New evidence of exchange rates predictability using the long swings model

Alex Nikolsko-Rzhevskyy
University of Memphis

Ruxandra Prodan
University of Houston

July 26, 2009

Using the two state Markov Switching “long swings” model and the recently developed Clark and West’s (2006) inference procedure for testing the equal predictive ability of two nested models, we present evidence of both, short run (one month) and long run (up to one year) predictability for monthly exchange rates over the post-Bretton Woods period. This is an important result as recent literature, mostly using economic models, fails to find significant evidence of consistent multi-horizon predictability. The results show that our model strongly outperforms the random walk model for 9 out of 12 exchange rate series at a short-horizon. In addition, for 7 out of 12 countries, we are able to find evidence of long-horizon predictability that declines as the forecast horizon increases. Our results remain robust to alternative out-of-sample test statistics and forecast windows.

JEL Classification: C5, F31, F37
Keywords: exchange rates, forecasting, Markov switching, long swings

1 We thank Tetyana Molodtsova, David Papell, and participants at the MEA 2009 Annual Meeting for helpful comments and discussions. This work was supported in full or in part by a grant from the Fogelman College of Business & Economics at the University of Memphis.
1. Introduction

There is a vast literature on forecasting asset prices. Specifically, studies have attempted to forecast nominal exchange rates, mostly using models based on macroeconomic fundamentals. Nevertheless, research has not yet reached a consensus regarding the forecasting performance of different models when modeling the exchange rate.

In their classic paper, Meese and Rogoff (1983) demonstrated that a random walk forecast of the post-Bretton Woods exchange rate generally outperforms alternative models drawn from economic theory, including purchasing power parity (PPP), uncovered interest rate parity (UIRP) and simple versions of the monetary and portfolio balance models. Later, a number of authors found evidence of some predictability of economic exchange rate models at long-horizon: MacDonald and Taylor (1992), Mark (1995), Chen and Mark (1996) assessed the out-of-sample performance of the three alternative fundamental models proposed in the literature, PPP, UIRP and the flexible price monetary model, and found that the monetary models have the greatest predictive power. Chinn and Meese (1995), using a variety of error-correction models, also found some evidence of long term exchange rate predictability.

The inference procedures and robustness of these papers’ results have been called into question by Killian (1999), Faust, Rogers and Wright (2003) and Berkowitz and Giorgianni (2001). In particular, a recent study by Cheung, Chinn and Pascual (2005) examined the out-of-sample performance of the UIRP, monetary, productivity-based and behavioral exchange rate models and concluded that none of them outperforms the random walk at any horizon.

The literature on exchange rate forecasting usually evaluates the out-of-sample predictability of two models (linear fundamental-based model and a random walk) on the basis of various measures. The most commonly used measure of forecasting is the mean squared prediction error (MSPE). In order to evaluate the out-of-sample performance of the models based on the MSPE comparison, tests for equal predictability of two non-nested models introduced by Diebold and Mariano (1995) and West (1996) are most often used (henceforth, DMW tests). While these tests are appropriate for non-nested models, when testing nested models, the use of standard normal critical values usually results in severely undersized tests. Consequently, while

---

using the standard critical values, the typical result is that the random walk null cannot be rejected in favor of the model-based alternative.

More recent studies take the above matters into consideration by using a newly developed inference procedure for testing the null of equal predictive ability of an econometric model and a martingale difference model proposed by Clark and West (2006, 2007), (henceforth the CW test). They argue that this methodology is preferable to the standard DMW procedure when the two models are nested. The test statistic takes into account that under the null, the sample MSPE of the alternative model is expected to be greater than that of the random walk model and adjusts for the upward shift in the sample MSPE of the alternative model. Previous papers have used the CW procedure to assess the exchange rate predictability. Clark and West (2006) find that the UIERP model significantly outpredicts the random walk model for two out of four countries, at a 1-month horizon. Gourinchas and Rey (2007) find that the ratio of net exports to net foreign assets forecasts movements in weighted exchange rates better than the random walk model at short and long-horizon. Alquist and Chinn (2007) examine the relative predictive power of the sticky-price monetary model, UIERP model and a transformation of net exports and net foreign assets and find that no single model uniformly outperforms the random walk forecast. Molodtsova and Papell (2008) re-examine the out-of-sample exchange rate predictability using Taylor rule fundamentals. They find that a symmetric Taylor rule model with heterogeneous coefficients and interest rate smoothing provides evidence of short-horizon predictability. Engel, Mark and West (2007) find some long-horizon evidence of predictability using panel data in the case of monetary and PPP fundamentals and single-equation models for Taylor rule fundamentals.

Rogoff and Stavrakeva (2009) examine the most popular exchange rate forecasting structural models and specifications (Gourinchas and Rey (2007), Engel, Mark and West (2007) and Molodtsova and Papell (2008)) and conclude that one of the sources of the overly optimistic results is the failure to check robustness with respect to alternative out-of-sample test statistics. They argue that the new asymptotic out-of-sample tests, such as the CW test, are easy to use but bootstrapped out-of-sample tests DMW remain more powerful and better sized. Furthermore, they argue that all of the structural models and specifications reviewed fail to produce robust

---

3 For simplicity, in the rest of the paper, we would loosely refer to the Martingale process as a random walk.
4 The simulations in Clark and West (2006) suggest that the inference made using asymptotically normal critical values results in properly-sized tests.
forecasts over different sample periods. Using real time data and both, CW and DMW tests, Molodtsova, Nikolsko-Rzhevskyy and Papell (2009) find evidence of short-horizon predictive ability for the Euro/US dollar exchange rate, partially addressing the above critique.

In all of the above papers, the typical route that researchers have taken when forecasting nominal exchange rates is to design more complicated economic models in an attempt to outperform the random walk. However, none of them provides consistent cross country evidence of both, short and long-horizon predictability. Our approach is different as we do not model exchange rate movements based on the macroeconomic fundamentals. Instead, we argue that the relative success attributed to the macroeconomic fundamentals models when forecasting nominal exchange rates is in part due to the drifting term – a constant – that is included in the majority of the economic models. We show that a simple random walk with drift has in many cases a superior predictive ability to the simple random walk model when forecasting nominal exchange rates and its performance is comparable to the performance of an economic model without a drifting term. Thus modeling the drifting term, while ignoring the economic models, should be just as promising.

It has been previously documented that a simple 2-state Markov-switching random walk model with drift (henceforth MS-RW), which allows the constant term as well as the variance of innovations to take two distinct values during times of appreciation and depreciation, is a good representation for the nominal exchange rate. In this model the exchange rate follows long swings, switching between an upward and a downward drift. These swings may originate from economic fundamentals; if economic growth is relevant for exchange rates, then business cycle differences between countries can lead to long swings (Kaminsky, 1993).

Several papers develop theoretical models that describe the exchange rate dynamics as following long swings: Bask (2007) constructs a DSGE model where the interest rate is set in response to the output gap to mimic the optimal monetary policy under commitment. Evans and Lewis (1995) develop a model that assumes that the exchange rate switches between appreciating and depreciating regimes are incorporated into rational traders’ forecasts of the future exchange rate. Shen and Chen (2004) show that developed countries experience symmetric long swings while developing countries may experience asymmetric speed of adjustment (long swings in appreciation and short swings in depreciation).
Among empirical studies, Engel and Hamilton (1990) found that the US dollar/German mark, US dollar/UK pound and US dollar/French franc exchange rates can be described well by Hamilton’s (1988, 1989) Markov switching model. Following this study, Engel (1994) investigated whether this model is a useful specification for the out-of-sample exchange rate predictability. Using a sample from 1973 to 1991, he fits the Markov-switching model for six exchange rate series at quarterly and monthly frequencies and finds that the model fits well in sample for many exchange rates but it is not able to generate forecasts superior to the random walk by using either the MSPE or the mean-absolute error (MAE) criterion. Klaassen (2004) enhances Engel’s model with a GARCH error structure, but also fails to find any nominal exchange rate predictability evaluated based on the MSPE criterion. However, the above studies document some degree of short-term predictability based on the direction of change measure.  

Our goal is to re-analyze the in-sample and the out-of-sample forecasting performance of the MS-RW model. The main contributions are a longer data set, more exchange rate series, a more accurate estimation method and a correctly sized predictability test. As a result, we are able to reverse Engel’s (1994) findings and provide evidence of both, short-term and long-term predictability of the MS-RW model, that is robust across various samples and different out-of-sample test statistics.

We examine the exchange rate series for 12 OECD countries versus the US dollar from March 1973 to January 2008, previously employed by Molodtsova and Papell (2008). Using Engel and Hamilton’s (1990) model, we first illustrate that for all the exchange rate series the one-state OLS models might be miss-specified and then provide evidence of a 2-state Markov process. We next check for the out-of-sample predictability of the MS-RW model. From the beginning, a clear distinction has to be made between forecastability and predictability. When evaluating forecasts from two non-nested models, one could compare the mean squared prediction errors (MSPE) from the two models, scaled to produce the tests of Diebold and Mariano (1995) and West (1996) (DMW statistic), and determine whether one model forecasts better than the other. In our case, the null hypothesis (a random walk) and the alternative model (MS-RW) are nested, resulting in severely undersized out-of-sample tests when using standard normal critical values. We therefore evaluate the out-of-sample predictability performance of the MS-RW model.

---

5 Rogoff and Stavrakeva (2008) point out that the “direction of change” criteria is known to select a model which performs well in predicting small changes but poorly at predicting major ones.
MS-RW model versus the random walk model by using the CW adjustment of the DMW statistic. The results show that, at the one-month horizon, our model strongly outperforms the random walk model for 9 out of 12 countries at 10% significance level or better. In addition, we find strong evidence of long-horizon predictability (more than one month horizon): 3 months horizon for 7 countries, 6 months horizon for 4 countries and 12 months horizon for 3 countries. A common characteristic for all countries is that the predictability evidence decreases as the forecast horizon increases.

As discussed above, we are aware of the Rogoff and Stavrakeva’s (2008) critique of the literature on the predictability of the exchange rates. Following their methodology, we check the robustness of our short-horizon results (one month) in several ways: First, when bootstrapping critical values for the CW test and using the bootstrapped Clark and McCracken (2005) (CM) test, we find that our results remain consistent. Second, we consider alternative out-of-sample tests statistics as the bootstrapped DMW and Theil’s U (TU) tests; we are able to provide evidence of forecasting ability (the MSPE from the alternative model is significantly smaller than the MSPE from the null model) for the MS-RW model for 3 out of 9 countries. Finally we explore the robustness of our short-horizon results with respect to different forecast windows, using the CW and the TU tests. We find that our results remain robust for 6 out of 9 countries when using the CW test and for 4 out of 9 countries when using the TU test.

2. Methodology

2.1. Data

The data consists of currencies for 12 countries: the Japanese yen, Swiss franc, Australian dollar, Canadian dollar, British pound, Swedish kronor, Danish kroner, Deutsche mark, French franc, Italian lira, Dutch guilder, and Portuguese escudo. The exchange rate is defined as the US dollar price per unit of foreign currency. The data is taken from the Federal Reserve Bank of Saint Louis database.

---

6 We also check for the robustness of our predictability results and provide some evidence of forecastability.
7 Using the alternative out-of-sample test statistics and comparing our results with Rogoff and Stavrakeva’s findings, we conclude that our evidence of forecasting ability significantly supersedes the previous literature results, using economic models.
8 The data is taken from the Federal Reserve Bank of Saint Louis database.
9 The only exceptions are U.K. and Australia, for which the definition is reversed.
ends in January 2008. The exchange rates for the 5 European Union (EU) countries are examined before December 1998, prior to the formation of the EU.

2.2. Basics: The evidence of long swings in the exchange rate data

Much of the previous research focuses on testing the predictability of the various exchange rate models (UIRP, Monetary, PPP and the Taylor rule model) versus the simple random walk model. These models include, by default, a constant term. We argue that the relative success attributed to the exchange rates’ predictability of the macroeconomic fundamentals models is in part due to the drifting term – a constant – that is included in the majority of the economic models.

In order to distinguish whether the superior predictability of the economic models is in fact due to the economic part or to the drifting term we consider the following exercise: Recall Clark and West’s (2006) model, based on interest parity, which relates the change in the spot exchange rate from the end of month $t$ to the end of month $t+1$ ($dy_t$) to a constant ($\alpha$) and the one-month interest rate differential ($i^*-i$) at the end of month $t$, as shown below:\(^{10}\)

\[
\begin{align*}
    dy_t &= \alpha + (i^*_t - i_t) \\
\end{align*}
\]

The data consists of monthly observations for four exchange rates, Canada, Japan, Switzerland, and the U.K against the US dollar, where the sample spans from March 1973 to October 2003. The CW test is used to test the out-of-sample predictability of the above model against the random walk model at a one step forecasting horizon. In Table 1, column 1, we reprint Clark and West’s (2006) results, which illustrate that based on the CW test, the UIRP model outperforms the random walk for two out of four countries (Canada and Japan).\(^ {11}\)

We re-estimate Clark and West’s (2006) model using our extended exchange rate data, which spans from March 1973 to January 2008, for the same countries. Our results are similar (Table 1, column 2), illustrating the superior predictability of the UIRP model. In order to make the distinction between the predictability of the interest rate differential and that of the drifting term we run two experiments: first, we test the predictability of the UIRP model (1) versus the

\(^{10}\) The exchange rate data were obtained from the Board of Governor’s FAME database; the one-month interest rates, which are averages of bid and ask rates on Eurocurrency deposits (London close), were obtained from Global Insight’s FACS database.

\(^{11}\) For Japan, however, the model is significant only at 20 percent.
random walk model running a regression where we do not include a constant term \( \alpha \), and second, we test the predictability of the UIRP model (1) excluding the interest rate differential \( i_t^* - i_t \) (a random walk with drift) versus a simple random walk model. The results are presented in Table 1, columns 3 and 4. Taking in consideration columns 2, 3 and 4, our observations are as follows: For Japan, where we found that the UIRP model outperforms the RW model at the 10% significance level, we argue that the predicting power comes from the drift term (the RW with drift model is significant at the 10% level) while the economic part is actually making the forecast worse (we find no significance when testing the UIRP model without drift versus the random walk model). For Canada, where we found that the UIRP model outperforms the RW model at the 5% significance level, we find that the predicting power comes from both, the economic part (significant at the 10% level) and the drift term (significant at the 20% level). For Switzerland, where we find that the UIRP model outperforms the RW model at the 1% significance level, we do not find any significance when we test each specification separately (the UIRP model without a drift and the RW with drift versus the RW model) but when combined, the result is significant. Finally, for the UK we do not find significant results with any of the specifications.

As shown above, the relative success attributed to the economic models when forecasting nominal exchange rates is partially due to the drifting term.\(^{12}\) In the above example the performance of the RW with drift is at least as good as the performance of the model without a constant term. Therefore, modeling the drifting term, while ignoring the economic models, should be just as promising.

2.3. The Markov Switching – Random Walk with Drift Model

As mentioned in the introduction, there is much theoretical and empirical evidence that long swings are present in the exchange rate data. Just by looking at the graph of Canada (representative non-EU country) and Germany (representative EU country) versus dollar exchange rates (Figure 1) we observe persistent appreciations and depreciations of the currency.

\(^{12}\) Similarly, Molodtsova and Papell (2008), when studying the predictability of various specifications of the Taylor rule, find that the significance of their results when including a constant considerably deteriorates when the constant term is excluded. Killian (1999) stresses the importance of differentiating between a random walk model and a random walk with drift.
Engel and Hamilton (1990) found that the dollar/German mark, dollar/UK pound and dollar/French franc exchange rates can be described well by Hamilton’s (1988, 1989) MS-RW model. Engel (1994) has investigated whether this model is a useful specification for the out-of-sample exchange rate predictability. He fits the MS-RW model for six exchange rates at quarterly and monthly frequencies. The estimation period is 1973-1986 and the post-sample forecast period is 1986-1991. While the model fits well in sample at the quarterly frequency, it is not able to generate superior forecasts to the random walk model by using either the MSPE or the MAE criteria.

In this section we outline the MS-RW model and we discuss its estimation. Following Engel and Hamilton (1990) and Engel (1994), we model the exchange rate behavior as a 2-state MS-RW with drift model, allowing the drift term as well as the variance of innovations to take two distinct values during times of appreciation and depreciation.\(^{13}\) The basic idea of the model is to decompose non-stationary time series into a sequence of stochastic, segmented time trends. We model any given month’s change in the exchange rate as being derived from one of the two regimes corresponding to episodes of rising or falling exchange rates. The regime at any given time is presumed to be the outcome of a Markov Chain whose realizations are unobserved. We characterize the two regimes and the law that governs the transition between them. The parameter estimates can then be used to infer which regime the process is in and provide forecasts for the future value of the series.

Let \( Y_t \) denote the spot exchange rate at time \( t \). We define the exchange rate change as 
\[
  dy_t = 100(\ln Y_t - \ln Y_{t-1}),
\]
so that \( dy_t \) is the percentage appreciation from time \( t-1 \) to time \( t \). The model then postulates the existence of an unobserved variable \( (s_t) \) that may be equal to one or two. The variable characterizes the “state” (or regime) that the process is in at time \( t \). When \( s_t = 1 \), the observed change in the exchange rate \( (dy_t) \) is distributed \( N(\mu_1, \sigma^2_1) \), whereas when \( s_t = 2 \), \( dy_t \) is distributed \( N(\mu_2, \sigma^2_2) \). Thus, the “long swings” model postulates that:
\[
  dy_t = \mu_{s_t} + \varepsilon_{s_t}
\]
where the unobserved state variable \( (s_t) \) is governed by the following transition probabilities:
\[
  \Pr[s_t = 1 | s_{t-1}] = p_{11}
\]
\(^{13}\) Contrary to Klaassen (2004), we do not impose a GARCH structure because we try to keep the model as parsimonious as possible. MS-GARCH model can only be computed using approximate Maximum Likelihood methods, which might often result in a local extrema.
This is a simple version of the more general MS-RW model. It is especially suitable for explaining the exchange rate behavior. Large values of $p_{ii}$ generate the “long swings” which seem to characterize the U.S. dollar exchange rates. However, they are not imposed by the model because nothing forbids $\mu_1$ from being equal to $\mu_2$. In addition, the model allows for asymmetry in the persistence of the two regimes – upward moves could be short but sharp (large and positive $\mu_1$, small $p_{11}$) whereas downward moves could be gradual and drawn out (negative and small in absolute value $\mu_2$, large $p_{22}$). Finally, the parsimonious parametrization (only six parameters need to be estimated for each exchange rate series) promises good forecasting properties of the model.

The parameter vector $\theta = \{\mu_1, \mu_2, \sigma_1, \sigma_2, p_{11}, p_{22}\}$ can be estimated by using a simple maximum likelihood estimation. The sample likelihood is a function of the observed values of the changes in the exchange rates. The states, $s_1$ and $s_2$, are unobserved and the inferences about their probabilities are based on the observed data. The maximum likelihood estimation in this paper is performed using the algorithm described by Hamilton (1989).

### 2.4. A test for the Markov Switching specification

We next investigate whether a two state MS-RW model provides a better representation than a one-state model (RW with drift). To formally test for the existence of a 2-state Markov process in the exchange rate dynamics, we employ the Garcia (1998) test. We test the null hypothesis of a RW with drift against the alternative hypothesis of a MS-RW with drift. The rejection of the single state null in favor of a two-state process is sufficient to justify the use of a MS-RW model for forecasting. (Refer to Appendix A for a more detailed explanation of the Garcia test). Garcia’s (1998) test results for all countries are presented in Table 2. We are able to identify two distinct states for all countries. The estimates are generally characterized by high values for transition probabilities, which are also statistically different from zero. The values for $p_{11}$ and $p_{22}$ range from 0.707 for Switzerland to 0.997 for Canada. The only exception is Portugal, where the

\[ \Pr(s_t = 2 \mid s_{t-1}) = p_{22} \]

In a more general Markov Switching model the exchange rate could be allowed to follow a more general stochastic process, there can be more than two states, or the exchange rate could be modeled as a multivariate Markov switching process. The simple univariate model described above was found by Engel and Hamilton to provide a good description of the exchange rate behavior.
transition probability, $p_{11}$, is 0.38, significant only at the 10% level. We obtain the highest value of the sup F statistic, 81.60, in the case of Australia, while the lowest sup F statistics, 21.62, belongs to Germany. The bootstrapped 1%, 5%, and 10% critical values for the sup F statistics are 18.14, 13.56, and 11.57, respectively. Thus we are able to reject the null hypothesis of a RW with drift versus the alternative hypothesis of the MS – RW with drift for all countries at the 1% significance level.

3. Predictability comparison

In this section we consider rolling regressions to obtain multi-step-ahead forecasts from every estimated model.\(^{15}\) We fix the window size at 120 month (10 years). We initially estimate a MS-RW model and a RW model using all observations from the start of the series through 1983:M1 (the first window). We then drop the first data point, add an additional data point at the end of the sample, and re-estimate the model. We repeat the estimation process by adding successive observations through 2007:M1 (or 1997:M12 for the European countries). Next, we compute 1- to 12-month ahead forecasts for our model, from all the forecasting origins, to 2008:M1 (1998:M12 for the European countries).\(^{16}\) Multistep forecasts constructed this way involve overlapping data, so we control for it using the Newey-West variance-covariance matrix adjustment. The forecasts are used to obtain the MSPE of the MS-RW and RW models, for each series at each forecasting horizon.

Using the MS-RW model for forecasting involves a few technical issues: First, the resultant likelihood function is known to be highly non-linear with an estimation resulting in a sub-optimal extrema. Second, the forecasting outcome is known to depend on the specific starting values chosen by a researcher, casting doubts on the robustness of the results. To address these problems, we propose the use of the multiple dynamic starting values (for the former) and what we call the “minimal interference approach” (for the latter). These procedures are presented in detail in Appendices B and C.

\(^{15}\) We use rolling rather than recursive regressions for comparison with the extensive literature following Meese and Rogoff (1983a), and we choose a rolling window of 120 observations to estimate alternative forecast models following the empirical exercise in Clark and West (2006).

\(^{16}\) We define a forecast of exchange rate $h$ steps ahead as a cumulative month-to-month change in its logarithm:

$$dy_{t+h} = 100(ln Y_{t+h} - ln Y_{t-1}).$$
In order to evaluate the out-of-sample performance of the MS-RW versus the RW model, previous literature has used different measures (such as the MSPE and the direction of change statistics). The evaluation of the out-of-sample performance is typically based on the MSPE comparison, using the Diebold and Mariano (1995) and West’ (1996) tests for equal predictability of two non-nested models (DMW). As Clark and McCraken (2001, 2005) and Corradi and Swanson (2007) argue, when comparing the MSPE of two nested models, the DMW procedure, using standard normal critical values, leads to very poorly sized tests. Thus, we apply a recently developed inference procedure for testing the null of equal predictive ability of a linear econometric model and a random walk model proposed by Clark and West (2006, 2007), (CW test). This methodology is preferable to the standard DMW procedure as the null hypothesis of a random walk and the alternative hypothesis of a MS-RW with drift model are nested. The test statistics takes into account that, under the null, the sample MSPE of the alternative model is expected to be greater than that of the RW model and adjusts for the upward shift in the sample MSPE of the alternative model. The simulations in CW (2006) suggest that the inference made using asymptotically normal critical values results in properly-sized tests for rolling regressions.

Table 3 presents the results for the multi-step forecasts for exchange rates. At one step-ahead forecast, our model significantly outperforms the RW model for 6 out of 12 countries at the 5% significance level and for 3 countries at the 10% significance level. In the case of Italy and Australia we do not find any significant results and, in the case of France, the RW model seems to outperform the MS-RW model: the test statistics has a negative sign and the RW model is significant at the 5% level. In the case of Italy and Australia, the MS-RW with drift model outperforms on average the RW model but the difference is not significant.

We next consider multi-step forecasting: At 2 and 3-month horizons, the MS-RW model outperforms the RW model for 7 out of 12 countries at the 10% significance level. We do not find and significant results for France, Italy and Australia (as previously) and Sweden and

---

17 CW test statistics is aimed to test the null that a series follows a martingale difference against the alternative that the series is linearly predictable. However, Clark and West (2006) argue that nonlinear models, specifically the Markov Switching model, are easily accommodated. Nikolsko-Rzhevskyy and Prodan (2009) provide simulations which show that CW is properly sized in the case of the nonlinear models.

18 An alternative strategy, used by Mark (1995) and Kilian (1999), is to calculate bootstrapped critical values for the DMW test to construct an accurately sized test. While this solves the most egregious problems with the application of the DMW test to nested models, the advantage of the CW test is that it has greater power. West (2006) provides a summary of recent literature on asymptotic inference about predicting ability.
Denmark. At 6 and 12 month horizons we find that on average, the MS-RW model outperforms the RW model at the 10% significance level for 4 countries: Japan, Canada (at both, 6 and 12 months), Switzerland (only at 12 months horizon) and Portugal (only at 6 months horizon). A common characteristic for all countries is that the significance of the MS-RW model predictability decreases as the forecast horizon increases.

Using the same set of exchange rates, Molodtsova and Papell (2008) have previously found that a symmetric Taylor rule model with heterogeneous coefficients and interest rate smoothing outperforms the RW model for 10 out of 12 countries at one step horizon, the only exception being Portugal and Sweden. For these countries, where single-state models seem to generally perform poorly, we argue that accounting for a second state leads to significant improvements in the predictive ability: we find evidence of predictability from up to one year horizon (Portugal) to one month horizon (Sweden).

We next examine the predictability of the exchange rates at one step ahead forecasting horizon when varying the forecast window. Figure 2 illustrates our approach to test our results’ robustness to different sample periods: we plot the asymptotic CW p-value on the y-axis and the starting date of the recursion on the x-axis (the first date for which a forecast is calculated). For instance, the CW value associated with April 1983 for a given country implies that the CW p-value is calculated using the forecast window from April 1983 to January 2008 (end of sample). If the CW p-value is below 0.1, we consider the result statistically significant at the 10% significance level. In order for a result to be considered robust, we would expect that the CW p-value is below 0.1 for almost all the plotted forecast windows. The graph shows that for two of the European Union countries (Germany and Netherlands) and for four of the non European Union countries (Canada, Switzerland, Sweden and Denmark) the results are pretty robust. Consequently, for 6 out of 9 countries where we found that the MS–and RW model outperforms the RW model at one step forecasting horizon, the results are robust across various samples. Overall, for the European and non-European Union countries, the MS-RW model performs relatively well in the 1980s, but its performance deteriorates for some of the countries in the 1990’s.

---

19 "Forecast window" refers to the part of the sample for which forecasts are calculated. For example, if we have a sample of 120 months and the first forecast is based on 30 months, then the forecast window is 90 months.

20 Similarly, when splitting the data into pre- and post-Euro introduction subsamples (January 1st, 1999), we find evidence that the predictability of the exchange rates declines after the introduction of the Euro in January 1999. The results are not reported in this paper.
4. Robustness check

Rogoff and Stavrakeva (2009) examine the most popular exchange rate forecasting structural models and specifications (Gourinchas and Rey (2007), Engel, Mark and West (2007) and Molodtsova and Papell (2008)) and conclude that one of the sources of the overly optimistic results is the failure to check robustness with respect to alternative out-of-sample test statistics. They argue that the new asymptotic out-of-sample tests such as the CW are easy to use but bootstrapped out-of-sample tests remain more powerful and better sized. Furthermore, they argue that all of the structural models and specifications reviewed fail to produce robust forecasts over different sample periods (with Gourinchas and Rey (2007) performing somewhat better than the rest of the specifications considered).

Following Rogoff and Stavrakeva’s (2009) critique, we check the robustness of our one-step ahead forecasting results over several dimensions. First, we consider alternative out-of-sample test statistics as the CM, DMW, TU and the sign test. Second, instead of relying on asymptotic results, we bootstrap critical values for all tests. Finally we explore the robustness of our results with respect to different forecast windows, using the TU test. All of the above are done for a one step forecasting horizon.

We first bootstrap critical values for the CW test. Using Germany as a representative Euro country, and Canada as a representative non-Euro country, we perform a non-parametric wild bootstrap. This method is based on re-sampling residuals: we generate the data by drawing residuals, with replacement, from the (demeaned) exchange rate series for the representative country and randomly multiply each value by either +1 or -1. We then re-fit the model, and record the statistics of interest. We repeat these steps 1,000 times. The results are presented in Table 4. The difference between our previous results using asymptotic critical values and the new results using bootstrapped critical values is insignificant; at the one step-ahead forecast horizon our model significantly outperforms the RW model for 8 out of 12 countries at the 5% significance level and for one country at the 10% significance level.

We consider four alternative test statistics: CM, DMW, TU and the sign test. Similar to Clark and West (2007), CM encompassing test is well suited for nested models. DMW and the
TU are designed to test the out-of-sample forecasting performance of non-nested models.\textsuperscript{21} In addition, we use the sign test as a comparison measure with previous studies that attempt to forecast nominal exchange rates using a MS-RW model. For a detailed description of how to calculate each test statistics and how to determine the statistical significance see Appendix D.

In Table 4 we report the bootstrapped p-values of the CM, DMW, TU and the sign test. Comparing our previous results (CW test) with the bootstrapped CM test, we observe an increase in the statistical significance: at one step-ahead forecast horizon, our model significantly outperforms the RW model for 10 out of 12 countries at the 5\% significance level (we add Italy to our previous evidence of out-of-sample forecasting). On the other hand, if we examine the statistical significance via the bootstrapped DMW and TU test statistics, we observe a decrease in the significance level: at one step-ahead forecasting horizon our model significantly outperforms the RW model for 3 out of 12 countries at the 10\% significance level (Canada, Germany and Netherlands, countries where we have previously found evidence of predictability with the CW test). In order to compare our results with the previous literature we consider the sign test. At one step-ahead forecast horizon our model significantly outperforms the RW model for 3 out of 12 countries at the 5\% significance level and for one country at the 10\% significance level.

We next examine the predictability of the exchange rates at one step ahead forecasting horizon, when varying the forecast window, using the bootstrapped TU test. Figure 3 illustrates our approach to test the robustness to different sample periods: we plot the TU p-value on the y-axis and the starting date of the recursion on the x-axis (the first date for which a forecast is calculated.)\textsuperscript{22}

The graph shows that for two of the European Union countries (Germany and Netherlands) and for 2 of the non European Union countries (Canada and Switzerland) the results are quite robust. Consequently, from the 9 countries where we found that the two-state MS-RW model outperforms the RW model at one step forecasting horizon, we find that for 4 countries the results are robust across the samples, using the bootstrapped TU test.

\textsuperscript{21} As argued by Rogoff and Stavrakeva’s (2008), one of the main problems related to using the new tests for nested models (CW and CM) as the main and only out-of-sample test statistics relates to the fact that they cannot be interpreted as minimum MSFE tests such as the TU and the DMW. On the other hand, the DMW and TU tests are known to have low power when applied to nested models. Thus, using the battery of tests instead of any single alternative might be beneficial.

\textsuperscript{22} This experiment is similar to the previous experiment of testing the robustness of the CW test to different sample periods.
5. Conclusion

There is a large literature, mostly based on macroeconomic fundamentals, that attempts to explain the movements in the exchange rate series. However, there has been mixed evidence regarding the forecasting performance of different models when modeling the nominal exchange rate. The use of various specifications of the Taylor rules and the recently developed inference procedure for testing the null of equal predictive ability proposed by Clark and West (2006, 2007) represent a step forward in assessing the predictability of the exchange rate models.

The most common route that researchers have taken when forecasting exchange rates is to design more complicated economic models in an attempt to outperform the random walk model. We argue that the relative success attributed to the exchange rates forecasting power of the macroeconomic fundamentals models is in part due to the drifting term and that a two-state MS-RW with drift model is a good representation for the nominal exchange rate. We use a modified version of the Engel and Hamilton (1990) model and nominal exchange rates for 12 OECD countries versus the US dollar from March 1973 to January 2008. In order to evaluate the out-of-sample performance of the Markov switching versus the random walk model based on the mean square prediction comparison, we apply the recently developed inference procedure proposed by Clark and West (2006, 2007) for testing the null of equal predictive ability of two nested models. We provide strong evidence of both short-term and long-term predictability. The results show that the model strongly outperforms the random walk for 9 out of 12 countries for one-month-ahead forecasts. In addition, we find strong evidence of long-horizon predictability (more than one month horizon): up to 3 months horizon for 7 countries, up to 6 months horizon for 4 countries and up to 12 months horizon for 3 countries. A common characteristic for all countries is that the significance of our model’s predictability decreases as the forecast horizon increases. Our results remain robust to different forecast samples for 6 out of 9 countries.

We address Rogoff and Stavrakeva’s (2008) critique and check the robustness of our short-horizon results. They remain consistent when bootstrapped Clark and West or Clark and McCracken tests statistics are employed. The significance of our results decreases when using the bootstrapped Diebold-Mariano/West our Theil’s U test, while remaining superior to the previous studies.
The above results are important as previous literature has failed to find any significant evidence that the MS-RW with drift model outperforms the random walk model, at a short or long horizon. Furthermore our evidence of predictability remains robust to alternative out-of-sample test statistics and forecast windows, significantly surpassing the previous literature results, using economic models.
References


West, Kenneth D., “Asymptotic Inference about Predictive Ability” Econometrica, 1996, 64, pp. 1067-1084
Table 1. The out-of-sample comparisons of the UIRP and the RW with drift vs the RW null models

<table>
<thead>
<tr>
<th></th>
<th>UIRP model vs RW, Clark and West (2006)</th>
<th>UIRP model vs RW, Full sample</th>
<th>UIRP w/o const vs RW, Full sample</th>
<th>RW with drift vs RW, Full sample</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Japan</strong></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
</tr>
<tr>
<td></td>
<td>1.24†</td>
<td>1.38*</td>
<td>-0.32</td>
<td>1.41*</td>
</tr>
<tr>
<td><strong>Canada</strong></td>
<td>1.78**</td>
<td>1.92**</td>
<td>1.39*</td>
<td>0.94†</td>
</tr>
<tr>
<td><strong>Switzerland</strong></td>
<td>1.88**</td>
<td>2.03***</td>
<td>0.48</td>
<td>0.68</td>
</tr>
<tr>
<td><strong>U.K.</strong></td>
<td>0.03</td>
<td>-0.13</td>
<td>0.68</td>
<td>-0.26</td>
</tr>
</tbody>
</table>

Table 2. Full sample MS-RW estimation results and test statistics

<table>
<thead>
<tr>
<th></th>
<th>$\mu_1$</th>
<th>$\mu_2$</th>
<th>$\sigma_1$</th>
<th>$\sigma_2$</th>
<th>$P_{11}$</th>
<th>$P_{22}$</th>
<th>Garcia test</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Japan</strong></td>
<td>2.893***</td>
<td>-0.58***</td>
<td>2.855***</td>
<td>2.112***</td>
<td>0.914***</td>
<td>0.713***</td>
<td>46.47***</td>
</tr>
<tr>
<td></td>
<td>(0.63)</td>
<td>(0.20)</td>
<td>(0.27)</td>
<td>(0.11)</td>
<td>(0.03)</td>
<td>(0.09)</td>
<td></td>
</tr>
<tr>
<td><strong>Canada</strong></td>
<td>0.661***</td>
<td>-0.126***</td>
<td>1.875***</td>
<td>-1.039***</td>
<td>0.997***</td>
<td>0.993***</td>
<td>44.81***</td>
</tr>
<tr>
<td></td>
<td>(0.23)</td>
<td>(0.05)</td>
<td>(0.16)</td>
<td>(0.03)</td>
<td>(0.00)</td>
<td>(0.01)</td>
<td></td>
</tr>
<tr>
<td><strong>Switzerland</strong></td>
<td>2.571***</td>
<td>-1.111***</td>
<td>2.485***</td>
<td>2.216***</td>
<td>0.823***</td>
<td>0.707***</td>
<td>31.65***</td>
</tr>
<tr>
<td></td>
<td>(0.52)</td>
<td>(0.28)</td>
<td>(0.20)</td>
<td>(0.13)</td>
<td>(0.04)</td>
<td>(0.08)</td>
<td></td>
</tr>
<tr>
<td><strong>UK</strong></td>
<td>0.185</td>
<td>-0.389</td>
<td>1.783***</td>
<td>3.054***</td>
<td>0.976***</td>
<td>0.985***</td>
<td>36.73***</td>
</tr>
<tr>
<td></td>
<td>(0.15)</td>
<td>(0.28)</td>
<td>(0.10)</td>
<td>(0.23)</td>
<td>(0.02)</td>
<td>(0.01)</td>
<td></td>
</tr>
<tr>
<td><strong>France</strong></td>
<td>1.036***</td>
<td>-1.553*</td>
<td>2.155***</td>
<td>2.531***</td>
<td>0.879***</td>
<td>0.912***</td>
<td>23.59***</td>
</tr>
<tr>
<td></td>
<td>(0.27)</td>
<td>(0.87)</td>
<td>(0.14)</td>
<td>(0.46)</td>
<td>(0.16)</td>
<td>(0.08)</td>
<td></td>
</tr>
<tr>
<td><strong>Germany</strong></td>
<td>1.745***</td>
<td>-1.547***</td>
<td>2.324***</td>
<td>2.142***</td>
<td>0.782***</td>
<td>0.799***</td>
<td>21.62***</td>
</tr>
<tr>
<td></td>
<td>(0.43)</td>
<td>(0.40)</td>
<td>(0.17)</td>
<td>(0.17)</td>
<td>(0.06)</td>
<td>(0.07)</td>
<td></td>
</tr>
<tr>
<td><strong>Italy</strong></td>
<td>0.255*</td>
<td>-0.487***</td>
<td>0.813***</td>
<td>2.806***</td>
<td>0.980***</td>
<td>0.910***</td>
<td>45.64***</td>
</tr>
<tr>
<td></td>
<td>(0.14)</td>
<td>(0.19)</td>
<td>(0.18)</td>
<td>(0.14)</td>
<td>(0.01)</td>
<td>(0.04)</td>
<td></td>
</tr>
<tr>
<td><strong>Sweden</strong></td>
<td>0.491</td>
<td>-2.521</td>
<td>1.997***</td>
<td>2.944***</td>
<td>0.759***</td>
<td>0.945***</td>
<td>50.57***</td>
</tr>
<tr>
<td></td>
<td>(0.49)</td>
<td>(2.34)</td>
<td>(0.22)</td>
<td>(0.54)</td>
<td>(0.31)</td>
<td>(0.03)</td>
<td></td>
</tr>
<tr>
<td><strong>Australia</strong></td>
<td>0.059</td>
<td>-0.229</td>
<td>1.172***</td>
<td>2.989***</td>
<td>0.946***</td>
<td>0.936***</td>
<td>81.60***</td>
</tr>
<tr>
<td></td>
<td>0.13</td>
<td>0.21</td>
<td>0.11</td>
<td>0.17</td>
<td>0.02</td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td><strong>Denmark</strong></td>
<td>1.477***</td>
<td>-1.741***</td>
<td>2.089***</td>
<td>1.906***</td>
<td>0.759***</td>
<td>0.812***</td>
<td>35.06***</td>
</tr>
<tr>
<td></td>
<td>(0.31)</td>
<td>(0.32)</td>
<td>(0.12)</td>
<td>(0.15)</td>
<td>(0.05)</td>
<td>(0.05)</td>
<td></td>
</tr>
<tr>
<td><strong>Netherlands</strong></td>
<td>1.735***</td>
<td>-1.569***</td>
<td>2.172***</td>
<td>2.116***</td>
<td>0.782***</td>
<td>0.792***</td>
<td>23.82***</td>
</tr>
<tr>
<td></td>
<td>(0.49)</td>
<td>(0.41)</td>
<td>(0.18)</td>
<td>(0.17)</td>
<td>(0.06)</td>
<td>(0.08)</td>
<td></td>
</tr>
<tr>
<td><strong>Portugal</strong></td>
<td>-0.337**</td>
<td>-5.963**</td>
<td>2.279***</td>
<td>4.511***</td>
<td>0.38*</td>
<td>0.966***</td>
<td>44.49***</td>
</tr>
<tr>
<td></td>
<td>(0.16)</td>
<td>(2.67)</td>
<td>(0.12)</td>
<td>(1.05)</td>
<td>(0.19)</td>
<td>(0.02)</td>
<td></td>
</tr>
</tbody>
</table>

Notes: Standard errors are in parenthesis. *, **, and *** indicate significance at 10, 5, and 1 percent level, respectively. Garcia test statistics for each country is obtained using 400 (n=20) grid points over uniformly distributed $p_{11}$ and $p_{22}$ values (see Garcia 1998 for details). Bootstrapped critical values for a representative country (Germany) were obtained using 1,000 iterations. Corresponding one-sided critical values for 1%, 5%, and 10% are 19.85, 15.77, and 10.11, respectively.
Table 3. Multistep MS-RW forecasting results for the full sample

<table>
<thead>
<tr>
<th>Forecast Horizon $h$</th>
<th>CW statistics/p-value</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>6</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Japan</strong></td>
<td></td>
<td>1.908**</td>
<td>2.053**</td>
<td>2.107**</td>
<td>1.208*</td>
<td>2.031**</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.03</td>
<td>0.02</td>
<td>0.002</td>
<td>0.10</td>
<td>0.02</td>
</tr>
<tr>
<td><strong>Canada</strong></td>
<td></td>
<td>2.225*</td>
<td>1.500*</td>
<td>1.355*</td>
<td>1.831**</td>
<td>1.546*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.01</td>
<td>0.07</td>
<td>0.09</td>
<td>0.03</td>
<td>0.06</td>
</tr>
<tr>
<td><strong>Switzerland</strong></td>
<td></td>
<td>2.483*</td>
<td>1.905**</td>
<td>1.605**</td>
<td>0.998</td>
<td>2.284***</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.01</td>
<td>0.03</td>
<td>0.005</td>
<td>0.016</td>
<td>0.01</td>
</tr>
<tr>
<td><strong>UK</strong></td>
<td></td>
<td>1.846**</td>
<td>2.032**</td>
<td>1.406*</td>
<td>0.788</td>
<td>-0.313</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.03</td>
<td>0.02</td>
<td>0.08</td>
<td>0.022</td>
<td>0.62</td>
</tr>
<tr>
<td><strong>France</strong></td>
<td></td>
<td>-1.611</td>
<td>-2.283</td>
<td>-2.132</td>
<td>-2.058</td>
<td>-1.401</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.95</td>
<td>0.99</td>
<td>0.98</td>
<td>0.98</td>
<td>0.92</td>
</tr>
<tr>
<td><strong>Germany</strong></td>
<td></td>
<td>2.486***</td>
<td>1.889**</td>
<td>1.624**</td>
<td>0.748</td>
<td>0.487</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.01</td>
<td>0.03</td>
<td>0.005</td>
<td>0.023</td>
<td>0.31</td>
</tr>
<tr>
<td><strong>Italy</strong></td>
<td></td>
<td>1.065</td>
<td>0.042</td>
<td>0.003</td>
<td>-0.712</td>
<td>-1.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.14</td>
<td>0.48</td>
<td>0.50</td>
<td>0.76</td>
<td>0.84</td>
</tr>
<tr>
<td><strong>Sweden</strong></td>
<td></td>
<td>1.261*</td>
<td>0.834</td>
<td>0.244</td>
<td>0.485</td>
<td>0.760</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.10</td>
<td>0.20</td>
<td>0.40</td>
<td>0.31</td>
<td>0.22</td>
</tr>
<tr>
<td><strong>Australia</strong></td>
<td></td>
<td>0.418</td>
<td>0.078</td>
<td>0.067</td>
<td>0.166</td>
<td>0.297</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.34</td>
<td>0.47</td>
<td>0.47</td>
<td>0.43</td>
<td>0.38</td>
</tr>
<tr>
<td><strong>Denmark</strong></td>
<td></td>
<td>1.743**</td>
<td>0.918</td>
<td>0.515</td>
<td>0.092</td>
<td>-0.598</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.04</td>
<td>0.18</td>
<td>0.30</td>
<td>0.46</td>
<td>0.73</td>
</tr>
<tr>
<td><strong>Netherlands</strong></td>
<td></td>
<td>3.01***</td>
<td>1.833**</td>
<td>1.47*</td>
<td>-0.390</td>
<td>-0.576</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.00</td>
<td>0.03</td>
<td>0.07</td>
<td>0.65</td>
<td>0.72</td>
</tr>
<tr>
<td><strong>Portugal</strong></td>
<td></td>
<td>1.703**</td>
<td>1.397*</td>
<td>1.284*</td>
<td>1.364*</td>
<td>0.739</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.04</td>
<td>0.08</td>
<td>0.10</td>
<td>0.09</td>
<td>0.23</td>
</tr>
</tbody>
</table>

Notes: CW stands for Clark and West (2006) statistics. Critical values are bootstrapped using n=1,000 repetitions. For Japan, Canada, Switzerland, UK, Sweden, Australia, and Denmark, the full sample corresponds to 1973:3 – 2008:1. For France, Germany, Italy, Netherlands, and Portugal, the full sample corresponds to 1973:3 – 1998:12. For $h>1$, the CW statistic is calculated using Newley-West adjustment. Forecasting is done using rolling window of 120 months. Standard errors are in parenthesis. *, ***, and *** indicate significance at 10, 5, and 1 percent level, respectively.
Table 4. Robustness With Respect to Alternative Test Statistics

<table>
<thead>
<tr>
<th>Country</th>
<th>CW</th>
<th>P-value</th>
<th>CM</th>
<th>P-value</th>
<th>DM</th>
<th>p-value</th>
<th>TU-test</th>
<th>p-value</th>
<th>Sign-test</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Japan</td>
<td>1.908</td>
<td>0.035</td>
<td>12.540</td>
<td>0.008</td>
<td>-2.038</td>
<td>0.637</td>
<td>1.052</td>
<td>0.897</td>
<td>2.194</td>
<td>0.007</td>
</tr>
<tr>
<td>Canada</td>
<td>2.225</td>
<td>0.017</td>
<td>17.990</td>
<td>0.001</td>
<td>0.127</td>
<td>0.050</td>
<td>0.997</td>
<td>0.040</td>
<td>0.462</td>
<td>0.206</td>
</tr>
<tr>
<td>Switzerland</td>
<td>2.483</td>
<td>0.015</td>
<td>13.340</td>
<td>0.007</td>
<td>-0.520</td>
<td>0.152</td>
<td>1.010</td>
<td>0.325</td>
<td>3.233</td>
<td>0.002</td>
</tr>
<tr>
<td>UK</td>
<td>1.846</td>
<td>0.035</td>
<td>10.890</td>
<td>0.015</td>
<td>-0.394</td>
<td>0.112</td>
<td>1.008</td>
<td>0.270</td>
<td>1.501</td>
<td>0.062</td>
</tr>
<tr>
<td>France</td>
<td>-1.611</td>
<td>0.917</td>
<td>-6.391</td>
<td>0.993</td>
<td>-3.761</td>
<td>0.968</td>
<td>1.092</td>
<td>0.977</td>
<td>-2.255</td>
<td>0.990</td>
</tr>
<tr>
<td>Germany</td>
<td>2.486</td>
<td>0.001</td>
<td>10.200</td>
<td>0.001</td>
<td>0.361</td>
<td>0.050</td>
<td>0.992</td>
<td>0.026</td>
<td>0.800</td>
<td>0.209</td>
</tr>
<tr>
<td>Italy</td>
<td>1.065</td>
<td>0.137</td>
<td>5.383</td>
<td>0.029</td>
<td>-1.781</td>
<td>0.652</td>
<td>1.053</td>
<td>0.943</td>
<td>-0.509</td>
<td>0.646</td>
</tr>
<tr>
<td>Sweden</td>
<td>1.261</td>
<td>0.096</td>
<td>11.730</td>
<td>0.010</td>
<td>-3.703</td>
<td>0.920</td>
<td>1.151</td>
<td>0.980</td>
<td>0.808</td>
<td>0.168</td>
</tr>
<tr>
<td>Australia</td>
<td>0.419</td>
<td>0.322</td>
<td>2.713</td>
<td>0.187</td>
<td>-3.013</td>
<td>0.842</td>
<td>1.085</td>
<td>0.955</td>
<td>-0.577</td>
<td>0.691</td>
</tr>
<tr>
<td>Denmark</td>
<td>1.743</td>
<td>0.047</td>
<td>9.316</td>
<td>0.022</td>
<td>-0.845</td>
<td>0.222</td>
<td>1.015</td>
<td>0.520</td>
<td>1.039</td>
<td>0.107</td>
</tr>
<tr>
<td>Netherlands</td>
<td>3.010</td>
<td>0.001</td>
<td>12.260</td>
<td>0.001</td>
<td>0.902</td>
<td>0.008</td>
<td>0.981</td>
<td>0.001</td>
<td>1.818</td>
<td>0.027</td>
</tr>
<tr>
<td>Portugal</td>
<td>1.703</td>
<td>0.036</td>
<td>11.960</td>
<td>0.001</td>
<td>-1.461</td>
<td>0.545</td>
<td>1.061</td>
<td>0.960</td>
<td>1.237</td>
<td>0.130</td>
</tr>
</tbody>
</table>

Notes: CW stands for Clark and West (2006) statistics, CM for Clark and McCraken (2005) encompassing test, DM for Diebold-Mariano/West (1995) test. Critical values are bootstrapped using n = 1,000 replications. For Japan, Canada, Switzerland, UK, Sweden, Australia, and Denmark, the full sample corresponds to 1973:3 – 2008:1. For France, Germany, Italy, Netherlands, and Portugal, the full sample corresponds to 1973:3 – 1998:12. Forecasting is done using rolling window of 120 months. Standard errors are in parenthesis. *, ***, and *** indicate significance at 10, 5, and 1 percent level, respectively.
Figure 1. Long swings: Canada and Germany

Note: Germany is used as a representative Euro country, and Canada is used as a representative non-Euro country. Both exchange rates are expressed as foreign currency per 1 USD.
Figure 2. Significance of the Clark and West test for different forecast windows

Euro countries

Non-Euro countries

Notes: For interpretation of the figures, refer to Section 3.
Figure 3: Significance of the Theil-U test for different forecast windows

Euro Countries

Non-Euro countries

Notes: For interpretation of the figures, refer to Section 3.
Appendix A: Garcia (1998) test

The test treats the transition probability parameters \( p_{11} \) and \( p_{22} \) as nuisance parameters and the remaining parameters are treated as parameters of interest. We test the null hypothesis of a random walk with drift model against the alternative hypothesis of a 2-state Markov-switching random walk with drift. Alternatively, if we refer to equation 2, we test the null versus the alternative hypothesis:

\[
H_0: \mu_1 = \mu_2, \sigma_1 = \sigma_2, p_{11} \text{ and } p_{22} \text{ are not identified}
\]

\[
H_1: \mu_1, \mu_2, \sigma_1, \sigma_2, \text{ and } p_{11}, p_{22} \text{ are identified}
\]

The test looks at the Sup F statistics of the difference between the likelihood of a single state and a two-state Markov switching models with transition probabilities running over a two-dimensional uniform grid of values, distributed between 0 and 1. We use a 20x20 grid, thus maximizing the F-statistics over 400 uniformly distributed \( p_{11} \) and \( p_{22} \) values (see Garcia 1998 for details). Bootstrap critical values are calculated as follows: we assume that the underlying process follows a random walk with drift, estimate the parameters in the case of Germany (as a representative country for the Euro country) and Canada (as a representative non-Euro country) and use this model to construct a pseudo-samples of size equal to the actual size of the data. We then calculate the bootstrap parameter estimates and compute the statistics of interest. The one-sided critical values for 1%, 5% and 10% are taken from the sorted vector of 1000 replicated statistics.

Appendix B: Multiple starting values and the minimal intervention approach

One drawback of the Markov switching model when used for forecasting is the significant non-linearity of the likelihood function, which often leads to suboptimal solutions. A common way to solve this problem is to use numerous starting values (e.g. Hamilton, 1989, Engel and Hamilton, 1990, Engel, 1994) that generally lead to a global maximum of the likelihood function when estimating the full sample or the first forecasting window. However, there is no assurance that this method will perform well in other subsamples when we "move" the forecasting window. One way to correct this is to increase the number of starting values up to several hundreds, but the disadvantage is that the estimation time would increase tremendously, making the moving
window forecasting unfeasible. Besides, any exogenously fixed set of starting values (especially if they are different from country to country) raises a question of the credibility of the results, as it might have been chosen as the one leading to "good" results.

To overcome the above problems we use an approach which minimizes the researcher's impact on the final results, but does not take the estimation beyond feasibility constraints. We allow the algorithm to automatically choose starting values for our estimation, making a dynamic choice, unbiased, data-dependent and unique in each subsample ("window"). In particular, as we move the estimation window to the right, we employ 7 (= 6+1) different (and unique) starting values constructed in the following way: First, we use a simple OLS to estimate the RW with drift model and obtain estimates of the intercept (drift) and the variance of residuals. Probabilistically, these OLS estimates should lie "between" true parameters values of the MS-RW model. We next construct the MS-RW model’s starting values by multiplying each of the OLS estimates by a scalar higher and lower than 1 to obtain starting values for each regime (for example, if we estimate $C_{OLS}=1.5$, we would use $C_0=1$ and $C_1=2.25$ as the starting values for the MS-RW model). In total, we use 6 different "multiplication vectors". Besides these 6 OLS-based starting vectors, we employ the "optimal estimates" starting vector from the previous step. As we move the window (of 120 months) by 1 observation, we re-estimate the model to make a new forecast, the 119 data points remaining the same. Thus, it is expected that "true" parameters should be close for both windows. In total, this gives us 7x300 unique starting values for non-EU country case, and 7x189 for an EU country case, which we use to test our 1-step-ahead forecasting models.

As past parameters estimates are not available for the very first window, we use a different technique: we use OLS to obtain single-state estimates of drift and variance, and then let the algorithm "shake" those estimates randomly 100 times, thus providing us with a hundred starting values. Therefore, besides pre-specifying 6 "multiplication vectors" (which are the same for all samples and all countries), our forecasting approach is completely "researcher-independent."
Appendix C: Constructing a one-step and a multi-step forecast.

For a one-step-ahead forecast, we estimate a sample for \( t \) observations and forecast the \((t+1)\) observation. First, we record the probability of the system to be in state 1 at time \( t \), \( s(t) \). Thus, \( s(t)p \) is the probability that the system is in state 1 at time \( t+1 \), and \( s(t)(1-p) \) is the probability that the system moves to state 0. Similarly, \((1-s(t))q \) is the probability that the system is in state 0 at time \( t+1 \), and \((1-s(t))(1-q) \) is the probability that the system is in state 1 at time \( t+1 \). (Note that \( p \) and \( q \) are conditional probabilities of being in state 1 and 0, respectively). Therefore, if \( c_i \) is the estimate of the drift in state \( i \), our forecast for \( t+1 \) takes the following form:

\[
y_f = \{s(t)(1-p) + (1-s(t))q\}c_0 + \{s(t)p + (1-s(t))(1-q)\}c_1,
\]

where the multipliers of the \( c_i \) term are unconditional probabilities of the system being in state 0 and 1, respectively, at time \( t+1 \).

For an \( h \)-step ahead forecast, defined as a percentage difference between the exchange rate at \((t+h)\) and that at \((t)\), we multiply the one-step ahead forecast, defined above, by the number of steps, \( h \). This is the result we would formally obtain if we were doing “direct forecasting” by estimating equation (1) with the leads of the one-step-ahead forecasts as the LHS variable and then summing up all of the forecasts.

Appendix D: Minimum MSFE Out-of-Sample Test Statistics

The Theil’s U-Test (TU)

If we define the sample forecast errors from the model under \( H_0 \) and \( H_1 \) as \( \hat{\epsilon}_{RW,t+1} = y_{t+1} \) and \( \hat{\epsilon}_{MS,t+1} = y_{t+1} - \hat{y}_{MS,t+1} \), respectively, then the TU test statistic is defined as

\[
TU = \sqrt{MSPE^{MS}/MSPE^{RW}}, \quad \text{where} \quad MSPE^i = P^{-1}\Sigma_{i=T-p+1}^T e_{i,t+1}^2
\]

“P” is the size of the mowing window, “Markov Switching” is defined as “MS” and “Random Walk” is defined as “RW”. A \( TU < 1 \) implies that our model outperforms the random walk model.
Diebold-Mariano/West (DMW) test
If we define the forecasting errors as $\hat{f}_t = e_{RW,t}^2 - e_{MS,t}^2$, the mean value as $\bar{f} = MSPE_{RW} - MSPE_{MS}$ and the variance as $V = P^{-1}\sum_{t=T-P+1}(\hat{f} - \bar{f})^2$, then the statistic is defined as $DMW = \bar{f} / \sqrt{P^{-1}V}$. A DMW$>0$ implies that our model achieves a lower MSPE than the RW.

Clark and West (CW) test
Clark and West (2006) propose an adjusted DMW statistic by correcting for the non-zero expected difference between the two MSPEs. If we define the forecasting errors as $\hat{f}_{t+1}^{ADJ} = e_{RW,t+1}^2 - [e_{MS,t+1}^2 - \hat{y}_{MS,t+1}^2]$, the mean value as $\bar{f}^{ADJ} = MSPE_{RW} - [MSPE_{MS} - P^{-1}\sum_{t=T-P+1}\hat{y}_{MS,t+1}^2]$ and the variance as $V^{ADJ} = P^{-1}\sum_{t=T-P+1}(\hat{f}^{ADJ} - \bar{f}^{ADJ})^2$, then the test statistic is defined as $CW = \bar{f}^{ADJ} / \sqrt{P^{-1}V^{ADJ}}$. A CW$>0$ implies that our model outperforms the RW.

Clark and McCracken (CM) test
Using the definitions above, the CM statistic is defined as $CM = \frac{P\sum_{t=T-P+1}^{T}y_{t+1}(\hat{y}_{MS,t+1} - \hat{y}_{MS,t+1})}{\sum_{t=T-P+1}^{T}y_{t+1}^2 - \hat{y}_{MS,t+1}^2}$. A positive value of the statistic implies that our model outperforms the RW.

Direction of change (Sign test)
If $1(\cdot)$ is an index function that equals $1$ is the variable is positive, and $0$ otherwise, then we define the sign statistic as $SIGNT = P^{-1}\{\sum_{t=T-P+1}^{T}[1(y_{t+1})l(\hat{y}_{MS,t+1})] + [1-l(y_{t+1})][1-l(\hat{y}_{MS,t+1})]\}$. A positive value of the statistic implies that our model predicts the direction of change better than the RW.