Can You Improve Upon the GDP Forecasts of Professional Forecasters Using Information About Monetary Policy?

By Dean Croushore*

November 12, 2024

In this paper, I examine the forecast errors of macroeconomic forecasters to see whether or not their forecasts are efficiently using information about monetary policy. The goal is to investigate, using real-time data, previous research that has found inefficiency in forecasts with respect to monetary policy. I use a real-time data set to investigate the relationship between GDP forecast errors and changes in monetary policy both in-sample and with out-of-sample methods. Out-of-sample results show that exploiting that inefficiency is difficult in real time.
JEL: E37
Keywords: real-time data, forecast efficiency, evaluating forecasts, macroeconomic forecasting, surveys

* Professor of Economics and Rigsby Fellow, University of Richmond, Robins School of Business, 102 UR Drive, University of Richmond, VA 23173, dcrousho@richmond.edu, 804-287-1961. I thank the University of Richmond for summer research support. Thanks for helpful comments to participants at the Society for Economic Measurement 2023, the Conference on Computing in Economics and Finance 2023, and the Monetary Policy sub-group of Macroeconomists at Liberal Arts Colleges 2024.

I. Introduction

Do forecasters optimally change their forecasts of GDP growth in response to changes in monetary policy? This question has been answered by a few papers in the literature but mostly using in-sample methods and based on final, revised data. In this paper, I examine the question in a more convincing manner, using real-time data to account more accurately for data revisions, using out-of-sample methods to examine the robustness of in-sample results, and exploring how inefficiency changes over time.

There is a vast literature on the evaluation of forecasts. Point forecasts are evaluated most often by examination of tests of unbiasedness and efficiency. The literature in this area was summed up most clearly by Clark and Mertens (2024), who suggest that forecasts from surveys of professional forecasters are "competitive (albeit not fully optimal) predictors of future outcomes." Recent research has suggested a number of problems with the forecasts of professionals. Bordalo et al. (2020) finds that the consensus of forecasts from a survey under-react to news, while individual forecasters over-react, and develop a model of dispersed information to explain it. Clements (2022) shows that individual forecasters are inefficient in their use of information. Bianchi, Ludvigson and Ma (2022) find that individual forecasters suffer from belief distortions but that artificial intelligence algorithms can be used to improve their forecasts. Eva and Winkler (2023), however, find that the research on forecast errors is not very robust and cannot be used to improve on the forecasts in a true real-time out-of-sample experiment. I follow the recent structure of Croushore (2024) to explore whether or not the forecasts can be improved out-of-sample in real-time using end-of-sample or real-time vintage methods (see Koenig, Dolmas and Piger (2003)), accounting for structural instability (see Rossi and Sekhposyan (2010).

The literature suggests that GDP forecasts may not respond appropriately to shocks to monetary policy. Several papers, Ball and Croushore (2003) and Rudebusch and Williams (2009), show that forecasters do not modify their GDP forecasts properly when monetary policy changes. We examine the results of these papers to see how their results hold up when we extend the analysis to include real-time out-of-sample tests.

II. Data

In this paper, I examine forecasts from the Survey of Professional Forecasters (SPF), which is widely studied.[1] I examine forecasts for real output growth, measured as GNP before 1992 and GDP from 1992 on. The forecasts are made quarterly and the survey asks the respondents to forecast the growth of real output in the current quarter and each of the following four quarters. I examine each of the quarterly annualized forecasts as well as the average output growth forecast over the next four quarters.

Quarterly forecasts for output growth (at an annualized rate) are calculated as in Equation ((1)):

$$y^e_{t,t+h} = (((\frac{Y^e_{t,t+h}}{Y^e_{t,t+h-1}})^4) - 1) \times 100\%, \tag{1}$$

where $h = 0$, 1, 2, 3, and 4, and $Y^e_{t,t+h}$ is the level of the output forecast made at date t for date $t + h$, using data on output through date $t - 1$.

For testing purposes, I compare those forecasts to realized values, which are calculated as

$$y_{t+h} = ((\frac{Y_{t+h}}{Y_{t+h-1}})^4 - 1) \times 100\%. \tag{2}$$

The definition of realized values is discussed below. The forecast error is the

[1]The SPF is the only quarterly survey of U.S. macroeconomic forecasters available at no charge, and has been produced on a quarterly basis since 1968. See Croushore and Stark (2019) for a historical discussion of the SPF and the research that uses it.

realized value of the growth rate minus the forecast

$$e_{t,t+h} = y_{t+h} - y^{e}_{t,t+h}. \tag{3}$$

The average annual output growth rate forecast over quarters $t+1$ to $t+4$ is calculated in Equation ((4)):

$$y^{e4}_{t,t+4} = \left(\frac{Y^{e}_{t,t+4}}{Y^{e}_{t,t}} - 1\right) \times 100\%. \tag{4}$$

Realized values over the same period are

$$y^{4}_{t+4} = \left(\frac{Y_{t+4}}{Y_{t}} - 1\right) \times 100\%. \tag{5}$$

Thus forecast errors for average annual forecasts are equal to

$$e^{4}_{t} = y^{4}_{t+4} - y^{e4}_{t,t+4}. \tag{6}$$

Similarly, we can calculate the average annual forecast growth rate from quarters t to $t+3$ by lagging Equation ((4)) by one quarter; similarly for the realized values and forecast errors.

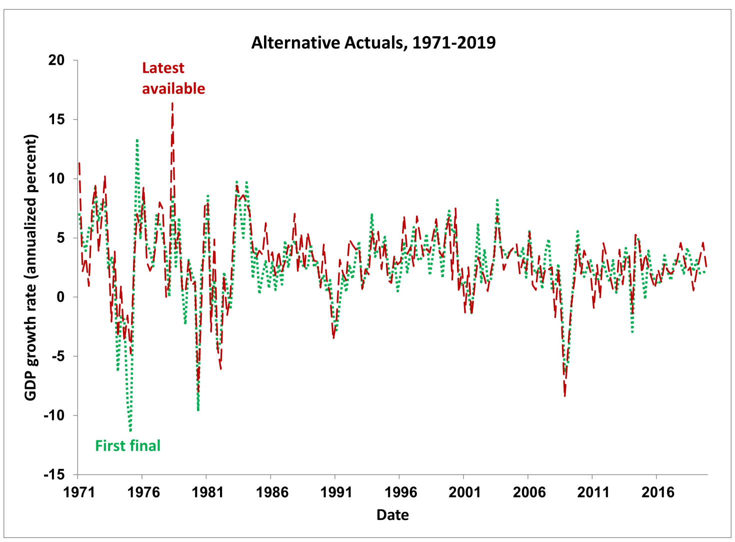
A key question in the forecasting literature is which vintage of the data to use as the realized value.[2] There are many alternatives and we explore differences across them, comparing initial realized values (the release at the end of the first month of the following quarter), to first final (FF) realized values (the release at the end of the third month of the following quarter), to first annual (A) realized values (the release at the end of July of the following year in most years), to pre-benchmark (B) realized values (the last release before a benchmark revision of the National Income and Product Accounts), to latest available (L) realized values

[2]See Croushore (2011) for a discussion of this issue.

(from the latest available vintage of data available when this research started, which was August 2024). I obtain the alternative realized values from the Real-Time Data Set for Macroeconomists (RTDSM), which was created by Croushore and Stark (2001) and made available on the website of the Federal Reserve Bank of Philadelphia. The RTDSM provides information on real output (GNP before 1992, GDP since 1992) and other major macroeconomic variables, as someone standing at the middle of any month from November 1965 to today would have viewed the data. The RTDSM lines up perfectly with the SPF in terms of data availability.

Figure 1 plots GDP growth rates for two of the alternative realized values, first final and latest available, from 1971Q1 to 2019Q4. You can see that the two series generally move together, but there are quarters when they differ substantially, in one case by over six percentage points. Thus, forecast evaluation conclusions potentially differ significantly depending on the choice of realized values.

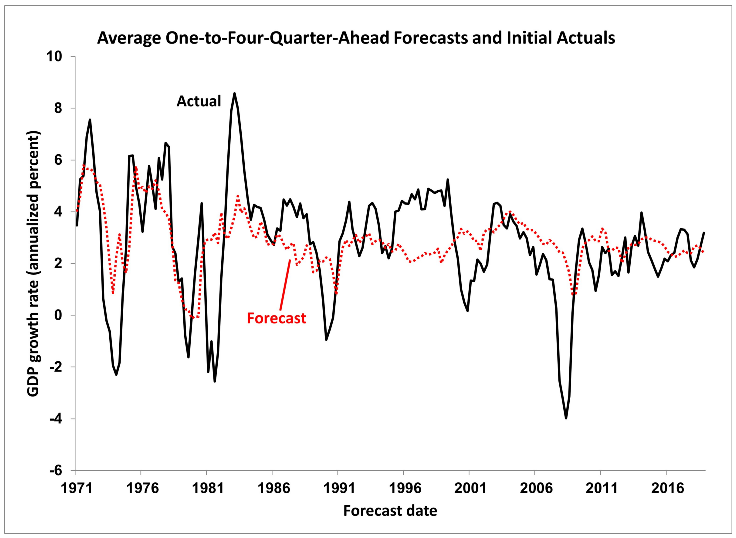
Figure 1. Alternative Realized Values



Note: The figure shows the quarterly realized values of GDP growth rates as calculated using Equation ((2)) based on two alternative concepts: first final and latest available. The graph ends prior to the COVID period, to avoid distortions caused by the large swings to GDP growth in 2020.

To visualize what the realized values and forecasts look like, Figure 2 shows a plot of the real GDP growth rate over the next four quarters and the initial realized value of GDP growth over the same horizon. The graph ends prior to the COVID period, to avoid distortions caused by the large swings to GDP growth in 2020; so it uses forecasts from 1971Q1 to 2018Q4, with the corresponding realized values (ending in 2019Q4).

Figure 2. Average One-to-Four Quarter Ahead Forecasts and Initial Realizations

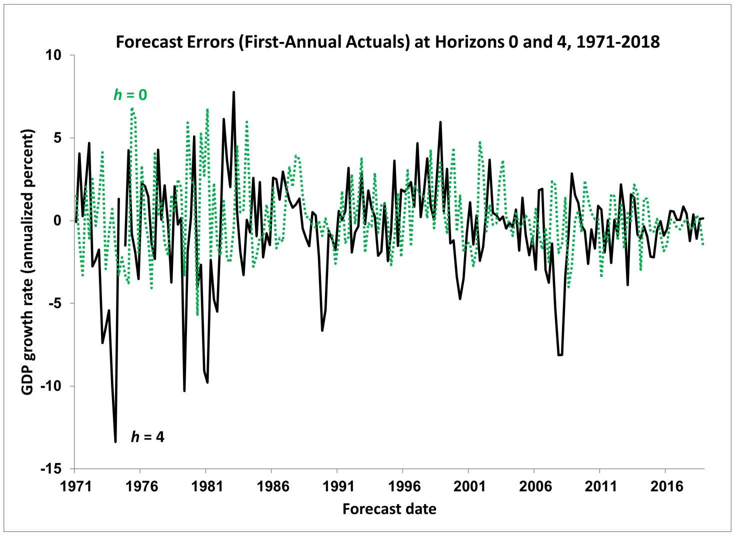


Note: The figure shows the forecast for the real GDP growth rate over the next four quarters and the initial realized value of GDP growth over the same horizon. The graph ends prior to the COVID period, to avoid distortions caused by the large swings to GDP growth in 2020; so it uses forecasts from 1971Q1 to 2018Q4, with the corresponding realized values (ending in 2019Q4 at the latest).

To provide a sense of the size of forecast errors, Figure 3 shows representative forecast errors based on the first-annual concept of realized values at quarterly horizons 0 and 4. The forecast errors are large and volatile, and they change signs frequently, making them difficult to predict. As we might expect, for $h = 4$, the forecast errors are more persistent because of the overlapping-observations problem, which induces a moving-average structure to the forecast errors.

To examine whether measures of monetary policy might be used to improve GDP forecasts, I consider three alternative measures of monetary policy: the

Figure 3. Forecast Errors at Horizons 0 and 4



Note: The figure shows the quarterly forecast errors for GDP growth rates as calculated using (3) for two horizons: current quarter ($h = 0$) and four quarters ahead ($h = 4$). The graph ends prior to the COVID period, to avoid distortions caused by the large swings to GDP growth in 2020; so it uses forecasts from 1971Q1 to 2018Q4, with the corresponding realized values (ending in 2019Q4).

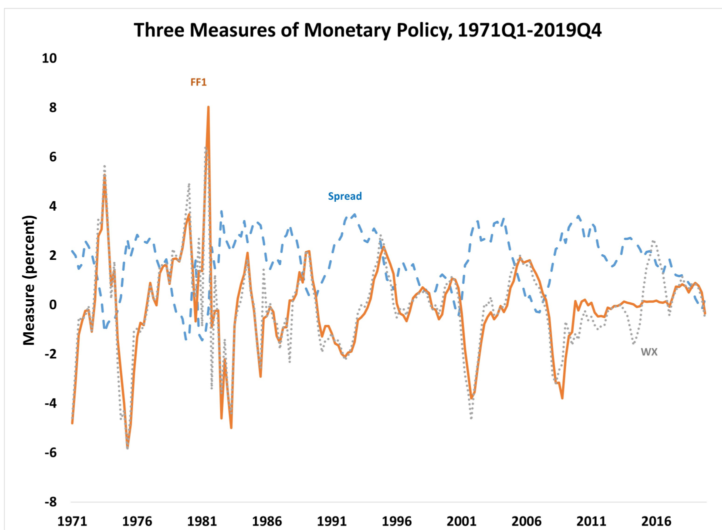
yield spread, changes in the real federal funds rate, and the Wu-Xia shadow real fed funds rate. For the yield spread, I use the measure of Rudebusch and Williams (2009), which is the interest rate on 10-year Treasury notes minus the interest rate on 3-month Treasury bills, using the constant-maturity series for each security. For the change in the real federal funds rate, I use the Ball and Croushore (2003) measure, which is the average federal funds rate in the previous quarter, minus the expected inflation rate over the coming year in the SPF.[3] Note that both the yield spread and the change in the real fed funds rate for the prior quarter are available to the SPF forecasters at the time they make their forecasts. I am careful to only use data available to the forecasters in these efficiency tests. However, in the late 2000s, the nominal federal funds rate became constrained by the effective lower bound on interest rates, so changes in the real federal funds rate may not be as useful as a measure of monetary policy as they were before. To

[3]Ball and Croushore (2003) examined alternatives to this measure and found that the results were not sensitive to the proxy used.

remedy that, we use the shadow real fed funds rate of Wu and Xia (2016), which accounts for nontraditional monetary policy tools and creates an effective federal funds rate based on the impact of those tools.[4] However, the Wu-Xia measure was not available to the forecasters in real time, so our results with this variable are indicative of what a researcher in real time might have found, but are not strictly a real-time result, so are valid only if the revisions to the variable are not large.

The three measures of monetary policy differ somewhat over time but their major movements are correlated, as you can see in Figure 4.

Figure 4. Three Measures of Monetary Policy



Note: The figure shows the three alternative measures of monetary policy that we use: the term spread between 10-year T-notes and 3-month T-bills (Spread), the change in the real fed funds rate over the previous year ($FF1$), and the change in the Wu-Xia measure of the real fed funds rate over the previous year (WX).

[4]Updated data on the Wu-Xia shadow rate are available online at www.atlantafed.org/cqer/research/wu-xia-shadow-federal-funds-rate.

III. In-Sample Results, Three Approaches

In this section, I investigate whether our three measures of monetary policy are correlated with forecast errors in-sample. The sample uses all SPF forecasts made from 1971Q1 to 2018Q4, so that four-quarter-ahead forecasts end before COVID begins in 2020.

First, we run a regression of each of the forecast errors for the seven horizons and four different measures of realizations for each of the three different measures of monetary policy. The regression is simply:

$$e_{t,t+h} = \alpha + \beta MP_{t-1} + \epsilon_t, \tag{7}$$

where MP_{t-1} is one of the measures of monetary policy through date $t-1$ (known to forecasters making their forecasts at date t) and $e_{t,t+h}$ is a forecast error from Equation (3) or (6). The results are summarized in Table 1.

Table 1—In-Sample Results for Monetary Policy

Horizon	0	1	2	3	4	1-4	0-3
Realized Value							
initial	x x x	x x x	M S M	S S S	S S S	S S S	x S S
first final	x x x	x x x	x M x	x S S	S S S	S S S	x S S
first annual	x x x	x M x	x S M	M S S	S S S	S S S	S S S
pre-benchmark	x x x	x M x	x S M	M S M	S S S	S S S	S S S
latest-available	S S S	x x x	x x x	S x x	S S S	S S x	S M x

Note:

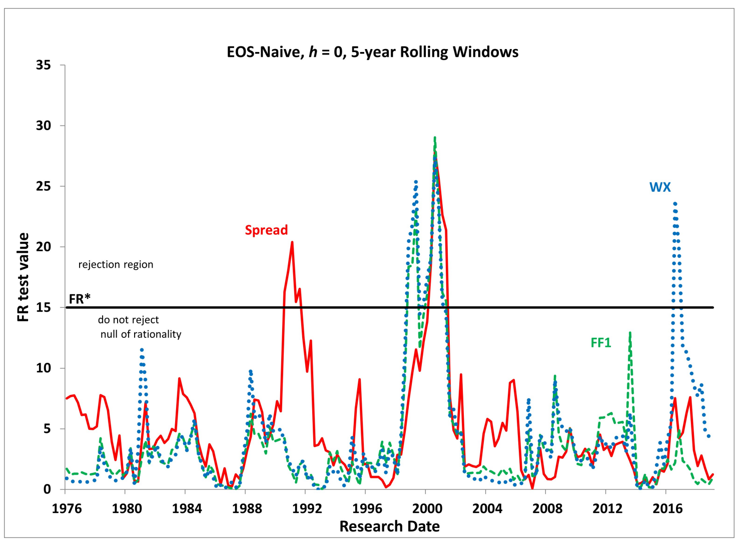
Results of joint test that $\alpha = 0$ and $\beta = 0$: x means p-value > 0.05; M means $0.05 < p$-value < 0.10; S means p-value ≤ 0.05
First term: yield spread; second term: lagged change in real fed funds rate; third term: lagged change in effective (Wu-Xia) real fed funds rate
The sample uses SPF forecasts from 1971Q1 to 2018Q4. Standard errors are adjusted following the Newey and West (1987) procedure.

In Table 1, we see that about half of all the cases show a statistically significant coefficient (p-value ≤ 0.05) in regression Equation (7), which suggests that forecasters are not using information about monetary policy efficiently in forming their forecasts. The coefficients on monetary policy are most often significant at longer horizons, which is consistent with the literature allowing for a lag in the effect of monetary policy on output. In terms of the alternative measures of realized values, the coefficients on monetary policy are more often significant for using first annual or pre-benchmark realized values. Coefficients using the Wu-Xia measure of monetary policy are somewhat less likely to be significant than the other two measures of monetary policy.

These in-sample results, however, are based on the full sample based on forecasts made from 1971Q1 to 2018Q4. They do not show how a researcher standing at different points in time would have perceived the inefficiency regressions. Croushore (2024) suggests three methods for viewing inefficiency in real time: End-Of-Sample naive (EOS-naive), End-Of-Sample benchmark-consistent (EOS benchmark-consistent), and Real-Time Vintages (RTV). We consider each of these in turn. For each of these methods, we look at the forecast-rationality statistics of Rossi and Sekhposyan (2016), using 5-year and 10-year rolling windows, with the idea being that even if the full-sample in-sample results do not show inefficiency, that may be because the inefficiencies in short periods offset each other. The method helps identify the periods of inefficiency.

In-Sample results for EOS-Naive Approach. In the End-Of-Sample-naive approach, we think about a researcher assuming that forecasters at each date use the latest data from FRED or a similar macroeconomic database. The assumption is that the researcher and forecasters ignore any effects of data revisions in evaluating and forming forecasts. So, imagine a researcher standing in 1976Q1, evaluating the current-quarter forecasts from the SPF from 1971Q1 to 1975Q4 (a five-year window), using the latest real GDP growth data from FRED to analyze the forecasts.[5] Then roll the exercise forward quarter by quarter, maintaining a five-year window each time. Do the same exercise for each of the three different measures of monetary policy and the seven different forecast horizons, allowing for a longer lag in data availability as the horizon lengthens. For each five-year window, calculate the forecast-rationality statistic and compare it with the critical value from Rossi and Sekhposyan (2016). The forecast-rationality statistic is calculated at each research date, and we reject the null hypothesis of forecast rationality if any of the values exceeds the critical value for any research date. The results of this exercise are shown in Figure 5.

Figure 5. Forecast-Rationality Test Results, EOS-Naive Approach, h = 0, 5-Year Rolling Windows



Note: The figure shows the forecast-rationality test results using the EOS-Naive approach with a 5-year rolling window and a horizon of zero. The forecast-rationality critical value is labeled FR*, and we reject forecast rationality if any value across the sample exceeds that threshold. Each line is labeled with the type of monetary policy used in Equation (7), where S is the yield spread, $FF1$ is the lagged change in real fed funds rate, and WX is the lagged change in the effective (Wu-Xia) real fed funds rate.

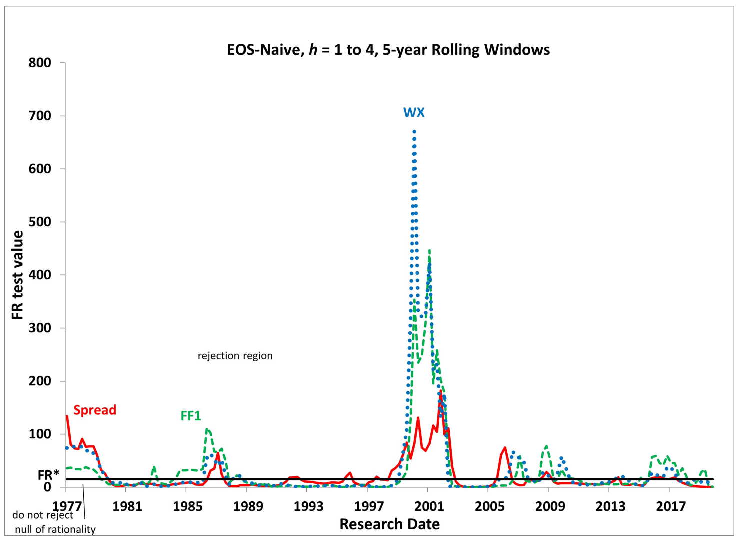
In Figure 5, we see a few rejections of the null hypothesis of forecast rationality, though not a huge number. We reject forecast rationality based on the yield

[5]The research date of 1976Q1 allows for a one-quarter lag for data availability from the last forecast made in 1975Q4.

spread in the early 1990s and based on all three measures in the late 1990s and the early 2000s. For the Wu-Xia version of FF1 there is a rejection in late 2016 to early 2017.

To compare with a longer-horizon forecast, Figure 6 shows the results of the forecast-rationality tests for the 1-to-4 quarter horizon.

Figure 6. Forecast-Rationality Test Results, EOS-Naive Approach, h = 1 to 4, 5-Year Rolling Windows



Note: The figure shows the forecast-rationality test results using the EOS-Naive approach with a 5-year rolling window and a horizon of 1 to 4 quarters. The forecast-rationality critical value is labeled FR*, and we reject forecast rationality if any value across the sample exceeds that threshold. Each line is labeled with the type of monetary policy used in Equation (7), where S is the yield spread, $FF1$ is the lagged change in real fed funds rate, and WX is the lagged change in the effective (Wu-Xia) real fed funds rate.

Figure 6 shows that for this longer horizon, rejections of forecast rationality occur much more frequently and are scattered throughout much of the sample.

Repeating this exercise for 10-year rolling windows leads to qualitatively similar results, though with fewer rejections of forecast rationality than for 5-year windows, as the figures in the Appendix show. For the current-quarter horizon, rejections occur only around 2000 to 2001 for all three measures of monetary policy. For the 1-to-4 quarter horizon, there are many rejections prior to 1990 but almost none after that. So, clearly rejections of forecast rationality depend on the horizon and the length of the rolling window. There are some differences across measures of monetary policy, as well.

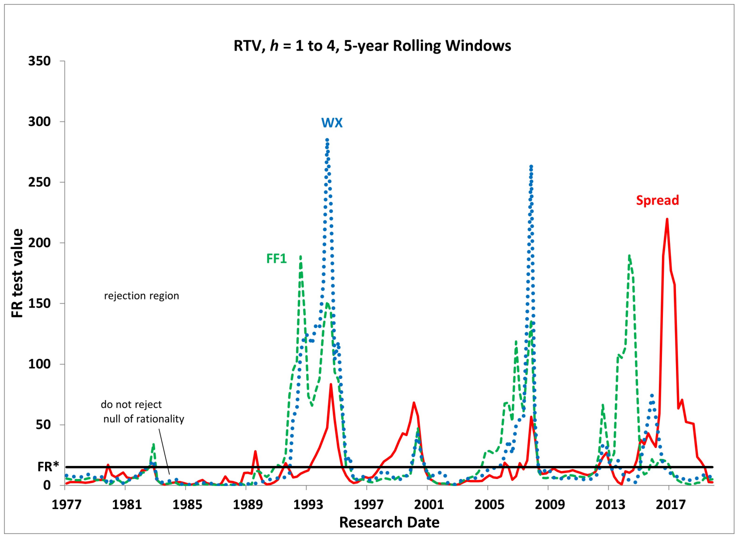
In-sample Results for EOS Benchmark-Consistent Approach. In the End-Of-Sample benchmark-consistent approach, we think about a researcher assuming that forecasters at each date string together pre-benchmark values of the data, with the idea that the DGP differs across benchmark revisions. The assumption is that the researcher and forecasters account for data revisions in evaluating and forming forecasts. The results of the exercises are discussed here; figures showing the results can be found in the Appendix to conserve space.

The forecast-rationality tests for the EOS benchmark-consistent approach show a similar pattern of rejections of forecast rationality as was the case for the EOS-naive approach for 5-year rolling windows with $h = 0$. For the $h = 1$ to 4 horizon, rejections of forecast rationality occur much more frequently than for $h = 0$ and are scattered throughout much of the sample, similar to the EOS-naive results. Results for the EOS benchmark-consistent approach in 10-year rolling windows are also similar to the EOS-naive approach.

In-sample Results for RTV Approach. In the real-time-vintage approach, we think about a researcher assuming that forecasters at each date look at data vintages with similar ages, with the idea that the DGP differs across concepts of realized values. In particular, especially for forecasting initial values, using just the initial release values in the forecasting model might be appropriate. The assumption is that the researcher and forecasters view the DGP as relating initial releases to each other over time. Results for the RTV approach differ somewhat from the EOS approaches, so I show the results here for the $h = 1$ to 4 horizon, with other results in the Appendix.

Figure 7 shows that for this longer horizon with the RTV approach, there are many rejections of forecast rationality. There are many more rejections than for the $h = 0$ horizon, and also many more than for the EOS approaches.

Figure 7. Forecast-Rationality Test Results, RTV Approach, h = 1 to 4, 5-Year Rolling Windows



Note: The figure shows the forecast-rationality test results using the RTV approach with a 5-year rolling window and a horizon of 1 to 4 quarters. The forecast-rationality critical value is labeled FR*, and we reject forecast rationality if any value across the sample exceeds that threshold. Each line is labeled with the type of monetary policy used in Equation (7), where S is the yield spread, $FF1$ is the lagged change in real fed funds rate, and WX is the lagged change in the effective (Wu-Xia) real fed funds rate.

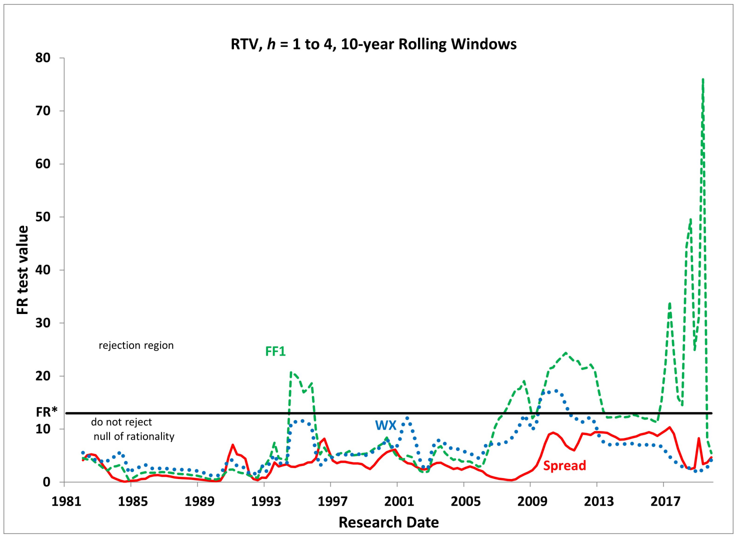
Repeating this exercise for 10-year rolling windows leads to many fewer rejections of forecast rationality than for 5-year windows, as Figure 8 shows.

IV. Forecast-Improvement Exercises for Inefficiency in Real Time

Given the in-sample results, we proceed to investigate the possibility of using the regression results from Equation (7) to improve upon the SPF forecasts in a simulated real-time out-of-sample exercise; we call this a forecast-improvement exercise (FIE). Taking the estimated $\hat{\alpha}$ and $\hat{\beta}$, and recalling from Equation (3) that $e_{t,t+h} = y^a_{t+h} - y^e_{t,t+h}$, we create, at each date t, an improved forecast $y^i_{t,t+h}$, where

$$y^i_{t,t+h} = y^e_{t,t+h} + (\delta^1_t \times \hat{\alpha}) + (\delta^2_t \times \hat{\beta} MP_{t-1}), \tag{8}$$

Figure 8. Forecast-Rationality Test Results, RTV Approach, h = 1 to 4, 5-Year Rolling Windows



Note: The figure shows the forecast-rationality test results using the RTV approach with a 10-year rolling window and a horizon of 1 to 4 quarters. The forecast-rationality critical value is labeled FR*, and we reject forecast rationality if any value across the sample exceeds that threshold. Each line is labeled with the type of monetary policy used in Equation (7), where S is the yield spread, $FF1$ is the lagged change in real fed funds rate, and WX is the lagged change in the effective (Wu-Xia) real fed funds rate.

where the δ terms will be described below. The baseline case has both δ terms equal to 1.

Using regression Equations (7) and (8), we simulate the activity of a real-time forecaster, forming improved forecasts at each date based only on the real-time data and past forecast errors available at each date.[6] We collect all the improved forecasts over that period and calculate RMSFEs for each different horizon and each different measure of monetary policy. We compare those RMSFEs to those of the SPF forecast, dividing the RMSFE of the attempt to improve on the survey by the RMSFE of the SPF, to generate a relative RRMSFE. An RRMSFE greater than one means the attempt to improve on the SPF forecasts actually made them worse, while an RRMSFE less than one means the attempt to improve on the SPF

[6]In the out-of-sample exercise, we use only data that the forecasters would have known in real time when the SPF survey results are released, using the data only up to the quarter prior to the SPF forecast. (The exception is for the Wu-Xia shadow rate because real-time data for it do not exist.) The coefficients of the regression are re-estimated at each date.

succeeded.[7]

I consider four different versions of Equation (8). The baseline case has $\delta_t^1 = 1$ and $\delta_t^2 = 1$ for all t. This method does not account for estimation error in the coefficients, however. To account for estimation error, we could shrink the estimated coefficients towards zero. As a simple first pass, we shrink the coefficients by half, so that $\delta_t^1 = 0.5$ and $\delta_t^2 = 0.5$ for all t. (I leave it for future research to determine optimal shrinkage in these forecast-improvement exercises.) An alternative is to use the information from the forecast-rationality tests to allow the δ terms to vary over time. One possibility is to set the δ terms to zero, if the FR-test value is less than the critical value at that date, or equal to one, if the FR-test value is greater than the critical value at that date. Another possibility is to use shrinkage with that method. So, I try all four adjustment methods to see how the results vary.

There are many results of these forecast-improvement exercises: 2 rolling window sizes (5 year and 10 year), 3 approaches (EOSn, EOSbc, RTV), 3 measures of monetary policy (S, FF1, WX), 7 horizons (0, 1, 2, 3, 4, 1 to 4, 0 to 3), and 4 adjustment methods (full, full with shrinkage, FR-test based, FR-test based with shrinkage, for a total of 504 sets of results. In what follows, I focus on the 1-to-4 quarter-ahead horizon, with other results shown in the Appendix.

Table 2 shows the results for the EOS-naive approach. In only 3 of the 24 cases is the RRMSFE below 1, and none of those cases are statistically significantly below 1. Most of the cases show that the attempt to improve the forecasts actually makes them worse, with 2 of them statistically significantly worse.

We can follow the same procedure for the EOS benchmark-consistent approach, as shown in Table 3, with results that are not too different from the EOS-naive approach.

I follow the same procedure using the RTV approach. For the 1-to-4 horizon,

[7]Statistical significance of differences between the surveys is tested using the Harvey, Leybourne and Newbold (1997) modified Diebold and Mariano (1995) test statistic of the corresponding null hypothesis.

TABLE 2—*RRMSFEs* AND P-VALUES FOR FORECAST IMPROVEMENT EXERCISES, EOS-NAIVE APPROACH WITH REALIZED VALUES = INITIAL, $h = 1 - 4$

Monetary Policy Measure	S		FF1		WX	
Window Size: Adjustment Method	5-yr	10-yr	5-yr	10-yr	5-yr	10-yr
All	1.048	1.005	1.048	1.013	1.060	1.007
	[0.02]	[0.80]	[0.13]	[0.08]	[0.29]	[0.41]
All, shrink	1.009	0.998	1.007	1.003	1.007	1.000
	[0.30]	[0.81]	[0.32]	[0.44]	[0.60]	[0.99]
$FR > cv$	1.030	0.999	1.047	1.000	1.063	1.000
	[0.05]	[0.16]	[0.14]	[0.32]	[0.27]	[0.32]
$FR > cv$, shrink	1.006	0.999	1.011	1.000	1.012	1.000
	[0.16]	[0.15]	[0.11]	[0.32]	[0.30]	[0.32]

Note: The table shows relative-root-mean-squared errors ($RRMSFE$) and p-values of the Diebold-Mariano test [in square brackets] for forecasts in forecast-improvement exercises, using the EOS-naive approach with realized values = initial. The sample consists of one-year-ahead SPF forecasts made from 1971Q1 to 2018Q4.

Table 3—*RRMSFEs* and *P*-values for Forecast Improvement Exercises, EOS benchmark-consistent approach with realized values = initial, h = 1 to 4

Monetary Policy Measure	S		FF1		WX	
Window Size: Adjustment Method	5-yr	10-yr	5-yr	10-yr	5-yr	10-yr
All	1.042	1.001	1.035	1.006	1.045	1.000
	[0.02]	[0.95]	[0.18]	[0.43]	[0.34]	[0.98]
All, shrink	1.007	0.997	1.002	1.000	1.002	0.997
	[0.36]	[0.67]	[0.70]	[0.96]	[0.82]	[0.40]
$FR > cv$	1.024	1.000	1.037	1.000	1.052	1.000
	[0.05]	[0.32]	[0.14]	[0.92]	[0.25]	[0.78]
$FR > cv$, shrink	1.005	1.000	1.008	1.000	1.010	1.000
	[0.25]	[0.32]	[0.15]	[0.26]	[0.28]	[0.36]

Note: The table shows relative-root-mean-squared errors ($RRMSFE$) and p-values of the Diebold-Mariano test [in square brackets] for forecasts in forecast-improvement exercises, using the EOS benchmark-consistent approach with realized values = initial. The sample consists of one-year-ahead SPF forecasts made from 1971Q1 to 2018Q4.

the results in Table 4 show a marginally better ability to improve on the SPF forecasts than is the case for the two EOS approaches. But of the 4 cases of forecast improvement, none are statistically significantly better. For the RTV approach, of the 16 cases of worse forecasts, none are statistically significantly worse.

TABLE 4—*RRMSFEs* AND *P*-VALUES FOR FORECAST IMPROVEMENT EXERCISES, RTV APPROACH WITH REALIZED VALUES = INITIAL, $h = 1 - 4$

Monetary Policy Measure	S		FF1		WX	
Window Size: Adjustment Method	5-yr	10-yr	5-yr	10-yr	5-yr	10-yr
All	1.090	1.056	1.032	1.009	1.038	1.019
	[0.14]	[0.07]	[0.29]	[0.66]	[0.10]	[0.16]
All, shrink	1.035	1.023	1.004	1.000	1.004	1.006
	[0.24]	[0.11]	[0.78]	[0.99]	[0.70]	[0.37]
$FR > cv$	1.028	1.000	1.021	0.983	1.019	0.999
	[0.13]	[1.00]	[0.20]	[0.48]	[0.08]	[0.28]
$FR > cv$, shrink	1.010	1.000	1.005	0.989	1.006	0.999
	[0.24]	[1.00]	[0.44]	[0.37]	[0.19]	[0.28]

Note: The table shows relative-root-mean-squared errors ($RRMSFE$) and p-values of the Diebold-Mariano test [in square brackets] for forecasts in forecast-improvement exercises, using the EOS benchmark-consistent approach with realized values = initial. The sample consists of one-year-ahead SPF forecasts made from 1971Q1 to 2018Q4.

To summarize the results for the 1-to-4-quarter horizon, Tables 2 to 4 show that the attempt to improve on the SPF forecasts is successful in fewer than 10 percent of the cases, and the improvement is never statistically significant. Improvement is more likely using the results of the forecast-rationality test as a guide to when to adjust the forecasts. For most cases, the attempt to improve on the SPF forecasts makes the forecasts worse, with a few cases in which the improved forecasts are statistically significantly worse than the original SPF forecasts.

For horizons other than the 1-to-4 quarter horizon, in what follows I show summary tables of the results, followed by some general description of which procedures are most likely to lead to better adjustments to the SPF forecasts.

We begin with an overall summary table, with the results of all 504 permutations of 2 rolling window sizes, 3 approaches, 3 measures of monetary policy, 7 horizons, and 4 adjustment methods. Table 5 shows the results.

TABLE 5—FORECAST IMPROVEMENT EXERCISES COUNTS OF RANGES, ALL PERMUTATIONS

RRMSE range	RRMSEs in range	p-value ≤ 0.05
1.10 to ∞	53	16
1.02 to 1.10	172	18
1.00 to 1.02	210	1
0.98 to 1.00	67	1
0.90 to 0.98	2	0
0.00 to 0.90	0	0

Note: The table shows the numbers of cases in which the relative-root-mean-squared error ($RRMSFE$) falls within the given range, and the number of cases with p-values of the Diebold-Mariano test that are less than or equal to 0.05, across all 504 permutations of 2 rolling window sizes, 3 approaches, 3 measures of monetary policy, 7 horizons, and 4 adjustment methods. The sample consists of SPF forecasts made from 1971Q1 to 2018Q4.

As Table 5 shows, there is only one statistically significant improvement to the forecasts; that one occurred with 10-year rolling windows, using the EOS benchmark-consistent approach, with annual forecasts from $h = 0$ to 3, adjusting all forecasts with shrinkage, and using the FF1 measure of monetary policy, which resulted in an $RRMSFE$ of 0.9905 and a p-value of 0.048. In two cases out of 504, there is a more than 2 percent improvement in the $RMSFE$, whereas in about 13 percent of the cases, there is a small improvement on the SPF forecasts of less than 2 percent of the SPF $RMSFE$. In the other 435 cases out of 504, the attempt to improve the forecasts makes them worse; they are statistically significantly worse in about 7 percent of all the cases. In about 10 percent of the cases, the attempt to improve on the forecasts made the RMSFE rise by more than 10 percent.

In the next set of tables, I generalize the results across different permutations, such as across horizons, approaches, rolling-window sizes, measures of monetary policy, and adjustment methods.

Table 6 shows the five quarterly horizons. Looking across the horizons, it appears most likely to find improved forecasts at the current-quarter horizon of $h = 0$. The most statistically significant cases of making the forecasts worse are for $h = 1$. Horizons $h = 3$ and $h = 4$ show the least likelihood of forecast improvement.

TABLE 6—FORECAST IMPROVEMENT EXERCISES COUNTS OF RANGES, ACROSS QUARTERLY HORIZONS

RRMSE	$h = 0$		$h = 1$		$h = 2$		$h = 3$		$h = 4$	
range	N	p	N	p	N	p	N	p	N	p
1.10 to ∞	3	3	9	4	14	6	19	3	7	0
1.02 to 1.10	19	2	31	6	31	3	29	2	25	3
1.00 to 1.02	39	0	32	0	22	0	21	0	39	1
0.98 to 1.00	11	0	0	0	5	0	0	0	1	0
0.90 to 0.98	0	0	0	0	0	0	0	0	0	0
0.00 to 0.90	0	0	0	0	0	0	0	0	0	0

Note: The table shows the numbers of cases in which the relative-root-mean-squared error ($RRMSFE$) falls within the given range, and the number of cases with p-values of the Diebold-Mariano test that are less than or equal to 0.05, across all 72 permutations of 2 rolling window sizes, 3 approaches, 3 measures of monetary policy, and 4 adjustment methods for each of the 5 quarterly forecast horizons. The sample consists of SPF forecasts made from 1971Q1 to 2018Q4.

Table 7 shows the two annual horizons. Looking across the horizons, it appears more likely to find improved forecasts at the annual horizon of $h = 0$ to 3, rather than $h = 1$ to 4. That horizon also has the only one of the 502 cases in which the $RMSFE$ is statistically significantly lower.

Table 8 shows results across the three different approaches (EOS-naive, EOS benchmark-consistent, and RTV). Looking across the approaches, it appears most likely to find improved forecasts using the EOSbc or RTV approaches, and less likely with the EOSn approach. The only statistically significant improvement to

TABLE 7—FORECAST IMPROVEMENT EXERCISES COUNTS OF RANGES, ACROSS ANNUAL HORIZONS

RRMSE range	h=1 to 4 N	h=1 to 4 p	h=0 to 3 N	h=0 to 3 p
1.10 to ∞	0	0	1	0
1.02 to 1.10	20	2	17	3
1.00 to 1.02	34	0	23	0
0.98 to 1.00	18	0	29	1
0.90 to 0.98	0	0	2	0
0.00 to 0.90	0	0	0	0

Note: The table shows the numbers of cases in which the relative-root-mean-squared error ($RRMSFE$) falls within the given range, and the number of cases with p-values of the Diebold-Mariano test that are less than or equal to 0.05, across all 72 permutations of 2 rolling window sizes, 3 approaches, 3 measures of monetary policy, and 4 adjustment methods for each of the 2 annual forecast horizons. The sample consists of SPF forecasts made from 1971Q1 to 2018Q4.

$RMSFE$ came from the EOSbs approach.

Table 9 shows the two different windows, 5-year and 10-year. Looking across the window sizes, it appears more likely to find improved forecasts using 10-year windows, and to find statistically significant worsening of forecasts with 5-year windows.

Table 10 shows results across the three different measures of monetary policy (S, FF1, and WX). Looking across the measures, it appears most likely to find improved forecasts using the $FF1$ or WX measures, and less likely with the spread measure. Using the spread measure is most likely to make the forecasts worse, often significantly so.

Table 11 shows the four different types of adjustments. Looking across the adjustment methods, it appears most likely to find improved forecasts using shrinkage. Full adjustment without shrinkage is particularly poor, and leads to statistically significant worsening of forecasts.

Table 8—Forecast Improvement Exercises Counts of Ranges, Across Approaches

RRMSE range	EOSn N	EOSn p	EOSbc N	EOSbc p	RTV N	RTV p
1.10 to ∞	20	8	17	5	16	3
1.02 to 1.10	61	8	58	8	53	2
1.00 to 1.02	69	0	66	0	75	1
0.98 to 1.00	18	0	27	1	22	0
0.90 to 0.98	0	0	0	0	2	0
0.00 to 0.90	0	0	0	0	0	0

Note: The table shows the numbers of cases in which the relative-root-mean-squared error ($RRMSFE$) falls within the given range, and the number of cases with p-values of the Diebold-Mariano test that are less than or equal to 0.05, across all 168 permutations of 2 rolling window sizes, 3 measures of monetary policy, 7 horizons, and 4 adjustment methods for each of the 3 approaches. The sample consists of SPF forecasts made from 1971Q1 to 2018Q4.

Table 9—Forecast Improvement Exercises Counts of Ranges, Across Window Sizes

RRMSE range	5-year N	5-year p	10-year N	10-year p
1.10 to ∞	52	15	1	1
1.02 to 1.10	123	4	49	4
1.00 to 1.02	63	0	147	1
0.98 to 1.00	4	0	53	1
0.90 to 0.98	0	0	2	0
0.00 to 0.90	0	0	0	0

Note: The table shows the numbers of cases in which the relative-root-mean-squared error ($RRMSFE$) falls within the given range, and the number of cases with p-values of the Diebold-Mariano test that are less than or equal to 0.05, across all 252 permutations of 7 horizons, 3 approaches, 3 measures of monetary policy, and 4 adjustment methods for each of the 2 window sizes: 5-year and 10-year. The sample consists of SPF forecasts made from 1971Q1 to 2018Q4.

Table 10—Forecast Improvement Exercises Counts of Ranges, Across Measures of Monetary Policy

RRMSE	S		$FF1$		WX	
range	N	p	N	p	N	p
1.10 to ∞	15	10	15	4	23	2
1.02 to 1.10	60	10	57	2	55	6
1.00 to 1.02	89	0	67	1	59	0
0.98 to 1.00	9	0	27	1	31	0
0.90 to 0.98	0	0	2	0	0	0
0.00 to 0.90	0	0	0	0	0	0

Note: The table shows the numbers of cases in which the relative-root-mean-squared error ($RRMSFE$) falls within the given range, and the number of cases with p-values of the Diebold-Mariano test that are less than or equal to 0.05, across all 168 permutations of 2 rolling window sizes, 3 approaches, 7 horizons, and 4 adjustment methods for each of the 3 measures of monetary policy. The sample consists of SPF forecasts made from 1971Q1 to 2018Q4.

Table 11—Forecast Improvement Exercises Counts of Ranges, Across Adjustment Methods

RRMSE	Full		Full, shrink		$FR > cv$		$FR > cv$, shrink	
range	N	p	N	p	N	p	N	p
1.10 to ∞	36	16	0	0	17	0	0	0
1.02 to 1.10	63	18	52	0	34	0	23	0
1.00 to 1.02	21	1	49	0	61	0	79	0
0.98 to 1.00	5	0	25	1	13	0	24	0
0.90 to 0.98	1	0	0	0	1	0	0	0
0.00 to 0.90	0	0	0	0	0	0	0	0

Note: The table shows the numbers of cases in which the relative-root-mean-squared error ($RRMSFE$) falls within the given range, and the number of cases with p-values of the Diebold-Mariano test that are less than or equal to 0.05, across all 72 permutations of 2 rolling window sizes, 3 approaches, 3 measures of monetary policy, and 7 horizons, for each of the 4 adjustment methods. The sample consists of SPF forecasts made from 1971Q1 to 2018Q4.

V. Summary and Conclusions

To summarize the results of this myriad of tests, I have shown that inefficiency holds in-sample, based on standard tests on the forecast errors. However, the attempt to improve on the SPF forecasts out of sample is generally not successful and sometimes makes the forecasts significantly worse. I accounted carefully for different approaches to thinking about the data-generating process, data revisions, and structural instability. The most promising avenues for forecast improvement seem to be using the RTV approach, the FF1 measure of monetary policy, a 10-year window size, adjusting using shrinkage, and using the current-quarter forecast or the annual forecast at the horizon from 0 to 3 quarters ahead.

Why might in-sample results show a relationship between macroeconomic variables and forecast errors, but out-of-sample results do not? It may be that forecasters do not recognize the importance of a variable for forecasting until some time passes, so there is an in-sample relationship that is not useful for forecasting for very long. Or it may take forecasters some time to adjust to structural shifts as they learn about the long run, as discussed by Farmer, Nakamura and Steinsson (2024). Or, as Cukierman, Lustenberger and Blinder (2018) suggest, a permanent-transitory confusion may lead to in-sample correlations, even if forecasters have rational expectations.

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 testing additional variables that might affect real GDP growth forecasts and modifying the degree of shrinkage or looking for optimal shrinkage.

REFERENCES

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

Bianchi, Francesco, Syndey C. Ludvigson, and Sai Ma. 2022. "Belief Distortions and Macroeconomic Fluctuations." *American Economic Review*, 112(7): 2269–2315.

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

Clark, Todd E., and Elmar Mertens. 2024. "Survey Expectations and Forecast Uncertainty." In *Handbook of Research Methods and Applications on Macroeconomic Forecasting.* , ed. M. Clements and A. Galvao. Edward Elger Publishing Ltd.

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.

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

Croushore, Dean. 2024. "Improving Forecasts in Real Time." University of Richmond working paper.

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 Blinder. 2018. “The Permanent-Transitory Confusion: Implications for Tests of Market Efficiency and For Expected Inflation During Turbulent and Tranquil Times.” Working Paper, Tel Aviv University.

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

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.

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.

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.

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

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.

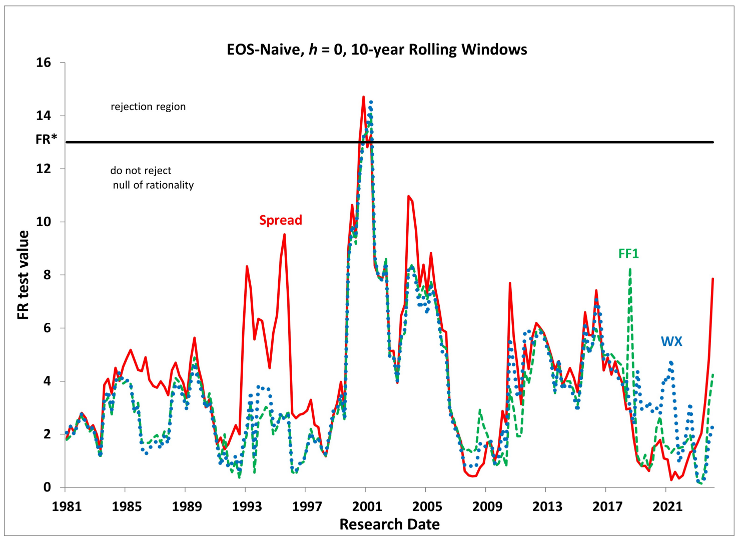
Rudebusch, Glenn D., and John C. Williams. 2009. "Forecasting Recessions: The Puzzle of the Enduring Power of the Yield Curve." *Journal of Business and Economic Statistics*, 27(4): 492–503.

Wu, Jing Cynthia, and Fan Dora Xia. 2016. "Measuring the Macroeconomic Impact of Monetary Policy at the Zero Lower Bound." *Journal of Money, Credit and Banking*, 48: 253–291.

VI. Appendix

This Appendix shows many of the graphs and tables discussed in the main body of the paper that were removed from the body of the paper to make it more succinct.

Figure 9. Forecast-Rationality Test Results, EOS-Naive Approach, h = 0, 10-Year Rolling Windows

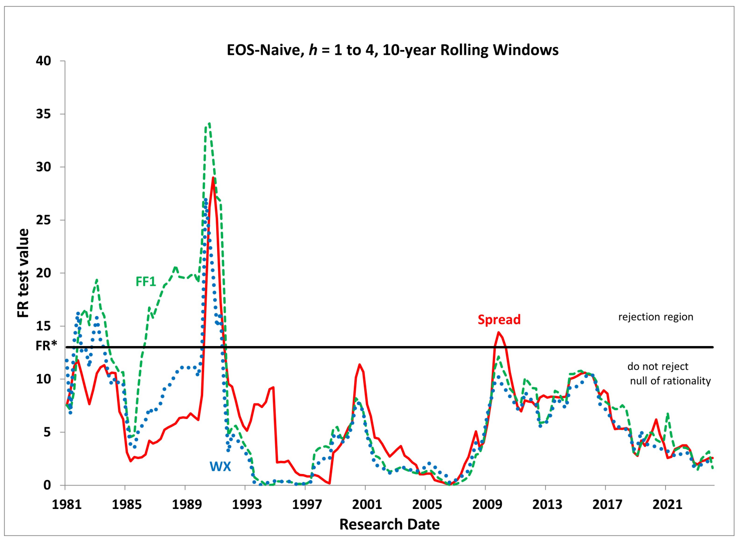


Note: The figure shows the forecast-rationality test results using the EOS-Naive approach with a 10-year rolling window and a horizon of zero. The forecast-rationality critical value is labeled FR*, and we reject forecast rationality if any value across the sample exceeds that threshold. Each line is labeled with the type of monetary policy used in Equation (7), where S is the yield spread, $FF1$ is the lagged change in real fed funds rate, and WX is the lagged change in the effective (Wu-Xia) real fed funds rate.

In Figure 15, for $h = 0$ in 5-year windows, we see very few rejections of forecast rationality, and many fewer than was the case for the EOS approaches.

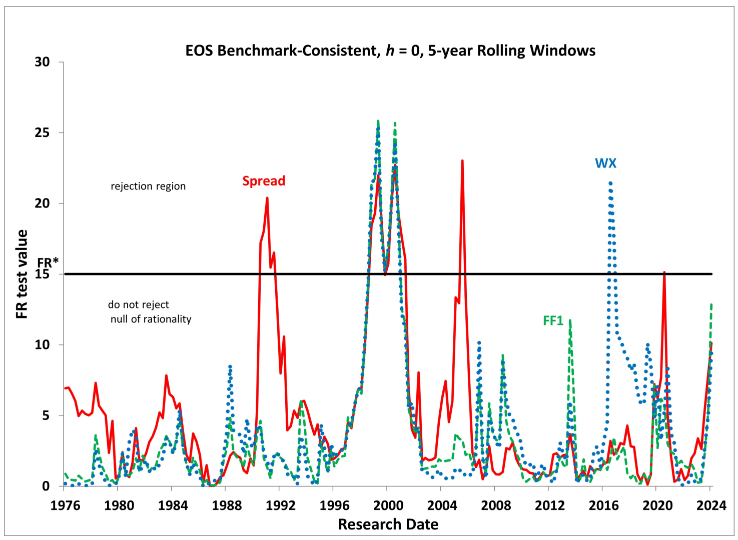
To compare with a longer-horizon forecast, Figure 7 shows the results of the forecast-rationality tests for the 1-to-4 quarter horizon.

Figure 10. Forecast-Rationality Test Results, EOS-Naive Approach, h = 1 to 4, 5-Year Rolling Windows



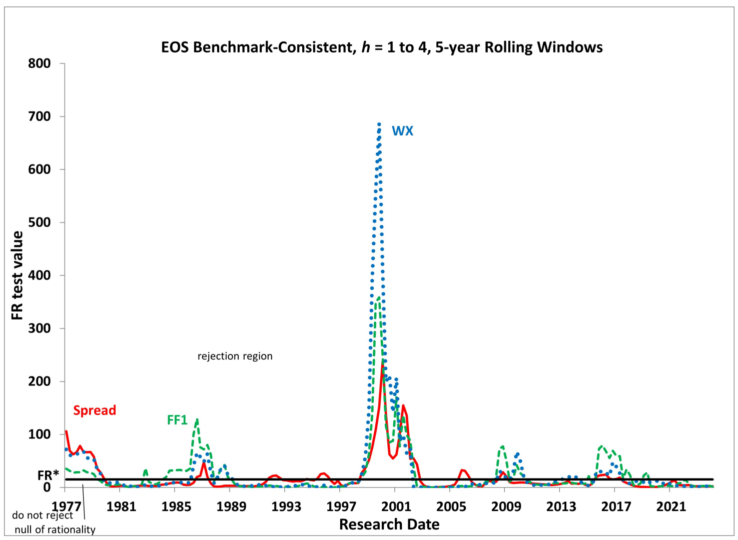
Note: The figure shows the forecast-rationality test results using the EOS-Naive approach with a 10-year rolling window and a horizon of 1 to 4 quarters. The forecast-rationality critical value is labeled FR*, and we reject forecast rationality if any value across the sample exceeds that threshold. Each line is labeled with the type of monetary policy used in Equation (7), where S is the yield spread, $FF1$ is the lagged change in real fed funds rate, and WX is the lagged change in the effective (Wu-Xia) real fed funds rate.

Figure 11. Forecast-Rationality Test Results, EOS Benchmark-Consistent Approach, h = 0, 5-Year Rolling Windows



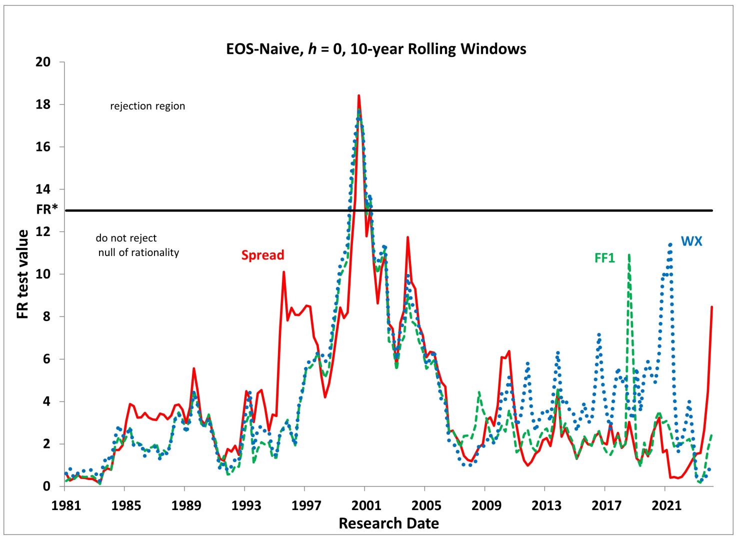
Note: The figure shows the forecast-rationality test results using the EOS Benchmark-Consistent approach with a 5-year rolling window and a horizon of zero. The forecast-rationality critical value is labeled FR*, and we reject forecast rationality if any value across the sample exceeds that threshold. Each line is labeled with the type of monetary policy used in Equation (7), where S is the yield spread, $FF1$ is the lagged change in real fed funds rate, and WX is the lagged change in the effective (Wu-Xia) real fed funds rate.

Figure 12. Forecast-Rationality Test Results, EOS Benchmark-Consistent Approach, h = 1 to 4, 5-Year Rolling Windows



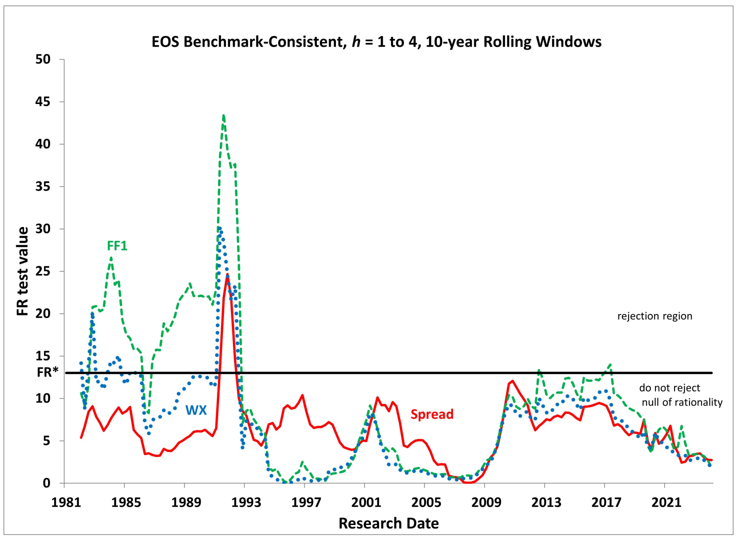
Note: The figure shows the forecast-rationality test results using the EOS Benchmark-Consistent approach with a 5-year rolling window and a horizon of 1 to 4 quarters. The forecast-rationality critical value is labeled FR*, and we reject forecast rationality if any value across the sample exceeds that threshold. Each line is labeled with the type of monetary policy used in Equation (7), where S is the yield spread, $FF1$ is the lagged change in real fed funds rate, and WX is the lagged change in the effective (Wu-Xia) real fed funds rate.

Figure 13. Forecast-Rationality Test Results, EOS Benchmark-Consistent Approach, h = 0, 10-Year Rolling Windows



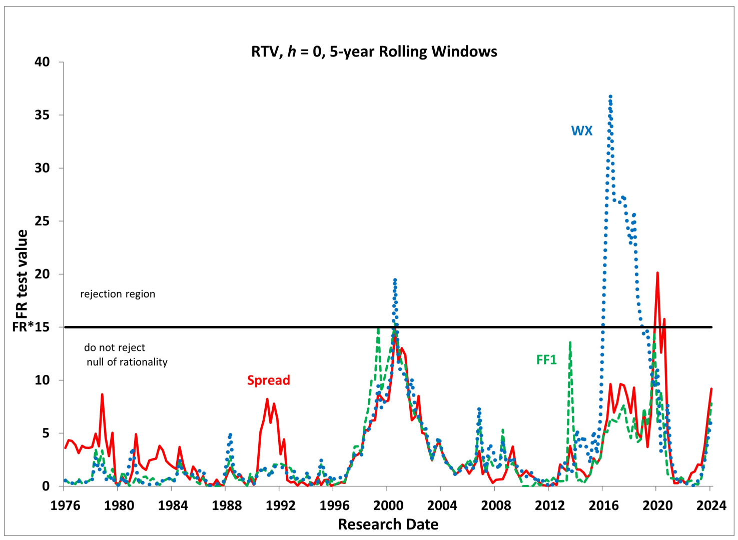
Note: The figure shows the forecast-rationality test results using the EOS Benchmark-Consistent approach with a 10-year rolling window and a horizon of zero. The forecast-rationality critical value is labeled FR*, and we reject forecast rationality if any value across the sample exceeds that threshold. Each line is labeled with the type of monetary policy used in Equation (7), where S is the yield spread, $FF1$ is the lagged change in real fed funds rate, and WX is the lagged change in the effective (Wu-Xia) real fed funds rate.

FIGURE 14. FORECAST-RATIONALITY TEST RESULTS, EOS BENCHMARK-CONSISTENT APPROACH, H = 1 TO 4, 5-YEAR ROLLING WINDOWS



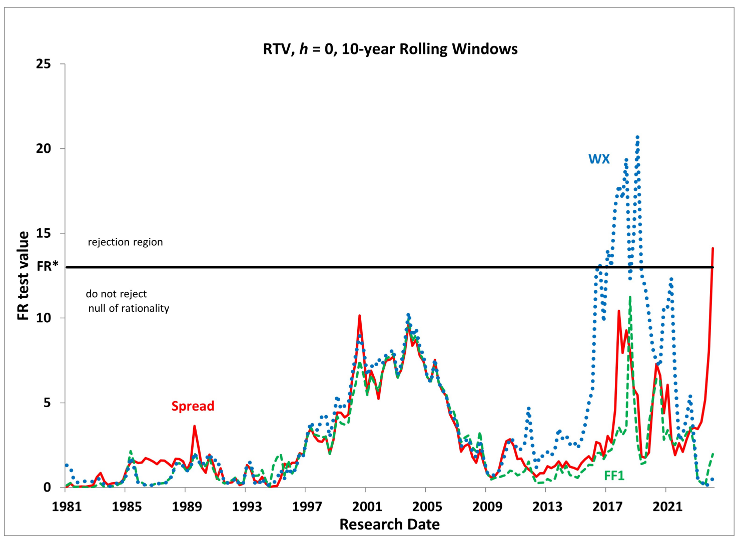
Note: The figure shows the forecast-rationality test results using the EOS Benchmark-Consistent approach with a 10-year rolling window and a horizon of 1 to 4 quarters. The forecast-rationality critical value is labeled FR*, and we reject forecast rationality if any value across the sample exceeds that threshold. Each line is labeled with the type of monetary policy used in Equation (7), where S is the yield spread, $FF1$ is the lagged change in real fed funds rate, and WX is the lagged change in the effective (Wu-Xia) real fed funds rate.

Figure 15. Forecast-Rationality Test Results, RTV Approach, h = 0, 5-Year Rolling Windows



Note: The figure shows the forecast-rationality test results using the RTV approach with a 5-year rolling window and a horizon of zero. The forecast-rationality critical value is labeled FR*, and we reject forecast rationality if any value across the sample exceeds that threshold. Each line is labeled with the type of monetary policy used in Equation (7), where S is the yield spread, $FF1$ is the lagged change in real fed funds rate, and WX is the lagged change in the effective (Wu-Xia) real fed funds rate.

Figure 16. Forecast-Rationality Test Results, RTV Approach, h = 0, 10-Year Rolling Windows



Note: The figure shows the forecast-rationality test results using the RTV approach with a 10-year rolling window and a horizon of zero. The forecast-rationality critical value is labeled FR*, and we reject forecast rationality if any value across the sample exceeds that threshold. Each line is labeled with the type of monetary policy used in Equation (7), where S is the yield spread, $FF1$ is the lagged change in real fed funds rate, and WX is the lagged change in the effective (Wu-Xia) real fed funds rate.