

Forecast Rationality Tests in the Presence of Instabilities, With Applications to Federal Reserve and Survey Forecasts

Barbara Rossi* and Tatevik Sekhposyan†

November 4, 2014

Abstract

This paper proposes a framework to implement regression-based tests of predictive ability in unstable environments, including, in particular, forecast unbiasedness and efficiency tests, commonly referred to as tests of forecast rationality. Our framework is general: it can be applied to model-based forecasts obtained either with recursive or rolling window estimation schemes, as well as to forecasts that are model-free. The proposed tests provide more evidence against forecast rationality than previously found in the Federal Reserve's Greenbook forecasts as well as survey-based private forecasts. It confirms, however, that the Federal Reserve has additional information about current and future states of the economy relative to market participants.

Keywords: Forecasting, forecast rationality, regression-based tests of forecasting ability, Greenbook forecasts, survey forecasts, real-time data

J.E.L. Codes: C22, C52, C53.¹

*ICREA-University of Pompeu Fabra, Barcelona GSE and CREI, c/Ramon Trias Fargas 25/27, Barcelona 08005, Spain; tel.: +34-93-542-1655; e-mail: barbara.rossi@upf.edu

†Texas A&M University, 3059 Allen Building, 4228 TAMU, College Station, TX 77843, USA; tel.: +1-979-862-8857; e-mail: tsekposyan@tamu.edu

¹**Acknowledgements:** We are grateful to M. McCracken, A. Patton, C. Vega and participants of seminars at

1 Introduction

Forecasting is a fundamental tool in economics, as well as statistics, business and other sciences. Judging whether forecasts are good is therefore of great importance, especially since forecasts are used everyday to guide policymakers' and practitioners' decisions. A large literature has provided important insights on how to test whether forecasts are optimal. For example, as the seminal works of Granger and Newbold (1986) and Diebold and Lopez (1996) show, under covariance stationarity and a mean square error loss, forecast errors are mean zero (conditionally and unconditionally) and the h -step-ahead forecast error has zero serial correlation after $(h - 1)$ lags. If a forecast is such that its forecast errors satisfy such properties, it is deemed optimal or rational.² If a forecast does not satisfy these properties, researchers conclude that the model underlying such forecast can be improved.

However, one of the fundamental assumptions tacitly underlying the existing literature is that of covariance stationarity. Only very recently researchers have become concerned about the consequences of relaxing stationarity assumptions in performing inference regarding predictive ability.³ For example, Giacomini and Rossi (2010) have developed methods to perform inference on forecast comparisons when the forecasting ability may be affected by instabilities. Besides forecast comparisons, another important issue that forecasters face in practice is to determine whether forecasts are rational or optimal, and that might also be affected by instabilities. In fact, several studies evaluate the robustness of forecast rationality in sub-samples (e.g. Croushore 1998, Patton and Timmermann, 2012, Croushore, 2012). However, while in some cases the choice of the sub-samples may be guided by economic considerations (e.g. sub-samples associated with structural breaks identified by the Great Moderation or Great Recession), in many cases the choice of sub-samples may be ad-hoc. Even when the choice is guided by economic considerations, it may be important to assess the robustness of the empirical results to other sub-samples, as there might be multiple breaks in the data, or the break date might be uncertain or completely unknown.

This paper proposes forecast rationality tests that are robust to the presence of instabilities. We consider a framework where forecasts are produced either with recursive or rolling estimation schemes, and the size of the estimation window is large relative to the sample size. We propose

the Bank of Canada, CREI, Trinity College Dublin, University of Barcelona, University of Montreal, Norges Bank, the 2011 IWH-CIREQ Macroeconometric Workshop, the 2011 JSM, the 2011 MEG, the 2013 Econometric Society Summer Meetings, the 2014 EABCN/Bank of England and ECB workshops and the 2014 NBER-NSF Time Series Conference for comments. B. Rossi gratefully acknowledges financial support from the European Research Agency's Marie Curie Grant 303434 and the ERC Grant 615608. A previous version was circulated under the title: "Forecast Optimality Tests in the Presence of Instabilities."

²Note that we use optimality and rationality interchangeably.

³See the discussion in Rossi (2013).

a “Fluctuation Rationality” test, which is based on testing forecast rationality in rolling windows over the out-of-sample forecast portion of the data. By using rolling windows we avoid averaging out instabilities, and our tests can have greater power to reject forecast rationality than traditional tests when rationality is present only in sub-samples of the data. Our “Fluctuation Rationality” test can be applied to study forecast unbiasedness, efficiency, rationality, encompassing, as well as serial uncorrelation, among other regression-based tests of forecasting ability.⁴

This paper is closely related to Giacomini and Rossi (2010) and West and McCracken (1998). Giacomini and Rossi (2010) propose a “Fluctuation test” to compare forecasting models in the presence of instabilities. While our “Fluctuation Rationality” test is inspired by their work, there are several differences between their framework and ours. Their framework compares models’ relative forecasting performance and is focused primarily on a rolling window estimation where the size of the window is fixed (i.e. finite). We are instead interested in measures of absolute predictive ability and tests for forecast rationality. Our framework focuses on an estimation window size that is a fixed fraction of the total sample size, which allows us to take into account parameter estimation error and considerably complicates the analysis. The latter framework is similar to that of West and McCracken (1998). The difference between our tests and West and McCracken’s (1998) is that the latter is based on measures of average forecasting ability in the out-of-sample portion of the data, and may lack power in certain directions when there are rationality breakdowns over time. Our tests can be used both when the forecasting model is known (and thus the researcher needs to correct for parameter estimation) and when it is not known (such as in the Greenbook and survey forecasts in our empirical application).

We demonstrate the usefulness of our procedures by evaluating the rationality of the Federal Reserve’s Greenbook forecast of inflation as well as the private sector’s forecasts provided by the Survey of Professional Forecasters (SPF) and Blue Chip Economic Indicators (BCEI). We revisit the empirical analysis in Romer and Romer (2000), Patton and Timmermann (2012), and Croushore (2012) in a framework that is more powerful in the presence of instabilities. We then reconsider Romer and Romer’s (2000) hypothesis that the Federal Reserve has an information advantage in forecasting inflation beyond what is known to the private forecasters, again using our framework is more powerful in the presence of time-variation. In both cases, our empirical results are very different than those in the literature: first, we find more empirical evidence against forecast rationality using our tests than using the traditional tests. In fact, the Fed was consistently under-estimating inflation in the 1970s, due to recurrent and unpredictable oil price shocks, and over-estimating

⁴It should be noted that, as shown in Rossi (2012), our framework can encompass the hypothesis of forecast optimality for more general definitions of optimality. As Patton and Timmermann (2010) suggest, under certain regularity conditions forecast optimality tests can reduce to those considered in this paper for arbitrary loss functions (symmetric or asymmetric) if one adheres to the definition of “generalized forecast error” and/or changes the probability measure.

inflation in the 1980s, during Volker’s disinflation period. Clearly, traditional forecast unbiasedness tests applied over the full sample do not reject forecast unbiasedness because under-predictions, on average, cancel out over-predictions. Similar issues affect tests of forecast rationality in general. Our test, instead, is capable of uncovering the lack of forecast rationality. Our findings are related to Croushore (2006), who uses forecast rationality tests and scatterplots in sub-samples and finds evidence of lack of forecast rationality in the Livingston and SPF surveys during the 1970s, but no evidence when considering a sample up to 2004. Our findings are also related to Sinclair et al. (2010), who similarly find systematic errors in Fed’s forecasts using other techniques. Furthermore, our test uncovers that the informational advantage of the Fed over private sectors’ forecasts, while confirmed in the data, has decreased over time.⁵

It is important to consider the trade-offs between our tests and the existing tests for forecast rationality. As our Monte Carlo simulations show, the test is capable of signalling lack of forecast rationality even if it is present in a sub-sample and therefore has higher power relative to the existing tests in these cases; however, because the test is implemented in rolling windows over the out-of-sample period, the number of observations used to implement the test is less than the total number of forecasts, thus its power may be lower than that of existing tests in small samples when there are no instabilities in the data.

The paper is structured as follows. Section 2 discusses the motivation that inspired the development of the techniques proposed in this paper, while Section 3 presents the econometric methodology. Sections 4 and 5 respectively present the results in the general framework and in special cases that are very relevant in practice. Section 6 studies the size and power of our “Fluctuation Rationality” test in small samples, while Section 7 discusses the empirical applications. The last section concludes.

2 Motivation

In a very influential paper, Romer and Romer (2000) analyzed the properties of forecasts of US inflation made by the Federal Reserve Board as well as by several private institutions. Their goal was to evaluate whether the forecasts were rational, that is, whether they were unbiased and efficient by using standard Mincer and Zarnowitz’s (1969) tests. Based on their empirical analysis, they find no evidence against the rationality of the Federal Reserve Board’s staff forecasts. Given that the forecasts are ultimately used by the central bank in guiding monetary policy, it is important that they are unbiased and efficient.

To motivate the methodologies developed in this paper, consider Figure 1. The figure reports two-quarter-ahead forecasts of US inflation made by several institutions. The dashed line reports

⁵This evidence is consistent with that in Gamber and Smith (2009).

the forecasts made by the Federal Reserve Board; the dotted line reports the forecasts made by the Blue Chip Economic Indicator and the dashed and dotted line reports forecasts made by the Survey of Professional Forecasters. These forecasts are discussed in detail in Section 7. Note that all forecasts *under-predict* inflation in the 1970s, a time period where the economy was constantly subject to unforecastable oil price shocks, which unexpectedly increased inflation relative to what forecasters predicted (see Croushore, 2006). Also, the forecasts constantly *over-predict* inflation in the 1980s and 1990s, a time period where the monetary authority was constantly fighting inflation. The over-prediction might be due to the fact that Greenspan’s policy turned out to be tougher than what forecasters anticipated. In addition, forecasts do poorly in the late 1990s: it is possible that they turned out to be upwardly biased because forecasters underestimated the growth rate of potential output or the disinflationary forces caused by the growth in Chinese imports. Thus, overall the forecasts appear not to be unbiased, nor rational: they under-predict the target in the first part of the sample, and over-predict it in the second part of the sample. We will investigate whether that is the case using formal statistical tests that we propose.

If the forecasts are biased, then, why did Romer and Romer (2000) conclude that the forecasts were unbiased? They applied their test over the full sample, which comprises periods of over-prediction as well as under-prediction. Thus, on average, the forecasts are unbiased. However, they may not be systematically so.

The goal of this paper is to develop techniques to help researchers detect situation where forecasts are not rational nor, in general, optimal, but the lack of rationality appears only in sub-samples of the data or presents itself in an unstable fashion. In fact, existing tests, that are based on stationarity assumptions under the alternative, should not be used in the presence of instabilities: they could lead to the wrong conclusion, as in the empirical example considered here.

3 The Econometric Framework

The main objective of this paper is to test whether the h -step ahead ($h > 0$), out-of-sample direct forecasts for the variable y_t , which we assume to be a scalar, are optimal. We assume that the forecasts are based on a model that is characterized by the $(k \times 1)$ parameter vector γ . The forecasts are obtained by dividing the sample of size $(T + h)$ observations into an in-sample portion of size R and an out-of-sample portion of size P , such that $R + P = T + h$. The sequence of P out-of-sample forecast errors depends on the realizations of the forecasted variable and on the in-sample parameter estimates, $\hat{\gamma}_{t,R}$. According to usual forecasting practices, we assume that these parameters are estimated in either one of two ways: (i) re-estimated at each $t = R, \dots, T$ over a window of R observations including data indexed $t - R + 1, \dots, t$ (rolling scheme); or (ii) re-estimated at each $t = R, \dots, T$ over a window of t observations including data indexed $1, \dots, t$

(recursive scheme).

Let the forecast error associated with the h-step-ahead forecast made at time t be denoted by $v_{t+h}(\hat{\gamma}_{t,R})$. For example, in a simple linear regression model with h-period lagged ($k \times 1$) vector of regressors x_t where $E_t y_{t+h} = x_t' \gamma$, the forecast at time t is $y_{t+h|t} = x_t' \hat{\gamma}_{t,R}$ and we have $v_{t+h}(\hat{\gamma}_{t,R}) = y_{t+h} - x_t' \hat{\gamma}_{t,R}$.

We focus on testing for forecast rationality in the framework developed by West and McCracken (1998). Consider the general regression:

$$v_{t+h}(\hat{\gamma}_{t,R}) = \hat{g}_t' \cdot \theta + \eta_{t+h}, \quad t = R, \dots, T, \quad (1)$$

where $\hat{g}_t \equiv g_t(\hat{\gamma}_{t,R})$ is an $(\ell \times 1)$ vector function of period t data (which can possibly be a function of the estimator $\hat{\gamma}_{t,R}$ – see examples below), θ is an $(\ell \times 1)$ parameter vector, and $v_{t+h}(\hat{\gamma}_{t,R})$ is the estimated forecast error. West and McCracken (1998) focus on testing the null hypothesis:

$$H_0 : \theta = \theta_0 \text{ vs. } H_A : \theta \neq \theta_0, \text{ where } \theta_0 = 0. \quad (2)$$

Let $\hat{\theta}_P$ denote the estimate of θ in regression (1). Consider the following Wald test:

$$\mathcal{W}_P = P \left(\hat{\theta}_P - \theta_0 \right)' \hat{V}_{\theta,P}^{-1} \left(\hat{\theta}_P - \theta_0 \right), \quad (3)$$

where $\hat{V}_{\theta,P}$ is a consistent estimate of the long run variance of $\sqrt{P} \hat{\theta}_P$. West and McCracken (1998) show that it is important to correct the estimate of the variance by parameter estimation error in order to estimate the long run variance consistently (cfr. their Theorem 4.1). We report the exact formula for West and McCracken's (1998) case, $\hat{V}_{\theta,P}$, at the end of Section 4.

The framework in equation (1) is quite general and includes the following leading cases:

(i) forecast unbiasedness tests, where $\hat{g}_t = 1$;

(ii) forecast efficiency, where $\hat{g}_t = y_{t+h|t}$.

(iii) forecast rationality (Mincer and Zarnowitz, 1969), where $\hat{g}_t = [1 \ y_{t+h|t}]$, and typically a researcher is interested in testing whether either or all regressors are insignificant; in general, \hat{g}_t may also contain any other variable known at time t which was not included in the forecasting model;

(iv) forecast encompassing tests, where \hat{g}_t is the forecast of the encompassed model;

(v) serial uncorrelation tests, where $\hat{g}_t = v_t(\hat{\gamma}_{t-h,R})$.

We refer to all these tests under the maintained assumption that $\theta_0 = 0$ as “tests for forecast rationality.” The zero restriction on the parameter under the null hypothesis ensures that the forecast errors are truly unpredictable given the information set available at the time when the forecast is made.

Our main interest is testing forecast optimality in the presence of instabilities. In fact, in the presence of instabilities, tests that focus on the average out-of-sample performance of a model may

be misleading, as they may average out instabilities. Instead, we consider the following rolling regression approach. Let $\widehat{\theta}_j$ be the parameter estimate in regression (1) computed at time j over rolling windows of size m .⁶ That is, $\widehat{\theta}_j$ is sequentially estimated in regression (1) for $j = R+m, \dots, T$ using the most recent m observations.⁷ Also, let the Wald test in the corresponding regressions be defined as:

$$\mathcal{W}_{j,m} = m\widehat{\theta}_j' \widehat{V}_\theta^{-1} \widehat{\theta}_j, \text{ for } j = R+m, \dots, T, \quad (4)$$

where, for example, in some special cases (such as forecast unbiasedness or efficiency), \widehat{V}_θ is a HAC estimator of the asymptotic variance of the parameter estimates in the rolling windows obtained as in West and McCracken (1998), that is, implemented by replacing P in their notation with m .⁸ Our proposed ‘‘Fluctuation Rationality’’ test is:

$$\max_{j \in \{R+m, \dots, T\}} \mathcal{W}_{j,m},$$

which we use to test the null hypothesis:⁹

$$H_0 : \theta_j = \theta_0 \text{ vs. } H_A : \theta_j \neq \theta_0, \forall j = R+m, \dots, T, \quad (5)$$

where $\theta_0 = 0$ and θ_j is the true parameter value.

4 General Results

Let the $(k \times 1)$ true parameter vector be denoted by γ^* , $v_{t+h} \equiv v_{t+h}(\gamma^*)$, $f_{t+h}(\widehat{\gamma}_{t,R}) \equiv \widehat{g}_t v_{t+h}(\widehat{\gamma}_{t,R})$ (an $(\ell \times 1)$ vector), $g_t \equiv g_t(\gamma^*)$, $f_{t+h} \equiv g_t v_{t+h} = f_{t+h}(\gamma^*)$, $f_{t+h,\gamma} \equiv \frac{\partial f_{t+h}(\gamma^*)}{\partial \gamma}$, $F \equiv E\left(\frac{\partial f_{t+h}(\gamma^*)}{\partial \gamma}\right)$ (an $(\ell \times k)$ matrix).

We make the following assumptions:

Assumption 1:

(i) *The estimate $\widehat{\gamma}_{t,R}$ satisfies $\widehat{\gamma}_{t,R} - \gamma^* = B_t H_t$ where B_t is $(k \times q)$ matrix and H_t is $(q \times 1)$ with (a) $B_t \xrightarrow{as} B$ with rank k ; (b) $H_t = t^{-1} \sum_{r=1}^t h_r(\gamma^*)$ for the recursive and $H_t = R^{-1} \sum_{r=t-R+1}^t h_r(\gamma^*)$ for the rolling estimation methods, for a $(q \times 1)$ orthogonality condition $h_r(\gamma^*)$; (c) $E(h_r(\gamma^*)) = 0$ and $h(\cdot)$ is measurable and continuously differentiable.*

⁶We removed the dependence of $\widehat{\theta}_j$ from m (m is the same for all $\widehat{\theta}_j$ s).

⁷E.g., $\widehat{\theta}_{R+m}$ is estimated in equation (1) using $v_{R+h}(\widehat{\gamma}_{R,R}), \dots, v_{R+m+h}(\widehat{\gamma}_{R+m,R})$; $\widehat{\theta}_{R+m+1}$ is estimated in equation (1) using $v_{R+1+h}(\widehat{\gamma}_{R+1,R}), \dots, v_{R+m+h+1}(\widehat{\gamma}_{R+m+1,R})$; ... and $\widehat{\theta}_T$ is estimated in equation (1) using $v_{T-m+1+h}(\widehat{\gamma}_{T-m+1,R}), \dots, v_{T+h}(\widehat{\gamma}_{T,R})$.

⁸Alternatively, one could design a Wald test similar to the One-time Reversal test in Giacomini and Rossi (2010), which may have more power against one-time breaks in θ ; its derivation follows directly from our results.

⁹In the construction of the test we associate the end of period date of the fixed window m with the parameter estimate $\widehat{\theta}_j$. In fact, that need not necessarily be the case. If one prefers, one can choose to associate the mid-period date of the fixed window m with the parameter estimate, for example.

(ii) In some neighborhood N around γ^* , and with probability 1, $v_t(\gamma)$ and $g_t(\gamma)$ are measurable and twice continuously differentiable, and $E(g_t g_t') \equiv G$ is an $(\ell \times \ell)$ of rank ℓ .

Assumption 2:

$$(i) \lim_{T \rightarrow \infty} \sup_j m^{-1/2} \sum_{t=j-m+1}^j (f_{t+h, \gamma} - F) B H_t = o_p(1);$$

$$(ii) \lim_{T \rightarrow \infty} \sup_j m^{-1/2} F \sum_{t=j-m+1}^j (B_t - B) H_t = o_p(1);$$

$$(iii) \lim_{T \rightarrow \infty} \sup_j m^{-1/2} \sum_{t=j-m+1}^j (f_{t+h, \gamma} - F) (B_t - B) H_t = o_p(1);$$

$$(iv) \lim_{T \rightarrow \infty} \sup_j \left[\left(m^{-1} \sum_{t=j-m+1}^j g_t g_t' \right)^{-1} - G^{-1} \right] = o_p(1).$$

(v) There is a finite constant D such that, for all t , $\sup_{\gamma \in N} |\partial^2 v_t(\gamma) / \partial \gamma \partial \gamma'| < m_t$ for a measurable m_t such that $E(m_t^4) < D$ and the same holds when v_t is replaced by any element of g_t .

Assumption 3. $m \rightarrow \infty$, $R \rightarrow \infty$ as $T \rightarrow \infty$, and $\lim_{T \rightarrow \infty} m/T = \mu \in (0, 1)$, $\lim_{T \rightarrow \infty} R/T = \rho \in (0, 1)$; let $j = \lceil \tau T \rceil$, $t = \lceil sT \rceil$, where $\lceil \cdot \rceil$ denotes the integer function and $\tau \in (0, 1)$, $s \in (0, 1)$, $h < \infty$.

Assumption 1(i) allows the in-sample parameter estimates to be obtained by general estimation procedures such as Ordinary Least Squares (OLS), maximum likelihood, and GMM, for example. Assumption 1(ii) imposes differentiability and full rank conditions. Assumption 2 guarantees that the remainder of the mean value expansion is asymptotically irrelevant. Assumption 3 requires R and m to be large relative to the sample size (in particular, relative to the finite horizon, h), be some fixed proportion of the total sample size and ensures the consistency of the out-of-sample test statistics.¹⁰ The assumption accommodates both rolling and recursive estimation schemes.

The Not-for-Publication Appendix shows that, under Assumptions 1, 2 and 3, we have:

$$m^{1/2} \widehat{\theta}_j = G^{-1} \left(\frac{T}{m} \right)^{1/2} [I_\ell, FB] \left\{ \frac{1}{\sqrt{T}} \sum_{t=R}^j \begin{pmatrix} f_{t+h} \\ H_t \end{pmatrix} - \frac{1}{\sqrt{T}} \sum_{t=R}^{j-m} \begin{pmatrix} f_{t+h} \\ H_t \end{pmatrix} \right\} + A_j, \quad (6)$$

where A_j is a remainder term such that $\lim_{T \rightarrow \infty} \sup_j A_j = o_p(1)$ and I_ℓ is an $(\ell \times \ell)$ identity matrix.

Let $\sum_{t=R}^j H_t = \sum_{t=1}^j a_{R,t,j} h_t$, where direct calculations show that:

(i) for the recursive estimation scheme $\left(H_t = t^{-1} \sum_{r=1}^t h_r \right)$:

$$a_{R,t,j} = (R^{-1} + \dots + j^{-1}) \cdot 1(t \leq R) + (t^{-1} + \dots + j^{-1}) \cdot 1(R < t \leq j); \quad (7)$$

¹⁰This assumption is not restrictive: $\widehat{\theta}_j$ can be estimated in recursive windows. The analytical calculations for this case are already provided in this paper. Critical values for select cases are reported in footnote 13.

(ii) for the rolling estimation scheme $\left(H_t = R^{-1} \sum_{r=t-R+1}^t h_r\right)$:
if $j - R \geq R$:

$$a_{R,t,j} = \left(\frac{1}{R}\right) \{t \cdot 1(t \leq R) + R \cdot 1(R < t \leq j - R) + (j - t) \cdot 1(j - R < t \leq j)\}; \quad (8)$$

whereas if $j - R < R$:

$$a_{R,t,j} = \left(\frac{1}{R}\right) \{t \cdot 1(t < j - R) + (j - R) \cdot 1(j - R \leq t \leq R) + (j - t) \cdot 1(R < t \leq j)\}. \quad (9)$$

In addition, for all estimation schemes let $b_{R,j,t} = 1(t \geq R)$.

Let I_q be a $(q \times q)$ identity matrix. It follows from equations (7), (8) and (9) and $b_{R,j,t} = 1(t \geq R)$ that

$$\frac{1}{\sqrt{T}} \sum_{t=R}^j \begin{pmatrix} f_{t+h} \\ H_t \end{pmatrix} = \frac{1}{\sqrt{T}} \sum_{t=1}^j \begin{pmatrix} b_{R,t,j} \cdot I_\ell & 0 \\ 0 & a_{R,t,j} \cdot I_q \end{pmatrix} \begin{pmatrix} f_{t+h} \\ h_t \end{pmatrix}.$$

We further approximate $b_{R,t,j}$ and $a_{R,t,j}$ as follows. Let $b_{R,t,j} = 1(t/T \geq R/T) \simeq 1(s \geq \rho) \equiv \sigma_f(s)$ and

(i) for recursive: given equation (7), we follow West (1996) to show that

$$a_{R,t,j} \simeq \left(\int_R^j \frac{1}{k} dk\right) \cdot 1(t \leq R) + \left(\int_t^j \frac{1}{k} dk\right) \cdot 1(R < t \leq j)$$

Consider $r = k/T$. Accordingly,

$$\begin{aligned} a_{R,t,j} &\simeq \left(\int_\rho^\tau \frac{1}{r} dr\right) \cdot 1(s \leq \rho) + \left(\int_s^\tau \frac{1}{r} dr\right) \cdot 1(\rho < s \leq \tau) \\ &= [\ln(\tau) - \ln(\rho)] \cdot 1(s \leq \rho) + [\ln(\tau) - \ln(s)] \cdot 1(\rho < s \leq \tau) \equiv \sigma_h(s, \tau); \end{aligned} \quad (10)$$

(ii) for rolling: when $j - R \geq R$, we can re-write equation (8) as

$$\begin{aligned} a_{R,t,j} &= \frac{t}{R} \cdot 1\left(\frac{t}{T} \leq \frac{R}{T}\right) + \frac{R}{R} \cdot 1\left(\frac{R}{T} < \frac{t}{T} \leq \frac{j-R}{T}\right) \\ &\quad + \frac{j-t}{R} \cdot 1\left(\frac{j-R}{T} < \frac{t}{T} \leq \frac{j}{T}\right); \end{aligned} \quad (11)$$

thus, when $\tau - \rho \geq \rho$,

$$\sigma_h(s, \tau) = \frac{s}{\rho} \cdot 1(s \leq \rho) + 1 \cdot 1(\rho < s \leq \tau - \rho) + \frac{\tau - s}{\rho} \cdot 1(\tau - \rho < s \leq \tau). \quad (12)$$

A similar argument shows that when $\tau - \rho < \rho$, equation (9) can be approximated as

$$\begin{aligned} \sigma_h(s, \tau) &= \frac{s}{\rho} \cdot 1(s \leq \tau - \rho) + \frac{\tau - \rho}{\rho} \cdot 1(\tau - \rho < s \leq \rho) \\ &+ \frac{\tau - s}{\rho} \cdot 1(\rho < s \leq \tau). \end{aligned} \quad (13)$$

The following table summarizes the approximations we use for the weights $a_{R,t,j}$, $b_{R,t,j}$:

Approximation for the weights $a_{R,t,j}$, $b_{R,t,j}$		
Weights	Estimation Scheme	Approximation
$b_{R,t,j}$	All	$\sigma_f(s) \equiv 1(s \geq \rho)$
$a_{R,t,j}$	Recursive	$\sigma_h(s, \tau) \equiv [\ln(\tau) - \ln(\rho)] \cdot 1(s \leq \rho)$ $+ [\ln(\tau) - \ln(s)] \cdot 1(\rho < s \leq \tau)$
	Rolling scheme:	
	(a) $\tau - \rho \geq \rho$	$\sigma_h(s, \tau) \equiv \frac{s}{\rho} \cdot 1(s \leq \rho) + 1 \cdot 1(\rho < s \leq \tau - \rho)$ $+ \frac{\tau - s}{\rho} \cdot 1(\tau - \rho < s \leq \tau)$
	(b) $\tau - \rho < \rho$	$\sigma_h(s, \tau) \equiv \frac{s}{\rho} \cdot 1(s \leq \tau - \rho) + \frac{\tau - \rho}{\rho} \cdot 1(\tau - \rho < s \leq \rho)$ $+ \frac{\tau - s}{\rho} \cdot 1(\rho < s \leq \tau)$

Define $\xi_j = \sum_{t=1}^j \begin{pmatrix} f_{t+h} \\ h_t \end{pmatrix}$ and the stochastic integral of interest as in Hansen (1992, p. 491):

$$\int_0^\tau \begin{pmatrix} \sigma_f(s) \cdot I_\ell & 0 \\ 0 & \sigma_h(s, \tau) \cdot I_q \end{pmatrix} d\xi_T = \frac{1}{\sqrt{T}} \sum_{t=1}^j \begin{pmatrix} b_{R,t,j} \cdot I_\ell & 0 \\ 0 & a_{R,t,j} \cdot I_q \end{pmatrix} \begin{pmatrix} f_{t+h} \\ h_t \end{pmatrix}.$$

The following assumption is based on Hansen (1992) and allows us to derive the asymptotic distribution of our parameter of interest, $\hat{\theta}_j$.

Assumption 4. For some $p > \beta > 2$, $(f'_{t+h}, h'_t)'$ is zero mean, strong mixing with mixing coefficients α_m of size $-p\beta/(p - \beta)$ and $\sup_{t \geq 1} \|(f'_{t+h}, h'_t)'\|_p = C < \infty$. In addition, $\lim_{T \rightarrow \infty} T^{-1} E(\xi_T \xi_T') = S \equiv \begin{pmatrix} S_{ff} & S_{fh} \\ S'_{fh} & S_{hh} \end{pmatrix}$ is an $(l + q) \times (l + q)$ positive definite and finite matrix as $T \rightarrow \infty$.

Proposition 1 (Preliminary Asymptotic Result) Under Assumptions 1-4 and $T^{-1/2} \xi_T \rightarrow \xi$ in $\mathcal{D}_{\mathbb{R}^{(l+q)}}[0, 1]$:

$$\frac{1}{\sqrt{T}} \sum_{t=1}^j \begin{pmatrix} b_{R,t,j} \cdot I_\ell & 0 \\ 0 & a_{R,t,j} \cdot I_q \end{pmatrix} \begin{pmatrix} f_{t+h} \\ h_t \end{pmatrix} \Rightarrow \int_0^\tau \Omega(s, \tau)^{1/2} d\xi(s),$$

where $\xi(s) = S^{1/2}\mathcal{B}_{\ell+q}(s)$, $\mathcal{B}_{\ell+q}(s)$ is an $(\ell+q) \times 1$ vector of independent standard Brownian motions, \mathcal{D} denotes the space of cadlag functions, “ \Rightarrow ” denotes weak convergence with respect to the Skorohod metric, and $\Omega(s, \tau)^{1/2} \equiv \begin{pmatrix} \sigma_f(s) \cdot I_\ell & 0 \\ 0 & \sigma_h(s, \tau) \cdot I_q \end{pmatrix}$.

The proof of the proposition builds on Cavaliere (2005) and is provided in Rossi and Sekhposyan (2014). We use the result in Proposition 1 to derive the asymptotic distribution of the parameter estimate, $\widehat{\theta}_j$, in the next Proposition.

Proposition 2 (Asymptotic Distribution of $\widehat{\theta}_j$) Under Assumptions 1-4 and $T^{-1/2}\xi_T \rightarrow \xi$ in $\mathcal{D}_{\mathbb{R}^{(\ell+q)}}[0, 1]$:

$$m^{1/2} \widehat{\theta}_j \Rightarrow \int_0^\tau \widetilde{\omega}(s, \tau) d\mathcal{B}_{\ell+q}(s) - \int_0^{\tau-\mu} \widetilde{\omega}(s, \tau - \mu) d\mathcal{B}_{\ell+q}(s) = \mathcal{B}_{\widetilde{\omega}}(\tau) - \mathcal{B}_{\widetilde{\omega}}(\tau - \mu) \quad (14)$$

$$= \int_0^\tau \omega(s, \tau) d\mathcal{B}_{\ell+q}(s) \equiv \mathcal{B}_\omega(\tau) = \mathcal{B}_\ell \left(\int_0^\tau \omega(s, \tau) \omega(s, \tau)' ds \right), \quad (15)$$

where

$$\widetilde{\omega}(s, \tau) = \mu^{-\frac{1}{2}} G^{-1} [I_\ell, FB] \Omega(s, \tau)^{1/2} S^{1/2}, \quad (16)$$

$$\begin{aligned} \omega(s, \tau) = \mu^{-\frac{1}{2}} G^{-1} [I_\ell, FB] \{ & \left[\Omega(s, \tau)^{1/2} - \Omega(s, \tau - \mu)^{1/2} \right] \cdot \mathbf{1}(s \leq \tau - \mu) \\ & + \Omega(s, \tau)^{1/2} \cdot \mathbf{1}(\tau - \mu < s \leq \tau) \} S^{1/2}, \end{aligned} \quad (17)$$

$\mathcal{B}_{\ell+q}(s)$ is an $(\ell+q) \times 1$ vector of independent standard Brownian motions and $\stackrel{d}{=}$ denotes equality in distribution.

Note that both $\mathcal{B}_\omega(\tau)$ as well as $\mathcal{B}_{\widetilde{\omega}}(\tau)$ are $(\ell \times 1)$ Gaussian processes with time-varying variances. $\mathcal{B}_\omega(\tau)$ is Gaussian with quadratic variation $\int_0^\tau \omega(s, \tau) \omega(s, \tau)' ds$.¹¹ Similarly, $\mathcal{B}_{\widetilde{\omega}}(\tau) \equiv \mathcal{B}_\ell \left(\int_0^\tau \widetilde{\omega}(s, \tau) \widetilde{\omega}(s, \tau)' ds \right)$ is Gaussian with quadratic variation $\int_0^\tau \widetilde{\omega}(s, \tau) \widetilde{\omega}(s, \tau)' ds$. The following Proposition calculates the quadratic variation of $\mathcal{B}_\omega(\tau)$ and $\mathcal{B}_{\widetilde{\omega}}(\tau)$ for the rolling and the recursive estimation schemes.

Proposition 3 (Calculation of $\int_0^\tau \omega(s, \tau) \omega(s, \tau)' ds$)

$$\begin{aligned} \int_0^\tau \omega(s, \tau) \omega(s, \tau)' ds = \mu^{-1} G^{-1} \{ & \left(\int_{\tau-\mu}^\tau \sigma_f^2(s) ds \right) S_{ff} + \left(\int_{\tau-\mu}^\tau \sigma_f(s) \sigma_h(s, \tau) ds \right) \\ & (FBS_{fh} + S_{fh}B'F') + \int_0^{\tau-\mu} \left[(\sigma_h(s, \tau) - \sigma_h(s, \tau - \mu))^2 \cdot \mathbf{1}(s \leq \tau - \mu) \right. \\ & \left. + \sigma_h^2(s, \tau) \cdot \mathbf{1}(\tau - \mu \leq s \leq \tau) \right] ds FBS_{hh}B'F' \} G^{-1}, \end{aligned}$$

¹¹We eliminated the vector dimension in the notation for $\mathcal{B}_\omega(\cdot)$, $\mathcal{B}_{\widetilde{\omega}}(\cdot)$ as they are always dimension $(\ell \times 1)$.

where

(i) $\int_{\tau-\mu}^{\tau} \sigma_f^2(s) ds = \mu$ for both rolling and recursive cases;

(ii) recursive: let $\tilde{\pi} \equiv \mu / (\tau - \mu)$;

$\int_{\tau-\mu}^{\tau} \sigma_f(s) \sigma_h(s, \tau) ds = \mu [1 - \tilde{\pi}^{-1} \ln(1 + \tilde{\pi})]$ and

$\int_0^{\tau} [(\sigma_h(s, \tau) - \sigma_h(s, \tau - \mu))^2 \cdot 1(s \leq \tau - \mu) + \sigma_h^2(s, \tau) \cdot 1(\tau - \mu \leq s < \tau)] ds = 2\mu [1 - \tilde{\pi}^{-1} \ln(1 + \tilde{\pi})]$;

(iii) rolling: let $\pi^\dagger \equiv \frac{\mu}{\rho}$;

(a) if $\mu \geq \rho$, then

$\int_{\tau-\mu}^{\tau} \sigma_f(s) \sigma_h(s, \tau) ds = \mu (1 - \frac{1}{2\pi^\dagger})$ and

$\int_0^{\tau} [(\sigma_h(s, \tau) - \sigma_h(s, \tau - \mu))^2 \cdot 1(s \leq \tau - \mu) + \sigma_h^2(s, \tau) \cdot 1(\tau - \mu \leq s < \tau)] ds = \mu (1 - \frac{1}{3\pi^\dagger})$;

(b) if $\mu < \rho$, then

$\int_{\tau-\mu}^{\tau} \sigma_f(s) \sigma_h(s, \tau) ds = \frac{1}{2}\mu\pi^\dagger$ and

$\int_0^{\tau} [(\sigma_h(s, \tau) - \sigma_h(s, \tau - \mu))^2 \cdot 1(s \leq \tau - \mu) + \sigma_h^2(s, \tau) \cdot 1(\tau - \mu \leq s < \tau)] ds = \mu\pi^\dagger (1 - \frac{1}{3}\pi^\dagger)$.

Proposition 4 (Calculation of $\int_0^{\tau} \tilde{\omega}(s, \tau) \tilde{\omega}(s, \tau)' ds$)

$$\begin{aligned} \int_0^{\tau} \tilde{\omega}(s, \tau) \tilde{\omega}(s, \tau)' ds &= \mu^{-1} G^{-1} \left\{ \left(\int_0^{\tau} \sigma_f^2(s) ds \right) S_{ff} + \left(\int_0^{\tau} \sigma_f(s) \sigma_h(s, \tau) ds \right) \times \right. \\ &\quad \left. \times (FBS_{fh} + S_{fh}B'F') + \left(\int_0^{\tau} \sigma_h^2(s, \tau) ds \right) FBS_{hh}B'F' \right\} G^{-1}, \end{aligned}$$

where

(i) $\int_0^{\tau} \sigma_f^2(s) ds = (\tau - \rho)$ for both rolling and recursive cases;

(ii) recursive:

$\int_{\tau-\mu}^{\tau} \sigma_h(s, \tau) \sigma_f(s) ds = (\tau - \rho) \left(1 - \frac{\rho}{\tau - \rho} \ln \left(\frac{\tau}{\rho} \right) \right)$ and

$\int_0^{\tau} \sigma_h^2(s, \tau) ds = 2(\tau - \rho) \left(1 - \frac{\rho}{\tau - \rho} \ln \left(\frac{\tau}{\rho} \right) \right)$;

(iii) rolling:

(a) if $\tau - \rho \geq \rho$, then

$\int_0^{\tau} \sigma_f(s) \sigma_h(s, \tau) ds = (\tau - \frac{3}{2}\rho)$ and $\int_0^{\tau} \sigma_h^2(s) ds = (\tau - \frac{4}{3}\rho)$;

(b) if $\tau - \rho < \rho$, then

$\int_0^{\tau} \sigma_f(s) \sigma_h(s, \tau) ds = \frac{1}{2\rho} (\tau - \rho)^2$ and $\int_0^{\tau} \sigma_h^2(s) ds = \frac{1}{3\rho^2} (\tau - \rho)^2 (4\rho - \tau)$.

The next Proposition discusses the asymptotic distribution of the $\mathcal{W}_{j,m}$ test statistic presented in equation (4).

Theorem 5 (Main Proposition) Under Assumption 1-4,

$$\begin{aligned} \mathcal{W}_{j,m} &= m\hat{\theta}'_j V_{\theta,\tau}^{-1} \hat{\theta}_j \\ &\Rightarrow \left[\mathcal{B}_\ell \left(\int_0^{\tau} \omega(s, \tau) \omega(s, \tau)' ds \right) \right]' V_{\theta,\tau}^{-1} \left[\mathcal{B}_\ell \left(\int_0^{\tau} \omega(s, \tau) \omega(s, \tau)' ds \right) \right], \end{aligned} \quad (18)$$

where

$$V_{\theta, \tau} = Avar \left(m^{1/2} \widehat{\theta}_j \right) \quad (19)$$

$$\begin{aligned} &= G^{-1} [I_\ell, FB] Avar \left(\frac{1}{\sqrt{m}} \sum_{t=j-m+1}^j \begin{pmatrix} f_{t+h} \\ H_t \end{pmatrix} \right) [I_\ell, FB]' G^{-1} \\ &= \int_0^\tau \omega(s, \tau) \omega(s, \tau)' ds, \end{aligned} \quad (20)$$

$j = \lceil \tau T \rceil$, $m = \lfloor \mu T \rfloor$ and $B_\ell(\cdot)$ is a standard ℓ -dimensional Brownian motion. Let θ_j be the true parameter value. We reject the null hypothesis:

$$H_0 : \theta_j = \theta_0, \theta_0 = 0 \text{ for all } j = R + m, \dots, T \quad (21)$$

if $\max_{j \in \{R+m, \dots, T\}} W_{j,m} > \kappa_{\alpha, \ell}$, where $\kappa_{\alpha, \ell}$ are the critical values at the $100\alpha\%$ significance level that can be simulated for given values of μ, ℓ, G, F, B and S .

$V_{\theta, \tau}$ can be estimated using Proposition 3 and replacing the population values of S_{ff}, S_{fh}, S_{hh} with a consistent estimate. For example, one could use Newey and West's (1987) covariance estimator of long run variance of $\left\{ (f'_{t+h}, h'_t)' \right\}_{t=R}^T$.

The asymptotic distribution of the test statistic $W_{j,m}$ is non-standard and depends on nuisance parameters. We obtain its critical values, $\kappa_{\alpha, \ell}$, via Monte Carlo simulations by the following steps:

1. Simulate $B_\ell \left(\int_0^\tau \widetilde{\omega}(s, \tau) \widetilde{\omega}(s, \tau)' ds \right)$ by using the approximation $\sqrt{T} \sum_{t=1}^j \widetilde{\omega} \left(\frac{t}{T}, \frac{j}{T} \right) \vartheta_\ell$, where ϑ_ℓ is an $(\ell \times 1)$ vector of independent standard Normal random variables;
2. Simulate $B_\ell \left(\int_0^{\tau-\mu} \widetilde{\omega}(s, \tau) \widetilde{\omega}(s, \tau)' ds \right)$ similarly by $\sqrt{T} \sum_{t=1}^{j-m} \widetilde{\omega} \left(\frac{t}{T}, \frac{j-m}{T} \right) \vartheta_\ell$;¹²
3. Then, we obtain

$$B_\ell \left(\int_0^\tau \omega(s, \tau) \omega(s, \tau)' ds \right) = B_\ell \left(\int_0^\tau \widetilde{\omega}(s, \tau) \widetilde{\omega}(s, \tau)' ds \right) - B_\ell \left(\int_0^{\tau-\mu} \widetilde{\omega}(s, \tau) \widetilde{\omega}(s, \tau)' ds \right);$$

4. Finally, conditional on the estimated value of $V_{\theta, \tau}$ obtained by equation (19), we use Proposition 4 to simulate the $W_{j,m}$ test statistic by using equation (18);
5. The critical values at significance level α can be obtained by the $(1 - \alpha)$ th quantile of the simulated distribution of $W_{j,m}$. Note that when S is full rank, $S^{1/2}$ can be calculated as the Cholesky factor of S ; when S is rank deficient, one can use a singular value decomposition in Rao (Section 8a.4) to approximate $\widetilde{\omega}(t/T, j/T)$.

¹²It is important that the random variable ϑ_ℓ used to simulate the two Brownian motions is the same.

The case of a linear regression model in equation (3) is the same as that considered in West and McCracken (1998). Note the similarity between the results in Proposition (3) and West and McCracken's (1998) variance. The latter define the variance in exactly the same way, except that in their case $m = P$, $\pi = \lim_{T \rightarrow \infty} (P/R)$ and $\tau = 1$. Consequently, by letting $\mu = \pi(1 + \pi)^{-1}$, $\rho = (1 + \pi)^{-1}$ and $(1 - \rho) = \pi(1 + \pi)^{-1}$ in Proposition 3 we recover their results:¹³

(i) recursive scheme:

$$V_{\theta,P} = G^{-1} \{ S_{ff} + (1 - \pi^{-1} \ln(1 + \pi)) [FBS'_{fh} + B'F'S_{fh}] + 2(1 - \pi^{-1} \ln((1 + \pi))) FBS_{hh}B'F' \} G^{-1};$$

(ii) rolling scheme:

$$V_{\theta,P} = \begin{cases} G^{-1} \left\{ S_{ff} + \left(1 - \frac{1}{2\pi}\right) [FBS'_{fh} + B'F'S_{fh}] - \left(1 - \frac{1}{3\pi}\right) FBS_{hh}F'B' \right\} G^{-1}; & \pi \geq 1 \\ G^{-1} \left\{ S_{ff} + \frac{\pi}{2} [FBS'_{fh} + B'F'S_{fh}] + \left(\pi - \frac{\pi^2}{3}\right) FBS_{hh}F'B' \right\} G^{-1}; & \pi < 1. \end{cases}$$

The difference between West and McCracken (1998) and our approach is that we aim at testing forecast rationality at each point in the out-of-sample period, based on rolling window averages, while they focus on optimality on average over the whole out-of-sample portion of the data. In the case of West and McCracken (1998), the tests take into account parameter estimation error by simple adjustment factors in the variance, which result in tests with asymptotic distributions that are nuisance parameter free. In our case, instead, we need to adjust the asymptotic distribution to take into account the parameter estimation error, which induces a time-varying variance; under very general conditions, this implies that the critical values depend on the data generating process and need to be simulated specifically for the individual application at hand.

5 Forecast Unbiasedness, Efficiency Tests, and Survey Forecasts

The general results presented in the previous section simplify considerably in two cases important for practitioners. The first important special case is when parameter estimation error is irrelevant ($F = 0$). This may be often of interest in practice when the model that generated the forecasts is not available and, thus, the correction for parameter estimation error is not applicable. Relevant examples are survey forecasts and judgemental forecasts produced by central banks; for instance, the Greenbook forecasts by the Federal Reserve Board or private sector forecasts such as those produced by the Blue Chip Economic Indicators. In addition, if a researcher were to consider a null hypothesis on the forecast errors evaluated at the estimated models' parameter, instead of at the pseudo-true parameter values, the asymptotic distribution of the test statistic would similarly be nuisance parameter free. For a discussion and implementation, see Rossi (2012, 2013).

¹³Recall that West and McCracken's (1998) test statistic is obtained by rescaling by $P^{1/2}$ rather than $T^{1/2}$ as shown in equation (3).

Proposition 6 (Special Case I: Irrelevant Parameter Estimation Error) *Under the condition $F = 0$, parameter estimation error becomes irrelevant and $\int_0^\tau \omega(s, \tau) \omega(s, \tau)' ds$ becomes $G^{-1} S_{ff} G^{-1}$ and $\int_0^\tau \tilde{\omega}(s, \tau) \tilde{\omega}(s, \tau)' ds = \frac{(\tau - \rho)}{\mu} G^{-1} S_{ff} G^{-1}$ for all estimation schemes.*

The second special case involves testing for forecast unbiasedness and efficiency using t-tests under general conditions, as well as several other tests under more specific assumptions. As discussed in West and McCracken (1998) Corollary 5, in such cases a special condition holds, which considerably simplifies the asymptotic distributions of our test statistic. As in West and McCracken (1998), the results in the special case below also hold for encompassing and serial correlation tests when the errors are conditionally homoskedastic.

Proposition 7 (Special Case II: Forecast Unbiasedness and Efficiency Tests) *Under the condition:*

$$S_{ff} = -\frac{1}{2}(FBS_{hf} + S_{fh}B'F') = FBS_{hh}B'F', \quad (22)$$

$\int_0^\tau \tilde{\omega}(s, \tau) \tilde{\omega}(s, \tau)' ds$ in Proposition 4 becomes:

(i) recursive case: $\frac{\tau - \rho}{\mu} G^{-1} S_{ff} G^{-1}$;

(ii) rolling case:

(a) $\frac{1}{\mu} \frac{2\rho}{3} G^{-1} S_{ff} G^{-1}$, if $\tau - \rho \geq \rho$; and

(b) $\frac{(\tau - \rho)}{\mu} \left(1 - \frac{(\tau - \rho)^2}{3\rho^2}\right) G^{-1} S_{ff} G^{-1}$, if $\tau - \rho < \rho$.

Furthermore, $\int_0^\tau \omega(s, \tau) \omega(s, \tau)' ds$ in Proposition 3 becomes $\lambda G^{-1} S_{ff} G^{-1}$, where:

(i') recursive case: $\lambda = 1$;

(ii') rolling case: let $\pi^\dagger \equiv \frac{\mu}{\rho}$; then,

(a) $\lambda = \frac{2}{3\pi^\dagger}$, if $\mu \geq \rho$; and

(b) $\lambda = \left(1 - \frac{1}{3}(\pi^\dagger)^2\right)$, if $\mu < \rho$.

Note that, when either Proposition 6 or 7 holds, $V_{\theta, \tau} = \int_0^\tau \omega(s, \tau) \omega(s, \tau)' ds$ does not depend on τ ; thus, the variance is not time-varying. The next proposition shows that, in these special cases, the distribution of the test statistic simplifies and its critical values can be tabulated.

Theorem 8 (Main Proposition in Special Cases) (a) *Under Assumption 1-4 and Condition (22), we have:*

$$\mathcal{W}_{j,m} \Rightarrow \mathcal{W}_{\tau,\mu}, \quad (23)$$

where $\mathcal{W}_{\tau,\mu}$ is: (i) Recursive estimation.¹⁴

$$\mathcal{W}_{\tau,\mu} = \mu^{-1} [\mathcal{B}_\ell(\tau - \rho) - \mathcal{B}_\ell(\tau - \rho - \mu)]' [\mathcal{B}_\ell(\tau - \rho) - \mathcal{B}_\ell(\tau - \rho - \mu)], \quad (24)$$

¹⁴If one implements the test by recursively estimating $\hat{\theta}_j$ as explained in Footnote 10 (that is, $\hat{\theta}_j$ is estimated sequentially in regression (1) for $j = R + m, \dots, T$ using all past $j - R$ observations), then the asymptotic distribution of the

(ii) *Rolling estimation:*

$$\begin{aligned}
\mathcal{W}_{\tau,\mu} = & \mu^{-1} \left\{ \left(\frac{2}{3\pi^\dagger} \right) \cdot 1(\mu \geq \rho) + \left(1 - \frac{1}{3} (\pi^\dagger)^2 \right) \cdot 1(\mu < \rho) \right\}^{-1} \times \tag{25} \\
& \left[\left\{ \mathcal{B}_\ell \left((\tau - \rho) \left(1 - \frac{(\tau - \rho)^2}{3\rho^2} \right) \right) - \mathcal{B}_\ell \left((\tau - \mu - \rho) \left(1 - \frac{(\tau - \mu - \rho)^2}{3\rho^2} \right) \right) \right\} \cdot 1(\mu + \rho \leq \tau < 2\rho) \right. \\
& + \left. \left\{ \mathcal{B}_\ell \left(\frac{2}{3}\rho \right) - \mathcal{B}_\ell \left((\tau - \mu - \rho) \left(1 - \frac{(\tau - \mu - \rho)^2}{3\rho^2} \right) \right) \right\} \cdot 1(2\rho < \tau \leq 2\rho + \mu) \right]' \\
& \times \left[\left\{ \mathcal{B}_\ell \left((\tau - \rho) \left(1 - \frac{(\tau - \rho)^2}{3\rho^2} \right) \right) - \mathcal{B}_\ell \left((\tau - \mu - \rho) \left(1 - \frac{(\tau - \mu - \rho)^2}{3\rho^2} \right) \right) \right\} \cdot 1(\mu + \rho \leq \tau < 2\rho) \right. \\
& + \left. \left\{ \mathcal{B}_\ell \left(\frac{2}{3}\rho \right) - \mathcal{B}_\ell \left((\tau - \mu - \rho) \left(1 - \frac{(\tau - \mu - \rho)^2}{3\rho^2} \right) \right) \right\} \cdot 1(2\rho < \tau \leq 2\rho + \mu) \right].
\end{aligned}$$

(b) *Furthermore, under Assumptions 1-4 and condition $F = 0$, equation (23) holds with*

$$\mathcal{W}_{\tau,\mu} = \mu^{-1} [\mathcal{B}_\ell(\tau - \rho) - \mathcal{B}_\ell(\tau - \rho - \mu)]' [\mathcal{B}_\ell(\tau - \rho) - \mathcal{B}_\ell(\tau - \rho - \mu)]. \tag{26}$$

We reject the null hypothesis:

$$H_0 : \theta_j = \theta_0, \theta_0 = 0 \text{ for all } j = R + m, \dots, T \tag{27}$$

if $\max_{j \in \{R+m, \dots, T\}} W_{j,m} > \kappa_{\alpha,\ell}$, where $\kappa_{\alpha,\ell}$ are the critical values at the 100 α % significance level and are reported for $\alpha = 0.05$ in Table 1, Panel A for equations (24) and (26) and in Table 1, Panel B for equation (25) for various combinations of $\rho = R/T, \mu = m/T$ and number of restrictions, ℓ .¹⁵

Under the special cases considered in Propositions 6 or 7, which are the ones more commonly used in the literature, the critical values do not depend on the data generating process and can be tabulated. More specifically, Theorem 8 shows that this is the case when either: (i) $F = 0$; or (ii) when testing forecast unbiasedness and rationality via t-tests (that is, when concerned about mean prediction errors and efficiency). Based on the considerations in West and McCracken (1998), results similar to (ii) apply when testing encompassing and serial correlation with conditionally homoskedastic errors. In these cases, our method results in an adjustment procedure similar to

Wald test (when the parameters are estimated recursively and θ is rescaled by \sqrt{T}) is $\mathcal{W}_{\tau,\mu} = \mathcal{B}_\ell(\tau - \rho)' \mathcal{B}_\ell(\tau - \rho)$. The 5% critical values for $\sup_{\tau \in \{\rho+\mu, \dots, 1\}} \mathcal{W}_{\tau,\mu}$ when $\mu = 0.25, \rho = \{0.3; 0.5\}$ and $l = 1(l = 2)$ are $\{3.3332; 2.3755\}$ ($\{5.1071; 3.6248\}$), respectively.

¹⁵The critical values can be obtained by Monte Carlo simulation. The critical values at significance level 100 α % are such that $\Pr \left\{ \sup_{\tau \in \{\rho+\mu, \dots, 1\}} \mathcal{W}_{\tau,\mu} > \kappa_{\alpha,\ell} \right\} = \alpha$. Critical values at the 1% and 10% as well as for $\ell > 5$ can be found in a Not-for-Publication Appendix (Rossi and Sekhposyan, 2014).

that of West and McCracken (1998), where the test statistics could be calculated similarly, by substituting P in their notation with m of our notation, provided inference is conducted using the critical values provided in this paper.¹⁶

Note that in the case of model-free forecasts, the only sample available to researchers is P : they do not have an available R ; therefore we define $\tilde{\mu}$ such that $m = [\tilde{\mu}P]$ (that is, we define m as a fraction of the number of observations P , as opposed to being a fraction of the total sample size T). We provide a separate table of critical values (Table 1, Panel C) for the test statistic $\max_{\tilde{j} \in \{m, \dots, P\}} \mathcal{W}_{\tilde{j}, m}^{\tilde{j}}$ as discussed in the following Corollary.

Corollary 9 (Main Proposition for Survey and Model-Free Forecasts) *Under Assumptions 1-4 and condition $F = 0$, the alternative test statistic*

$$\mathcal{W}_{\tilde{j}, m}^{\tilde{j}} = m \hat{\theta}_{\tilde{j}}^t \hat{V}_{\theta}^{-1} \hat{\theta}_{\tilde{j}}, \text{ for } \tilde{j} = m, \dots, P,$$

implemented over the sequence of P forecasts is such that:

$$\max_{\tilde{j} \in \{m, \dots, P\}} \mathcal{W}_{\tilde{j}, m}^{\tilde{j}} \implies \sup_{\tilde{\tau} \in \{\tilde{\mu}, \dots, 1\}} \mathcal{W}_{\tilde{\tau}, \tilde{\mu}}^{\tilde{j}}, \quad (28)$$

where $\tilde{\tau} \equiv \tilde{j}/P$, $\tilde{\mu} \equiv m/P$ and

$$\mathcal{W}_{\tilde{\tau}, \tilde{\mu}}^{\tilde{j}} = \mu^{-1} [\mathcal{B}_{\ell}(\tilde{\tau}) - \mathcal{B}_{\ell}(\tilde{\tau} - \tilde{\mu})]' [\mathcal{B}_{\ell}(\tilde{\tau}) - \mathcal{B}_{\ell}(\tilde{\tau} - \tilde{\mu})]. \quad (29)$$

We reject the null hypothesis:

$$H_0 : \theta_j = \theta_0, \theta_0 = 0 \text{ for all } \tilde{j} = m, \dots, P \quad (30)$$

if $\max_{\tilde{j} \in \{m, \dots, P\}} \mathcal{W}_{\tilde{j}, m}^{\tilde{j}} > \kappa_{\alpha, \ell}$, where $\kappa_{\alpha, \ell}$ are the critical values at the 100 α % significance level and are reported in Table 1, Panel C for various values of $\tilde{\mu} = m/P$ and number of restrictions, ℓ .

6 Monte Carlo Analysis

We study the small sample performance of the methods that we propose in a series of Monte Carlo experiments inspired by West and McCracken (1998). Let the Data Generating Process (DGP) be: $y_t = \gamma y_{t-1} + \varepsilon_t$, where $\gamma = 0.5$, $\varepsilon_t \sim iid(0, 1)$, and y_0 is drawn from its unconditional distribution, a normal with zero mean and variance $(1 - \gamma^2)^{-1}$. In each sample, $t = 1, \dots, T + 1$, we split the data into $T + 1 = P + R$, and we use either a rolling or a recursive scheme to generate P

¹⁶While West and McCracken (1998) formally demonstrate that their condition holds for one-step-ahead forecasts, conversations with the authors, as well as comments in their paper (p. 829) suggest that it holds for $h > 1$ as well.

one-step ahead out-of-sample forecasts, $y_{R+1|R}, \dots, y_{T+1|T}$. The forecasting model for the recursive estimation scheme is $y_{t+1|t} = \hat{\gamma}_t y_t$, while for the rolling estimation scheme it is $y_{t+1|t} = \hat{\gamma}_{0,t} + \hat{\gamma}_{1,t} y_t$. The forecast errors are denoted by $v_{t+1|t} \equiv y_{t+1} - y_{t+1|t}$, for $t = R, \dots, T$.¹⁷

First, we consider the size properties of our test and compare it to the ‘‘Traditional tests’’ typically used in the literature. Consider the following regression model:

$$E_t v_{t+1|t} = \theta_0 + \theta_1 Z_t. \quad (31)$$

The ‘‘Traditional tests’’ include testing:

(i) Zero mean prediction error (or forecast unbiasedness). We test whether the mean of the sequence of forecast errors is zero. The test is implemented by a two-sided t-test for $\theta_0 = 0$ in regression (31), where there are no regressors other than the constant. Let $\hat{\theta}_0 = P^{-1} \sum_{t=R}^T v_{t+1|t}$ and $\hat{\sigma}_{\theta_0}^2 = P^{-1} \sum_{t=R}^T (v_{t+1|t})^2$. We consider a t-test with West and McCracken’s (1998) variance correction: $t_{\alpha}^c = P^{1/2} \hat{\theta}_0 \hat{\sigma}^{-1} \lambda_{WM}^{-1/2}$, where $\lambda_{WM} = 1$ for the recursive scheme and, for the rolling scheme, $\lambda_{WM} = 1 - \pi^2/3$ when $\pi < 1$ and $\lambda_{WM} = 2/(3\pi)$ when $\pi \geq 1$.

(ii) Forecast efficiency. The test is implemented by a two-sided t-test for $\theta_1 = 0$ in the regression (31), where $Z_t = y_{t+1|t}$, and a constant is included in the regression. Let $\hat{\theta}_1$ be the OLS estimate of the slope coefficient in the regression (31), and $\hat{\sigma}_{\theta_1}^2$ be its estimated standard error. We consider a t-test that utilizes West and McCracken’s (1998) correction: $t_{\theta_1}^c = \hat{\beta} \hat{\sigma}_{\theta_1}^{-1} \lambda_{WM}^{-1/2}$, for the same values of λ_{WM} as in (i).

In addition, we consider our proposed ‘‘Fluctuation Rationality’’ test, equation (4), implemented in rolling regressions over the out-of-sample period with a rolling window size $m = 50$. The estimate of the asymptotic variance is obtained by using a simple homoskedastic covariance estimate of S , as in West and McCracken (1998). The number of Monte Carlo replications is 1,000.

Table 2 reports results for the recursive and the rolling estimation schemes, respectively. Panel A reports results for testing forecast unbiasedness and panel B for forecast efficiency. The tables shows that the empirical rejection frequencies of our proposed tests (reported in the column labeled ‘‘Fluctuation Test’’) as well as those of the traditional tests (reported in the column labeled ‘‘Traditional Test’’) are close to the nominal value except in very small sample sizes. The size distortions in small samples are mild for the recursive scheme for both tests and for the rolling scheme for the mean prediction error test, and a little bit larger for the rolling case in the efficiency test.¹⁸ In

¹⁷The advantage of using the same DGP as West and McCracken (1998) is that we can directly compare our results to theirs. In addition, in order for Condition (22) to hold, one would need to include a constant in the estimation equation for the rolling case. For a reference, see West and McCracken (1998), proof of Theorem 7.1.

¹⁸This result is consistent with West and McCracken (1998). As a referee pointed out, it might be possible that the finite sample correction in Giacomini and Rossi (2009) could improve the size distortions in our context. We leave this for future research.

general, for the small samples, i.e. $R \leq 100$, the recursive estimation scheme results in a better sized tests than the rolling estimation scheme.

In order to evaluate the power of our test in the presence of time variation, we consider experiments based on three DGPs. All DGPs are based on the model: $y_t = \gamma y_{t-1} + \varepsilon_t + b_t$, where $\gamma = 0.5$, $\varepsilon_t \sim iid(0, 1)$, y_0 is drawn from the unconditional distribution of y_t when $b_t = 0$. In DGP A (labeled as “A. Non-stationary”), $b_t = b \cdot 1(1 < t \leq 345) - b \cdot 1(345 < t \leq T + 1)$. In DGP B (labeled as “B. Non-stationary”), $b_t = 0 \cdot b \cdot 1(1 < t \leq 345) + 2 \cdot b \cdot 1(345 < t \leq T + 1)$. DGP C (labeled as “C. Stationary”) considers $b_t = b$, for all t . Furthermore, $b = \{0, 0.1, \dots, 1\}$ for the power exercises in the case of mean prediction error and $b = \{0, 0.5, \dots, 5\}$ for the power exercises in the case of efficiency. Predictions are based on an AR(1) model. DGP A is used to assess the power of the “Fluctuation Rationality” test. DGP B is a non-stationary DGP where traditional tests can also have some power. DGP C is a stationary model that we use to study the power loss from using our test that is robust in the presence of instabilities. The power loss occurs since our test uses fewer observations than the traditional tests, i.e. $m < P$. In all cases parameters are estimated with a recursive scheme, and $T = 400$, $R = 300$ and $m = 60$.

The power comparisons are reported in Table 3. The table shows that, in DGP A, the traditional tests do not have power to reject the null hypothesis, and, in fact, their rejection frequencies approach zero under the alternative hypothesis, whereas our proposed tests do have substantial power (see panel A. Non-stationary). In DGP B, the traditional test has some power, although our test has a higher power to detect lack of rationality (see panel B. Non-stationary). Finally, DGP C illustrates the loss of power in our test relative to the traditional test when there are no instabilities in the data (see panel C. Stationary). Clearly, there is a trade-off between the proposed tests and the traditional ones: if one is certain that the forecast environment is stable, the traditional tests would have more power to detect lack of rationality in small samples; however, when the forecast environment is unstable, the traditional tests, even asymptotically, may have no power at all in certain situations, while our proposed test would have power.

7 Are Federal Reserve and Private Sector’s Forecasts Rational?

The quality of private sector’s forecasts relative to the internal forecasts of the Federal Reserve has been frequently considered in the literature. As anticipated in Section 2, in important contribution, Romer and Romer (2000) showed that the Federal Reserve has more information relative to the private sector when forecasting inflation. Hence, it would be optimal for a third party with access to both forecasts to put all the weight on the forecasts provided by the Federal Reserve and zero weight on the ones provided by the commercial forecasters.

We revisit the existing empirical evidence from two points of view. First, we consider the

rationality of private sector's as well as the Federal Reserve's Greenbook forecasts, as in Romer and Romer (2000), Faust and Wright (2010), Patton and Timmermann (2012) and Croushore (2012), among others. These papers have found that forecast rationality tests for the various, competing inflation forecasts are sensitive to the sub-sample period used for forecast evaluation. The novelty of our approach is to study whether forecast rationality holds by using our "Fluctuation Rationality" test, which is more powerful in the presence of instabilities. One of the advantages of our approach is that it does not require researchers to know or impose a sub-sample date a-priori. Second, we evaluate whether Romer and Romer's (2000) finding that Federal Reserve forecasts are superior to private sectors' forecasts continues to hold when we allow for instabilities.

We consider the Federal Reserve's inflation forecasts provided in the Greenbook and compare them with two commercial forecasts: the Blue Chip Economic Indicators (BCEI) and the Survey of Professional Forecasters (SPF). In what follows, we describe the data from each of the sources.

Greenbook forecasts are made by the staff of the Federal Reserve Board of Governors prior to each Federal Open Market Committee (FOMC) meeting. The Greenbook provides quarterly forecasts for a variety of economic indicators and for several forecast horizons (from contemporaneous up to nine quarters) under a maintained assumption about monetary policy. We consider only forecasts up to five quarters to ensure a sample large enough for inference. We focus on inflation forecasts provided by the Greenbook, which are measured by (annualized) quarter-over-quarter GNP deflator growth rates from 1965 to 1991 and by (annualized) quarter-over-quarter GDP deflator growth rates afterwards. Greenbook forecasts are available only with a five-year lag. Thus, our current sample includes data up to 2008:IV. The data are provided by the Federal Reserve Bank of Philadelphia, which matches the timing of the Greenbook forecasts with that of the SPF: the dataset includes forecasts from four of the annual FOMC meetings whose date is the closest to the middle of the quarter.¹⁹ In order to make the two datasets comparable, we omit the first 3 years of observations and start the series in 1968:IV.

The Survey of Professional Forecasters (SPF) provides forecasts of inflation as well as a variety of economic fundamentals at the quarterly frequency. These include nowcasts (forecasts of the current quarter) as well as forecasts up to four quarters ahead. We use the forecasts of (annualized) quarterly GNP/GDP deflator growth rates. The survey is conducted roughly at the end of every second month in the quarter and includes 34 professional forecasts. The series start in 1968:IV. We use the median forecast and end our series in 2008:IV to obtain a dataset spanning the same time period as the Greenbook.

The Blue Chip Economic Indicators (BCEI) provides monthly forecasts of quarterly economic

¹⁹Greenbook forecasts can be obtained from the Philadelphia Fed web-site at <http://www.philadelphiafed.org/research-and-data/real-time-center/greenbook-data/>, while the SPF forecasts are at <http://www.philadelphiafed.org/research-and-data/real-time-center/survey-of-professional-forecasters/>.

series starting from 1980. It is a survey-based forecast dataset where about 50 US business economists participate each month. Though this is a monthly series, in order to match the Greenbook and SPF forecasts we take only four forecasts per year corresponding to the mid-quarter, i.e. February, May, August, and November, from 1980 to 2008.²⁰

To evaluate the Greenbook, SPF and Blue Chip forecasts, we use realized values of (annualized) quarter-over-quarter growth rates of the GNP/GDP deflator constructed from the quarterly vintages in the real-time dataset discussed by Croushore and Stark (2001). Our forecast evaluation approach is consistent with that in Romer and Romer (2000), who use the second revision as the benchmark for forecast evaluation, i.e. the specific quarter data available at the last month of the consecutive quarter. Given the real-time nature of the dataset, the way we construct the (annualized) quarter-over-quarter GNP/GDP deflator based inflation rate is as follows. For example, the deflator for 1968:IV uses the 1969:I vintage and applies the following transformation: $100((PGDP68 : IV/PGDP68 : III)^4 - 1)$. We do so for all the vintages up to 2010:II, then take the diagonal elements of the resulting matrix. This way we obtain a real-time, annualized measure of the quarter-over-quarter inflation rate of the previous quarter, which we use to evaluate the nowcast or the corresponding h -quarter-ahead forecast over time.²¹

Figure 1 compares the Greenbook forecasts with those of the SPF and BCEI at the two-quarter-ahead forecast horizon. The figure also plots the realized values for inflation at each horizons (reported by the solid line, labeled “actual”). In the figure, not all forecasts have the same starting point. In addition, there are several missing values, depending on when the forecast has been made. Overall, the forecasts appear to be correlated with each other. Panel A in Table 4 reports the mean squared forecast errors (MSFE) for the forecast at various horizons. It appears that the SPF forecasts are inferior to those of the Greenbook at all horizons whereas the BCEI forecasts appear to be superior. However, as we show further, there is substantial evidence of instabilities, and the difference is most likely explained by the different sample period that BCEI covers.

To evaluate the forecast performance, we consider the following regression:

$$\pi_{t+h} - \pi_{t+h|t} = \alpha + \beta\pi_{t+h|t} + \epsilon_{t+h}, \quad (32)$$

where π_{t+h} is the realized inflation rate, $\pi_{t+h|t}$ is the inflation expectation for $t+h$ based on the information available at the time t , h is the forecast horizon and ϵ_{t+h} is a forecast error. We consider $h = 0, 1, \dots, 5$ for the Greenbook forecast, $h = 0, 1, \dots, 5$ for the BCEI and $h = 0, 1, \dots, 4$ for the SPF. For the Greenbook, the choice of h is constrained by the need to have a sample size large enough for inference. The choice of h for the BCEI and SPF is dictated by data availability.

²⁰ Although the BCEI forecasts are available from August 1976, the forecasts for the initial four years are for annual changes in key economic variables as opposed to quarterly, thus we omit the earlier period.

²¹ In the real-time dataset provided by the Philadelphia Fed, the observation for 1995:IV is missing in the vintage of 1996:I. We use the value available in the vintage of 1996:II as a substitute value.

Table 4, Panel B, presents results for both traditional forecast rationality tests as well as the “Fluctuation Rationality” test that we propose. The former relies on the maintained assumption that the parameters of the regression are time invariant and it is implemented using a simple Wald-type test in equation (32), where the parameters are estimated by OLS; we use a HAC variance estimate (Newey and West, 1987) with a bandwidth equal to $P^{1/4}$. Our proposed test instead assesses whether the parameters equal the values implied by forecast rationality at any given point in time, and it is robust in the presence of instabilities. The test is implemented as in equation (4), where $\hat{\theta}_j$ are the OLS estimates of α and β from equation (32) in rolling regressions with a window of size $m = 60$.

The choice of m may affect the empirical results. In fact, m operates as a smoothing parameter and its choice balances a usual trade-off between bias and efficiency: in the absence of instabilities, a larger m would guarantee the most efficient estimate, but in the presence of instabilities a smaller m would reduce the bias. An optimal choice of m would require assumptions on the nature of the instabilities in the unknown data generating process. To avoid that, we recommend researchers to check the sensitivity of their results to the choice of m , while making sure that the values of m are large enough to ensure that Assumption 3 holds. In fact, following this recommendation, we also examine the robustness of our results to $m = 25$ further below.

The column labeled “Traditional” in Panel B of Table 4 reports the test statistic W_P in equation (3) and the column labeled “Fluctuation” reports $\max_{j \in \{m, \dots, P\}} W_{j,m}^{\sim}$ in Proposition 9; both are reported for several horizons h , listed in the first column. Asterisks denote significance at the 5% significance level. The panel suggests that traditional forecast rationality tests fail to reject the null hypothesis of forecast rationality at the 5% significance level for the Greenbook and SPF forecasts, whereas they reject forecast rationality for the BCEI forecasts. However, as we show later, this difference in the results highly depends on the evaluation period, as the sample for the BCEI forecasts starts much later. It is, in fact, during a period of time when the Greenbook and SPF forecasts fail the rationality test as well. In contrast, the “Fluctuation Rationality” test rejects the null hypothesis of rationality for all forecasts.

Figure 2 plots $W_{j,m}^{\sim}$ for several horizons together with the critical values for the $\max_{j \in \{m, \dots, P\}} W_{j,m}^{\sim}$ test statistic at the 5% significance level.²² The dates on the horizontal axis provide useful information about the timings of the forecast rationality breakdowns. The dashed line plots the test for the Greenbook forecasts. The figure shows three substantial breakdowns: the first two are associated with the beginning and the end of 1990s. It appears that the forecasts deteriorate over the 1990s and rationality tends to recover by the 2000s. However, for almost all forecast horizons, forecast rationality breaks down again around 2005. Overall, it appears that the empirical evidence

²²Given that the sample size for various forecasts at different horizons is not the same, the critical values for 5% significance level could be different. We plot the maximum of these critical values.

in favor of forecast rationality supported by the traditional forecast rationality tests, reported in Table 4, is driven mainly by the good performance of the Greenbook forecasts at the beginning of our sample. The dotted line plots the “Fluctuation Rationality” test for the BCEI forecasts while the dashed and dotted line reports the same test for the SPF forecasts. The test suggests that the empirical evidence on forecast rationality for SPF forecasts is qualitatively similar to that of the Greenbook. However, the recovery of forecast rationality during the first half of 2000s is less pronounced for SPF than for the Greenbook. By comparing the BCEI and the SPF, we note that they behave similarly in the overlapping part of the evaluation period with some differences in the timing of the rationality breakdown that occurs in the second half of 2000s. This suggests that the traditional forecast rationality test results for the BCEI reported in Table 4 are different from the other forecasts solely due to the different sample period. In general, the empirical evidence in Figure 2 does not support forecast rationality for any of the forecasts at any horizons.

An important question is why professional forecasters and the Federal Reserve both made systematic forecast errors for so long. It could be that they were focused on forecasting a different inflation definition (such as core inflation), or that it was difficult to distinguish between permanent and transitory effects of the shocks, or it could be that their forecasts are targeted at a different horizon. Without knowing the actual forecasting model that they used, it is difficult to answer this question. It is important to note that violations of forecast efficiency are not necessarily evidence that forecasters’ performance is sub-optimal if there are occasional structural breaks that forecasters learn about slowly. For example, if the process for inflation switches between different regimes, forecasters’ models may adjust only slowly to the switch, as it would be hard to distinguish between an outlier and a regime change; during the learning process, forecasts may look irrational.

Figure 3 reports robustness results to the choice of m . When we choose a smaller value ($m = 25$), the test has higher power to reject rationality in the Federal Reserve’s forecasts even earlier in the sample, emphasizing the systematic downward forecast errors in the 1970s.

Our second objective is to assess whether the Federal Reserve has an information advantage over private sector’s forecasts. To do so, we consider the following regression:

$$\pi_{t+h} - \pi_{t+h|t}^i = \delta + \beta_g \pi_{t+h|t}^g + \beta_i \pi_{t+h|t}^i + \eta_{t+h}, \quad (33)$$

where $\pi_{t+h|t}^g$ is the Greenbook forecast and $\pi_{t+h|t}^i$, $i = SPF, BCEI$ denote the SPF and BCEI forecasts, respectively. The Federal Reserve forecasts are useful beyond that of the private sector in predicting inflation if and only if $\beta_g \neq 0$. We test this hypothesis both with the traditional tests as well as with our Fluctuation-type test. The latter test is implemented as in equation (4), where $\hat{\theta}_j$ are the OLS estimates of δ and β_g from equation (33) in rolling regressions with a window of size $m = 60$.

The results are reported in Table 4, Panel C. The table reports the traditional test statistics

(column labeled “Traditional”) and the Fluctuation-type test statistic (column labeled “Fluctuation”); asterisks denote significance at the 5% level. According to the table, both the traditional tests and the Fluctuation-type test suggest statistically significant evidence that the Federal Reserve has additional information relative to the private sector’s forecasts. Figure 4 sheds additional light on this conclusion. The figure plots the Fluctuation-type test statistics over time and shows that the information advantage of the Federal Reserve has deteriorated after 2003. In fact, the rejections of the hypothesis of no information advantage of the Federal Reserve based on the Fluctuation test appear mostly at the beginning of the sample. The result holds for both commercial forecasts, that is the BCEI and the SPF.

Finally, Figure 5 plots the coefficients on Federal Reserve’s Greenbook forecasts, β_g , in equation (33) implemented with either the BCEI (dotted line) or the SPF (dashed-dotted line) forecasts as additional explanatory variables in the rolling regressions. The figure suggests that the coefficient for most horizons averages around unity. However, the coefficient seems to have been decreasing over time across all horizons. The picture also shows a mild revival of the information advantage between 1995-2001: it appears there is evidence in favor of the presence of additional explanatory power of the Greenbook forecasts around 1995, which starts diminishing around 2001.

8 Conclusion

This paper proposes new forecast rationality tests that can be used in unstable environments. The tests we propose can be applied to test forecast unbiasedness, efficiency, encompassing, serial uncorrelation and, in general, regression-based tests of forecasting ability. Our test statistics have non-standard limiting distributions and depend on nuisance parameters; in special cases that are relevant in practice, the critical values can be tabulated, thus making the test easily implementable. Our paper also analyzes the size properties of the test that we propose in small samples, as well as the power of our tests relative to traditional tests in the presence of instabilities. We show that traditional tests may fail to reject forecast optimality in the presence of instabilities whereas our test performs well in that regard.

The methods we propose are robust to data mining because the critical values are based on the supremum of the statistics across all samples. The test is not robust to data mining due to the choice of the out-of-sample window size; to resolve the latter issue, the reader is referred to Inoue and Rossi (2012) and references therein. Furthermore, we should note that our test is designed to signal lack of rationality in sub-samples of the data: it might still be that our test signals lack of rationality in sub-samples but forecasts are, on average, rational, as in the empirical analysis. It is also possible that a researcher may want to allow for a learning period or for some violations of optimality at specific points in time; the test can be adapted to these situations by examining a

plot of the test statistic $W_{j,m}$ over time.

The empirical analysis compares various private sector forecasts to those of the Federal Reserve Greenbook. We reject the forecast rationality of all these forecasts at some point in time. However, even after allowing for time-variation, we find significant evidence in favor of the Fed's additional information advantage over the private sector when predicting future inflation.

9 References

- Cavaliere, G. 2004. Unit root tests under time-varying variances. *Econometric Reviews* 23(3): 259-292. DOI: 10.1081/ETC-200028215
- Croushore D. 1998. Evaluating inflation forecasts. *FRB of Philadelphia Working Paper* No 98-14.
- Croushore D. 2006. An evaluation of inflation forecasts from surveys using real-time data. *FRB of Philadelphia Working Paper* No. 06-19.
- Croushore D. 2012. Forecast bias in two dimensions. *FRB of Philadelphia Working Paper* No 12-9.
- Croushore D, Stark T. A real-time data set for macroeconomists. *Journal of Econometrics* 105(1): 111-130. DOI: 10.1016/S0304-4076(01)00072-0
- Diebold FX, Lopez JA. 1996. Forecast evaluation and combination. In *Handbook of Statistics* 14, Maddala GS, Rao CR (eds.). North-Holland: Amsterdam.
- Faust J, Wright JH. 2009. Comparing greenbook and reduced form forecasts using a large real-time dataset. *Journal of Business and Economic Statistics* 27(4): 468-479. DOI: 10.1198/jbes.2009.07214
- Gamber EN, Smith J. 2009. Are the Fed's forecasts still superior to the private sector's? *Journal of Macroeconomics* 31(2): 240-251. DOI: 10.1016/j.jmacro.2008.09.005
- Giacomini R, Rossi B. 2009. Detecting and predicting forecast breakdowns *Review of Economic Studies* 76(2): 669-705. DOI: 10.1111/j.1467-937X.2009.00545.x
- Giacomini R, Rossi B. 2010. Forecast comparisons in unstable environments. *Journal of Applied Econometrics* 25(4): 595-620. DOI: 10.1002/jae.1177
- Granger CWJ, Newbold P. 1986. *Forecasting Economic Time Series*, 2nd ed. Academic Press: New York.
- Hansen BE. 1992. Convergence to stochastic integrals for dependent heterogeneous processes. *Econometric Theory* 8(4): 489-500. DOI: 10.1017/S0266466600013189
- Inoue A, Rossi B. 2012. Out-of-sample forecast tests robust to the window size choice. *Journal of Business and Economics Statistics* 30(3): 432-453. DOI: 10.1080/07350015.2012.693850
- Mincer J, Zarnowitz V. 1969. The evaluation of economic forecasts. In *Economic Forecasts*

and Expectations: Analysis of Forecasting Behavior and Performance, Mincer JA (ed.), National Bureau of Economic Research: New York.

Newey WK, West KD. 1987. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* 55(3): 703-708. DOI: 10.2307/1913610

Patton A, Timmermann A. 2010. Generalized forecast errors, a change of measure, and forecast optimality conditions. In *Volatility and Time Series Econometrics: Essays in Honor of Robert F. Engle*, Bollerslev T, Russell JR, Watson MW (eds.). Oxford University Press.

Patton A, Timmermann A. 2012. Forecast rationality tests based on multi-horizon bounds. *Journal of Business and Economic Statistics* 30(1): 1-17. DOI: 10.1080/07350015.2012.634337

Rao CR. 1965. *Linear Statistical Inference and its Applications*, 2nd ed. John Wiley & Sons.

Romer CD, Romer DH. 2000. Federal Reserve information and the behavior of interest rates. *American Economic Review* 90(3): 429-457. DOI: 10.1257/aer.90.3.429

Rossi B. 2012. Comment to: Forecast rationality tests based on multi-horizon bounds. *Journal of Business and Economic Statistics* 30(1): 25-29. DOI:10.1080/07350015.2012.634343

Rossi B. 2013. Advances in forecasting under instabilities. In *Handbook of Economic Forecasting* 2B, Elliott G, Timmermann A. (eds.), *Handbook of Economic Forecasting* 2B. North-Holland: Amsterdam.

Rossi B, Sekhposyan T. 2014. Not-for-publication appendix to: Forecast rationality tests in the presence of instabilities, with applications to the Federal Reserve and survey forecasts. Available at <http://www.barbararossi.eu> and <http://www.tateviksekhposyan.org/>.

Sinclair TM, Joutz F, Stekler HO. 2010, Can the Fed predict the state of the economy? *Economics Letters* 108(1): 28-32. DOI: 10.1016/j.econlet.2010.04.010

West KD. 1996. Asymptotic inference about predictive ability. *Econometrica* 65(5): 1067-1084. DOI: 10.2307/2171956

West KD, McCracken MW. 1998. Regression-based tests of predictive ability. *International Economic Review* 39(4): 817-840. DOI: 10.2307/2527340

Tables

Table 1. Critical Values for the Fluctuation Rationality Test

Panel A. Recursive Case									
ℓ	$\rho = 0.5; \mu = 0.25$	$\rho = 0.5; \mu = 0.3$	$\rho = 0.3; \mu = 0.25$	$\rho = 0.3; \mu = 0.3$					
1	7.6103	6.8123	8.1529	8.0414					
2	10.7828	10.3909	12.1409	11.1946					
3	12.6497	11.8263	14.2097	13.1495					
4	14.8763	14.2381	15.8727	15.4504					
5	16.4838	16.1415	17.9421	17.4355					
Panel B. Rolling Case									
ℓ	$\rho = 0.5; \mu = 0.25$	$\rho = 0.5; \mu = 0.3$	$\rho = 0.3; \mu = 0.25$	$\rho = 0.3; \mu = 0.3$					
1	7.7122	6.9621	8.8102	8.5989					
2	10.5702	10.0698	12.4778	12.1265					
3	13.2956	12.3069	14.5513	13.9501					
4	14.8771	14.2805	16.6307	15.6392					
5	16.8451	16.6441	19.0969	18.6127					
Panel C. Survey and Model-Free Forecasts									
ℓ	$\tilde{\mu} = 0.1$	$\tilde{\mu} = 0.2$	$\tilde{\mu} = 0.3$	$\tilde{\mu} = 0.4$	$\tilde{\mu} = 0.5$	$\tilde{\mu} = 0.6$	$\tilde{\mu} = 0.7$	$\tilde{\mu} = 0.8$	$\tilde{\mu} = 0.9$
1	11.8290	10.5637	8.9252	8.1468	8.1409	7.2803	6.4978	6.0837	5.4695
2	14.9966	13.0846	12.8141	10.9084	11.1314	9.9386	9.1724	9.0589	7.8305
3	17.6768	15.7548	15.0608	13.4383	13.2113	12.6018	10.9597	10.8426	9.4727
4	19.8434	17.6051	17.0158	16.3186	15.1404	14.7573	13.5928	13.1087	10.8243
5	21.7091	20.4659	18.7186	18.2152	17.1092	15.6317	15.4842	13.9418	13.6335

Note. The table reports the critical values, $\kappa_{\alpha, \ell}$, for several restrictions (ℓ) at $\alpha = 5\%$ significance level for $\max_{j \in \{R+m, \dots, T\}} \mathcal{W}_{j,m}$ for the: (i) recursive scheme under condition (22) (Panel A); (ii) rolling scheme under condition (22) (Panel B); (iii) case when parameter estimation error is irrelevant as in Corollary 9 (Panel C). Critical values are based on 1000 Monte Carlo simulations. For Panels A and B, $T = 1000$, $\rho = R/T$ and $\mu = m/T$, while $P = 1000$ and $\tilde{\mu} = m/P$ in Panel C.

Table 2. Size

		Recursive Case				Rolling Case			
		Traditional Test		Fluctuation Test		Traditional Test		Fluctuation Test	
R/P		100	200	100	200	100	200	100	200
Panel A. Mean Prediction Error									
25		0.068	0.050	0.054	0.064	0.113	0.155	0.112	0.179
50		0.052	0.041	0.070	0.075	0.069	0.078	0.077	0.103
100		0.038	0.036	0.064	0.074	0.067	0.062	0.078	0.095
200		0.056	0.054	0.070	0.056	0.063	0.055	0.071	0.097
300		0.046	0.054	0.065	0.084	0.054	0.053	0.091	0.085
400		0.043	0.056	0.073	0.075	0.051	0.045	0.056	0.076
Panel B. Efficiency Test									
25		0.045	0.058	0.071	0.101	0.924	0.999	0.546	0.784
50		0.049	0.055	0.073	0.097	0.226	0.721	0.092	0.128
100		0.042	0.054	0.072	0.125	0.063	0.129	0.050	0.056
200		0.050	0.047	0.086	0.068	0.045	0.060	0.054	0.055
300		0.060	0.059	0.079	0.111	0.060	0.036	0.048	0.047
400		0.070	0.041	0.085	0.083	0.047	0.051	0.048	0.050

Note. Table 2 reports empirical rejection frequencies of the traditional test statistics (column labeled “Traditional Test”) and the test statistics $\max_{j \in \{R+m, \dots, T\}} \mathcal{W}_{j,m}$ (column labeled “Fluctuation Test”) under the recursive and rolling estimation schemes (see DGP in Section 3). The first column provides the R values; the columns under the header give the P values. Nominal size is 0.05 and $m = 50$.

Table 3. Power Analysis

Panel A. Mean Prediction Error						
	A. Non-Stationary		B. Non-Stationary		C. Stationary	
b	I. Traditional	II. Fluctuation	I. Traditional	II. Fluctuation	I. Traditional	II. Fluctuation
0	0.0500	0.0510	0.0500	0.0510	0.0500	0.0510
0.1	0.0500	0.0630	0.0970	0.1040	0.0880	0.0760
0.2	0.0560	0.0890	0.1920	0.2090	0.1570	0.1170
0.3	0.0420	0.0840	0.3960	0.4370	0.2820	0.2030
0.4	0.0320	0.1117	0.5430	0.6600	0.3900	0.2730
0.5	0.0280	0.1480	0.7720	0.8610	0.5540	0.3590
0.6	0.0210	0.1460	0.8820	0.9570	0.6770	0.4490
0.7	0.0160	0.1720	0.9560	0.9940	0.7830	0.5050
0.8	0.0120	0.2010	0.9810	0.9990	0.8290	0.5350
0.9	0.0080	0.1740	0.9980	1.0000	0.8700	0.5490
1.0	0.0030	0.1890	0.9980	1.0000	0.9020	0.5720

Panel B. Efficiency Test						
	A. Non-Stationary		B. Non-Stationary		C. Stationary	
b	I. Traditional	II. Fluctuation	I. Traditional	II. Fluctuation	I. Traditional	II. Fluctuation
0	0.0610	0.0700	0.0610	0.0700	0.0610	0.0700
0.50	0.0710	0.0820	0.1000	0.0960	0.1900	0.1730
1.00	0.0370	0.2020	0.4420	0.3170	0.8030	0.6620
1.50	0.0080	0.3600	0.7580	0.7000	0.9960	0.9560
2.00	0.0020	0.4930	0.8550	0.9250	1.0000	0.9900
2.50	0	0.6380	0.8730	0.9840	1.0000	1.0000
3.00	0	0.7630	0.8180	0.9920	1.0000	1.0000
3.50	0	0.8450	0.7140	0.9960	1.0000	1.0000
4.00	0	0.9090	0.5680	0.9970	1.0000	1.0000
4.50	0	0.9560	0.3830	0.9930	1.0000	1.0000
5.00	0	0.9740	0.2230	0.9960	1.0000	1.0000

Note. The table reports empirical rejection frequencies of the traditional test statistics (column labeled “Traditional”) and the test statistics $\max_{j \in \{R+m, \dots, T\}} W_{j,m}$ (column labeled “Fluctuation”) under the recursive estimation scheme. Non-Stationary cases A and B refer to DGP A and DGP B, Stationary C refers to DGP C in Section 6, respectively. Nominal size is 0.05; T = 400; R = 300; m = 60.

Table 4. Empirical Results

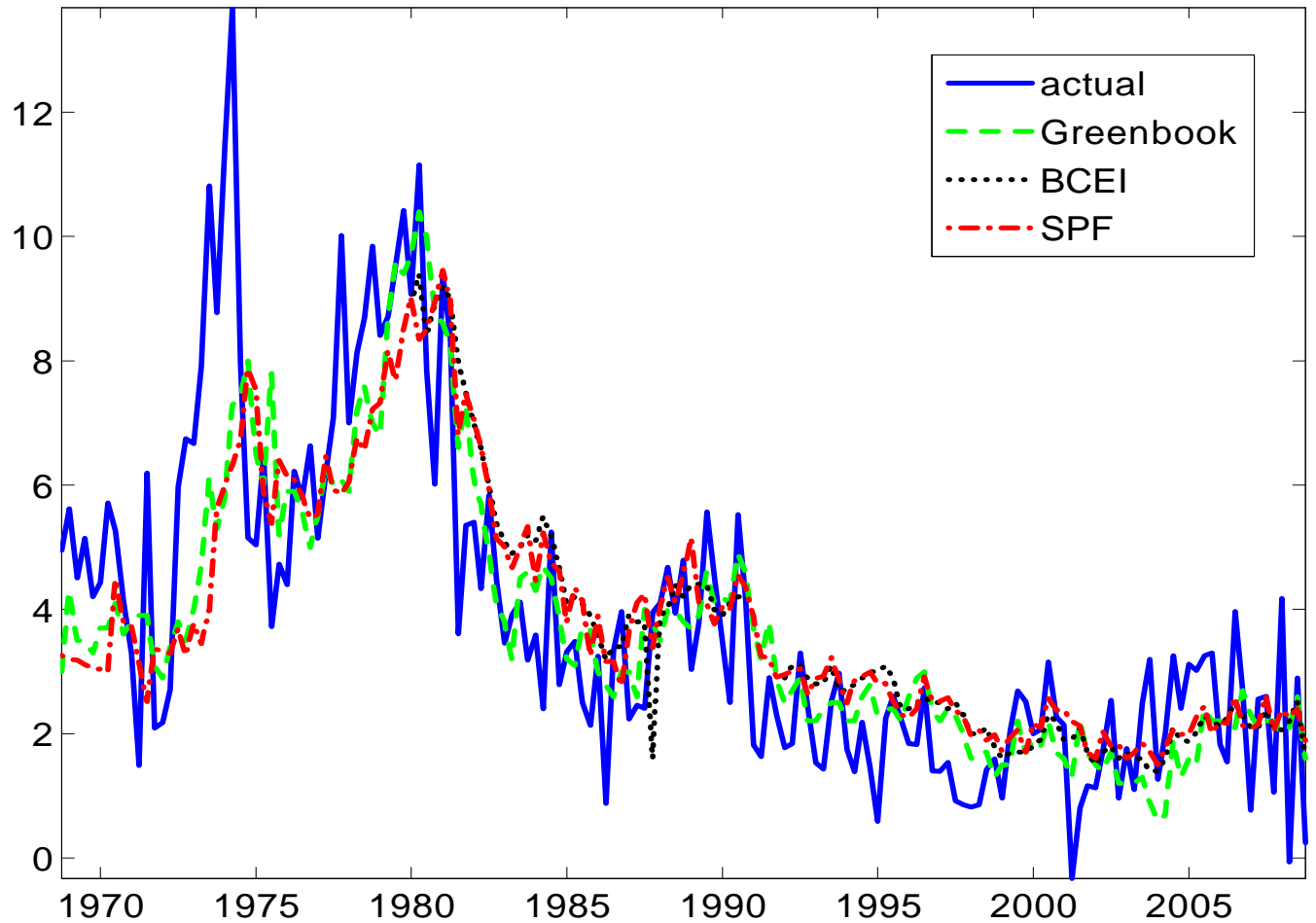
<i>h</i>	A. MSFE	B. Inflation Forecast Rationality Tests		C. Fed's Information Advantage Over Private Sector's Forecasts	
		Traditional	Fluctuation	Traditional	Fluctuation
Greenbook					
0	1.15	1.02	40.17*	--	--
1	1.70	0.16	77.36*	--	--
2	2.28	0.42	46.89*	--	--
3	2.42	0.15	49.41*	--	--
4	2.64	0.04	41.59*	--	--
5	2.60	0.22	37.88*	--	--
BCEI					
0	0.98	8.59*	25.79*	6.20*	95.46*
1	1.32	11.31*	43.75*	17.99*	49.81*
2	1.55	16.10*	51.98*	32.29*	93.01*
3	1.89	22.85*	74.23*	19.64*	52.00*
4	2.25	45.19*	135.80*	11.58*	28.10*
5	2.56	67.95*	167.51*	5.25*	2.37
SPF					
0	1.34	1.82	27.29*	29.08*	51.04*
1	2.23	0.12	45.12*	39.77*	39.71*
2	2.91	0.30	66.35*	16.99*	47.67*
3	3.52	0.14	77.79*	18.39*	38.46*
4	4.13	0.24	158.49*	36.90*	55.40*

Note. Panel A reports the MSFE, calculated as $\frac{1}{P} \sum_{t=1}^P (\pi_{t+h} - \pi_{t+h,t})^2$, for various forecast horizons, h . Panels B and C report the traditional test statistics, \mathcal{W}_P , (column labeled “Traditional”) and the test statistics $\max_{j \in \{m, \dots, P\}} \mathcal{W}_{j,m}$ (column labeled “Fluctuation”) for the null hypotheses of inflation forecast rationality and the Fed having no informational advantage over the private sector, respectively. The Fluctuation test results are based on $m = 60$ and the significance of the test statistics at the 5% significance level is indicated by asterisks.

Figures

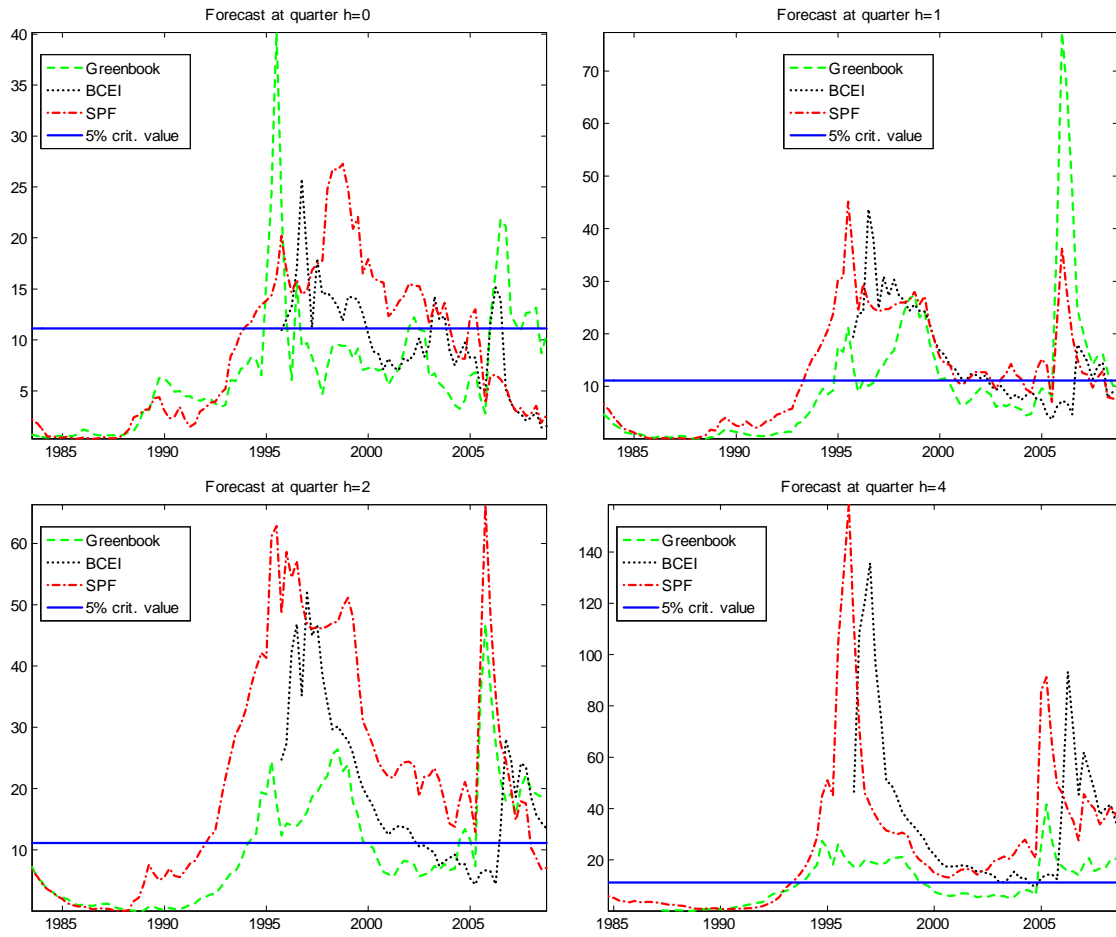
Figure 1: Inflation Forecasts

Inflation at quarter $h=2$



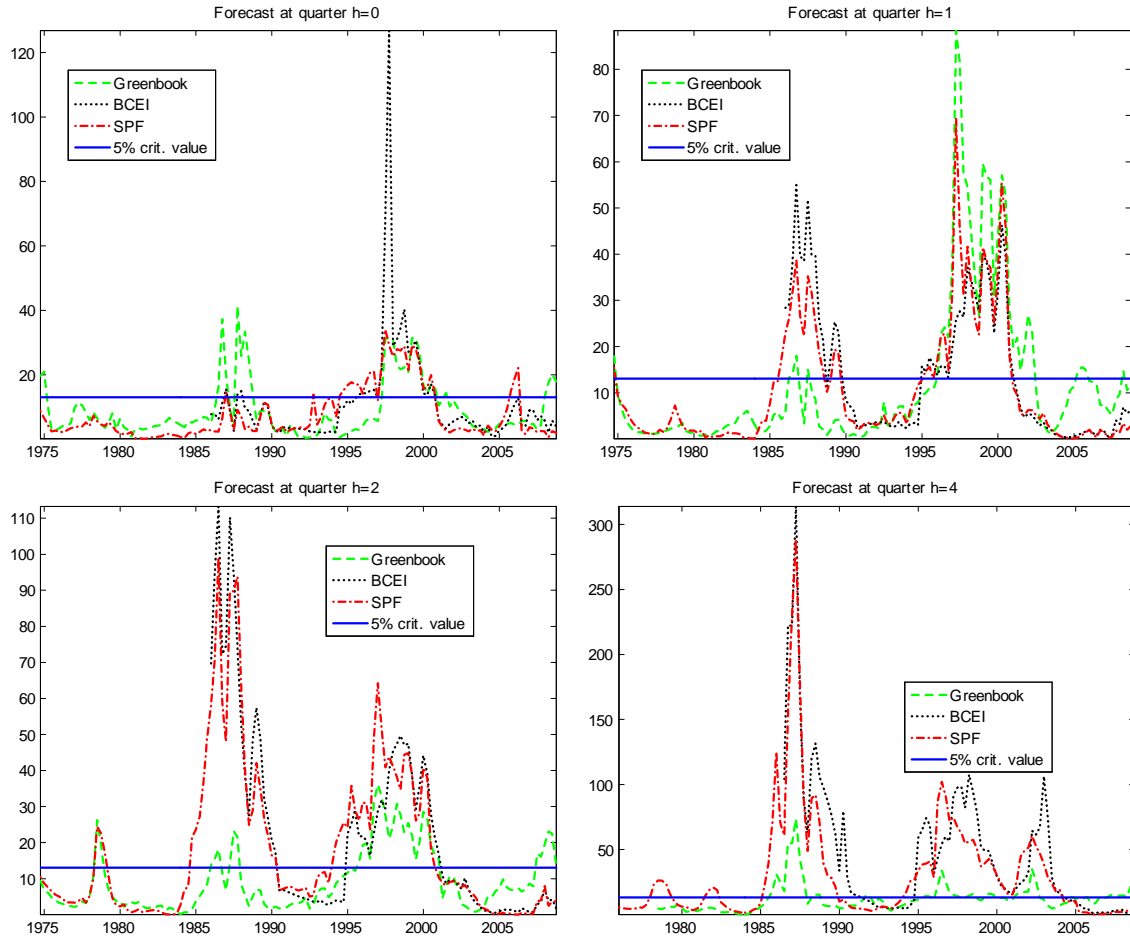
Note. The figure plots Greenbook, BCEI and SPF forecasts of inflation for two-quarter-ahead as well as the realized values of inflation, labeled as “actual,” for the same horizon. If a forecast at a particular time is not available, it is depicted as a missing value.

Figure 2: Fluctuation Rationality Test



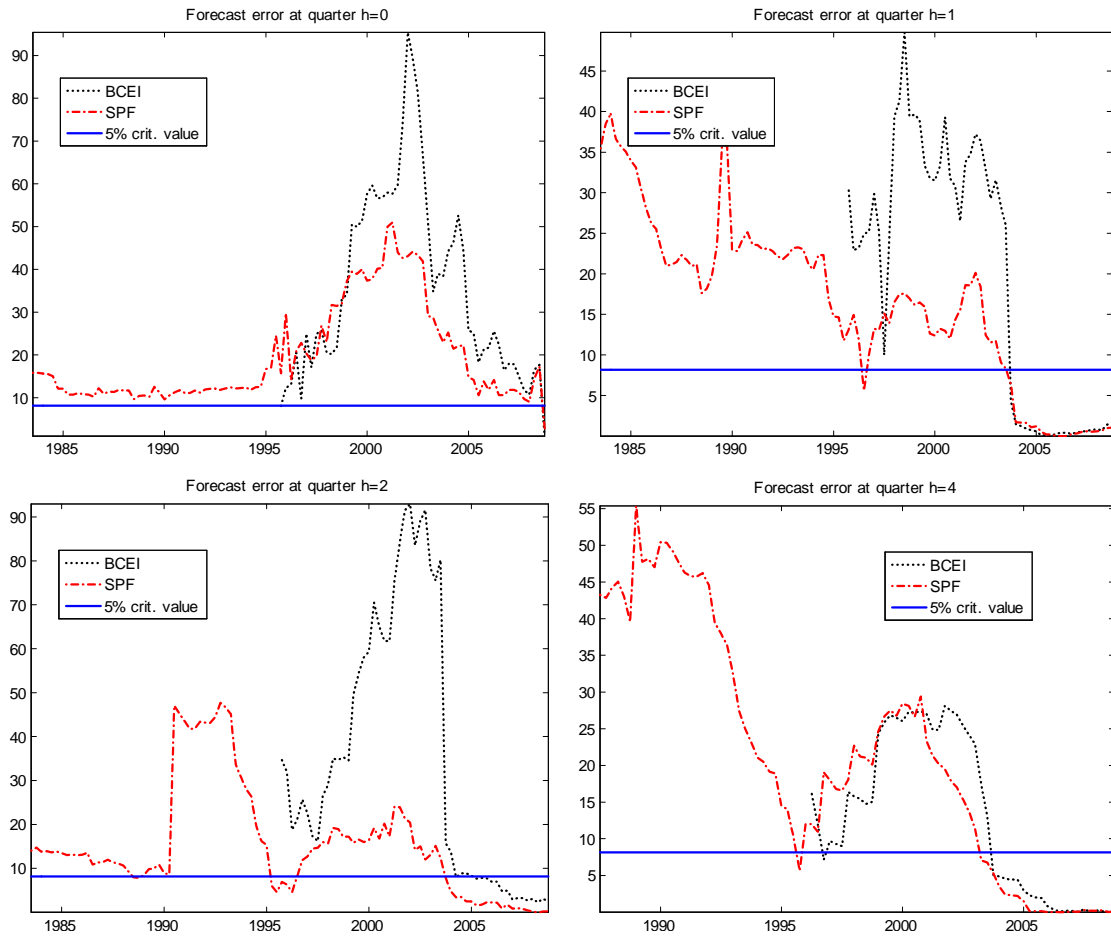
Note. The figure reports the time path of the test statistics $W_{j,m}$ for the null hypothesis of forecast rationality for Greenbook (dashed line), BCEI (dotted line) and SPF (dashed and dotted line), $m = 60$ and the solid line reports the critical value of the Fluctuation Rationality test at the 5% significance level. If the test statistic is above the solid line, we reject the null hypothesis of rationality at any point in time. The dates in the horizontal axis suggest a particular break-date.

Figure 3: Robustness to the Choice of m for Greenbook Forecasts



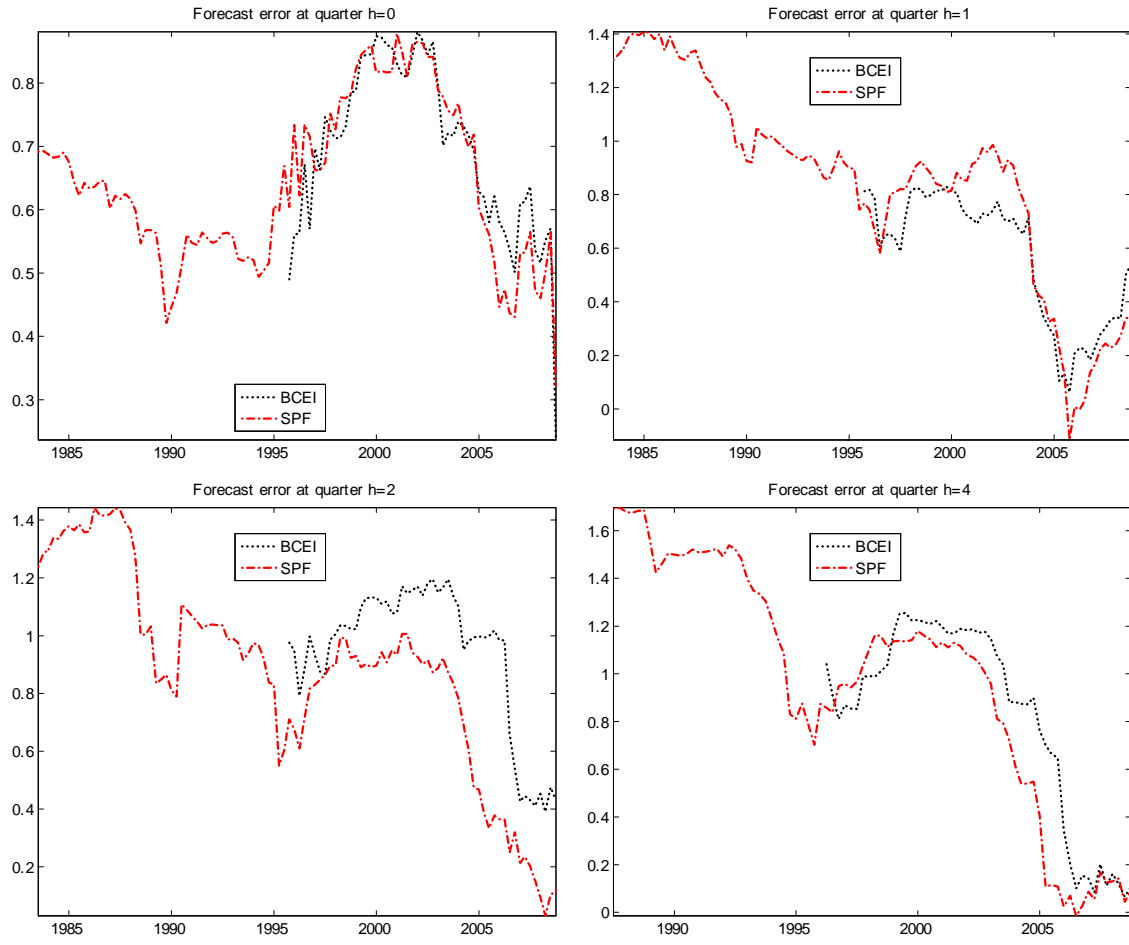
Note. The figure reports the time path of the test statistics $W_{j,m}$ for the null hypothesis of forecast rationality for Greenbook (dashed line), BCEI (dotted line) and SPF (dashed and dotted line), $m = 25$ and the solid line reports the critical value of the Fluctuation Rationality test at the 5% significance level. If the test statistic is above the solid line, we reject the null hypothesis of rationality at any point in time. The dates in the horizontal axis suggest a particular break-date.

Figure 4: Fed's Informational Advantage



Note. The figure reports the test statistics $W_{j,m}$ for the null hypothesis $\beta_g = 0$ based on $m = 60$.

Figure 5: Fed's Informational Coefficients



Note. The figure shows the rolling estimate of β_g based on $m = 60$.