NBER WORKING PAPER SERIES

DO MACRO VARIABLES, ASSET MARKETS OR SURVEYS FORECAST INFLATION BETTER?

Andrew Ang
Geert Bekaert
Min Wei

Working Paper 11538
http://www.nber.org/papers/w11538

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
August 2005

We thank Jean Boivin for kindly providing data. Andrew Ang acknowledges support from the National Science Foundation. The opinions expressed in this paper do not necessarily reflect those of the federal Reserve Board or the Federal Reserve system. The views expressed herein are those of the author(s) and do not necessarily reflect the views of the National Bureau of Economic Research.

Do Macro Variables, Asset Markets or Surveys Forecast Inflation Better?
Andrew Ang, Geert Bekaert and Min Wei
NBER Working Paper No. 11538
August 2005
JEL No. E31, E37, E43, E44

ABSTRACT

Surveys do! We examine the forecasting power of four alternative methods of forecasting U.S. inflation out-of-sample: time series ARIMA models; regressions using real activity measures motivated from the Phillips curve; term structure models that include linear, non-linear, and arbitrage-free specifications; and survey-based measures. We also investigate several optimal methods of combining forecasts. Our results show that surveys outperform the other forecasting methods and that the term structure specifications perform relatively poorly. We find little evidence that combining forecasts using means or medians, or using optimal weights with prior information produces superior forecasts to survey information alone. When combining forecasts, the data consistently places the highest weights on survey information.

Andrew Ang
Columbia Business School
805 Uris Hall
3022 Broadway
New York, NY 10027
and NBER
aa610@columbia.edu

Geert Bekaert
Columbia Business School
805 Uris Hall
3022 Broadway
New York, NY 10027
and NBER
gb241@columbia.edu

Min Wei
Federal Reserve Board of Governors
Division of Monetary Affairs
Washington, DC 20551
min.wei@frb.gov

1 Introduction

Obtaining reliable and accurate forecasts of future inflation is crucial for policymakers conducting monetary and fiscal policy; for investors hedging the risk of nominal assets; for firms making investment decisions and setting prices; and for labor and management negotiating wage contracts. Consequently, it is no surprise that a considerable academic literature evaluates different inflation forecasts and forecasting methods. In particular, econometricians use four main methods to forecast inflation. First, forecasts from benchmark time-series models of the ARIMA variety are used. Second, we can forecast inflation using information filtered through theoretical models of economic causation, such as regression models motivated from the Phillips curve that use real activity measures. Third, forecasts can be made using term structure models, where the econometrician uses information filtered indirectly through asset prices. Finally, we can forecast inflation using information filtered directly through agents by employing survey-based measures.

In this article, we comprehensively compare and contrast the ability of these four methods to forecast inflation out of sample. Our approach makes four main contributions to the literature. First, our analysis is the first to comprehensively compare time-series forecasts, forecasts based on the Phillips curve, forecasts from the yield curve, and survey data (from three different surveys). The previous literature has concentrated on only one or two of these different forecasting methodologies. For example, Stockton and Glassman (1987) show that pure time-series models out-perform more sophisticated macro models, but do not consider term structure models or surveys. Fama and Gibbons (1984) compare term structure forecasts with the Livingston survey, but they do not consider forecasts from macro factors. Whereas Grant and Thomas (1999), Thomas (1999) and Mehra (2002) show that surveys out-perform simple time-series benchmarks for forecasting inflation, all these studies do not compare the performance of survey measures with forecasts from Phillips curve models or term structure models.

The lack of a study comparing these four methods of inflation forecasting implies that there is no well-accepted set of findings regarding the superiority of a particular forecasting method. The most comprehensive study to date, Stock and Watson (1999) finds that Phillips curve-based forecasts produce the most accurate out-of-sample forecasts of U.S. inflation compared with other macro series and asset prices, using data up to 1996. However, Stock and Watson only briefly compare the Phillips-curve forecasts to the Michigan survey and to simple regressions using term structure information. Stock and Watson do not consider no-arbitrage term structure models, non-linear forecasting models, or combined forecasts from all four forecasting

methods. Recent work also casts doubts on the robustness of the Stock-Watson findings. In particular, Atkeson and Ohanian (2001), Sims (2002), and Cecchetti, Chu and Steindel (2000) show that the accuracy of Phillips curve-based forecasts depends crucially on the sample period and Clark and McCracken (2003) address the issue of how instability in the output gap coefficients of the Phillips curve affects forecasting power. To assess the stability of the inflation forecasts across different samples, we specifically consider the out-of-sample forecasts over post-1985 and post-1995 U.S. data.

Our second contribution is to evaluate inflation forecasts implied by arbitrage-free asset pricing models. Previous studies employing term structure data mostly use only the term spread in simple OLS regressions and usually do not use all available term structure data (see, for example, Mishkin (1990, 1991), Jorion and Mishkin (1991), and Stock and Watson (2003)). Frankel and Lown (1994) use a simple weighted average of different term spreads, but they do not impose no-arbitrage restrictions. In contrast to these approaches, we develop forecasting models that use all available data and impose no-arbitrage restrictions. Our no-arbitrage term structure models also incorporate inflation as a state variable because inflation is an integral component of nominal yields. Importantly, the no-arbitrage framework allows us to extract forecasts of inflation jointly from inflation and asset prices by taking into account time-varying risk premia.

No-arbitrage constraints are reasonable in a world where hedge funds and investment banks routinely eliminate arbitrage opportunities from fixed income prices. Imposing theoretical no-arbitrage restrictions may also lead to more efficient estimation. Just as Ang, Piazzesi and Wei (2004) show that no-arbitrage models produce superior forecasts of GDP growth, no-arbitrage restrictions may also produce more accurate forecasts of inflation. In addition, this is the first article to investigate non-linear, no-arbitrage models of inflation. We investigate both an empirical regime-switching model incorporating term structure information and a no-arbitrage, non-linear term structure model following Ang and Bekaert (2004) with inflation as a state variable.

Our third contribution is that we thoroughly investigate combined forecasts. Stock and Watson (2002a, 2003), among others, show that the use of aggregate indices of many macro series measuring real activity produces better forecasts of inflation than individual macro series. To investigate this further, we also include the (Phillips curve-based) index of real activity constructed by Bernanke, Boivin and Eliasz (2005) from 65 macroeconomic series. In addition, several authors (see, for example, Stock and Watson (1999) and Wright (2004)) advocate combining several alternative models to forecast inflation. We investigate five different methods of combining forecasts: simple means or medians, OLS based combinations, and Bayesian

estimators with equal or unit weight priors.

Finally, our main focus is forecasting inflation rates. While a long-standing debate in macroeconomics focuses on whether inflation rates are stationary or non-stationary. Economically, non-stationary inflation is hard to interpret and the recent working paper version of Bai and Ng (2004) very strongly rejects the null of a unit root in inflation data. Nevertheless, we explicitly contrast the predictive power of some non-stationary models with stationary models and consider whether forecasting inflation changes alters the relative forecasting ability of different models.

Our major empirical results can be summarized as follows. The first major result is that survey forecasts outperform the other three methods in forecasting inflation. That the median Livingston and SPF survey forecasts do well is perhaps not surprising, because presumably many of the best analysts use time-series and Phillips Curve models. But, even participants in the Michigan survey who are consumers, not professionals, produce accurate out-of-sample forecasts in line with the professionals in the Livingston and SPF surveys. We also find that the best survey forecasts are the unadjusted survey median forecast themselves; adjustments to take into account both linear and non-linear bias yield worse out-of-sample forecasts than raw survey forecasts.

Second, term structure information does not generally lead to better forecasts and often leads to inferior forecasts than models using only aggregate activity measures. Whereas this confirms the results in Stock and Watson (1999), our investigation of term structure models is much more comprehensive. The relatively poor forecasting performance of term structure models extends to simple regression specifications, iterated long-horizon VAR forecasts, no-arbitrage affine models, and non-linear no-arbitrage models. These results point to the conclusion that while inflation is very important for explaining the dynamics of the term structure (see, for example, Ang and Bekaert (2004)), yield curve information is less important for forecasting future inflation.

Our third major finding is that combining forecasts does not generally lead to better out-of-sample forecasting performance than single forecasting models. In particular, simple averaging, like using the mean or median of a number of forecasts, does not necessarily improve the forecast performance, whereas linear combinations of forecasts with optimal weights computed based on past performance and prior information generate the biggest gains. We find that even the use of the Bernanke, Boivin and Eliasz (2005) forward-looking aggregate measure of real activity does not perform well relative to simpler Phillips curve models and survey forecasts. The strong success of the surveys in forecasting inflation out-of-sample extends to

surveys dominating other models in forecast combining methods. The data consistently place the highest weights on the survey forecasts and little weight on other forecasting methods.

The rest of this paper is organized as follows. Section 2 describes the data set. In Section 3, we describe the time-series models, predictive macro regressions, term structure models, and forecasts from survey data, and detail the forecasting methodology. Section 4 contains the empirical out-of-sample results. We examine the robustness of our results to a non-stationary inflation specification in Section 5. Finally, Section 6 concludes.

2 Data

2.1 Inflation

We consider four different measures of inflation. The first three are consumer price index (CPI) measures, including CPI-U for All Urban Consumers, All Items (*PUNEW*), CPI for All Urban Consumers, All Items Less Shelter (*PUXHS*) and CPI for All Urban Consumers, All Items Less Food and Energy (*PUXX*), which is also called core CPI. The fourth measure is the Personal Consumption Expenditure deflator (*PCE*). All measures are seasonally adjusted and obtained from the Bureau of Labor Statistics website. The sample period is 1952:Q2 to 2002:Q4 for PUNEW and PUXHS, 1958:Q2 to 2002:Q4 for PUXX, and 1960:Q2 to 2002:Q4 for PCE.

We define the quarterly inflation rate, π_t, from $t-1$ to t as:

$$\pi_t = \log\left(\frac{P_t}{P_{t-1}}\right), \tag{1}$$

where P_t is the level of one of the four inflation indices at time t. We use the terms "inflation" and "inflation rate" interchangeably as defined in equation (1). We take one quarter to be our base unit.

In our main analysis, we assume that the inflation rate is stationary. Economically, it is hard to interpret non-stationary inflation and difficult to generate non-stationary inflation in standard rational models. In particular, non-stationary inflation can only arise in standard overlapping contract models of inflation by the presence of non-stationary excess demand (see, for example, comments by Fuhrer and Moore (1995)). The empirical work on inflation forecasting has either assumed that inflation is stationary (see Bryan and Cecchetti (1993)), or that inflation has a unit root (see, for example, Quah and Vahey (1995) and Stock and Watson (1999)). In finance, there is also a tradition of assuming that inflation is non-stationary (see, for instance, Nelson and Schwert (1977)). While standard unit root tests sometimes fail to reject the null of a unit root

for inflation, more powerful tests like those developed by Bai and Ng (2004) strongly reject the null that the inflation rate is a unit root process and conclude that it is stationary. Even using standard unit root tests, Stock and Watson (1999) reject the null of a unit root for inflation pre-1982. Nevertheless, we also consider the robustness of our results to considering non-stationary inflation in Section 5.

Table 1 reports summary statistics for all four measures of inflation for the full sample in Panel A, and the post-1985 sample and the post-1995 sample in Panels B and C, respectively. The inflation rate data are annual horizon but at a quarterly frequency. We report the fourth quarterly autocorrelation, which corresponds to the annual horizon. Table 1 shows that all four inflation measures are lower and more stable during the last two decades, in common with many other macroeconomic series, including output (see Kim and Nelson (1999), McConnell and Perez-Quiros (2000), and Stock and Watson (2002b)). Core CPI (PUXX) has the lowest volatility of all the inflation measures. PUXX volatility ranges from 2.56% per annum over the full sample to only 0.24% per annum post-1996, dramatically showcasing the fall in food and energy shocks in the later part of the sample. As is well known, PCE inflation is, on average, lower than CPI inflation, particularly in the later sample periods, because it uses chain weighting in contrast to the other CPI measures which use a fixed basket (see Stock and Watson (1999)).

Inflation is somewhat persistent (0.79% for PUNEW over the full sample), but its persistence decreases over time, as can be seen from the lower autocorrelation coefficients for the PUNEW and the PUXHS measures after 1986, and for all measures after 1995. The correlations of the four measures of inflation with each other are all over 75% over the full sample. The comovement can be clearly seen in the top panel of Figure 1. Inflation is lower prior to 1969 and after 1983, but reaches a high of around 14% during the oil crisis of 1973–1983. PUXX tracks both PUNEW and PUXHS closely, except during the 1973–1975 period, where it is about 2% lower than the other two measures, and after 1985, where it appears to be more stable than the other two measures. During the periods when inflation is decelerating, such as in 1955–1956, 1987–1988, 1998–2000 and most recently 2002–2003, PUNEW declines more gradually than PUXHS, suggesting that housing prices are less volatile than the prices of other consumption goods during these periods.

2.2 Real Activity Measures

We consider six individual series for real activity along with one composite real activity factor. We compute GDP growth (*GDPG*) using the seasonally adjusted data on real GDP in billions of chained 2000 dollars. The unemployment rate (*UNEMP*) is also seasonally adjusted and

computed for the civilian labor force aged 16 years and over. Both real GDP and the unemployment rate are from the Federal Reserve Economic Data (FRED) database. We compute the GDP gap either as the detrended log real GDP by removing a quadratic trend as in Gali and Gertler (1999), which we term *GAP1*, or by using the Hodrick-Prescott (1997) filter (with the standard smoothness parameter of 1,600), which we term *GAP2*. At time t, both measures are constructed using only current and past GDP values, so the filters are run recursively. We also use the labor income share (*LSHR*), defined as the ratio of nominal compensation to total nominal output in the U.S. nonfarm business sector. For forward-looking indicators, we take the Stock-Watson (1989) Experimental Leading Index (*XLI*) and the Alternative Nonfinancial Experimental Leading Index-2 (*XLI-2*).

Motivated from studies like Stock and Watson (2002a), who show that aggregating the information from many factors has good forecasting power, we also use a single factor aggregating the information from 65 individual series. This single real activity series, which we term *FAC*, aggregates real output and income, employment and hours, consumption, housing starts and sales, real inventories, and average hourly earnings, and is constructed by Bernanke, Boivin and Eliasz (2005). The sample period for all the real activity measures is from 1952:Q2 to 2001:Q4, except the Bernanke-Boivin-Eliasz real activity factor, which spans 1959:Q1 to 2001:Q3. We use the composite real activity factor at the end of each quarter for forecasting inflation over the next year.

The real activity measures have the disadvantage that they may be using information that is not actually available at the time of the forecast, either through data revisions, or because of full sample estimation in the case of the Bernanke-Boivin-Eliasz measure. This biases the forecasts from Phillips curve models to be better than what could be actually forecasted using a real-time data set. Orphanides and van Norden (2001) show that real-time economic activity measures provide much less accurate forecasts of inflation than revised economic series. Bernanke and Boivin (2003) also find that forecasts from factors extracted from many macro economic series are clearly inferior to forecasts based on macro factors using revised series. Thus, our forecast errors using real activity measures are biased downwards.

2.3 Term Structure Data

The term structure variables are zero-coupon yields for the maturities of 1, 4, 8, 12, 16, and 20 quarters from CRSP spanning 1952:Q2 to 2001:Q4. The 1-quarter rate is from the CRSP Fama risk-free rate file, while all other bond yields are from the CRSP Fama-Bliss discount bond file. All yields are continuously compounded and expressed at a quarterly frequency. We define the

short rate (*RATE*) to be the 1-quarter yield and define the term spread (*SPD*) to be the difference between the 20-quarter yield and the short rate. Some of our term structure models also use 4-quarter and 12-quarter yields for estimation.

2.4 Surveys

We examine three inflation expectation surveys: the Livingston survey, the Survey of Professional Forecasters (SPF), and the Michigan survey.[1] There are some reporting lags between the time the surveys are taken and the public dissemination of their results. For the Livingston and the SPF surveys, there is a lag of up to four and three weeks, respectively, between the time the survey is conducted and their publication. For the Michigan survey, the lag is up to three weeks. This reporting delay does not mean that using survey information entails the use of forward-looking information not in the current information set. Indeed, because of the time taken to conduct the surveys, survey forecasts must use less up-to-date information than either macroeconomic or term structure forecasts. Together with the slight data advantages present in revised, fitted macro data, we are in fact biasing the results against survey forecasts. The information contained in survey data can be collected in real time with sufficient resources. However, the reporting lag for the Livingston, SPF, and Michigan surveys does mean that forecasts for the next year from these surveys are only available with a small delay of, at most, four weeks already into the year.

The Livingston survey is conducted twice a year, in June and in December, and polls economists from industry, government, and academia. The Livingston survey records participants' forecasts of non-seasonally-adjusted CPI levels 6 and 12 months in the future and is usually conducted in the middle of the month. Unlike the Livingston survey, participants in the SPF and the Michigan survey forecast inflation rates. Participants in the SPF are drawn primarily from business, and forecast changes in the quarterly average of seasonally-adjusted CPI-U levels. The SPF is conducted in the middle of every quarter and the sample period for the SPF median forecasts is from 1981:Q3 to 2002:Q4. In contrast to the Livingston survey and SPF, the Michigan survey is conducted monthly and asks households (consumers), rather

[1] We obtain data for the Livingston survey and SPF data from the Philadelphia Fed website (http://www.phil.frb.org/econ/liv and http://www.phil.frb.org/econ/spf, respectively). We take the Michigan survey data from the St. Louis Federal Reserve FRED database (http://research.stlouisfed.org/fred2/series/MICH/). Median Michigan survey data is also available from the University of Michigan's website (http://www.sca.isr.umich.edu/main.php. However, there are small discrepancies between the two sources before September 1996. We choose to use data from the FRED because it is consistent with the values reported in Curtin (1996).

than professionals, to estimate expected price changes over the next twelve months. We use the median Michigan survey forecast from 1978:Q1 to 2002:Q4.

The Livingston survey is the only survey available for our full sample. In the top panel of Figure 1, which graphs the full sample of inflation data, we also include the unadjusted median Livingston forecasts. We plot the survey forecast lagged one year, so that in December 1990, we plot inflation from December 1989 to December 1990 together with the survey forecasts at December 1989. The Livingston forecasts broadly track the movements of inflation, but there are several large movements that the Livingston survey fails to track, for example the pickup in inflation in 1956–1959, 1967–1971, 1972–1975, and 1978–1981. In the bottom panel of Figure 1, we graph all three survey forecasts of future one-year inflation together with the annual PUNEW inflation, where the survey forecasts are lagged one year for direct comparison. After 1981, all survey forecasts move reasonably closely together and track inflation movements relatively well. Nevertheless, there are still some notable failures, like the slowdowns in inflation in the early 1980s and in 1996.

3 Forecasting Models and Methodology

In this section, we describe the forecasting models and describe our statistical tests. In all our out-of-sample forecasting exercises, we forecast future annual inflation. Hence, for all our models, we compute annual inflation forecasts of:

$$\mathrm{E}_t(\pi_{t+4,4}) = \mathrm{E}_t\left(\sum_{i=1}^{4}\pi_{t+i}\right), \tag{2}$$

where π_t is the quarterly inflation rate defined in equation (1) and $\pi_{t+4,4}$ is annual inflation from t to $t+4$:

$$\pi_{t+4,4} = \pi_{t+1} + \pi_{t+2} + \pi_{t+3} + \pi_{t+4} \tag{3}$$

In Sections 3.1 to 3.4, we describe the forecasting models. Table 2 contains a full nomenclature of these 38 forecasting models. Section 3.1 focuses on time-series models of inflation, which serve as our benchmark forecasts; Section 3.2 summarizes our OLS regression models using real activity macro variables; Section 3.3 describes the term structure models incorporating inflation data; and finally, Section 3.4 describes our survey forecasts. In Section 3.5, we define the out-of-sample periods and list the criteria that we use to assess the performance of out-of-sample forecasts. Finally, Section 3.6 describes our methodology to combine model forecasts.

For all models except OLS regressions, we compute implied long-horizon forecasts from single-period (quarterly) models, as Marcellino, Stock and Watson (2004) show that iterated forecasts are superior to direct forecasts from horizon-specific models. For OLS regressions, we compute the forecasts directly from the long-horizon regression estimates since the OLS models do not specify a unique, underlying single-period model.

3.1 Time-Series Models

ARIMA Models

If inflation is stationary, the Wold theorem suggests that a parsimonious ARMA(p, q) model may perform well in forecasting. We consider two ARMA(p, q) models: an ARMA(1,1) model and a pure autoregressive model with p lags, AR(p). The optimal lag length for the AR model is selected for each forecasting period using the Schwartz criterion (BIC) on the in-sample data. The motivation for the ARMA(1,1) model derives from a long tradition in rational expectations macroeconomics (see Hamilton (1985)) and finance (see Fama (1975)) that models inflation as the sum of expected inflation and noise. If expected inflation follows an AR(1) process, then the reduced-form model for inflation is given by an ARMA(1,1) model. The ARMA(1,1) model also nicely fits the slowly decaying autocorrelogram of inflation.

The specifications of the ARMA(1,1) model,

$$\pi_{t+1} = \mu + \phi\pi_t + \psi\varepsilon_t + \varepsilon_{t+1}, \tag{4}$$

and the AR(p) model,

$$\pi_{t+1} = \mu + \phi_1\pi_t + \phi_2\pi_{t-1} + \ldots + \phi_p\pi_{t-p+1} + \varepsilon_{t+1}, \tag{5}$$

are entirely standard. The ARMA(1,1) model is estimated by maximum likelihood, conditional on a zero initial residual. We compute the implied inflation level forecast over the next year expressed at a quarterly frequency. For the ARMA(1,1) model, the forecast is:

$$\mathrm{E}_t(\pi_{t+4,4}) = \frac{1}{1-\phi}\left[1 - \frac{\phi\left(1-\phi^4\right)}{(1-\phi)}\right]\mu + \frac{\phi\left(1-\phi^4\right)}{(1-\phi)}\pi_t + \frac{\left(1-\phi^4\right)\psi}{(1-\phi)}\varepsilon_t,$$

while the forecast for the AR(p) model is:

$$\mathrm{E}_t(\pi_{t+4,4}) = e_1'\left(I-\Phi\right)^{-1}\left(I - \Phi\left(I-\Phi\right)^{-1}\left(I-\Phi^4\right)\right)A + e_1'\Phi\left(I-\Phi\right)^{-1}\left(I-\Phi^4\right)X_t,$$

where e_1 is a $p \times 1$ selection vector containing a one in the first row and zeros elsewhere, and A and Φ represent the companion form representation of the AR(p) process:

$$X_{t+1} = A + \Phi X_t + U_{t+1},$$

in which

$$X_t = \begin{bmatrix} \pi_t \\ \pi_{t-1} \\ \vdots \\ \pi_{t-p+1} \end{bmatrix}, \; A = \begin{bmatrix} \mu \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \; \Phi = \begin{bmatrix} \phi_1 & \phi_2 & ... & \phi_p \\ 1 & 0 & ... & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & ... & 0 \end{bmatrix} \text{ and } U_t = \begin{bmatrix} \varepsilon_t \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Since inflation is persistent, our third ARIMA benchmark is a random walk (*RW*) forecast where $\pi_{t+1} = \pi_t + \varepsilon_{t+1}$, and $\mathrm{E}_t(\pi_{t+4,4}) = 4\pi_t$. We also comment on the forecasting results of a random walk model on annual inflation, where the forecast is given by $E_t(\pi_{t+4,4}) = \pi_{t,4}$.

A Regime-Switching Model

Evans and Wachtel (1993), Evans and Lewis (1995), and Ang and Bekaert (2004), among others, document regime-switching behavior in inflation. A regime-switching model may potentially account for non-linearities and structural changes, where, for example, a change in regime may occur from a sudden shift in inflation expectations after a supply shock.

We estimate a univariate regime-switching model in inflation, which we term *RGM*:

$$\pi_{t+1} = \mu\left(s_{t+1}\right) + \phi\left(s_{t+1}\right)\pi_t + \sigma\left(s_{t+1}\right)\varepsilon_{t+1} \tag{6}$$

The regime variable $s_t = 1, 2$ follows a Markov chain with constant transition probabilities $P = Pr(s_{t+1} = 1|s_t = 1)$ and $Q = Pr(s_{t+1} = 2|s_t = 2)$. The model can be estimated using the Bayesian filter algorithms of Hamilton (1989) and Gray (1996). We compute the implied annual horizon forecasts of inflation from equation (6), assuming that the current regime is the regime that maximizes the probability $Pr(s_t|I_t)$. This is a byproduct of the Hamilton-Gray estimation algorithm.

3.2 Regression Forecasts Based on the Phillips Curve

In standard Phillips curve models of inflation, expected inflation is linked to some measure of the output gap. There are both forward- and backward-looking Phillips curve models, but ultimately even forward-looking models link expected inflation to the current information set. According to the Phillips curve, measures of real activity should be an important part of this information set. We avoid the debate regarding regarding the actual measure of the output gap (see, for instance, Gali and Gertler (1999)) by taking an empirical approach and use a large number of real activity measures. We choose not to estimate structural models because the BIC criterion is likely to choose the empirical model best suitable for forecasting. Previous

work often finds that models with the clearest theoretical justification often have poor predictive content (see the literature summary by Stock and Watson (2003)).

The empirical specification we estimate is:

$$\pi_{t+4,4} = \alpha + \beta(L)' X_t + \varepsilon_{t+4,4} \tag{7}$$

where X_t combines π_t and one or two real activity measures. The lag length in the lag polynomial $\beta(L)$ is selected by BIC for each forecasting period and is set to be equal across all the regressors in X_t. The chosen specification tends to have two or three lags in our forecasting exercises. We list the complete set of real activity regressors in Table 2 as *PC1* to *PC10*.

In our next section, we extend the information set to include term structure information. Regression models where term structure information is included in X_t along with inflation and real activity are potentially consistent with a forward-looking Phillips curve that includes inflation and real activity measures in the information set. Such models can proxy the reduced form of a more sophisticated, forward-looking rational expectations Phillips curve model of inflation (see, for instance, Bekaert, Cho and Moreno (2005)).

3.3 Models Using Term Structure Data

We consider a variety of term structure forecasts, including augmenting the simple Phillips Curve OLS regressions with short rate and term spread variables; long-horizon VAR forecasts; a regime-switching specification; affine term structure models; and term structure models incorporating regime switches. We outline each of these specifications in turn.

Linear Non-Structural Models

We begin by augmenting the OLS Phillips Curve models in equation (7) with the short rate, RATE, and the term spread, SPD, as regressors in X_t. Specifications *TS1–TS8* add RATE to the Phillips Curve Curve specifications *PC1–PC8*. *TS9* and *TS10* only use inflation and term structure variables as predictors. *TS9* uses inflation and the lagged term spread, producing a forecasting model similar to the specification in Mishkin (1990, 1991). *TS10* adds the short rate to this specification. Finally, *TS11* adds GDP growth to the *TS10* specification.

We also consider forecasts with a VAR(1) in X_t, where X_t contains RATE, SPD, GDPG, and π_t:

$$X_{t+1} = \mu + \Phi X_t + \varepsilon_{t+1}. \tag{8}$$

Although the VAR is specified at a quarterly frequency, we compute the annual horizon forecast of inflation implied by the VAR. We denote this forecasting specification as *VAR*. As Ang,

Piazzesi and Wei (2004) and Cochrane and Piazzesi (2005) note, a VAR specification can be economically motivated from the fact that a reduced from VAR is equivalent to a Gaussian term structure model where the term structure factors are observable yields and certain assumptions on risk premia apply. Under these restrictions, a VAR coincides with a no-arbitrage term structure model only for those yields included in the VAR. However, the VAR does not impose over-identifying restrictions generated by the term structure model for yields not included as factors in the VAR.

An Empirical Non-Linear Regime-Switching Model

A large empirical literature has documented the presence of regime switches in interest rates (see, among others, Hamilton (1988), Gray (1996), and Bekaert, Hodrick and Marshall (2001)). In particular, Ang and Bekaert (2002) show that regime-switching models forecast interest rates better than linear models. Thus, capturing the regime-switching behavior in interest rates may help in forecasting potentially regime-switching dynamics of inflation.

We estimate a regime-switching VAR:

$$X_{t+1} = \mu(s_{t+1}) + \Phi X_t + \Sigma(s_{t+1})\varepsilon_{t+1}, \tag{9}$$

where X_t contains RATE, SPD and π_t. Similar to the time-series univariate regime-switching model in equation (6), we also use two regimes $s_t = 1, 2$ that follow a Markov chain with constant transition probabilities. We compute out-of-sample forecasts from equation (9) assuming that the current regime is the regime with the highest probability $Pr(s_t|I_t)$. We denote the regime-switching VAR in equation (9) as *RGMVAR*.

No-Arbitrage Term Structure Models

We estimate two no-arbitrage term structure models. Because such models have implications for the complete yield curve, it is straightforward to incorporate additional information from the yield curve into the estimation. Such additional information is absent in the empirical VAR specified in equation (8). Concretely, both no-arbitrage models have two latent variables and quarterly inflation as factors in X_t. We estimate the models by maximum likelihood, and following Chen and Scott (1993), assume that the 1- and 20-quarter yields are measured without error, and the other 4- and 12-quarter yields are measured with error. The estimated models build on Ang and Bekaert (2004), who formulate a real pricing kernel as:

$$\widehat{M}_{t+1} = \exp\left(-r_t + \lambda_t'\lambda_t - \lambda_t\varepsilon_{t+1}\right), \tag{10}$$

and model the real short rate as an affine function of the state variables. The nominal pricing kernel is defined in the standard way as $M_{t+1} = \widehat{M}_{t+1} P_t / P_{t+1}$. Bonds are priced using the recursion:

$$\exp(-n y_t^n) = E_t[M_{t+1} \exp(-(n-1) y_{t+1}^{n-1})],$$

where y_t^n is the n-quarter zero-coupon bond yield.

The first no-arbitrage model (*MDL1*) is an affine model from the class of Duffie and Kan (1996) with affine, time-varying risk premia (see Dai and Singleton (2002) and Duffee (2002)). The real price of risk vector, λ_t, is modelled as:

$$\lambda_t = \lambda_0 + \lambda_1 X_t. \tag{11}$$

where λ_0 is a 3×1 vector and λ_1 a 3×3 diagonal matrix, and the state variables follow a linear VAR:

$$X_t = \mu + \Phi X_{t-1} + \Sigma \varepsilon_{t+1}. \tag{12}$$

The second model (*MDL2*) incorporates regime switches and represents Model IV of Ang and Bekaert (2004). Ang and Bekaert show that this model has an amazing fit to the moments of yields and inflation and almost exactly matches the autocorrelogram of inflation. MDL2 replaces equation (12) with the regime-switching VAR:

$$X_t = \mu(s_{t+1}) + \Phi X_{t-1} + \Sigma(s_{t+1}) \varepsilon_{t+1}, \tag{13}$$

and also incorporates regime switches in the prices of risk, replacing equation (11) with

$$\lambda_t = \lambda_0(s_{t+1}) + \lambda_1 X_t. \tag{14}$$

In estimating *MDL1* and *MDL2*, we impose the same parameter restrictions necessary for identification as Ang and Bekaert (2004) do. For both *MDL1* and *MDL2*, we compute recursive out-of-sample forecasts of annual inflation, but the models are estimated and specified using quarterly data.

3.4 Survey Forecasts

We produce estimates of $\mathrm{E}_t(\pi_{t+4,4})$ from the Livingston survey, SPF, and the Michigan survey. We denote the actual forecasts from the SPF, Livingston and Michigan surveys as *SPF1*, *LIV1*, and *MCH1*, respectively.

Producing Forecasts from Survey Data

Participants in the Livingston survey are asked to forecast a CPI level (not an inflation rate). Given the timing of the survey, Carlson (1977) carefully studies the forecasts of individual participants in the Livingston survey and finds that the participants generally forecast inflation over the next 14 months. We follow Thomas (1999) and Mehra (2002) and derive the implied 12-month inflation forecast, assuming that inflation is expected to stay constant during the 14-month forecasting interval. That is, the raw Livingston forecasts are adjusted by a factor of 12/14.

Participants in both the SPF and the Michigan survey do not forecast log year-on-year CPI levels according to the definition of inflation in equation (1). Instead, the surveys record simple expected inflation changes, $\mathrm{E}_t(P_{t+4}/P_t - 1)$. This differs from $\mathrm{E}_t(\log P_{t+4}/P_t)$ by a Jensen's term. In addition, the SPF participants are asked to forecast changes in the quarterly average of seasonally-adjusted PUNEW (CPI-U), as opposed to end-of-quarter changes. In both the SPF and the Michigan survey, we cannot directly recover forecasts of expected log changes in CPI levels. Instead, we directly use the SPF and Michigan survey forecasts to represent forecasts of future annual inflation as defined in equation (2). We expect that the effects of the Jensen's term and the use of changes in quarterly averages in the SPF, as opposed to changes in end of quarter CPI levels, are small. In any case, the presence of the small Jensen's term biases our survey forecasts upwards and, thus, imparts a conservative upward bias to our Root Mean Squared Error (RMSE) statistics.[2]

Adjusting Surveys for Bias

Several authors, including Thomas (1999) and Mehra (2002), document that survey forecasts are biased. We take into account the survey bias by estimating α_1 and β_1 in the regressions:

$$\pi_{t+4,4} = \alpha_1 + \beta_1 f_t^S + \varepsilon_{t+4,4}, \tag{15}$$

where f_t^S is the forecast from the candidate survey S. A test of an unbiased forecasting model is $\alpha_1 = 0$ and $\beta_1 = 1$. We denote survey forecasts that are adjusted this way as *SPF2*, *LIV2*, and

[2] In the data, the correlation between log CPI changes, $\log(P_{t+4}/P_t)$ and simple inflation, $P_{t+4}/P_t - 1$ is 1.000 for all four measures of inflation across our full sample period. The correlation between end-of-quarter log CPI changes and quarterly average CPI changes is also above 0.994. The differences in log CPI changes, simple inflation, and changes in quarterly average CPI are very small, and an order of magnitude smaller than the forecast RMSEs. As an illustration, for PUNEW, the means of $\log(P_{t+4}/P_t)$, $P_{t+4}/P_t - 1$, and changes in quarterly average CPI-U are 3.83%, 3.82%, and 3.86%, respectively, while the volatilities are 2.87%, 2.86%, and 2.91%, respectively.

MCH2 for the SPF, Livingston, and Michigan surveys, respectively. The bias adjustment occurs recursively, that is, we update the regression with new data points each quarter and re-estimate the coefficients.

Table 3 provides empirical evidence regarding these biases using the full sample. For each inflation measure, the first three rows report the results from regression (15). The SPF and Livingston survey forecasts produce β_1s that are smaller than 1 for all inflation measures, and in the case of the SPF forecasts significantly so. However, the point estimates of α_1 are also positive, although mostly not significant, which implies that at low levels of inflation, the surveys under-predict future inflation and at high levels of inflation the surveys over-predict future inflation. For the Livingston measure, the turning point is rather high in the 3 to 4% annual inflation range, but for the SPF measure, the turning point is at most 2.75%, so that it mostly over-predicts future inflation. The Michigan survey produces largely unbiased forecasts because the slope coefficients are insignificantly different from one and the constants are insignificantly different from zero.

Thomas (1999) and Mehra (2002) suggest that the bias in the survey forecasts may vary across accelerating versus decelerating inflation environments, or across the business cycle. To take account of this possible asymmetry in the bias, we augment the survey forecasts in equation (15) with a dummy variable if current inflation is greater than a two-year moving average of past inflation:

$$\pi_{t+4,4} = \alpha_1 + \alpha_2 D_t + \beta_1 f_t^S + \beta_2 \left(D_t f_t^S\right) + \varepsilon_{t+4,4}. \tag{16}$$

The dummy variable D_t equals one if inflation at time t exceeds its past two-year average,

$$\pi_t - \frac{1}{8}\sum_{j=0}^{7} \pi_{t-j} > 0,$$

otherwise D_t is set equal to zero. We denote the survey forecasts that are non-linearly bias-adjusted using equation (16) as *SPF3*, *LIV3*, and *MCH3* for the SPF, Livingston, and Michigan surveys, respectively.

The bottom three rows of each panel in Table 3 report results from regression (16). There is little evidence of asymmetric bias for forecasting PUXX or PCE. When we use the SPF or the Michigan survey forecasts, for all inflation measures, there is only weak evidence of non-linearity in the coefficients. In contrast, for the Livingston survey forecasts and the PUNEW and PUXHS inflation measures, the β_2 estimates are significantly positive, implying quite different biases depending on whether inflation is rising or falling relative to a moving average. Note that overall the α_2 coefficients are negative and the β_2 coefficients are positive. Hence,

the SPF and Livingston forecasts are closer to being unbiased when inflation is rising. For the Michigan survey, the economic magnitudes of both α_2 and β_2 are large (except for PUXX) and imply very different behavior of the forecasts in rising inflation environments relative to other periods. When inflation has increased recently, the Michigan survey over- (under-) estimates future inflation at low (high) inflation levels, whereas the opposite occurs in decelerating inflation environments.

3.5 Assessing Forecasting Models

Out-of-Sample Periods

We select two starting dates for our out-of-sample forecasts, 1985:Q4 and 1995:Q4. All our out-of-sample forecasts use all the data available at time t to forecast annual future inflation from t to $t+4$. Hence, the windows used for estimation lengthen through time. All our annual forecasts are computed at a quarterly frequency, with the exception of forecasts from the Livingston survey, where forecasts are only available for the second and fourth quarter each year. The out-of-sample periods end in 2002:Q4, except for forecasts with the composite real activity factor, which end in 2001:Q3.

Measuring Forecast Accuracy

We assess forecast accuracy with the Root Mean Squared Error (RMSE) of the forecasts produced by each model and also report the ratio of RMSEs relative to a time-series ARMA(1,1) benchmark that uses only information in the past series of inflation. We show below that the ARMA(1,1) model produces the lowest RMSE among all of the ARIMA time-series models that we examine.

To compare the out-of-sample forecasting performance of the various models, we perform a forecast comparison regression, following Stock and Watson (1999):

$$\pi_{t+4,4} = \lambda f_t^{ARMA} + (1-\lambda) f_t^x + \varepsilon_{t+4,4}, \tag{17}$$

where f_t^{ARMA} is the forecast of $\pi_{t+4,4}$ from the ARMA(1,1) time-series model, f_t^x is the forecast from the candidate model x, and $\varepsilon_{t+4,4}$ is the forecast error associated with the combined forecast. The forecast error follows an MA(3) process because of the overlapping annual observations taken at a quarterly frequency. Therefore, we compute Hansen and Hodrick (1980) standard errors. If $\lambda = 0$, then forecasts from the ARMA(1,1) model add nothing to the forecasts from candidate model x, and we thus conclude that model x out-performs the

ARMA(1,1) benchmark. If $\lambda = 1$, then forecasts from model x add nothing to forecasts from the ARMA(1,1) time-series benchmark.

3.6 Combining Models

A long statistics literature has often found that forecast combinations typically provide better forecasts than individual forecasting models.[3] In particular, Stock and Watson (1999) and Wright (2004), among others, show that combined forecasts of future inflation using real activity and financial indicators are usually more accurate than individual forecasts. To examine if combining the information in different forecasts lead to gains in out-of-sample forecasting accuracy, we examine five different methods of combining forecasts. All these methods involve placing different weights over n individual forecasting models. The five model combining methods can be summarized as follows:

Combination Methods	**Post-1985 sample**	**Post-1995 sample**
1. Mean	ex-ante	ex-ante
2. Median	ex-ante	ex-ante
3. OLS	ex-post	ex-post/ex-ante
4. Equal Weight Prior	ex-post	ex-post/ex-ante
5. Unit Weight Prior	ex-post	ex-post/ex-ante

We distinguish between *ex-ante* and *ex-post* model combinations. Ex-ante optimal weights are computed using the history of out-of-sample forecasts up to time t. For example, the weights used to construct the ex-ante combined forecast in 2000:Q4 is based on a regression of realized annual inflation over 1985:Q4 to 2000:Q4 on the constructed out-of-sample forecasts for the same period. We examine ex-ante model combinations for the 1995:Q4 to 2002:Q4 period. Hence, the ex-ante method assesses actual out-of-sample forecasting power of combination methods.

We also ask the question whether ex-post, a particular combination of models would have performed better than individual forecasts. In the ex-post exercise, we use all the information in the sample to construct a single set of optimal weights. The ex-post analysis cannot be used for actual forecasting, but it provides us a picture of which models would have been most successful ex-post forecasting inflation out-of-sample. We examine ex-post model combinations for the

[3] See the literature reviews by, among others, Clemen (1989), Diebold and Lopez (1996), and more recently Timmermann (2004).

two samples 1985:Q4 to 2002:Q4 and 1995:Q4 to 2002:Q4.

In the first two model combining methods, we simply look at the overall mean and median, respectively, over n different forecasting models. These are simple ex-ante forecasts with fixed weights. Equal weighting of many forecasts has been used as early as Bates and Granger (1969) and, in practice, simple equal-weighting forecasting schemes are hard to beat. In particular, Stock and Watson (2003) show that this method produces superior out-of-sample forecasts of inflation. In the last three combination methods, we compute different individual model weights that vary over time. These weights are estimated as slope coefficients in a regression of realized inflation on model forecasts

$$\pi_{t+4,4} = \sum_{i=1}^{n} \omega_t^i f_t^i + \varepsilon_{t,t+4}, \;\; t = 1, \ldots, T, \tag{18}$$

where f_s^i is the i-th model forecast at time s. The $n \times 1$ weight vector $\omega_t = \{\omega_t^i\}$ is estimated either by OLS, as in our third combining model specification, or using the mixed regressor method proposed by Theil and Goldberger (1961) and Theil (1963), as in Combination Methods 4 and 5.

To describe the last two combination methods, we set up some notation. Suppose we have T forecast observations with n individual models. Let F be the $T \times n$ matrix of forecasts and π the $T \times 1$ vector of actual future inflation levels that are being forecast. Consequently, the s-th row of F is given by $F_s = \{f_s^1, ...f_s^n\}$. The mixed regression estimator can be viewed as a Bayesian estimator with the prior $\omega \sim N(\mu, \sigma_\omega^2 I)$, where σ_ω^2 is a scalar and I the $n \times n$ identity matrix. The estimator can be derived as:

$$\widehat{\omega} = (F'F + \gamma I)^{-1} (F'\pi + \gamma\mu), \tag{19}$$

where the parameter γ controls the amount of shrinkage towards the prior. In particular, when $\gamma = 0$, the estimator simplifies to standard OLS, and when $\gamma \to \infty$, the estimator approaches the prior. It is instructive to re-write the estimator as a weighted average of the OLS estimator and the prior:

$$\widehat{\omega} = \theta_{OLS}\, \omega_{OLS} + \theta_{prior}\, \mu$$

with $\theta_{OLS} = (F'F + \gamma I)^{-1} (F'F)$ and $\theta_{prior} = (F'F + \gamma I)^{-1} (\gamma I)$, so that the weights add up to the identity matrix.

We use empirical Bayes and estimate the shrinkage parameter as:

$$\widehat{\gamma} = \widehat{\sigma}^2 / \widehat{\sigma}_\omega^2, \tag{20}$$

where

$$\widehat{\sigma}^2 = \frac{1}{T}\pi' \left[I - F \left(F'F \right)^{-1} F' \right] \pi$$

and

$$\widehat{\sigma}_\omega^2 = \frac{\pi'\pi - T\widehat{\sigma}^2}{trace\left(F'F \right)}.$$

To interpret the shrinkage parameter, observe that $\widehat{\sigma}^2$ is simply the residual variance of the regression; the numerator of $\widehat{\sigma}_\omega^2$ is the fitted variance of the regression and the denominator is the average variance of the independent variables (the forecasts) in the regression. Consequently, the shrinkage parameter, γ, in equation (20) increases when the variance of the independent variables becomes larger, and decreases as the R^2 of the regression increases. In other words, if forecasts are (not) very variable and the regression R^2 is small (large), we trust the prior (the regression).

We examine the effect of two priors. In Model Combination 4, we use an equal-weight prior where each element of μ, $\mu_i = 1/n, i = 1, \ldots, n$, which leads to the Ridge regressor used by Stock and Watson (1999). In the second prior (Model Combination 5), we assign unit weight to one type of forecast, for example, $\mu = \{0 \ldots 1 \ldots 0\}'$. One natural choice for a unit weight prior would be to choose the best performing univariate forecast model.

When we compute the model weights, we impose the constraint that the weight on each model is positive and the weights sum to one. This has the natural interpretation that the weights represent the best combination of models that give good forecasts in their own right, rather than placing negative weights on models that give consistently wrong forecasts. This is also very similar to shrinkage methods of forecasting (see Stock and Watson (2005)). For example, Bayesian Model Averaging uses posterior probabilities as weights (which are, by construction, positive and sum to one).[4] The positivity constraint is imposed by minimizing the usual loss function, L, associated with OLS:

$$L = \left(\pi - F\omega \right)' \left(\pi - F\omega \right),$$

and a loss function for the mixed regressor estimations:

$$L = \frac{\left(\pi - F\omega \right)' \left(\pi - F\omega \right)}{\widehat{\sigma}^2} + \frac{\left(\omega - \mu \right)' \left(\omega - \mu \right)}{\widehat{\sigma}_\omega^2},$$

subject to the positivity constraints.

[4] Diebold (1988) shows that when the target is persistent, as in the case of inflation, the forecast error from the combination regression will typically be serially correlated and hence predictable, unless the constraint that the weights sum to one is imposed.

4 Empirical Results

Section 4.1 lays out our main empirical results for the forecasts of time-series models, OLS Phillips curve regressions, term structure models, and survey forecasts. We provide interpretations of our results in Section 4.2. In Section 4.3, we examine forecast combinations.

4.1 Forecast Accuracy

Time-Series Models

In Table 4, we report RMSE statistics, in annual percentage terms, for the ARIMA model out-of-sample forecasts over the the post-1985 and post-1995 periods. The RMSEs generally range from around 0.6-0.7% for PUXX to around 1.5% for PUXHS. Among the ARIMA models, the ARMA(1,1) model generates the lowest RMSE in forecasting all inflation series except PUNEW and core inflation (PUXX) post-1995. The random walk model also outperforms the ARMA model for PUXX because this measure is less variable than the other inflation measures over this sample period (see Table 1).[5] In the remainder of the paper, we use the ARMA(1,1) model as the benchmark model.

Table 4 also reports the results of the non-linear regime-switching model, RGM. Over the post-1985 period, RGM generally performs in line with, and slightly worse than, a standard ARMA model. There is some evidence that non-linearities are important for forecasting in the post-1995 sample, where the regime-switching model outperforms the ARIMA models in forecasting PUNEW and PUXHS. Both these inflation series become much less persistent post-1995, and the RGM model captures this by transitioning to a regime of less persistent inflation. However, the Hamilton (1989) RGM model performs worse than a linear ARMA model for forecasting PUXX and PCE.

OLS Phillips Curve Forecasts

Table 5 reports the out-of-sample RMSEs and the model comparison regression estimates (equation (17)) for the Phillips curve models described in Section 3.2, relative to the benchmark of

[5] We find that a random walk model on annual inflation performs better than the random walk model on quarterly inflation for all measures. However, this model still fails to beat the best time-series models for PUNEW and PUXHS, and still fails to beat the surveys for all three of the CPI inflation measures, PUNEW, PUXHS and PUXX (see below). It performs best when forecasting PCE inflation, generating lower RMSEs than the best quarterly models in both samples.

the ARMA(1,1) model. The overall picture in Table 5 is that the ARMA(1,1) model typically outperforms any Phillips curve forecast. Of the 80 comparisons (10 models, 2 out-samples, and 4 inflation measures), the model comparison regression coefficient $(1-\lambda)$ is significantly positive in only 9 out of 80 cases. A Phillips curve forecast also beats the ARMA(1,1) model in terms of RMSE in only 9 out of 80 cases. Hence, the predictive ability of the Phillips curve models is generally weak, relative to the time-series forecasts.

The OLS Phillips curve regressions are most successful in forecasting core inflation, PUXX. Of the 9 cases where the Phillips curve beats the ARMA(1,1) model, 5 occur for PUXX. The best model forecasting PUXX inflation uses the composite Bernanke-Boivin-Eliasz aggregate real activity factor (PC8), which strongly rejects the null that forecasts from the Phillips curve add nothing to the ARMA(1,1) benchmark. However, the aggregate macro-economic factor (constructed with look-ahead bias!) fares rather poorly in forecasting the other inflation measures (PUNEW, PUXHS, and PCE).

Another relatively successful Phillips curve specification is the PC7 model that uses the Stock-Watson nonfinancial Experimental Leading Index-2. This index does not embed asset pricing information. PC7 generates significantly positive $(1-\lambda)$ coefficients for PUNEW, PUXHS, and PCE in the post-1985 sample, but does not produce significant $(1-\lambda)$ coefficients in the post-1995 sample. For the post-1985 sample, the RMSEs of PC7 are also all higher than the RMSE of an ARMA(1,1) model. In contrast, the PC1 model, which simply uses past inflation and past GDP growth, delivers 5 of the 9 relative RMSEs below one and beats PC7 in all but one case.

Term Structure Forecasts

In Table 6, we report the out-of-sample forecasting results for the various term structure models (see Section 3.3). Generally, the term structure based forecasts perform worse than the Phillips-curve based forecasts. Over a total of 120 statistics (15 models, 4 inflation measures, 2 sample periods), term structure based-models beat the ARMA(1,1) model in only 10 cases in terms of producing smaller RMSE statistics.

The coefficient $(1-\lambda)$ is significantly positive in 16 out of 120 horse race regressions. Of these 16 cases, 10 occur for forecasting core inflation, PUXX, which is the inflation measure most successfully forecasted by the term structure models. In particular, the model TS1 that includes inflation, GDP growth, and the short rate beats an ARMA(1,1) and has a very positive $(1-\lambda)$ coefficient in both the post-1985 and post-1995 samples. The other models with term structure information that are successful at forecasting PUXX are TS6 and TS8, both of which

also include short rate information. The performance of TS6 is also impressive as it succeeds in beating the RMSE of a random walk in both out-samples.

The finance literature has instead typically used term spreads, not short rates, to predict future inflation changes (see, for example, Mishkin (1990, 1991)). In contrast to the relative success of the models with short rate information, models TS9-TS11 that incorporate information from the term spread perform badly and produce higher RMSE statistics than the benchmark ARMA(1,1) model. In fact, using term spreads in unconstrained regressions leads to poor forecasting performance for all four inflation measures. The poor inflation forecasts for the term spread is consistent with Estrella and Mishkin (1997) and Kozicki (1997), who find that the forecasting ability of the term spread is diminished after controlling for lagged inflation. After controlling for lagged inflation, the short rate still contains modest predictive power. Thus, the short rate, not the term spread, contains the most predictive power in simple forecasting regressions.

Table 6 shows that the performance for iterated VAR forecasts is mixed. VARs perform well, in producing lower RMSE than an ARMA(1,1), for PUNEW and PUXHS over the post-1995 sample, but otherwise deliver worse RMSEs than an ARMA(1,1). The relatively poor performance of long-horizon VAR forecasts for inflation contrasts with the good performance for VARs in forecasting GDP (see Ang, Piazzesi and Wei (2004)) and for forecasting other macroeconomic time series (see Marcellino, Stock and Watson (2004)). The non-linear empirical regime-switching VAR (RGMVAR) fares much worse than the VAR and is always beaten by an ARMA(1,1). This result stands in contrast to the relatively strong performance of the univariate regime-switching model using only inflation data (RGM in Table 4) for forecasting PUNEW and PUXX. This implies that the non-linearities in term structure data have no marginal value for forecasting inflation above the non-linearities already present in inflation itself.

The last two lines of each panel in Table 6 shows that there is some evidence that no-arbitrage forecasts (MDL1-2) are useful for forecasting PUXX by their significant $(1 - \lambda)$ coefficients. However, disappointingly, both no-arbitrage term structure models always fail to beat the ARMA(1,1) forecasts in terms of RMSE. While the finance literature shows that inflation is a very important determinant of yield curve movements, our results show that the no-arbitrage cross-section of yields appears to provide little marginal forecasting ability for the dynamics of future inflation over simple time-series models.

Surveys

Table 7 reports the results for the survey forecasts, and shows several notable results. First,

surveys perform very well in forecasting PUNEW, PUXHS, and PUXX. With only one exception, the unadulterated survey forecasts SPF1, LIV1 and MICH1 have lower RMSEs than ARMA(1,1) forecasts over both the post-1985 and the post-1995 samples (the exception is MICH1 for PUXX over the post-1985 sample). For example, for the post-1985 (post-1995) sample, the RMSE ratio of the raw SPF forecasts relative to an ARMA(1,1) is 0.779 (0.861) when predicting PUNEW. For PUNEW, PUXHS, and PUXX, the horse races always assign large, positive $(1 - \lambda)$ weights to the pure survey forecasts (the lowest one is 0.383) in both out-of-sample periods. In none of the cases can we reject the hypothesis that the ARMA(1,1) time-series model adds nothing to the predictive power of the raw survey forecasts.

Second, while the SPF and Livingston surveys do a good job at forecasting all three measures of CPI inflation (PUNEW, PUXHS, and PUXX) out-of-sample, the Michigan survey is relatively unsuccessful at forecasting core inflation, PUXX. It is not surprising that consumers in the Michigan survey fail to forecast PUXX, since PUXX excludes food and energy which are integral components of the consumer's basket of goods. Note that while the PUNEW and PUXHS measures have the highest correlations with each other, (over 98% in both out-samples and over 95% over the full sample), core inflation is less correlated with the other CPI measures. In particular, post-1995, the correlation of PUXX with PUNEW (PUXHS) is only 36% (26%).

Third, surveys do less well at forecasting PCE inflation, although there are a few significant positive coefficients on the SPF survey forecasts in the horse races. For PCE inflation, surveys almost always produce worse forecasts in terms of RMSE than an ARMA(1,1). This result is expected because the survey participants are asked to forecast CPI inflation, not the consumption deflator PCE. The PCE series is a deflator index, which is quite different to the fixed basket CPI index.

Fourth, the raw survey forecasts outperform the linear or non-linear bias adjusted forecasts (with the only notable exception being the bias-adjusted forecasts for PCE). As a specific example, for PUNEW, the relative RMSE ratios are always higher for the models with suffix 2 (linear bias adjustment) or the models with suffix 3 (non-linear bias adjustment) compared to the raw survey forecasts across all three surveys. This result is perhaps surprising due to the evidence of non-linear survey bias, but consistent with the weak evidence of linear bias, in the entire 1952-2002 sample (see Table 3). This implies the non-linear bias in survey forecasts is small, relative to the total amount of forecast error in predicting inflation.

Finally, we might expect that the Livingston and SPF surveys produce good forecasts because they are conducted among professionals. In contrast, participants in the Michigan survey are consumers, not professionals, yet the Michigan forecasts are of the same order of magnitude

as the Livingston and SPF surveys. For example, for PUNEW over the post-1995 sample, the Michigan RMSE ratio is 0.862, just slightly above the SPF RMSE ratio of 0.861. Hence, information aggregated over non-professionals also produces accurate forecasts that beats ARIMA time-series models!

The Livingston survey is the only survey available over our full sample, from 1952-2002. As McConnell and Perez-Quiros (2000) and Stock and Watson (2002b), among others, note, a notable feature of the post-1985 period is declining macro-economic volatility. Campbell (2004) finds that professionals were considerably more adept at forecasting GDP prior to 1985 relative to a simple AR(1) model than after 1985. In fact, post-1985, SPF forecasts of GDP perform worse than an AR(1) model. Thus, Campbell attributes a significant proportion of the total decline in the volatility of GDP to a decline in predictability as well as uncertainty. To see if the predictable components in inflation exhibit lower volatility than simple time-series models, we compute the RMSE ratio of the out-of-sample forecasts for the Livingston survey relative to an ARMA(1,1) model for 1960-1985 and 1986-2002, where the first 8 years are used as an in-sample estimation period for the ARMA(1,1) model. Over the pre-1985 sample, the Livingston RMSE ratio is 1.046 (with a RMSE level of 2.324), while over the post-1985 sample, the RMSE ratio is 0.789 (with a RMSE level of 0.896). In contrast to GDP forecasts, professionals are more adept at forecasting inflation in the post-1985 period.

4.2 Summary and Interpretation

Let us summarize the results so far. First, the ARMA (1,1) model is the overall best ARIMA time-series model and is relatively hard to beat across all the models. Nevertheless, quite often, some models that incorporate real activity information, term structure information, or, especially, survey information beat the ARMA(1,1) model, even when ARMA(1,1) forecasts are put in a forecast comparison regression. Second, the simplest Phillips curve model using only past inflation and GDP growth is a good performer. Third, adding term structure information often leads to an improvement in inflation forecasts, but generally only for core inflation. No-arbitrage restrictions actually generally lead to deterioration in fit. Fourth, the survey forecasts do very well in forecasting all inflation measures except PCE.

To get an overall picture of the relative forecasting power of the various models, Table 8 reports the relative RMSE ratios of the best models from each of the first three categories (pure time-series, Phillips-curve, and term structure models) and of each raw survey forecast. The most remarkable result in Table 8 is that for CPI inflation (PUNEW, PUXHS, and PUXX), the survey forecasts completely dominate the Phillips curve or term structure models in both out-of-

sample periods. For the post-1985 sample, the RMSEs are around 20% smaller for the survey forecasts compared to forecasts from Phillips-curve or term structure models. The exception is PCE inflation, which is hard to forecast. In fact, the best model for PCE in both out-samples is just the ARMA(1,1)!

With the exception of PCE, the surveys consistently deliver the RMSEs that are among the lowest for both the post-1985 and post-1995 periods. For the post-1985 sample, the best forecast is always a survey. The performance of the survey forecasts remains impressive in the post-1995 sample, but the Hamilton (1989) regime-switching model (RGM) has a slightly lower RMSE for PUNEW and PUXHS. Nevertheless, the survey RMSE are very similar to this best model. Impressively, the Livingston survey continues to deliver the most accurate forecast of PUXX post-1995.

Among the Phillips curve and term structure forecasts, the most simple PC1 and TS1 regressions frequently outperform more complicated models, especially for PUNEW. These regressions only use standard GDP growth. Other measures of economic growth are more successful at forecasting other measures of inflation. For PUXX inflation, PC8 produces forecasts that beat an ARMA(1,1) model for both the post-1985 and post-1995 sample. The PC8 forecasting model uses the Bernanke et al. (2005) composite indicator. More structured no-arbitrage approaches deliver better forecasts than unrestricted OLS regressions with term structure data only for MDL2 for PUXHS in the post-1985 sample. But, this specification still fails to beat an ARMA(1,1) model. The final important result in Table 8 is that non-linearities are important in forecasting inflation. For PUNEW and PUXHS, the univariate regime-switching model (RGM) delivers the best individual performance across all models, including surveys, over the post-1995 sample.

4.3 Combining Model Forecasts

Table 9 investigates how we can improve our forecasts by combining different models. We first combine models within each of the four categories (time-series, Phillips curve, term structure, and survey models), and then combine across all the models in the last column labelled “All Models.” The models in the survey category are only the SPF and Michigan survey because the Livingston survey is conducted at a semiannual frequency, as opposed to a quarterly frequency for all the other models. Since Table 7 shows the Livingston forecasts to be very similar to the SPF and Michigan surveys for PUNEW and PUXHS, and the best single forecaster for PUXX, excluding the Livingston survey places a conservative higher bound on our RMSEs for the forecast combinations involving surveys.

We use four methods of model combination: means or medians over all the models, and linear combinations using weights that are recursively computed using OLS or the mixed combination regression with an equal-weight prior. We start the model combination regressions at 1995:Q4 using realized inflation and the out-of-sample forecasts over 1985:Q4 to 1995:Q4. At each subsequent period, we advance the data sample by one quarter and re-run the model combination regression to obtain the slope coefficient estimates. We do not include the unit prior as it requires finding the best model at each step. Below, we will show that even if we pretend to know the best model (using look-ahead biased, full sample information), the unit weight regressions do not significantly improve on the regressions reported here.

There are three main findings in Table 9. First, using mean or median forecasts mostly does not improve the forecast performance relative to the best individual model. Taking the mean only improves out-of-sample forecasts for the term structure models for PUNEW, PUXHS, and PCE, but even here the improvements are tiny. Thus, simple methods of combining forecasts provide little additional predictive power relative to the best model (observed ex-post). But, ex-ante, model combinations can produce lower RMSEs than simple ARMA(1,1) models, as seen by the simple means of all model forecasts for PUNEW and PUXHS, which produce RMSE ratios less than one.

Second, updating the model weights based on previous model performance does not always lead to superior performance. For the term structure models, OLS model combinations outperform means and medians for all inflation measures except for PUXHS inflation. However, only for the PCE measure is using an OLS-based combination forecast better than the best individual model when all models are considered. Finally, the mixed equal-weight prior combination generally outperforms the OLS forecast combination, but, it when it outperforms, its RMSE is very close to the RMSE of the OLS forecast combination. Nevertheless, the performance of the best model is still better, sometimes substantially better, than all the model combinations. When all models are combined, the OLS and equal-weight combination methods only beat the best individual model for PCE inflation.

To help interpret the results, we investigate the ex-ante OLS weights on some selected models. In Figure 2, we plot the OLS slope estimates of regression (18) for various inflation measures over the period of 1995:Q4 to 2002:Q4. For clarity, rather than showing the weights on all models, we combine only the ex-post best model within each category (time-series, Phillips Curve, and term structure) with the SPF in the regression. Note that by choosing the best models, we handicap the survey forecasts. We compute the weights in the regression recursively like the forecasts in Table 9; that is, we start in 1995:Q4, and recursively compute forecasts from

1985:Q4 to 1995:Q4. This is a quasi out-of-sample exercise in the sense that all regressors in this regression are prior information, but we choose the best model to combine in each category based on information from a full-sample comparison.

Figure 2 shows that when forecasting all the CPI inflation measures (PUNEW, PUXHS, and PUXX), the data consistently places the highest ex-ante weights on survey forecasts and very little weight on the other models. The weights on the SPF1 forecasts are generally constant and around 0.8 for PUNEW and PUXX, and 0.9 for PUXHS. While the best time-series model, RGM, is the single best forecaster over all models for PUNEW and PUXHS over the post-1995 sample (see Table 8), the weights on RGM are almost zero. The highest weights for RGM occur over 1998-1999 for PUXHS where inflation started to increase. The weights on the Phillips curve and term structure forecasts are also close to zero and in fact become less important over time for PUXX.

For PCE inflation, surveys contain little information. The weights on the SPF1 start at 0.2 in 1995 but decline quickly and remain close to zero after 1997. Among the other categories of models, the ARMA(1,1) forecast stands out, with weights ranging from 0.4 to 0.7. The Phillips curve forecast also receives a relatively high weight of 0.4, but always smaller than the ARMA(1,1) model. These results are consistent with the poor forecasting performance of all the models in Table 8, where individual models or combinations of models barely improve on, and usually do worse than, an ARMA(1,1) forecast.

We can also ask the question whether ex-post, combinations of models would have performed better than individual forecasts. In this exercise, we run the regression over the full sample so that the weights are not recursively updated. This exercise shows which combination of models would have provided the best forecasts ex-post. Apart from combining models within each category and across all models, we also look at combining the ex-post best models from each category listed in Table 8. Table 10 reports the results.

Table 10 shows that using future information to compute the ex-post weights often does not beat the best individual forecasting model in terms of RMSE. The best performance of ex-post model combination across all models occurs for forecasting PUNEW in the post-1995 sample, where OLS (equal-weight priors) produce a RMSE ratio relative to an ARMA(1,1) of 0.722 (0.726), while the best individual model (RGM) produces a RMSE ratio of 0.764. In all other cases, the combined forecast does not beat the best individual model forecast, and where the combined forecast does beat the best individual model, the RMSEs are very similar. This is strong evidence that combining forecasts, at least with the techniques explored here, is not a

very useful forecasting tool.[6] We also find that imposing a unit prior on the best forecasting model also does not necessarily lead to lower RMSEs, even when ex-post information is used. In fact, for many cases, like PUNEW for the post-1995 sample, the unit prior underperforms an OLS or an equal-weight prior in terms of lower RMSEs.

5 Robustness to Non-Stationary Inflation

5.1 Definition and Models

In this section we investigate the robustness of our results to the alternative assumption that quarterly inflation is difference stationary. Our exercise is now to forecast four-quarter ahead inflation changes:

$$\begin{aligned} \mathrm{E}_t\left(\pi_{t+4,4} - \pi_{t,4}\right) &= \mathrm{E}_t\left[\sum_{i=-3}^{3}(4-|i|)\Delta\pi_{t+1+i}\right] \\ &= \mathrm{E}_t\left[\sum_{i=0}^{3}(4-i)\Delta\pi_{t+1+i}\right] + 4\pi_t - \pi_{t,4}, \end{aligned} \tag{21}$$

where $\pi_{t+4,4}$ is annual inflation defined in equation (3).

We now replace quarterly inflation, π_t, by quarterly inflation changes, $\Delta\pi_{t+1} = \pi_{t+1} - \pi_t$ in all the models considered in Sections 3.1 to 3.3. For example, we estimate an ARMA(1,1) on first differences of inflation:

$$\Delta\pi_{t+1} = \mu + \phi\Delta\pi_t + \psi\varepsilon_t + \varepsilon_{t+1}$$

and an AR(p) on first differences of inflation:

$$\Delta\pi_{t+1} = \mu + \phi_1\Delta\pi_t + \phi_2\Delta\pi_{t-1} + \ldots + \phi_p\Delta\pi_{t-p+1} + \varepsilon_{t+1}.$$

The OLS Phillips Curve and term structure regressions are performed by including quarterly inflation changes rather than quarterly inflation as one of the regressors. From the models estimated on $\Delta\pi_t$, we compute forecasts of inflation changes over the next year, $\mathrm{E}_t(\pi_{t+4,4}-\pi_{t,4})$.

There are three models for which we do not estimate a counterpart using quarterly inflation differences: We do not consider a random walk model for inflation changes and do not specify

[6] In unreported results, we also consider unconstrained regressions, that is, regression where the weights are not constrained to lie between 0 and 1. Given the poor performance of the forecasting models in the recent period, it is not surprising that some of the combined models with full-sample information significantly outperform the best single model forecasts in cases where many models (particularly term structure forecasts) have significantly negative weights.

the no-arbitrage term structure models to have non-stationary inflation dynamics (MLD1-2), although we still consider the forecasts of annual inflation changes implied by the original stationary models. In all other cases, we examine the forecasts of both the original stationary models and the new non-stationary models that use first differences of inflation. Note that the original models estimated on inflation levels generate RMSEs for forecasting annual inflation changes that are identical to the RMSEs for forecasting annual inflation levels. Hence, the question is whether models estimated on differences provide superior forecasts to models estimated on levels. We maintain the ARMA(1,1) model estimated on inflation rate levels as a benchmark.

5.2 Performance of Individual Models

Over both the post-1985 and post-1995 out-samples, the RMSE statistics are very similar for the models specified in inflation levels or inflation changes. For example, post-1985 for PUNEW, the RMSE of forecasting annual inflation changes by an ARMA(1,1) model estimated on inflation levels is 1.136%, compared to the RMSE of 1.217% for forecasting inflation differences for the ARMA(1,1) estimated on inflation differences. The RMSEs for the other inflation measures are also similar for inflation levels or differences. Thus, the magnitudes of the errors are similar for forecasting in levels or differences.

Table 11 reports the RMSE ratios of the best performing models on levels or differences within each category for forecasting inflation changes. Table 11 shows that with the exception of PUXX, time-series models estimated on levels provide lower RMSEs than time-series models estimated on differences. For the PUNEW and PUXHS measures, the best time-series model estimated on levels (the ARMA(1,1) model over the post-1985 sample and the regime-switching model (RGM) over the post-1995 sample) also out-performs the more complicated Phillips Curve and term structure models in all cases but one. However, for the PUXX and PCE measures, Phillips curve and term structure regressions using past inflation changes are slightly more accurate than regressions with past inflation levels.

Our major finding that surveys generally out-perform other model forecasts is robust to specifying the models in inflation differences. For the CPI inflation measures (PUNEW, PUXHS, PUXX) over the post-1985 sample, surveys deliver lower RMSEs than the best time-series, Phillips curve, and term structure forecasts. First difference models help the most for lowering RMSEs for PUXX over the post-1995 sample, where the best term structure model estimated on differences (TS6) produces a relative RMSE ratio of 0.655. This is still beaten by the raw Livingston survey, with a RMSE ratio of 0.557.

In unreported results available upon request, we find that in the model comparison regression

(17) against a stationary ARMA(1,1) model and with annual inflation changes on the LHS, models specified in differences do not fare any better than models specified in levels. For example, over the post-1985 sample for PUNEW, only one Phillips curve model (PC7) and two term structure models (RGMVAR and MDL1), all estimated on differences, provide additional information about future inflation over a stationary ARMA(1,1) model. For PUXX post-1985, we can reject the null that the models add nothing to the ARMA(1,1) forecast only for three Phillips curve models (PC2, PC4, and PC9) and two term structure models (TS2 and VAR) estimated on differences. In contrast, surveys consistently provide significant improvement in forecasting inflation changes above an ARMA(1,1) model, especially for PUNEW, PUXHS, and PUXX in the post-1985 sample period.

5.3 Performance of Combining Models

Similar to Section 3.6, we also run forecast combination regressions to determine both ex-ante and ex-post the best combination of models to forecast inflation changes. The model weights are computed from the regression:

$$\pi_{s+4,4} - \pi_{s,4} = \sum_{i=1}^{n} \omega_s^i f_s^i + \varepsilon_{s,s+4}, \quad s = 1, \ldots, t. \tag{22}$$

We repeat the exercise of Table 9 and compute ex-ante recursive weights over 1995:Q4-2002:Q4 using the best forecasting models over the full-sample in each category. In unreported results available upon request, we find that our original results for forecasting inflation levels also extends to forecasting inflation changes. Specifically, there is generally no improvement in combining model forecasts, or when model combinations result in out-performance, the improvement is small. Specifically, for PUNEW and PUXHS, using means, medians, OLS, or an equal-weight prior produces higher RMSEs than the best individual model. For these inflation measures, all model combinations produce RMSEs that are higher than the survey forecasts. This result is robust to both combining models in levels and also combining models in differences. There are some improvements for forecasting PCE inflation using models in differences, but the forecasting gains are very small.

In Figures 3 and 4, we plot the OLS coefficient estimates of equation (22) for the models specified in differences and the models specified in levels, respectively, together with the raw SPF forecast. Similar to Figure 2, we compute the OLS ex-ante weights recursively over 1995:Q4 to 2004:Q4, but choose the best performing time-series, Phillips Curve, and term structure models from full-sample information. Both Figures 3 and 4 confirm the robustness of our findings of the superior survey forecasts of inflation changes.

In Figure 3 the weight on the SPF survey for PUNEW and PUXHS changes is above or around 0.8. The surveys clearly dominate the I(1) time-series, Phillips Curve, and term structure models. For PUXX changes, the model combinations still place the largest weight on the SPF survey, but the weight is around 0.5. In contrast, for forecasting PUXX inflation levels, the weights on SPF1 range from 0.6 to above 0.9. Thus, there is now additional information in the other models for forecasting PUXX changes, most particularly the Phillips Curve PC1 model. Nevertheless, surveys still have the highest weight on model combination regressions. Consistent with the results of forecasting inflation levels, surveys provide little information to forecast PCE changes. For PCE changes, the largest ex-ante weight in the forecast combination regression is for the ARMA(1,1) estimated on differences.

Like Figure 3, Figure 4 examines the performance of the SPF combined with other models in forecasting inflation changes, except that it considers stationary models. While Table 11 shows that the RGM model on levels gives the lowest RMSE over the post-1995 sample for PUNEW and PUXHS differences, there appears to be little additional value in the RGM forecast once surveys are included. Figure 4 shows that the forecast combination regression has almost zero ex-ante weight on the RGM model. The weights on the other I(0) models are also low, whereas the SPF weights are around 0.8 or higher. Compared to the other stationary model categories, the SPF also has an edge at forecasting PUXX inflation. Again, surveys do not perform well relative to I(0) models for forecasting PCE changes.

6 Conclusion

We have conducted a very comprehensive analysis of different inflation forecasting methods using four inflation measures and two different out-of-sample periods (post-1985 and post-1995). We investigated forecasts based on time-series models, Phillips curve inspired forecasts, and forecasts embedding information from the term structure, through linear regressions, non-linear regime switching models or filtered through arbitrage-free term structure models. We also investigated the forecasting performance of three different survey measures (the SPF, Livingston, and Michigan surveys), examining both raw and bias-adjusted survey measures.

Our results can be summarized as follows. First, the best time series model is mostly a simple ARMA(1,1) model, which can be motivated by thinking of inflation comprising stochastic expected inflation following an AR(1) process, and shocks to inflation. Second, while the ARMA(1,1) model is hard to beat in terms of RMSE forecast accuracy, it is never the best model. For CPI measures, the survey measures consistently deliver better forecasts than

ARMA(1,1) models, and in fact, much better forecasts than Phillips curve-based regressions, or term structure models. However, surveys do a relatively poor job at forecasting PCE inflation, as do all the inflation forecasting models.

Third, term structure information does not generally lead to better forecasts and often leads to inferior forecasts than models using only aggregate activity measures. This result extends to simple term structure models that are based on unrestricted OLS regressions, non-linear models, iterated VAR forecasts, and also no-arbitrage term structure models that use information from the entire cross-section of yields. Whereas this seems to confirm the results in Stock and Watson (1999), our investigation of term structure models is much more comprehensive.

Finally, we also examined forecasts that combine information from various models or from various data sources. Our real activity measures included the Bernanke et al. (2005) measure of aggregate activity based on 65 separate time-series of various macro factors measuring real activity. This forecast is dominated by surveys. We find that model combinations do not generally lead to better performance. Simple means or medians of forecasts, or forecast combination regressions that use prior information often produce inferior forecasts than, and when they out-perform, their performance is similar to, the best individual performing forecasts. In both ex-ante and ex-post model combination exercises, almost all the weight is placed on survey forecasts for forecasting CPI inflation.

Two conclusions stand out from these results and provide a clear suggestion for future research. First, survey forecasts have at times provided very high quality forecasts that beat simple economic models, suggesting that the surveys have information absent in extant models. Since surveys aggregate information from many different sources, the superior information in median survey forecasts may be due to an effect similar to Bayesian Model Averaging, or averaging across potentially hundreds of different individual forecasts and taking common components (see Stock and Watson (2002a) and Timmermann (2004)). While the Michigan survey, which is conducted among relatively unsophisticated consumers, produces aggregate forecasts of CPI inflation that are worse than the Livingston and SPF surveys, which are conducted among professionals. But, the Michigan survey also generally betas time-series, Phillips curve, and term structure forecasts with errors similar in magnitude to the Livingston and SPF surveys.

Surveys may be capturing information that is orthogonal to information that can be obtained by averaging across large numbers of models. Hence, one avenue for future research is to investigate whether alternative techniques for combining forecasts perform better (see Inoue and Killian (2005) for a survey and study of one promising technique). At the very least, our results strongly suggest that there would be additional information in including survey forecasts

in the large datasets used to construct a small number of composite factors, which are designed to summarize aggregate macroeconomic dynamics (see Bernanke et. al. (2005) and Stock and Watson (2005), among others.)

Second, extant sophisticated no-arbitrage term structure models, while performing well in sample, seem to provide relatively poor forecasts relative to simpler term structure or Phillips curve models out-of-sample. A potential solution is to introduce the information present in the surveys as additional state variables in the term structure models. Pennacchi (1991) was an early attempt in that direction and Kim (2004) is a recent attempt to build survey expectations into a no-arbitrage quadratic term structure model.

References

[1] Atkeson, A., and L. E. Ohanian, 2001, "Are Phillips Curves Useful for Forecasting Inflation?" *Federal Reserve Bank of Minneapolis Quarterly Review*, 25, 2-11.

[2] Ang, A., and G. Bekaert, 2002, "Regime Switches in Interest Rates," *Journal of Business and Economic Statistics*, 20, 163-182.

[3] Ang, A., and G. Bekaert, 2004, "The Term Structure of Real Rates and Expected Inflation," working paper, Columbia University.

[4] Ang, A., M. Piazzesi, and M. Wei, 2004, "What does the Yield Curve Tell us about GDP Growth?" forthcoming *Journal of Econometrics*.

[5] Bai, J., and S. Ng, 2004, "A Panic Attack on Unit Roots and Cointegration," *Econometrica*, 72, 4, 1127-1177.

[6] Bates, J. M., and C. W. J. Granger, 1969, "The Combination of Forecasts," *Operations Research Quarterly*, 20, 451-468.

[7] Bekaert, G., S. Cho, and A. Moreno, 2005, "New Keynesian Macroeconomics and the Term Structure," Working Paper, Columbia University.

[8] Bekaert, G., R. J. Hodrick, and D. Marshall, 2001, "Peso Problem Explanations for Term Structure Anomalies," *Journal of Monetary Economics*, 48, 2, 241-270.

[9] Bernanke, B. S., and J. Boivin, 2003, "Monetary Policy in a Data-Rich Environment," *Journal of Monetary Economics*, 50, 3, 525-546.

[10] Bernanke, B. S., J. Boivin, and P. Eliasz, 2005, "Measuring the Effects of Monetary Policy: A Factor-Augmented Vector Autoregressive (FAVAR) Approach," *Quarterly Journal of Economics*, 120, 1, 387-422.

[11] Bryan, M., and S. G. Cecchetti, 1993, "The Consumer Price Index as a Measure of Inflation," *Economic Review of the Federal Reserve Bank of Cleveland*, 29, 15-24.

[12] Campbell, S. D., 2004, "Volatility, Predictability and Uncertainty in the Great Moderation: Evidence from the Survey of Professional Forecasters," working paper, Board of Governors.

[13] Carlson, J. A., 1977, "A Study of Price Forecasts," *Annals of Economic and Social Measurement*, 1, 27–56.

[14] Clark, T. E., and M. W. McCracken, 2003, "The Predictive Content of the Output Gap for Inflation: Resolving In-Sample and Out-of-Sample Evidence," working paper, University of Missouri-Columbia.

[15] Clemen, R. T., 1989, "Combining Forecasts: A Review and Annotated Bibliography," *International Journal of Forecasting*, 5, 559-581.

[16] Chen, R. R., and L. Scott, 1993, "Maximum Likelihood Estimation for a Multi-factor Equilibrium Model of the Term Structure of Interest Rates," *Journal of Fixed Income*, 3, 14-31.

[17] Cecchetti, S., R. Chu, and C. Steindel, 2000, "The Unrealiability of Inflation Indicators," *Federal Reserve Bank of New York Current Issues in Economics and Finance*, 6, 1-6.

[18] Cochrane, J., and M. Piazzesi, 2005, "Bond Risk Premia," *American Economic Review*, 95, 1, 138-160.

[19] Curtin, R. T., 1996, "Procedure to Estimate Price Expectations," Manuscript, University of Michigan Survey Research Center.

[20] Dai, Q., and K. J. Singleton, 2002, "Expectation Puzzles, Time-Varying Risk Premia, and Affine Models of the Term Structure," *Journal of Financial Economics*, 63, 415-41.

[21] Diebold, F. X., 1989, "Forecast Combination and Encompassing: Reconciling Two Divergent Literatures," *International Journal of Forecasting*, 5, 589-92.

[22] Diebold, F. X., and J. A. Lopez, 1996, "Forecasting Evaluation and Combination," in G. S. Maddala and C. R. Rao (eds.), *Handbook of Statistics*, 241-268, Elsevier, Amsterdam.

[23] Duffee, G. R., 2002, "Term Premia and the Interest Rate Forecasts in Affine Models," *Journal of Finance*, 57, 1, 405–443.

[24] Duffie, D., and R. Kan, 1996, "A Yield-Factor Model of Interest Rates," *Mathematical Finance*, 6, 379-406.

[25] Estrella, A., and F. 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," *European Economic Review*, 41, 1375-401.

[26] Evans, M. D. D., and K. K. Lewis, 1995, "Do Expected Shifts in Inflation Affect Estimates of the Long-Run Fisher Relation?," *Journal of Finance*, 50, 1, 225-253.

[27] Evans, M. D. D., and P. Wachtel, 1993, "Inflation Regimes and the Sources of Inflation Uncertainty," *Journal of Money, Credit and Banking*, 25, 3, 475-511.

[28] Fama, E. F., 1975, "Short-Term Interest Rates as Predictors of Inflation," *American Economic Review*, 65, 3, 269-282.

[29] Fama, E. F., and M. R. Gibbons, 1984, "A Comparison of Inflation Forecasts," *Journal of Monetary Economics*, 13, 3, 327-348.

[30] Frankel, J. A., and C. S. Lown, 1994, "An Indicator of Future Inflation Extracted from the Steepness of the Interest Rate Yield Curve along Its Entire Length," *Quarterly Journal of Economics*, 59, 517-30.

[31] Fuhrer, J., and G. Moore, 1995, "Inflation Persistence," *Quarterly Journal of Economics*, 110, 1, 127-159.

[32] Gali, J., and M. Gertler, 1999, "Inflation Dynamics: A Structural Econometrics Analysis," *Journal of Monetary Economics*, 44, 2, 195-222.

[33] Grant, A. P., and L. B. Thomas, 1999, "Inflation Expectations and Rationality Revisited," *Economics Letters*, 62, 3, 331-8.

[34] Gray, S. F., 1996, "Modeling the Conditional Distribution of Interest Rates as a Regime-Switching Process," *Journal of Financial Economics*, 42, 27-62.

[35] Hamilton, J. D., 1985, "Uncovering Financial Market Expectations of Inflation," *Journal of Political Economy*, 93, 6, 1224-1241.

[36] Hamilton, J., 1988, "Rational-Expectations Econometric Analysis of Changes in Regime: An Investigation of the Term Structure of Interest Rates," *Journal of Economic Dynamics and Control*, 12, 385-423.

[37] Hamilton, J., 1989, "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle," *Econometrica*, 57, 357-384.

[38] Hansen, L. P., and R. J. Hodrick, 1980, "Forward exchange rates as optimal predictors of future spot rates: an econometric analysis," *Journal of Political Economy*, 88, 5, 829-853.

[39] Hodrick, R. J., and E. C. Prescott, 1997, "Postwar U.S. Business Cycles: An Empirical Investigation," *Journal of Money, Credit and Banking*, 29, 1, 1-16.

[40] Inoue, A., and L. Kilian, 2005, "How Useful is Bagging in Forecasting Economic Time Series? A Case Study of U.S. CPI Inflation," working paper, University of Michigan.

[41] Jorion, P., and F. S. Mishkin, 1991, "A Multi-Country Comparison of Term Structure Forecasts at Long Horizons," *Journal of Financial Economics*, 29, 59-80.

[42] Kim, D. H., 2004, "Inflation and the Real Term Structure," working paper, Federal Reserve Board of Governor.

[43] Kim, C. J., and C. R. Nelson, 1999, "Has the U.S. Economy Become More Stable? A Bayesian Approach Based on a Markov Switching Model of the Business Cycle," *Review of Economics and Statistics*, 81, 608-616.

[44] Kozicki, S., 1997, "Predicting Real Growth and Inflation with the Yield Spread," *Federal Reserve Bank of Kansas City Economic Review*, 82, 39-57.

[45] Marcellino, M., J. H. Stock, and M. W. Watson, 2004, "A Comparison of Direct and Iterated Multistep AR Methods for Forecasting Macroeconomic Time Series," working paper, Harvard University.

[46] Mehra, Y. P., 2002, "Survey Measures of Expected Inflation: Revisiting the Issues of Predictive Content and Rationality," *Federal Reserve Bank of Richmond Economic Quarterly*, 88, 3, 17-36.

[47] McConnell, M. M., and G. Perez-Quiros, 2000, "Output Fluctuations in the United States: What Has Changed Since the Early 1950's," *American Economic Review*, 90, 5, 1464-1476.

[48] Mishkin, F. S., 1990, "What does the Term Structure Tell us about Future Inflation?" *Journal of Monetary Economics*, 25, 77-95.

[49] Mishkin, F. S., 1991, "A Multi-Country Study of the Information in the Term Structure about Future Inflation," *Journal of International Money and Finance*, 19, 2-22.

[50] Nelson, C. R., and G. W. Schwert, 1977, "On Testing the Hypothesis that the Real Rate of Interest is Constant," *American Economic Review*, 67, 478-486.

[51] Orphanides, A., and S. van Norden, 2003, "The Reliability of Inflation Forecasts Based on Output Gap Estimates in Real Time," working paper, CIRANO.

[52] Pennacchi, G. G., 1991, "Identifying the Dynamics of Real Interest Rates and Inflation: Evidence Using Survey Data," *Review of Financial Studies*, 4, 53-86.

[53] Quah, D., and S. P. Vahey, 1995, "Measuring Core Inflation," *Economic Journal* 105, 1130-1144.

[54] Sims, C. A., 2002, "The Role of Models and Probabilities in the Monetary Policy Process," *Brookings Papers on Economic Activity*, 2, 1-40.

[55] Stock, J. H., and M. W. Watson, 1989, "New Indexes of Coincident and Leading Economic Indicators," *NBER Macroeconomics Annual*, 351-394.

[56] Stock, J. H., and M. W. Watson, 1999, "Forecasting Inflation," *Journal of Monetary Economics*, 44, 293-335.

[57] Stock, J. H., and M. W. Watson, 2002a, "Forecasting Using Principal Components from a Large Number of Predictors," *Journal of the American Statistical Association*, 97, 1167-1179.

[58] Stock, J. H., and M. W. Watson, 2002b, "Has the Business Cycle Changed and Why?," in Gertler M. and K. Rogoff (eds.), *NBER Macroeconomics Annual 2002*, MIT Press.

[59] Stock, J. H., and M. W. Watson, 2003, "Forecasting Output and Inflation: The Role of Asset Prices," *Journal of Economic Literature*, 41, 788-829.

[60] Stock J.H., and M. W. Watson, 2005, "An Empirical Comparison of Methods for Forecasting Using Many Predictors," working paper, Harvard University.

[61] Stockton, D., and J. Glassman, 1987, "An Evaluation of the Forecast Performance of Alternative Models of Inflation," *Review of Economics and Statistics*, 69, 1, 108-117.

[62] Theil, H., 1963, "On the Use of Incomplete Prior Information in Regression Analysis," *Journal of the American Statistical Association*, 58, 401-14.

[63] Theil, H., and A. S. Goldberger, 1961, "On Pure and Mixed Estimation in Economics," *International Economic Review*, 2, 65-78.

[64] Thomas, L. B., 1999, "Survey Measures of Expected U.S. Inflation," *Journal of Economic Perspectives*, 13, 4, 125-44.

[65] Timmermann, A., 2004, "Forecast Combinations," forthcoming in G. Elliot, C. W. J. Granger and A. Timmermann (eds.), *Handbook of Economic Forecasting*, Elsevier, Amsterdam.

[66] Wright, J. H., 2004, "Forecasting U.S. Inflation by Bayesian Model Averaging," working paper, Federal Reserve Board of Governors.

Table 1: Summary Statistics

	PUNEW	*PUXHS*	*PUXX*	*PCE*
Panel A: 1952:Q2 – 2002:Q4*				
Mean	3.84	3.60	4.24	3.84
	(0.20)	(0.20)	(0.19)	(0.19)
Standard Deviation	2.86	2.78	2.56	2.45
	(0.14)	(0.14)	(0.14)	(0.13)
Autocorrelation	0.79	0.76	0.79	0.80
	(0.10)	(0.10)	(0.11)	0.11)
Correlations				
PUXHS	0.99			
PUXX	0.94	0.91		
PCE	0.98	0.98	0.93	
Panel B: 1986:Q1–2002:Q4				
Mean	3.09	2.87	3.21	2.58
	(0.14)	(0.17)	0.12)	(0.14)
Standard Deviation	1.12	1.37	0.97	1.08
	(0.10)	(0.12)	(0.09)	(0.10)
Autocorrelation	0.51	0.41	0.81	0.69
	(0.10)	(0.13)	(0.09)	(0.09)
Correlations				
PUXHS	0.99			
PUXX	0.85	0.79		
PCE	0.95	0.93	0.90	
Panel C: 1996:Q1–2002:Q4				
Mean	2.27	1.84	2.32	1.70
	(0.17)	(0.25)	(0.05)	(0.13)
Standard Deviation	0.81	1.19	0.24	0.62
	(0.12)	(0.17)	(0.03)	(0.09)
Autocorrelation	−0.03	0.07	−0.15	−0.01
	(0.22)	(0.21)	(0.23)	(0.19)
Correlations				
PUXHS	0.99			
PUXX	0.33	0.21		
PCE	0.89	0.88	0.19	

This table reports various moments of different measures of annual inflation at a quarterly frequency for different sample periods. PUNEW is CPI-U All Items; PUXHS is CPI-U Less Shelter; PUXX is CPI-U All Items Less Food and Energy, also called core CPI; and PCE is the Personal Consumption Expenditure deflator. All measures are in annual percentage terms. Standard errors reported in parentheses are computed by GMM.

* For PUXX, the start date is 1958:Q2 and for PCE, the start date is 1960:Q2.

Table 2: Forecasting Models

	Abbreviation	Specification
Time-Series Models	ARMA	ARMA(1,1)
	AR	Autoregressive model
	RW	Random Walk
	RGM	Univariate regime-switching model
Phillips Curve (OLS)	PC1	INFL + GDPG
	PC2	INFL + GAP1
	PC3	INFL + GAP2
	PC4	INFL + LSHR
	PC5	INFL + UNEMP
	PC6	INFL + XLI
	PC7	INFL + XLI-2
	PC8	INFL + FAC
	PC9	INFL + GAP1 + LSHR
	PC10	INFL + GAP2 + LSHR
OLS Term Structure Models	TS1	INFL + GDPG + RATE
	TS2	INFL + GAP1 + RATE
	TS3	INFL + GAP2 + RATE
	TS4	INFL + LSHR + RATE
	TS5	INFL + UNEMP + RATE
	TS6	INFL + XLI + RATE
	TS7	INFL + XLI-2 + RATE
	TS8	INFL + FAC + RATE
	TS9	INFL + SPD
	TS10	INFL + RATE + SPD
	TS11	INFL + GDPG + RATE + SPD
Empirical Term Structure Models	VAR	VAR(1) on RATE, SPD, INFL, GDPG
	RGMVAR	Regime-switching model on RATE, SPD, INFL
No-Arbitrage Term Structure Models	MDL1	Three-factor affine model
	MDL2	General three-factor regime-switching model
Inflation Surveys	SPF1	Survey of Professional Forecasters
	SPF2	Linear bias-corrected SPF
	SPF3	Non-linear bias-corrected SPF
	LIV1	Livingston Survey
	LIV2	Linear bias-corrected Livingston
	LIV3	Non-linear bias-corrected Livingston
	MICH1	Michigan Survey
	MICH2	Linear bias-corrected Michigan
	MICH3	Non-linear bias-corrected Michigan

INFL refers to the inflation rate over the previous quarter; GDPG to GDP growth; GAP1 to detrended log real GDP using a quadratic trend; GAP2 to detrended log real GDP using the Hodrick-Prescott filter; LSHR to the labor income share; UNEMP to the unemployment rate; XLI to the Stock-Watson Experimental Leading Index; XLI-2 to the Stock-Watson Experimental Leading Index-2; FAC to an aggregate composite real activity factor constructed by Bernanke, Boivin and Eliasz (2004); RATE to the 1-quarter yield; and SPD to the difference between the 20-quarter and the 1-quarter yield.

Table 3: Bias of Survey Forecasts

		α_1	α_2	β_1	β_2
PUNEW	SPF	0.330		0.482**	
		(0.173)		(0.190)	
	Livingston	0.159		0.993	
		(0.094)		(0.161)	
	Michigan	-0.206		1.276	
		(0.164)		(0.205)	
	SPF	0.359*	-0.047	0.414**	0.128
		(0.168)	(0.146)	(0.180)	(0.140)
	Livingston	0.147**	-0.074	0.806**	0.461**
		(0.044)	(0.126)	(0.067)	(0.153)
	Michigan	0.010	-0.315	0.959	0.482
		(0.107)	(0.206)	(0.099)	(0.249)
PUXHS	SPF	0.160		0.601*	
		(0.201)		(0.199)	
	Livingston	0.140		0.942	
		(0.084)		(0.130)	
	Michigan	-0.185		1.167	
		(0.155)		(0.166)	
	SPF	0.153	-0.067	0.580*	0.147
		(0.179)	(0.271)	(0.164)	(0.279)
	Livingston	0.142**	-0.048	0.765**	0.389**
		(0.051)	(0.144)	(0.070)	(0.129)
	Michigan	-0.067	-0.181	1.002	0.262*
		(0.153)	(0.143)	(0.143)	(0.132)
PUXX	SPF	0.213		0.694	
		(0.153)		(0.179)	
	Livingston	0.095		1.055	
		(0.107)		(0.133)	
	Michigan	-0.070		1.194	
		(0.117)		(0.124)	
	SPF	0.242	-0.050	0.643	0.100
		(0.166)	(0.124)	(0.192)	(0.123)
	Livingston	0.108	0.031	0.931	0.165
		(0.077)	(0.133)	(0.106)	(0.118)
	Michigan	-0.040	-0.011	1.137	0.059
		(0.145)	(0.210)	(0.146)	(0.245)
PCE	SPF	0.010		0.728*	
		(0.125)		(0.125)	
	Livingston	0.059		0.949	
		(0.120)		(0.136)	
	Michigan	-0.137		1.058	
		(0.130)		(0.139)	
	SPF	0.031	-0.143	0.689**	0.213
		(0.120)	(0.188)	(0.108)	(0.187)
	Livingston	0.070	-0.023	0.785*	0.399**
		(0.113)	(0.120)	(0.087)	(0.085)
	Michigan	-0.015	-0.172	0.900	0.228
		(0.145)	(0.140)	(0.145)	(0.117)

This table reports the coefficient estimates in equations (15) and (16). Numbers in the drift columns (α_1 and α_2) are reported in quarterly percentages. We denote values of α_1, α_2 and β_2 that are different from zero, and values of β_1 that are different from one at the 95% and 99% level by * and **, respectively, based on Hansen and Hodrick (1980) standard errors (reported in parentheses). For the SPF survey, the sample is 1981:Q3 to 2002:Q4; for the Livingston survey, the sample is 1952:Q2 to 2002:Q4 for PUNEW and PUXHS, 1958:Q2 to 2002:Q4 for PUXX, and 1960:Q2 to 2002:Q4 for PCE; and for the Michigan survey, the sample is 1978:Q1 to 2002:Q4.

Table 4: Time-Series Forecasts of Annual Inflation

		Post-1985 Sample		Post-1995 Sample	
		RMSE	ARMA=1	RMSE	ARMA=1
PUNEW	ARMA	1.136	1.000	1.144	1.000
	AR	1.140	1.003	1.130	0.988
	RW	1.626	1.431	1.529	1.337
	RGM	1.420	1.250	0.873	0.764
PUXHS	ARMA	1.490	1.000	1.626	1.000
	AR	1.515	1.017	1.634	1.005
	RW	2.172	1.458	2.146	1.320
	RGM	1.591	1.068	1.355	0.833
PUXX	ARMA	0.630	1.000	0.600	1.000
	AR	0.644	1.023	0.593	0.988
	RW	0.675	1.072	0.549	0.915
	RGM	0.677	1.075	0.727	1.211
PCE	ARMA	0.878	1.000	0.944	1.000
	AR	0.942	1.073	1.014	1.074
	RW	1.140	1.298	1.215	1.288
	RGM	0.945	1.077	1.081	1.146

We forecast annual inflation out-of-sample over 1985:Q4 to 2002:Q4 and from 1995:Q4 to 2002:Q4 at a quarterly frequency. Table 2 contains full details of the time-series models. Numbers in the RMSE columns are reported in annual percentage terms. The column labeled ARMA = 1 reports the ratio of the RMSE relative to the ARMA(1,1) specification.

Table 5: OLS Phillips Curve Forecasts of Annual Inflation

		Post-1985 Sample			Post-1995 Sample		
		Relative RMSE	$1-\lambda$	SE	Relative RMSE	$1-\lambda$	SE
PUNEW	PC1	0.979	0.638	0.392	0.977	0.673	0.624
	PC2	1.472	0.066	0.145	1.956	-0.117	0.199
	PC3	1.166	0.269	0.233	1.295	0.171	0.349
	PC4	1.078	-1.043	0.632	1.025	0.046	0.890
	PC5	1.032	0.354	0.288	1.115	-0.174	0.222
	PC6	1.103	-0.304	0.575	1.086	-0.633	0.488
	PC7	1.022	0.460**	0.161	1.040	0.367	0.406
	PC8	1.039	0.319	0.477	0.993	0.468	0.793
	PC9	1.576	0.006	0.119	1.994	-0.121	0.174
	PC10	1.264	0.146	0.205	1.426	0.119	0.246
PUXHS	PC1	1.000	0.497	0.458	0.992	0.616	0.814
	PC2	1.328	-0.022	0.218	1.586	-0.192	0.317
	PC3	1.113	0.200	0.310	1.105	0.238	0.522
	PC4	1.096	-0.989*	0.497	1.029	0.006	0.745
	PC5	1.083	-0.080	0.299	1.077	-0.412	0.357
	PC6	1.131	-1.074*	0.519	1.061	-1.317**	0.510
	PC7	1.001	0.498**	0.186	1.070	0.084	0.529
	PC8	1.094	-0.325	0.466	1.007	0.100	1.258
	PC9	1.394	-0.056	0.186	1.624	-0.204	0.290
	PC10	1.165	0.125	0.273	1.202	0.150	0.340
PUXX	PC1	0.866	1.432**	0.340	0.825	1.182**	0.120
	PC2	2.463	-0.120	0.072	3.257	-0.227*	0.093
	PC3	1.664	0.054	0.213	2.076	-0.063	0.275
	PC4	1.234	0.126	0.143	1.330	0.187	0.214
	PC5	1.024	0.460*	0.207	1.185	0.134	0.445
	PC6	1.005	0.479	0.477	0.916	1.009**	0.277
	PC7	1.074	0.381	0.277	1.089	0.293	0.500
	PC8	0.862	0.809**	0.297	0.767	1.127**	0.275
	PC9	2.485	-0.076	0.069	3.262	-0.168*	0.069
	PC10	1.873	0.079	0.136	2.562	0.038	0.150
PCE	PC1	1.053	0.029	0.469	1.088	-0.240	0.434
	PC2	1.698	-0.136	0.141	1.997	-0.240	0.223
	PC3	1.274	-0.031	0.280	1.407	-0.239	0.354
	PC4	1.027	0.343	0.392	1.031	0.339	0.535
	PC5	1.125	-0.080	0.327	1.215	-0.635	0.389
	PC6	1.053	0.035	0.484	1.020	0.272	0.508
	PC7	1.033	0.436*	0.175	1.116	0.033	0.334
	PC8	1.040	0.269	0.476	1.044	0.043	1.100
	PC9	1.518	-0.100	0.166	1.786	-0.282	0.258
	PC10	1.247	0.120	0.201	1.432	-0.069	0.235

We forecast annual inflation out-of-sample over 1985:Q4 to 2002:Q4 and over 1995:Q4 to 2002:Q4 at a quarterly frequency. Table 2 contains full details of the Phillips Curve models. The column labelled "Relative RMSE" reports the ratio of the RMSE relative to the ARMA(1,1) specification. The columns titled "1-λ" and "SE" report the coefficient $(1-\lambda)$ and its standard error, respectively, from equation (17). We denote values of $(1-\lambda)$ significantly different from zero at the 95% (99%) level by * (**), based on Hansen and Hodrick (1980) standard errors.

Table 6: Term Structure Forecasts of Annual Inflation

		Post-1985 Sample			Post-1995 Sample		
		Relative RMSE	$1-\lambda$	SE	Relative RMSE	$1-\lambda$	SE
PUNEW	TS1	1.096	0.137	0.332	1.030	0.362	0.410
	TS2	1.444	0.019	0.145	1.826	-0.147	0.229
	TS3	1.176	0.193	0.229	1.226	0.156	0.335
	TS4	1.166	-0.108	0.249	1.018	0.370	0.474
	TS5	1.134	0.088	0.186	1.122	0.006	0.187
	TS6	1.194	-0.241	0.326	1.112	-0.162	0.406
	TS7	1.091	0.308	0.252	1.039	0.373	0.434
	TS8	1.119	0.116	0.332	1.010	0.380	0.816
	TS9	1.363	0.086	0.085	1.229	-0.008	0.083
	TS10	1.196	-0.024	0.143	1.043	0.132	0.557
	TS11	1.198	-0.124	0.431	1.052	0.286	0.318
	VAR	1.415	0.287**	0.108	0.913	0.584**	0.202
	RGMVAR	1.647	0.050	0.050	1.518	-0.170	0.198
	MDL1	1.323	0.161*	0.064	1.345	-0.088	0.174
	MDL2	1.192	0.225	0.117	1.329	-0.118	0.251
PUXHS	TS1	1.080	-0.025	0.413	1.014	0.373	0.553
	TS2	1.345	-0.017	0.205	1.584	-0.197	0.328
	TS3	1.116	0.186	0.278	1.118	0.195	0.435
	TS4	1.085	-0.276	0.499	0.996	0.541	0.593
	TS5	1.113	-0.082	0.214	1.094	-0.191	0.264
	TS6	1.140	-0.566	0.342	1.069	-0.361	0.419
	TS7	1.081	0.161	0.298	1.070	0.088	0.409
	TS8	1.083	-0.054	0.411	0.975	0.558	1.057
	TS9	1.173	0.114	0.105	1.130	-0.123	0.211
	TS10	1.140	-0.595	0.468	1.032	-0.036	0.082
	TS11	1.102	-0.121	0.423	1.049	0.092	0.163
	VAR	1.802	0.175	0.096	0.973	0.523*	0.250
	RGMVAR	1.363	0.070	0.085	1.285	-0.149	0.366
	MDL1	1.225	0.127	0.081	1.186	-0.048	0.247
	MDL2	1.047	0.395	0.203	1.156	0.000	0.406
PUXX	TS1	0.945	0.667*	0.322	0.945	0.665*	0.317
	TS2	2.262	-0.092	0.084	2.982	-0.225*	0.099
	TS3	1.399	0.121	0.260	1.698	-0.057	0.344
	TS4	1.232	0.260	0.156	1.268	0.319	0.225
	TS5	1.081	0.392	0.203	1.258	0.085	0.407
	TS6	0.969	0.567	0.294	0.866	0.788**	0.078
	TS7	1.068	0.419*	0.203	1.118	0.342	0.289
	TS8	0.948	0.568**	0.197	0.958	0.520*	0.253
	TS9	1.372	0.050	0.239	1.282	-0.101	0.457
	TS10	1.034	0.433	0.284	1.208	-0.048	0.548
	TS11	1.017	0.474	0.246	1.192	0.099	0.502
	VAR	1.379	0.301*	0.123	1.762	-0.119	0.274
	RGMVAR	1.572	0.120	0.138	1.622	-0.211	0.340
	MDL1	1.506	0.253**	0.091	1.593	-0.004	0.280
	MDL2	1.833	0.262**	0.039	1.329	0.355**	0.069

Table 6 Continued

		Post-1985 Sample			Post-1995 Sample		
	Model	Relative RMSE	$1-\lambda$	SE	Relative RMSE	$1-\lambda$	SE
PCE	TS1	1.075	-0.073	0.453	1.078	-0.208	0.432
	TS2	1.670	-0.149	0.145	1.966	-0.247	0.226
	TS3	1.279	-0.053	0.288	1.374	-0.245	0.376
	TS4	1.075	0.017	0.372	1.059	0.234	0.442
	TS5	1.126	-0.115	0.331	1.202	-0.645	0.383
	TS6	1.094	-0.149	0.428	1.100	-0.359	0.397
	TS7	1.018	0.443	0.272	1.106	0.033	0.303
	TS8	1.027	0.373	0.414	1.025	0.345	1.056
	TS9	1.141	-0.024	0.192	1.121	-0.825	0.584
	TS10	1.087	-0.569	0.549	1.110	-0.851	0.639
	TS11	1.086	0.006	0.418	1.132	-0.396	0.288
	VAR	2.083	0.195^{*}	0.080	1.095	0.440^{**}	0.155
	RGMVAR	1.507	-0.242	0.131	1.461	-0.356	0.233
	MDL1	1.169	0.143	0.235	1.271	-0.374	0.284
	MDL2	1.314	-0.205	0.159	1.339	-0.331^{**}	0.120

We forecast annual inflation out-of-sample over 1985:Q4 to 2002:Q4 and over 1995:Q4 to 2002:Q4 at a quarterly frequency. Table 2 contains full details of the term structure models. The column labelled "Relative RMSE" reports the ratio of the RMSE relative to the ARMA(1,1) specification. The remaining columns report the coefficient $(1-\lambda)$ in equation (17) together with its standard error. We denote values of $(1-\lambda)$ significantly different from zero at the 95% (99%) level by * (**), based on Hansen and Hodrick (1980) standard errors.

Table 7: Survey Forecasts of Annual Inflation

		Post-1985 Sample			Post-1995 Sample		
		Relative RMSE	$1-\lambda$	SE	Relative RMSE	$1-\lambda$	SE
PUNEW	SPF1	0.779	1.051**	0.177	0.861	0.869*	0.407
	SPF2	0.964	0.564**	0.216	0.902	0.745*	0.377
	SPF3	0.976	0.541**	0.207	0.915	0.728	0.414
	LIV1	0.789	1.164**	0.102	0.792	1.140**	0.203
	LIV2	1.180	0.335	0.177	1.092	0.403	0.437
	LIV3	1.299	0.251	0.163	1.152	0.275	0.517
	MICH1	0.902	0.771*	0.324	0.862	1.113*	0.520
	MICH2	0.961	0.674*	0.327	0.930	0.861	0.644
	MICH3	0.968	0.655	0.347	0.947	0.776	0.653
PUXHS	SPF1	0.819	0.939**	0.171	0.914	0.772*	0.394
	SPF2	0.924	0.666**	0.227	0.888	0.825*	0.357
	SPF3	1.348	0.103	0.183	0.958	0.582	0.323
	LIV1	0.844	1.098**	0.099	0.856	1.072**	0.214
	LIV2	1.054	0.554**	0.176	1.031	0.550	0.366
	LIV3	1.299	0.327*	0.157	1.152	0.502	0.444
	MICH1	0.881	0.876**	0.273	0.937	0.749	0.434
	MICH2	0.918	0.814**	0.290	0.932	0.813	0.516
	MICH3	0.970	0.607*	0.251	0.953	0.684	0.492
PUXX	SPF1	0.691	0.968**	0.140	0.699	1.260**	0.225
	SPF2	1.145	0.125	0.362	1.104	0.091	0.852
	SPF3	1.179	0.035	0.373	1.180	-0.358	0.956
	LIV1	0.655	0.803**	0.193	0.557	1.227**	0.134
	LIV2	1.355	-0.185	0.177	1.387	-0.423	0.415
	LIV3	1.289	-0.095	0.259	1.278	-0.496	0.735
	MICH1	1.185	0.383*	0.159	0.822	1.041**	0.208
	MICH2	1.343	-0.153	0.248	1.566	-0.385	0.286
	MICH3	1.360	-0.242	0.253	1.617	-0.493	0.273
PCE	SPF1	1.199	0.147	0.267	1.250	0.090	0.395
	SPF2	0.980	0.537**	0.206	0.924	0.655*	0.325
	SPF3	1.034	0.453*	0.180	1.040	0.453	0.234
	LIV1	1.082	0.175	0.325	1.101	0.132	0.412
	LIV2	1.397	-0.050	0.189	1.303	-0.027	0.265
	LIV3	1.380	-0.123	0.149	1.341	-0.191	0.272
	MICH1	1.217	0.108	0.216	1.338	-0.030	0.327
	MICH2	1.194	0.039	0.253	1.205	0.055	0.415
	MICH3	1.248	-0.022	0.239	1.255	-0.003	0.399

We forecast annual inflation out-of-sample over 1985:Q4 to 2002:Q4 and from 1995:Q4 to 2002:Q4 at a quarterly frequency for the SPF survey (SPF1-3) and the Michigan survey (MICH1-3). The frequency of the Livingston survey (LIV1-3) is biannual and forecasts are made at the end of the second and end of the fourth quarter. Table 2 contains full details of the survey models. The column labelled "Relative RMSE" reports the ratio of the RMSE relative to the ARMA(1,1) specification. The remaining columns report the coefficient $(1-\lambda)$ in equation (17) together with its standard error. We denote values of $(1-\lambda)$ significantly different from zero at the 95% (99%) level by * (**), based on Hansen and Hodrick (1980) standard errors.

Table 8: Best Models in Forecasting Annual Inflation

	PUNEW		PUXHS		PUXX		PCE	
Panel A: Post-1985 Sample								
Best Time-Series Model	ARMA	1.000	ARMA	1.000	ARMA	1.000	ARMA	1.000*
Best Phillips-Curve Model	PC1	0.979	PC1	1.000	PC8	0.862	PC4	1.027
Best Term-Structure Model	TS7	1.091	MDL2	1.047	TS1	0.945	TS7	1.018
Raw Survey Forecasts	SPF1	0.779*	SPF1	0.819*	SPF1	0.691	SPF1	1.199
	LIV1	0.789	LIV1	0.844	LIV1	0.655*	LIV1	1.082
	MICH1	0.902	MICH1	0.881	MICH1	1.185	MICH1	1.217
Panel B: Post-1995 Sample								
Best Time-Series Model	RGM	0.764*	RGM	0.833*	RW	0.915	ARMA	1.000*
Best Phillips-Curve Model	PC1	0.977	PC1	0.992	PC8	0.767	PC6	1.020
Best Term-Structure Model	VAR	0.913	VAR	0.973	TS6	0.866	TS8	1.025
Raw Survey Forecasts	SPF1	0.861	SPF1	0.914	SPF1	0.699	SPF1	1.250
	LIV1	0.792	LIV1	0.856	LIV1	0.557*	LIV1	1.101
	MICH1	0.862	MICH1	0.937	MICH1	0.822	MICH1	1.338

The table reports the best time-series model, the best OLS Phillips Curve model, the best model using term structure data, along with SPF1, LIV1, and MCH1 forecasts for out-of-sample forecasting of annual inflation at a quarterly frequency. Each entry reports the ratio of the model RMSE to the RMSE of an ARMA(1,1) forecast. Models with the smallest RMSEs are marked with an asterisk.

Table 9: Ex-Ante Combined Forecasts of Annual Inflation

	Model Combination Method	Pure Time-Series	Phillips Curve	Term Structure	Surveys	All Models
PUNEW	Mean	0.888	1.123	0.991	0.851	0.973
	Median	0.907	1.093	1.047	0.851	1.031
	OLS	0.971	1.007	0.858	0.858	0.827
	Equal Weight Prior	0.955	1.007	0.862	0.858	0.820
	Best Individual Model	0.764	0.977	0.913	0.861	0.764
PUXHS	Mean	0.947	1.065	0.951	0.921	0.966
	Median	0.935	1.083	1.040	0.921	1.036
	OLS	0.962	1.001	0.937	0.917	0.859
	Equal Weight Prior	0.950	1.008	0.931	0.918	0.867
	Best Individual Model	0.833	0.992	0.973	0.914	0.833
PUXX	Mean	0.926	1.547	1.289	0.719	1.252
	Median	0.985	1.167	1.215	0.719	1.078
	OLS	0.881	0.885	1.104	0.699	0.821
	Equal Weight Prior	0.845	0.878	1.092	0.699	0.789
	Best Individual Model	0.915	0.767	0.866	0.699	0.699
PCE	Mean	1.012	1.160	1.031	1.285	1.070
	Median	1.020	1.136	1.106	1.285	1.114
	OLS	1.028	0.974	0.988	1.288	0.955
	Equal Weight Prior	1.035	0.984	0.983	1.287	0.961
	Best Individual Model	1.000	1.020	1.025	1.250	1.000

The table reports the relative RMSEs for forecasting annual inflation at a quarterly frequency out-of-sample from 1995:Q4 to 2002:Q4 by combining models within each category or over all models. Forecasts reported include the mean and median forecasts, and linear combinations of forecasts using recursively-computed weights computed from OLS, or model combination regressions using equally-weighted priors. We consider unadjusted SPF and Michigan survey forecasts only in the survey category. For comparison, the last row in each panel reports the relative RMSE of the best performing single forecast model (see Table 8).

Table 10: Ex-Post Combined Forecasts of Annual Inflation

	Model Combination Method	Time-Series	Phillips Curve	Term Structure	Surveys	Best Models	All Models
Panel A: Post-1985 Sample							
PUNEW	OLS	0.969	0.937	0.834	0.775	0.773	0.873
	Equal Weight Prior	0.968	0.940	0.835	0.775	0.773	0.873
	Unit Weight Prior	0.973	0.943	0.847	0.775	0.772	0.874
	Best Individual Model	1.000	0.979	1.091	0.779	0.779	0.779
PUXHS	OLS	0.965	0.946	0.883	0.816	0.809	0.864
	Equal Weight Prior	0.965	0.953	0.885	0.816	0.810	0.865
	Unit Weight Prior	0.972	0.954	0.903	0.816	0.809	0.864
	Best Individual Model	1.000	1.000	1.047	0.819	0.819	0.819
PUXX	OLS	0.914	0.834	0.857	0.691	0.698	0.819
	Equal Weight Prior	0.911	0.832	0.856	0.692	0.698	0.819
	Unit Weight Prior	0.921	0.835	0.860	0.691	0.697	0.820
	Best Individual Model	1.000	0.862	0.945	0.691	0.691	0.691
PCE	OLS	0.978	0.924	0.852	1.157	0.937	0.962
	Equal Weight Prior	0.979	0.930	0.853	1.157	0.937	0.960
	Unit Weight Prior	0.979	0.932	0.861	1.158	0.938	0.962
	Best Individual Model	1.000	1.027	1.018	1.199	1.000	1.000
Panel B: Post-1995 Sample							
PUNEW	OLS	0.743	0.931	0.676	0.862	0.725	0.722
	Equal Weight Prior	0.746	0.934	0.676	0.851	0.730	0.726
	Unit Weight Prior	0.743	0.933	0.679	0.855	0.747	0.735
	Best Individual Model	0.764	0.977	0.913	0.861	0.764	0.764
PUXHS	OLS	0.833	0.934	0.698	0.914	0.735	0.735
	Equal Weight Prior	0.843	0.943	0.714	0.919	0.743	0.743
	Unit Weight Prior	0.833	0.940	0.712	0.914	0.767	0.759
	Best Individual Model	0.833	0.992	0.973	0.914	0.833	0.833
PUXX	OLS	0.801	0.786	0.802	0.695	0.659	0.702
	Equal Weight Prior	0.793	0.781	0.787	0.696	0.658	0.692
	Unit Weight Prior	0.794	0.767	0.788	0.695	0.656	0.693
	Best Individual Model	0.915	0.767	0.866	0.699	0.699	0.699
PCE	OLS	1.018	0.949	0.631	1.250	1.005	1.005
	Equal Weight Prior	1.007	0.950	0.635	1.257	0.993	0.993
	Unit Weight Prior	1.000	0.967	0.687	1.250	1.007	1.006
	Best Individual Model	1.000	1.020	1.025	1.250	1.000	1.000

The table reports the relative RMSEs for forecasting annual inflation out-of-sample at a quarterly frequency over 1985:Q4 to 2002:Q4 and 1995:Q4 to 2002:Q4 by combining models within each category (time-series, Phillips curve, term structure, surveys), using the best models in each category, or over all models. We compute the model weights using the full sample by OLS, a mixed regression with an equal-weight prior, and a mixed regression with a unit-weight prior placed on the best model. We consider unadjusted SPF and Michigan survey forecasts only in the survey category. For comparison, the last row in each panel reports the relative RMSE of the best performing individual forecasting model (see Table 8).

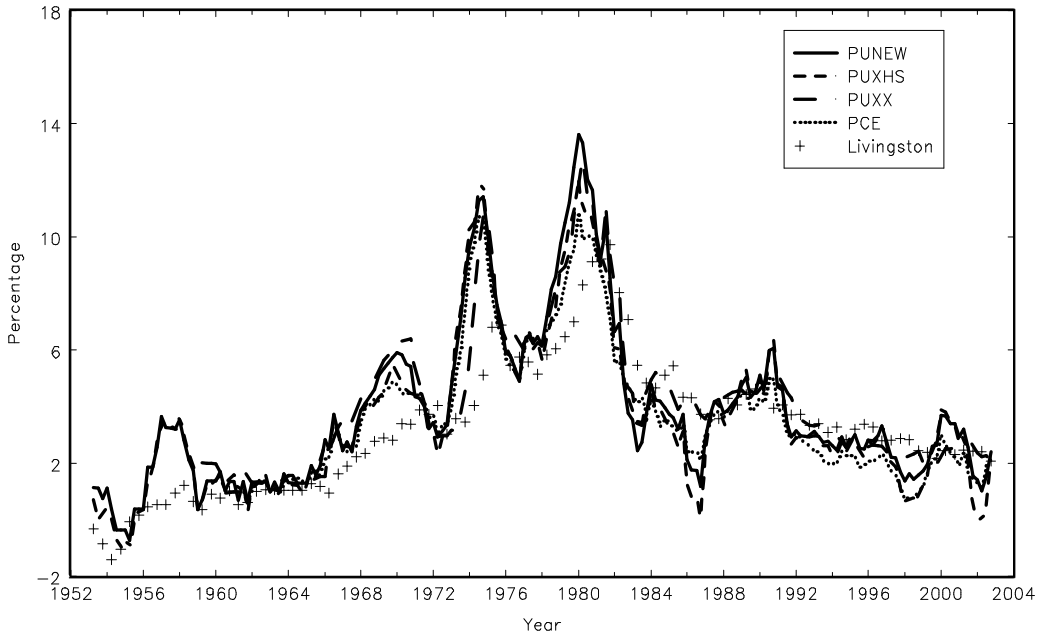
Table 11: Best Models in Forecasting Annual Inflation Changes

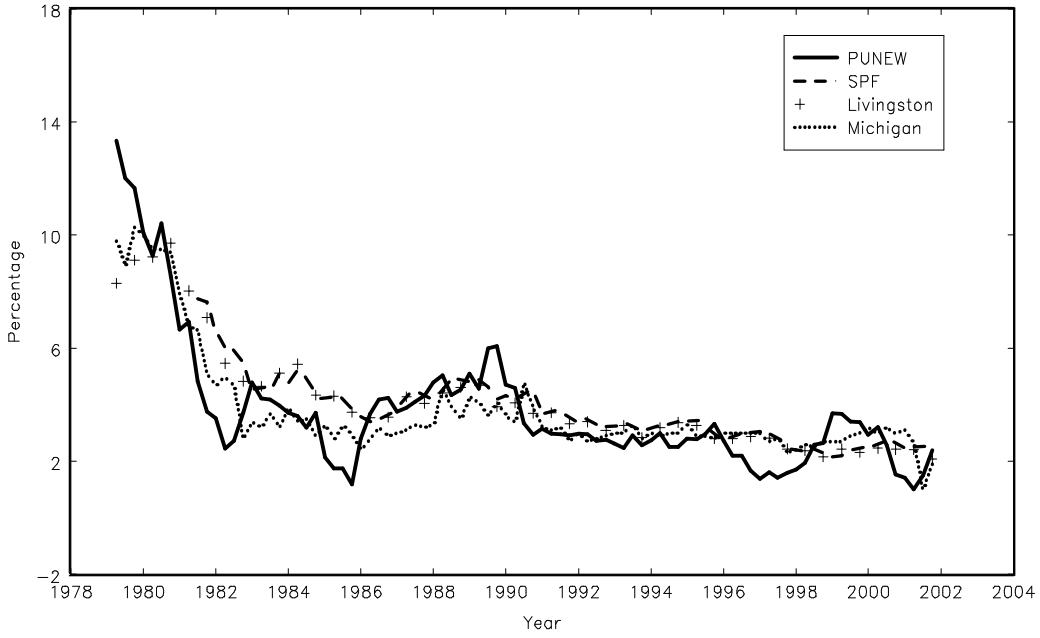
	Post-1985 Sample				Post-1995 Sample			
	Estimated on Levels		Estimated on Differences		Estimated on Levels		Estimated on Differences	
	Model	RMSE	Model	RMSE	Model	RMSE	Model	RMSE
PUNEW								
Best Time-Series Model	ARMA	1.000	ARMA	1.071	RGM	0.764*	ARMA	1.025
Best Phillips-Curve Model	PC1	0.979	PC7	1.005	PC1	0.977	PC7	0.976
Best Term-Structure Model	TS7	1.091	TS7	1.023	VAR	0.913	TS1	0.968
Raw Survey Forecasts	SPF1	0.779*			SPF1	0.861		
	LIV1	0.789			LIV1	0.792		
	MICH1	0.902			MICH1	0.862		
PUXHS								
Best Time-Series Model	ARMA	1.000	ARMA	1.098	RGM	0.833*	ARMA	1.046
Best Phillips-Curve Model	PC1	1.000	PC7	1.027	PC1	0.992	PC1	1.023
Best Term-Structure Model	MDL2	1.047	TS7	1.004	VAR	0.973	TS7	0.987
Raw Survey Forecasts	SPF1	0.819*			SPF1	0.914		
	LIV1	0.844			LIV1	0.856		
	MICH1	0.881			MICH1	0.937		

Table 11 Continued

	Post-1985 Sample				Post-1995 Sample			
	Estimated on Levels		Estimated on Differences		Estimated on Levels		Estimated on Differences	
	Model	RMSE	Model	RMSE	Model	RMSE	Model	RMSE
PUXX								
Best Time-Series Model	ARMA	1.000	ARMA	0.837	RW	0.915	ARMA	0.649
Best Phillips-Curve Model	PC8	0.862	PC1	0.722	PC8	0.767	PC1	0.652
Best Term-Structure Model	TS1	0.945	TS8	0.861	TS6	0.866	TS6	0.655
Raw Survey Forecasts	SPF1	0.691			SPF1	0.699		
	LIV1	0.655*			LIV1	0.557*		
	MICH1	1.185			MICH1	0.822		
PCE								
Best Time-Series Model	ARMA	1.000	ARMA	1.029	ARMA	1.000	ARMA	1.004
Best Phillips-Curve Model	PC4	1.027	PC8	0.978	PC6	1.020	PC6	1.019
Best Term-Structure Model	TS7	1.018	TS8	0.945*	TS8	1.025	TS4	0.951*
Raw Survey Forecasts	SPF1	1.199			SPF1	1.250		
	LIV1	1.082			LIV1	1.101		
	MICH1	1.217			MICH1	1.338		

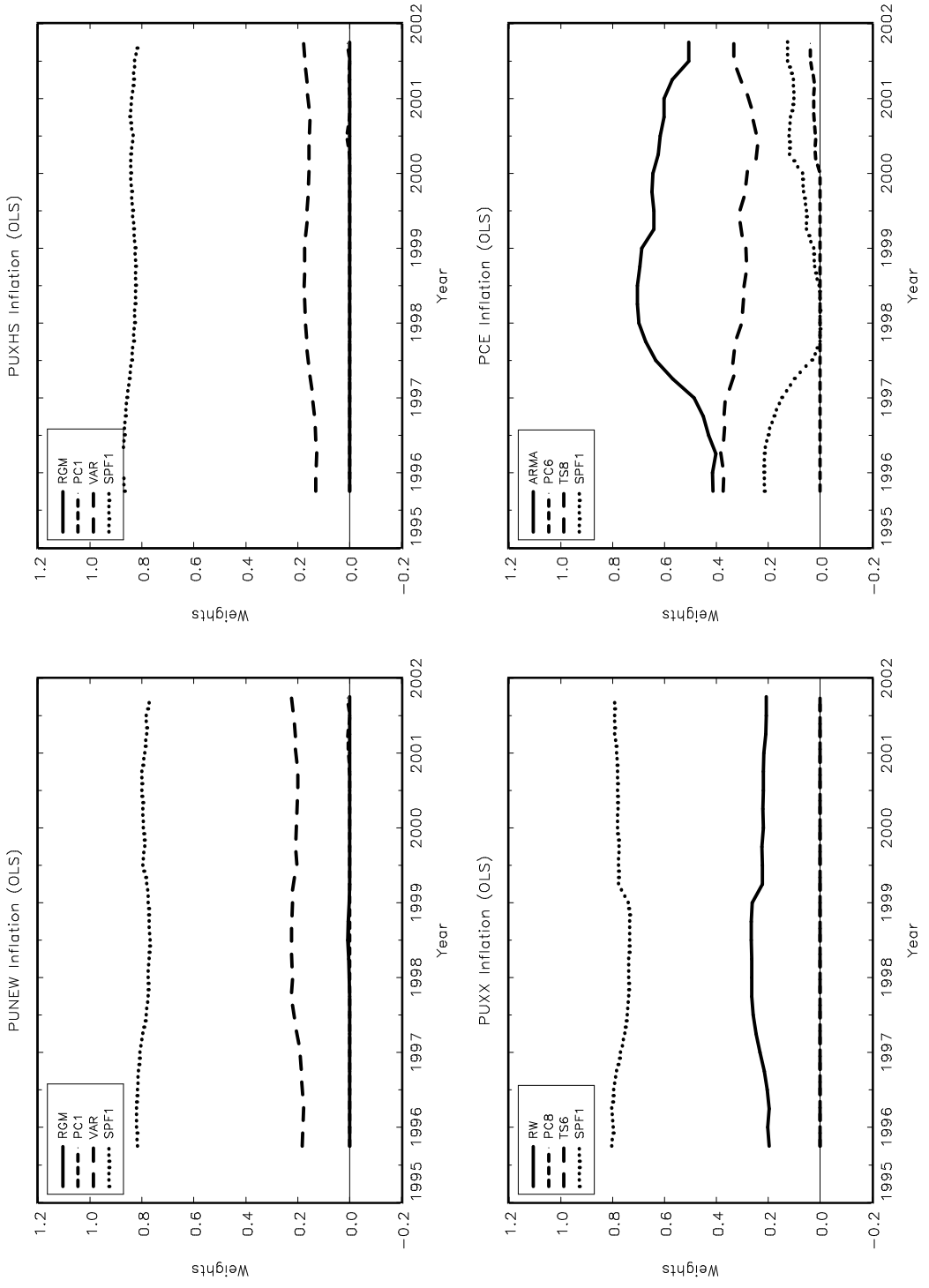
This table reports the relative RMSE of the best performing out-of-sample forecasting model in each of the first three categories of models (time-series, Phillips Curve, and term structure models) and those of raw surveys. The models are estimated in either inflation levels of inflation differences. Table 2 contains full details of all the forecasting models. We report the RMSE ratios relative to an ARMA(1,1) specification estimated on levels. Models with the smallest RMSEs are marked with an asterisk.





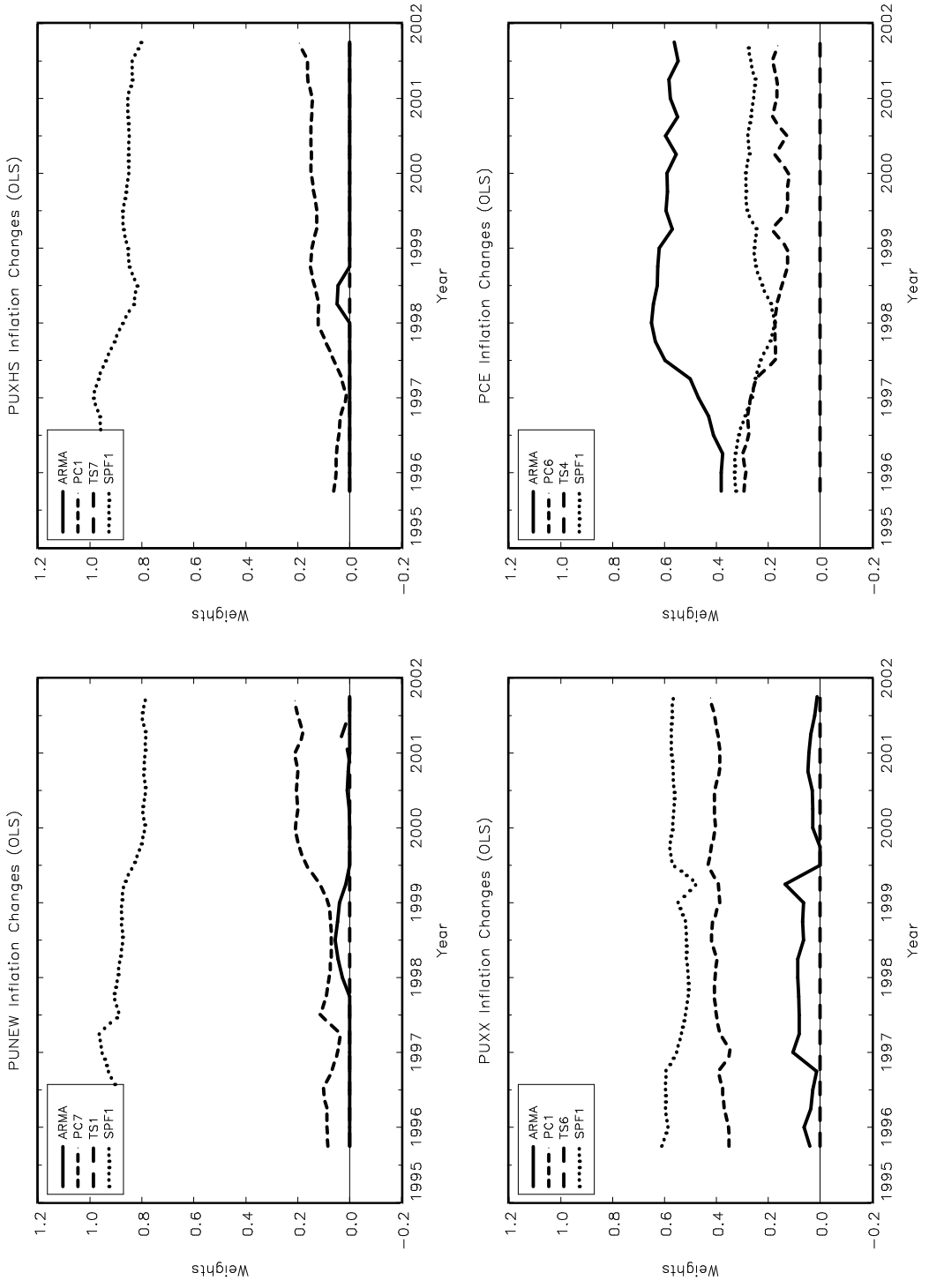
In the top panel, we graph the four inflation measures *PUNEW*, CPI-U All Items; *PUXHS*, CPI-U Less Shelter; *PUXX*, CPI-U All Items Less Food and Energy, or core CPI; and *PCE*, the Personal Consumption Expenditure deflator, together with the Livingston survey forecast. The survey forecast is lagged one year, so that in December 1990, we plot inflation from December 1989 to December 1990 together with the survey forecasts at December 1989. In the bottom panel, we plot all three survey forecasts (SPF, Livingston, and the Michigan surveys), together with PUNEW inflation. The survey forecasts are also lagged one year for comparison.

Figure 1: Annual Inflation and Survey Forecasts



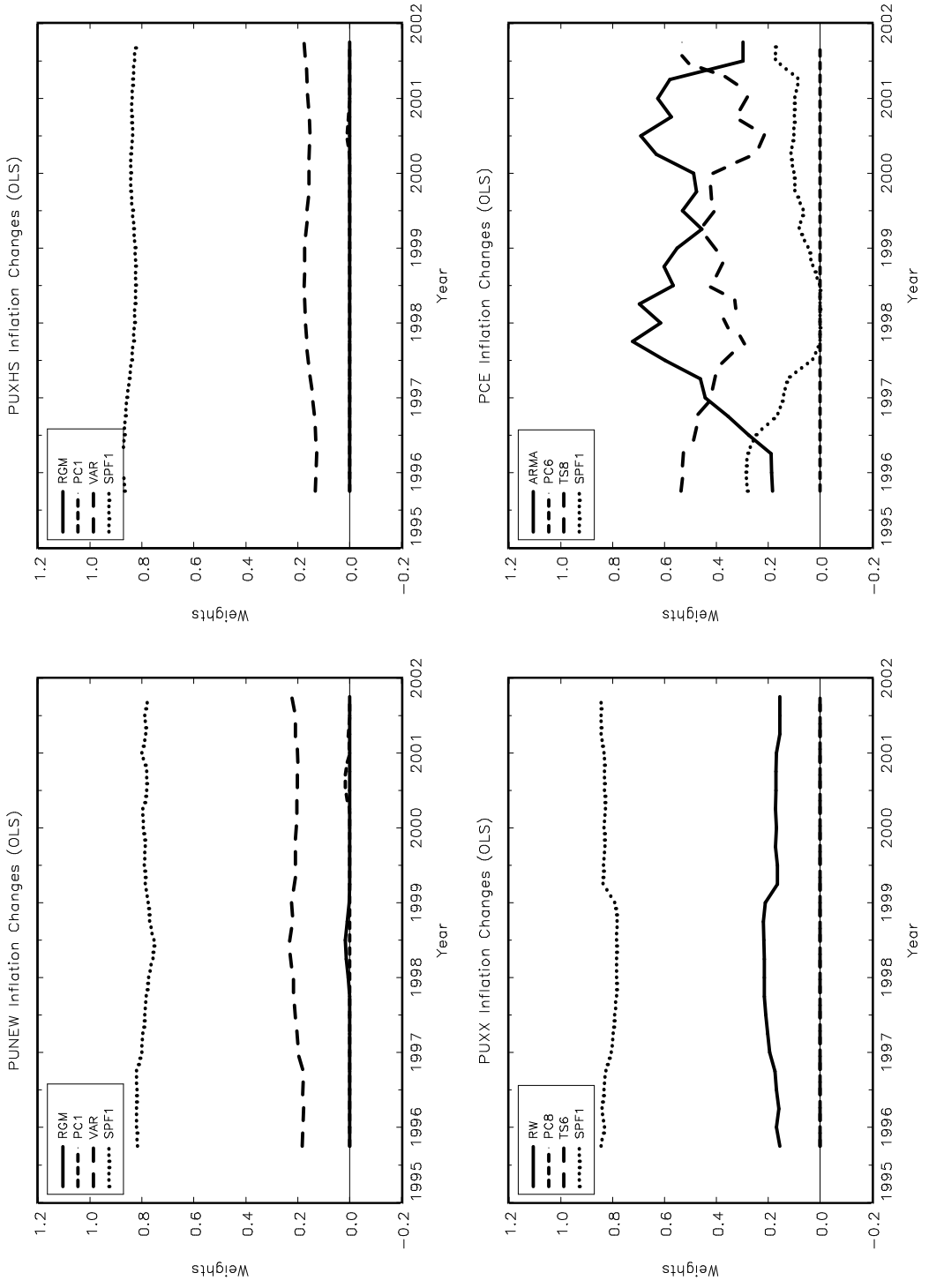
We graph the ex-ante OLS weights on models from regression (18) over the period 1995:Q4 to 2002:Q4. We combine the best model within each category (time-series, Phillips Curve, term structure, and survey) from Table 9. The ex-ante weights are computed recursively through the sample.

Figure 2: Ex-Ante Weights on Best Models for Forecasting Annual Inflation



We graph the ex-ante OLS weights on models from regression (22) over the period 1995:Q4 to 2002:Q4. We combine the best non-stationary model within the time-series, Phillips Curve, and term structure classes together with the raw SPF forecast. The ex-ante weights are computed recursively through the sample.

Figure 3: Ex-Ante Weights on Best I(1) Models for Forecasting Annual Inflation Changes



We graph the ex-ante OLS weights on models from regression (22) over the period 1995:Q4 to 2002:Q4. We combine the best stationary model within the time-series, Phillips Curve, and term structure classes together with the raw SPF forecast. The ex-ante weights are computed recursively through the sample.

Figure 4: Ex-Ante Weights on Best I(0) Models for Forecasting Annual Inflation Changes