Improving Forecasts in Real Time

By Dean Croushore*

August 5, 2024

In this paper, I suggest a method for improving upon forecasts of macroeconomic variables. I consider forecasts that have been found to be biased or inefficient in-sample, and show how to improve them. Two major complications include data revisions and structural instability. I illustrate the ideas with tests for bias and inefficiency in expectations of inflation, measured using the Survey of Professional Forecasters.

JEL: E37, E17
Keywords: real-time data, inflation forecasts, forecast improvement

* University of Richmond, Robins School of Business, dcrousho@richmond.edu.

Economists are constantly looking for stylized facts. One of the most important stylized facts that economists have tried to establish (or disprove) is that forecasts are rational. The theory of rational expectations depends on it, yet the evidence is mixed. Whether a set of forecasts is found to be rational or not seems to depend on many things, including the sample, the source of data on the expectations being examined, and the empirical technique used to investigate rationality.

Early papers in the rational-expectations literature used surveys of expectations, such as the Livingston Survey and the Survey of Professional Forecasters (SPF), to test whether the forecasts made by professional forecasters were consistent with the theory. A number of the tests in the 1970s and 1980s cast doubt on the rationality of the forecasts, with notable results by Su and Su (1975) and Zarnowitz (1985). But later results, such as Croushore (2010), found no bias over a longer sample. Similarly, some papers found inefficiency in the forecasts, such as Ball and Croushore (2003), Rudebusch and Williams (2009), and Coibion and Gorodnichenko (2015). The question is: is the bias or inefficiency they found exploitable in real time?

In this paper, I develop methods for how, in principle, to improve upon a forecast that is biased or inefficient. The two practical considerations that make forecast improvement difficult to determine are: (1) Is the variable subject to data revisions? (2) Did the bias or inefficiency arise because of a structural change that forecasters did not anticipate?

I. Theory

Suppose we have a set of forecasts generated by a forecaster, or from a survey of forecasters, and we wish to investigate whether the forecasts have desirable properties. We can calculate the forecast errors over time, and test them to see if they are unbiased and efficient, as discussed by Elliott and Timmermann (2008). The forecast error at each forecast date t is:

$$e_{t,h} = Y_{t+h} - Y^f_{t,h}, \tag{1}$$

where Y_{t+h} is the realized value of the variable being forecasted, and $Y^f_{t,h}$ is the forecast made at date t for the variable Y at time $t+h$.

Bias can be tested by regressing the forecast errors on a constant:

$$e_{t,h} = k^1_h + \epsilon^1_{t,h}. \tag{2}$$

The test for unbiasedness comes from testing the null hypothesis that $k^1_h = 0$, for each horizon h.

An alternative is to test for efficiency, which examines the correlation between the forecast errors and a variable that was known contemporaneously at each forecast date. A typical test to see if variable Z is correlated with the forecast error would be to run the regression:

$$e_{t,h} = k^2_h + \gamma_h Z_{t-j} + \epsilon^2_{t,h}, \tag{3}$$

where the lag j on variable Z depends on data availability. Inefficiency only occurs if data available at the time a forecast was made is correlated with the forecast error, so Z_{t-j} must be in the information set of the forecaster when the forecast is made at date t.

The point of departure for this paper comes from the question of how to empirically implement a finding of bias or inefficiency. As Elliott and Timmermann (2008) note, a finding of bias or inefficiency suggests "that improved forecasts are possible given the available data." (p. 34) I develop several methods to test the extent to which, in practical circumstances, it is possible to improve upon forecasts.

Improving Upon a Biased Forecast. If a forecast is biased, we can estimate Equation (2) and use the regression results to improve the forecast out of sample. So, if we have an information set, Ω_{T-1}, with data on variable Y from date $T-s$ to date $T-1$, we can forecast out of sample using the equation

$$Y^I_{T,h} = Y^f_{T,h} + \hat{k}^1_h, \tag{4}$$

where the superscript I stands for "improved", though more precisely, we should perhaps say "potentially improved."

Improving Upon an Inefficient Forecast. If a forecast is inefficient, we can estimate Equation (3) and use the regression results to try to improve the forecast out of sample, using the equation

$$Y^I_{T,h} = Y^f_{T,h} + \hat{k}^2_h + \hat{\gamma}_h Z_{T-j}. \tag{5}$$

Testing Improvement. Suppose we test a set for forecasts for bias by estimating Equation (2) and generate improved forecasts using Equation (4). Alternatively, we test for inefficiency by estimating Equation (3) and generate improved forecasts using Equation (5). In both cases, suppose we run the bias or inefficiency tests at the start of each quarter, and repeat the same exercise over time. Of course, as we roll over time, the estimated coefficients in Equations (2) and (3) change.

Does the attempt to improve upon the forecasts work? We can test the original forecast with the "improved" forecast using a standard Diebold and Mariano (1995) test, as modified by Harvey, Leybourne and Newbold (1997).

In a typical application, rather than running these tests and trying to improve the forecasts in real time, which might take many years, a researcher might instead opt to consider forecasts from a forecaster or from a survey over a period of time, simulating how a researcher might test for bias or inefficiency over time.

For example, I might want to test if there is bias in the Survey of Professional Forecasters' forecasts of inflation. I could take a first sample, say SPF surveys from 1971Q1 to 1975Q4, estimate the bias using Equation (2), and make an improved forecast for 1976Q1. Then roll both dates forward one quarter at a time (both the end date of the sample and the forecast date). Finally, gather the simulated forecasts from 1976Q1 to 2023Q4 and test them against the original SPF survey forecasts to see which is more accurate.

Two Difficult Issues: Data Revisions and Instability. The methods described above are difficult to complete satisfactorily because of two problems. First, data may be revised, so what is a researcher to consider to be the realized value of the variable from which to compute the forecast error? And what relationship between data with different degrees of revision should a researcher use? Second, bias or inefficiency might not occur over the entire sample because of structural instability in the data-generating process or in the forecasting process.

Data Revisions. To test for bias and efficiency requires data on the realized value of the variable being forecast. But as Croushore (2011) and others have noted, data may be revised substantially. So, what value does a researcher use as the realized value in Equation (1)? There is no right answer to that question because data may be revised forever. So, researchers often make a choice of one particular concept of the vintage of data they use, and seldom check the robustness of that choice. But what if data appear biased using one concept, but not biased using others? What if forecasts can be improved using one concept, but not using others? And, what can a researcher do if the data-generating process is different between data that have been recently released compared to those that have been revised multiple times based on different source data used by the government statistical agency?

So, researchers must make a choice about what to assume about how data revisions affect the data-generating process. There are two main choices: looking at data at the end of the sample (EOS) or using real-time vintages (RTV). I

consider two different EOS approaches: naive and benchmark-consistent. A traditional data-generating process suggests the use of an EOS approach, while a revision-based data-generating process suggests the use of the RTV approach. To explain those approaches more clearly, here is a description of various measures of realized values.

If revisions to the data were small and white noise, the use of different concepts for realized values would be inconsequential.[1] But the literature on real-time data analysis suggests that the revisions are neither small nor innocuous. I consider six different concepts for realized values for all National Income and Product Account (NIPA) data: (1) the initial release, which comes out at the end of the first month following the end of a quarter; (2) the first revision, which occurs one month after the initial release; (3) the first-final release, also called the second revision, which comes out at the end of the third month following the end of a quarter; (4) the first annual release, which is usually produced each year at the end of July and usually includes revisions to data from the prior three calendar years; (5) the pre-benchmark release, which is the last release of the data prior to a benchmark revision that makes major changes in the data construction process; and (6) the last release, which is the most recent vintage of the data at the time of writing this paper, which incorporates many benchmark revisions.[2] In years in which a benchmark revision occurs, such as 2003, there is often no annual revision, so I take the benchmark revision of the data as the annual release. The pre-benchmark release is an important concept because it shows the last data following a consistent methodology. For example, before 1996, macroeconomic forecasters all based their forecasts on fixed-weighted GDP. But in early 1996, when the government introduced chain-weighted GDP in a benchmark revision, the entire past history of GDP changed substantially. A forecaster who made a forecast

[1] The assumption that data revisions were trivial and not worth considering was common prior to the development of the real-time datasets described below. That assumption was convenient but not correct.

[2] I use the date March 2024 in this paper; it corresponds to the vintage of April 2024 in the Philadelphia Fed's Real-Time Data Set for Macroeconomists (RTDSM), the timing of which is in the middle of the month. So, the data released at the end of March 2024 are recorded in the April vintage of the RTDSM.

of GDP growth in 1994 would not have produced forecasts of chain-weighted GDP, so it seems appropriate to compare those forecasts to the last release of the data, in the pre-benchmark release, containing fixed-weighted GDP. As another example, it is difficult to imagine that a forecaster in 1971 would account for the future change of the output concept to include intellectual property products, which caused GDP for most periods to be revised up after the benchmark revision of July 2013, when the concept of intellectual property products was introduced. For complete details on these concepts and the revision process, see Croushore (2011).[3]

Because there is no clear best vintage of data to use in empirical exercises, some researchers, such as Zarnowitz (1985), prefer to use a concept like the pre-benchmark release, while others, such as Croushore (2019), focus on the first annual revision. Others prefer to use the first-final (third) release, such as Romer and Romer (2000) and Rudebusch and Williams (2009). The real-time literature has shown that some empirical results are sensitive to the choice of concept to use as the realized value. The Appendix to this paper provides precise definitions and notation for the realized values.

In addition to the choice of realized values, different vintages may need to be used to get an accurate portrayal of the data-generating process. Most prominently, Kishor and Koenig (2012) show that the correct relationship across vintages may depend on the vintage concept; for example, the sequence of initial releases may have a separate data-generating process than later releases of the data. I explore three different possibilities in this paper: EOS-naive, EOS benchmark-consistent, and RTV.

EOS-Naive. A researcher gathers data in an information set that would have existed at each point in time in the out-of-sample evaluation period and uses it assuming a particular equation describes the data-generating process. For

[3]The Appendix shows the dates of both first-annual revisions in Table A1 and benchmark revisions in Table A2.

example, suppose a forecaster in the SPF is forecasting inflation, using the full data set available for inflation at each date in real time. Suppose we wish to evaluate forecasts made at each quarterly date, starting in 1971Q1, then rolling forward one quarter at a time. So, the researcher would assume the forecaster is generating forecasts with a sequence of data sets, pulled from a data source like FRED at each date, which would be exactly the data set known to SPF forecasters for each survey. I call this sequence of data sets "EOS-naive" because it ignores data revisions completely. This would be a reasonable approach if forecasters indeed paid no attention to the revision process and just used the same forecasting model with the most recent data available to them. The problem for a researcher is that this sequence of data sets assumes that revisions, in particular benchmark revisions to the data, are innocuous and do not change the data-generating process.

EOS Benchmark-Consistent. Benchmark revisions may change the data-generating process. Croushore and Stark (2001) show that the revision process cannot possibly be represented in a mathematically convenient ARIMA process, which means we cannot simply add a measurement equation to a state equation for forecasting. Benchmark revisions often redefine variables, especially real GDP and other NIPA variables, thus distorting the data-generating process. At the same time, recognizing the value of additional source data is important, so the ideal vintage to use for evaluating forecasts is the pre-benchmark release, which is the last vintage before a benchmark revision. The idea is that forecasters make their forecasts using a data series based on current statistical methodologies, and do not know how benchmark revisions might redefine the data. Even if they did (as in the switch from fixed-weighting to chain-weighting in 1996), the Bureau of Economic Analysis usually does not release past values under the new methodology until the benchmark release date, so forecasters have no choice but to use the older methodology for their forecasts.

RTV. Under the RTV (real-time vintage) approach, as proposed initially by

Koenig, Dolmas and Piger (2003) and expanded upon by Kishor and Koenig (2012), the data-generating process is most accurately described as a relationship between data that have been revised to similar extents. So, the RTV approach says that an appropriate model to use in one in which data that has not yet gone through an annual revision follows one data-generating process, while data that has been revised many times may follow a very different process. Under the RTV approach, for example, a researcher might argue that the initial vintage of the data should be used to evaluate forecasts and assume that forecasters do not use EOS vintages in forming forecasts, but rather they divide data into vintages of different "maturities".

Instability. A second difficult issue is that the forecasts might be unbiased and/or efficient for some period of time, but a structural shift might occur that the forecaster does not understand immediately. This may cause a string of forecast errors for a period of time until the forecaster begins to understand it and improve the forecasting method. These issues are addressed in research most notably by Barbara Rossi and coauthors: Rossi and Sekhposyan (2010), Rossi and Sekhposyan (2016), Rossi (2021). They develop a number of tests for instability in forecasts. The empirical question I try to answer is, does identifying such periods help us to improve forecasts?

Suppose, for example, that a forecaster estimates a forecasting model based on the equation:

$$Y_t = \alpha + \beta y_t + \epsilon_t. \tag{6}$$

But suppose the true data generating process is

$$Y_t = \alpha_t + \beta_t y_t + \epsilon_t. \tag{7}$$

Time variation in either the α or β terms will lead to apparent bias or inconsis-

tency in the forecasts based on Equation (6). I will use the fluctuation-rationality tests of Rossi and Sekhposyan (2016) to investigate whether they can be used to improve the forecasts.

Putting both the stability question and analysis of data revisions together, Croushore (2010) found substantial instability across subsamples in evaluations of survey forecasts of inflation in a manner similar to that found by Giacomini and Rossi (2010) for model forecasts of exchange rates. No global stylized facts appear to hold. Forecasters go through periods in which they forecast well, then there is a deterioration of the forecasts, and then they respond to their errors and improve their models, leading to lower forecast errors again. This pattern may explain why Stock and Watson (2003) find that many variables lose their predictive power as leading indicators. Perhaps parameters are changing in economic models, as Rossi (2006) suggests for models of exchange rates.

I begin by looking at bias, investigating bias tests in real time, accounting for instability, such as that shown in Equation (7), where α_t is changing over time. Then, I look at inefficiency, investigating different tests for inefficiency in real time, accounting for instability, such as the shown in Equation (7), where β_t is changing over time.

This analysis is unique in two aspects. First, it is one of few analyses to compare and contrast forecast evaluations using the end-of-sample (EOS) and the real-time vintage (RTV) approaches. Second, it is the only paper to use and compare EOS and RTV results, along with the fluctuation-rationality test of Rossi and Sekhposyan (2016), in the context of forecast-improvement exercises.

II. An Illustration of Testing for Bias in Real Time

Data. I examine the U.S. inflation rate, based on the implicit price deflator for output, which is the longest-running series on inflation. I handle the complication of data revisions by using the real-time data set of Croushore and Stark (2001). Data are available from data vintages beginning in the third quarter of 1965,

when quarterly real output was reported for the first time on a regular basis by the U.S. Bureau of Economic Analysis.[4]

To study the ability of forecasters to provide accurate forecasts, I use the Survey of Professional Forecasters (SPF), which records the forecasts of a large number of private-sector forecasters.[5] The literature studying the SPF forecasts has found that the SPF forecasts outperform macroeconomic models, even fairly sophisticated ones, as shown by Ang, Bekaert and Wei (2007). The SPF has also been found to influence household expectations, as shown by Carroll (2003).

While some arguments can be made that testing bias is best done by examining the forecasts of individual forecasters,[6] a more compelling argument is that the most accurate forecasts are provided by taking the mean across the forecasters, as illustrated by Aiolfi, Capistran and Timmermann (2011). An additional problem with using the forecasts of individual forecasters is that the SPF survey has many missing observations, which is problematic. Data on mean and median forecasts of the deflator are reported in the SPF beginning with the fourth quarter of 1968.[7] However, the deflator forecasts in the early years of the survey were not reported to enough significant digits, and four-quarter-ahead forecasts were sometimes not reported in the early years of the survey. To avoid these problems, I begin the analysis using surveys beginning from the first quarter of 1971.

There are many horizons for the SPF, and in this section I choose to focus on the longest forecasting horizon that is consistently available in the survey, which is the average inflation rate over the next year (four quarters). The one-year-ahead forecast is subject to less noise and presumably more economic causes than would be the case for studying the forecasts for a particular quarterly horizon.

I begin by looking at the forecasts and forecast errors in Figure 1. The fig-

[4]See the documentation on the Federal Reserve Bank of Philadelphia Real-Time Data Set for Macroeconomists at `www.philadelphiafed.org/research-and-data/real-time-center/`.

[5]Details on the SPF can be found in Croushore and Stark (2019).

[6]See Keane and Runkle (1990).

[7]The mean and median across forecasters are almost identical in the SPF. This paper reports results based on the mean forecast but all tests reported in the paper have also been done with median forecasts, with no material differences in results.

ure is based on using the initial data release as the realized value; of course, other concepts of the realized value could be used. The figure shows some periods of persistent forecast errors, especially in the 1970s, but also at other times. However, this persistence is overstated by the figures because of the overlapping-observations problem: we are observing the forecasts quarterly, but they are four quarters ahead from the forecast date, and five quarters ahead of the last observation in the forecasters' data set. The overlapping-observations problem leads to the correlation of forecast errors. In the empirical work, I will use standard techniques to overcome this problem, adjusting the variance-covariance matrix using techniques developed by Newey and West (1987).

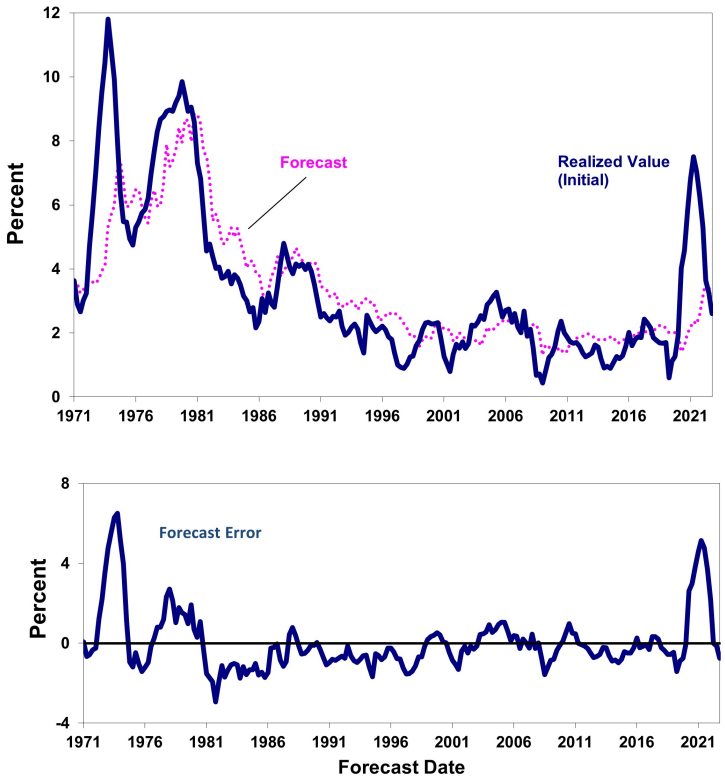
A. Results of Tests for Unbiasedness over Full Sample

In this paper, our focus is on tests for the unbiasedness of forecasts. In the literature on forecast bias, the standard test is the Mincer and Zarnowitz (1969) test, which regresses realized values on forecasts. However, the Mincer-Zarnowitz test may be inaccurate in small samples, as Mankiw and Shapiro (1986) show. Because I am using small samples, and because some of the tests I perform will be sensitive to parameter uncertainty, I modify the test for unbiasedness to a simpler version, which tests whether the forecast error has a mean of zero.[8]

I run the zero-mean-forecast-error test for inflation using all six versions of realized values. The results of this exercise are shown in Table 1. In each case, I show the mean forecast error, the standard error, and the p-value from the t-test for whether the mean forecast error is significantly different from zero. Table 1 shows that for all versions of realized values and for both variables, we never reject the null hypothesis of zero-mean forecast error, with all p-values well above 0.05.

[8]I follow most of the forecasting literature in testing for bias under the assumption of a loss function for which bias is undesirable. A few papers, such as Elliott, Komunjer and Timmermann (2008), allow for the possibility that the loss function of forecasters may be asymmetric, which implies that bias in the forecasts may be optimal.

Figure 1. Mean One-Year-Ahead Inflation Forecasts, Realized Values, and Forecast Errors



Note: The upper panel shows one-year-ahead inflation forecasts from the SPF (Forecast) and realized values based on the initial data release, labeled Realized Value (Initial). The bottom panel shows the forecast error, measured as realized value minus forecast. The date shown on the horizontal axis is the date on which the forecasts were made, ranging from 1971Q1 to 2022Q4. Note some large forecast errors and some persistent errors.

Table 1—Test for bias, one-year ahead, based on mean SPF inflation forecast, full sample

Realized Value	Mean Error	Standard Error	p-value
Initial	0.025	0.21	0.91
First revision	0.040	0.21	0.85
First final	0.048	0.21	0.82
First annual	0.139	0.22	0.52
Pre-benchmark	0.124	0.23	0.58
Last	0.043	0.21	0.84

Note: The table shows the results of the zero-mean forecast-error test for inflation forecasts using the six different alternative measures of realized values. The sample uses SPF forecasts from 1971Q1 to 2022Q4. The p-value is a standard t-test for the null hypothesis that the mean forecast error is zero. Standard errors are adjusted following the Newey and West (1987) procedure.

As Figure 1 suggests, however, the COVID period represented a huge shock that forecasters could not have possibly forecast well, so perhaps the results in Table 1 are distorted by COVID. To test that, I rerun the bias tests so that they end before the COVID period, as shown in Table 2. The results are consistent with those in Table 1, with no rejection (at the 0.05 level) of the null hypothesis of zero-mean-forecast errors. But notice that the mean errors, p-values, and standard errors all differ from the period that includes COVID. For the remainder of this paper, I will analyze the pre-COVID period.

Table 2—Test for bias, one-year ahead, based on mean SPF inflation forecast, pre-COVID sample

Realized Value	Mean Error	Standard Error	p-value
Initial	−0.105	0.20	0.59
First revision	−0.091	0.20	0.65
First final	−0.082	0.20	0.68
First annual	0.0072	0.20	0.97
Pre-benchmark	−0.012	0.21	0.96
Last	−0.101	0.20	0.61

Note: The table shows the results of the zero-mean forecast-error test for inflation forecasts using the six different alternative measures of realized values. The sample is 1971Q1 to 2018Q4. The p-value is a standard t-test for the null hypothesis that the mean forecast error is zero. Standard errors are adjusted following the Newey and West (1987) procedure.

B. Tests for Unbiasedness in Sub-Samples

Tests Using the EOS-Naive Approach. Croushore (2010) shows that results like the tests for bias shown in Tables 1 and 2 tend to be fragile: they change dramatically depending on the precise beginning and ending dates of the sample. One way to investigate this is to consider how researchers might have perceived the bias at various points in (vintage) time. Suppose a researcher had run the zero-mean test in the second quarter of 1979, with data and one-year-ahead forecasts made from 1971Q1 to 1978Q1. What conclusion about bias would she have drawn? We can ask the same question for a researcher standing at any date between 1979Q3 and 2020Q1. But doing so is a bit difficult because we must be careful to consider the exact information set a researcher would have at each date. For example, a researcher would most likely use the latest-available vintage at each date to evaluate the past forecasts; so the researcher might use latest-available realized values at each date. But, of course, a researcher standing in the second quarter of 1979 would have had a very different version of the latest-available data than the March 2024 vintage that is the last one I use in this paper. So, I collect a sequence of latest-available data sets at each date. I call the date at which a researcher would observe those data as the "research date."

Running the bias test regressions at each date from 1979Q2 and 2020Q1, leads to the results shown in Figure 2 labeled "EOS-naive". The solid red line in the upper panel of the figure shows the p-value for the test of the null hypothesis of unbiasedness, that is, testing whether k_h^1 in Equation (2) $= 0$, while the solid red line in the lower panel shows the estimated value of k_h^1. The horizontal axis shows the research date at which each test was performed in our simulated experiment. The results show some periods early in the sample period in which the test rejects the null hypothesis of unbiasedness. The estimated bias was positive in samples that ended in the late 1970s and early 1980s, but moved closer to zero over time. The p-values for the null of unbiasedness are below 0.05 in the late 1970s and early 1980s, but never after that. So, there is not much evidence using the EOS-naive

approach for bias in the inflation forecasts.

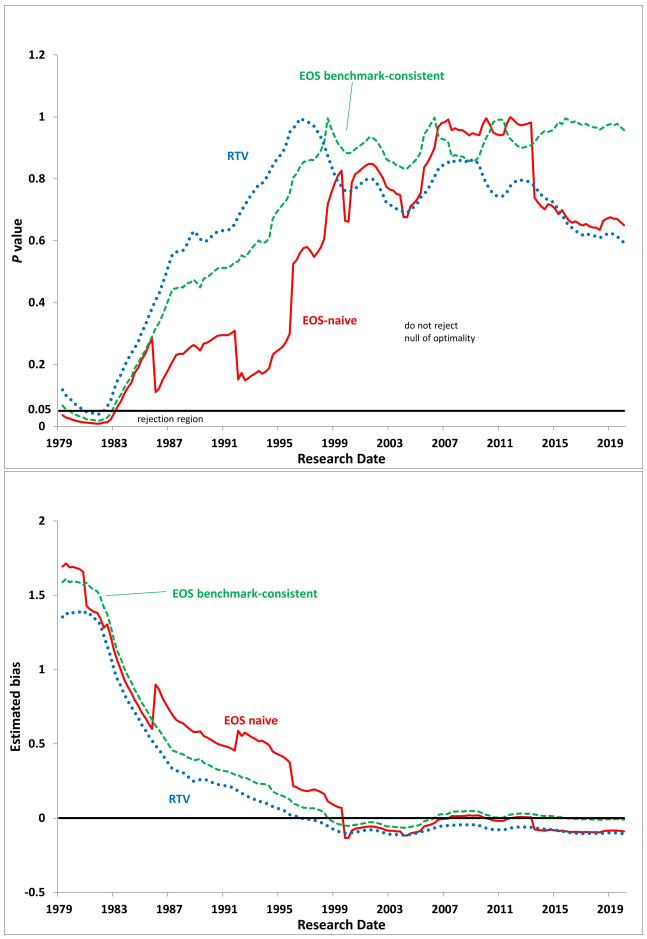
Tests Using the EOS Benchmark-Consistent Approach. The idea of being benchmark-consistent means that a researcher judging the quality of the forecasts uses data available at each date, but adjusts for benchmark revisions, by using the pre-benchmark release as the realized value for each forecast. Suppose a researcher wants to test for bias over the entire sample at each date, but understands data revisions and wants to be benchmark-consistent. Then, the researcher would use pre-benchmark realized values for evaluating forecasts for which a benchmark revision has occurred, but would use the latest-available data for evaluating forecasts for which a new benchmark revision has not yet occurred. For example, consider evaluating the forecast made in 1982Q1, when our latest-available data vintage is from the end of April 1983. We would use data from the end of December 1975 to evaluate the forecasts made from 1971Q1 to 1974Q3, then use the data from the end of November 1980 to evaluate the forecasts made from 1974Q4 to 1979Q3, and use the current vintage of data from the end of April 1983 to evaluate the forecasts made from 1979Q4 to 1982Q1.

Following this procedure and simulating what a researcher would have done in testing for bias at every research date from 1979Q2 to 2020Q1, gives the results shown in the dashed green line Figure 2. As was the case with the EOS-naive approach, the EOS benchmark-consistent approach also shows only a small bit of evidence of bias in the forecasts. As the sample advances through time, the estimated bias gets close to zero in the early 2000s, but then drifts down over time.

Tests Using the RTV Approach. The idea of using real-time vintages (RTV) means that a researcher judging the quality of the forecasts uses data of the same vintage type as the realized value for each forecast. Under this view, a researcher at each date assumes the data-generating process relates all initial releases to each other.

Following this procedure and simulating what a researcher would have done in

FIGURE 2. *P*-VALUES FOR BIAS AND ESTIMATED BIAS IN INFLATION FORECASTS FOR SAMPLE OBSERVED BY RESEARCHER AT ALTERNATIVE RESEARCH DATES USING ALL THREE APPROACHES



Note: The upper panel shows the p-values from the zero-mean forecast error test that would have been calculated by a researcher using data at the date shown on the horizontal axis using the EOS-naive approach, the EOS benchmark-consistent approach, and the RTV approach. The horizontal line in the upper panel shows where the p-value = 0.05. The lower panel shows the estimated bias in the forecast at each date. The research dates are 1979Q2 to 2020Q1, based on expanding windows of SPF forecasts starting in 1971Q1 and ending from 1978Q1 to 2018Q4. (The five-quarter lag is because the forecasts are for four-quarters ahead and the realized values are not known until another quarter after that.)

testing for bias at every research date from 1979Q2 to 2020Q1, gives the results shown in the dotted blue lines in Figure 2. As was the case with the two EOS approaches, the RTV approach also shows little evidence of bias in the forecasts. The early years of the sample show positive estimated bias terms, but they are rarely statistically significantly positive.

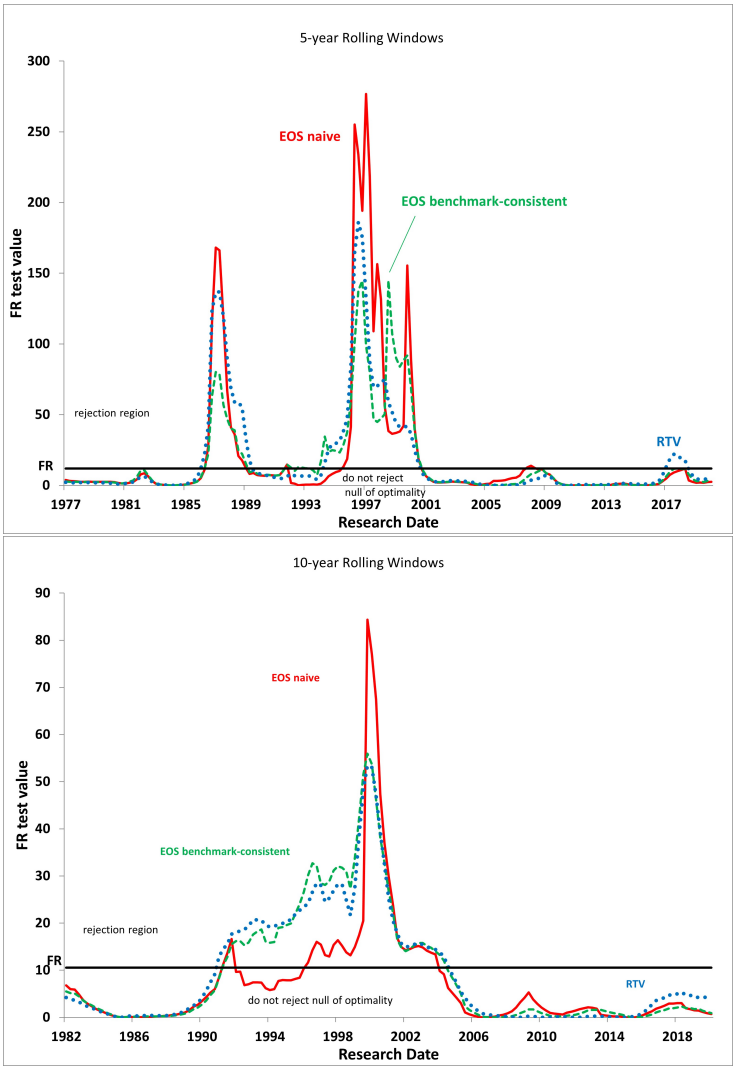
Accounting for Instability. The three approaches that I used showed relatively little evidence of bias in the samples that all began in 1971Q1 that I used in the previous section. But Rossi (2021) suggests that the lack of bias in our findings is because positive bias in some parts of the sample may offset negative bias in other parts. To account for the instability in the bias, she suggests rolling tests and in the paper Rossi and Sekhposyan (2016) the authors develop a fluctuation-rationality (FR) test, which provides a better test of bias and is robust to the presence of instabilities. I implement their test here using both 5-year rolling windows and 10-year rolling windows. The advantage of their test is that it accounts for sequential testing bias.

Running all three approaches (EOS-naive, EOS benchmark-consistent, and RTV) for rolling 5-year windows and 10-year windows gives us the fluctuation-rationality tests shown in Figure 3. The results are consistent with the Rossi (2021) suggestion: the apparent lack of rejection of bias over the entire sample shown in the results above arises because of offsetting biases in sub-samples. The fluctuation-rationality tests reject the null of unbiasedness. Rejections are fewer for the EOS-naive approach than for the other two approaches. The rejection of rationality suggests that there is scope for improving the forecasts in real time.

III. Forecast-Improvement Exercises for Bias in Real Time

A problem in the literature on forecast evaluation is that many researchers find bias or inefficiency in-sample, but that bias cannot be exploited out of sample. I would like to be able to use the results of the bias tests to show that, in real time, a better forecast could have been constructed. In the early rational-expectations

FIGURE 3. FLUCTUATION-RATIONALITY TESTS IN FORECASTS FOR INFLATION IN ROLLING 5-YEAR AND 10-YEAR WINDOWS



Note: The upper panel shows the values of the Fluctuation-Rationality test for inflation using rolling 5-year samples of data. Each line corresponds to a different approaches: the EOS-naive approach, the EOS benchmark-consistent approach, or the RTV approach using initial realized values. The lower panel shows the same concept for 10-year rolling windows. The research dates (the dates at which a research would have data on realized values at the end of the rolling window) are shown on the horizontal axis.

literature, the bias that was found in the forecasts was clear, and the prescription for researchers and policymakers was that they could improve on published forecasts by adjusting the forecasts by the amount of the bias.

A. Forecast Improvement for Bias Using EOS-Naive Approach

To improve the forecasts, given that the EOS-naive approach showed bias in numerous sub-samples, I estimate the bias in rolling samples, then create a new and improved forecast from the survey forecast, as in Equation (4).

The results of this exercise are shown in Table 3. The rows of the tables show alternative experiments, described below. The first column of numbers shows the relative-root-mean-squared forecast error ($RRMSFE$) for estimating the bias using 5-year rolling windows and Equation (4), where $RRMSFE$ is the $RMSFE$ of the improved forecast divided by the $RMSFE$ of the original survey. Thus, an $RRMSFE$ less than one means that estimating the bias and using Equation (4) leads to a lower $RMSFE$ and an improved forecast; an $RRMSFE$ greater than one means that the attempt to improve the forecast failed. The second column of numbers shows the p-value for the test of a significantly difference in $RMSFEs$, based on the Diebold and Mariano (1995) test.[9] The next two columns repeat this exercise for 10-year rolling windows.

The first row in Table 3 labeled "Adjust every period" shows the results of the basic experiment in which I use Equation (4) to attempt to improve on the survey forecasts based on the estimated bias each period. In both cases, the forecasts are worse, as the $RRMSFE$ is greater than one, so the $RMSFE$ is higher than for the original survey. However, the p-values are all above 0.05, meaning that the difference in $RMSFEs$ is not statistically significant.

Part of the reason for the poor performance of these attempts at forecast improvement is that we are trying to use the estimated bias even in periods when the

[9]The test is subject to the caveat that with real-time data, subject to revisions, the test is not completely valid, per Clark and McCracken (2009), for which there is not yet a satisfactory solution. For that reason, all the p-values reported should be viewed as a general guide, not as precise estimates.

TABLE 3—*RRMSFEs* AND *P*-VALUES FOR FORECAST IMPROVEMENT EXERCISES, EOS-NAIVE APPROACH WITH REALIZED VALUES = INITIAL

Window Size:	5-year	10-year
Adjust every period	1.134	1.227
	[0.35]	[0.23]
Adjust when FR test rejects	1.083	0.989
	[0.29]	[0.82]
	With Shrinkage	
Adjust every period	0.995	1.072
	[0.94]	[0.44]
Adjust when FR test rejects	1.000	0.978
	[0.99]	[0.36]

Note: The table shows relative-root-mean-squared errors ($RRMSFE$) and p-values of the Diebold-Mariano test [in square brackets] for inflation forecasts in forecast-improvement exercises, using the EOS-naive approach with realized values = initial. The sample consists of one-year-ahead SPF forecasts evaluated at research dates from 1982Q1 to 2020Q1. Note that the inflation $RMSFE = 0.792$.

bias is not statistically significant. However, more likely someone estimating bias in real time would adjust the forecast using Equation (4) only if the fluctuation-rationality test showed rejection. I will apply Equation (4) only in periods when the fluctuation rationality test is rejected, as shown in Figure 3.

The results of this exercise are shown in the rows for each variable in Table 3, labeled "Adjust when FR test rejects." Compared with the first row, the results here suggest some ability to improve upon the SPF forecasts for inflation in 10-year rolling windows, though not statistically significantly so.

One final possibility is to recognize that the bias is estimated with error, so it makes sense to use shrinkage methods to reduce the error introduced by parameter estimation. Suppose I adjust for bias, but only adjust for the bias by a factor of one-half:

$$\hat{X}_t = (0.5 \times \hat{\alpha}) + X_t^f, \tag{8}$$

Using Equation (8) instead of Equation (4), I get the results shown in Table 3 under the header "With Shrinkage". I can use shrinkage, adjusting every period, or only when the FR test shows rejection.

The results show that shrinkage always helps. Although I could search for the optimal degree of shrinkage, this would violate the concept of a researcher being able to adjust for the bias in real time. Overall, using the EOS-naive approach, there is scope for improving the inflation forecasts, though in no cases is the reduction in $RMSFE$ is significant.

B. Forecast Improvement for Bias Using the EOS Benchmark-Consistent Approach

If I repeat the steps above, but use the EOS benchmark-consistent approach, I obtain similar results to using the EOS-naive approach, as can be seen in Table 4. In about half of the cases, the $RRMSFE$ is less than one, with as much as a 6 percent improvement in $RMSFE$ (though not statistically significantly so).

Shrinkage and using the fluctuation-rationality test results both help to reduce the $RMSFEs$.

TABLE 4—$RRMSFEs$ AND P-VALUES FOR FORECAST IMPROVEMENT EXERCISES, EOS BENCHMARK-CONSISTENT APPROACH WITH REALIZED VALUES = INITIAL

Window Size:	5-year	10-year
Adjust every period	1.145	1.262
	[0.35]	[0.24]
Adjust when FR test rejects	1.045	0.957
	[0.63]	[0.57]
	With Shrinkage	
Adjust every period	0.992	1.076
	[0.92]	[0.48]
Adjust when FR test rejects	0.969	0.941
	[0.42]	[0.17]

Note: The table shows relative-root-mean-squared errors ($RRMSFE$) and p-values of the Diebold-Mariano test [in square brackets] for inflation forecasts in forecast-improvement exercises, using the EOS benchmark-consistent approach with realized values = initial. The sample consists of one-year-ahead SPF forecasts evaluated at research dates from 1982Q1 to 2020Q1.

C. Forecast Improvement for Bias Using RTV Approach

Finally, I use the RTV approach, with the initial release of the data to determine the forecast error, with results in Table 5. It might be possible to use a later release of the data as well, but that creates problems in a real-time forecast-improvement exercise because concepts other than the initial release mean longer lags in data availability. For example, using pre-benchmark data as realized values to determine the forecast error means that in real time there might be five years that pass before you get any new observations to use.

Overall, using the RTV approach, there is scope for improving the inflation

TABLE 5—*RRMSFEs* AND *P*-VALUES FOR FORECAST IMPROVEMENT EXERCISES, RTV APPROACH WITH REALIZED VALUES = INITIAL

Window Size:	5-year	10-year
Adjust every period	1.073	1.182
	[0.59]	[0.35]
Adjust when FO test rejects	1.035	0.947
	[0.67]	[0.49]
	With Shrinkage	
Adjust every period	0.961	1.038
	[0.58]	[0.69]
Adjust when FO test rejects	0.966	0.937
	[0.35]	[0.15]

Note: The table shows relative-root-mean-squared errors ($RRMSFE$) and p-values of the Diebold-Mariano test [in square brackets] for inflation forecasts in forecast-improvement exercises using the RTV approach. The sample consists of one-year-ahead SPF forecasts evaluated at research dates from 1982Q1 to 2020Q1.

forecasts, similar to the EOS benchmark-consistent approach. The forecast improvement found here is stronger than that found by Eva and Winkler (2023), in their recent study. But they work on many more variables than in this paper, over a different sample period.

Overall, testing numerous approaches to improve on the forecasts shows that forecast improvement is not easy, though not impossible.

D. Conclusions About Bias Tests, Allowing for Instability

Our analysis of the variation in results across subsamples and alternate versions of realized values can explain many of the results about bias in survey forecasts of output growth in the literature.

The conclusions of this section are that (1) there are no simple stylized facts about bias in survey forecasts of inflation; (2) many subsamples of survey data show evidence of bias, even though no bias is apparent in the full sample; (3) it may be possible to improve on the survey forecasts in real time, especially using forecast-rationality tests and shrinkage; and (4) the conclusions we can draw about bias in survey forecasts are heavily dependent on the choice of realized values for data that are subject to revisions. The main contributions of this paper to the literature on the rationality of forecasts are to provide more evidence about the sub-sample variation in estimates of bias and to provide a more-detailed examination of forecast-improvement exercises than has been done before, including the use of forecast-rationality tests and shrinkage.

IV. Testing for Inefficiency in Real Time, Accounting for Instability

Researchers have tested for inefficiency in inflation forecasts in a number of ways. For example, researchers such as Roberts (1997) and Schuh (2001) have suggested that the unemployment rate might be correlated with inflation forecast errors. Since the goal here is to illustrate how testing for inefficiency works and how to do forecast-improvement exercises, I test the unemployment rate as a

possible predictor of inflation forecast errors.

A. Forecast Errors and Unemployment

In this section, I investigate whether data on the unemployment rate could be used to improve inflation forecasts in real time. This might occur if the Phillips Curve is a good model of inflation but if forecasters do not use the model appropriately. I begin with in-sample results to see if unemployment is related to inflation forecast errors in the data set, then I move to out-of-sample forecasting to see if the in-sample relationship can be used to improve inflation forecasts. An advantage of testing the unemployment rate is that the unemployment rate for the first month of each quarter is released by the Bureau of Labor Statistics early in the second month of the quarter, shortly after the initial NIPA release but before the SPF forecasters have made their new round of forecasts. I use only real-time, initial releases for the unemployment rate in the following analysis.

First, I run a regression of the forecast error for inflation over the next year and the unemployment rate observed when the SPF survey is taken. The regression is simply:

$$e_{t,t+4} = \alpha + \beta U_t + \epsilon_t, \tag{9}$$

where U_t is the unemployment rate available at date t and $e_{t,t+h}$ is the forecast error.

The results are summarized in Table 6. About half of all the cases show a statistically significant coefficient in regression Equation (9) on the unemployment rate, which suggests that forecasters are not using information about unemployment efficiently in forming their forecasts. In addition, most of the slope coefficients show statistical significance, so using the unemployment rate might help improve inflation forecasts. In terms of the alternative measures of realized values, the

coefficients on the unemployment rate are more often significant for using earlier versions of realized values: initial, first revision, or first final.

Table 6—In-Sample Results for Inflation Forecast Errors Regressed on the Unemployment Rate

	$\hat{\alpha}$	$\hat{\beta}$	p-value
Initial	1.336	-0.230	0.017
	(0.79)	(0.11)	
First revision	1.366	-0.232	0.024
	(0.81)	(0.11)	
First final	1.372	-0.232	0.027
	(0.81)	(0.11)	
First annual	1.519	-0.241	0.055
	(0.78)	(0.11)	
Pre-benchmark	1.462	-0.235	0.083
	(0.85)	(0.12)	
Last	1.236	-0.213	0.062
	(0.76)	(0.11)	

Note: The table shows the results of the efficiency test from Equation (9) for inflation forecasts using the six different alternative measures of realized values. The sample uses SPF forecasts from 1971Q1 to 2018Q4. The p-value comes from a χ-squared test for the null hypothesis that both coefficients in Equation (9) equal zero. Standard errors are adjusted following the Newey and West (1987) procedure.

As with the bias results, the results of the tests for inefficiency change substantially over the sample period. So, in what follows, I will use all three approaches (EOS-naive, EOS benchmark-consistent, and RTV) with rolling windows to examine the inefficiency tests, followed by forecast-improvement exercises.

Inefficiency Tests. Following a similar procedure that I used in the bias tests, I find periods in which rolling 5- and 10-year windows show significant p-values for the three approaches (EOS-naive, EOS benchmark-consistent, and RTV). Similarly, I run the fluctuation-rationality tests. Given the in-sample results of the fluctuation-rationality tests, I proceed to investigate the possibility of

using the regression results from Equation (9) to improve upon the SPF forecasts in forecast-improvement exercises. Taking the estimated $\hat{\alpha}$ and $\hat{\beta}$ coefficients over rolling samples, I create, at each date t, an improved forecast $Y^I_{T,h}$, where

$$Y^I_{T,h} = Y^f_{T,h} + \hat{\alpha} + \hat{\beta} U_T. \tag{10}$$

Using regression Equations (9) and (10), I simulate the activity of a real-time forecaster standing at different dates, using rolling 5-year and 10-year windows and proceeding to 2018Q4 with one-year-ahead SPF forecasts, allowing for the lag in data availability, and forming improved forecasts at each date based only on the real-time data and past forecast errors available at each date.

Table 7 reports the results. The most promising method for forecast improvement is the RTV approach with shrinkage, which has $RRMSFEs$ less than 1, though none of the improvements are statistically significant. In a few other cases, with 10-year rolling windows, the attempt at forecast improvement actually leads to statistically significantly worse forecasts.

B. Summary and Conclusions About Forecast-Improvement Exercises for Inefficiency in Inflation Forecasts

I have tested the ability of a researcher to improve on SPF forecasts of inflation by using data on the unemployment rate. Even though inefficiency appears to hold in sample, it is quite difficult to improve on the forecasts. The most promising approach seems to be the RTV approach with shrinkage. But the forecast improvement, if any, is modest. In fact, most attempts to improve upon the SPF forecasts in terms of inefficiency make forecasts worse, not better. It is hard, though not impossible, to improve on SPF forecasts of inflation using data on the unemployment rate.

TABLE 7—*RRMSFEs* AND *P*-VALUES FOR FORECAST IMPROVEMENT EXERCISES, 3 APPROACHES WITH REALIZED VALUES = INITIAL

Window Size:	5-year	10-year
EOS-naive		
Adjust every period	1.558 [0.20]	1.389 [0.04]
Adjust when FR test rejects	1.451 [0.20]	1.028 [0.53]
EOS-naive with shrinkage		
Adjust every period	1.080 [0.66]	1.051 [0.60]
Adjust when FR test rejects	1.043 [0.78]	1.000 [0.98]
EOS benchmark-consistent		
Adjust every period	1.584 [0.16]	1.427 [0.03]
Adjust when FR test rejects	1.467 [0.18]	0.996 [0.95]
EOS benchmark-consistent with shrinkage		
Adjust every period	1.082 [0.64]	1.044 [0.61]
Adjust when FR test rejects	1.036 [0.81]	0.970 [0.37]
RTV		
Adjust every period	1.302 [0.11]	1.275 [0.10]
Adjust when FR test rejects	1.195 [0.15]	0.972 [0.67]
RTV with shrinkage		
Adjust every period	0.950 [0.62]	0.971 [0.66]
Adjust when FR test rejects	0.907 [0.24]	0.958 [0.23]

Note: The table shows relative-root-mean-squared errors ($RRMSFE$) and p-values of the Diebold-Mariano test [in square brackets] for inflation forecasts in forecast-improvement exercises, using all three approaches for five- and ten-year rolling windows realized values = initial. The sample consists of one-year-ahead SPF forecasts evaluated at research dates from 1982Q1 to 2020Q1.

V. Summary and Conclusions

The goal of this paper was to create a systematic method for improving forecasts to reduce bias and inefficiency. I examined three different approaches for analysis, with differing assumptions about the data-generating process related to how data are revised: End-of-Sample (EOS)-naive, End-of-Sample (EOS) benchmark-consistent, and real-time vintages (RTV). I considered optimal ways to account for data revisions and instability. I developed forecast-improvement exercises, employing shrinkage. When forecasts exhibit bias and inefficiency, I found some ability to improve forecasts out-of-sample, but the improvement was never statistically significant.

Why might in-sample results show a relationship between macroeconomic variables and forecast errors, but out-of-sample results do not? It may be that forecasters do not recognize the importance of a variable for forecasting until some time passes, so there is an in-sample relationship that is not useful for forecasting for very long. Or, as Cukierman, Lustenberger and Meltzer (2020) suggest, a permanent-transitory confusion may lead to in-sample correlations, even if forecasters have rational expectations.

Why might forecasters show periodic bouts of bias in their forecasts? As Farmer, Nakamura and Steinsson (2024) suggest, forecasters may not know the data-generating process at a given date but learn more about it over time. Our results are consistent with their theoretical model—forecasters do the best they can with a changing structure of the economy, and biases appear from time to time but disappear once forecasters understand the structural change.

The structure of the forecast-improvement exercises in this paper is based on the in-sample results reported by others in the literature, cited in the Introduction. Some possible future extensions of this work include: (1) Looking at forecasts errors and their relationship to forecast revisions, as in Coibion and Gorodnichenko (2015); (2) Testing additional variables that might affect inflation forecasts; (3) Modifying the degree of shrinkage used in forecast-improvement exercises; and (4)

Applying these methods to early data releases to see if they are optimal forecasts of later vintages. This paper should serve as a guide for future research.

I suggest that we focus on the question of whether or not we, as forecasting researchers, can identify flaws in forecasts made by forecasters and help them make better forecasts. That is the objective of this paper and I have suggested ways to do that.

REFERENCES

Aiolfi, Marco, Carlos Capistran, and Allan Timmermann. 2011. "Forecast Combinations." In *The Oxford Handbook of Economic Forecasting.* , ed. Michael P. Clements and David F. Hendry, Chapter 12, 355–388. Oxford University Press.

Ang, Andrew, Geert Bekaert, and Min Wei. 2007. "Do Macro Variables, Asset Markets, or Surveys Forecast Inflation Better?" *Journal of Monetary Economics*, 54: 1163–1212.

Ball, Laurence, and Dean Croushore. 2003. "Expectations and the Effects of Monetary Policy." *Journal of Money, Credit and Banking*, 35: 473–484.

Carroll, Christopher D. 2003. "Macroeconomic Expectations of Households and Professional Forecasters." *Quarterly Journal of Economics*, 118: 269–298.

Clark, Todd E., and Michael W. McCracken. 2009. "Tests of Equal Predictive Ability with Real-Time Data." *Journal of Business and Economic Statistics*, 27: 441–454.

Coibion, Olivier, and Yuriy Gorodnichenko. 2015. "Information Rigidity and the Expectations Formation Process: A Simple Framework and New Facts." *American Economic Review*, 105(8): 2644–2678.

Croushore, Dean. 2010. "An Evaluation of Inflation Forecasts from Surveys using Real-Time Data." *B.E. Journal of Macroeconomics*, 10(1): article 10.

Croushore, Dean. 2011. "Frontiers of Real-Time Data Analysis." *Journal of Economic Literature*, 49(1): 72–100.

Croushore, Dean. 2019. "Revisions to PCE Inflation Measures: Implications for Monetary Policy." *International Journal of Central Banking*, 15(4): 241–265.

Croushore, Dean, and Tom Stark. 2001. "A Real-Time Data Set for Macroeconomists." *Journal of Econometrics*, 105: 111–130.

Croushore, Dean, and Tom Stark. 2019. "Fifty Years of the Survey of Professional Forecasters." *Federal Reserve Bank of Philadelphia Economic Insights*, 1–11.

Cukierman, Alex, Thomas Lustenberger, and Allan Meltzer. 2020. "The Permanent-Transitory Confusion: Implications for Tests of Market Efficiency and for Expected Inflation During Turbulent and Tranquil Times." *Expectations: Theory and Applications from Historical Perspectives*, , ed. Arie Arnon, Warren Young and Karine van der Beek, 215–238. Cham:Springer International Publishing.

Diebold, Francis X., and Roberto S. Mariano. 1995. "Comparing Predictive Accuracy." *Journal of Business and Economic Statistics*, 13: 253–263.

Elliott, Graham, and Allan Timmermann. 2008. "Economic Forecasting." *Journal of Economic Literature*, 46: 3–56.

Elliott, Graham, Ivana Komunjer, and Allan Timmermann. 2008. "Biases in Macroeconomic Forecasts: Irrationality or Asymmetric Loss?" *Journal of the European Economic Association*, 6: 122–157.

Eva, Kenneth, and Fabian Winkler. 2023. "A Comprehensive Empirical Evaluation of Biases in Expectation Formation." Working Paper, Federal Reserve Board.

Farmer, Leland E., Emi Nakamura, and Jon Steinsson. 2024. "Learning About the Long Run." *Journal of Political Economy*.

Giacomini, Raffaella, and Barbara Rossi. 2010. "Forecast Comparisons in Unstable Environments." *Journal of Applied Econometrics*, 25: 595–620.

Harvey, David S., Stephen J. Leybourne, and Paul Newbold. 1997. "Testing the Equality of Prediction Mean Squared Errors." *International Journal of Forecasting*, 13: 281–291.

Keane, Michael P., and David E. Runkle. 1990. "Testing the Rationality of Price Forecasts: New Evidence From Panel Data." *American Economic Review*, 80: 714–735.

Kishor, N. Kundan, and Evan F. Koenig. 2012. "VAR Estimation and Forecasting When Data Are Subject to Revision." *Journal of Business and Economic Statistics*, 30: 181–190.

Koenig, Evan, Sheila Dolmas, and Jeremy Piger. 2003. "The Use and Abuse of 'Real-Time' Data in Economic Forecasting." *Review of Economics and Statistics*, 85: 618–628.

Mankiw, N. Gregory, and Matthew D. Shapiro. 1986. "Do We Reject Too Often? Small Sample Bias in Tests of Rational Expectations Models." *Economics Letters*, 20: 139–145.

Mincer, Jacob A., and Victor Zarnowitz. 1969. "The Evaluation of Economic Forecasts." In *Economic Forecasts and Expectations.* , ed. Jacob Mincer. New York:National Bureau of Economic Research.

Newey, Whitney K., and Kenneth D. West. 1987. "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix." *Econometrica*, 55: 703–708.

Roberts, John. 1997. "Is Inflation Sticky?" *Journal of Monetary Economics*, 39(2): 173–196.

Romer, Christina D., and David H. Romer. 2000. "Federal Reserve Information and the Behavior of Interest Rates." *American Economic Review*, 90(3): 429–457.

Rossi, Barbara. 2006. "Are Exchange Rates Really Random Walks? Some Evidence Robust to Parameter Instability." *Macroeconomic Dynamics*, 10: 20–38.

Rossi, Barbara. 2021. "Forecasting in the Presence of Instabilities: How We Know Whether Models Predict Well and How to Improve Them." *Journal of Economic Literature*, 59(4): 1135–1190.

Rossi, Barbara, and Tatevik Sekhposyan. 2010. "Have Economic Models' Forecasting Performance for US Output Growth and Inflation Changed Over Time, and When?" *International Journal of Forecasting*, 26(4): 808–835.

Rossi, Barbara, and Tatevik Sekhposyan. 2016. "Forecast Rationality Tests in the Presence of Instabilities, with Applications to Federal Reserve and Survey Forecasts." *Journal of Applied Econometrics*, 31(3): 507–532.

Rudebusch, Glenn D., and John C. Williams. 2009. "Forecasting Recessions: The Puzzle of the Enduring Power of the Yield Curve." *Journal of Business and Economic Statistics*, 27(4): 492–503.

Schuh, Scott. 2001. "An Evaluation of Recent Macroeconomic Forecast Errors." *New England Economic Review*, 35–56.

Stock, James H., and Mark W. Watson. 2003. "Forecasting Output and Inflation: The Role of Asset Prices." *Journal of Economic Literature*, 41: 788–829.

Su, Vincent, and Josephine Su. 1975. "An Evaluation of ASA/NBER Business Outlook Survey Forecasts." *Explorations in Economic Research*, 2: 588–618.

Zarnowitz, Victor. 1985. "Rational Expectations and Macroeconomic Forecasts." *Journal of Business and Economic Statistics*, 3: 293–311.

Appendix

This Appendix contains some notation to clearly define the concepts of realized values, as well as showing the dates of the first annual vintages and pre-benchmark revisions.

In general terms, we use a subscript to denote the quarter for which the data apply and a superscript to denote the date of the vintage, where a subscript has two terms: the quarter of the vintage, and the month. For example, consider the observations in our sample that were released at the end of March 2024 and the last quarter for which data exist is the fourth quarter of 2023. So, the value of variable X for that date is denoted as:

$$X_{2023Q4}^{2024Q1,3}.$$

We will denote all the data in the last release (March 2024), which contains data from 1947Q1 to 2023Q4, as:[10]

$$X^{last} = \{X_{1947Q1}^{2024Q1,3}, X_{1947Q2}^{2024Q1,3}, X_{1947Q3}^{2024Q1,3}, ..., X_{2023Q2}^{2024Q1,3}, X_{2023Q3}^{2024Q1,3}, X_{2023Q4}^{2024Q1,3}\}.$$

Similarly, any other vintage of data can be described as:

$$X^{Q,M} = \{X_{1947Q1}^{Q,M}, X_{1947Q2}^{Q,M}, ..., X_{Q-1}^{Q,M}\}.$$

For example, the data release at the end of January 1999 is:

$$X^{1999Q1,1} = \{X_{1947Q1}^{1999Q1,1}, X_{1947Q2}^{1999Q1,1}, ..., X_{1998Q4}^{1999Q1,1}\}.$$

[10]The data come from the Real-Time Data Set for Macroeconomists (RTDSM), the vintages of which are dated mid-month. So, the data released at the end of March 2024 are called the vintage of 2024M4 (April 2024) in the RTDSM.

Thus, based on our earlier definition, $X^{last} = X^{2024Q1,3}$.

The first regular monthly release of quarterly GDP data occurred at the end of October 1965 and the last observation in that release was for 1965Q2. Almost always,[11] the first release for output and the price level occurred in the first month of the following quarter, so we denote a collection of all the initial releases as:

$$X^{initial} = \{X^{1965Q3,1}_{1965Q2}, X^{1965Q4,1}_{1965Q3}, X^{1966Q1,1}_{1965Q4}, ..., X^{2023Q3,1}_{2023Q2}, X^{2023Q4,1}_{2023Q3}, X^{2024Q1,1}_{2023Q4}\}.$$

The first-revision realized values are similar to the initial realized values but use the data vintage from the second month of the following quarter. Similarly, the first-final realized values use the data vintage from the third month of the following quarter.

$$X^{initial} = \{X^{1965Q3,2}_{1965Q2}, X^{1965Q4,2}_{1965Q3}, X^{1966Q1,2}_{1965Q4}, ..., X^{2023Q3,2}_{2023Q2}, X^{2023Q4,2}_{2023Q3}, X^{2024Q1,2}_{2023Q4}\}.$$

$$X^{FirstFinal} = \{X^{1965Q3,3}_{1965Q2}, X^{1965Q4,3}_{1965Q3}, X^{1966Q1,3}_{1965Q4}, ..., X^{2023Q3,3}_{2023Q2}, X^{2023Q4,3}_{2023Q3}, X^{2024Q1,3}_{2023Q4}\}.$$

Annual revisions usually occur every year at the end of July, with some exceptions (including in July 2023), which we note in Table A1. The first annual revision occurred at the end of July each year (and are thus in the August monthly vintage of the RTDSM) pertaining to the prior calendar year every year except for the following. For example, the first annual revision of the data for 1965 was at the end of July 1966 and recorded in the 1966M8 vintage of the RTDSM. The

[11] The exception was the first release of 1995Q4, which was delayed because of the federal government shutdown.

exceptions are given in this table:

Table A1—Annual Revision Dates for Quarterly National Accounts

Year Revised	First Annual Revision Date in RTDSM
1974	1976M2
1979	1981M1
1980	1981M8
1984	1986M1
1990	1991M12
1994	1996M1
1998	1999M11
2002	2003M12
2022	2023M10

Collecting all the first annual revisions gives us the following vector:

$$\begin{aligned} X^{annual} = \{&X_{1965Q2}^{1966Q3,1}, X_{1965Q3}^{1966Q3,1}, X_{1965Q4}^{1966Q3,1}, \\ &X_{1966Q1}^{1967Q3,1}, X_{1966Q2}^{1967Q3,1}, X_{1966Q3}^{1967Q3,1}, X_{1966Q4}^{1967Q3,1}, \\ &\ldots, \\ &X_{2020Q1}^{2021Q3,1}, X_{2020Q2}^{2021Q3,1}, X_{2020Q3}^{2021Q3,1}, X_{2020Q4}^{2021Q3,1}, \\ &X_{2021Q1}^{2022Q3,1}, X_{2021Q2}^{2022Q3,1}, X_{2021Q3}^{2022Q3,1}, X_{2021Q4}^{2022Q3,1}, X_{2022Q4}^{2023Q3,3}\}. \end{aligned}$$

The pre-benchmark values are more difficult to generate, as their pattern is irregular. Dates for the pre-benchmark vintages are given in Table A2.

The first benchmark revision was in late January 1976, so the pre-benchmark values came from the December 1975 (1975Q4,3) vintage. If there has not yet been a benchmark revision for some observations, we use the last vintage available. The overall vector looks like:

Table A2—Pre-Benchmark-Revision RTDSM Monthly Dates

SPF Forecast Dates	Pre-Benchmark Vintage	Last Observation
1971Q1 to 1974Q3	1975Q4,3	1975Q3
1974Q4 to 1979Q3	1980Q4,2	1980Q3
1979Q4 to 1984Q3	1985Q4,2	1985Q3
1984Q4 to 1990Q3	1991Q4,1	1991Q3
1990Q4 to 1994Q3	1995Q4,2	1995Q3
1994Q4 to 1998Q2	1999Q3,3	1999Q2
1998Q3 to 2002Q3	2003Q4,2	2003Q3
2002Q4 to 2008Q1	2009Q2,3	2009Q1
2008Q2 to 2012Q1	2013Q2,3	2013Q1
2012Q2 to 2017Q1	2018Q2,3	2018Q1
2017Q2 to 2022Q2	2023Q3,2	2023Q2

Note: The table shows the pre-benchmark vintage date in the Real-Time Data Set for Macroeconomists that provides the last observations for one-year-ahead output forecasts available prior to a benchmark revision of the National Income and Product Accounts. The benchmark revision vintage is one month after the pre-benchmark date.

$$
\begin{aligned}
X^{pre-benchmark} = \{ & X_{1965Q2}^{1975Q4,3}, X_{1965Q3}^{1975Q4,3}, X_{1965Q4}^{1975Q4,3}, \\
& X_{1966Q1}^{1975Q4,3}, X_{1966Q2}^{1975Q4,3}, X_{1966Q3}^{1975Q4,3}, X_{1966Q4}^{1975Q4,3}, \\
& \ldots, \\
& X_{2022Q1}^{2023Q3,2}, X_{2022Q2}^{2023Q3,2}, X_{2022Q3}^{2023Q3,2}, X_{2022Q4}^{2023Q3,2}, \\
& X_{2023Q1}^{2023Q3,2}, X_{2023Q2}^{2023Q3,2}, X_{2023Q3}^{2024Q1,3}, X_{2023Q4}^{2024Q1,3} \}.
\end{aligned}
$$