Improving Biased Forecasts in Real Time

By Dean Croushore*

March 4, 2026

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 inflation, using the Survey of Professional Forecasters. Even when bias is clear in-sample, it is often difficult to improve upon the forecasts out-of-sample. However, in some cases the methods show promise and lead to lower root-mean-squared forecast errors by 5 percent or more, though the improvement is never statistically significant.

JEL: E37, E17

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

* Professor of Economics and Rigsby Fellow, University of Richmond, Robins School of Business, 102 UR Drive, University of Richmond, VA 23173, dcrousho@richmond.edu. 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.

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 of unbiasedness 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?

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. What value does a researcher use as the realized value in determining the forecast error? 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 with those that have been revised multiple times based on different source data used by the government statistical agency?

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.[1]

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

In the forecasting literature, prior to the development of real-time data sets, most researchers 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 real-time data based on just the initial release, second release, or first-final release. These

[1]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. The Appendix to this paper provides precise definitions and notation for the realized values.

[2]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.

include Croushore (2010), who used first-final data 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 various different data vintages, using the last release as the realized value; Coibion and Gorodnichenko (2015), who used realized values as those from one year after the end of the forecast horizon; Bordalo et al. (2020), who used initial-release realized values; Clements (2022), who used realized values as first- or second-release values to create efficiency-corrected forecasts; and Eva and Winkler (2023), who use initial-release realized values to analyze whether forecasts can be improved. The current paper is more general, analyzing approaches that depend on the data-revision process, considering a variety of other realized values, and accounting for structural instability, as described next.

A difficult issue in research on analyzing forecasts 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), and 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?

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 approaches to test the extent to which, in practical circumstances, it is possible to improve upon forecasts. The first implementation of this in the literature is Faust, Rogers and Wright (2005), who find bias in initial releases of GDP data (which could be viewed as similar to forecasts)

in G7 countries and show that the in-sample bias estimate can be used to forecast revisions and improve upon the initial release of the GDP data. Croushore (2010) did a similar exercise for bias in forecasts, which he labeled Forecast-Improvement Exercises, for inflation forecasts from the Livingston and SPF surveys, but could not improve upon the survey forecasts. For bias in initial release data, Croushore (2019) found that initial releases of inflation based on the PCE price index were biased and could be improved upon. He also introduced the use of shrinkage in the forecast-improvement exercise. More recently, Eva and Winkler (2023) looked at many different in-sample findings of bias for many different forecast variables and found no ability to improve upon them. In these studies, the researcher must use real-time data and take the viewpoint of a forecaster standing at different points in time, running a bias test, then using the results of that test to make a better forecast, or in the case of initial data releases, to forecast the revision to the data. To test whether the attempt to improve upon the forecast works, one 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).

To illustrate the theory of how to improve biased forecasts, I examine forecasts from the U.S. Survey of Professional Forecasters (SPF) for inflation in the GDP deflator, for which the SPF has forecasts since it began in 1968. The survey records the forecasts of a large number of private-sector forecasters.[3] 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). 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

[3]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*).

U.S. Bureau of Economic Analysis.[4]

The main contribution of this paper is to develop a theoretical framework for handling data revisions, depending on the structure of the revision process. The paper proceeds in this way. In section 1, I develop three alternative structures: a continuously updated approach, a benchmark-consistent approach, and a vintage-specific approach. I show how the data production process determines when each approach should be used. I also show how the handle possible structural shifts. Following that, in section 2, I illustrate the theory using inflation data, testing for bias in real time, accounting for structural shifts. In section 3, I run forecast-improvement exercises based on those bias tests to see if the forecasts of inflation can be improved upon in real time. Section 4 provides conclusions and relates the results of the paper to the literature on why forecasts might be biased and how structural shifts might lead to such bias.

I. Theory of the Structure of Data Revisions

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.

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

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.

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".

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 coefficient in Equation (2) changes.

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.

Dealing with Data Revisions and Instability. Because data may be revised, a researcher must make a choice about which concept to use as the realized value of the variable from which to compute a forecast error. In addition, bias might not occur over the entire sample because of structural instability in the data-generating process or in the forecasting process.

A Structure of Data Revisions.

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:

(4)

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

If revisions to the data are small and white noise, the use of different concepts for realized values would be inconsequential.[5] 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.[6] 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

[5]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.

[6]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.

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 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).[7]

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 because more source data become available to the government statistical agency. In this case, it would be optimal for forecasters to use the latest-available data at each date, such as data downloaded from FRED or some similar database. I call this the Continuously Updated Approach.

If forecasters use this approach, a researcher evaluating their forecasts 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

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

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, and if the government statistical agency is getting new source data all the time, and if their benchmark revisions are not forecastable and do not change the data-generating process.

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.

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, 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 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, forecasters do not use other vintage concepts in forming forecasts, but rather they divide data into vintages of different maturities.

Instability. In addition to handling data revisions, attempts to improve biased forecasts must deal with structural instability. Suppose, for example, that a forecaster estimates a forecasting model based on the equation:

$$Y_t = \alpha + \beta y_t + \epsilon_t. \tag{5}$$

But suppose the true data-generating process is

$$Y_t = \alpha_t + \beta_t y_t + \epsilon_t. \tag{6}$$

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

in the forecasts based on Equation (5). 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, the paper by 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. 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.

II. Testing for Bias in Inflation Forecasts in Real Time

Data. As discussed in the introduction, I am testing SPF forecasts of the GDP deflator. Data on mean forecasts of the deflator are reported in the SPF beginning with the fourth quarter of 1968.[8] However, the deflator forecasts in the early years

[8]While some arguments can be made that testing bias is best done by examining the forecasts of individual forecasters, (see Keane and Runkle (1990)) 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

of the survey were not reported to enough significant digits,[9] 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.[10]

There are many horizons for the SPF, and in this paper 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 short quarterly horizon.

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

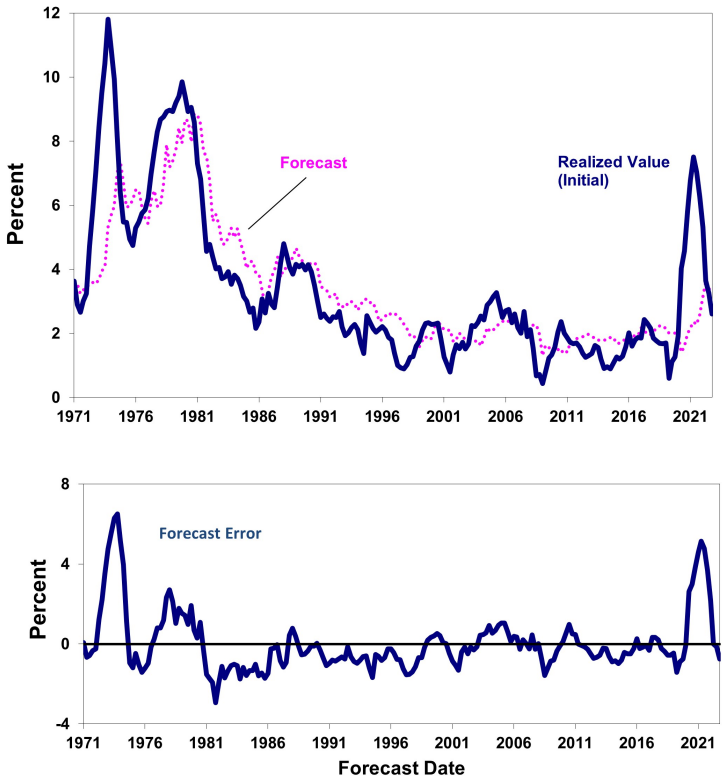
In this paper, our focus is on tests for the unbiasedness of forecasts. In the literature on forecast bias, the standard test is the Mincer and Zarnowitz (1969)

that the SPF survey has many missing observations, which is problematic. The mean and median across forecasters are almost identical in the SPF. This paper reports results based on the mean forecast.

[9] In particular, in the early years of the survey, forecasts for the deflator were reported to the nearest whole number, with no decimal points. This error, which leads to a sawtooth pattern of inflation forecasts, was not rectified until the survey taken in 1970Q4.

[10] Doing so leaves only one four-quarter-ahead forecast missing and eliminates the period with not enough decimal places in the forecasts. To handle the one missing four-quarter-ahead forecast, I extrapolate the three-quarter-ahead forecast for that one forecast that was made in 1974Q3.

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.

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.[11]

I run the zero-mean-forecast-error test for inflation using 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.

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
Second	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,

[11] 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, Komunjer and Timmermann (2008), if the loss function of forecasters is asymmetric.

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
Second	−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 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. Standard errors are adjusted following the Newey and West (1987) procedure.

B. 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. Their Fluctuation-Rationality (FR) Test is robust to 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 FR 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.[12] 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.

[12]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 11.75 for 5-year windows, and 9.96 for 10-year windows, based on Table II in Rossi and Sekhposyan (2016).

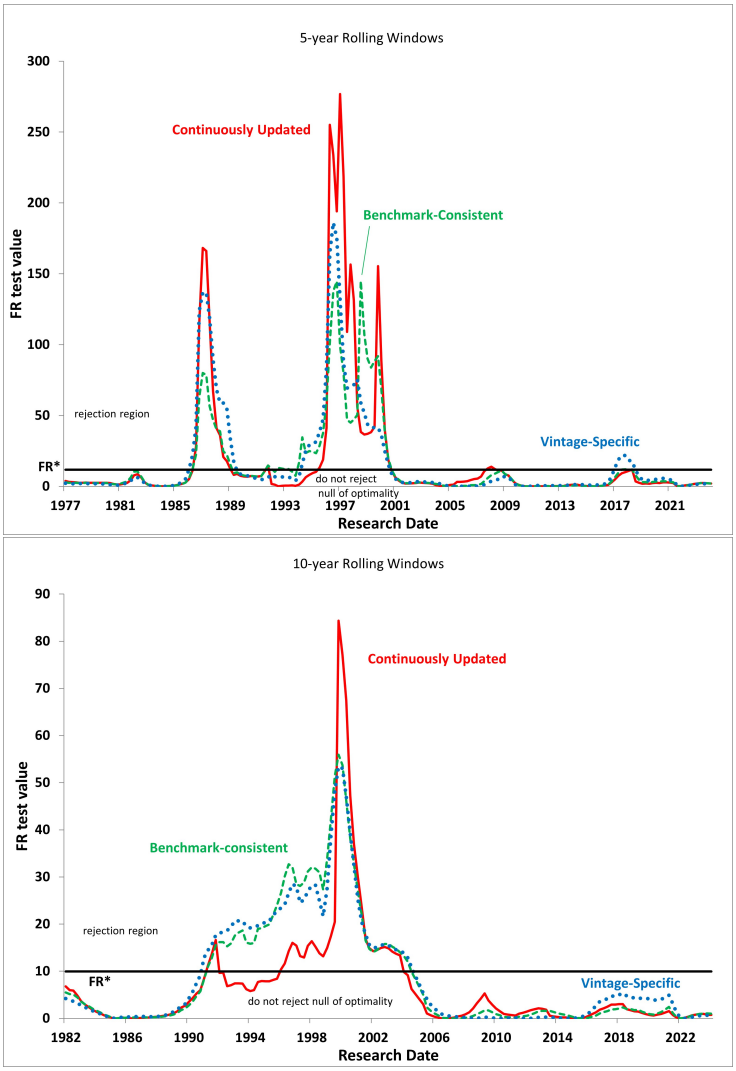
Imagine a researcher standing at different points of time and trying to improve on the SPF forecast. The researcher would need to pick one of the three approaches and a choice of realized value. Note that each different approach (Continuously Updated, Benchmark-Consistent, Vintage-Specific) will have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

Running all three approaches (Continously Updated, Benchmark-Consistent, and Vintage-Specific) for rolling 5-year windows and 10-year windows gives us the forecast-rationality tests shown in Figure 2. If the FR test value exceeds the critical value anywhere in the sample period, the forecast is not rational. The figure shows numerous cases in which the FR test value exceeds the critical value, for both 5-year windows in the top panel and 10-year windows in the bottom panel, for all three methods. 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 forecast-rationality tests reject the null of unbiasedness. Rejections are fewer for the Continuously Updated 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 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.

FIGURE 2. FORECAST-RATIONALITY TESTS IN FORECASTS FOR INFLATION IN ROLLING 5-YEAR AND 10-YEAR WINDOWS



Note: The upper panel shows the values of the Forecast-Rationality test for inflation using rolling 5-year samples of data. Each line corresponds to a different approaches: the Continuously Updated approach, the Benchmark-Consistent approach, or the Vintage-Specific 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 researcher would have data on realized values at the end of the rolling window) are shown on the horizontal axis.

A. 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 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, Leybourne and Newbold (1997) modification of the Diebold and Mariano (1995) test.[13] 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. However, the p-values are all above 0.05, meaning that the difference in $RMSFEs$ is not statistically significant. Still, the attempt to improve on the forecasts made them about 13 percent worse using 5-year rolling windows and 23 percent worse using 10-year rolling windows.

Part of the reason for the poor performance of these attempts at forecast improve-

[13]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 = INITIAL

Window Size:	5-year	10-year
Adjust every period	1.122	1.215
	[0.38]	[0.24]
Adjust when FR test rejects	1.080	0.990
	[0.29]	[0.82]
	With Shrinkage	
Adjust every period	0.991	1.066
	[0.89]	[0.46]
Adjust when FR test rejects	1.000	0.979
	[0.99]	[0.35]

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 Continuously Updated approach with realized values = initial. The sample consists of one-year-ahead SPF forecasts from 1980Q4 to 2018Q4. Note that the inflation $RMSFE$ = 0.794.

ment 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 $RMSEs$.

To remedy this, consider estimating bias in real time but adjusting the forecast using Equation (3) only if the forecast-rationality test showed rejection.[14] 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:

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

But, more generally, we modify this equation to:

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

where $\delta_t = 1$ 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 quite a bit lower. In 5-year windows, the $RRMSFE$ falls about 5 percentage points, so the forecasts were only 8 percent worse instead of 13 percent worse. For 10-year windows, the $RRMSFE$ falls about 24 percentage points, moving from a 23 percent worsening to a 1 percent forecast improvement. 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

[14] 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.

estimation.[15] Suppose I adjust for bias, but in Equation (8), set $\delta_t = 0.5$.[16] 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. Every value from the upper half of the table falls when I use shrinkage, and in two of the four cases, the new $RRMSFE$ is below one. The best case is a 2 percent forecast improvement for 10-year windows, adjusting only when the FR test rejects, using shrinkage. Overall, using the Continuously Updated approach, there is scope for improving the inflation forecasts, though in no case is the reduction in $RMSFE$ is significant.

B. 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 worse results than for the Continuously Updated case. But basing adjustment on the FR test, or using shrinkage, is helpful. 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 forecast-rationality test results both help to reduce the $RMSFEs$.

C. Forecast Improvement for Bias Using Vintage-Specific Approach

Finally, I use the Vintage-Specific 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

[15] This was first done by Croushore (2019).

[16] 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. Finding the optimal degree of shrinkage is part of my ongoing research.

TABLE 4—*RRMSFEs* AND *P*-VALUES FOR FORECAST IMPROVEMENT EXERCISES BASED ON ESTIMATES OF BIAS, BENCHMARK-CONSISTENT APPROACH WITH REALIZED VALUES = INITIAL

Window Size:	5-year	10-year
Adjust every period	1.134	1.249
	[0.38]	[0.25]
Adjust when FR test rejects	1.045	0.958
	[0.62]	[0.57]
	With Shrinkage	
Adjust every period	0.989	1.070
	[0.88]	[0.50]
Adjust when FR test rejects	0.971	0.943
	[0.43]	[0.17]

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 = initial. The sample consists of one-year-ahead SPF forecasts from 1980Q4 to 2018Q4.

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.

TABLE 5—*RRMSFEs* AND *P*-VALUES FOR FORECAST IMPROVEMENT EXERCISES BASED ON ESTIMATES OF BIAS, VINTAGE-SPECIFIC APPROACH WITH REALIZED VALUES = INITIAL

Window Size:	5-year	10-year
Adjust every period	1.062	1.168
	[0.64]	[0.37]
Adjust when FR test rejects	1.034	0.949
	[0.66]	[0.49]
	With Shrinkage	
Adjust every period	0.958	1.032
	[0.53]	[0.73]
Adjust when FR test rejects	0.967	0.940
	[0.35]	[0.15]

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 from 1980Q4 to 2018Q4.

With the Vintage-Specific approach, adjusting every period gives better results than it did for the Continuously Updated or Benchmark-Consistent approaches, but the forecasts are still worse by 7 percent for 5-year windows or 18 percent for 10-year windows. The outcome improves when using shrinkage, or adjusting only when the FR test rejects, or both. The best case, using 10-year windows, adjusting only when the FR test rejects, and using shrinkage, leads to a more than 6 percentage point improvement in $RMSFE$. 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.

IV. 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, the methods show promise and lead to lower root-mean-squared forecast errors by 5 percent or more, though the improvement is never statistically significant.

Why might 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, 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 to see if their forecasts are biased or inefficient; (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. Using these methods could be fruitful in helping forecasters make better forecasts.

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.

Aruoba, S. Boragan. 2008. "Data Revisions Are Not Well Behaved." *Journal of Money, Credit, and Banking*, 40: 319–340.

Bordalo, Pedro, Nicola Gennaioli, Yueran Ma, and Andrei Shleifer. 2020. "Overreaction in Macroeconomic Expectations." *American Economic Review*, 110(9): 2748–2782.

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.

Clements, Michael 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, 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: Contributions*, 10(1).

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*, 132(10): 3334–3377.

Faust, Jon, John H. Rogers, and Jonathan H. Wright. 2005. "News and Noise in G-7 GDP Announcements." *Journal of Money, Credit, and Banking*, 37: 403–419.

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, 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.

Kishor, N. Kundan, and Evan F. Koenig. 2014. "Credit Indicators as Predictors of Real Activity: A Real-Time VAR Analysis." *Journal of Money, Credit and Banking*, 46: 545–564.

Kishor, N. Kundan, and Evan F. Koenig. 2022. "Finding a Role for Slack in Real-Time Inflation Forecasting." *International Journal of Central Banking*, 18: 245–282.

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.

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.

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, 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 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:

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

I will denote all the data in the last release (March 2024), which contains data from 1947Q1 to 2023Q4, as:[17]

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

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:

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

[17] 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,[18] 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:

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

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.

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

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

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

Annual revisions usually occur every year at the end of July, with some exceptions, which I note in Table A1. 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

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

the RTDSM. The exceptions are given in Table A1.

Table A1—First Annual Revision Dates for Quarterly National Accounts, Exceptions to Normal Revision Dates

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

Note: For all other years, the first annual revision was released at the end of July of the following year, so appears in the August RTDSM.

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

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

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.

Table A2—Pre-Benchmark-Revision RTDSM Monthly Dates

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

Note: The table shows the pre-benchmark vintage date in the Real-Time Data Set for Macroeconomists. The benchmark revision vintage is one month after the pre-benchmark date.

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:

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