the relative MSFEs for CPI inflation forecasts based on nominal and real interest rates can (and do) differ.

[5] Full empirical results are available at http://www.wws.princeton.edu/~mwatson.

6.1 *Forecasts of Inflation*

The performance of the various individual indicators relative to the autoregressive benchmark is summarized in table 2 for four-quarter ahead forecasts of CPI inflation. The first row provides the root mean squared forecast error of the pseudo out-of-sample benchmark univariate autoregressive forecasts in the two sample periods. For the subsequent rows, each cell corresponds to an indicator/country pair, where the two entries are for the two sample periods. The second and third rows report the relative MSFEs of the no-change (random walk) forecast and of the seasonal no-change forecast, which is the Atkeson-Ohanian (2001) forecast at a quarterly sampling frequency.

Inspection of table 2 reveals that some variables forecast relatively well in some countries in one or the other subsamples. For example, the inflation forecast based on the nominal short rate has a relative MSFE of 0.68 in the first subsample in France, indicating a 32-percent improvement in this period relative to the benchmark autoregression; in Japan and the United Kingdom, stock prices produce a relative MSFE of .86 and .85 in the first period. Real activity measures are also useful in some country/variable/subsample cases, for example the capacity utilization rate works well for the United States during both subsamples, and M2 predicted inflation well for Germany in the first period.

These forecasting successes, however, are isolated and sporadic. For example, housing price inflation predicts CPI inflation in the first period in the United States, but it performs substantially worse than the AR benchmark in the second period in the United States and in the other countries. The short rate works well in France in the first period, but quite poorly in the second. The rate of increase of the price of gold occasionally is a useful predictor. Monetary aggregates rarely improve upon the AR model except for M2 and real M2 in the first period for Germany. Commodity price inflation works well in the

TABLE 2

PSEUDO OUT-OF-SAMPLE MEAN SQUARE FORECAST ERRORS: 1971–84 AND 1985–99, CPI INFLATION, 4 QUARTERS AHEAD

Indicator	Transformation	Canada		France		Germany		Italy		Japan		U.K.		U.S.	
		71–84	85–99	71–84	85–99	71–84	85–99	71–84	85–99	71–84	85–99	71–84	85–99	71–84	85–99
		Root Mean Square Forecast Error													
Univ. Autoregression		2.10	1.67	2.37	1.02	1.28	1.73	4.65	1.38	4.95	1.32	5.25	2.06	2.50	1.28
Univariate Forecasts		MSFE Relative to Univariate Autoregression													
$(1-L)^2 p_t = \varepsilon_t$		0.92	1.20	0.97	1.09	1.17	1.74	0.94	0.89	0.77	1.91	0.97	1.14	0.92	1.20
$(1-L^4)^2 p_t = \varepsilon_t$		1.17	0.76	1.04	1.06	0.97	0.85	0.99	1.03	0.90	0.92	0.91	1.01	1.19	0.79
Bivariate Forecasts		MSFE Relative to Univariate Autoregression													
rovnght	level		1.12	0.68	1.47	1.01	1.03		2.76	1.01	2.03		0.98	0.99	1.07
rtbill	level	1.08	1.07		2.12		1.15		1.80			1.71	0.90	0.92	1.03
rbnds	level								1.86				0.90	0.99	1.03
rbndm	level							1.65	0.94					1.01	0.96
rbndl	level	1.24	0.99	1.26	1.00	0.82	1.11	1.37	1.02		5.34	1.08	1.00	1.06	0.98
rovnght	Δ		1.03	1.07	1.05	0.99	0.97		2.10	1.00	0.97		1.15	1.05	0.99
rtbill	Δ	1.03	0.99		1.00		1.02		1.12			1.07	0.92	1.13	0.98
rbnds	Δ								1.04				0.90	1.02	0.99
rbndm	Δ							1.16	1.53					1.02	1.18
rbndl	Δ	1.27	0.98	1.07	1.05	0.94	1.02	1.20	1.15		2.41	1.06	1.09	0.98	1.17
rrovnght	level		1.06	1.16	0.97	1.21	0.97		1.46	1.54	1.30		1.43	1.30	1.07
rrtbill	level	1.49	1.01		1.22		1.24		1.16			0.89	1.74	1.38	0.96
rrbnds	level								0.96				1.48	1.27	0.98
rrbndm	level							1.26	1.62					1.35	1.11
rrbndl	level	1.24	0.93	1.30	1.48	1.12	0.82	1.33	1.74		1.26	0.96	1.34	1.32	1.12
rrovnght	Δ		0.89	1.06	0.87	0.99	0.97		0.92	1.01	1.03		1.18	1.14	0.98
rrtbill	Δ	1.04	0.87		0.98		1.10		1.08			0.88	0.97	1.10	0.97
rrbnds	Δ								0.90				0.88	0.99	0.98
rrbndm	Δ							1.15	1.06					0.96	1.02
rrbndl	Δ	1.18	0.88	0.92	1.04	0.93	1.04	1.18	1.16		1.17	1.00	1.06	0.97	1.05
rspread	level		1.07	1.10	1.46	1.13	1.01		2.55		1.24		1.06	0.91	1.40
exrate	Δln		0.98		1.24		1.21		1.03		1.77		1.23		2.12
rexrate	Δln		0.93		1.32		1.08		0.92		1.88		1.06		2.12
stockp	Δln	0.99	1.12	1.18	1.01	1.02	1.01	1.35	1.07	0.86	2.64	0.85	1.21	0.95	1.20
rstockp	Δln	1.00	1.14	1.11	1.01	1.01	1.01	1.26	1.14	0.83	2.83	0.93	1.17	0.94	1.22
divpr	ln		1.54		1.96		1.20		1.05		4.33		2.01	1.09	1.22
house	Δln		1.16								6.60		0.97	0.86	1.11
rhouse	ln		1.26								4.53		2.02	0.91	1.11
rhouse	Δln		1.20								3.91		1.08	0.70	1.04
gold	Δln	1.02	0.95	1.06	0.91	1.19	0.99	1.14	0.95	2.02	0.93	0.90	0.92	1.43	1.03
gold	Δ^2ln	1.30	1.01	1.00	0.99	1.05	1.00	0.95	1.01	1.01	0.99	1.03	1.00	1.02	1.10
rgold	ln	1.19	0.93	2.03	0.98	1.16	1.02	1.54	1.05	1.51	1.26	1.18	1.00	2.20	0.93
rgold	Δln	0.94	0.91	1.24	0.92	1.17	0.98	1.06	1.18	1.67	0.89	0.94	0.92	1.31	0.90
silver	Δln		1.06		1.09		1.06		1.08		1.11		0.96		1.13
silver	Δ^2ln		1.05		1.03		1.06		0.98		1.17		1.08		1.17
rsilver	ln		1.11		1.65		1.11		2.93		2.47		1.45		1.39
rsilver	Δln		1.01		1.10		1.06		1.15		1.09		0.95		1.12
rgdp	Δln	1.00	0.89		1.00	0.94	0.99	0.90	0.89	1.03	1.77	1.04	0.91	0.82	0.84
rgdp	gap	0.99	0.84		1.30	0.82	0.94	0.92	1.13	1.07	0.84	1.06	0.88	0.85	0.94
ip	Δln	0.99	0.84	1.00	1.04	0.95	1.07	0.94	0.81	0.95	1.43	1.05	0.96	0.83	0.86
ip	gap	1.00	0.91	1.00	1.07	0.85	1.06	0.98	1.16	1.05	1.00	0.92	0.91	0.77	0.97
capu	level	1.03	0.70		2.21		1.03		1.96		2.55			0.74	0.80
emp	Δln	0.98	0.95		1.63	0.81	1.06			1.00	1.86	0.85	0.90	0.74	0.89

TABLE 2 *(cont.)*

Indicator	Transformation	Canada		France		Germany		Italy		Japan		U.K.		U.S.	
		71–84	85–99	71–84	85–99	71–84	85–99	71–84	85–99	71–84	85–99	71–84	85–99	71–84	85–99
emp	gap	0.89	0.78		2.53	0.82	1.05			1.04	1.19	0.91	1.38	0.65	1.04
unemp	level	1.16	0.81		3.69	1.02	0.98	1.15	1.30	1.19	2.32	1.06	0.88	0.76	0.89
unemp	Δln	0.97	0.91		0.83	0.82	0.98	1.01	1.26	0.98	2.06	0.89	1.14	0.78	0.97
unemp	gap	0.92	0.76		1.14	0.83	0.96	1.08	1.13	1.13	1.17	0.88	0.88	0.75	1.02
pgdp	Δln	1.08	1.02		2.36	1.00	0.99	1.10	1.02	1.16	1.49	0.99	1.19	1.06	1.08
pgdp	Δ^2ln	1.02	1.00		1.02	0.98	1.00	1.03	0.99	0.98	1.10	0.99	1.01	1.00	0.98
ppi	Δln	1.36	0.92			1.82	0.98			1.34	1.39	1.03	1.02	1.22	0.91
ppi	Δ^2ln	1.00	0.91			1.06	0.98			0.75	1.29	0.90	0.99	1.05	0.93
earn	Δln	1.09	1.03	1.07	1.11	1.03	0.95			1.29	1.05	1.28	0.95	1.10	1.03
earn	Δ^2ln	1.03	1.00	1.00	0.99	0.99	1.00			1.03	1.03	1.05	0.99	1.00	0.99
oil	Δln	1.16	0.93	2.04	1.01	1.23	1.00	0.91	1.62	2.40	1.47	0.99	1.08	1.09	0.99
oil	Δ^2ln	1.22	0.96	1.49	0.99	1.29	0.98	0.92	1.60	0.97	0.98	1.05	1.00	1.03	0.89
roil	ln	1.57	0.95	1.14	0.71	1.10	0.99	1.78	0.96	1.44	1.77	1.11	1.66	2.81	0.86
roil	Δln	1.11	0.92	1.89	1.04	1.05	1.00	1.08	1.47	2.06	1.23	0.95	1.38	1.01	0.99
comod	Δln	1.12	0.91	1.20	1.02	1.05	0.99	1.03	0.97	1.36	1.98	1.02	0.84	0.79	1.26
comod	Δ^2ln	1.00	1.01	1.13	1.34	1.02	1.00	0.99	1.48	1.05	2.06	1.05	1.00	0.99	1.64
rcomod	ln	1.23	0.89	1.28	1.12	1.21	1.14	1.08	1.38	1.13	2.26	0.96	1.60	0.79	1.44
rcomod	Δln	1.03	0.85	1.14	1.07	1.03	0.97	0.90	1.18	0.97	2.05	0.94	0.82	0.68	1.34
m0	Δln										1.49			1.05	1.12
m0	Δ^2ln										2.81			1.00	1.05
m1	Δln	1.28	1.03		1.23	1.06	1.08	0.95	0.96	1.31	1.39			0.95	1.20
m1	Δ^2ln	1.09	1.02		1.23	1.01	1.05	1.01	0.94	1.03	1.32			1.01	1.05
m2	Δln		1.24			0.75	1.22		2.37		3.20			1.04	1.01
m2	Δ^2ln		1.30			0.99	1.05		1.61		1.78			1.02	1.03
m3	Δln		1.24	0.99	0.97		1.08	1.01	0.90		3.17			1.03	1.02
m3	Δ^2ln		1.18	0.98	1.00		1.03	1.07	0.94		3.09			1.00	0.96
rm0	Δln										2.71			0.80	1.39
rm1	Δln	1.14	1.12		1.79	1.05	1.01	0.90	1.02	1.36	1.44			0.83	1.65
rm2	Δln		1.23			0.65	1.22		2.32		2.73			0.97	0.95
rm3	Δln		1.30	0.92	1.07		0.95	0.92	1.45		2.20			0.89	1.13

Notes: The two entries in each cell are results for first and second out-of-sample forecast periods (1971–84 and 1985–99). The first row shows the root mean square forecast error for the univariate autoregression. All other entries are the mean square forecast errors (MSFE) relative to the MSFE for the univariate autoregression. The second and third row present relative MSFEs for alternative univariate benchmarks, the no-change forecast of inflation (inflation follows a random walk) and the seasonal (four-lag) no-change forecast of inflation. For the entries labeled *Bivariate Forecasts,* the first column lists the indicator and the second column lists the transformation used for the indicator. Let S_t denote the original series, and X_t denote the series used in the regression (3). The transformations are:

level	$X_t = S_t$
Δ	$X_t = S_t - S_{t-1}$
ln	$X_t = \ln S_t$
Δln	$X_t = \ln S_t - \ln S_{t-1}$
Δ^2ln	$X_t = (\ln S_t - \ln S_{t-1}) - (\ln S_{t-1} - \ln S_{t-2})$.

United States in the first period but not in the second; in Canada, it works well in the second period but not in the first; and in some country/period combinations it works much worse than the AR benchmark. This instability is also evident in the two other univariate forecasts, the no-change and seasonal no-change. For example, the seasonal no-change forecast works well in the United States in the second period but poorly in the first, a similar pattern as in Canada (but the opposite pattern as in the United Kingdom).

The only set of predictors that usually improves upon the AR forecasts is the measures of aggregate activity. For example, the IP and unemployment gaps both improve upon the AR (or are little worse than the AR) for both periods for Canada, Germany, and the United States. Even for these predictors, however, the improvement is neither universal nor always stable.

One possible explanation for this apparent instability is that it is a statistical artifact: after all, these relative MSFEs have a sampling distribution, so there is sampling uncertainty (estimation error) associated with the estimates in table 2. To examine this possibility we used the results in Clark and McCracken (2001) to test the hypothesis that the relative MSFE is one, against the alternative that it is less than one. The Clark-McCracken (2001) null distribution of the relative MSFE cannot be computed for the statistics in table 2 because the number of lags in the models changes over time, so instead we computed it for pseudo out-of-sample forecasts from models with fixed lag lengths of four (complete results are available in the web results supplement). With fixed lag lengths, the 5-percent critical value ranges from 0.92 to 0.96 (the null distribution depends on consistently estimable nuisance parameters so it was computed by simulation on a series-by-series basis, yielding series-specific critical values). Most of the improvements over the AR model are statistically significant. In this sense, it appears that the observed temporal instability of the MSFEs is not a consequence of sampling variability alone for series that in population have no predictive content.

6.2 *Forecasts of Output Growth*

Table 3, which has the same format as table 2, summarizes the performance of the individual forecasts of real GDP growth at the four quarter horizon (results for the other horizons for GDP, and for all horizons for IP, are given in the results supplement described in footnote 5). The results in table 3 are consistent with the literature surveyed in section 3. The forecasts based on the term spread are of particular interest, given their prominence in that literature. In the United States, these forecasts improve upon the AR benchmark in the first period, but in the second period they are much worse than the AR forecasts (the relative MSFE is 0.48 in the first period but 2.51 in the second). This is consistent with the literature reviewed in section 3.1, which found a deterioration of the forecasting performance of the term spread as a predictor of output growth in the United States since 1985. In Germany, the term spread is useful in the first period but not in the second (the relative MSFEs are 0.51 and 1.09). The cross-country evidence on usefulness is also mixed: the term spread beats the AR benchmark in the second period in Canada and Japan, but not in France, Germany, Italy, the United Kingdom, or the United States.

The picture for other asset prices is similar. The level of the short rate is a useful predictor in Japan in the second period but not the first, and in Germany and the United States in the first period but not the second. The nominal exchange rate outperforms the AR benchmark in the second period in Canada, Germany, Italy, and Japan, but not in France, the United Kingdom, or the United States. Real stock returns improve upon the AR benchmark in the first period, but not the second, in Canada, Germany, Japan, and the United States.

Predictors that are not asset prices fare no better, or even worse. Forecasts based on

TABLE 3

PSEUDO OUT-OF-SAMPLE MEAN SQUARE FORECAST ERRORS:
1971–84 AND 1985–99, REAL GDP GROWTH, 4 QUARTERS AHEAD

Indicator	Transformation	Canada		France		Germany		Italy		Japan		U.K.		U.S.	
		71–84	85–99	71–84	85–99	71–84	85–99	71–84	85–99	71–84	85–99	71–84	85–99	71–84	85–99
		Root Mean Square Forecast Error													
Univ. Autoregression		2.91	2.55	1.90	1.56	2.83	1.84	3.47	1.88	3.59	2.46	2.96	1.89	3.19	1.31
Univariate Forecasts		MSFE Relative to Univariate Autoregression													
$(1-L)y_t = \alpha + \varepsilon_t$		0.97	0.99		1.13	1.04	1.04	1.05	1.37	1.51	2.88	1.03	0.98	0.98	1.09
Bivariate Forecasts		MSFE Relative to Univariate Autoregression													
rovnght	level		0.74		1.57	0.30	1.48		1.48	1.21	0.89		1.13	0.78	1.42
rtbill	level	0.59	0.72		1.63		1.28		0.86			1.19	0.79	0.85	1.06
rbnds	level								0.92				0.87	0.92	1.29
rbndm	level							1.56	1.40					1.11	1.47
rbndl	level	0.80	0.94		1.59	0.46	2.12	1.16	1.55		0.88	1.00	0.96	1.18	1.66
rovnght	Δ		0.68		0.91	1.09	1.17		0.85	1.05	0.98		1.21	1.11	1.57
rtbill	Δ	1.05	1.03		0.98		1.37		0.48			1.24	1.11	1.32	1.63
rbnds	Δ								0.59				1.04	1.19	1.85
rbndm	Δ							1.11	1.42					1.01	2.17
rbndl	Δ	1.08	1.25		1.12	0.91	1.64	1.24	1.26		0.89	0.97	0.92	0.90	2.38
rrovnght	level		0.56		0.99	0.64	1.42		0.78	1.27	1.07		1.18	1.29	1.00
rrtbill	level	1.09	0.64		0.99		1.34		0.57			0.98	1.07	1.35	1.11
rrbnds	level								0.60				0.95	1.38	1.10
rrbndm	level							1.26	1.37					1.27	1.41
rrbndl	level	1.06	0.99		1.04	1.09	1.29	1.23	1.45		0.87	1.18	0.84	1.38	1.50
rrovnght	Δ		0.77		0.98	1.00	1.16		0.92	1.03	1.02		1.21	1.05	1.01
rrtbill	Δ	1.06	1.03		0.99		1.20		0.66			1.31	1.03	1.50	1.01
rrbnds	Δ								0.60				1.04	1.50	1.01
rrbndm	Δ							1.03	1.00					1.54	1.02
rrbndl	Δ	1.03	1.02		1.01	1.08	1.20	1.03	1.00		0.84	1.37	1.00	1.52	1.02
rspread	level		0.67		1.04	0.51	1.09		1.11		0.83		1.35	0.48	2.51
exrate	Δln		0.72		1.08		0.95		0.83		0.95		1.31		1.24
rexrate	Δln		0.72		1.08		1.15		0.77		0.94		1.27		1.24
stockp	Δln	0.96	1.02		0.99	0.92	1.50	1.11	1.04	0.98	1.01	0.98	1.00	0.90	1.27
rstockp	Δln	0.91	1.05		1.00	0.87	1.62	1.06	1.14	0.97	1.03	0.91	0.88	0.82	1.65
divpr	ln		0.83		1.15		1.44		1.54		1.41		1.10	1.01	1.47
house	Δln		0.84								1.19		1.34	1.06	0.93
rhouse	ln		0.69								1.00		1.76	1.47	1.33
rhouse	Δln		0.81								1.17		1.23	0.98	0.93
gold	Δln	1.27	0.96		1.12	1.28	1.04	1.09	1.05	1.56	1.05	1.21	1.24	1.39	1.01
gold	Δ^2ln	1.09	1.01		1.00	1.04	1.00	1.01	1.01	1.09	1.03	1.10	1.00	1.10	1.01
rgold	ln	1.25	0.73		2.12	1.24	1.30	1.33	1.31	1.23	1.52	1.02	1.25	1.61	1.05
rgold	Δln	1.28	0.95		1.08	1.24	1.03	1.08	1.01	1.47	1.04	1.18	1.15	1.32	1.01
silver	Δln		0.79		1.35		0.91		0.77		0.97		1.33		1.00
silver	Δ^2ln		0.82		1.00		0.90		0.72		1.02		1.08		0.97
rsilver	ln		1.15		2.88		1.66		1.24		0.96		2.02		1.35
rsilver	Δln		0.78		1.33		0.91		0.80		0.97		1.37		1.01
ip	Δln	0.97	0.89		1.04	0.96	0.93	0.98	1.11	0.99	1.02	1.15	1.00	1.00	1.00
ip	gap	1.03	0.99		1.20	1.04	1.00	0.97	1.01	1.09	0.95	1.08	1.08	1.12	1.07
capu	level	1.22	1.00		1.29		1.06		0.55		0.92			0.87	1.13
emp	Δln	1.17	0.95		0.97	0.98	0.82			1.02	1.01	1.20	0.99	1.00	1.00
emp	gap	1.10	1.04		1.20	1.09	1.36			1.00	1.01	1.10	1.04	1.52	1.06
unemp	level	1.28	0.96		1.49	1.47	0.97	1.17	0.97	1.09	0.84	1.52	0.93	1.14	1.06
unemp	Δln	1.08	1.03		1.08	0.92	1.12	1.09	1.05	1.05	1.01	1.39	1.07	0.97	1.05

TABLE 3 *(cont.)*

Indicator	Transformation	Canada		France		Germany		Italy		Japan		U.K.		U.S.	
		71–84	85–99	71–84	85–99	71–84	85–99	71–84	85–99	71–84	85–99	71–84	85–99	71–84	85–99
unemp	gap	1.05	1.07	1.39		0.96	1.00	0.88	1.09	1.02	0.96	1.01	1.19	1.01	1.22
pgdp	Δln	0.81	1.34	1.39		1.00	1.07	0.76	1.89	1.35	1.10	0.95	0.89	0.84	1.60
pgdp	Δ^2ln	1.01	1.01	1.00		1.01	0.99	0.99	0.98	1.01	0.98	1.02	1.00	1.27	1.02
cpi	Δln	0.77	1.31	1.98		1.21	1.95	0.87	2.08	1.13	1.10	1.18	0.82	0.85	1.41
cpi	Δ^2ln	1.03	1.01	1.04		1.00	1.03	1.05	1.01	1.02	1.03	1.23	1.04	1.29	1.22
ppi	Δln	0.92	1.31			0.43	2.15			1.51	1.22	1.03	0.94	0.90	1.78
ppi	Δ^2ln	1.03	1.02			1.03	0.99			1.14	1.01	1.13	1.00	1.02	1.01
earn	Δln	0.91	1.24	1.13		1.14	0.98			1.23	0.97	0.86	0.95	0.94	2.04
earn	Δ^2ln	1.02	1.02	1.00		0.99	0.95			0.98	1.03	0.98	1.03	1.04	1.01
oil	Δln	2.23	1.02	1.42		1.27	1.75	0.99	1.65	1.74	1.01	1.65	1.14	1.90	1.54
oil	Δ^2ln	1.61	1.02	1.01		1.41	1.00	1.06	1.01	2.04	1.02	1.71	1.04	1.86	1.02
roil	ln	1.92	0.98	1.86		1.48	1.30	1.46	1.55	1.75	1.30	1.39	1.16	5.07	1.38
roil	Δln	2.42	1.01	1.38		1.55	1.51	1.00	1.53	1.64	1.01	1.89	1.15	3.79	1.37
comod	Δln	1.04	1.00	1.50		1.21	1.69	1.15	1.59	1.62	1.16	1.29	1.22	1.24	1.45
comod	Δ^2ln	1.02	1.02	1.02		1.02	1.02	0.99	1.04	1.02	1.02	1.03	1.01	1.07	1.06
rcomod	ln	0.90	1.37	1.36		1.14	0.96	1.44	1.56	1.29	1.19	1.40	1.51	1.19	2.34
rcomod	Δln	1.02	1.00	1.38		1.27	1.35	1.13	1.03	1.64	1.09	1.23	0.99	1.32	1.01
m0	Δln										1.05			1.05	1.02
m0	Δ^2ln										1.04			1.05	1.03
m1	Δln	0.99	0.85	1.76		0.80	1.64	0.91	0.75	1.25	0.98			1.01	1.08
m1	Δ^2ln	1.02	0.97	0.85		1.04	1.01	0.94	0.89	0.80	0.95			1.03	1.18
m2	Δln		0.87			1.02	0.90		0.68		0.67			0.98	1.33
m2	Δ^2ln		0.82			1.01	0.99		0.63		0.87			1.14	0.99
m3	Δln		0.84	1.80			1.05	1.01	0.55		1.02			1.20	0.91
m3	Δ^2ln		0.82	1.04			0.90	0.95	0.91		1.09			1.15	1.00
rm0	Δln										1.14			0.65	2.81
rm1	Δln	0.65	1.13	0.74		0.58	1.96	0.69	0.93	0.68	1.14			0.62	3.51
rm2	Δln		0.89			1.18	0.91		0.82		0.80			0.57	1.41
rm3	Δln		0.89	1.67			1.23	0.58	0.68		0.91			0.79	1.06

Notes: The second row presents the relative MSFE of a constant-change forecast of GDP (GDP follows a random walk with drift). See the notes to Table 2.

money growth sometimes outperform the AR benchmark but usually do not. In some cases, entire classes of predictors fail to improve upon the AR forecast. For example, oil prices and commodity prices typically produce forecasts much worse than the AR forecast, and forecasts based on output gaps generally perform slightly worse than the AR forecasts.

As was the case for the inflation forecasts, the Clark-McCracken (2001) critical value for the fixed-length models with four lags indicate that many of the improvements evident in table 3 are statistically significant, so the apparent instability cannot be explained just by sampling (estimation) uncertainty of the sample relative MSFE.

6.3 *Forecast Stability*

The foregoing discussion highlighted some examples in which forecasts made using a given asset price predictor in a certain country did well in one period, but poorly in another. This could, of course, simply reflect the examples we chose. We therefore look more systematically at the stability of forecasts made using a given predictor/country/horizon

TABLE 4
SUMMARY OF PSEUDO OUT-OF-SAMPLE FORECAST ACCURACY FOR TWO PERIODS: ASSET PRICE PREDICTORS, 4 QUARTER HORIZON

A. Inflation (N = 211)

		1971–84 Out-of-Sample Period		
		Relative MSFE < 1	Relative MSFE > 1	Total
1985–99 Out-of-Sample Period	Relative MSFE < 1	0.06	0.29	0.36
	Relative MSFE > 1	0.18	0.46	0.64
	Total	0.25	0.75	1.00

B. Output (N = 211)

		1971–84 Out-of-Sample Period		
		Relative MSFE < 1	Relative MSFE > 1	Total
1985–99 Out-of-Sample Period	Relative MSFE < 1	0.10	0.16	0.26
	Relative MSFE > 1	0.21	0.53	0.74
	Total	0.31	0.69	1.00

Notes: Each table shows the fraction of relative means square forecast errors (MSFE) less than 1 or greater than 1 for each sample period, relative to the univariate autoregressive benchmark. Results shown are pooled for all pairs of asset price predictors and inflation measures (part A) or output measures (part B) for all countries at horizon $h = 4$.

combination, as measured by the relative MSFE in the two periods.

Summary evidence on the stability of forecasting relations based on asset prices is given in table 4. Table 4a presents a cross-tabulation of 211 four-quarter ahead forecasts of inflation for all possible predictor-dependent variable pairs for the different countries (this table includes results for both the GDP deflator and the CPI). Of the 211 asset price/country/dependent variable combinations, 6 percent have relative MSFEs less than one in both the first and second period, that is, 6 percent outperform the AR benchmark in both periods; 18 percent outperform the benchmark in the first but not the second period, 29 percent in the second but not the first, and 46 percent are worse than the benchmark in both periods. Table 4b presents analogous results for output (the table includes results for both IP and real GDP growth).

The binary variables cross-tabulated in table 4 appear to be approximately independently distributed: the joint probabilities are very nearly the product of the marginal probabilities. For example, in panel A, if the row and column variables were independent then the probability of an indicator/country/horizon/dependent variable combination outperforming the benchmark would be $.25 \times .36 = .09$; the empirically observed probability is slightly less, .06. In panel B, the analogous predicted probability of outperforming the benchmark in both periods, computed under independence, is $.31 \times .26$

equals .08; the observed probability is slightly more, .10. Because the draws are not independent, the chi-squared test for independence of the row and column variables is inappropriate. Still, these calculations suggest that whether an asset price/country/horizon/dependent variable combination outperforms the benchmark in one period is nearly independent of whether it does so in the other period.

This lack of a relation between performance in the two subsamples is also evident in figures 1a (inflation) and 1b (output), which are scatterplots of the logarithm of the relative MSFE in the first vs. second periods for the 211 asset price-based forecasts analyzed in table 4a and 4b, respectively. An asset price forecast that outperforms the AR benchmark in both periods appears as a point in the southwest quadrant.

One view of the universe of potential predictors is that there are some reliable ones (perhaps the term spread or the short rate), while many others (perhaps gold and silver prices) have limited value and thus have regression coefficients near zero. If so, the points in figure 1 would be scattered along a forty-five degree line in the southwest quadrant, with many points clustered near the origin. But this is not what figure 1 looks like: instead, there are very few predictors near the forty-five degree line in the southwest quadrant, and there are too many points far from the origin, especially in the northwest and southeast quadrants. If anything, the view that emerges from figure 1 is that performance in the two periods is nearly unrelated or, if it is related, the correlation is *negative*: asset prices that perform well in the first period tend to perform poorly in the second period (the correlations in figures 1a and 1b are $-.22$ and $-.21$).

This pattern of instability is evident whether we aggregate or disaggregate the forecasts, and is also present at shorter and longer forecast horizons. Figure 2 presents the comparable scatterplot of first v. second period log relative MSFEs for all 1080 available combinations of predictors (asset prices and otherwise), countries, and dependent variables at the four-quarter horizon; the pattern is similar to that in figure 1, also showing a negative correlation (the correlation is $-.08$). Similarly, as seen in table 5, this instability is evident when the forecasts are disaggregated by country. As reported in table 5, the product of the marginal probabilities of beating the AR in the first period, times that for the second period, very nearly equals the joint probability of beating the AR in both periods, for all countries, for either output or inflation, and for the two- and eight-quarter horizons as well as the four-quarter horizon. A predictor that worked well in the first period—asset price or otherwise—is no more (and possibly less) likely to beat the AR in the second period than a predictor drawn at random from our pool.

These findings of forecast instability are also present in forecasts based on fixed lag lengths (see the results supplement), so the instability appears not to be an artifact of recursive BIC lag length selection. Interestingly in many cases the relative RMSFEs of the BIC- and fixed-lag forecasts differ substantially, by .10 or more. Although the overall conclusions based on tables 2–5 and figures 1 and 2 are the same for fixed- and recursive-lag forecasts, the conclusions for individual indicators sometimes differ by a surprising amount, and this sensitivity to lag selection methods could be another indication of instability in the forecasting relations.

In short, there appear to be no subsets of countries, predictors, horizons, or variables being forecast that are immune to this instability. Forecasting models that outperform the AR in the first period may or may not outperform the AR in the second, but whether they do appears to be random.

6.4 *In-Sample Tests for Predictive Content and Instability*

Because the literature surveyed in section 2 mainly uses in-sample statistics, this

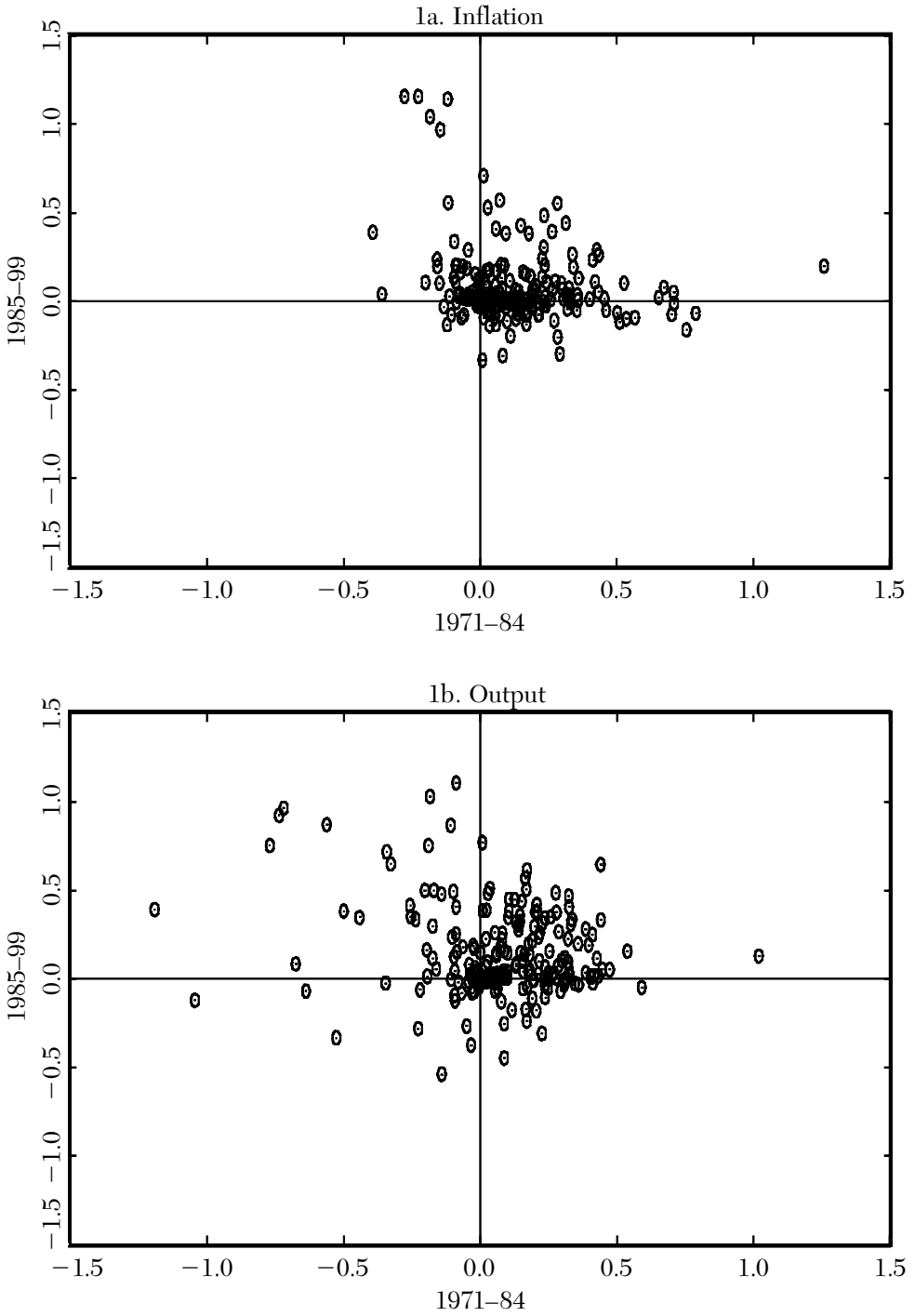

Figure 1. Scatterplot of Pseudo Out-of-Sample Log Relative Mean Squared Forecast Errors: 4-Quarter Ahead Forecasts, Asset Price Predictors

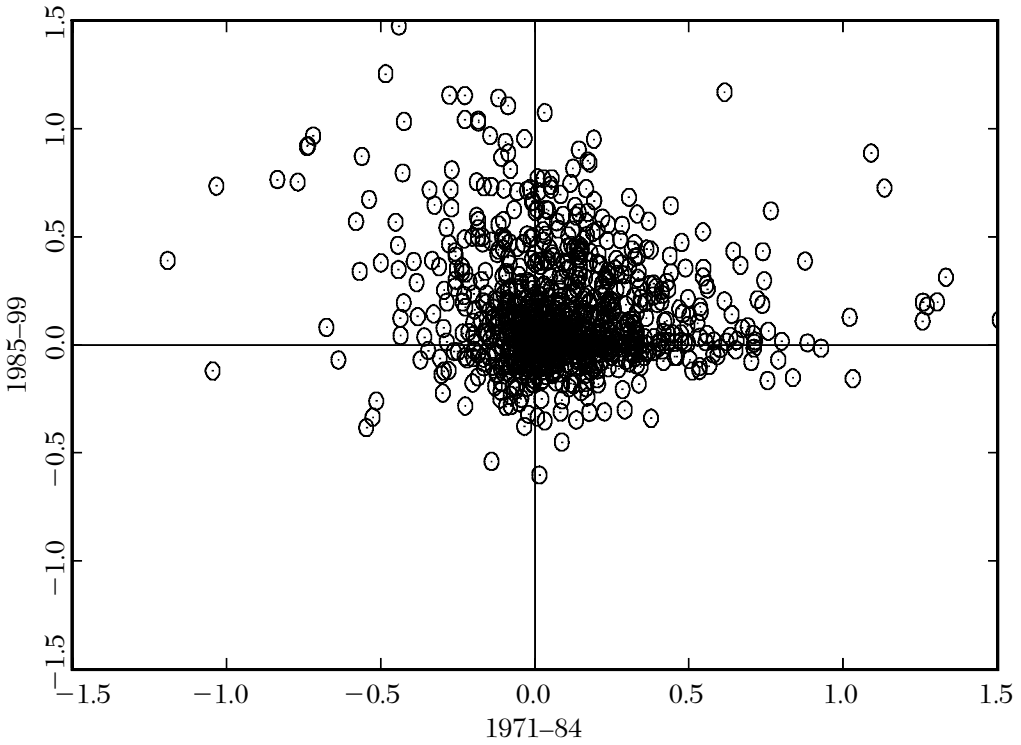

Figure 2. Scatterplot of Pseudo Out-of-Sample Log Relative Mean Squared Forecast Errors: 4-Quarter Ahead Forecasts, All Predictors

section turns to in-sample measures of predictive content and stability for these asset-price based forecasts. The results of full-sample Granger causality tests for predictive content and QLR tests for instability suggest three conclusions.

First, the Granger causality tests frequently reject, indicating that a large fraction of these relations have substantial in-sample predictive content. The Granger causality test results are summarized in the first entry in each cell in table 6. For both inflation and output forecasts, the Granger causality test rejects the null hypothesis of no predictive content for 35 percent of the asset prices. This evidence of predictive content is not surprising: after all, these variables were chosen in large part because they have been identified in the literature as useful predictors. Inspection of the results for each individual indicator/country/dependent variable combination (given in the results supplement) reveals individual Granger causality results that are consistent with those in the literature. For example, the term spread is a statistically significant predictor of GDP growth at the 1-percent level in Canada, France, Germany, and the United States, but not (at the 10-percent level) in Italy, Japan, or the United Kingdom. Exchange rates (real or nominal) are not significant predictors of GDP growth at the 5-percent level for any of the countries, but short-term interest rates are significant for most of the countries. The Granger causality tests suggest that housing prices have predictive content for IP growth, at least in some countries. Real activity variables (the IP gap, the unemployment rate, and capacity utilization) are significant in

TABLE 5
SUMMARY OF PSEUDO OUT-OF-SAMPLE FORECAST ACCURACY FOR TWO PERIODS: ALL PREDICTORS

	Inflation					Output				
Country	1st	2nd	1&2	1x2	N	1st	2nd	1&2	1x2	N
Canada										
2Q Ahead	0.36	0.34	0.10	0.12	80	0.20	0.54	0.13	0.11	80
4Q Ahead	0.46	0.33	0.10	0.15	80	0.31	0.66	0.24	0.21	80
8Q Ahead	0.34	0.38	0.14	0.13	80	0.31	0.47	0.20	0.15	80
France										
2Q Ahead	0.21	0.24	0.03	0.05	29	0.21	0.28	0.07	0.06	29
4Q Ahead	0.34	0.31	0.14	0.11	29	0.28	0.38	0.14	0.10	29
8Q Ahead	0.31	0.41	0.14	0.13	29	0.14	0.21	0.00	0.03	29
Germany										
2Q Ahead	0.45	0.28	0.14	0.13	86	0.35	0.35	0.15	0.12	86
4Q Ahead	0.47	0.24	0.12	0.11	86	0.36	0.31	0.14	0.11	86
8Q Ahead	0.45	0.27	0.15	0.12	86	0.40	0.40	0.17	0.16	86
Italy										
2Q Ahead	0.24	0.44	0.18	0.10	72	0.28	0.33	0.11	0.09	72
4Q Ahead	0.35	0.32	0.21	0.11	72	0.32	0.29	0.11	0.09	72
8Q Ahead	0.35	0.39	0.17	0.14	72	0.28	0.38	0.17	0.10	72
Japan										
2Q Ahead	0.37	0.24	0.09	0.09	70	0.19	0.30	0.01	0.06	70
4Q Ahead	0.17	0.30	0.09	0.05	70	0.23	0.16	0.04	0.04	70
8Q Ahead	0.31	0.27	0.17	0.09	70	0.19	0.13	0.03	0.02	70
United Kingdom										
2Q Ahead	0.14	0.43	0.07	0.06	72	0.24	0.35	0.06	0.08	72
4Q Ahead	0.24	0.35	0.14	0.08	72	0.43	0.39	0.14	0.17	72
8Q Ahead	0.36	0.36	0.17	0.13	72	0.56	0.29	0.17	0.16	72
United States										
2Q Ahead	0.30	0.20	0.02	0.06	132	0.37	0.44	0.20	0.16	132
4Q Ahead	0.38	0.17	0.05	0.07	132	0.33	0.35	0.10	0.12	132
8Q Ahead	0.39	0.18	0.07	0.07	132	0.52	0.33	0.17	0.17	132
All										
2Q Ahead	0.31	0.30	0.09	0.09	541	0.28	0.39	0.12	0.11	541
4Q Ahead	0.35	0.27	0.11	0.10	541	0.33	0.36	0.13	0.12	541
8Q Ahead	0.37	0.30	0.13	0.11	541	0.38	0.33	0.15	0.12	541

Notes: The four numbers in each cell show the fraction of relative MSFEs less than 1 in the first out-of-sample period (column label 1st), in the second out-of-sample period (column label 2nd), in both the first and second periods (column label 1&2), and the product of the first and the second (column label 1×2). Results are pooled for all predictors; the inflation results are the pooled results for the GDP deflator and CPI inflation; the output results are the pooled results for IP growth and real GDP growth.

TABLE 6
SUMMARY OF GRANGER CAUSALITY AND QLR TEST STATISTICS

A. Summarized by Predictor Category

	Inflation					Output				
Predictor Category	GC	QLR	G&Q	G×Q	N	GC	QLR	G&Q	G×Q	N
Asset Prices	0.35	0.78	0.29	0.28	420	0.35	0.71	0.28	0.25	420
Activity	0.63	0.78	0.48	0.49	106	0.56	0.46	0.28	0.26	134
G&C Prices	0.32	0.79	0.25	0.26	216	0.56	0.77	0.47	0.43	188
Money	0.52	0.68	0.38	0.35	114	0.40	0.53	0.30	0.21	114
All	0.40	0.77	0.32	0.31	856	0.44	0.66	0.32	0.29	856

B. Summarized by Country

	Inflation					Output				
Country	GC	QLR	G&Q	G×Q	N	GC	QLR	G&Q	G×Q	N
Canada	0.52	0.74	0.40	0.38	124	0.50	0.68	0.33	0.34	124
France	0.41	0.84	0.35	0.34	108	0.52	0.68	0.38	0.35	108
Germany	0.43	0.70	0.32	0.30	118	0.35	0.67	0.25	0.23	118
Italy	0.33	0.83	0.25	0.27	126	0.35	0.62	0.25	0.22	126
Japan	0.34	0.85	0.29	0.29	122	0.41	0.66	0.31	0.27	122
United Kingdom	0.33	0.54	0.21	0.18	112	0.36	0.58	0.24	0.21	112
United States	0.4	0.82	0.38	0.37	146	0.56	0.73	0.47	0.41	146
All	0.40	0.77	0.32	0.31	856	0.44	0.66	0.32	0.29	856

Notes: The Granger causality and QLR test statistics are heteroskedasticity-robust and were computed for a one-quarter ahead bivariate in-sample (full-sample) regression ($h = 1$ in equation (3)). The five numbers in each cell are the fraction of bivariate models with significant (5%) GC statistics (column label GC), significant (5%) QLR statistics (column label QLR), significant GC and QLR statistics (column label G&Q), the product of the first and second (column label G × Q), and the number of models in each cell. The models making up each cell are the pooled results using CPI, the GDP deflator, real GDP, and IP.

most of the inflation equations. Over all categories of predictors, 40 percent reject Granger noncausality for inflation, and 44 percent reject Granger noncausality for output growth. The term spread has limited predictive content for inflation, based on the Granger causality statistic: it enters the GDP inflation equation significantly (at the 5-percent level) for only France and Italy, and it enters the CPI inflation equation signifi-cantly for only France, Italy, and the United States. This finding is consistent with Bernanke and Mishkin (1992), Kozicki (1997), and Estrella and Mishkin (1997), and suggests that the predictive content of the term spread for inflation found elsewhere in the literature is a consequence of omitting lagged values of inflation.

Second, the QLR statistic detects widespread instability in these relations. The results in table 6 indicate that, among forecasting equations involving asset prices, the QLR statistic rejects the null hypothesis of stability (at the 5-percent level) in 78 percent of the inflation forecasting relations and in 71 percent of the output forecasting relations. This further suggests that the instability revealed by the analysis of the relative

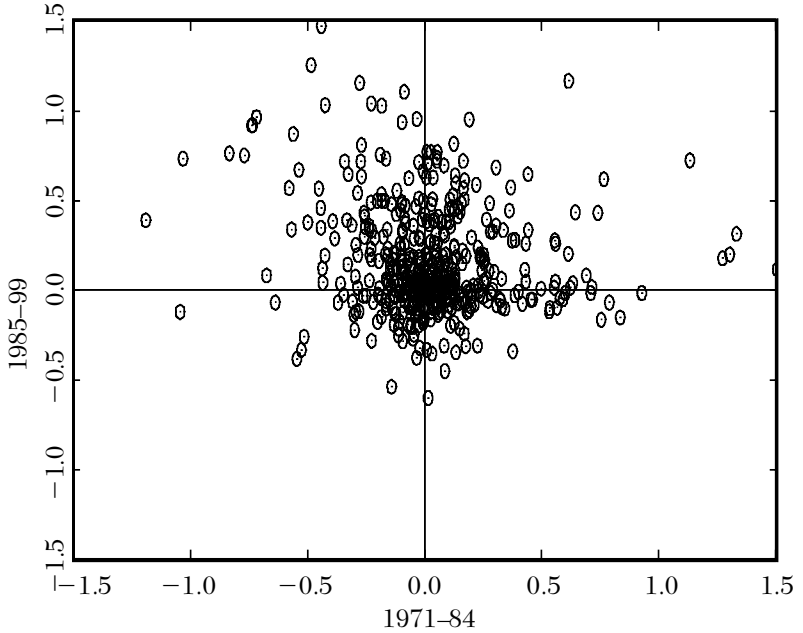

Figure 3. Scatterplot of Pseudo Out-of-Sample Log Relative Mean Squared Forecast Errors: 4-Quarter Ahead Forecasts, Predictors with Significant Granger Causality Statistics

MSFEs in the two subsamples is not a statistical artifact but rather is a consequence of unstable population relations.

Third, a statistically significant Granger causality statistic conveys little if any information about whether the forecasting relation is stable. This can be seen in several ways. For example, the frequent rejection of Granger noncausality contrasts with the findings of section 6.3: table 5 (last panel) reports that only 9 percent of all predictors improve upon the AR benchmark forecast of inflation two quarters ahead in both the first and second period, yet table 6 reports that Granger noncausality is rejected for 40 percent of all predictors. Similarly, only 12 percent of the predictors of output growth improve upon the AR benchmark in both periods at two quarters ahead, but Granger noncausality is rejected in 44 percent of the relations. Figure 3 is a scatterplot of the log relative MSFE, restricted to predictors (asset prices and otherwise) and dependent variables (inflation and output) for which the Granger causality test rejects at the 5-percent significance level. If relations that show in-sample predictive content were stable, then the points would lie along the 45° line in the southwest quadrant, but they do not. A significant Granger causality statistic makes it no more likely that a predictor outperforms the AR in both periods.

Related evidence is reported in the third and fourth entries of each cell of table 6, which respectively contain the product of the marginal rejection probabilities of the Granger causality and QLR tests and the joint probability of both rejecting. The joint probability is in every case very close to the product of the marginal probabilities: rejection of Granger noncausality appears to be approximately unrelated to whether or not the QLR statistic rejects. These findings hold, with some variation, for all the predictor category/country/dependent variable combinations examined in table 6. The QLR statistics suggest a greater amount of insta-

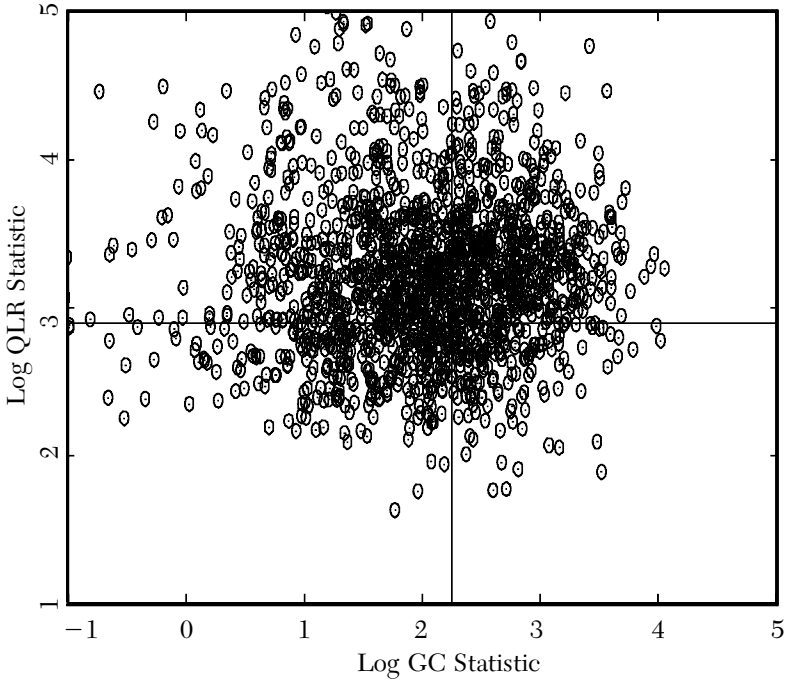

Figure 4. Scatterplot of Granger Causality and QLR Statistics, All Predictors

bility in the inflation forecasts than in the output forecasts. Across countries, inflation forecasts for Japan are most frequently unstable and output forecasts for the United Kingdom are the least frequently unstable. Among predictor category/dependent variable pairs, the greatest instability is among goods and commodity prices as predictors of inflation, and the least is among activity variables as predictors of output. In all cases, however, the QLR and Granger causality statistics appear to be approximately independently distributed. These findings are recapitulated graphically in figure 4, a scatterplot of the log of the QLR statistic vs. the log of the Granger causality *F*-statistic (all cases) that evinces little relation between the two statistics.

6.5 *Monte Carlo Check of Sampling Distribution of Relative MSFEs*

As an additional check that the instability in the relative MSFEs is not just a consequence of sampling variability, we undertook a Monte Carlo analysis designed to match an empirically plausible null model with stable but heterogeneous predictive relations. Specifically, for each of the 788 indicator/country/dependent variable combinations for which we have complete data from 1959 to 1999, the full available data set was used to estimate the VAR, $W_t = \mu + A(\mathrm{L})W_{t-1} + V_t$, where $Wt = (Y_t, X_t)$, where Y_t is the variable to be forecast and X_t is the candidate predictor, and V_t is an error vector. For the i^{th} such pair, this produced estimates of the VAR parameters θ_i, where $\theta = (\mu, A(\mathrm{L}), \Sigma_v)$, for $i = 1,\ldots, 788$. This collection of VAR coefficients constitutes the distribution of models used for the Monte Carlo experiment.

With this empirical distribution in hand, the artificial data were drawn as follows:

1. VAR parameters θ were drawn from their joint empirical distribution.
2. Artificial data on $W_t = (Y_t, X_t)$ were

TABLE 7
Differences in First and Second Period Relative MSFE
Four-Quarter Ahead Forecasts

	Median	75%–25% Range	90%–10% Range
Data	0.01	0.34	1.00
Simulations	0.02	0.12	0.27

Notes: The entries summarize the distribution of the difference between first and second period relative MSFEs for 4-quarter ahead forecasts. The first row summarizes results for the 788 country/variable pairs for which complete data from 1959–99 are available. The second row summarizes results from 5000 simulated country/variable pairs using a Monte Carlo design described in the text.

generated according to a bivariate VAR with these parameters and Gaussian errors, with the number of observations matching the full sample used in the empirical analysis.

3. Benchmark and bivariate forecasts of Y_t were made using the recursive multistep ahead forecasting method outlined in sections 2 and 4.
4. Relative MSFEs for the two periods (simulated 1971–84 and 1985–99) were computed as described in section 4, and the change between the relative MSFEs in the two periods was computed.

In this design, the distributions of the change in the relative MSFEs incorporates both the sampling variability of these statistics, conditional on the VAR parameters, and the (empirical) distribution of the estimated VAR parameters.

The results are summarized in table 7. The main finding is that the distribution of the change in the relative MSFEs is much tighter in the Monte Carlo simulation than in the actual data—approximately three times tighter at the quartiles, and four times tighter at the outer deciles. We conclude that sampling variation is insufficient to explain the dramatic shifts in predictive content observed in the data, even after accounting for heterogeneity in the predictive relations. Said differently, if the predictive relations were stable, it is quite unlikely that we would have observed as many cases as we actually did with small relative MSFEs in one period and large relative MSFEs in the other period.

6.6 *Trivariate Models*

In addition to the bivariate models, we considered forecasts based on trivariate models of the form (5). The trivariate models for inflation included lags of inflation, the IP gap, and the candidate predictor. The trivariate models for output growth included lags of output growth, the term spread, and the candidate predictor.

Relative MSFEs are given for all indicators/countries/dependent variables/horizons in the results supplement. The main conclusions drawn from the bivariate models also hold for the trivariate models. In some countries and some time periods, some indicators perform better than the bivariate model. For example, in Canada it would have been desirable to use the unemployment rate in addition to the IP gap for forecasting CPI inflation in the second period (but not the first); in Germany it would have been desirable to use M2 growth in addition to the IP gap in the first period (but not the second).

There are, however, no clear systematic patterns of improvement when candidate indicators are added to the bivariate model. Rather, the main pattern is that the trivariate

relative MSFEs show subsample instability similar to those of the bivariate relative MSFEs. This instability is, presumably, in part driven by the instability of the bivariate relation that the trivariate relation extends. For example, all the trivariate models of output growth perform poorly in the United States in the second period, reflecting the poor performance of the term spread over this period.

7. *Results for Combination Forecasts*

This section examines the possibility that combining the forecasts based on the individual indicators can improve their performance. The standard logic of combination forecasts is that, by pooling forecasts based on different data, the combined forecast uses more information and thus should be more efficient than any individual forecast. Empirical research on combination forecasts has established that simple combinations, such as the average or median of a panel of forecasts, frequently outperform the constituent individual forecasts; see the review in R. T. Clemen (1989) and the introductions to combination forecasts in Diebold (1998) and Paul Newbold and David Harvey (2002). The theory of optimal linear combination forecasts (J. M. Bates and Granger 1969; Granger and Ramu Ramanathan 1984) suggests that combination forecasts should be weighted averages of the individual forecasts, where the optimal weights correspond to the population regression coefficients in a regression of the true future value on the various forecasts. One of the intriguing empirical findings in the literature on combination forecasts, however, is that theoretically "optimal" combination forecasts often do not perform as well as simple means or medians.

The combination forecasts considered here are the trimmed mean of a set of forecasts, where the lowest and highest forecasts were trimmed to mitigate the influence of occasional outliers; results for combination forecasts based on the median are similar and are given in the Results Supplement. The relative MSFEs of these combination forecasts are summarized in table 8, and the results are striking. First consider the results for asset-price based forecasts of inflation (the first three rows of parts A and B). The trimmed mean of all the individual forecasts of CPI inflation outperforms the benchmark AR in most of the country/horizon/period combinations, and the relative RMSFE never exceeds 1.09. When all forecasts (all predictor groups) are combined, the combination CPI inflation forecast improves upon the AR forecast in 41 of 42 cases, and in the remaining case the relative RMSFE is 1.00. The overall combination GDP inflation forecasts improve upon the benchmark AR in 35 of the 39 country/horizon cases, and its worst RMSFE is 1.04.

Inspection of the results for different groups of indicators reveals that these improvements are realized across the board. For Canada, Germany, the United Kingdom, and the United States, the greatest improvements tend to obtain using the combination inflation forecasts based solely on the activity indicators, while for France, Italy, and Japan the gains are typically greatest if all available forecasts are used. In several cases, the combination forecasts have relative MSFEs under 0.80, so that these forecasts provide substantial improvements over the AR benchmark.

The results for combination forecasts of output growth are given in parts C and D of table 8. The combination forecasts usually improve upon the AR benchmark, sometimes by a substantial amount. Interestingly, the combination forecasts based only on asset prices often have a smaller MSFE than those based on all predictors. It seems that, for forecasting output growth, adding predictors beyond asset prices does not reliably improve upon the combination forecasts based on asset prices. Even though the individual forecasts based on asset prices are unstable, com-

TABLE 8
PSEUDO OUT-OF-SAMPLE FORECAST ACCURACY OF COMBINATION FORECASTS

A. Consumer Price Index

	Canada		France		Germany		Italy		Japan		U.K.		U.S.	
	71–84	85–99	71–84	85–99	71–84	85–99	71–84	85–99	71–84	85–99	71–84	85–99	71–84	85–99
Asset Prices														
2Q Horizon	0.93	0.99	0.93	0.89	1.01	1.02	1.09	0.90	1.09	0.90	0.99	0.90	0.86	0.93
4Q Horizon	0.85	0.96	0.92	0.88	1.01	1.01	1.02	0.89	1.06	0.91	0.84	0.86	0.89	0.94
8Q Horizon	0.80	0.94	0.88	0.93	0.98	1.00	0.94	0.95	0.91	0.81	0.89	0.74	0.84	0.85
Activity														
2Q Horizon	0.97	0.91		1.00	0.91	1.00	0.94	0.91	0.99	0.94	0.97	0.81	0.75	0.89
4Q Horizon	0.96	0.77		0.95	0.83	0.97	0.93	0.90	0.99	1.07	0.86	0.74	0.72	0.82
8Q Horizon	0.90	0.59		1.31	0.68	0.88	1.04	0.93	1.07	0.97	0.69	0.67	0.72	0.69
All														
2Q Horizon	0.93	0.95	0.94	0.90	0.97	1.00	0.99	0.78	0.92	0.82	0.96	0.85	0.83	0.89
4Q Horizon	0.90	0.89	0.93	0.86	0.92	0.99	0.93	0.80	0.91	0.78	0.84	0.78	0.84	0.88
8Q Horizon	0.87	0.83	0.93	0.88	0.88	0.94	0.93	0.85	0.91	0.65	0.81	0.66	0.81	0.78

B. GDP Deflator

	Canada		France		Germany		Italy		Japan		U.K.		U.S.	
	71–84	85–99	71–84	85–99	71–84	85–99	71–84	85–99	71–84	85–99	71–84	85–99	71–84	85–99
Asset Prices														
2Q Horizon	1.04	1.02		0.93	0.97	0.97	1.00	0.98	1.02	1.07	1.01	1.04	1.10	0.94
4Q Horizon	0.98	0.97		0.89	0.98	1.01	1.00	0.92	0.98	1.09	1.07	0.96	0.97	0.90
8Q Horizon	0.88	0.92		0.81	0.96	0.98	0.83	0.91	0.90	1.02	0.97	0.84	0.83	0.86
Activity														
2Q Horizon	0.92	0.92		1.08	0.82	0.90	0.89	1.03	1.06	0.94	0.93	1.05	0.94	0.95
4Q Horizon	0.92	0.86		1.21	0.76	0.86	1.01	0.94	1.05	0.77	0.95	0.94	0.88	0.89
8Q Horizon	1.01	0.75		1.22	0.80	0.78	1.09	0.94	1.30	0.78	0.67	1.00	0.76	0.83
All														
2Q Horizon	1.00	0.96		0.92	0.88	0.95	0.95	0.96	0.99	0.99	0.95	0.99	1.03	0.94
4Q Horizon	0.94	0.91		0.83	0.80	0.95	0.94	0.89	0.91	0.93	1.00	0.89	0.94	0.87
8Q Horizon	1.04	0.86		0.81	0.71	0.92	0.82	0.89	0.95	0.85	0.86	0.69	0.82	0.77

bined they perform well across the different horizons and countries. Notably, in the United States the relative mean squared forecast error for 8-quarter ahead forecasts of industrial production growth based on the combined asset price forecast is 0.55 in the first period and 0.90 in the second period.

Results for combination forecasts based on the trivariate models are presented in the results supplement. The trivariate forecasts typically improve upon the benchmark AR forecasts, however the improvements are not as reliable, nor are they as large, as for the bivariate forecasts. We

TABLE 8 *(cont.)*

C. Real GDP

	Canada		France		Germany		Italy		Japan		U.K.		U.S.	
	71–84	85–99	71–84	85–99	71–84	85–99	71–84	85–99	71–84	85–99	71–84	85–99	71–84	85–99
Asset Prices														
2Q Horizon	0.92	0.75	0.98		0.78	0.98	0.93	0.92	0.92	0.94	1.04	0.95	0.84	0.96
4Q Horizon	0.89	0.76	1.01		0.71	1.04	0.83	0.83	1.04	0.90	0.83	0.98	0.76	1.01
8Q Horizon	0.82	0.75	0.96		0.82	0.97	0.78	0.66	1.19	0.94	0.79	1.06	0.63	1.00
All														
2Q Horizon	0.96	0.83	0.99		0.86	0.97	0.95	0.87	0.91	0.94	1.01	0.93	0.86	0.97
4Q Horizon	0.94	0.87	1.04		0.80	1.02	0.84	0.82	1.01	0.90	0.87	0.97	0.82	1.01
8Q Horizon	0.90	0.87	0.98		0.90	0.98	0.78	0.70	1.03	0.91	0.89	1.04	0.76	0.99

D. Industrial Production

	Canada		France		Germany		Italy		Japan		U.K.		U.S.	
	71–84	85–99	71–84	85–99	71–84	85–99	71–84	85–99	71–84	85–99	71–84	85–99	71–84	85–99
Asset Prices														
2Q Horizon	0.88	0.93	0.80	0.93	0.79	0.89	0.91	0.94	0.90	0.96	0.98	0.92	0.87	0.93
4Q Horizon	0.83	0.84	0.81	0.89	0.72	0.90	0.89	0.83	0.94	0.91	0.85	0.91	0.66	0.95
8Q Horizon	0.80	0.75	0.85	0.86	0.82	0.82	0.82	0.78	1.07	0.79	0.76	1.00	0.55	0.90
All														
2Q Horizon	0.93	0.94	0.83	0.94	0.80	0.94	0.91	0.92	0.86	0.94	0.96	0.90	0.86	0.93
4Q Horizon	0.87	0.94	0.81	0.94	0.78	0.98	0.88	0.83	0.87	0.89	0.83	0.91	0.74	0.96
8Q Horizon	0.86	0.87	0.85	0.88	0.88	0.90	0.84	0.79	0.94	0.82	0.81	0.99	0.70	0.93

Notes: Entries are the relative mean square forecast errors for the combined forecasts constructed from the variables in the categories listed in the first column, relative to the AR benchmark.

interpret this as arising because the trivariate models all have a predictor in common (the IP gap for inflation, the term spread for output). This induces common instabilities across the trivariate models, which in turn reduces the apparent ability of the combination forecast to "average out" the idiosyncratic instability in the individual forecasts.

8. *Discussion and Conclusions*

This review of the literature and our empirical analysis lead us to four main conclusions.

1. *Some asset prices have been useful predictors of inflation and/or output growth in some countries in some time periods.* For example, the term spread was a useful predictor of output growth in the United States and Germany prior to the mid-1980s. The empirical analysis in section 6, like those in the literature, occasionally found large forecast improvements using an asset price. This said, no single asset price is a reliable predictor of output growth across countries over multiple decades. The term spread perhaps comes closest to achieving this goal, but its good performance in some periods and coun-

tries is offset by poor performance in other periods and/or countries. As for inflation forecasts, after controlling for lagged inflation, individual asset prices provide improvements that are sometimes modest but rarely large, relative to the AR benchmark. For the United States in the first period, the term spread helped to predict inflation in our pseudo out-of-sample forecasting exercise, but this was not the case in other countries or in the second period in the United States. Still, even if the improvements are small, in many cases our study (like previous ones) finds these improvements to be statistically significant. Said differently, when asset prices improve forecasts of inflation or output growth, these improvements often appear to be real in the sense that they are unlikely to have arisen just from the sampling variability of the relative MSFE under the null hypothesis that the population regression coefficients on the asset price are zero.

2. *There is considerable instability in bivariate and trivariate predictive relations involving asset prices and other predictors.* In our pseudo out-of-sample forecast comparison, we found that whether a predictor forecasts better than an autoregression in the first out-of-sample period is essentially unrelated to whether it will do so in the second period. This finding of instability in predictive relations is confirmed by widespread rejections of the null hypothesis of constant coefficients by the (in-sample) QLR statistic.

This empirical finding of instability is consistent with our reading of the literature on asset prices as predictors of output and inflation, in which an initial series of papers identifies what appears to be a potent predictive relation that is subsequently found to break down in the same country, not to be present in other countries, or both. On the one hand, this finding of instability is surprising, for the logic behind using asset prices for forecasting includes some cornerstone ideas of macroeconomics: the Fisher hypothesis, the idea that stock prices reflect expected future earnings, and the notion that temporarily high interest rates lead to an economic slowdown. On the other hand, it makes sense that the predictive power of asset prices could depend on the nature of the shocks hitting the economy and the degree of development of financial institutions, which differ across countries and over time. Indeed, several of the papers reviewed in section 3 underscore the situational dependence of the predictive content of asset prices; Cook (1981) and Duca (1999) provide detailed institutional interpretations of the predictive power of specific asset prices, and Smets and Tsatsaronis (1997) (among others) emphasize that different combinations of shocks and policies can lead to different degrees of predictive performance for asset prices. These considerations suggest that asset prices that forecast well in one country or in one period might not do so in another. Of course, this interpretation of these results is not very useful if these indicators are to be used prospectively for forecasting: according to this argument one must know the nature of future macroeconomic shocks and institutional developments that would make a particular candidate indicator stand out. It is one thing to understand ex post why a particular predictive relation broke down; it is quite another to know whether it will ex ante.

Looked at more broadly, the instability present in forecasts based on asset prices is consistent with other evidence of instability in the economy. The U.S. productivity slowdown of the mid-1970s and its revival in the late 1990s represent structural shifts. Recent research on monetary policy regimes provides formal empirical support for the conventional wisdom that there have been substantial changes in the way the Federal Reserve Bank has conducted monetary policy over the past forty years (Bernanke and Ilian Mihov 1998; Cogley and Sargent 2001, 2002; Richard Clarida, Jordi Galí, and Mark Gertler 2000; Sims and Tao Zha 2002). Europe has seen sweeping institutional changes in monetary policy and trade integration over the past forty years. Other research on forecasting

using low-dimensional models emphasizes the widespread nature of parameter instability (Stock and Watson 1996; Michael Clements and David Hendry 1999). In addition, there has been a reduction in the volatility of many macroeconomic variables, including output growth and inflation in the United States (and, it appears, other countries) that has persisted since the mid-1980s (C. J. Kim and Charles Nelson 1999, and Margaret McConnell and Gabriel Perez-Quiros 2000; for recent reviews, see Olivier Blanchard and John Simon 2001, and Stock and Watson 2002). Work on this moderation in volatility has identified a number of possible explanations: changes in methods of inventory management; shifting sectoral composition of the economy; changes in the financial sector that reduce the cyclical sensitivity of production of durables and residential housing; changes in the size and nature of the shocks to the economy; and changes in monetary policy. For example, a successful shift of monetary policy to an inflation targeting regime, in which future deviations from the target were unexpected, would have the effect of making previously potent predictive relations no longer useful, although such a shift generally would not eliminate the predictability of output fluctuations. In principal, any of these shifts could result in changes to the reduced-form forecasting relations examined in this article.

These observations suggest a number of important directions for future research. In a practical sense, the previous paragraph provides too many potential explanations for these shifting relations, and it seems important to obtain a better economic understanding of the nature and sources of these changes on a case-by-case basis. From the perspective of forecasting methods, this evidence of sporadic predictive content poses the challenge of developing methods that provide reliable forecasts in the face of time-varying relations. Conventional time-varying parameter models and forecasts based on trun-cated samples or windows are a natural approach, but these methods do not seem to lead to useful forecasts, at least for low-dimensional models fit to U.S. postwar data (Stock and Watson 1996, 1999c); these models reduce bias at the expense of increasing variance, so MSFEs frequently increase. Other possibilities include intercept corrections and overdifferencing (Clements and Hendry 1999). Alternatively, the evidence presented here does not rule out the possibility that some fixed-parameter nonlinear models might produce stable forecasts, and that linear models simply represent state-dependent local approximations; empirically, however, nonlinear models can produce even wilder pseudo out-of-sample forecasts than linear models (Stock and Watson 1999c). In any event, the challenge of producing reliable forecasts using low-dimensional models based on asset prices remains open.

3. *In-sample Granger causality tests provide a poor guide to forecast performance.* The distribution of relative MSFEs in the two periods for the subset of predictive relations with a statistically significant Granger causality statistic is similar to the distribution of relative MSFEs for all the predictive relations; in this sense, rejection of Granger noncausality does not provide useful information about the predictive value of the forecasting relation. Similarly, the Granger causality statistic is essentially uncorrelated with the QLR statistic, so the Granger causality statistic provides no information about whether the predictive relation is stable. In short, we find that rejection of Granger noncausality is, to a first approximation, uninformative about whether the relation will be useful for forecasting.

The conclusion that testing for Granger noncausality is uninformative for assessing predictive content is not as counterintuitive as it initially might seem. One model consistent with this finding is that the relation has non-zero coefficients at some point but those coefficients change at an unknown date. Clark and McCracken (2002) provide theoretical and Monte Carlo evidence that the single-break model is capable of generating

results like ours. Their theoretical results, combined with our empirical findings, suggest that future investigations into predictive content need to use statistics, such as break tests and pseudo out-of-sample forecasting tests, that can detect instability in predictive relations. Additional theoretical work on which set of tests has best size and power is in order.

4. *Simple combination forecasts reliably improve upon the AR benchmark and forecasts based on individual predictors*. Forecasts of output growth constructed as the median or trimmed mean of the forecasts made using individual asset prices regularly exhibit smaller pseudo out-of-sample MSFEs than the autoregressive benchmark and typically perform nearly as well as or better than the combination forecast based on all predictors. In this sense, asset prices, taken together, have predictive content for output growth. Moreover, the combination forecasts are stable even though the individual predictive relations are unstable. The value of asset prices for forecasting inflation is less clear: although combination forecasts of inflation using real activity indicators improve upon the autoregressive benchmark, further forecasting gains from incorporating asset prices are not universal but instead arise only for selected countries and time periods. In most cases, the combination forecasts improve upon the Atkeson-Ohanian (2001) seasonal random walk forecast, sometimes by a substantial margin.

The finding that averaging individually unreliable forecasts produces a reliable combination forecast is not readily explained by the standard theory of forecast combination, which relies on information pooling in a stationary environment. Rather, it appears that the instability is sufficiently idiosyncratic across series for the median forecast to "average out" the instability across the individual forecasting relations. Fully articulated statistical or economic models consistent with this observation could help to produce combination forecasts with even lower MSFEs. Developing such models remains a task for future research.

Appendix

The sample dates and sampling frequencies are listed for each variable, by country, in table 9. The original sources for the series are given in the web results supplement.

Some of the variables exhibited large outliers due to strikes, redefinitions, etc. As discussed in the text, these observations were replaced by the median of the three observations on either side of the observation(s) in question. The observations with interpolated values are: France, IP, March 1963 and May–June 1968; United Kingdom, PPI, January 1974; Germany, M3, July 1990; France, M3, December 1969 and January 1978; Germany, unemployment rate, January 1978, January 1984, and January 1992; Germany, M1, January 1991; Germany, real and nominal GDP, 1991:1; Italy, real and nominal GDP, 1970:1; and Japan, real GDP, 1979:1.

The gap variables were constructed as the deviation of the series from a one-sided version of the Hodrick-Prescott (1981) (HP) filter. The one-sided HP gap estimate is constructed as the Kalman filter estimate of ε_t from the model $y_t = \tau_t + \varepsilon_t$ and $\Delta^2\tau_t = \eta_t$, where y_t is the observed series, τ_t is its unobserved trend component, and ε_t and η_t are mutually uncorrelated white noise sequences with relative variance $\text{var}(\varepsilon_t)/\text{var}(\eta_t)$. As discussed in Andrew Harvey and Albert Jaeger (1993) and Robert King and Sergio Rebelo (1993), the HP filter is the minimum mean square error linear two-sided trend extraction filter for this model. Because our focus is on forecasting, we use the optimal one-sided analogue of this filter, so that future values of y_t (which would not be available for real time forecasting) are not used in the detrending operation. The filter is implemented with $\text{var}(\varepsilon_t)/\text{var}(\eta_t) = .00675$, which corresponds to the usual value of the HP smoothing parameter ($\lambda = 1600$).

References

Akerlof, George A.; William T. Dickens and George L. Perry. 2000. "Near-Rational Wage and Price Setting and the Optimal Rates of Inflation and Unemployment," *Brookings Papers Econ. Act.* 2000:1, pp.1–44.

Andrews, Donald W.K. 1993. "Tests for Parameter Instability and Structural Change with Unknown Change Point," *Econometrica* 614, pp. 821–56.

Ang, Andrew; Monika Piazzesi and Min Wei. 2003. "What Does the Yield Curve Tell Us About GDP Growth?" manuscript, Columbia Business School.

Atkeson, Andrew and Lee E. Ohanian. 2001. "Are Phillips Curves Useful for Forecasting Inflation?" *Fed. Reserve Bank Minneapolis Quart. Rev.* 25:1, pp. 2–11.

Atta-Mensah, Joseph and Greg Tkacz. 2001. "Predicting Recessions and Booms Using Financial Variables," *Can. Bus. Econ.* 8:3, pp. 30–36.

Barr, David G. and John Y. Campbell. 1997. "Inflation, Real Interest Rates, and the Bond Market: A Study

TABLE 9
Data Sample Periods

Series	Canada	France	Germany	Italy	Japan	U.K.	U.S.
Interest rate: overnight	75:1–99:12,m	64:1–99:3,m	60:1–99:12,m	71:1–99:12,m	59:1–99:12,m	72:1–99:12,m	59:1–99:12,m
short-term gov't bills	59:1–99:12,m	70:1–99:12,m	75:7–99:12,m	74:5–99:6,m		64:1–99:12,m	59:1–99:12,m
short-term gov't bonds				70:2–99:6,m		66:1–99:12,m	59:1–99:12,m
medium-term gov't bonds				59:1–99:12,m			59:1–99:12,m
long-term gov't bonds	59:1–99:12,m	59:1–99:12,m	59:1–99:12,m	59:1–99:12,m	66:10–99:9,m	59:1–99:12,m	59:1–99:12,m
Exchange rate	73:1–99:12,m	73:1–99:12,m	73:1–99:12,m	73:1–99:12,m	73:1–99:12,m	73:1–99:12,m	73:1–99:12,m
Stock prices	59:1–99:12,m	59:1–99:12,m	59:1–99:12	59:1–99:12,m	59:1–99:12,m	59:1–99:3,m	59:1–99:12,m
Dividend price index	70:1–97:1,q	70:1–97:1,q	70:1–97:1,q	70:1–97:1,q	70:1–97:1,q	70:1–97:1,q	59:1–96:4,q
Housing price index	70:1–98:4,q				70:1–98:4,q	70:1–98:4,q	70:1–98:4,q
Gold price	59:1–99:12,m	59:1–99:12,m	59:1–99:12,m	59:1–99:12,m	59:1–99:12,m	59:1–99:12,m	59:1–99:12,m
Silver price	59:1–99:12,m	59:1–99:12,m	59:1–99:12,m	59:1–99:12,m	59:1–99:12,m	59:1–99:12,m	59:1–99:12,m
Real GDP	59:1–99:4,q	70:1–99:4,q	60:1–99:4,q	60:1–99:4,q	59:1–99:4,q	59:1–99:4,q	59:1–99:4,q
Nominal GDP	59:1–99:4,q	65:1–99:4,q	60:1–99:4,q	60:1–99:4,q	59:1–99:4,q	59:1–99:4,q	59:1–99:4,q
Industrial production	59:1–99:12,m	59:1–99:12,m	59:1–99:12,m	59:1–98:12,m	59:1–99:12,m	59:1–99:12,m	59:1–99:12,m
Capacity utilization rate	62:1–99:4,q	76:1–99:4,q	70:1–99:4,q	62:1–98:4,m	68:1–99:12		59:1–99:12,m
Employment	59:1–99:12,m	70:1–99:4,q	60:1–99:12,m	60:1–90:4,q	59:1–99:1,q	60:1–99:4,q	59:1–99:12,m
Unemployment rate	59:1–99:12,m	74:4–99:1,q	62:1–99:12,m	60:1–99:4,q	59:1–99:1,q	60:1–99:4,q	59:1–99:12,m
CPI	59:1–99:12,m	59:1–99:12,m	59:1–99:12,m	59:1–99:12,m	59:1–99:12,m	59:1–99:12,m	59:1–99:12,m
PPI	59:1–99:12,m		59:1–99:11,m	81:1–99:11,m	59:1–99:12,m	59:1–99:12,m	59:1–99:12,m
Earnings	59:1–99:12,m	60:1–99:4,m	62:1–99:12,m		59:1–99:12,m	63:1–99:12,m	59:1–99:12,m
Oil price	59:1–99:12,m	59:1–99:12,m	59:1–99:12,m	59:1–99:12,m	59:1–99:12,m	59:1–99:12,m	59:1–99:12,m
Commodity price index	59:1–99:12,m	59:1–99:12,m	59:1–99:12,m	59:1–99:12,m	59:1–99:12,m	59:1–99:12,m	59:1–99:12,m
Money supply: m0					70:1–99:12,m		59:1–99:12,m
Money supply: m1	59:1–99:12,m	77:1–98:4,m	60:1–98:4,m	62:1–98:4,m	63:1–99:12,m		59:1–99:12,m
Money supply: m2	59:1–99:12,m		60:1–98:4,m	74:1–98:4,m	67:1–99:12,m		59:1–99:12,m
Money supply: m3	59:1–99:12,m	60:1–98:4,m	69:1–98:12,m	62:1–98:4,m	71:12–99:12,m		59:1–99:12,m

Notes: The table entries show the sample periods of each data series for each country. Blank cells indicate missing data; m means the data series is monthly, and q means quarterly.

of U.K. Nominal and Index-Linked Government Bond Prices," *J. Monet. Econ.* 39:3, pp. 361–83.

Barro, Robert J. 1990. "The Stock Market and Investment," *Rev. Finan. Stud.* 3:1, pp. 115–31.

Bates, J.M. and Clive W.J Granger. 1969. "The Combination of Forecasts," *Operations Res. Quart.* 20, pp. 319–25.

Bernanke, Ben S. 1983. "Nonmonetary Effects of the Financial Crisis in the Propagation of the Great Depression," *Amer. Econ. Rev.* 73:3, pp. 257–76.

———. 1990. "On the Predictive Power of Interest Rates and Interest Rate Spreads," *New England Econ. Rev.* Nov.:Dec. pp. 51–68.

Bernanke, Ben S. and Alan S. Blinder. 1992. "The Federal Funds Rate and the Channels of Monetary Transmission," *Amer. Econ. Rev.* 82:4, pp. 901–21.

Bernanke, Ben S. and Ilian Mihov. 1998. "Measuring Monetary Policy," *Quart. J. Econ.* 113:3, pp. 869–902.

Bernanke, Ben S. and Frederic S. Mishkin. 1992. "The Predictive Power of Interest Rate Spreads: Evidence from Six Industrialized Countries," manuscript, Princeton U.

Bernard, Henri and Stefan Gerlach. 1998. "Does the Term Structure Predict Recessions? The International Evidence," *Int. J. Finance Econ.* 3:3, pp. 195–215.

Blanchard, Olivier. 1993. "Consumption and the Recession of 1990–1991," *Amer. Econ. Rev.* 83:2, pp. 270–74.

Blanchard, Olivier and John Simon. 2001. "The Long and Large Decline in U.S. Output Volatility,"

Brookings Papers Econ. Act. 2001:1, pp. 135–64.
Bonser-Neal, Catherine and Timothy R. Morley. 1997. "Does the Yield Spread Predict Real Economic Activity? A Multicountry Analysis," *Fed. Reserve Bank Kansas City Econ. Rev.* 82:3, pp. 37–53.
Brainard, William C. and George L. Perry. 2000. "Making Policy in a Changing World," in *Economic Events, Ideas, and Policies: The 1960s and After*. G.L. Perry and J. Tobin, eds. Brookings Institution: Washington, D.C.
Campbell, John Y. 1999. "Asset Prices, Consumption and the Business Cycle," in *The Handbook of Macroeconomics*, Vol. 1. John B. Taylor and Michael Woodford, eds. Amsterdam: Elsevier, pp. 1231–303.
Campbell, John Y.; Martin Lettau; Burton G. Malkiel and Yexaio Xu. 2001. "Have Individual Stocks Become More Volatile? An Empirical Exploration of Idiosyncratic Risk," *J. Finance* 56:1, pp. 1–43.
Canova, Fabio and Gianni De Nicolo. 2000. "Stock Return, Term Structure, Real Activity, and Inflation: An International Perspective," *Macroecon. Dynam.* 4:3, pp. 343–72.
Cecchetti, Stephen G.; Rita S. Chu and Charles Steindel. 2000. "The Unreliability of Inflation Indicators," *Fed. Reserve Bank New York Current Issues Econ. Finance* 6:4, pp. 1–6.
Chen, Nai-Fu. 1991. "Financial Investment Opportunities and the Macroeconomy," *J. Finance* 46:2, pp. 529–54.
Clarida, Richard; Jordi Galí and Mark Gertler. 2000. "Monetary Policy Rules and Macroeconomic Stability: Evidence and Some Theory," *Quart. J. Econ.* 115:1, pp. 147–80.
Clark, Todd E. and Michael W. McCracken. 2001. "Tests of Equal Forecast Accuracy and Encompassing for Nested Models," *J. Econometrics* 105, pp. 85–100.
———. 2002. "Forecast-Based Model Selection in the Presence of Structural Breaks," Fed. Reserve Bank Kansas City discuss. paper RWP 02-05.
Clemen, R.T. 1989. "Combining Forecasts: A Review and Annotated Bibliography," *Int. J. Forecasting* 5, pp. 559–83.
Clements, Michael P. and David F. Hendry. 1999. *Forecasting Non-stationary Economic Time Series*. Cambridge, MA: MIT Press.
Cogley, Timothy and Thomas J. Sargent. 2001. "Evolving Post-World War II U.S. Inflation Dynamics," *NBER Macroeconomics Annual*, pp. 331–72.
———. 2002. "Drifts and Volatilities: Monetary Policies and Outcomes in the Post WWII U.S.," manuscript, Stanford U.
Cook, Timothy. 1981. "Determinants of the Spread between Treasury Bill Rates and Private Sector Money Market Rates," *J. Econ. Bus.* 33, pp. 177–87.
Croushore, Dean and Tom Stark. 2003. "A Real-Time Data Set for Macroeconomists: Does the Data Vintage Matter?" forthcoming, *Rev. Econ. Statist.*.
Davis, E. Philip and Gabriel Fagan. 1997. "Are Financial Spreads Useful Indicators of Future Inflation and Output Growth in EU Countries?" *J. Applied Econometrics* 12, pp. 701–14.
Davis, E. Philip and S.G.B. Henry. 1994. "The Use of Financial Spreads as Indicator Variables: Evidence for the United Kingdom and Germany," *IMF Staff Papers* 41, pp. 517–25.
Diebold, Francis X. 1998. *Elements of Forecasting*. Cincinnati: Southwestern Publishing.
Diebold, Francis X. and Robert S. Mariano. 1995. "Comparing Predictive Accuracy," *J. Bus. Econ. Statist.* 13, pp. 253–63.
Dotsey, Michael. 1998. "The Predictive Content of the Interest Rate Term Spread for Future Economic Growth," *Fed. Reserve Bank Richmond Econ. Quart.* 84:3, pp. 31–51.
Duca, John V. 1999. "What Credit Market Indicators Tell Us," *Fed. Reserve Bank Dallas Econ. Finan. Rev.* 1999:Q3, pp. 2–13.
Dueker, Michael J. 1997. "Strengthening the Case for the Yield Curve as a Predictor of U.S. Recessions," *Fed. Reserve Bank St. Louis Rev.* 79:2, pp. 41–50.
Emery, Kenneth M. 1996. "The Information Content of the Paper–Bill Spread," *J. Econ. Business* 48, pp. 1–10.
Estrella, Arturo and Frederic S. Mishkin. 1997. "The Predictive Power of the Term Structure of Interest Rates in Europe and the United States: Implications for the European Central Bank," *Europ. Econ. Rev.* 41, pp. 1375–401.
———. 1998. "Predicting U.S. Recessions: Financial Variables as Leading Indicators," *Rev. Econ. Statistics* 80, pp. 45–61.
Estrella, Arturo and Gikas Hardouvelis. 1991. "The Term Structure as a Predictor of Real Economic Activity," *J. Finance* 46:2, pp. 555–76.
Estrella, Arturo; Anthony P. Rodrigues and Sebastian Schich. 2003. "How Stable is the Predictive Power of the Yield Curve? Evidence from Germany and the United States," *Rev. Econ. Statistics* 85:3, forthcoming.
Fama, Eugene F. 1981. "Stock Returns, Real Activity, Inflation and Money," *Amer. Econ. Rev.* 71, pp. 545-65.
———. 1990. "Term-Structure Forecasts of Interest Rates, Inflation, and Real Returns," *J. Monet. Econ.* 25:1, pp. 59–76.
Feldstein, Martin and James H. Stock. 1994. "The Use of a Monetary Aggregate to Target Nominal GDP," in N.G. Mankiw (ed.), *Monetary Policy*. Chicago: U. Chicago Press for the NBER, pp. 7–70.
Fischer, Stanley and Robert C. Merton. 1984. "Macroeconomics and Finance: The Role of the Stock Market," *Carnegie-Rochester Conference Series on Public Policy* 21, pp. 57–108.
Frankel, Jeffrey A. 1982. "A Technique for Extracting a Measure of Expected Inflation from the Interest Rate Term Structure," *Rev. of Econ. and Statistics* 64:1, pp. 135–142.
Frankel, Jeffrey A. and Cara S. Lown. 1994. "An Indicator of Future Inflation Extracted from the Steepness of the Interest Rate Yield Curve Along Its Entire Length," *Quart. J. Econ.* 59, pp. 517–30.
Friedman, Benjamin M. and Kenneth N. Kuttner. 1992. "Money, Income, Prices and Interest Rates," *Amer. Econ. Rev.* 82:June, pp. 472–92.
———. 1993a. "Why Does the Paper-Bill Spread Predict Real Economic Activity?" in *Business Cycles, In-dicators, and Forecasting*. James H. Stock and

Mark W. Watson, eds. Chicago: U. Chicago Press.
———. 1993b. "Economic Activity and the Short-term Credit Markets: An Analysis of Prices and Quantities," *Brookings Papers Econ. Act.* 1993:2, pp. 193–266.
———. 1998. "Indicator Properties of the Paper-Bill Spread: Lessons from Recent Experience," *Rev. Econ. Statistics* 80, pp. 34–44.
Galbraith, John W. and Greg Tkacz. 2000. "Testing for Asymmetry in the Link Between the Yield Spread and Output in the G-7 Countries," *J. Int. Money Finance* 19, pp. 657–72.
Gerlach, Stefan. 1997. "The information Content of the Term Structure: Evidence for Germany," *Empirical Econ.* 22:2, pp. 161-80.
Gertler, Mark and Cara S. Lown. 2000. "The Information in the High Yield Bond Spread for the Business Cycle: Evidence and Some Implications," NBER work. paper 7549.
Goodhart, Charles and Boris Hofmann. 2000a. "Do Asset Prices Help to Predict Consumer Price Inflation," *Manchester School Supplement* 68, pp. 122–40.
———. 2000b. "Financial Variables and the Conduct of Monetary Policy," Sveriges Riksbank work. paper 112.
Gordon, Robert J. 1982. "Inflation, Flexible Exchange Rates, and the Natural Rate of Unemployment," in M.N. Baily (ed.), *Workers, Jobs, and Inflation*, Brookings Institution.
———. 1997. "The Time-Varying NAIRU and its Implications for Economic Policy," *J. Econ. Perspectives* 11:1, pp. 11–32.
———. 1998. "Foundations of the Goldilocks Economy: Supply Shocks and the Time-Varying NAIRU," *Brookings Papers Econ. Activity* 1998:2, pp. 297–333.
Granger, Clive W.J. and Ramu Ramanathan. 1984. "Improved Methods for Combining Forecasts," *J. Forecasting* 3, pp. 197–204.
Guo, Hui. 2002. "Stock Market Returns, Volatility, and Future Output," *Fed. Reserve Bank St. Louis Rev.* 84:5, pp. 75–85.
Hafer, Rik W. and Ali M. Kutan. 1992. "Money, Interest Rates and Output: Another Look," Southern Illinois U. Edwardsville econ. dept. work. paper 92-0303.
Hamilton, James D. and Dong Heon Kim. 2002. "A Re-Examination of the Predictability of Economic Activity using the Yield Spread," *J. Money, Credit and Banking* 34, pp. 340–360.
Harvey, Andrew C. and Albert Jaeger. 1993. "Detrending, Stylized Facts and the Business Cycle," *J. Applied Econometrics* 8:3, pp. 231–48.
Harvey, Campbell R. 1988. "The Real Term Structure and Consumption Growth," *J. Finan. Econ.* 22, pp. 305–333.
———. 1989. "Forecasts of Economic Growth from the Bond and Stock Markets," *Finan. Analysts J.* 45:5, pp. 38–45.
———. 1991. "The Term Structure and World Economic Growth," *J. Fixed Income*, June, pp. 7–19.
———. 1993. "The Term Structure Forecasts Economic Growth," *Finan. Analysts J.* 49:3, pp. 6–8.
Haubrich, Joseph G. and Ann M. Dombrosky. 1996. "Predicting Real Growth Using the Yield Curve," *Fed. Reserve Bank Cleveland Econ. Rev.* 32:1, pp. 26–34.
Hodrick, Robert and Edward Prescott. 1981. "Post-war U.S. Business Cycles: An Empirical Investigation," work. paper, Carnegie-Mellon U.; printed in J. *Money, Credit and Banking* 29 (1997), pp. 1-16.
Hu, Zuliu. 1993. "The Yield Curve and Real Activity," *IMF Staff Papers* 40, pp. 781–806.
Jaditz, Ted; Leigh A. Riddick and Chera L. Sayers. 1998. "Multivariate Nonlinear Forecasting: Using Financial Information to Forecast the Real Sector," *Macroeconomic Dynamics* 2, pp. 369–382.
Jorion, Philippe and Frederic S. Mishkin. 1991. "A Multi-Country Comparison of Term Structure Forecasts at Long Horizons," *J. Finan. Econ.* 29, pp. 59–80.
Katz, Lawrence F. and Alan B. Krueger. 1999. "The High-Pressure U.S. Labor Market of the 1990s," *Brookings Papers on Econ. Activity* 1999:1, pp. 1–87.
Kim, C.-J. and Charles R. Nelson. 1999. "Has the U.S. Economy Become More Stable? A Bayesian Approach Based on a Markov-Switching Model of the Business Cycle," *The Rev. of Econ. and Statistics* 81, pp. 608–616.
King, Robert G. and Sergio T. Rebelo. 1993. "Low Frequency Filtering and Real Business Cycles," *J. Econ. Dynamics Control* 17, pp. 207–31.
Kozicki, Sharon. 1997. "Predicting Real Growth and Inflation with the Yield Spread," *Fed. Reserve Bank Kansas City Econ. Rev.* 82, pp. 39–57.
Lahiri, Kajal and Jiazhou G. Wang. 1996. "Interest Rate Spreads as Predictors of Business Cycles," in *Handbook of Statistics: Methods in Finance Vol 14*. G.S. Maddala and C.R. Rao eds. Amsterdam: Elsevier.
Larocque, Guy. 1977. "Analyse d'une methode de desaisonnalisation: le programme X11 du US Bureau of Census, version trimestrielle," *Annale de l'INSEE*, 28.
Laubach, Thomas. 2001. "Measuring the NAIRU: Evidence from Seven Economies," *Rev. Econ. Statistics* 83, pp. 218–31.
Laurent, Robert D. 1988. "An Interest Rate-Based Indicator of Monetary Policy," *Fed. Reserve Bank Chicago Econ. Perspectives* 12:1, pp. 3–14.
———. 1989. "Testing the Spread," *Fed. Reserve Bank Chicago Econ. Perspectives* 13, pp. 22–34.
Lettau, Martin and Sydney Ludvigson. 2000. "Understanding Trend and Cycle in Asset Values," manuscript, research dept., Fed. Reserve Bank New York.
———. 2001. "Consumption, Aggregate Wealth and Expected Stock Returns," *J. Finance* 56, pp. 815–49.
Ludvigson, Sydney and Charles Steindel. 1999. "How Important is the Stock Market Effect on Consumption," *Fed. Reserve Bank New York Econ. Policy Rev.* 5:2, pp. 29–51.
Marcellino, Massimiliano; James H. Stock and Mark W. Watson. 2000. "Macroeconomic Forecasting in the Euro Area: Country Specific vs. Area-Wide Information," forthcoming, *Europ. Econ. Rev.*
McConnell, Margaret M. and Gabriel Perez-Quiros. 2000. "Output Fluctuations in the United States: What has Changed Since the Early 1980's," *Amer. Econ. Rev.* 90:5, pp. 1464–76.

McCracken, Michael W. 1999. "Asymptotics for Out-of-Sample Tests of Causality," manuscript, U. Missouri-Columbia.

McCracken, Michael W. and Kenneth D. West. 2002. "Inference about Predictive Ability," in *Companion to Economic Forecasting*. Michael P. Clements and David F. Hendry, eds. Oxford: Blackwell, pp. 299–321.

Mishkin, Frederic S. 1990a. "What Does the Term Structure Tell Us About Future Inflation?" *J. Monet. Econ.* 25, pp. 77–95.

———. 1990b. "The Information in the Longer-Maturity Term Structure About Future In-flation," *Quart. J. Econ.* 55, pp. 815–28.

———. 1991. "A Multi-Country Study of the Information in the Term Structure About Future Inflation," *J. Int. Money Finance* 19, pp. 2–22.

Mitchell, Wesley C. and Arthur F. Burns. 1938. *Statistical Indicators of Cyclical Revivals*, NBER Bulletin 69, NY. Reprinted in *Business Cycle Indicators*. G.H. Moore, ed. 1961. Princeton: Princeton U. Press.

Newbold, Paul and David I. Harvey. 2002. "Forecast Combination and Encompassing," in *A Companion to Economic Forecasting*. M.P. Clements and D.F. Hendry, eds. Oxford: Blackwell Press.

Pesaran, M. Hashem and Spyros Skouras. 2002. "Decision-Based Methods for Forecast Evaluation," in *A Companion to Economic Forecasting*. M. P. Clements and D. F. Hendry, eds. Oxford: Blackwell Press.

Pivetta, Frederic and Ricardo Reis. 2002. "The Persistence of Inflation in the United States," manuscript, Harvard U.

Plosser, Charles I. and K. Geert Rouwenhorst. 1994. "International Term Structures and Real Economic Growth," *J. Monet. Econ.* 33, pp. 133–56.

Poon, Ser-Huang and Clive W.J. Granger. 2003. "Forecasting Volatility in Financial Markets: A Review," *J. Econ. Lit.* 41:2, pp. 478–539.

Quandt, Richard E. 1960. "Tests of the Hypothesis that a Linear Regression Obeys Two Separate Regimes," *J. Amer. Statist. Assoc.* 55, pp. 324–30.

Sims, Christopher A. 1980. "A Comparison of Interwar and Postwar Cycles: Monetarism Reconsidered," *Amer. Econ. Rev.* 70:May, pp. 250–57.

Sims, Christopher A. and Tao Zha. 2002. "Macroeconomic Switching," manuscript, Princeton U.

Smets, Frank and Kostas Tsatsaronis. 1997. "Why Does the Yield Curve Predict Economic Activity? Dissecting the Evidence for Germany and the United States?" BIS work. paper 49.

Staiger, Douglas; James H. Stock and Mark W. Watson. 1997a. "The NAIRU, Unemployment, and Monetary Policy," *J. Econ. Perspect.* 11:Winter, pp. 33–51.

———. 1997b. "How Precise Are Estimates of the Natural Rate of Unemployment?" in *Reducing Inflation: Motivation and Strategy*. C. Romer and D. Romer, eds. U. Chicago Press for NBER, pp. 195–242.

———. 2001. "Prices, Wages, and U.S. NAIRU in the 1990s," in *The Roaring Nineties*. A. Krueger and R. Solow, eds. NY: Russell Sage Foundation/ Century Fund, pp. 3–60

Stock, James H. 1998. "Discussion of Gordon's 'Foundations of the Goldilocks Economy'," *Brookings Papers Econ. Act.* 1998:2, pp. 334–41.

Stock, James H. and Mark W. Watson. 1989. "New Indexes of Coincident and Leading Economic Indicators," in *NBER Macroeconomics Annual 1989*. O.J. Blanchard and S. Fischer, eds., pp. 352–94.

———. 1996. "Evidence on Structural Instability in Macroeconomic Time Series Relations," *J. Business Econ. Statistics* 14, pp. 11–29.

———. 1998. "Median Unbiased Estimation of Coefficient Variance in a Time Varying Parameter Model," *J. Amer. Statistical Assoc.* 93, pp. 349–58.

———. 1999a. "Business Cycle Fluctuations in U.S. Macroeconomic Time Series," in *Handbook of Macroeconomics, Vol. 1*. J.B. Taylor and M. Woodford, eds. pp. 3–64.

———. 1999b. "Forecasting Inflation," *J. Monet. Econ.* 44, pp. 293–335.

———. 1999c. "A Comparison of Linear and Nonlinear Univariate Models for Forecasting Macroeconomic Time Series," in *Cointegration, Causality and Forecasting: A Festschrift for Clive W.J. Granger*. R. Engle and H. White eds. Oxford: Oxford U. Press, pp. 1–44.

———. 2001. "Forecasting Output and Inflation: The Role of Asset Prices," NBER work. paper 8180.

———. 2002. "Has the Business Cycle Changed and Why?" *NBER Macroeconomics Annual* 2002. Cambridge: MIT Press.

———. 2003a. *Introduction to Econometrics*. Boston: Addison Wesley Longman.

———. 2003b. "How Did Leading Indicator Forecasts Do During the 2001 Recession?" *Fed. Reserve Bank Richmond Econ. Quart.* 89:3, forthcoming.

Swanson, Norman R. and Halbert White. 1995. "A Model Selection Approach to Assessing the Information in the Term Structure Using Linear Models and Artificial Neural Networks," *J. Business Econ. Statistics* 13, pp. 265–80.

———. 1997. "A Model Selection Approach to Real-Time Macroeconomic Forecasting Using Linear Models and Artificial Neural Networks," *Rev. Econ. Statistics* 79, pp. 540–50.

Thoma, Mark A. and Jo Anna Gray. 1994. "On Leading Indicators: Is There a Leading Contender?" manuscript, U. Oregon.

Tkacz, Greg. 2001. "Neural Network Forecasting of Canadian GDP Growth," *Int. J. Forecasting* 17, pp. 57–69.

Wallis, Kenneth F. 1974. "Seasonal Adjustment and Relations Between Variables," *J. Amer. Statistical Assoc.* 69, pp. 18–31.

West, Kenneth D. 1996. "Asymptotic Inference about Predictive Ability," *Econometrica* 64, pp. 1067–84.