Improving Biased Forecasts in Real Time

Dean Croushore[1]

Abstract

I develop three approaches to improve forecasts of macroeconomic variables in real time, dealing with complications including data revisions and structural instability. I consider forecasts that have been found to be biased in-sample, and I illustrate the ideas with forecasts of corporate profits as a share of GDP, using the Survey of Professional Forecasters. Even when bias is clear in-sample, the time-varying nature of the bias makes it difficult to improve upon the forecasts out-of-sample. Only in forecasting the most recent vintage of the data is there a significant reduction in root-mean-squared forecast errors.

Keywords: real-time data, forecast bias, forecast improvement

[1] dcrousho@richmond.edu, Economics Department, Robins School of Business, 102 UR Drive, University of Richmond, VA, 23173, USA

1. Introduction

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. In a related vein of work, forecasts may be biased over some periods, with offsetting bias in other periods, but the bias may last long enough to be exploitable, as Rossi and Sekhposyan (2016) suggest. The question is: Could a researcher use results from the bias tests to improve the forecasts in real time?

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

2. 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, 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 $t+h$, where h is the horizon.

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

$$e_{t,h} = k_h + \epsilon_{t,h}. \tag{2}$$

The test for unbiasedness comes from testing the null hypothesis that $k_h = 0$, for each horizon h.

The point of departure for this paper comes from the question of how to empirically implement a finding of bias. As Elliott and Timmermann (2008) note, a finding of bias suggests "that improved forecasts are possible given the available data." (p. 34) I develop several apporoaches 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

$$Y^I_{T,h} = Y^f_{T,h} + \hat{k}_h, \tag{3}$$

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

Testing Improvement. Suppose we test a set for forecasts for bias by estimating Equation (2) and generate improved forecasts using Equation (3). Suppose we run the bias 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 Equation (2) 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 et al. (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 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 2024Q4 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 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 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?

Consider a time-series variable Y_t. Suppose the true value of it is Y_t^* but the variable is imperfectly measured and undergoes revisions over time, with the measured value at date $t + j$ denoted as Y_t^{t+j}. Now suppose the government data agency that reports the data sees differing sets of sample data for the variable at different times, denoted S_t^{t+j}.

The process by which the data agency releases data is that it follows a set of instructions, or functions, using its sample data. Following the structure of the National Income and Product Accounts, the structure of data releases is:

initial: $Y_t^i = F_1(S_t^{t+1})$

second: $Y_t^2 = F_2(S_t^{t+2})$

first final: $Y_t^{ff} = F_3(S_t^{t+3})$

first annual: $Y_t^{A1} = F_{A1}(S_t^{A1})$

second annual: $Y_t^{A2} = F_{A2}(S_t^{A2})$

third annual: $Y_t^{A3} = F_{A3}(S_t^{A3})$

first benchmark: $Y_t^{B1} = G_{B1}(S_t^{B1})$

second benchmark: $Y_t^{B2} = G_{B2}(S_t^{B2})$

. . .

Nth benchmark: $Y_t^{BN} = G_{BN}(S_t^{BN})$

Using this structure, the latest data that we observe in February 2025, with

$N = 11$ when this was written, is:

$$\{Y^{B11}_{1947Q1}, Y^{B11}_{1947Q2}, ..., Y^{B11}_{2023Q2}, Y^{A1}_{2023Q3}, Y^{A1}_{2023Q4}, \\ Y^{ff}_{2024Q1}, Y^{ff}_{2024Q2}, Y^{ff}_{2024Q3}, Y^{i}_{2024Q4}\}$$

If revisions to the data are small and white noise, the use of different concepts for realized values would be inconsequential.[2] But the literature on real-time data analysis suggests that the revisions are neither small nor innocuous. 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.[3] 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 and the data release in the previous month as the pre-benchmark 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

[2] 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.

[3] I use the date January 2025 in this paper; it corresponds to the vintage of February 2025 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 January 2025 are recorded in the February vintage of the RTDSM.

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 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).[4]

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.[5] 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;[6] for example, the sequence of initial releases may have a separate data-generating process than later releases of the

[4] The Appendix shows the dates of both first-annual revisions and pre-benchmark revisions.

[5] Given that the goal of this paper is to improve forecasts in real time, I am going to assume that it is not possible to forecast data revisions, so that early releases of the data are optimal forecasts of later releases. That is not always true for every variable, as Aruoba (2008) shows, and can be tested using the methods presented in this paper.

[6] For applications of these concepts, see Kishor and Koenig (2014) and Kishor and Koenig (2022), which show how to use information on data revisions to improve upon the forecasting performance of professional forecasters in predicting GDP growth, employment growth, and headline PCE inflation in real time.

data.

So, researchers must make a choice about what to assume about how data revisions affect the data-generating process. A key issue is that data revisions never end because of changes in data concepts (such as the introduction of intellectual property products in 2013). The possibilities for dealing with revisions depend on the structure of those revisions. I consider three hypotheses about how data are revised, each of which leads to a different empirical approach.

Hypothesis 1: Continuously-Updated Approach. Suppose Y_t^* is the truth and later measures of the data get successively closer to the truth, on average. In this case, it would be optimal for forecasters to use the latest-available data to them at each date, such as data downloaded from FRED or some similar database. I call this the Continuously-Updated Approach.

To use this approach, a researcher must gather data in an information set that would have existed at each point in time in the out-of-sample evaluation period and use it assuming a particular equation describes the data-generating process. For 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 moving 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 "Continuously-Updated" because forecasters always use the latest version of the data at each date, and they ignore 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.

Hypothesis 2: Benchmark-Consistent Approach. Suppose the G functions from benchmark revisions redefine the truth conceptually, as if it were a different variable. In that case, the Y^* vector might look like:

$$Y^* = \{Y^{B1}_{1947Q1}, Y^{B1}_{1947Q2}, ..., Y^{B1}_{1975Q3}, Y^{B2}_{1975Q4}, Y^{B2}_{1976Q1}, ..., Y^{B2}_{1980Q3}, ..., \\ Y^{B11}_{2018Q2}, Y^{B11}_{2018Q3}, ..., Y^{B11}_{2023Q2}\},$$

where we stack all the data from within each benchmark period and the last observation date for which there has been a benchmark release is 2023Q2. For observation dates after that, I would use the latest-available data in empirical exercises. I call this the Benchmark-Consistent Approach.

Benchmark revisions seem to 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 later 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. A researcher attempting to improve the forecasts would need to use this approach.

Hypothesis 3: Vintage-Specific Approach. Suppose both the F and G functions disrupt the data-generating process, but the F_1, F_2, and F_3 functions are similar over time. Then forecasters would optimally relate initial, second, and first-final releases to each other. I call this the Vintage-Specific Approach.

Under the Vintage-Specific approach, as proposed initially by Koenig et al. (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 Vintage-Specific approach says that an appropriate model to use is one in which data that have not yet gone through an annual revision follow one data-generating process, while data that have been revised many times may follow a very different process. Under the Vintage-Specific 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 other vintage concepts in forming forecasts, but rather they divide data into vintages of different maturities.

Approaches Used in the Literature. In the forecasting literature, prior to the development of real-time data sets, most researchers used the Continuously-Updated Approach and did not account for data revisions at all. After the publication of Croushore and Stark (2001), researchers began considering issues of data revision. After that, for analyzing forecast bias, most of the literature used the Vintage-Specific method with initial releases, second releases, or first-final releases. These include Croushore (2010), who used first-final data with the Vintage-Specific approach, and showed how to improve forecasts with that approach only when the p-value of the bias test was less than 0.05; Kishor and Koenig (2012), who used the Vintage-Specific approach with various different data vintages, using the last release as the realized value; Coibion and Gorodnichenko (2015), who used the Vintage-Specific approach with realized values as one-year later than one-year-ahead forecasts, so generally they used A1 annual revisions with some benchmark revisions included, depending on the exact date; Bordalo et al. (2020), who used the Vintage-Specific approach with initial release realized values; Clements (2022), who used the Vintage-Specific approach with realized values as first- or second-release values to create efficiency-corrected forecasts, similar to what I do in this paper; and Eva and Winkler (2023), who use the Vintage-Specific approach with initial-release realized values to analyze whether forecasts can be improved. Thus the current paper is more general, analyzing approaches other than the Vintage-Specific approach and considering a variety of other realized values, and accounting for instability with the FR test, as described next.

Instability. A second difficult issue is that the forecasts might be unbiased 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:

$$Y_t = \alpha + \beta y_t + \epsilon_t. \tag{4}$$

But suppose the true data generating process is

$$Y_t = \alpha_t + \beta_t y_t + \epsilon_t. \tag{5}$$

Time variation in either the α or β terms will lead to apparent bias or inefficiency in the forecasts based on Equation (4). I will use the forecast-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. In both cases, the researchers used only the Vintage-Specific Approach. 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.

The analysis in this paper is unique in two aspects. First, it is one of few analyses to compare and contrast forecast evaluations using the three different approaches: Continuously-Updated, Benchmark-Consistent, and Vintage-Specific, each of which is based on a different hypothesis about how data revisions change the data-generating process. Second, it is the only paper to use and compare these approaches based on the forecast-rationality test of Rossi and Sekhposyan (2016), in the context of forecast-improvement exercises.

3. Testing for Bias in Real Time

Data. To illustrate the theory of how to improve biased forecasts, I examine forecasts from the U.S. Survey of Professional Forecasters (SPF) for corporate profits as a share of output. The survey records the forecasts of a large number of private-sector forecasters.[7] The literature studying the SPF forecasts has found that the SPF forecasts outperform macroeconomic models, even fairly sophisticated ones, as shown by Ang et al. (2007). The SPF has also been found to influence household expectations, as shown by Carroll (2003). I handle the complication of data revisions by using the real-time data set (RTDSM) 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.[8] Corporate profits and output have been included in the SPF since its inception in 1968. However, in early years, the data were not reported accurately for all horizons, so I begin the analysis using the SPF forecast for the first quarter of 1971. There are many horizons for the SPF, and in this paper I choose to focus on

[7] Details on the SPF can be found in Croushore and Stark (2019). Data can be found at Federal Reserve Bank of Philadelphia (1990-2024*b*).

[8] 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/`. Data can be found at Federal Reserve Bank of Philadelphia (1999-2024*a*).

the current-quarter horizon; that is, forecasts of the share of corporate profits in GDP for the same quarter in which the SPF survey is taken. This avoids problems of overlapping observations, as longer horizons entail adjustment because multiple forecasts are susceptible to the same shock. The SPF variable for corporate profits changed in the 2006Q1 survey from the overall corporate profits measure (not including IVA and CCAdj), to the one including IVA and CCAdj, so I collect data on both from the RTDSM. Also, because corporate profits are reported with a one-month or two-month lag after the initial output release in the NIPAs, I cannot use the initial or second release of the NIPA data, but use the first-final release of output combined with the release of corporate-profits data at the same time as the earliest vintage of the corporate-profits share.

I begin by looking at the forecasts and realizations in Figure 1, followed by a graph of the forecast errors in Figure 2. The figures are based on using the first-final 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.

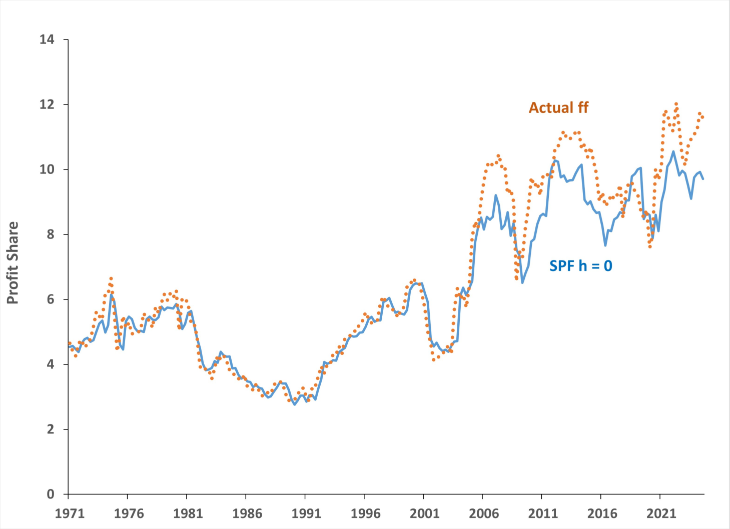
3.1. Results of Tests for Unbiasedness over Full Sample

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.[9]

I run the zero-mean-forecast-error test for inflation using all four versions of realized values, using data from the most recent data set available in January 2025, that is, following the Continuously-Updated Appraoch. The results of this

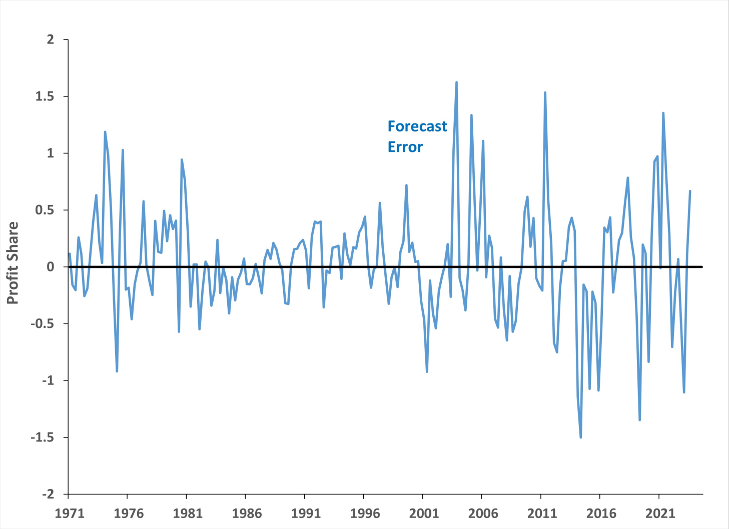
[9] I follow most of the forecasting literature in testing for bias under the assumption of a loss function for which bias is undesirable. Bias could be optimal, as in Elliott et al. (2008), if the loss function of forecasters is asymmetric.

Figure 1: Mean Current-Quarter Corporate Profits Share Forecasts and Realized Values



Note: The graph shows current-quarter corporate profit share forecasts from the SPF (labeled "SPF h=0") and realized values based on the first-final data release (labeled "Actual ff"). The dates shown on the horizontal axis are the dates on which the forecasts were made, ranging from 1971Q1 to 2024Q3. Note some large and persistent differences between forecasts and realized values.

Figure 2: Mean Current-Quarter Corporate Profits Share Forecast Errors



Note: The graph shows current-quarter corporate profit share forecast errors. The dates shown on the horizontal axis are the dates on which the forecasts were made, ranging from 1971Q1 to 2024Q3. Note some large forecast errors and some persistent errors.

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 first-final realized values, there is no evidence of statistically significant bias in the forecasts, but for other measures of realized values, the bias is statistically significant. In terms of magnitudes, for first annual realized values and pre-benchmark realized values, the bias is modest, about 0.2 percentage points, over a period when the corporate profits share ranged from about 4 percent to 12 percent. However, using the last realized value, the bias is very large, at 1.2 percentage points.

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, however the mean errors, p-values, and standard errors all differ slightly from the period

Table 1: Test for Bias, Based on Mean SPF Current-Quarter Corporate Profit Share Forecasts, Full Sample, Continuously-Updated Approach at Sample End

Realized Value	Mean Error	Standard Error	p-value
First final	0.046	0.032	0.157
First annual	0.203	0.044	0.000
Pre-benchmark	0.197	0.047	0.000
Last	1.211	0.060	0.000

Note: The table shows the results of the zero-mean forecast-error test for corporate profit share forecasts using the four different alternative measures of realized values. The sample uses SPF forecasts from 1971Q1 to 2023Q4, which is the last forecast date for which all four measures of realized values are available. The p-value is a standard t-test for the null hypothesis that the mean forecast error is zero.

that includes COVID. Mean forecast errors are lower for most measures of realizations (except last) in the pre-COVID sample. Not surprisingly, standard errors are lower for all realizations in the pre-COVID sample. But the differences across the pre-COVID and full samples are not nearly as large as for other variables, so for the remainder of this paper, I will analyze the full period.

Table 2: Test for Bias, Based on Mean Current-quarter SPF Corporate Profit Share Forecasts, Continuously-Updated Approach for Pre-COVID Sample

Realized Value	Mean Error	Standard Error	p-value
First final	0.046	0.032	0.148
First annual	0.163	0.044	0.000
Pre-benchmark	0.177	0.047	0.000
Last	1.232	0.064	0.000

Note: The table shows the results of the zero-mean forecast-error test for corporate profit share forecasts using the four different alternative measures of realized values. The sample uses SPF forecasts from 1971Q1 to 2018Q4. The p-value is a standard t-test for the null hypothesis that the mean forecast error is zero.

3.2. Tests for Unbiasedness in Sub-Samples

To implement tests for unbiasedness in sub-samples, we use the approach of Rossi and Sekhposyan (2016). The idea is that bias measures in the full sample in Table 1 might be masking bias that could be forecast across sub-samples. The Fluctuation-Rationality Test is robust the the presence of instabilities across sub-samples. The FR test statistic is the test statistic from the bias test regression of the forecast error on a constant, but using modified critical values that account for multiple testing in rolling windows.

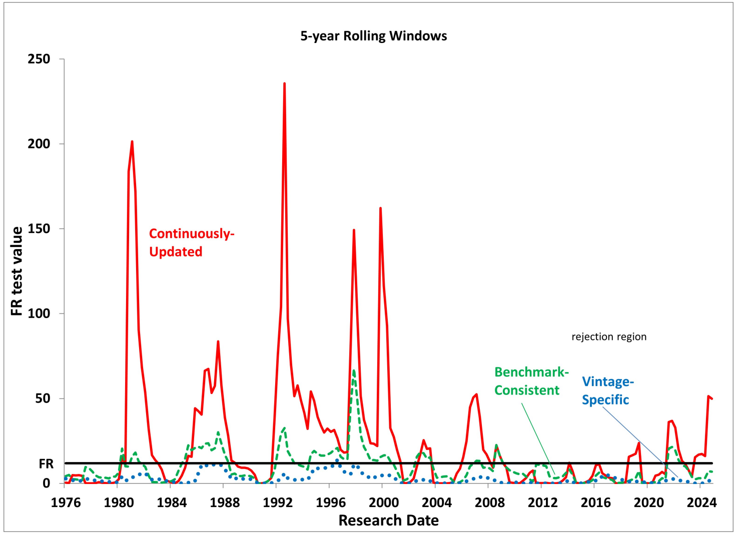
The plan here is to run the Forecast-Rationality test in 5-year and 10-year rolling windows for all three approaches (Continuously-Updated, Benchmark-Consistent, Vintage-Based). Critical values for the test account for the rolling nature of the test windows and adjust for multiple testing across windows. The critical values depend on the sample size and length of the window. In the data I use, the critical value of the FR test is 11.83 for 5-year windows, and 10.56 for 10-year windows, based on the table in Rossi and Sekhposyan (2016). An FR test value greater than the critical value in any rolling window means a lack of forecast rationality. Based on that, my goal is to see if I can exploit the lack of forecast rationality to improve on the SPF forecast.

Imagine a researcher standing at different points of time and trying to improve on the SPF forecast. The researcher would need to pick one of the three approaches and a choice of realized value. Note that each different approach (Continuously-Updated, Benchmark-Consistent, Vintage-Based) will have a different measure of bias estimates over time because the past realized values will differ, and in some cases the researcher might consider alternatives measures of realized values, as well.

Tests Using the Continuously-Updated Approach. Running the bias test regressions with rolling 5-year windows at each date from 1976Q1 to 2024Q4, leads to the results shown in Figure 3 labeled “Continuously-Updated”. The solid red line shows the p-value for the test of the null hypothesis of unbiasedness, that is, testing whether k_h in Equation (2) $= 0$. The horizontal axis shows the research date at which each test was performed in our simulated experiment. The results show many periods in which the test rejects the null hypothesis of unbiasedness.

If we repeat this exercise for ten-year rolling windows, as in Figure 4, we see similar results, consistent with bias in sub-samples and we should be able to

Figure 3: FR Test in Rolling Five-Year Windows, Continuously-Updated Approach

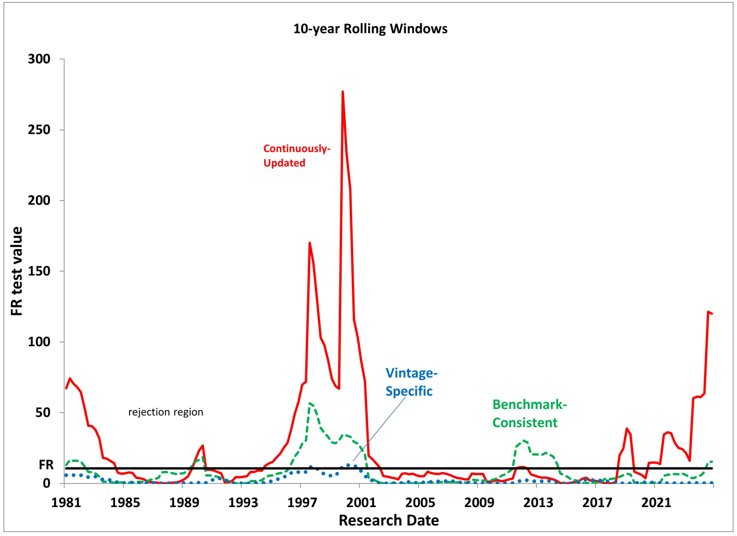


Note: The graph shows results of the FR test in rolling 5-year windows, based on all three approaches. The date shown on the horizontal axis is the research date, at which a researcher is standing, using the past five years of data, ranging from 1976Q1 to 2024Q4.

improve the forecasts.

Tests Using the 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 1982. 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

Figure 4: FR Test in Rolling Ten-Year Windows, Continuously-Updated Approach



Note: The graph shows results of the FR test in rolling ten-year windows, based on all three approaches. The date shown on the horizontal axis is the research date, at which a researcher is standing, using the past ten years of data, ranging from 1981Q1 to 2024Q4.

forecasts made from 1974Q4 to 1979Q3, and use the current vintage of data from the end of April 1982 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 1976Q1 to 2024Q4, gives the results shown in the dashed green line Figure 3. As was the case with the Continuously-Updated approach, the Benchmark-Consistent approach shows bias in many parts of the sample, and thus we reject unbiasedness overall. It should be possible to improve on the forecasts, though the degreee of unbiasedness is much less than was the case with the Continuously-Updated Approach.

Following the same procedure for ten-year windows gives results shown in Figure 4. As was the case with five-year windows, ten-year windows show bias and it should be possible to improve on the forecasts, though again showing fewer periods of bias than was the case with the Continuously-Updated Approach.

Tests Using the Vintage-Specific Approach. The idea of using real-time vintages (Vintage-Specific) 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 first-final releases to each other.

Following this procedure and simulating what a researcher would have done in testing for bias at every research date from 1976Q1 to 2024Q4, using first-final releases at each date, gives the results shown in the dotted blue lines in Figure 3. As was the case with the other two approaches, the Vintage-Specific approach finds bias. In this case, however, there is only one very short period in which the FR tests rejects the null hypothesis of unbiasedness.

Following the same procedure for ten-year windows gives results shown in Figure 4. As was the case with five-year windows, ten-year windows show bias and it should be possible to improve on the forecasts. As with five-year windows, the Vintage-Specific approach shows many fewer rejections than for the other two approaches. Thus, the Vintage-Specific Approach is much less likely to be useful for forecast improvement than the other two approaches.

4. Forecast-Improvement Exercises for Bias in Real Time

A problem in the literature on forecast evaluation is that many researchers find bias 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 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.[10]

4.1. Forecast Improvement for Bias Using Continuously-Updated Approach

To improve the forecasts, given that the Continuously-Updated approach showed bias in numerous sub-samples, I estimate the bias in rolling samples, then create

[10] For an early example, see Faust et al. (2003).

a new and improved forecast from the survey forecast, as in Equation (3).

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 (3), 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 (3) leads to a lower $RMSFE$ and an improved forecast; an $RRMSFE$ greater than one means that the attempt to improve the forecast failed. The p-value for the test of a significant difference in $RMSFEs$, shown in square brackets, is based on the Harvey et al. (1997) modification of the Diebold and Mariano (1995) test.[11] The second column repeats 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 (3) 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. In fact, in both cases, the $RMSE$ is over 30 percent worse. The p-values for the Diebold-Mariano test of equal forecast accuracy are both 0.000, so the improved forecasts are statistically significantly worse than the SPF forecasts.

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 bias is not statistically significant. It may be that the attempt to improve upon the forecasts, even in periods when the bias is not statistically significant, introduces noise into the forecast-improvement attempt, leading to higher $RMSEs$. A plot of the forecasts, realized values, and improved forecasts, in

[11] This test is valid for fixed rolling windows, despite the presence of parameter estimation error. For other methods, such as using expanding windows, the ideal test has not been fully developed, as suggested by Clark and McCracken (2009).

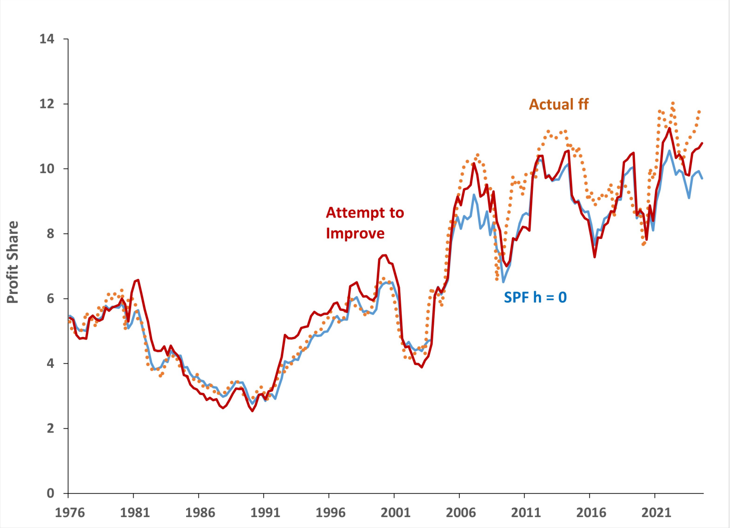
Table 3: *RRMSFEs* and *P*-values for Forecast Improvement Exercises Based on Estimates of Bias, Continuously-Updated Approach with Realized Values = First Final

Window Size:	5-year	10-year
Adjust every period	1.390 [0.000]	1.333 [0.000]
Adjust when FR test rejects	1.342 [0.000]	1.292 [0.000]
	With Shrinkage	
Adjust every period	1.111 [0.001]	1.093 [0.005]
Adjust when FR test rejects	1.097 [0.001]	1.089 [0.002]

Note: The table shows relative-root-mean-squared errors ($RRMSFE$) and p-values of the Harvey et al. modification of the Diebold-Mariano test [in square brackets] for corporate-profit share forecasts in forecast-improvement exercises, using the Continuously-Updated approach with realized values = first final. The sample consists of one-year-ahead SPF forecasts made at dates from 1971Q1 to 2023Q4.

Figure 5 shows that this seems to be the case.

Figure 5: Plot of Forecast Improvement



Note: The graph shows current-quarter corporate profit share forecasts from the SPF (labeled “SPF h=0”) and realized values based on the first-final data release (labeled “Actual ff”), along with the results of the attempt to improve on the SPF forecasts (labeled “Attempt to Improve”). The dates shown on the horizontal axis are the dates on which the forecasts were made, ranging from 1976Q1 to 2023Q4.

To remedy this, consider estimating bias in real time but adjusting the forecast using Equation (3) only if the forecast-rationality test showed rejection.[12] I will apply Equation (3) only in periods when the forecast rationality test is rejected. That is, the row in the table labeled “Adjust every period” uses the equation:

$$Y^I_{T,h} = Y^f_{T,h} + \hat{k}_h. \tag{6}$$

But, more generally, we modify this equation to:

$$Y^I_{T,h} = Y^f_{T,h} + \delta_t \hat{k}_h, \tag{7}$$

[12] An alternative is to adjust only when we reject the null hypothesis of zero-mean forecast error. In my experiments, that procedure generally reduces the $RRMSFE$. But basing the adjustment on the FR test instead leads to much lower $RRMSFE$s, so in the interest of space, I only report the latter.

where $\delta_t = 1$ when $FR_t >$ c.v., else $\delta_t = 0$.

The results of this exercise are shown in the row in Table 3 labeled "Adjust when FR test rejects." Compared with the first row, the $RRMSFE$s are slightly lower, but the attempt to improve on the forecasts still makes them significantly worse.

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.[13] But, more generally, we modify this equation to:

$$Y^I_{T,h} = Y^f_{T,h} + \delta_t \hat{k}_h, \tag{8}$$

where $\delta_t = 0.5$ when I "adjust every period, with shrinkage"; or $\delta_t = 0.5$ when $FR_t >$ c.v., else $\delta_t = 0$, when I "adjust when FR test rejects, with shrinkage".

The results show that shrinkage always helps. The attempt to improve on the forecasts still makes them statistically significantly worse, but they are only about 10 percent worse (rather than about 30 percent worse) when I use shrinkage.

Overall, using the Continuously-Updated approach, the attempt to improve the SPF forecasts makes them significantly worse.

4.2. Forecast Improvement for Bias Using the Benchmark-Consistent Approach

If I repeat the steps above, but use the Benchmark-Consistent approach, I obtain similar results to using the Continuously-Updated approach, as can be seen in Table 4. Adjusting the forecasts every period gives slightly better results than for the Continuously-Updated case. But basing adjustment on the FR test, or using shrinkage, is helpful. In the case of five-year windows with shrinkage, based on the FR test, the forecasts are only worse by about 3 percent, and the difference in $RMSEs$ is not statistically significant at the 5 percent level.

[13] 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.

Table 4: $RRMSFEs$ and P-values for Forecast Improvement Exercises Based on Estimates of Bias, Benchmark-Consistent Approach with Realized Values = First Final

Window Size:	5-year	10-year
Adjust every period	1.227 [0.000]	1.168 [0.000]
Adjust when FR test rejects	1.106 [0.004]	1.123 [0.003]
	With Shrinkage	
Adjust every period	0.067 [0.009]	1.055 [0.018]
Adjust when FR test rejects	1.030 [0.063]	1.041 [0.037]

Note: The table shows relative-root-mean-squared errors ($RRMSFE$) and p-values of the Harvey et al. modification of the Diebold-Mariano test [in square brackets] for inflation forecasts in forecast-improvement exercises, using the Benchmark-Consistent approach with realized values = first final. The sample consists of one-year-ahead SPF forecasts made at dates from 1971Q1 to 2023Q4.

4.3. Forecast Improvement for Bias Using Vintage-Specific Approach

Finally, I use the Vintage-Specific approach, with the first-final release of the data to determine the forecast error, with results in Table 5.

Table 5: *RRMSFEs* and *P*-values for Forecast Improvement Exercises Based on Estimates of Bias, Vintage-Specific Approach with Realized Values = First Final

Window Size:	5-year	10-year
Adjust every period	1.048 [0.010]	1.033 [0.003]
Adjust when FR test rejects	1.001 [0.319]	1.0017 [0.654]
	With Shrinkage	
Adjust every period	1.016 [0.088]	1.013 [0.017]
Adjust when FR test rejects	1.000 [0.319]	1.000 [0.968]

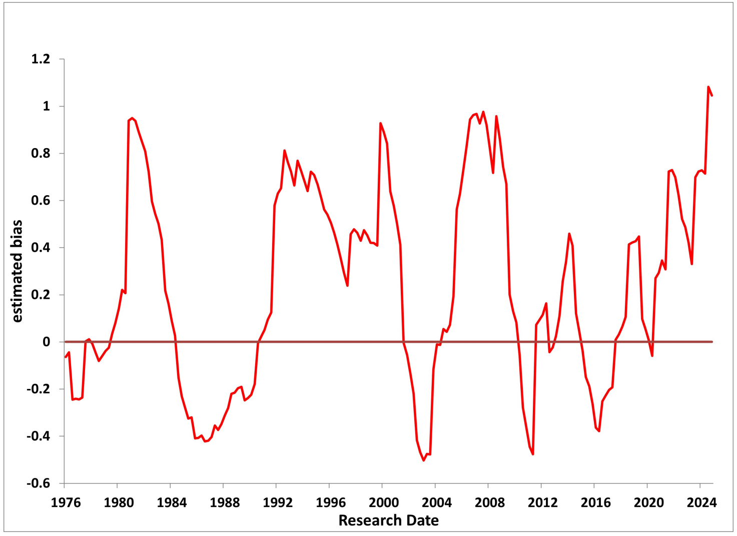
Note: The table shows relative-root-mean-squared errors ($RRMSFE$) and p-values of the Harvey et al. modification of the Diebold-Mariano test [in square brackets] for inflation forecasts in forecast-improvement exercises using the Vintage-Specific approach. The sample consists of one-year-ahead SPF forecasts made at dates from 1971Q1 to 2023Q4.

With the Vintage-Specific approach, adjusting every period gives better results than it did for the Continuously-Updated or Benchmark-Consistent methods, but the forecasts are still worse by 5 percent for five-year windows or 3 percent for ten-year windows. The outcome improves when using shrinkage, or adjusting only when the FR test rejects, or both. But in no cases is the $RRMSE$ less than 1, so there is no improvement.

4.4. Why Is Forecast Improvement for Bias So Difficult?

Though adjusting forecasts only when the FR test is violated, and using shrinkage methods, both help the forecast-adjustment process, the clear violation of unbiasedness in-sample should mean it is possible to improve the forecasts. The difficult problem to solve, however, is that the bias is changing dramatically over time, as Figure 6 shows.

Figure 6: Estimated Bias in Rolling Five-Year Windows, Continuously-Updated Approach



Note: The graph shows the estimated bias over time in rolling 5-year windows, based on the Continuously-Updated Approach. The date shown on the horizontal axis is the research date, at which a researcher is standing, using the past five years of data, ranging from 1976Q1 to 2024Q4.

4.5. Forecast Improvement for Bias Using Different Realized Values

So far, I used only first-final realized values as the object being forecast. But what if our goal is to forecast a later realized value? Could we then make improvements on SPF forecasts? For example, under Hypothesis 1 above, when each subsequent measure of the data gets closer to the truth, using the last realized value is the ideal measure.

If we use last realized values with the Vintage-Specific Approach, that would be equivalent to the Continuously-Updated Approach, as we would be taking the last data as the best measure at each date. So, I repeat here the exercises shown in Tables 3 and 4, but using the last realized values.

Using the Continuously-Updated Approach, with realized values equal to the last vintage, gives the results shown in Table 6.

Table 6: *RRMSFEs* and P-values for Forecast Improvement Exercises Based on Estimates of Bias, Continuously-Updated approach with realized values = last

Window Size:	5-year	10-year
Adjust every period	1.018 [0.311]	0.934 [0.002]
Adjust when FR test rejects	1.004 [0.790]	0.981 [0.262]
	With Shrinkage	
Adjust every period	0.997 [0.716]	0.954 [0.000]
Adjust when FR test rejects	0.992 [0.290]	0.981 [0.019]

Note: The table shows relative-root-mean-squared errors ($RRMSFE$) and p-values of the Harvey et al. modification of the Diebold-Mariano test [in square brackets] for corporate-profit share forecasts in forecast-improvement exercises, using the Continuously-Updated approach with realized values = last. The sample consists of one-year-ahead SPF forecasts made at dates from 1971Q1 to 2023Q4.

In Table 6, most of the $RRMSFEs$ are less than one, so the attempt to improve on the forecasts works. For ten-year rolling windows, adjusting every period, the improvement is nearly 7 percent of the $RMSFE$, and is statistically significant. Other statistically significant forecast improvements occur in ten-year windows with shrinkage.

Repeating the same exercise using the Benchmark-Consistent approach leads to similar results, as Table 7 shows, though the degree of forecast improvement is not as large as it was using the Continuously-Updated Approach.

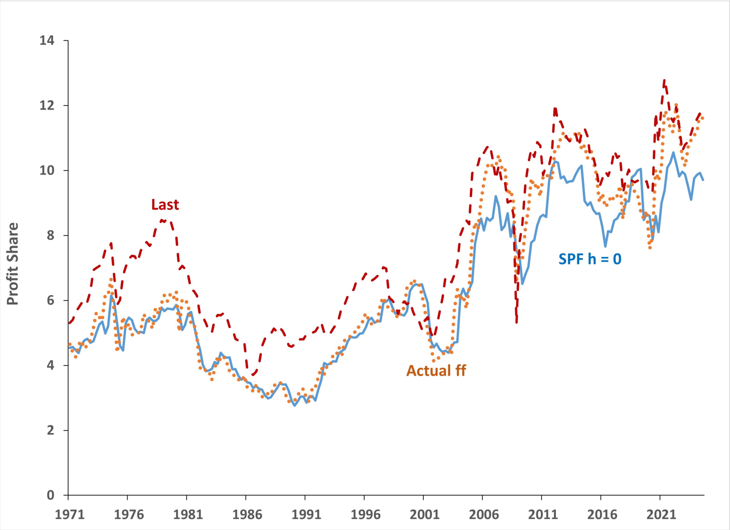
Table 7: $RRMSFEs$ and P-values for Forecast Improvement Exercises Based on Estimates of Bias, Benchmark-Consistent Approach with Realized Values = Last

Window Size:	5-year	10-year
Adjust every period	0.996 [0.807]	0.977 [0.030]
Adjust when FR test rejects	1.106 [0.102]	0.997 [0.626]
	With Shrinkage	
Adjust every period	0.992 [0.284]	0.983 [0.003]
Adjust when FR test rejects	1.006 [0.241]	0.995 [0.135]

Note: The table shows relative-root-mean-squared errors ($RRMSFE$) and p-values of the Harvey et al. modification of the Diebold-Mariano test [in square brackets] for inflation forecasts in forecast-improvement exercises, using the Benchmark-Consistent approach with realized values = last. The sample consists of one-year-ahead SPF forecasts made at dates from 1971Q1 to 2023Q4.

Why might forecast improvement be possible using the last release of the realized values, but not when using the first-final release? It may be that early releases of the corporate profits data are not optimal forecasts of later data. Figure 7 supports that idea by repeating Figure 1 but adding a line showing the last release realized values. As the figure shows, the corporate profits share has been broadly revised up over time, often long after the first-final release was made. Figuring out the source of those long-term revisions is worth exploring in additional research but beyond the scope of this paper.

Figure 7: Plot of Forecasts, First-Final Realized Values, and Last Realized Values



Note: The graph shows current-quarter corporate profit share forecasts from the SPF (labeled "SPF h=0") and realized values based on the first-final data release (labeled "Actual ff") and last release (labeled "Last"). The dates shown on the horizontal axis are the dates on which the forecasts were made, ranging from 1971Q1 to 2023Q4.

5. Summary and Conclusions

The goal of this paper was to create a systematic method for improving forecasts to reduce bias. I examined three different approaches for analysis, with differing assumptions about the data-generating process related to how data are revised: Continuously-Updated, Benchmark-Consistent, and Vintage-Specific. I considered optimal ways to account for data revisions and instability. I developed forecast-improvement exercises, employing shrinkage. When forecasts exhibit bias, I found no ability to improve forecasts out-of-sample based on first-final realized values. However, when using last vintage realized values, forecast improvement is possible and sometimes statistically significant.

Why might some in-sample results show a relationship between macroeconomic variables and forecast errors, but out-of-sample results often 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 et al. (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 et al. (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 for bias and testing for the efficiency of forecasts with respect to other information in the forecasters' information sets; (3) Determining the optimal degree of shrinkage to use in forecast-improvement exercises; (4) Finding methods to help forecasters find and understand structural breaks; and (5) 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.

Acknowledgments

Thanks for helpful suggestions to Evan Koenig, Mike McCracken, Sergey Slobodyan, Julie Smith, and Shaun Vahey, as well as participants from Lafayette College, Conference on Computing in Economics and Finance, Society for Economic Measurement, and the Conference on Real-Time Data Analysis, Methods, and Applications.

References

Ang, A., Bekaert, G. and Wei, M. (2007), 'Do macro variables, asset markets, or surveys forecast inflation better?', *Journal of Monetary Economics* **54**, 1163–1212.

Aruoba, S. B. (2008), 'Data revisions are not well behaved', *Journal of Money, Credit, and Banking* **40**, 319–340.

Bordalo, P., Gennaioli, N., Ma, Y. and Shleifer, A. (2020), 'Overreaction in macroeconomic expectations', *American Economic Review* **110**(9), 2748–2782.

Carroll, C. D. (2003), 'Macroeconomic expectations of households and professional forecasters', *Quarterly Journal of Economics* **118**, 269–298.

Clark, T. E. and McCracken, M. W. (2009), 'Tests of equal predictive ability with real-time data', *Journal of Business and Economic Statistics* **27**, 441–454.

Clements, M. P. (2022), 'Forecaster efficiency, accuracy, and disagreement: Evidence using individual-level survey data', *Journal of Money, Credit and Banking* **54**(2-3), 537–568.

Coibion, O. and Gorodnichenko, Y. (2015), 'Information rigidity and the expectations formation process: A simple framework and new facts', *American Economic Review* **105**(8), 2644–2678.

Croushore, D. (2010), 'An evaluation of inflation forecasts from surveys using real-time data', *B.E. Journal of Macroeconomics: Contributions* **10**(1).

Croushore, D. (2011), 'Frontiers of real-time data analysis', *Journal of Economic Literature* **49**(1), 72–100.

Croushore, D. (2019), 'Revisions to pce inflation measures: Implications for monetary policy', *International Journal of Central Banking* **15**(4), 241–265.

Croushore, D. and Stark, T. (2001), ‘A real-time data set for macroeconomists’, *Journal of Econometrics* **105**, 111–130.

Croushore, D. and Stark, T. (2019), ‘Fifty years of the survey of professional forecasters’, *Federal Reserve Bank of Philadelphia Economic Insights* pp. 1–11.

Cukierman, A., Lustenberger, T. and Meltzer, A. (2020), The permanent-transitory confusion: Implications for tests of market efficiency and for expected inflation during turbulent and tranquil times, *in* A. Arnon, W. Young and K. van der Beek, eds, ‘Expectations: Theory and Applications from Historical Perspectives’, Springer International Publishing, pp. 215–238.

Diebold, F. X. and Mariano, R. S. (1995), ‘Comparing predictive accuracy’, *Journal of Business and Economic Statistics* **13**, 253–263.

Elliott, G., Komunjer, I. and Timmermann, A. (2008), ‘Biases in macroeconomic forecasts: Irrationality or asymmetric loss?’, *Journal of the European Economic Association* **6**, 122–157.

Elliott, G. and Timmermann, A. (2008), ‘Economic forecasting’, *Journal of Economic Literature* **46**, 3–56.

Eva, K. and Winkler, F. (2023), A comprehensive empirical evaluation of biases in expectation formation. Working Paper, Federal Reserve Board.

Farmer, L. E., Nakamura, E. and Steinsson, J. (2024), ‘Learning about the long run’, *Journal of Political Economy* **132**(10), 3334–3377.

Faust, J., Rogers, J. H. and Wright, J. H. (2003), ‘Exchange rate forecasting: the errors we’ve really made’, *Journal of International Economics* **60**, 35–59.

Federal Reserve Bank of Philadelphia (1990-2024*b*), ‘Survey of professional forecasters’, https://www.philadelphiafed.org/surveys-and-data/real-time-data-research/survey-of-professional-forecasters.

Federal Reserve Bank of Philadelphia (1999-2024*a*), 'Real-time data set for macroeconomists', https://www.philadelphiafed.org/surveys-and-data/real-time-data-research/real-time-data-set-for-macroeconomists.

Giacomini, R. and Rossi, B. (2010), 'Forecast comparisons in unstable environments', *Journal of Applied Econometrics* **25**, 595–620.

Harvey, D. S., Leybourne, S. J. and Newbold, P. (1997), 'Testing the equality of prediction mean squared errors', *International Journal of Forecasting* **13**, 281–291.

Kishor, N. K. and Koenig, E. F. (2012), 'Var estimation and forecasting when data are subject to revision', *Journal of Business and Economic Statistics* **30**, 181–190.

Kishor, N. K. and Koenig, E. F. (2014), 'Credit indicators as predictors of real activity: A real-time var analysis', *Journal of Money, Credit and Banking* **46**, 545–564.

Kishor, N. K. and Koenig, E. F. (2022), 'Finding a role for slack in real-time inflation forecasting', *International Journal of Central Banking* **18**, 245–282.

Koenig, E., Dolmas, S. and Piger, J. (2003), 'The use and abuse of 'real-time' data in economic forecasting', *Review of Economics and Statistics* **85**, 618–628.

Mankiw, N. G. and Shapiro, M. D. (1986), 'Do we reject too often? small sample bias in tests of rational expectations models', *Economics Letters* **20**, 139–145.

Mincer, J. A. and Zarnowitz, V. (1969), The evaluation of economic forecasts, *in* J. Mincer, ed., 'Economic Forecasts and Expectations', National Bureau of Economic Research, New York.

Romer, C. D. and Romer, D. H. (2000), 'Federal reserve information and the behavior of interest rates', *American Economic Review* **90**(3), 429–457.

Rossi, B. (2006), 'Are exchange rates really random walks? some evidence robust to parameter instability', *Macroeconomic Dynamics* **10**, 20–38.

Rossi, B. (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, B. and Sekhposyan, T. (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, B. and Sekhposyan, T. (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, G. D. and Williams, J. C. (2009), 'Forecasting recessions: The puzzle of the enduring power of the yield curve', *Journal of Business and Economic Statistics* **27**(4), 492–503.

Stock, J. H. and Watson, M. W. (2003), 'Forecasting output and inflation: The role of asset prices', *Journal of Economic Literature* **41**, 788–829.

Su, V. and Su, J. (1975), 'An evaluation of asa/nber business outlook survey forecasts', *Explorations in Economic Research* **2**, 588–618.

Zarnowitz, V. (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 prebenchmark revisions.

In general terms, I 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 January 2025 and the last quarter for which data exist is the fourth quarter of 2024. So, the value of variable X for that date is denoted as:

$$X_{2024Q4}^{2025Q1,1}.$$

I will denote all the data in the last release (January 2025), which contains data from 1947Q1 to 2024Q4, as:[14]

$$X^{last} = \{X_{1947Q1}^{2025Q1,1}, X_{1947Q2}^{2025Q1,1}, X_{1947Q3}^{2025Q1,1}, ..., X_{2024Q2}^{2025Q1,1}, X_{2024Q3}^{2025Q1,1}, X_{2024Q4}^{2025Q1,1}\}.$$

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:

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

Thus, based on our earlier definition, $X^{last} = X^{2025Q1,1}$.

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

[14] 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 January 2025 are called the vintage of 2025M2 (February 2025) in the RTDSM.

always,[15] the first release for output and the price level occurred in the first month of the following quarter, so I denote a collection of all the initial releases as:

$$X^{initial} = \{X_{1965Q2}^{1965Q3,1}, X_{1965Q3}^{1965Q4,1}, X_{1965Q4}^{1966Q1,1}, ..., X_{2024Q2}^{2024Q3,1}, X_{2024Q3}^{2024Q4,1}, X_{2024Q4}^{2025Q1,1}\}.$$

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.

$$\begin{aligned} X^{FirstRevision} = \{&X_{1965Q2}^{1965Q3,2}, X_{1965Q3}^{1965Q4,2}, X_{1965Q4}^{1966Q1,2}, ..., \\ &X_{2024Q2}^{2024Q3,2}, X_{2024Q3}^{2024Q4,2}\}. \end{aligned}$$

Similarly, the first-final realized values use the data vintage from the third month of the following quarter.

$$\begin{aligned} X^{FirstFinal} = \{&X_{1965Q2}^{1965Q3,3}, X_{1965Q3}^{1965Q4,3}, X_{1965Q4}^{1966Q1,3}, ..., \\ &X_{2024Q2}^{2024Q3,3}, X_{2024Q3}^{2024Q4,3}\}. \end{aligned}$$

Annual revisions usually occur every year at the end of July. For example, the first annual revision of the data for 1965 was released at the end of July 1966 and recorded in the 1966M8 vintage of the RTDSM. The exceptions to this normal pattern are:

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

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
2023	2024M10

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

$$\begin{aligned}
X^{annual} = \{&X_{1965Q2}^{1966Q3,1}, X_{1965Q3}^{1966Q3,1}, X_{1965Q4}^{1966Q3,1},\\
&X_{1966Q1}^{1967Q3,1}, X_{1966Q2}^{1967Q3,1}, X_{1966Q3}^{1967Q3,1}, X_{1966Q4}^{1967Q3,1},\\
&...,\\
&X_{2020Q1}^{2021Q3,1}, X_{2020Q2}^{2021Q3,1}, X_{2020Q3}^{2021Q3,1}, X_{2020Q4}^{2021Q3,1},\\
&X_{2021Q1}^{2022Q3,1}, X_{2021Q2}^{2022Q3,1}, X_{2021Q3}^{2022Q3,1}, X_{2021Q4}^{2022Q3,1},\\
&X_{2022Q1}^{2023Q3,3}, X_{2022Q2}^{2023Q3,3}, X_{2022Q3}^{2023Q3,3}, X_{2022Q4}^{2023Q3,3},\\
&X_{2023Q1}^{2024Q3,3}, X_{2023Q2}^{2024Q3,3}, X_{2023Q3}^{2024Q3,3}, X_{2023Q4}^{2024Q3,3}\}.
\end{aligned}$$

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

Observation Dates	Pre-Benchmark Vintage
1965Q2 to 1975Q3	1975Q4,3
1975Q4 to 1980Q3	1980Q4,2
1980Q4 to 1985Q3	1985Q4,2
1985Q4 to 1991Q3	1991Q4,1
1991Q4 to 1995Q3	1995Q4,2
1995Q4 to 1999Q2	1999Q3,3
1999Q3 to 2003Q3	2003Q4,2
2003Q4 to 2009Q1	2009Q2,3
2009Q2 to 2013Q1	2013Q2,3
2013Q2 to 2018Q1	2018Q2,3
2018Q2 to 2023Q2	2023Q3,2

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, I use the last vintage available. The overall vector looks like:

$$\begin{aligned}
X^{pre-benchmark} = \{&X^{1975Q4,3}_{1965Q2}, X^{1975Q4,3}_{1965Q3}, X^{1975Q4,3}_{1965Q4}, \\
&X^{1975Q4,3}_{1966Q1}, X^{1975Q4,3}_{1966Q2}, X^{1975Q4,3}_{1966Q3}, X^{1975Q4,3}_{1966Q4}, \\
&..., \\
&X^{2023Q3,2}_{2023Q1}, X^{2023Q3,2}_{2023Q2}, X^{2025Q1,1}_{2023Q3}, X^{2025Q1,1}_{2023Q4}, \\
&X^{2025Q1,1}_{2024Q1}, X^{2025Q1,1}_{2024Q2}, X^{2025Q1,1}_{2024Q3}, X^{2025Q1,1}_{2024Q4}\}.
\end{aligned}$$