JOURNAL OF APPLIED ECONOMETRICS
J. Appl. Econ. **25**: 595–620 (2010)
Published online 13 April 2010 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/jae.1177

FORECAST COMPARISONS IN UNSTABLE ENVIRONMENTS

RAFFAELLA GIACOMINI AND BARBARA ROSSI*
UCL/CEMMAP and Department of Economics, Duke University, Durham, NC, USA

SUMMARY

We propose new methods for comparing the out-of-sample forecasting performance of two competing models in the presence of possible instabilities. The main idea is to develop a measure of the relative local forecasting performance for the two models, and to investigate its stability over time by means of statistical tests. We propose two tests (the Fluctuation test and the One-Time Reversal test) that analyze the evolution of the models' relative performance over historical samples. In contrast to previous approaches to forecast comparison, which are based on measures of global performance, we focus on the entire time path of the models' relative performance, which may contain useful information that is lost when looking for the model that forecasts best on average. We apply our tests to the analysis of the time variation in the out-of-sample forecasting performance of monetary models of exchange rate determination relative to the random walk.

Received 30 March 2008; Revised 28 October 2008

1. INTRODUCTION

This paper proposes new techniques for comparing the out-of-sample forecasting performance of competing models in the presence of instabilities. The main insight of the paper is that, in unstable environments, it is plausible that the relative forecasting performance of models may itself change over time. Existing techniques for forecast comparison do not account for this possibility, in spite of the mounting empirical evidence (e.g., Stock and Watson, 2003a) suggesting instability in the forecasting performance of econometric models relative to naïve benchmarks.[1] For example, Stock and Watson (2003a) report that a model using housing price inflation as a predictor for consumer price inflation worked quite well in 1971–1984, but it performed significantly worse than an autoregressive model in 1985–1999 in the USA as well as in other countries. Similarly, the short-term interest rate helped predict inflation in France before 1984 but its forecasting ability disappears when considering the period 1985–1999. In short, the forecasting success of a model relative to a competitor seems to be linked to specific periods in time, and there are numerous situations in which there has been a reversal in the relative forecasting ability of two competing models.

Existing econometric techniques are inadequate for conducting forecast evaluation in an environment characterized by instability. In fact, it is common in forecasting to select the model with the best global forecasting performance, which in practice amounts to selecting the model that

* Correspondence to: Barbara Rossi, Department of Economics, 213 Social Science Building, Duke University, PO Box 90097, Durham, NC 27708, USA. E-mail: brossi@econ.duke.edu
[1] In the authors' words: 'Forecasts based on individual indicators are unstable. Finding an indicator that predicts well in one period is no guarantee that it will predict well in later periods. It appears that instability of predictive relations based on asset prices (like many other candidate leading indicators) is the norm' (Stock and Watson, 2003a, p. 789).

forecasts best on average over the in-sample period or over the (simulated) out-of-sample period (see, for example, Rissanen, 1986; Wei, 1992; Inoue and Kilian, 2006). The latter approach has also motivated the development of tests of overall predictive ability such as Diebold and Mariano (1995), West (1996), McCracken (2000), Clark and McCracken (2001), Clark and West (2006), and Giacomini and White (2006). In the presence of structural instability, however, the relative performance of the two models may itself be time-varying, and thus averaging this evolution over time will result in a loss of information. For example, a forecaster may select the model that performed best on average over a particular historical sample, ignoring the fact that the competing model produced more accurate forecasts when considering only the recent past.

This paper proposes two techniques that are useful for forecasters interested in analyzing the evolution in the performance of two competing forecasting models over historical samples. The first technique introduces a measure of the local relative forecasting performance of the models, and tests whether it equals zero at each point in time by means of an out-of-sample Fluctuation test. The test is easily implemented by plotting the (standardized) sample path of the relative measure of local performance, together with critical values which, if crossed, signal that one of the models outperformed its competitor at some point in time. The Fluctuation test, although convenient to obtain, does not, however, specify an alternative hypothesis and therefore might have lower power than a test designed for a specific alternative hypothesis. We thus further provide a test of the null hypothesis that the two models perform equally well at each point in time against the alternative that there is a one-time break in the relative performance, and propose a method for estimating the timing of the break. We call this the One-Time Reversal test.

We illustrate the usefulness of our techniques in the analysis of the out-of-sample forecasting performance of exchange rate models driven by economic fundamentals relative to a random walk benchmark. Since the seminal papers by Meese and Rogoff (1983a,b), it is well known that the random walk forecasts exchange rates better than any model with economic fundamentals, such as money, output, or interest rate differentials. As shown by Rossi (2006), the estimates of exchange rate models with economic fundamentals are plagued by parameter instabilities. Using in-sample Granger causality tests that are robust to parameter instability, she shows that it is possible to reject the null hypothesis of a random walk for selected countries and fundamentals. We examine the implications of this finding for forecasting exchange rates out-of-sample in unstable environments. We consider two models of exchange rate determination: the uncovered interest rate parity (UIRP) model and a model with Taylor rule fundamentals. We find widespread evidence that the relative forecasting performance has changed over time. The general pattern revealed by our methods is that the British pound and Deutsche Mark exchange rates were predictable in the late 1980s, but such predictability has disappeared in more recent years. We find that conventional out-of-sample forecast comparison tests (such as the test proposed by Clark and West (2006)) do find empirical evidence in favor of models with economic fundamentals for selected countries, as reported by Molodtsova and Papell (2007). However, we also find that the relative forecasting performance has changed over time. In fact, our procedures indicate that the Deutsche Mark and the British pound exchange rates were predictable in the late 1980s, but such predictability has disappeared in the 1990s. We show that conventional out-of-sample tests would have been unable to uncover such evidence in favor of models with economic fundamentals.

The paper is organized as follows. Section 2 discusses a simple example that motivates our procedure. Section 3 presents the econometric methodologies. Section 4 shows some Monte Carlo evidence on the performance of our procedures in small samples, and Section 5 presents the empirical results. Section 6 concludes.

DOI: 10.1002/jae

2. MOTIVATING EXAMPLE

Consider a researcher who is interested in assessing whether exchange rates are forecastable by using macroeconomic fundamentals. For example, UIRP implies that currencies' appreciations/depreciations should reflect interest rate differentials between countries. Therefore interest rate differentials should predict exchange rate changes. One might be interested in testing whether such a model provides any forecasting improvements relative to a simple random walk benchmark, according to which exchange rate changes are unpredictable.

Focusing on monthly data from 1973 : 3 to 2008 : 1 for the dollar/British pound exchange rate, for example, one would find that the square root of the out-of-sample mean square forecast error (MSFE) equals 0.0245 for the random walk and 0.0249 for the UIRP model. One would thus conclude that the UIRP model produces less accurate forecasts than the random walk, when considering the models' average performance over the whole out-of-sample period.[2]

However, the relative forecasting performance of the two models has changed considerably over the sample. Figure 1(a) depicts a sequence of differences between the MSFE of the random walk and the MSFE of the UIRP model computed over rolling windows of 50 observations. Each MSFE difference is rescaled by its standard deviation, to abstract from unit of measurement issues. Positive (negative) values of such differences indicate that the economic model produces better (worse) forecasts than the random walk. Interestingly, in the late 1980s, the UIRP model's forecasts were more accurate than the random walk forecasts. However, during the 1990s as well as in most recent years, the random walk produced consistently better forecasts than the UIRP model. That is why, when we consider the relative MSFE over the whole out-of-sample period, we find that the random walk is better on average: the negative MSFE differences observed during the 1990s more than offset the positive MSFE differences observed in the late 1980s. This highlights one of the most important points of this paper: looking at global (or average) relative forecasting performance may hide important information about the relative forecasting performance of the two models over time.

This paper proposes two techniques for extracting information about the time variation in the models' relative forecasting performance. The first involves measuring the models' local relative performance as the out-of-sample MSFE differences computed over rolling windows (the local relative MSFE). We provide critical values for testing the null hypothesis that the local relative MSFE equals zero at each point in time (rather than on average over the whole sample, which is the null hypothesis considered by previous literature, such as Diebold and Mariano, 1995, and West, 1996). We call this the Fluctuation test, which emphasizes the analogy between our procedure and the fluctuation tests for parameter stability proposed by Ploberger and Kramer (1992); see also Brown *et al*. (1975) and Chu *et al*. (1995). Figure 1(b) shows how to implement the Fluctuation test in the simple example considered in this section. It reports the (standardized) local relative MSFE for the UIRP and the random walk models, as well as the critical value for testing the null hypothesis that the two models have equal out-of-sample performance at each point in time, against the alternative that the UIRP performs better at least at one point in time. Since the local relative MSFE exceeds the critical value in the early part of the sample, we reject the null hypothesis, and conclude that there were periods during which the UIRP produced better forecasts than the random walk (from Figure 1(b), this seems to have occurred primarily in the late 1980s).

[2] These results are based on the actual empirical application of this paper. See Section 5 for more details. The standard deviation of the difference of the MSFE is such that one would not reject the null hypothesis that the two models have equal predictive ability.

DOI: 10.1002/jae

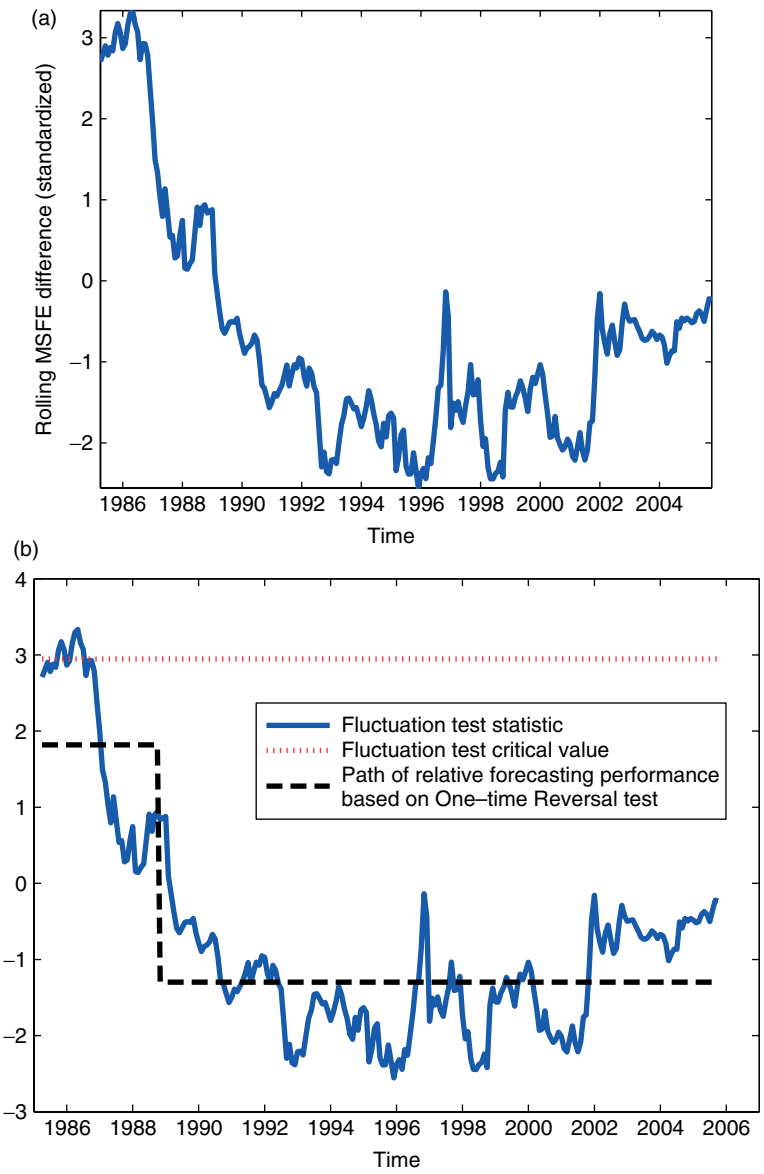

Figure 1. (a) Fluctuation test statistic, obtained as the (standardized) difference between the MSFE of the random walk and the MSFE of the UIRP model calculated over rolling windows (b) Fluctuation test statistic as well as one-sided critical value and the path of relative performance implied by the One-Time Reversal test. This figure is available in color online at www.interscience.wiley.com/journal/jae

The second technique that we propose is a test for the null hypothesis that the relative forecasting performance is equal at each point in time against the joint alternative that either one of the two forecasts was always better or that there was a reversal in the relative forecasting performance at one (unknown) point in time. We call this the One-Time Reversal test. When the test rejects

DOI: 10.1002/jae

the null hypothesis, our technique allows the researcher to estimate the time of the reversal. For the data discussed above, the One-Time Reversal test rejects the null hypothesis that the relative forecasting performance is equal at each point in time, and finds evidence of a reversal. The dashed line in Figure 1(b) shows the path of the estimated relative performance, suggesting that the reversal occurred around 1989.

3. ECONOMETRIC METHODOLOGY

3.1. Notation and Definitions

We first introduce the notation and discuss the assumptions about the data, the models and the estimation procedures. We are interested in comparing two h-step-ahead forecasts for the variable y_t, which we assume for simplicity to be a scalar. The first model is characterized by parameters θ and the second model by parameters γ.

We assume that the researcher has divided the sample of size T into an in-sample portion of size R and an out-of-sample portion of size P, and obtained two competing sequences of h-step-ahead out-of-sample forecasts. For a general loss function L, we thus have a sequence of P out-of-sample forecast loss differences, $\{\Delta L_t(\widehat{\theta}_{t-h,R}, \widehat{\gamma}_{t-h,R})\}_{t=R+h}^{T} \equiv \{L^{(1)}(y_t, \widehat{\theta}_{t-h,R}) - L^{(2)}(y_t, \widehat{\gamma}_{t-h,R})\}_{t=R+h}^{T}$, which depend on the realizations of the variable and on the in-sample parameter estimates for each model, $\widehat{\theta}_{t-h,R}$ and $\widehat{\gamma}_{t-h,R}$. These parameters are estimated only once, using a sample including data indexed $1, \ldots, R$ (fixed scheme) or re-estimated at each $t = R + h, \ldots, T$ over a window of R data including data indexed $t - h - R + 1, \ldots, t - h$ (rolling scheme).

We define the local relative loss for the two models as the sequence of out-of-sample loss differences computed over centered rolling windows of size m (without loss of generality, we assume m to be an even number):

$$m^{-1} \sum_{j=t-m/2}^{t+m/2-1} \Delta L_j(\widehat{\theta}_{j-h,R}, \widehat{\gamma}_{j-h,R}), \quad t = R + h + m/2, \ldots, T - m/2 + 1$$

3.2. The Fluctuation Test

We make the following assumptions.

Assumption 1 *Let* $\tau \in [0, 1]$.

(a) $\left\{P^{-1/2} \sum_{t=R+h}^{R+h+[\tau P]} \Delta L_t(\widehat{\theta}_{t-h,R}, \widehat{\gamma}_{t-h,R})\right\}$ obeys a Functional Central Limit theorem;
(b) $\sigma^2 = \lim_{P\to\infty} E\left(P^{-1/2} \sum_{t=R+h}^{T} \Delta L_t(\widehat{\theta}_{t-h,R}, \widehat{\gamma}_{t-h,R})\right)^2 > 0$;
(c) $m/P \to \mu \in (0, \infty)$ as $m \to \infty$, $P \to \infty$, whereas $R < \infty$, $h < \infty$.

Note that we do not impose restrictions on the estimation methods used to produce the forecasts for the two models. This is because we adopt the same asymptotic framework as Giacomini and White (2006), which allows the competing models to be nested or non-nested and estimated by a general estimation procedure. The only requirement is the use of a rolling or fixed estimation

DOI: 10.1002/jae

window scheme in producing the out-of-sample forecasts. Giacomini and White (2006) also provide primitive conditions for Assumption 1(a), which allow the data to be mixing and heterogeneous.

Proposition 1 describes the procedure for deriving the out-of-sample Fluctuation test.

Proposition 1 Fluctuation test *Suppose Assumption 1 holds. Let*

$$F_{t,m}^{\mathrm{OOS}} = \widehat{\sigma}^{-1} m^{-1/2} \sum_{j=t-m/2}^{t+m/2-1} \Delta L_j(\widehat{\theta}_{j-h,R}, \widehat{\gamma}_{j-h,R}), \tag{1}$$

$t = R + h + m/2, \ldots, T - m/2 + 1$, where $\widehat{\sigma}^2$ is a HAC estimator of σ^2; for example:

$$\widehat{\sigma}^2 = \sum_{i=-q(P)+1}^{q(P)-1} (1 - |i/q(P)|) P^{-1} \sum_{j=R+h}^{T} \Delta L_j(\widehat{\theta}_{j-h,R}, \widehat{\gamma}_{j-h,R}) \Delta L_{j-i}(\widehat{\theta}_{j-i-h,R}, \widehat{\gamma}_{j-i-h,R}) \tag{2}$$

and $q(P)$ is a bandwidth that grows with P (e.g., Newey and West, 1987). Under the null hypothesis $H_0 : E[\Delta L_t(\widehat{\theta}_{t-h,R}, \widehat{\gamma}_{t-h,R})] = 0$ for all $t = R + h, \ldots, T$:

$$F_{t,m}^{\mathrm{OOS}} \Longrightarrow [\mathcal{B}(\tau + \mu/2) - \mathcal{B}(\tau - \mu/2)]/\sqrt{\mu} \tag{3}$$

where $t = [\tau P]$, $m = [\mu P]$ and $\mathcal{B}(\cdot)$ is a standard univariate Brownian motion. The critical values for a significance level α are $\pm\, k_\alpha$, where k_α solves

$$\Pr\{\sup_{\tau} |[\mathcal{B}(\tau + \mu/2) - \mathcal{B}(\tau - \mu/2)]/\sqrt{\mu}| > k_\alpha\} = \alpha \tag{4}$$

The null hypothesis is rejected against the two-sided alternative $E[\Delta L_t(\widehat{\theta}_{t-h,R}, \widehat{\gamma}_{t-h,R})] \neq 0$ when $\max_t |F_{t,m}^{\mathrm{OOS}}| > k_\alpha$.

Critical values for testing H_0 against the one-sided alternative $E[\Delta L_t(\widehat{\theta}_{t-h,R}, \widehat{\gamma}_{t-h,R})] > 0$ can be similarly obtained as a solution to $\Pr\{\sup_\tau [\mathcal{B}(\tau + \mu/2) - \mathcal{B}(\tau - \mu/2)]/\sqrt{\mu} > k_\alpha\} = \alpha$, in which case the null is rejected when $\max_t F_{t,m}^{\mathrm{OOS}} > k_\alpha$. Simulated values of (α, k_α) for both the one-sided and the two-sided case are reported in Table I for various choices of μ.

The test statistic $F_{t,m}^{\mathrm{OOS}}$ in (1) is equivalent to Diebold and Mariano's (1995) and Giacomini and White's (2006) (unconditional) test statistic, computed over rolling out-of-sample windows of size m. Similar reasonings to those in the proof of Proposition 1 can be used to show that any other test statistic commonly used for out-of-sample predictive ability testing could be used in (1), as long as its asymptotic distribution is normal. In particular, one could substitute $F_{t,m}^{\mathrm{OOS}}$ with the test statistics proposed by West (1996) or by Clark and West (2006), which are respectively applicable to non-nested and nested models. The fundamental differences in the two approaches is that they test two different null hypotheses: the null hypothesis in West (1996) and Clark and West (2006) concerns forecast losses that are evaluated at the population parameters, whereas in Giacomini and White (2006) the losses depend on estimated in-sample parameters. This reflects the different focus of the two approaches on comparing forecasting models (West, 1996, and Clark and West, 2006) versus comparing forecasting methods (Giacomini and White, 2006). The adoption of West's

DOI: 10.1002/jae

Table I. Asymptotic critical values for the Fluctuation test (k_α)

μ	Two-sided test α		One-sided test α	
	0.05	0.10	0.05	0.10
0.1	3.393	3.170	3.176	2.928
0.2	3.179	2.948	2.938	2.676
0.3	3.012	2.766	2.770	2.482
0.4	2.890	2.626	2.624	2.334
0.5	2.779	2.500	2.475	2.168
0.6	2.634	2.356	2.352	2.030
0.7	2.560	2.252	2.248	1.904
0.8	2.433	2.130	2.080	1.740
0.9	2.248	1.950	1.975	1.600

Note: The table reports critical values for the Fluctuation test of Proposition 1. α denotes the nominal size of the test and $\mu = m/P$, where m denotes the size of the rolling window and P the out-of-sample size.

(1996) asymptotic framework would involve replacing $\widehat{\sigma}$ in (2) with an estimator of the asymptotic variance that reflects the contribution of estimation uncertainty (see Theorem 4.1 of West, 1996). Also note that West's (1996) approach allows the parameters to be estimated using a recursive scheme, in addition to a rolling or fixed scheme. For the nested case, the use of the Clark and West (2006) test statistic instead of (1) in practice amounts to replacing $\Delta L_j(\widehat{\theta}_{j-h,R}, \widehat{\gamma}_{j-h,R})$ in (1) with its corrected version (see their equation (3.1)).

Algorithm 1 Clark and West (2006) and West (1996) Fluctuation test

I. Rolling window case Let $W_{t,m}^{\mathrm{OOS}}$ denote a sequence of either West's (1996) test statistic (Theorem 4.1) (for non-nested models) or the statistic in equation (3.1) of Clark and West (2006) (for nested models). Both statistics are for h-step-ahead forecasts computed over rolling windows of size m and centered at time t (that is, on observations $t - m/2, \ldots, t + m/2 - 1$), for $t = R + h + m/2, \ldots, T - m/2 + 1$.

(a) For West's (1996) test the null hypothesis is rejected when $\max_t |m^{1/2} W_{t,m}^{\mathrm{OOS}}| > k_\alpha$ (using two-sided critical values from Table I).
(b) For Clark and West's (2006) test the null hypothesis is rejected when $\max_t m^{1/2} W_{t,m}^{\mathrm{OOS}} > k_\alpha$ (using one-sided critical values from Table I).

II. Recursive window case Let W_t^{OOS} denote a sequence of West's (1996) test statistics for h-step-ahead forecasts calculated over recursive windows (with an initial window of size R) for $t = R + h + m/2, \ldots, T - m/2 + 1$. The null hypothesis is rejected when $\max_t |W_t^{\mathrm{OOS}}| > k_\alpha^{rec} \sqrt{\frac{T-R}{t}} \left(1 + 2\frac{t-R}{T-R}\right)$, where (α, k_α^{rec}) are (0.01,1.143), (0.05,0.948) and (0.10,0.850).[3]

[3] The proofs follow from an argument similar to that of Proposition 1 and are therefore omitted. The critical values for the recursive window case follow from Brown *et al*. (1975).

DOI: 10.1002/jae

3.3. The One-Time Reversal Test

The assumptions that guarantee validity of the test against a one-time reversal in the forecasting performance are the same as those for the Fluctuation test. The following proposition gives the justification for this test.

Proposition 2 One-Time Reversal test *Suppose Assumption 1 holds. Let* $QLR_P^* = \sup_t \Phi_P^*(t)$, $t \in \{[0.15P], \ldots [0.85P]\}$, $\Phi_P^*(t) = LM_1 + LM_2(t)$, *where*

$$LM_1 = \widehat{\sigma}^{-2} P^{-1} \left[\sum_{j=R+h}^{T} \Delta L_j(\widehat{\theta}_{j-h,R}, \widehat{\gamma}_{j-h,R}) \right]^2$$

$$LM_2(t) = \widehat{\sigma}^{-2} P^{-1} (t/P)^{-1} (1 - t/P)^{-1} \left[\sum_{j=R+h}^{t} \Delta L_j(\widehat{\theta}_{j-h,R}, \widehat{\gamma}_{j-h,R}) - (t/P) \sum_{j=R+h}^{T} \Delta L_j(\widehat{\theta}_{j-h,R}, \widehat{\gamma}_{j-h,R}) \right]^2,$$

$\widehat{\sigma}^2$ is a HAC estimator of the asymptotic variance $\sigma^2 = \text{var}\left(P^{-1/2} \sum_{j=R+h}^{T} \Delta L_j(\widehat{\theta}_{j-h,R}, \widehat{\gamma}_{j-h,R})\right)$, for example:

$$\widehat{\sigma}^2 = \sum_{i=-q(P)+1}^{q(P)-1} (1 - |i/q(p)|) P^{-1} \sum_{j=R+h}^{T} \Delta L_j(\widehat{\theta}_{j-h,R}, \widehat{\gamma}_{j-h,R}) \Delta L_{j-i}(\widehat{\theta}_{j-i-h,R}, \widehat{\gamma}_{j-i-h,R}) \tag{5}$$

Consider the null hypothesis:

$$H_0 : E[\Delta L_t(\widehat{\theta}_{t-h,R}, \widehat{\gamma}_{t-h,R})] = 0$$

for every $t = R + h, \ldots, T$. We have $QLR_P^* \Longrightarrow \sup_\tau \left[\frac{\mathcal{BB}(\tau)^2}{\tau(1-\tau)} + \mathcal{B}(1)^2 \right]$, where $t = [\tau P]$, and $\mathcal{B}(\cdot)$ and $\mathcal{BB}(\cdot)$ are, respectively, a standard univariate Brownian motion and a Brownian bridge. The null hypothesis is rejected when $QLR_P^* > k_\alpha$. The critical values (α, k_α) are: (0.05,9.8257), (0.10,8.1379), (0.01,13.4811).[4]

The intuition behind this test is that it jointly tests whether the relative forecasting performance is stable over time and equal to zero. It can be thought of as a test of globally equal forecasting ability that detects situations in which there is a one-time reversal. This approach is reminiscent of that in Rossi (2005b), who proposed optimal tests for these joint hypotheses when comparing the in-sample relative performance of two nested models. The results in this paper are different because

[4] The test against a one-time reversal is implemented with trimming values 0.15 and 0.85. Such trimming values are a conventional choice for the implementation of Andrews' (1993) test (cf. Stock and Watson, 2003b).

DOI: 10.1002/jae

we focus on the relative out-of-sample forecasting performance of either nested or non-nested models.

Among the advantages of this approach, we have that: (i) when the null hypothesis is rejected, it is possible to evaluate whether the rejection is due to instabilities in the relative performance or to a model being constantly better than its competitor; (ii) if such instability is found, it is possible to estimate the time of the reversal in the relative performance; (iii) the test is designed to have power against one-time breaks in the relative performance. This is achieved by using the following procedure for a test with overall significance level α:

(i) Test the hypothesis of equal performance at each time by using the statistic QLR_P^* from Proposition 2 for a significance level α.
(ii) If the null is rejected, compare LM_1 and $\sup_t LM_2(t)$, $t \in \{[0.15P], \ldots [0.85P]\}$, with the following critical values: (3.84,8.85) for $\alpha = 0.05$, (2.71,7.17) for $\alpha = 0.10$, and (6.63,12.35) for $a = 0.01$. If only LM_1 rejects then there is evidence in favor of the hypothesis that one model is constantly better than its competitor. If only LM_2 rejects, then there is evidence that there are instabilities in the relative performance of the two models but neither is constantly better over the full sample. If both reject, then it is not possible to attribute the rejection to a unique source.[5]
(iii) Estimate the time of the break by $t^* = \arg\max_{t \in \{0.15P, \ldots, 0.85P\}} LM_2(t)$.
(iv) To extract information on which model to choose, we suggest plotting the time path of the underlying relative performance as

$$\begin{cases} \frac{1}{t^*} \sum_{j=R+h}^{t^*} \Delta L_j(\widehat{\theta}_{j-h,R}, \widehat{\gamma}_{j-h,R}), & \text{for } t \leq t^* \\ \frac{1}{(P - t^*)} \sum_{j=t^*+1}^{T} \Delta L_j(\widehat{\theta}_{j-h,R}, \widehat{\gamma}_{j-h,R}), & \text{for } t > t^* \end{cases}$$

The estimator of the timing of the break in point (iii) above is analogous to the estimator proposed by Bai (1997) for estimating the timing of a break in the unconditional mean of a variable y_t, where in our case $y_t = \Delta L_t(\widehat{\theta}_{t-h,R}, \widehat{\gamma}_{t-h,R})$. Using similar reasonings to those in Giacomini and White (2006), it is easy to show that this choice of y_t satisfies Bai's (1997) assumptions. For example, if the data are mixing, $\widehat{\theta}_{t-h,R}$ and $\widehat{\gamma}_{t-h,R}$ are also mixing because they are measurable functions of the finite (because R is kept fixed) history of a mixing process, and thus y_t satisfies Assumption A6(b) of Bai (1997). By the same reasoning, it is also easy to see how the One-Time Reversal test could be generalized to detect multiple changes in relative performance by following, for example, the sequential procedure suggested by Bai and Perron (1998).

The Fluctuation test and the One-Time Reversal test have trade-offs. If the researcher is willing to specify the alternative of interest (in this case, a one-time break in the relative performance), then the latter test can be implemented. Furthermore, it allows the researcher to estimate the time of the break. The Fluctuation test, on the other hand, does not require the researcher to specify an alternative, and therefore might be preferable for researchers who do not have one. The two tests also have trade-offs in terms of their power. For example, if there are multiple breaks, the

[5] This procedure is justified by the fact that the two components LM_1 and LM_2 are asymptotically independent—see Rossi (2005b). Performing two separate tests does not result in a test with equal power against all deviations from the null hypothesis, but it is nevertheless useful to heuristically disentangle the causes of rejection of the null hypothesis of equal performance at each point in time. The critical values for LM_1 are from a chi-square with one degree of freedom, whereas those for LM_2 are from Andrews (1993).

DOI: 10.1002/jae

Fluctuation test should reveal their presence, whereas the One-Time Reversal test would only identify the largest break. The extension of the latter test to the case of multiple breaks requires the researcher to determine the number of breaks under the alternative hypothesis, in which case the test can be expected to have greater power than the Fluctuation test. Overall, one can view the Fluctuation test as a 'robust' method, but its robustness may, however, come at the cost of possible power losses.

4. MONTE CARLO EVIDENCE

In this section, we analyze the size and power properties of the Fluctuation and One-Time Reversal tests, relative to the full-sample Giacomini and White (2006) (henceforth GW) and the full-sample Clark and West (2006) (henceforth CW) tests, which focus on average performance over the whole out-of-sample period. Our goal is threefold: first, to understand whether our tests have comparable size to the GW and CW tests when the competing models have equal performance over time; second, to compare the power properties of the tests when the relative performance is not equal but is constant over time, and to illustrate situations in which the Fluctuation and One-Time Reversal tests, unlike the GW and CW tests, have the ability to detect time variation in relative performance; third, to investigate how the size and power of the Fluctuation test depend on the choice of the parameters R (in-sample size), P (out-of-sample size) and μ ($= m/P$, with m the size of the rolling window used to construct the test statistics).

4.1. Size Properties

Suppose the data-generating process (DGP) is

$$\begin{aligned} Y_t &= \beta_t X_t + \varepsilon_t, \\ X_t &= .5X_{t-1} + \nu_t, \\ \nu_t, \varepsilon_t &\sim \text{i.i.d. } N(0, 1), \text{ independent of each other} \end{aligned} \tag{6}$$

We compare one-step-ahead out-of-sample forecasts from the model (6), estimated in-sample under the assumption that β is constant, with forecasts from a model that assumes Y_t to be a zero-mean white noise. This setup is meant to represent a plausible comparison between a fundamental-based model for the exchange rate and a random walk benchmark, which is the case considered in this paper's empirical application (Y_t should be interpreted as the log-exchange rate first differences). The time-t forecasts implied by the two models are

$$\begin{aligned} f_{t,R}^{(1)} &= \widehat{\beta}_{t,R} X_{t+1} \text{ and} \\ f_{t,R}^{(2)} &= 0 \end{aligned}$$

where $\widehat{\beta}_{t,R}$ is the in-sample parameter estimate based on a rolling window of size R, and where we assume for simplicity that X_{t+1} is known at time t.

We analyze the size properties of the Fluctuation test in both the GW and the CW frameworks, which amount to imposing different null hypotheses. In the GW case, it is easy to show that values

DOI: 10.1002/jae

of β_t that satisfy the null hypothesis $H_0 : E[(Y_{t+1} - f_{t,R}^{(1)})^2] = E[(Y_{t+1} - f_{t,R}^{(2)})^2]$ can be obtained by setting[6]

$$\beta_{t+1} = \frac{\dfrac{\left(\sum_{j=t-R+1}^{t} \beta_j X_j^2\right)^2}{\sum_{j=t-R+1}^{t} X_j^2} + \sigma^2}{2\sum_{j=t-R+1}^{t} \beta_j X_j^2}, \quad t = R, \ldots, T-1. \tag{7}$$

Note that in this situation the two models have equal relative performance at each point in time in spite of the fact that the DGP parameters are time-varying. We first generate a time series X_t as in (6), and initialize the time series of β_t by letting $\beta_t = 0.05$ for $t = 1, \ldots, R$. For each pair of in-sample and out-of-sample sizes (R, P) with $R, P = 20, 50, 150$, we generate $T = R + P$ observations for Y_t that satisfy equations (6) and (7). We then implement the Fluctuation test using $\mu = m/P = 0.1, 0.3, 0.7, 0.9$, where m is the window size, and the One-Time Reversal test.

In the CW case, the null hypothesis is $H_0 : \beta_t = 0$ for all t, and therefore we generate data that satisfy the null hypothesis by letting $Y_t = \varepsilon_t$.

The rejection frequencies over 5000 Monte Carlo replications are contained in Table II.

The GW Fluctuation test has a mild tendency to under-reject for most values of μ, with the exception of $\mu = 0.1$, in which case the test is oversized. The One-Time Reversal test is slightly undersized when R is small relative to P, but correctly sized when P and R are similar. All tests perform best when the in-sample and out-of-sample sizes are of comparable magnitude. The CW Fluctuation test has no size distortions for samples that are sufficiently large, but exhibits considerable size distortions in small samples when $\mu = 0.1$ (which is due to the estimation window being too small for the normal approximation to be accurate).

4.2. Power Properties

We now investigate the power of the three tests above in relation to the full-sample tests. We consider two scenarios. In the first, the performance of the two models is not equal but is constant over time, whereas in the second scenario there is time variation in the relative performance. In all cases the power curves are obtained over 5000 Monte Carlo replications. Both the Fluctuation and the full-sample tests are derived in either the GW or the CW framework, which correspond to two different Monte Carlo designs.

[6] This is obtained by first showing that $E[(Y_{t+1} - f_{t,R}^{(1)})^2] = \left(\beta_{t+1} - \dfrac{\sum_{j=t-R+1}^{t} \beta_j X_j^2}{\sum_{j=t-R+1}^{t} X_j^2}\right)^2 X_{t+1}^2 + \dfrac{\sigma^2 X_{t+1}^2}{\sum_{j=t-R+1}^{t} X_j^2} + \sigma^2$, and $E[(Y_{t+1} - f_{t,R}^{(2)})^2] = \beta_{t+1}^2 X_{t+1}^2 + \sigma^2$, setting the two expressions equal to each other, and then solving for β_{t+1}.

DOI: 10.1002/jae

Table II. Empirical size of Fluctuation and One-Time Reversal tests: nominal size 0.05

A. Fluctuation test

		GW Fluctuation P				CW Fluctuation P		
	R	20	50	150	R	20	50	150
$\mu = 0.1$	20	0.18	0.15	0.13	20	0.97	0.47	0.11
	50	0.17	0.17	0.17	50	0.98	0.47	0.11
	150	0.16	0.16	0.19	150	0.98	0.49	0.13
		GW Fluctuation P				CW Fluctuation P		
	R	20	50	150	R	20	50	150
$\mu = 0.3$	20	0.04	0.04	0.04	20	0.19	0.08	0.05
	50	0.04	0.04	0.04	50	0.20	0.07	0.05
	150	0.04	0.05	0.06	150	0.21	0.08	0.06
		GW Fluctuation P				CW Fluctuation P		
	R	20	50	150	R	20	50	150
$\mu = 0.5$	20	0.03	0.03	0.02	20	0.08	0.05	0.05
	50	0.03	0.03	0.02	50	0.08	0.05	0.04
	150	0.04	0.04	0.04	150	0.09	0.05	0.05
		GW Fluctuation P				CW Fluctuation P		
	R	20	50	150	R	20	50	150
$\mu = 0.7$	20	0.03	0.02	0.02	20	0.05	0.04	0.05
	50	0.03	0.03	0.02	50	0.06	0.05	0.05
	150	0.03	0.03	0.03	150	0.06	0.05	0.05
		GW Fluctuation P				CW Fluctuation P		
	R	20	50	150	R	20	50	150
$\mu = 0.9$	20	0.03	0.02	0.02	20	0.05	0.04	0.05
	50	0.03	0.03	0.03	50	0.05	0.05	0.05
	150	0.04	0.04	0.04	150	0.05	0.05	0.05

B. One-Time Reversal test

	P		
R	20	50	150
20	0.04	0.04	0.04
50	0.04	0.05	0.05
150	0.03	0.05	0.07

The table reports empirical rejection frequencies for the GW and CW Fluctuation tests (for various values of $\mu = m/P$, where m denotes the size of the rolling window used to construct the Fluctuation test statistic and P the out-of-sample size) and for the One-Time Reversal test. R denotes the in-sample size. The DGP is described in Section 4.1.

DOI: 10.1002/jae

4.2.1 Unequal but Constant Performance

For the GW Fluctuation test, we generate data under the alternative hypothesis by following the procedure explained in Section 4.1 for $(R, P) = (150, 150)$ and by letting σ^2 decrease from its value of 1 under the null hypothesis to $\sigma^2 = 0.1$. The effect of this is a reduction in the variance of the parameter estimate $\widehat{\beta}_{t,R}$, which results in a more accurate forecast for the larger model. Figure 2(a) shows the power curves for the GW, One-Time Reversal and GW Fluctuation tests (the latter for $\mu = 0.3$ in the top panel and $\mu = 0.7$ in the bottom panel).

For the CW Fluctuation test, we generate data under the alternative hypothesis by letting β_t be constant over the sample but with values increasing from 0 to 1. Figure 2(d) reports the results for this case.

Figure 2(a) shows that the GW Fluctuation and One-Time Reversal tests have lower power than the GW test when the relative performance is constant over time, but that the power loss for the Fluctuation test is smaller for larger μ. Figure 2(d) shows that similar conclusions hold for the CW Fluctuation test relative to the full-sample CW test.

4.2.2 Time-Varying Relative Performance

We consider the situation in which there is one break in the relative performance of the two models during the out-of-sample period, induced by a break in the DGP parameter. For both the GW Fluctuation and the CW Fluctuation tests, we generate the data as

$$Y_t = -\delta X_t \cdot 1(t \leq R + \tau P) + \delta X_t \cdot 1(t > R + (1 - \tau)P) + \varepsilon_t, \ \ \varepsilon_t \sim \text{i.i.d. } N(0, 1)$$

where X_t is as in (6) and $(R, P) = (150, 150)$. In this situation, the relative performance of the models changes at $t = R + \tau P$. We consider $\tau = 1/3$ and $\tau = 2/3$ for the GW case and $\tau = 1/2$ for the CW case. These parameter choices ensure that the two models perform equally well on average over the out-of-sample period. We obtain power curves for the various tests by letting the size of the break increase from $\delta = 0$ to $\delta = 1$ in increments of 0.05. Figure 2(b) shows the power curves for the full-sample GW, the One-Time Reversal and the GW Fluctuation test (the latter for $\mu = 0.3$, in the top panel and $\mu = 0.7$ in the bottom panel) when the break occurs at $\tau = 1/3$; Figure 2(c) shows the power curves for a break occurring at $\tau = 2/3$; Figure 2(e) shows the power curves for the CW Fluctuation test and the full-sample CW test.

The power curves bear out the prediction that the Fluctuation and One-Time Reversal tests are able to detect the change in relative performance for the two models, whereas the full-sample tests may incorrectly conclude that the models are equally accurate, regardless of the magnitude of the break. The Fluctuation test has higher power than the One-Time Reversal test for small values of μ and for breaks occurring towards the beginning of the out-of-sample period. Note that the power of the Fluctuation test diminishes (and converges to that of the full-sample GW test) as μ increases. Figure 2(e) similarly shows that the Fluctuation test implemented with the CW test statistic has power to detect changes in the relative performance, whereas the full-sample CW test may have no power at all.

4.3. Summary of Monte Carlo Results

The simulation results suggest that the Fluctuation test has good size and power properties when implemented using a rolling window size that is a small—but not too small—fraction of the

DOI: 10.1002/jae

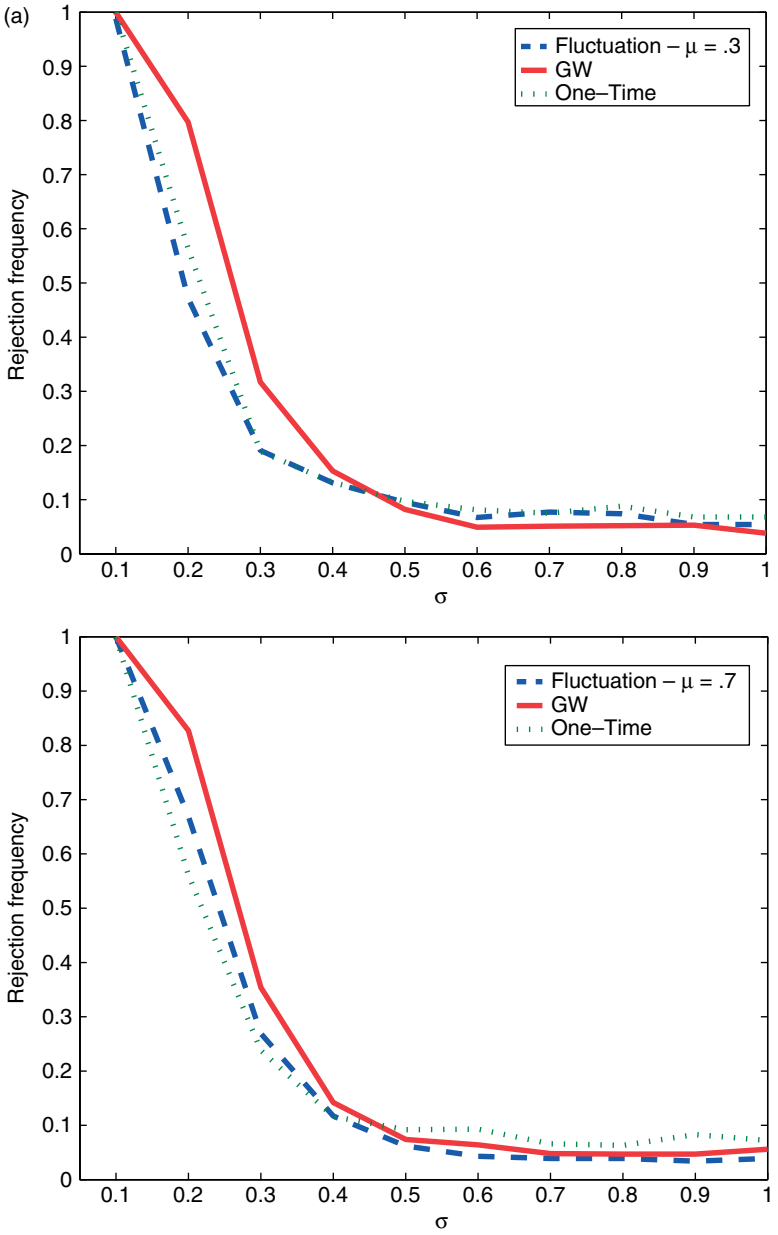

Figure 2(a). Power of Fluctuation, full-sample GW and One-Time Reversal tests. Unequal but constant relative performance. This figure is available in color online at www.interscience.wiley.com/journal/jae

out-of-sample size (e.g., $\mu = m/P = 0.3$). In such cases, the test has comparable properties to the full-sample tests when the two models perform equally well, it involves a relatively small loss of power relative to the full-sample test when the relative performance is unequal but constant over

DOI: 10.1002/jae

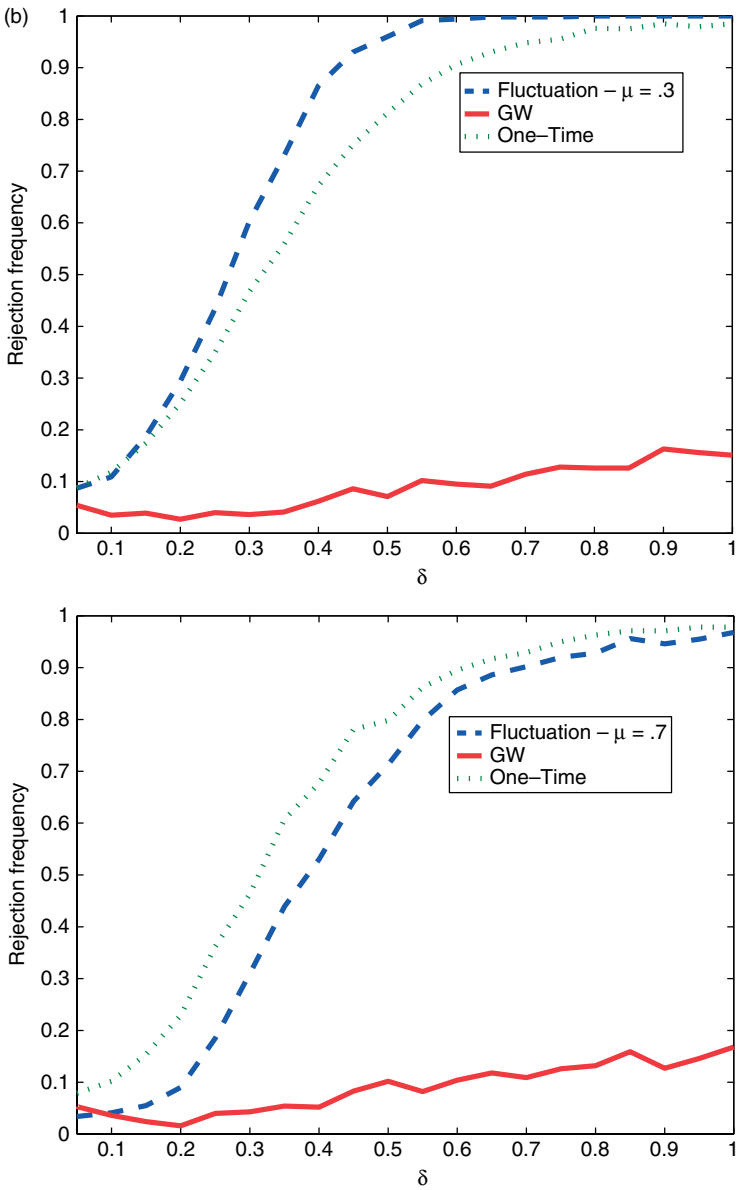

Figure 2(b). Power of Fluctuation, full-sample GW and One-Time Reversal tests. Break in relative performance at $R + \frac{1}{3}P$. This figure is available in color online at www.interscience.wiley.com/journal/jae

time, and is able to detect time variation in relative performance, whereas the full-sample test may incorrectly conclude that the models are equally accurate.

The One-Time Reversal test can also detect time variation in relative performance. It has comparable power to that of the Fluctuation test against the alternative of unequal but constant

DOI: 10.1002/jae

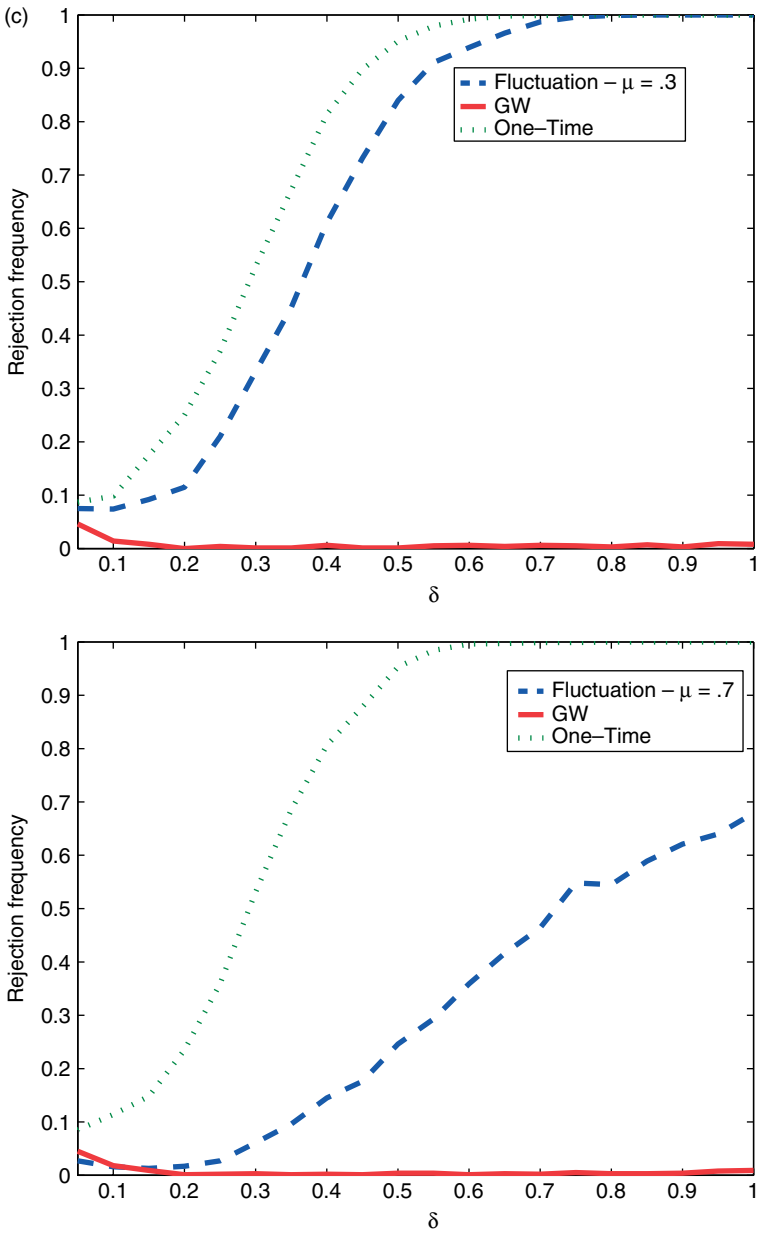

Figure 2(c). Power of Fluctuation, full-sample GW and One-Time Reversal tests. Break in relative performance at $R + \frac{2}{3}P$. This figure is available in color online at www.interscience.wiley.com/journal/jae

relative performance. It has higher power than the Fluctuation test when the latter is implemented using large values of μ and when the break occurs towards the end of the sample.

DOI: 10.1002/jae

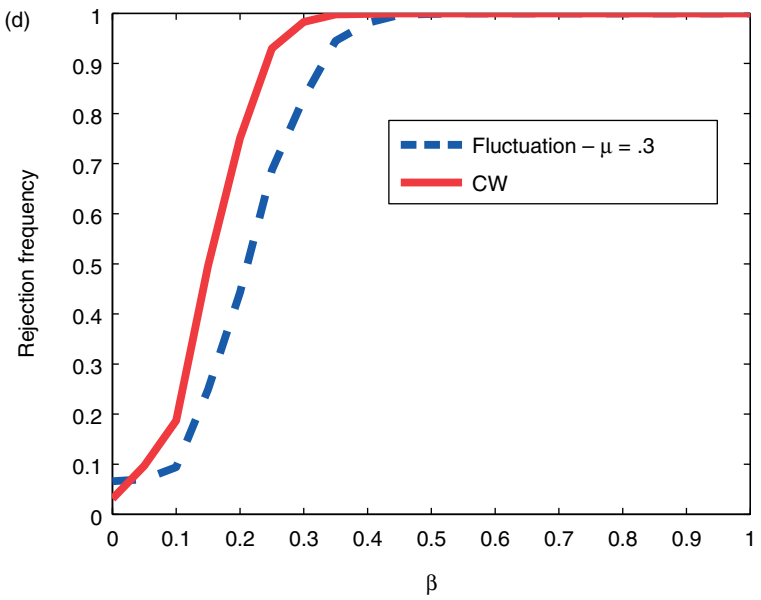

Figure 2(d). Power of CW Fluctuation test and full-sample CW test. Unequal but constant relative performance. This figure is available in color online at www.interscience.wiley.com/journal/jae

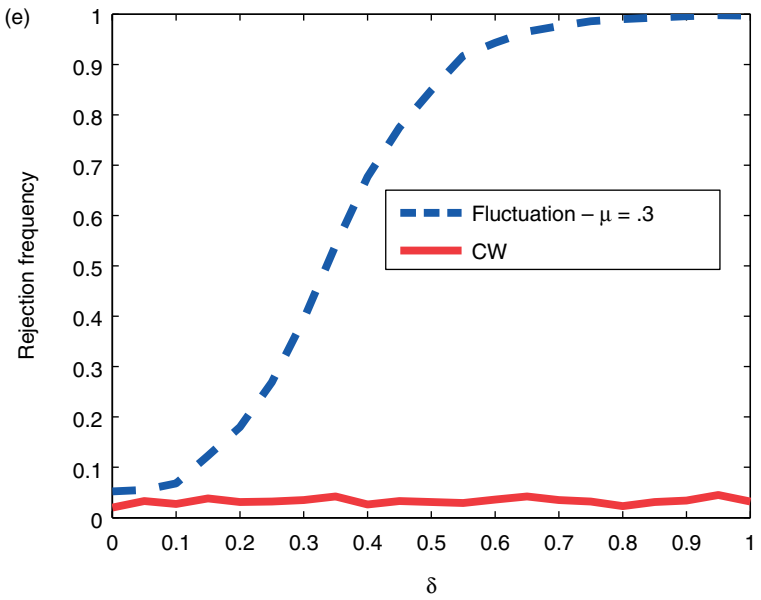

Figure 2(e). Power of CW Fluctuation test and full-sample CW test. Break in relative performance. This figure is available in color online at www.interscience.wiley.com/journal/jae

DOI: 10.1002/jae

5. EMPIRICAL EVIDENCE ON THE TIME VARIATION IN THE OUT-OF-SAMPLE RELATIVE FORECASTING PERFORMANCE OF EXCHANGE RATE MODELS

A vast literature has analyzed the out-of-sample forecasting performance of exchange rate models since the seminal papers by Meese and Rogoff (1983a,b). Even though the Meese–Rogoff stylized fact that a random walk predicts exchange rates better than conventional macroeconomic models is still alive, there are a variety of conjectures regarding why that might be the case. These include the presence of parameter instabilities in predictive regressions. As shown by Rossi (2006), parameter instability plagues the estimation of exchange rate models. Such instability might confound results of in-sample Granger causality tests of whether the macroeconomic fundamentals predict future exchange rate changes. By using Granger causality tests that are robust to parameter instability, Rossi (2006) rejects the hypothesis that exchange rates are random walks in-sample. Kilian and Taylor (2003) arrive at the same conclusion on the basis of an in-sample test that allows for nonlinearities in the DGP (which is equivalent to parameter instabilities in predictive regressions). As an additional economic motivation for our analysis, Timmermann (2008) suggests that, as a result of efficient markets where investors are constantly searching for arbitrage opportunities, one would not expect to find constant predictability patterns.

Given the widespread instabilities detected by in-sample tests and the promising finding that, when such instabilities are correctly taken into account, it is possible to reject the random walk model, we proceed to examine the implications of these findings for forecasting exchange rates out-of-sample by using the techniques developed in this paper.

We consider two models of exchange rate determination: the conventional uncovered interest rate parity (UIRP) model and the model with Taylor rule fundamentals considered by Molodtsova and Papell (2007). The latter report that evidence of short-term predictability of exchange rates appears to be stronger with the Taylor rule model than with the UIRP model. The two models are specified as follows. Let the logarithm of the bilateral nominal exchange rate (determined as the domestic price of foreign currency) be denoted by s_t. The one-step-ahead change in s_t can be modeled as a function of its deviation from the current level of the macroeconomic fundamental:

$$s_{t+1} - s_t = \alpha + \beta z_t + \varepsilon_{t+1} \tag{8}$$

where $z_t = f_t - s_t$, and f_t is the long-run equilibrium level of the nominal exchange rate determined by the macroeconomic fundamental.[7] In the UIRP model:

$$f_t = (i_t - i_t^*) + s_t \tag{9}$$

where $(i_t - i_t^*)$ is the short-term interest differential between the home and the foreign countries. In the model with Taylor fundamentals, the home country interest rate follows a Taylor rule (see Taylor, 1993):

$$i_t = \pi_t + \delta(\pi_t - \pi^T) + \gamma y_t^{\text{gap}} + r \tag{10}$$

where π_t is the inflation rate, π^T is the target level of inflation, y_t^{gap} is the output gap[8] and r is the equilibrium level of the real interest rate. A similar condition applies to the foreign country.

[7] We do not consider multi-step-ahead changes in the exchange rate because the tests of out-of-sample forecast comparisons do have a non-normal distribution when the number of steps ahead is non-negligible and the regressor z_t is highly persistent (as it is in our data). See Rossi (2005a).

[8] The output gap is the percentage difference between actual and potential output at time t, where potential output is measured by the linear time trend in output. The coefficients of the linear time trend are re-estimated as the parameters

DOI: 10.1002/jae

Let asterisks denote the variables in the foreign country. If the coefficients of the home and foreign Taylor rule are similar (the 'symmetric Taylor rule with homogeneous coefficients and no smoothing' case considered in Molodtsova and Papell, 2007) then, by taking their differences:

$$i_t - i_t^* = (1+\delta)(\pi_t - \pi_t^*) + \gamma(y_t^{\text{gap}} - y_t^{\text{gap}*}) \tag{11}$$

Therefore, in the exchange rate model with Taylor rule fundamentals, by substituting (11) into (9) we have

$$f_t = (1+\delta)(\pi_t - \pi_t^*) + \gamma(y_t^{gap} - y_t^{gap*}) + s_t \tag{12}$$

We estimate the models using monthly data for output, interest rates, and inflation from the IMF's International Financial Statistics database from 1973 : 3 to 2008 : 1.[9] The exchange rate series are from the Federal Reserve Bank of St Louis. The countries that we consider are Japan, Switzerland, Australia, Canada, Great Britain, Sweden, Denmark, Germany, France, Italy, the Netherlands, and Portugal. We recursively estimate the parameters of the two models over rolling windows of 50 observations starting from 1983 : 2.[10] All tests are one-sided: the null hypothesis is that, for each country, the model with fundamentals has the same MSFE as the random walk; the alternative is that the model with fundamentals forecasts better than the random walk.

Table III reports the *p*-values of the average tests of equal predictive ability via the GW and the CW tests. For completeness, Table IV reports the corresponding average out-of-sample

Table III. *p*-Values of full-sample tests

	GW				CW			
	1973 : 3–2004 : 10		1973 : 3–2008 : 1		1973 : 3–2004 : 10		1973 : 3–2008 : 1	
	Taylor	UIRP	Taylor	UIRP	Taylor	UIRP	Taylor	UIRP
Japan	0.75	0.69	0.80	0.69	0.07	0.06	0.07	0.06
Canada	0.15	0.20	0.23	0.37	0.00	0.00	0.01	0.02
Switzerland	—	0.84	—	0.85	—	0.18	—	0.17
UK	0.77	0.72	0.80	0.74	0.27	0.08	0.29	0.08
France	0.77	0.98	0.82	0.98	0.19	0.83	0.34	0.83
Germany	0.89	0.83	0.88	0.83	0.57	0.20	0.56	0.20
Italy	0.92	0.75	0.93	0.75	0.14	0.31	0.14	0.31
Sweden	0.96	1.00	0.93	1.00	0.35	0.95	0.16	0.94
Australia	—	0.76	0.42	0.76	—	0.19	0.28	0.19
Denmark	—	—	—	—	—	—	—	—
Netherlands	0.96	—	0.95	—	0.67	—	0.65	—
Portugal	0.99	—	0.99	—	0.52	—	0.52	—

The table reports *p*-values of the full-sample GW and CW tests. The tests compare the models with fundamentals, either the model with Taylor-rule fundamentals or the UIRP model, to a random walk benchmark.

of the model are re-estimated through time, and their estimation is based only on variables available in the information set of the forecaster at the time in which the forecast is made.

[9] The data are the same as in Molodtsova and Papell (2007), and are the seasonally adjusted industrial production index for output and the 12-month difference of the CPI for the annual inflation rate.

[10] Our theoretical results hold as long as the number of out-of-sample forecast error differences within the estimation window is large enough for the asymptotic theory to apply. However, our sample is fairly small. In order to strike a balance between the two, we choose a window of 50 observations, which should allow our approximations to be sufficiently precise.

DOI: 10.1002/jae

Table IV. Out-of-sample MSFE differences (standardized)

	1973 : 3–2004 : 10		1973 : 3–2008 : 01	
	Taylor	UIRP	Taylor	UIRP
Japan	−0.66	−0.49	−0.85	−0.49
Canada	1.04	0.85	0.72	0.33
Switzerland	—	−1.01	—	−1.01
UK	−0.74	−0.57	−0.85	−0.65
France	−0.74	−2.01	−0.90	−2.01
Germany	−1.23	−0.96	−1.18	−0.96
Italy	−1.43	−0.66	−1.45	−0.66
Sweden	−1.80	−3.34	−1.48	−3.51
Australia	—	−0.70	—	−0.70
Denmark	—	—	—	—
Netherlands	−1.72	—	−1.69	—
Portugal	−2.22	—	−2.22	—

The table reports the MSFE of the random walk minus the MSFE of the economic model. The difference has been rescaled by the standard deviation of the MSFE differences so that it is comparable to the test statistic considered by Diebold and Mariano (1995).

MSFE differences (divided by their standard deviation). We consider both the full sample (1973 : 3–2008 : 1) and the subsample considered by Molodtsova and Papell (2007), namely 1973 : 1–2004 : 10. We note that the GW test does not reject the null hypothesis that the models with fundamentals and the random walk have the same predictive ability. The CW test instead rejects the null in favor of the model with fundamentals for Japan and Canada at the 10% significance level, and for the UIRP for the UK. The latter results provide interesting evidence in favor of the model with fundamentals, and are Molodtsova and Papell's (2007) most important piece of evidence in favor of short-horizon predictive ability of exchange rates.

How robust are these findings? Rogoff and Stavrakeva (2008) show that the results may depend on the initial estimation point as well as the size of the rolling window. In other words, the relative forecasting performance of the models might have changed over time, and we address this issue by using our tests.

We focus on the UIRP model. Figures 3(a) and 4(a) report results for the GW Fluctuation test for Germany and the UK. Figures 3(b) and 4(b) report results for the CW Fluctuation test. The figures report both the Fluctuation test statistic (constructed using a centered moving window) as well as the one-sided critical value at 5% (the constant line). Positive values of the test statistic indicate that the model with fundamentals is better than the random walk. Our procedure points out that there have been periods in which the Deutsche Mark and the British pound have been predictable, and this happened at the beginning of the out-of-sample period, in the late 1980s. However, such evidence has disappeared in the 1990s. Interestingly, by comparing these results with Table IV, note that average tests of predictive ability would have been incapable of uncovering such favorable evidence in favor of the UIRP model in the Deutsche Mark case.

For the British Pound, Figure 5 shows results for the One-Time Reversal test. The test clearly identifies a reversal in the relative forecasting ability of the two models around 1989 from a situation where the model with fundamentals forecasts best to a situation where the random walk forecasts best. The pattern is similar to that reported in Figure 4, except that the Fluctuation test 'smooths out' the measure of relative performance over time.

DOI: 10.1002/jae

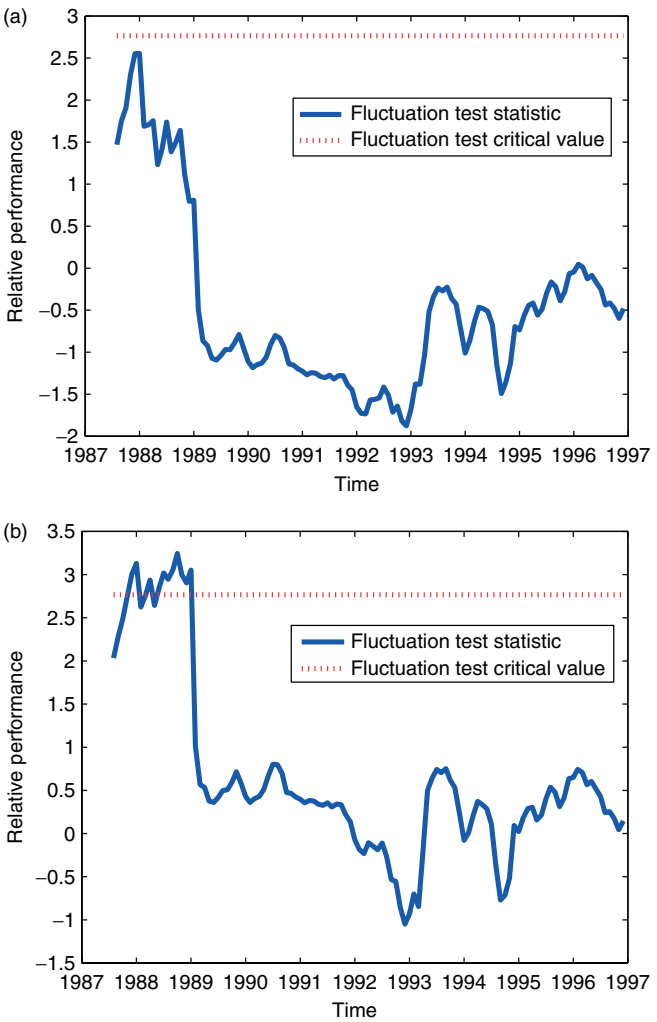

Figure 3. (a) GW Fluctuation test, Deutsche Mark. (b) CW Fluctuation test, Deutsche Mark. The figure shows Fluctuation test statistics and the critical value of the Fluctuation test. Positive values of the Fluctuation statistic imply that the economic model is better than the random walk. This figure is available in color online at www.interscience.wiley.com/journal/jae

Overall, we interpret our empirical results as pointing towards a worsening of the performance of the models with fundamentals relative to the random walk in the most recent years, to the point that measures of average performance would overstate the recent predictive ability of the economic models.

6. CONCLUSIONS

We introduce new methods for assessing the possible presence of time variation in the relative forecast performance of two models. A companion paper (Giacomini and Rossi, 2007) considers the

DOI: 10.1002/jae

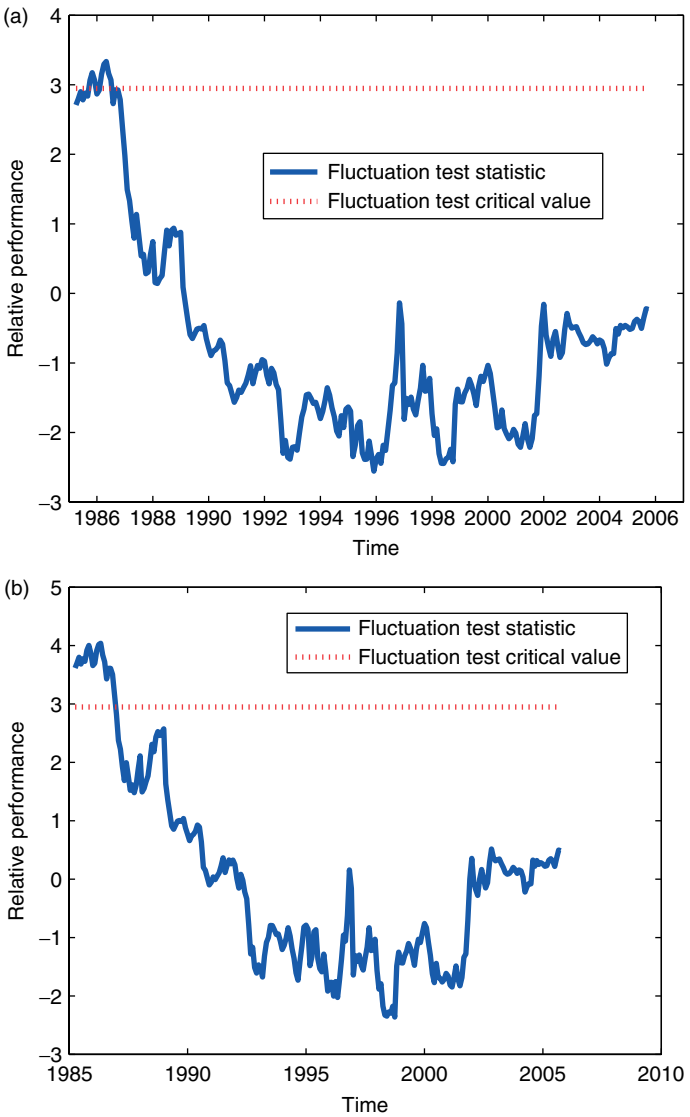

Figure 4. (a) GW Fluctuation test, UK pound. (b) CW Fluctuation test, UK pound. The figure shows Fluctuation test statistics and the critical value of the Fluctuation test. Positive values of the Fluctuation statistic imply that the economic model is better than the random walk. This figure is available in color online at www.interscience.wiley.com/journal/jae

problem of comparing the in-sample performance of competing models in unstable environments. Our techniques can be generally applied to nonlinear, dynamic, nested or non-nested forecasting models.

We proposed two tests: a Fluctuation test, which does not require specifying the nature of the instability under the alternative hypothesis, and a One-Time Reversal test, when the alternative is of a single, permanent break in the relative performance of the two models.

DOI: 10.1002/jae

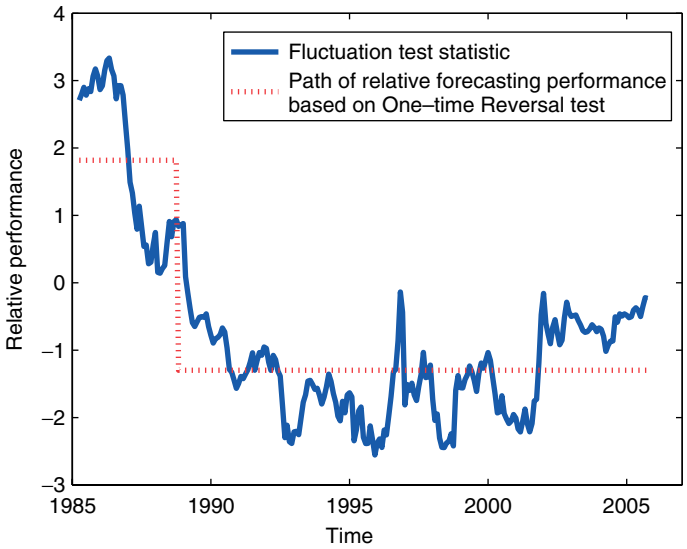

Figure 5. One-time Reversal test, UK pound. The figure shows the Fluctuation test statistics and the path of relative performance implied by the One-Time Reversal test. Positive values imply that the economic model forecasts better than the random walk. This figure is available in color online at www.interscience.wiley.com/journal/jae

A natural question to ask is what a forecaster should do if the tests find instability in the relative performance of competing models. The paper does not investigate this issue in depth, but possible strategies can be devised if one is willing to specify the nature of the instability. For example, in the case of a one-time permanent break (or a finite number of such breaks), the forecast strategy suggested by our One-Time Reversal test is to select the forecast that is most accurate in the period after the break. Alternatively, the Fluctuation test may reveal that one model performs better in certain periods and the competing model is more accurate in other periods, in which case a combination forecast may be more robust to structural instability than either of the individual forecasts. A forecast combination with time-varying weights (e.g., Elliott and Timmermann, 2005) would in this case be a natural way to accommodate underlying instability in the relative forecast performance of the models.

We illustrate the usefulness of our techniques by analyzing the time variation in the relative forecasting performance of exchange rate models with economic fundamentals relative to the random walk. Our techniques uncover a sharp worsening in the forecasting ability of the UIRP model around 1989. Existing tests of equal predictive ability, that consider only average predictive ability over the out-of-sample period, would miss this interesting stylized fact.

ACKNOWLEDGEMENTS

We thank Kirstin Hubrich and two anonymous referees for detailed comments, as well as Taebong Kim for assistance in collecting the data used in the empirical analysis, and Tanya Molodtsova, Ulrich Mueller, and seminar participants at the 2007 European Central Bank Workshop on Forecast Uncertainty in Macroeconomics and Finance, Oxford University, Warwick University, Manchester

University, the Tinbergen Institute, and the 2007 NBER Summer Institute for useful comments and suggestions. Support by NSF grant 0647770 is gratefully acknowledged.

REFERENCES

Andrews DWK. 1991. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica* **59**: 817–858.

Andrews DWK. 1993. Tests for parameter instability and structural change with unknown change point. *Econometrica* **61**: 821–856.

Bai J. 1997. Estimation of a change point in multiple regression models. *Review of Economics and Statistics* **79**(4): 551–563.

Bai J, Perron P. 1998. Estimating and testing linear models with multiple structural changes. *Econometrica* **66**(1): 47–78.

Brown RL, Durbin J, Evans JM. 1975. Techniques for testing the constancy of regression relationships over time with comments. *Journal of the Royal Statistical Society, Series B* **37**: 149–192.

Chu CJ, Hornik K, Kuan C. 1995. MOSUM tests for parameter constancy. *Biometrika* **82**(3): 603–617.

Clark T, McCracken M. 2001. Tests of Equal forecast accuracy and encompassing for nested models. *Journal of Econometrics* **105**(1): 85–110.

Clark T, West KD. 2006. Using out-of-sample mean squared prediction errors to test the martingale difference hypothesis. *Journal of Econometrics* **135**: 155–186.

Diebold FX, Mariano RS. 1995. Comparing predictive accuracy. *Journal of Business and Economic Statistics* **13**: 253–263.

Elliott G, Timmermann A. 2005. Optimal forecast combination under regime switching. *International Economic Review* **46**: 1081–1102.

Giacomini R, Rossi B. 2007. Model comparisons in unstable environments. *ERID Working Paper* #30, Duke University.

Giacomini R, White H. 2006. Tests of conditional predictive ability. *Econometrica* **74**: 1545–1578.

Inoue A, Kilian L. 2006. On the selection of forecasting models. *Journal of Econometrics* **130**(2): 273–306.

Kilian L, Taylor MP. 2003. Why is it so difficult to beat the random walk forecast of exchange rates? *Journal of International Economics* **60**(1): 85–107.

McCracken MW. 2000. Robust out-of-sample inference. *Journal of Econometrics* **99**: 195–223.

Meese R, Rogoff K. 1983a. Exchange rate models of the seventies: do they fit out of sample? *Journal of International Economics* **14**: 3–24.

Meese R, Rogoff K. 1983b. The out of sample failure of empirical exchange rate models. In *Exchange Rates and International Macroeconomics*, Frankel J (ed.) University of Chicago Press for NBER: Chicago, IL; 67–105.

Molodtsova T, Papell DH. 2007. *Out-of-sample exchange rate predictability with Taylor rule fundamentals*. Mimeo, University of Houston.

Newey W, West K. 1987. A simple, positive Semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* **55**: 703–708.

Ploberger W, Kramer W. 1992. The CUSUM test with OLS residuals. *Econometrica* **60**(2): 271–285.

Rissanen J. 1986. Stochastic complexity and modeling. *Annals of Statistics* **14**: 1080–1100.

Rogoff KS, Stavrakeva V. 2008. *The continuing puzzle of short horizon exchange rate forecasting*. NBER Working Paper W14071.

Rossi B. 2005a. Testing long-horizon predictive ability, and the Meese–Rogoff puzzle. *International Economic Review* **46**(1): 61–92.

Rossi B. 2005b. Optimal tests for nested model selection with underlying parameter instabilities. *Econometric Theory* **21**(5): 962–990.

Rossi B. 2006. Are exchange rates really random walks? Some evidence robust to parameter instability. *Macroeconomic Dynamics* **10**(1): 20–38.

Stock JH, Watson MW. 2003a. Forecasting output and inflation: the role of asset prices. *Journal of Economic Literature* **41**: 788–829.

Stock JH, Watson MW. 2003b. *Introduction to Econometrics*. Addison-Wesley: Reading, MA.

DOI: 10.1002/jae

Taylor JB. 1993. Discretion versus policy rules in practice. *Carnegie-Rochester Conference Series on Public Policy* **39**: 195–214.
Timmermann A. 2008. Elusive return predictability. *International Journal of Forecasting* **24**: 1–18.
Wei CZ. 1992. On predictive least squares principles. *Annals of Statistics* **20**: 1–42.
West KD. 1996. Asymptotic inference about predictive ability. *Econometrica* **64**: 1067–1084.

DOI: 10.1002/jae

APPENDIX: PROOFS

Proof of Proposition 1 *Let* $\sum_j \equiv \sum_{j=t-m/2}^{t+m/2-1}$ *for* $t = R + h + m/2, \ldots, T - m/2 + 1$. *We have*

$$\begin{aligned}
&\sigma^{-1}m^{-1/2}\sum_j \Delta L_j(\widehat{\theta}_{j-h,R}, \widehat{\gamma}_{j-h,R}) \\
&= (m/P)^{-1/2}\left(\sigma^{-1}P^{-1/2}\sum_{j=R+h}^{t+m/2-1}\Delta L_j(\widehat{\theta}_{j-h,R}, \widehat{\gamma}_{j-h,R}) - \sigma^{-1}P^{-1/2}\sum_{j=R+h}^{t-m/2-1}\Delta L_j(\widehat{\theta}_{j-h,R}, \widehat{\gamma}_{j-h,R})\right)
\end{aligned}$$

By Assumption 1(a), we have

$$\sigma^{-1}m^{-1/2}\sum_j \Delta L_j(\widehat{\theta}_{j-h,R}, \widehat{\gamma}_{j-h,R}) \Longrightarrow [\mathcal{B}(\tau + \mu/2) - \mathcal{B}(\tau - \mu/2)]/\sqrt{\mu}$$

The statement in the proposition then follows from the fact that, under H_0, $\widehat{\sigma}$ in (2) is a consistent estimator of σ (Andrews, 1991; Newey and West, 1987). □

Proof of Proposition 2 *Note that, by Assumption 1(a), under the null hypothesis:*

$$\sigma^{-1}P^{-1/2}\sum_{j=R+h}^{T}\Delta L_j(\widehat{\theta}_{j-h,R}, \widehat{\gamma}_{j-h,R}) \Longrightarrow \mathcal{B}(1) \tag{13}$$

$$\begin{aligned}
&\sigma^{-1}(t/P)^{-1/2}(1 - t/P)^{-1/2}\left[P^{-1/2}\sum_{j=R+h}^{t}\Delta L_j(\widehat{\theta}_{j-h,R}, \widehat{\gamma}_{j-h,R})\right. \\
&\left.-(t/P)P^{-1/2}\sum_{j=R+h}^{T}\Delta L_j(\widehat{\theta}_{j-h,R}, \widehat{\gamma}_{j-h,R})\right] \\
&\Longrightarrow \tau^{-1/2}(1-\tau)^{-1/2}[\mathcal{B}(\tau) - \tau\mathcal{B}(1)] = \tau^{-1/2}(1-\tau)^{-1/2}\mathcal{BB}(\tau)
\end{aligned} \tag{14}$$

where (13) and (14) are asymptotically independent since $\mathrm{cov}(\mathcal{B}(1), \mathcal{BB}(\tau)) = 0$. Then, by the continuous mapping theorem, we have

$$LM_1 + LM_2(t) \Longrightarrow \mathcal{B}(1)^2 + \tau^{-1}(1-\tau)^{-1}\mathcal{BB}(\tau)^2$$

and the result follows. □

DOI: 10.1002/jae