Evaluating non-linear models on *point* and *interval* forecasts: an application with exchange rates

GIANNA BOERO and EMANUELA MARROCU

1. Introduction

Since the adoption of the floating-rates regime in 1973 numerous efforts have been made to understand exchange rate dynamics. By providing evidence on the superiority of the random walk forecasts, the seminal paper by Meese and Rogoff (1983) gave rise to a long series of papers aimed at proving the superiority of exchange rate determination models based on economic theory. The evidence so far is mixed, some authors finding that some simple specifications including prices, money supplies and output as fundamentals are able to improve forecast accuracy (among others Mark and Sul 2001), while in other papers the Meese and Rogoff's results find new support. Recently attention has focused on the relevance of non-linear dependence in the first and second moments of exchange rate log-differences (Meese and Rose 1991); the presence of non-linearity features might well have important implications in terms of model adequacy, predictability and market efficiency.

The analysis presented in this paper is an attempt to contribute to this avenue of research by exploiting recent developments in non-linear time series econometrics and in forecasting evaluation methods within a univariate framework. In the context of univariate models the most commonly applied non-linear models are the GARCH

□ Università degli Studi di Cagliari, Dipartimento di Ricerche Economiche e Sociali, Cagliari (Italy); e-mail:boero@unica.it;
Università degli Studi di Cagliari, CRENoS, Cagliari (Italy); e-mail: emarrocu@unica.it

BNL Quarterly Review, vol. LVIII, no. 232, March 2005, pp. 91-120.

(generalised autoregressive conditional heteroscedastic) and the SETAR (self-exciting threshold autoregressive) models, which have proved successful in describing the dynamic behaviour of many economic and financial variables. With the GARCH models it is possible to specify the process governing both the mean and the variance of the series, and they are particularly suitable to describe the typical behaviour of financial time series, namely the fact that large (small) price changes tend to be followed by large (small) price changes of either sign; this kind of dependency can be exploited to improve interval forecasts. The SETAR models represent a stochastic process generated by the alternation of different regimes. This class of model has been used with impressive success to forecast certain natural phenomena, such as Canadian lynx data and Wolf's sunspot numbers (Tong 1995); they have also provided significant gains in forecasting economic and financial variables; reference here is, among others, to Kräger and Kugler (1993), Peel and Speight (1994), Tiao and Tsay (1994), Potter (1995) and Clements and Smith (1999).

Related models are the Markov-switching autoregressive (MS-AR) model and the Smooth transition autoregressive (STAR) model, which have also received great attention in the empirical literature on non-linearity in exchange rate movements. The main feature of the MS-AR model is that the switch between regimes is entirely governed by an unobservable variable (for application to exchange rate dynamics see the well-known studies by Hamilton 1989, Engel and Hamilton 1990, Engel 1994, and the very recent application by Clarida *et al.* 2003). The STAR model, on the other hand, is a variant of the SETAR model and can be obtained when the parameters are allowed to change smoothly over time (Granger and Teräsvirta 1993). In this study we limit our investigation to the case of SETAR models in order to facilitate comparison with the existing empirical literature on univariate exchange rate dynamics. Moreover, recent findings (see Clements *et al.* 2003) suggest that the MS-AR and the STAR model provide qualitative similar results to those obtained from SETAR specifications.

Although there have been extensive applications of new techniques to describe the non-linearities and asymmetries which characterise exchange rate dynamics, there are still few studies on the forecasting performance of the different models for historical time series data. Comparisons have been carried out typically with respect to the random walk model or, more recently, by means of simulated data based

on Monte Carlo experiments (see, for example, Clements and Smith 1999). In general, the significant presence of mean-non-linearities for the in-sample period has only rarely provided better out-of-sample forecasts than those obtained from a simple linear or a random walk model. Furthermore, the results are often sensitive to the length of the forecast horizon and to the metric adopted to measure the forecasting accuracy.

Diebold and Nason (1990) suggest four different reasons why non-linear models cannot provide better out-of-sample forecasts than the simpler linear model even when linearity is significantly rejected in-sample: 1) non-linearities concern the even-ordered conditional moments and are therefore not useful for improving forecasts; 2) in-sample non-linearities are due to structural breaks or outliers which cannot be exploited to improve out-of-sample forecasts; 3) conditional means non-linearities are a feature of the DGP but are not large enough to offer better forecasts; 4) non-linearities are present but they are captured by the wrong type of nonlinear model.

Dacco and Satchell (1999) and Clements and Smith (2001) argue that the alleged *poor* forecasting performance of non-linear models can also be due to the evaluation and measurement method adopted. Clements and Smith show, on the basis of a Monte Carlo study, that evaluation of the whole forecast density may reveal gains to the non-linear models which are systematically masked if the comparison is carried out only in terms of *mean square forecast error* (MSFE). This result was confirmed by Boero and Marrocu (2002) in an application with actual data. Dacco and Satchell (1999) suggest that methods based on the profitability criterion should prove more adequate in the case of financial variables.

The aim of the present paper is to compare the forecasting performance of SETAR and GARCH models against the AR benchmark by using weekly log-differences of the Japanese yen and the daily log-differences of the British pound, both quoted against the US dollar. Building on the results in Boero and Marrocu (2000, 2002 and 2003), we pursue the evaluation of alternative models along different lines. We conduct the forecast analysis by using different evaluation criteria based on *point forecasts* and *interval forecasts*. The measure adopted for the evaluation of point forecasts is the MSFE. Interval forecasts are evaluated by means of the Likelihood Ratio (LR) tests of correct conditional coverage, as recently proposed by Christoffersen (1998). The use of interval forecasts is becoming increasingly common in

practical applications, as they provide a description of forecast uncertainty which is not available from point forecasts alone. Models of conditional variance, such as GARCH, are particularly suitable to provide some indication of the uncertainty around the forecast, and when evaluated on interval forecasts they can exhibit accuracy gains which may be systematically masked in MSFE comparisons.

For both point and interval forecast we conduct a multi-period evaluation in order to assess the sensitivity of the results to the selection of the forecast origin and to the specific period considered (Fildes 1992 and Tashman 2000).

The rest of the paper is organised as follows. In section 2 we describe the models adopted. In section 3 we present the statistical properties of the data and the results of the tests performed to detect the presence of non-linearities. The findings from the modelling and forecasting exercises are reported in sections 4 and 5, respectively. Finally, in section 6 we summarise the main results and make some concluding remarks.

2. The models

2.1. *The threshold autoregressive models*

Threshold autoregressive models were first proposed by Tong (1978), Tong and Lim (1980) and Tong (1983). The essential idea of this class of non-linear models is that the behaviour of a process can be described by a finite set of linear autoregressions. The appropriate AR model that generates the value of the time series at each point in time is determined by the relation of a conditioning variable to the threshold values; if the conditioning variable is the dependent variable itself after some delay, *d*, then the model is known as *self-exciting*, hence the acronym SETAR.

The SETAR model is piecewise-linear in the space of the threshold variable, rather than in time. If the process is in the j^{th} regime, the p^{th} order linear autoregression is formally defined as:

$$y_t = \phi_0^{(j)} + \phi_1^{(j)} y_{t-1} + ... + \phi_p^{(j)} y_{t-p} + \varepsilon_t^{(j)} \quad \text{for } r_{j-1} \leq y_{t-d} < r_j \qquad (1)$$

where $\varepsilon_t^{(j)} \sim IID(0,\sigma^{2(j)})$, r_{j-1} *and* r_j are threshold values, p is the lag order and d is the delay parameter.

In order to allow for different autoregressive structures across regimes, p can be seen as the maximum lag order. An interesting feature of SETAR models is that the stationarity of y_t does not require the model to be stationary in each regime; on the contrary, the limit cycle behaviour that this class of models is able to describe arises from the alternation of explosive and contractionary regimes.

In this study we choose two-regime (SETAR-2) and three-regime (SETAR-3) SETAR models, which can be represented as follows:

$$\text{SETAR-2: } y_t = \begin{cases} \phi_0^{(1)} + \sum_{i=1}^{p^{(1)}} \phi_i^{(1)} y_{t-i} + \varepsilon_t^{(1)} & if\ y_{t-d} \le r \\ \phi_0^{(2)} + \sum_{i=1}^{p^{(2)}} \phi_i^{(2)} y_{t-i} + \varepsilon_t^{(2)} & if\ y_{t-d} > r \end{cases}$$

$$\text{SETAR-3: } y_t = \begin{cases} \phi_0^{(1)} + \sum_{i=1}^{p^{(1)}} \phi_i^{(1)} y_{t-i} + \varepsilon_t^{(1)} & if\ y_{t-d} \le r_1 \\ \phi_0^{(2)} + \sum_{i=1}^{p^{(2)}} \phi_i^{(2)} y_{t-i} + \varepsilon_t^{(2)} & if\ r_1 < y_{t-d} \le r_2 \\ \phi_0^{(3)} + \sum_{i=1}^{p^{(3)}} \phi_i^{(3)} y_{t-i} + \varepsilon_t^{(3)} & if\ y_{t-d} > r_2 \end{cases}$$

where $\varepsilon_t^{(j)}$ is assumed IID$(0,\sigma^{2(j)})$ and r_j represents the threshold values.

When the structural parameters, r and d, are known, a SETAR model can be estimated by fitting an AR model to the appropriate subset of observations determined by the relationship of the threshold variable to the value of the threshold (*arranged autoregression*).

In cases where the threshold parameter (r) and the delay parameter (d) are unknown, Tong (1983) suggests an empirical procedure which selects as 'best' the model that yields the minimum Akaike Information Criteria (AIC). However, as stressed by Priestley (1988), such a procedure is to be seen as a guide in choosing a small subclass of non-linear models featuring desirable economic and statistical properties.

For the case of a SETAR (p_1, p_2; d) model, Tong (1983) proposes a three-stage procedure: for given values of d and r, separate AR models are fitted to the appropriate subsets of data, the order of each

model being chosen according to the usual AIC criteria. In the second stage r can vary over a set of possible values while d has to remain fixed and is determined as the parameter for which AIC(d, $\hat{r}$) attains its minimum value. In stage three the search over d is carried out by repeating both stage 1 and stage 2 for $d=d_1, d_2, ..., d_p$. The selected value of d is, again, the value which minimise AIC(d).

2.2. *GARCH models*

An ARCH process can be defined in terms of the error distribution of a model in which the variable y_t is generated by:

$$y_t = x_t\beta + \varepsilon_t \quad t=1, ..., T \tag{2}$$

where x_t is a vector of *kx1* explanatory variables, which in our study includes only lagged values of y_t, and β is a *kx1* vector of autoregressive coefficients. The ARCH model proposed by Engle (1982) specifies the distribution of ε_t conditioned on the information set Ψ_{t-1}, which includes the actual values for the variables $y_{t-1}, y_{t-2}, ..., y_{t-k}$. In particular, the model is based on the assumption that:

$$\varepsilon_t \,|\Psi_{t-1} \sim N(0,h_t) \quad \text{where } h_t = \alpha_0 + \alpha_1\varepsilon^2_{t-1} + ... + \alpha_q\varepsilon^2_{t-q} \tag{3}$$

with $\alpha_0 > 0$ and $\alpha_i \geq 0$, $i=1, ..., q$, in order to constrain the conditional variance to be positive. Thus, the error variance is time-varying and depends on the magnitude of past errors.

Bollerslev (1986) proposes a generalisation of the ARCH model, which leads to the following specification of the conditional variance:

$$h_t = \alpha_0 + \alpha_1\varepsilon^2_{t-1} + ... + \alpha_q\varepsilon^2_{t-q} + \beta_1 h_{t-1} + ... + \beta_p h_{t-p} \tag{4}$$

This process is known as GARCH(p, q). To guarantee that the conditional variance assumes only positive values, the following restrictions have to be imposed: $\alpha_0 > 0$, $\alpha_i \geq 0$ for $i=1, ..., q$, and $\beta_i \geq 0$ for $i=1, ..., p$. In practice, the value of q in the GARCH model is much smaller than the corresponding value of q in the ARCH representation. Usually, a simple GARCH(1,1) model offers an adequate description of most economic and financial time series.

3. Preliminary data analysis and linearity tests

The empirical analysis was carried out on the log-differences of the end-of-week quotation of the Japanese yen exchange rate series and the daily quotation for the British pound exchange rate series. The log-levels and the log-differences of the series for the period 1973.1-1997.7 are depicted in figure 1. The log-differences series are mean-stationary, while the variance features the typical *volatility clustering* phenomenon with periods of high volatility followed by periods of low volatility. Table 1 outlines the descriptive statistics for the exchange rate log-differences. The series are characterised by excessive kurtosis and asymmetry, while the Jarque-Bera test strongly rejects the normality hypothesis.

In order to detect the presence of nonlinear components in the differenced series, we apply the RESET test and the S_2 test proposed by Luukkonen, Saikkonen and Teräsvirta (1988). These tests are devised for the null hypothesis of linearity. The RESET test is applied in its Lagrange Multiplier variant (Granger and Teräsvirta 1993): a linear autoregression of order p is run, followed by an auxiliary regression in which powers of the fitted values obtained in the first stage are included along with the initial regressors up to the power $k = 2, 3, 4$. The test is distributed as a χ^2 with $k-1$ degrees of freedom. While the RESET test is devised for a generic form of misspecification, the S_2 test is formulated for a specific alternative hypothesis, i.e. STAR-type non-linearity; the authors show that the S_2 test has reasonable power even when the true model is a SETAR one. The test is calculated as $S_2 = T(SSE_0 - SSE_1)/SSE_0$, where SSE_0 is the residual sum of squares from a linear autoregression of order p for y_t, and SSE_1 is the residual sum of squares from the auxiliary regression in which the initial regressors enter linearly and multiplied by the transition variable y_{t-d} raised up to the third power.[1] In this analysis we perform the test selecting the value of the delay parameter, d, in the range [1,5]; under the null hypo-

---

[1] The auxiliary regression is specified as:

$$\hat{\varepsilon}_t = \beta_0 + \sum_{i=1}^{p} \beta_i y_{t-i} + \sum_{i=1}^{p} \xi_i y_{t-i} y_{t-d} + \sum_{i=1}^{p} \psi_i y_{t-i} y_{t-d}^2 + \sum_{i=1}^{p} \kappa_i y_{t-i} y_{t-d}^3 + \nu_t,$$

where $\hat{\varepsilon}_t$ are the residuals from the linear AR(p) model.

TABLE 1

DESCRIPTIVE STATISTICS FOR THE LOG-DIFFERENCED SERIES

	Japanese yen	British pound
Frequency	weekly	daily
Mean	–0.000740	0.000058
Median	0.000000	–0.000100
Maximum	0.063120	0.038400
Minimum	–0.105679	–0.045900
Standard deviation	0.014186	0.006278
Skewness	–0.702024	0.152252
Kurtosis	7.815579	6.972260
Jarque-Bera	1342.976	4078.99
Probability	0.000000	0.000000
Period	03.01.73-31.07.97	03.01.73-31.07.97
Observations	1281	6168

FIGURE 1

EXCHANGE RATE LOG-LEVELS AND LOG-DIFFERENCES 1973.1-1997.7
WEEKLY JAPANESE YEN SERIES

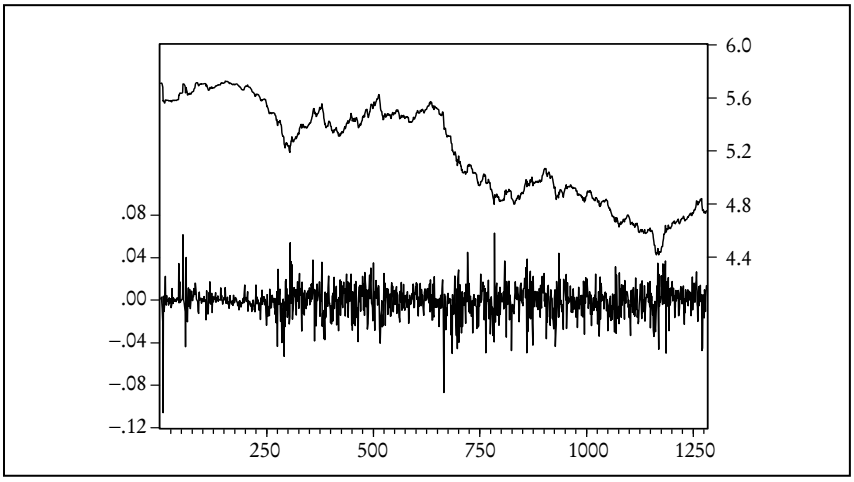
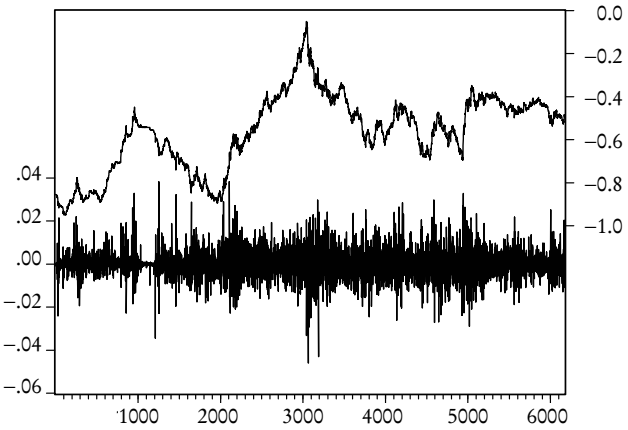


FIGURE 1 (*cont.*)

DAILY BRITISH POUND SERIES



thesis of linearity the test has a χ^2 distribution with $3p$ degrees of freedom.

In table 2 we report the probability values for the tests computed for the whole sample period, the estimation period and the forecast period. For each test the linear model under the null hypothesis was estimated assuming different lag structures ($p=2$, ..., 6). The table shows results only for $p=3$, 4 and 5. As we can see, when the tests are applied to the whole sample, they lead to the rejection of the null in a large number of cases, indicating that there is strong evidence of non-linear components in the data. However, by splitting the sample into the estimation period and the forecast period we find that there is less evidence of non-linearities in the latter for the weekly Japanese yen. When the tests are applied to the forecast period, in fact, we obtain clear evidence of non-linearity only from the S_2 test with $d=1$, indicating some type of threshold behaviour. The daily British pound series, on the other hand, exhibits a high degree of non-linearity in both estimation and forecasting period.

4. Model estimation

The models are estimated over the period 1973.2-1991.6. The estimates of the models are set out in table 3. The linear model selected for

TABLE 2

LINEARITY TESTS – *P*-VALUES

Japanese yen	Entire sample observations = 1281 02.01.73-31.07.97			Estimation sample observations = 964 02.01.73-30.06.91			Forecasting sample observations = 313 01.07.91-31.07.97		
p	*3*	*4*	*5*	*3*	*4*	*5*	*3*	*4*	*5*
RESET $h=2$	0.878	0.712	0.958	0.697	0.522	0.756	0.684	0.644	0.613
RESET $h=3$	**0.025**	**0.098**	**0.036**	0.118	0.165	**0.094**	0.435	0.870	0.464
RESET $h=4$	**0.018**	**0.085**	**0.016**	**0.083**	0.109	**0.043**	0.621	0.877	0.445
S_2, $d=1$	0.141	**0.067**	**0.009**	0.259	0.152	**0.030**	**0.005**	**0.000**	**0.000**
S_2, $d=2$	0.422	0.227	0.172	0.551	0.173	0.141	0.477	0.709	0.789
S_2, $d=3$	**0.017**	**0.037**	0.101	0.155	0.300	0.524	0.317	0.449	0.602
S_2, $d=4$	0.139	**0.071**	**0.087**	**0.056**	**0.020**	**0.024**	0.456	0.556	0.375
S_2, $d=5$	**0.013**	**0.002**	**0.007**	**0.016**	**0.008**	**0.049**	0.374	0.280	**0.080**
British pound	Entire sample observations = 6168 02.01.73-31.07.97			Estimation sample observations = 4636 02.01.73-30.06.91			Forecasting sample observations = 1532 01.07.91-31.07.97		
p	*3*	*4*	*5*	*3*	*4*	*5*	*3*	*4*	*5*
RESET $h=2$	0.3541	0.3668	**0.0172**	0.7403	0.6254	0.1675	**0.0000**	**0.0000**	**0.0000**
RESET $h=3$	0.5688	0.5784	**0.0583**	**0.0342**	**0.0428**	0.1307	**0.0000**	**0.0000**	**0.0000**
RESET $h=4$	0.6382	0.6267	**0.0328**	**0.0688**	**0.0665**	0.2476	**0.0000**	**0.0001**	**0.0001**
S_2, $d=1$	**0.0027**	**0.0004**	**0.0001**	**0.0012**	**0.0001**	**0.0001**	**0.0000**	**0.0000**	**0.0000**
S_2, $d=2$	0.3573	0.3846	**0.0018**	**0.0022**	**0.0028**	**0.0000**	**0.0000**	**0.0000**	**0.0000**
S_2, $d=3$	**0.0956**	0.1695	0.1662	**0.0060**	**0.0080**	**0.0022**	**0.0004**	**0.0037**	**0.0014**
S_2, $d=4$	0.1063	0.0073	**0.0078**	**0.0006**	**0.0000**	**0.0000**	0.1066	**0.0609**	0.1011
S_2, $d=5$	**0.0025**	0.0046	**0.0047**	**0.0000**	**0.0000**	**0.0002**	**0.0355**	**0.0798**	**0.0364**

p is the autoregressive lag order under the null hypothesis of linearity.
Numbers in bold indicate rejections of the linearity hypothesis up to 10% level of significance.

the forecast exercise is an AR(2) process for the Japanese yen series and an AR(9) process for the British pound series. The AR models are compared in the forecasting exercise with a GARCH(1,1) model and with a two-regime and three-regime SETAR model. All model specifications are reported in table 3. As we can see from table 3 Panel A, GARCH components are strongly present in the data, thus capturing the evident volatility clustering illustrated in figure 1. Moreover, in both GARCH specification shocks to volatility have markedly persistent effects.

TABLE 3

MODELS SPECIFICATIONS OVER THE ESTIMATION SAMPLE

Panel A

	Japanese yen						British pound					
	AR			GARCH(1,1)			AR			GARCH(1,1)		
Model	$y_t = c + \rho_1 y_{t-1} + \rho_2 y_{t-2} + e_t$			$y_t = c + b y_{t-1} + e_t$ $h_t = \alpha_0 + \alpha_1 e_t^2 + \beta_1 h_{t-1}$			$y_t = \rho_1 y_{t-1} + \rho_2 y_{t-5} +$ $+ \rho_3 y_t + 9 + e_t$			$y_t = b_1 y_{t-1} + b_2 y_{t-5} +$ $+ b_3 y_{t-9} + e_t$ $h_t = \alpha_0 + \alpha_1 e_t^2 + \beta_1 h_{t-1}$		
		estimate	t – value		estimate	t – value		estimate	t – value		estimate	t-value
	C	–0.0007	–1.471	c	–0.0005	–1.336	ρ_1	0.0405	2.083	b_1	0.0648	3.039
	P_1	0.0705	2.196	b	0.0991	2.971	ρ_2	0.0599	3.079	b_2	0.0684	3.530
	P_2	0.1038	3.230	α_0	6.84E-07	2.527	ρ_3	0.0646	3.323	b_3	0.0456	2.365
				α_1	0.0726	13.171				α_0	9.08E-07	22.187
				β_1	0.9301	234.329				α_1	0.1360	18.622
	σ	0.0063					σ	0.0073		β_1	0.8482	140.95

Panel B

		Japanese yen				British pound			
		SETAR-2		SETAR-3		SETAR-2		SETAR-3	
		estimate	t-value	estimate	t-value	estimate	t-value	estimate	t-value
Regime 1	$\phi_0^{(1)}$	–0.001	–2.139	–0.003	–3.804	–		–0.0003	–1.500
	$\phi_1^{(1)}$	0.096	2.698			–			
	$\phi_2^{(1)}$	0.137	3.989			–0.1582	–3.202		
	$\sigma^{(1)}$	0.0134		0.0164		0.0084		0.0071	
	$T^{(1)}$	736		332		503		1428	
Regime 2	$\phi_0^{(2)}$	0.001	1.441	0.001	0.283	–		–	
	$\phi_1^{(2)}$	–		0.407	1.708	0.0810		0.0095	3.785
	$\phi_2^{(2)}$	–		0.220	5.176	0.0235	5.329	–0.0739	–2.875
	$\phi_3^{(2)}$	–		–		–0.0200	1.546	–	
	$\phi_4^{(2)}$	–		–		–	–1.307	–	
	$\phi_5^{(2)}$	–		–		0.0531		–	
	$\phi_6^{(2)}$						2.87	0.0829	2.961
	$\sigma^{(2)}$	0.0150		0.0103		0.0060		0.0048	
	$T^{(2)}$	222		361		4120		1343	
Regime 3	$\phi_0^{(3)}$	–		0.001	1.160			0.0004	2.000
	$\phi_1^{(3)}$	–						0.0932	4.070
	$\phi_2^{(3)}$	–						0.0486	2.122
	$\sigma^{(3)}$	–		0.0140				0.0065	
	$T^{(3)}$	–		265				1859	
Model	$\sigma^{(model)}$	0.0138		0.0137		0.0063		0.0063	
	d	3		1		5		5	
	r_1	0.0072		–0.0032		–0.0066		–0.0020	
	r_2	–		0.0057				0.0008	

With regard to the SETAR models, we estimated specifications with one threshold (2 regimes) and two thresholds (3 regimes), following the estimation procedure suggested by Tong (1983). Model selection was conducted on the basis of the AIC criterion; however, when it appeared that the AIC overestimated the autoregressive order of the model, we selected the model with the most parsimonious dynamic structure. We considered models with a maximum lag order $p=6$ for the weekly yen series and $p=9$ for the daily British pound series. The models selected are indicated in Panel B of Table 3. In general, the dynamic structure, the estimated coefficients and the error variance differ significantly across regimes, thus indicating that the data are strongly characterised by non-linearities. Moreover, it is interesting to note that the dynamics of the three-regime SETAR model for the yen series are in line with the theoretical model described in Hsieh (1989) and with the empirical evidence reported by Kräger and Kugler (1993): the evidence of non-linearity in the mean is probably due to the existence of a managed floating exchange rate regime, in which the central banks intervene in order to avoid excessive depreciation or appreciation.

5. The forecasting exercise

The forecasting performance of the models is evaluated in different ways. First we compute MSFE for the various models for different steps ahead (1 to 5), and compare the relative performance of the models by means of the Diebold and Mariano test. This exercise is first conducted over the entire period, then on different sub-samples where non-linearities may be present with varying intensities. Fildes (1992) and Tashman (2000) point out the importance of carrying out this kind of investigation, arguing that selection of the forecasting origin might affect the accuracy of the results. Selection of the forecasting origin is, in fact, often arbitrary, there being no clear-cut criterion to decide the splitting of the entire sample into estimation and forecasting periods, apart from the need to have enough observations in the first period to obtain sensible and robust estimates, and to reserve enough observations to the second period to be able to evaluate the out-of sample performance of the model.

Diversifying into "multiple test period" (Tashman 2000) is particularly wise when some important characteristics of the data may be masked on carrying out analysis over the whole forecasting period. Our sub-sample analysis may provide some fruitful insights into the capacity of the models to discriminate between periods featuring high non-linearity and periods characterised by linearity.

We then extend the evaluation of the models to cover interval predictions. Models of conditional variance are particularly useful when the object of the analysis is to provide some indication of the uncertainty around the mean. Evaluation of interval forecasts could reveal gains to the non-linear models, particularly the GARCH models, which may not be apparent on MSFE measures.

5.1. *Point forecasts evaluation*

5.1.1. MSFEs over the entire forecast period

In this comparative exercise the forecasting ability of the models is assessed by means of the MSFE. The forecasts for the two series were calculated recursively from 1 to 5 steps-ahead. The models were identified and specified only once, over the first estimation periods, 1973.2-1991.6. The models were then re-estimated (but not re-specified) by expanding the sample with one observation each time, over the period 1991.7-1997.7, obtaining for each forecasting horizon (h) 313 point forecasts for the weekly yen and 1532 for the daily British pound. Computation of multi-step-ahead forecasts ($h>1$) from non-linear models (SETAR) involves complex analytical calculations and the use of numerical integration techniques or, alternatively, the use of simulation methods. In this study the forecasts are obtained by applying the Monte Carlo[2] method. In Table 4 we report the MSFEs and MSFEs normalised with respect to the linear model, which represents our benchmark. The values are calculated as the ratio $MSFE_{NL}/MSFE_{L}$; a number less than one means that the non-linear model provides more accurate forecasts than the simple linear model. Furthermore, in order to assess whether this superiority is statistically significant, we perform the Diebold-Mariano (1995) test; values

[2] Each point forecast is obtained as the average over 500 replications.

TABLE 4

FORECASTING PERFORMANCE – MSFE AND NORMALISED MSFE

	Number of steps-ahead									
	1		2		3		4		5	
Japanese yen	MSFE	N-MSFE	MSFE	N-MSFE	MSFE	N-MSFE	MSFE	N-MSFE	MSFE	N-MSFE
Naïve	1.8917	–	1.8944	–	1.8915	–	1.8985	–	1.8992	–
Linear AR(2)	1.8929	–	1.8892	–	1.8837	–	1.8948	–	1.8992	–
GARCH	1.8980	1.003	1.8903	1.001	1.8785	0.997	1.8913	0.998	1.8965	0.999
SETAR-2	1.8643	0.985	1.8777	0.994	1.8569	0.986	1.9306	1.019**	1.8686	0.984**
SETAR-3	1.9324	1.021	1.9384	1.026**	1.8847	1.001	1.9153	1.011	1.8747	0.987
British pound										
Linear AR(9)	0.3886	–	0.3884	–	0.3881	–	0.3881	–	0.3892	–
GARCH	0.3883	0.9920*	0.3883	0.997	0.388	0.997	0.3880	0.997	0.3891	0.997
SETAR-2	0.3906	1.0051	0.3950	1.0170**	0.3954	1.0188**	0.3898	1.0044	0.3895	1.0008
SETAR-3	0.3903	1.0045	0.3951	1.0173**	0.3906	1.0065	0.3889	1.0022	0.3948	1.0145**

Values are calculated for 313 forecasts for the yen and 1532 forecasts for the British pound. Note that the value of MSFE has been rescaled by multiplying by 10^4. The normalised MSFE is calculated as the ratio $MSFE_{NL}/MSFE_{L}$; *, ** denotes significance of the Diebold-Mariano test at 10% and 5% level of significance.

leading to rejection of the null hypothesis of equality of forecast accuracy are indicated with stars. Table 4 also shows for the yen series the MSFE obtained from a *naïve* forecast by assuming that the levels of the exchange rates follow a *random walk* with drift process.

We note that in terms of MSFE the models exhibit in general similar values. The SETAR-2 model for the yen produces point forecasts which are marginally better than the AR forecasts for 4 horizons out of 5, although only in one case do the forecasting gains prove significant according to the Diebold-Mariano test. In the case of the daily British pound the few significant values seem to favour the linear model.

As mentioned in the introduction, such results may, as argued by Diebold and Nason (1990), be due either to the fact that non-linearity is weaker over the forecast period or to the fact that, if non-linearities concern the even-ordered conditional moments, they are of no use in improving point forecasts and cannot be revealed in terms of MSFE. In our application the former explanation may apply to the weekly yen for which, as we have seen, the linearity tests detected fewer rejections in the forecasting period than in the estimation period. On the other hand, for the daily British pound, which is strongly charac-

terised by second order non-linearities, we expect the GARCH model to outperform the linear AR in terms of interval forecasts rather than point forecasts.

5.1.2. MSFEs over different sub-samples

Following the recommendations in Tashman (2000), in this section we further articulate evaluation of the models by conducting multi-period tests to assess the sensitivity of the results to specific sub-samples.

We start from examination of the linearity properties of the series over different sub-samples. The results of this analysis are reported in table 5. For the Japanese yen the forecast period is divided into six sub-samples of equal length, each containing approximately 50 observations. As can be seen from Table 5, there is some suggestion of non-linearity of varying degrees across sub-samples: linearity is rejected in sub-samples 1, 2 and 5, while there appears to be very little evidence of non-linearity for sub-samples 3 and 4. Non-linear models are expected to perform better in periods characterised by non-linearity. However, a recent Monte Carlo study (Clements *et al*. 2003) has indicated that the data need to exhibit a significantly high degree of non-linearity for the Diebold-Mariano test to reveal that a SETAR model outperforms a linear AR specification.

Application of the Diebold-Mariano test to distinct sub-samples yields the results summarised in table 6A, where values of normalised MSFE greater than one indicate the superiority of the benchmark AR model. The picture obtained from table 6A is only marginally more informative on the relative performance of the models than that obtained from the analysis over the full forecast sample. In particular, we observe that cases in which the SETAR model outperforms the linear model are still rare and, as expected, coincide with the sub-periods characterised by more prominent non-linearity.

With regard to the British pound, the multi-period evaluation was conducted over 15 sub-samples, each containing approximately 100 observations. Again, only a minority of the cases indicated in table 6B show some gains from the non-linear models in terms of MSFEs. The most striking forecasting gains are obtained from the GARCH model in sub-samples S4 and S10, from the SETAR-2 model in sub-samples S9 and S10 and from the SETAR-3 model in sub-samples

TABLE 5

LINEARITY TESTS BY SUB-SAMPLES – P-VALUES

Japanese yen	S1			S2			S3			S4			S5			S6		
p	**3**	**4**	**5**	3	4	5	3	4	5	3	4	5	3	4	5	3	4	5
RESET $h=2$	**0.008**	**0.070**	0.236	0.971	0.864	0.763	0.169	0.989	0.979	0.437	0.303	0.791	0.698	0.636	0.134	**0.008**	**0.009**	**0.008**
RESET $h=3$	**0.014**	**0.070**	0.287	0.522	0.438	0.463	0.358	0.888	0.872	0.228	0.231	0.109	0.511	0.884	0.322	**0.019**	**0.017**	**0.025**
RESET $h=4$	**0.036**	**0.068**	0.233	0.523	0.348	0.342	0.372	0.746	0.734	0.328	0.387	0.212	**0.027**	0.942	**0.053**	**0.021**	**0.033**	**0.053**
S_2, $d=1$	0.637	0.788	0.666	0.147	0.210	0.207	0.617	0.704	0.314	0.499	0.583	0.450	**0.045**	**0.002**	**0.010**	**0.075**	0.214	0.327
S_2, $d=2$	**0.073**	**0.028**	**0.032**	0.206	0.284	0.213	0.584	0.848	0.718	0.210	0.387	0.232	**0.012**	**0.063**	0.138	0.522	0.625	0.603
S_2, $d=3$	**0.001**	**0.002**	**0.005**	**0.091**	**0.011**	**0.030**	0.835	0.942	0.685	0.781	0.738	0.775	0.576	0.638	0.703	0.712	0.439	0.651
S_2, $d=4$	0.680	0.342	0.281	0.194	**0.095**	0.125	0.484	0.730	0.263	0.355	0.177	0.205	**0.015**	**0.087**	**0.051**	0.269	0.562	0.304
S_2, $d=5$	**0.026**	**0.016**	**0.033**	**0.038**	0.169	0.253	**0.065**	0.102	**0.018**	0.209	0.302	0.112	**0.016**	**0.005**	**0.038**	0.104	0.173	0.334

p is the autoregressive lag order under the null hypothesis of linearity
Numbers in bold indicate rejections of the linearity hypothesis up to 10% level of significance.

TABLE 6A

NORMALISED MSFE BY SUB-SAMPLES

Japanese yen		Number of steps-ahead				
		1	2	3	4	5
GARCH-M	S	1.003	1.001	0.997	0.998	0.999
	S1	1.017	1.011	1.007	1.007	1.006
	S2	1.022	1.017	1.007	1.009*	1.006
	S3	1.000	1.010	0.999	0.995	1.001
	S4	1.009	1.010	1.009	1.005	1.003
	S5	0.976	0.980	0.999	0.997	0.989**
	S6	0.977	0.985	0.999	1.000	0.995
SETAR-2	S	0.985	0.994	0.986	1.019**	0.984**
	S1	0.957*	0.987	0.937**	1.040*	0.946**
	S2	1.008	1.001	0.939**	1.010	1.024
	S3	1.022	1.047	1.070**	1.000	0.974
	S4	0.983	1.016	1.000	0.983	0.973
	S5	0.929**	0.992	1.012	1.021	0.991
	S6	1.010	1.012	1.012	1.024**	1.019
SETAR-3	S	1.021	1.026**	1.001	1.011	0.987
	S1	1.058	1.021	0.985	1.067**	0.930**
	S2	1.017	1.056	0.889**	0.979	1.013
	S3	1.038	1.007	1.053**	1.021	0.992
	S4	1.032	1.002	0.990	0.995	0.997
	S5	0.982	1.051**	1.039*	1.012	0.987
	S6	0.995	1.034	1.031	1.016	1.006

S refers to the whole forecasting period, S1, S2, S3, S4, S5 and S6 refer to the six sub-periods.
*, ** denotes significance of the Diebold-Mariano test at 10% and 5% level of significance.

TABLE 6B

NORMALISED MSFE BY SUB-SAMPLES – BRITISH POUND LOG-DIFFERENCES

British pound		Number of steps ahead				
		1	2	3	4	5
GARCH	S2	1.0014	1.0012	1.0010	1.0009	1.0011*
	S4	**0.9957****	**0.9973***	**0.9977****	**0.9977****	**0.9978****
	S9	0.9996	1.0003**	1.0003	1.0000	1.0002
	S10	**0.9981****	**0.9996****	**0.9995****	**0.9994****	**0.9995***
	S13	1.0002	0.9994	0.9991	**0.9992***	0.9993
	S14	**0.9984****	0.9999	1.0001	0.9999	1.0001
SETAR-2	S2	1.0060	1.0262	1.0138	0.9911	1.0007
	S4	0.9969	1.0275	1.0274	0.9835	0.9939
	S9	0.9851	**0.9485****	1.0145*	0.9484	0.9950
	S10	1.0043	**0.9343****	1.0177**	**0.9443***	0.9926
	S13	0.9974	1.0035	1.0007	1.0203	1.0236*
	S14	1.0014	**0.9560****	1.0209	0.9966	0.9873
SETAR-3	S2	0.9873	1.0273*	0.9980	**0.9909***	1.0233**
	S4	1.0161	1.0182	0.9891	**0.9710****	1.0113
	S9	0.9701	0.9879	1.0085	0.9843	1.0067
	S10	0.9923	**0.9374****	1.0357**	0.9856	1.0032
	S13	1.0313*	1.0087	0.9901	1.0255	1.0328
	S14	1.0014	0.9780	1.0024	1.0083	1.0036

*, ** denotes significance at 10% and 5% level of the Diebold-Mariano test.

S2, S4 and S10. Unreported results of the linearity tests over distinct sub-samples showed the highest number of rejections in sub-samples S9 and S10, confirming, in part, the results obtained for the Japanese yen indicating that the few cases of forecasting gains from the non-linear models coincide with periods of stronger non-linearity.

Overall, the results in this section demonstrate that, even in the presence of in-sample evidence of SETAR type non-linearity, these kinds of models only rarely outperform the linear AR benchmark on MSFE comparisons, therefore supporting the conclusion that the AR model provides a simple and robust tool for point forecast.

5.2. *Interval forecasts evaluation*

In this section we extend the forecast comparison by evaluating the models in terms of their ability to produce interval forecasts. An interval forecast, or prediction interval, for a variable specifies the probability that the future outcome will fall within a stated interval. The lower and upper limits of the interval forecast are given as the corresponding percentiles. We use central intervals so that, for example, the 90% prediction interval is formed by the 5^{th} and 95^{th} percentiles. Evaluation of interval forecasts is conducted by means of the likelihood ratio test of correct conditional coverage, as recently proposed by Christoffersen (1998). We are interested in detecting whether this comparison reveals gains to the non-linear models, particularly the GARCH models, which were not apparent on MSFE measures.

Christoffersen (1998) shows that a conditional interval forecast correctly calibrated will provide a hit sequence I_t (for $t=1, 2, ..., T$), with value 1 if the realisation is contained in the forecast interval, and 0 otherwise, that is distributed i.i.d. Bernoulli, with desired success probability p. The likelihood ratio test statistic for correct conditional coverage combines a test of unconditional coverage, LR_{UC}, with a test of independence. A sequence of interval forecasts is said to have correct unconditional coverage if $E[I_t]=p$, for all t. Denoting p the nominal coverage, n_1 and n_0 the realisations respectively inside and outside the forecast interval, and $\pi=n_1/(n_0+n_1)$ the sample proportion of successes, the test for unconditional coverage is given by:

$$LR_{UC} = -2\log\left(\frac{L(p;.)}{L(\pi;.)}\right) \text{ asy} \sim \chi_1^2$$

The likelihood under the null hypothesis $E[I_t]=p$ is $L(p; I_1, I_2, ..., I_n) = (1-p)^{n_0} p^{n_1}$, and under the alternative hypothesis $E[I_t]\neq p$ is $L(\pi; I_1, I_2, ..., I_n) = (1-\pi)^{n_0} \pi^{n_1}$. Thus, the test is computed as

$$LR_{UC} = 2[n_0 \log(1-\pi)/(1-p) + n_1 \log(\pi/p)]$$

This test does not have power against the alternative that the zeros and ones are clustered in time-dependent fashion. As stressed by Christoffersen, a simple test for correct unconditional coverage is insufficient in the presence of higher-order moments dynamics (conditional heteroscedasticity, for example). In order to overcome this limitation, Christoffersen proposes a test for independence and a joint test for independence and correct coverage (LR_{IND}).

The LR test for independence assumes a binary first-order Markov chain for the indicator function I_t with transition probability matrix given by:

$$\Pi_1 = \begin{bmatrix} 1-\pi_{01} & \pi_{01} \\ 1-\pi_{11} & \pi_{11} \end{bmatrix}$$

where $\pi_{ij}=Pr(I_t = j/I_{t-1}=i)$. Under independence $\pi_{ij}=\pi_j$, $i,j=0,1$ where $\pi_j = Pr(I_t=j)$. Thus, under the null hypothesis the transition probability matrix is restricted to:

$$\Pi_2 = \begin{bmatrix} 1-\pi_1 & \pi_1 \\ 1-\pi_1 & \pi_1 \end{bmatrix}$$

The π_{ij} and π_i are estimated by their sample frequencies. The unrestricted likelihood for the LR test is given by:

$$L(\Pi_1) = (1-\pi_{01})^{n_{00}} \pi_{01}^{n_{01}} (1-\pi_{11})^{n_{10}} \pi_{11}^{n_{11}}$$

and the restricted likelihood by:

$$L(\Pi_2) = (1-\pi_1)^{(n_{00}+n_{10})} \pi_1^{(n_{01}+n_{11})}$$

where n_{ij} is the number of times event i is followed by event j. LR_{IND} is asymptotically χ^2 with one degree of freedom under the null hypothesis of independently distributed indicator function values. A combina-

tion of these two tests will give a test of 'correct conditional coverage'. The joint test (LR_{CC}) is obtained as the sum of the two LR tests and is asymptotically χ^2 distributed with two degrees of freedom.

In our evaluation analysis of interval forecasts we have considered intervals with nominal coverage in the range 0.90-0.50. In table 7 we present results for a selected choice of intervals corresponding to probabilities 90, 75 and 50%. This choice enables us to investigate the accuracy of the model forecasts over different regions of the distribution, for example a model that does well in predicting the 90% interval also implies correct forecasts of events in the 5% left and right tails of the distribution; it is also possible that some models do better than others in predicting the tails of the distribution, but worse in predicting other aspects.

The models considered for our evaluation are the AR, GARCH and SETAR-2 estimated for the weekly log-differences of the Japanese yen and for the daily log-differences of the British pound. We also obtained results for the SETAR-3 model, but they are omitted from the following discussion as they were similar to those obtained from the SETAR-2 model. Table 7 reports, for each nominal coverage (p), the actual unconditional coverage (π) and the P-values of the three LR tests presented above.

The first set of results in table 7 refers to the Japanese yen (see panel A). The results reveal that for levels of coverage 90 and 75% the GARCH model is the only model to pass all three tests. The SETAR-2 model shows similar performance to the GARCH model in terms of correct unconditional coverage at these intervals, but fails the independence test. The worst performance is obtained from the AR model, which fails the correct conditional coverage test for all intervals. We also observe that all the models fail to capture some aspects of the underlying data generating process, as indicated by the highly significant test for correct unconditional coverage at the 50% interval. In particular, all the models appear to generate interval forecasts with actual coverage (π) greater than the nominal coverage ($p=0.50$), that is, the interval forecasts corresponding to 50% probability are too large, as more than 50% observations actually fall into that interval. This result may be attributed either to an overestimate of the standard errors used in the calculation of the forecast intervals or to inappropriate error distribution (see, for examples of possible alternative distributions not pursued here, Ryden, Teräsvirta and Asbrink 1998).

In applications with exchange rates and other financial variables accuracy seems more important in predicting events in the tails of the distribution (large losses or gains) rather than values in the middle of the distribution (small losses or gains). The more accurate performance of the GARCH model at wider intervals (90 to 75%) implies correct forecasts of events in the tails of the distribution, which is an interesting result, suggesting that the GARCH model can be more useful, especially for risk management.

In figure 2 we present plots of the 90% interval forecasts obtained for the log-differences of the yen from the three competing models. As can be seen from the figure, the GARCH model gives interval forecasts that are wider in volatile periods and narrower in tranquil periods. In this case, observations outside the intervals are evenly spread over the periods, while in the case of the AR model, and to a lesser extent for the SETAR models, observations outside the intervals are clustered in volatile periods and largely absent from tranquil periods. Thus, a fixed width conditional confidence interval, such as that obtained from AR models, is not correctly calibrated, since it fails to widen when the conditional variance rises and narrow when the conditional variance falls.

We now evaluate the interval forecasts generated by the three models for the daily log differences of the British pound, for which we present in table 7 different sets of results: for the entire *ex post* forecast period (Panel B) and for various distinct sub-periods (panels C and D). The results of the evaluation of forecast accuracy over the entire forecast period, consisting of 1532 observations, show that in terms of correct unconditional coverage, there is little to choose in the performance of the three models, in that all the models pass the test for the 90% intervals, but fail at the other coverage rates. However, when we examine the independence property of the forecasts, only the GARCH model passes the test at all levels of coverage, which reflects the ability of GARCH model to capture higher-ordered moment dynamics, and confirms the results previously obtained for the weekly log differences of the yen. The AR and SETAR models fail the independence and correct conditional coverage tests at all levels, while the GARCH model passes the correct conditional coverage test at the 90% interval, but fails at the 75 and 50% intervals, due to the highly significant test for correct unconditional coverage. Summarising the results obtained so far, again the GARCH model appears to have an advantage

TABLE 7

FORECAST INTERVALS EVALUATION FOR 1-STEP-AHEAD FORECASTS – *P*-VALUES

		AR				GARCH				SETAR-2			
	p	π	LR_{UC}	LR_{IND}	LR_{CC}	π	LR_{UC}	LR_{IND}	LR_{CC}	π	LR_{UC}	LR_{IND}	LR_{CC}
		Panel A											
Japanese yen, nf=313 01.07.91-31.07.97	0.90	0.904	0.805	**0.004**	**0.016**	0.904	0.805	0.202	0.430	0.901	0.955	**0.027**	0.087
	0.75	0.805	**0.021**	**0.002**	**0.001**	0.773	0.339	0.066	0.117	0.789	0.103	**0.047**	**0.037**
	0.50	0.597	**0.001**	0.793	**0.002**	0.569	**0.015**	0.308	**0.031**	0.578	**0.006**	0.789	**0.021**
		Panel B											
British pound, nf=1532 01.07.91-31.07.97	0.90	0.908	0.293	**0.000**	**0.001**	0.907	0.335	0.740	0.594	0.909	0.255	**0.008**	**0.016**
	0.75	0.817	**0.000**	**0.000**	**0.000**	0.802	**0.000**	0.775	**0.000**	0.819	**0.000**	**0.000**	**0.000**
	0.50	0.622	**0.000**	**0.000**	**0.000**	0.597	**0.000**	0.205	**0.000**	0.619	**0.000**	**0.000**	**0.000**
		Panel C											
British pound, nf=700 01.07.91-11.04.94	0.90	0.86	**0.002**	**0.008**	**0.00**	0.900	0.707	0.88	0.921	0.870	**0.008**	0.063	**0.005**
	0.75	0.74	0.386	0.122	0.208	0.780	0.093	0.13	0.077	0.740	0.435	**0.042**	0.093
	0.50	0.49	0.762	0.966	0.954	0.550	**0.008**	0.765	**0.029**	0.500	0.880	0.911	0.982
British pound, nf=832 12.04.94-31.07.97	0.90	0.96	**0.000**	**0.045**	**0.00**	0.920	**0.021**	0.667	0.064	0.960	**0.000**	0.905	**0.000**
	0.75	0.91	**0.000**	0.123	**0.00**	0.840	**0.000**	0.169	**0.000**	0.900	**0.000**	**0.019**	**0.000**
	0.50	0.76	**0.000**	**0.002**	**0.00**	0.650	**0.000**	0.090	**0.000**	0.760	**0.000**	**0.007**	**0.000**

nf indicates the number of one-step ahead forecasts; *p* indicates the nominal coverage; π indicates the actual unconditional coverage; numbers in bold represent rejections at 5% level of significance for the unconditional coverage test (LR_{UC}), the independence test (LR_{IND}) and the conditional coverage test (LR_{CC}). See text for a detailed description of these tests.

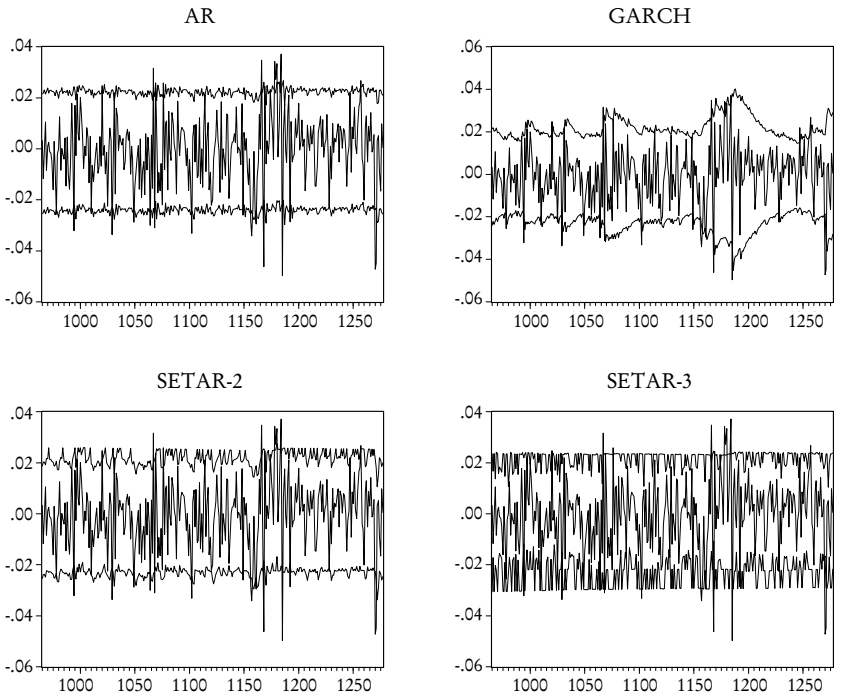
TABLE 7 (*cont.*)

British pound		p	AR				GARCH				SETAR-2			
			π	LR_{UC}	LR_{IND}	LR_{CC}	π	LR_{UC}	LR_{IND}	LR_{CC}	π	LR_{UC}	LR_{IND}	LR_{CC}
Sub-period	*Forecast obs*		Panel D											
S3	4837-4936	0.90	0.90	1.000	0.991	1.000	0.90	1.000	0.991	1.000	0.90	1.000	0.991	1.000
		0.75	0.81	0.153	0.259	0.191	0.77	0.641	0.166	0.344	0.82	0.094	0.087	0.057
		0.50	0.56	0.230	0.650	0.438	0.56	0.230	0.441	0.361	0.55	0.317	1.000	0.606
S4	4937-5036	0.90	0.68	**0.000**	0.226	**0.000**	0.83	**0.032**	0.500	0.080	0.70	**0.000**	0.368	**0.000**
		0.75	0.57	**0.000**	0.895	**0.000**	0.74	0.818	0.332	0.608	0.58	**0.000**	0.627	**0.001**
		0.50	0.37	**0.009**	0.968	**0.033**	0.51	0.841	0.614	0.863	0.37	**0.009**	0.968	**0.033**
S7	5237-5336	0.90	0.99	**0.000**	n.a.	n.a.	0.93	0.293	0.488	0.453	0.98	**0.001**	n.a.	n.a.
		0.75	0.92	**0.000**	0.654	**0.000**	0.82	0.094	0.367	0.164	0.92	**0.000**	0.654	**0.000**
		0.50	0.70	**0.000**	0.132	**0.000**	0.59	0.071	0.097	0.050	0.70	**0.000**	0.132	**0.000**
S9	5437-5536	0.90	0.94	0.153	0.342	0.229	0.89	0.742	0.817	0.922	0.94	0.153	n.a.	n.a.
		0.75	0.87	**0.003**	0.800	**0.012**	0.82	0.094	0.541	0.204	0.88	**0.001**	0.621	**0.005**
		0.50	0.74	**0.000**	0.457	**0.000**	0.69	**0.000**	0.094	**0.000**	0.73	**0.000**	0.335	**0.000**
S10	5537-5836	0.90	0.86	0.206	0.987	0.449	0.88	0.517	n.a.	n.a.	0.85	0.118	0.829	0.287
		0.75	0.76	0.817	0.524	0.794	0.78	0.482	0.949	0.780	0.77	0.641	0.361	0.591
		0.50	0.58	0.109	0.085	0.063	0.54	0.423	0.211	0.332	0.56	0.230	0.161	0.182
S12	5737-5836	0.90	0.98	**0.001**	n.a.	n.a.	0.93	0.293	0.409	0.409	0.96	**0.024**	0.084	**0.018**
		0.75	0.93	**0.000**	0.409	**0.000**	0.86	**0.007**	0.891	**0.026**	0.92	**0.000**	0.091	**0.000**
		0.50	0.83	**0.000**	0.384	**0.000**	0.66	**0.001**	0.881	**0.005**	0.81	**0.000**	0.324	**0.000**
S14	5937-6036	0.90	0.93	0.293	0.409	0.409	0.90	1.000	0.917	0.995	0.93	0.293	0.409	0.409
		0.75	0.87	**0.003**	0.232	**0.006**	0.80	0.237	0.395	0.346	0.88	**0.001**	0.536	**0.004**
		0.50	0.71	**0.000**	0.383	**0.000**	0.55	0.317	0.526	0.496	0.70	**0.000**	0.563	**0.000**

p indicates the nominal coverage; π indicates the actual unconditional coverage; numbers in bold represent rejections at 5% level of significance for the unconditional coverage test (LR_{UC}), the independence test (LR_{IND}) and the conditional coverage test (LR_{CC}). See text for a detailed description of these tests.

FIGURE 2

DAILY BRITISH POUND SERIES



over the other two models in predicting the more extreme values (left and right 5% tails), although all the models produce inaccurate interval forecasts at the other coverage levels.

The large number of daily observations has enabled us to articulate evaluation analysis for the British pound further by comparing the performance of the forecast models over distinct sub-periods. This analysis highlights some interesting features of the models in terms of the sensitivity of their forecast performance to the period considered and their ability to produce accurate interval forecasts at different levels of coverage. In the first stage, the multi-period evaluation is conducted by splitting the entire forecast period (1532 observations) into two broadly equal sub-periods, the first consisting of 700 observations, the second of the remaining 832 observations. These are then further divided into smaller sub-samples of approximately 100 observations each, for a total of 15 sub-samples. The results for the two

major sub-periods are set out in table 7, panel C, while in panel D we summarise the main findings for some selected smaller sub-periods.

First of all it is interesting to observe that the performance of the models is sensitive to the period over which they are evaluated: more specifically, all the models perform overall better in the first half of the forecast period examined (first 700 observations). On the other hand, in the second half of the forecast period (last 832 observations), all the models fail the correct unconditional coverage test. This failure leads to strong rejection of the combined test for correct conditional coverage for the AR and SETAR models, while rejection is only marginal (P-value=0.064) for the GARCH 90% interval, since the GARCH as usual passes the independence test. As can be seen from figure 3, the behaviour of the series changes substantially in the second half of the forecast period, when the series exhibits less volatility than previously. All the models seem to be affected in the same way by this break in the variance: evaluation over the second half of the forecast period reveals that the actual coverage (π) is higher than the nominal level used to construct the forecast intervals. We conjecture that these results can be attributed mainly to over-estimated standard errors. This conjecture is supported by comparison of the standard deviation exhibited by the series in different sub-periods with the standard errors used by the models to construct the interval forecasts, averaged over the relevant sub-periods (see Table 8).

We now turn our attention to panel D of table 7, where we report some selected results of the evaluation of interval forecasts over a number of distinct smaller sub-periods, offering further support to our conjecture above. As can be seen, in periods S9, S10, S12 and S14, which are characterised by particularly low standard deviation of the series, all the models produce excessively wide interval forecasts. These results clearly reflect the deficiency of fixed width confidence intervals of the AR model and also the limitation of the two-regime interval width of the SETAR-2 model. With regard to the GARCH model, its inability to produce accurate forecast intervals in those specific periods of reduced volatility may be attributed to the highly persistent conditional variance estimates which prevent immediate adjustments to the new volatility regime. In panel D we also present results for some selected sub-samples belonging to the first half of the forecast period (S3, S4 and S7), which highlight other interesting features of the models. In particular, we notice that while the GARCH produces accurate

TABLE 8

SUB-SAMPLES DESCRIPTIVE STATISTICS

British pound	period	obs	st-dev	skewness	kurtosis	Jarque-Bera (P-value)	Estimated standard deviations*		
							AR	GARCH	SETAR-2
Estimation sample	03.01.73-28.06.91	1-4636	0.00629	0.08139	7.30715	0.00000	0.00627	0.00732	0.00630
Entire forecasting sample	01.07.91-31.07.97	4637-6168	0.00624	0.03720	5.93367	0.00000	0.00638	0.00607	0.00634
S1	01.07.91-21.11.91	4637-4736	0.00752	−0.50711	4.58428	0.00063	0.00630	0.00753	0.00641
S2	22.11.91-16.04.92	4737-4836	0.00787	0.56257	3.50707	0.04188	0.00633	0.00765	0.00646
S3	17.04.92-04.09.92	4837-4936	0.00602	−0.04277	4.68446	0.00267	0.00633	0.00619	0.00646
S4	08.09.92-02.02.93	4937-5036	0.01125	0.09294	3.21896	0.84210	0.00642	0.01043	0.00653
S5	03.02.93-24.06.93	5037-5136	0.00723	−0.02145	3.73136	0.32688	0.00648	0.00778	0.00661
S6	25.06.93-17.11.93	5137-5236	0.00711	−0.20581	2.62645	0.52535	0.00650	0.00735	0.00657
S7	18.11.93-11.04.94	5237-5336	0.00403	0.41481	3.01503	0.23827	0.00648	0.00466	0.00625
S8	12.04.94-31.08.94	5337-5436	0.00386	0.09468	4.01721	0.10749	0.00644	0.00445	0.00628
S9	01.09.94-27.01.95	5437-5536	0.00461	0.04585	4.33491	0.02399	0.00641	0.00502	0.00629
S10	30.01.95-20.06.95	5537-5636	0.00667	0.14529	4.09594	0.06869	0.00640	0.00662	0.00640
S11	21.06.95-10.11.95	5637-5736	0.00397	0.21627	4.47846	0.00713	0.00639	0.00474	0.00624
S12	13.11.95-05.04.96	5736-5836	0.00389	0.42307	4.86244	0.00016	0.00636	0.00439	0.00617
S13	08.04.96-27.08.96	5837-5936	0.00328	0.29339	4.67223	0.00144	0.00632	0.00398	0.00609
S14	28.08.96-23.01.97	5937-6036	0.00549	1.28747	7.88290	0.00000	0.00628	0.00506	0.00616
S15	24.01.97-31.07.97	6037-6168	0.00513	0.41659	4.65228	0.00008	0.00627	0.00539	0.00617

* calculated as the average across relevant observations.

interval forecasts throughout all the sub-samples, the AR and SETAR models are unable to perform well in sub-samples where the standard deviation of the observed series is either above or below its unconditional value of 0.006.

6. Conclusions

In this study we have compared the forecasting performance of alternative univariate time series models for the Japanese yen (weekly log-differences) and the British pound (daily log-differences). Three non-linear models, namely a two-regime SETAR, a three-regime SETAR and a GARCH model, were contrasted with simple linear alternatives (AR processes).

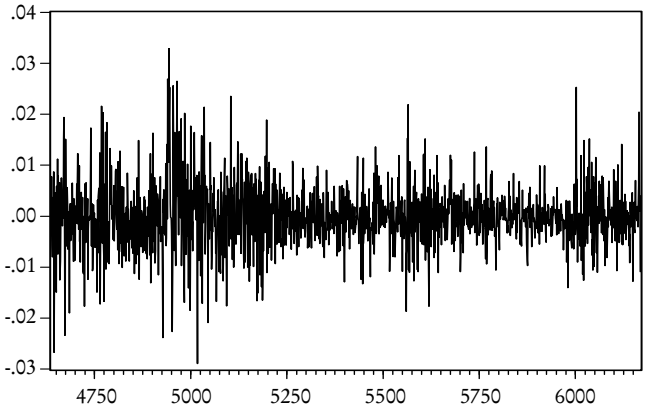
The SETAR and GARCH models proved successful in describing non-linear features of the data. In particular, the SETAR models provided strong in-sample evidence for the existence of different regimes, in which the exchange rates log-differences exhibit quite different dynamics, while the GARCH models successfully captured the volatility clustering of the log-differenced series.

The forecast performance of the models was assessed by means of different forecasting evaluation criteria. First, we evaluated the models in terms of their ability to produce point forecasts, by comparing MSFEs over the entire forecasting period (1991.6-1997.7) and over different sub-samples. Differences in MSFE between models were evaluated by means of the Diebold and Mariano test. This analysis did not show significant forecast gains of the non-linear models over the linear benchmark, with only few exceptions coinciding with periods of more prominent non-linearity. These results support the conclusion that, even in the presence of in-sample non-linearity, AR models can provide a simple and robust tool for point forecasts.

Next, the models were evaluated in terms of their ability to produce interval forecasts. Following Christoffersen (1998), for evaluation of the interval forecasts we computed LR tests for unconditional coverage and independence. These tests were then combined into joint tests of conditional coverage. Multi-period analysis showed that the results are somewhat sensitive to the period chosen for the evaluation of the models and also vary markedly with the level of coverage considered.

FIGURE 3

DAILY BRITISH POUND LOG-DIFFERENCES
FORECASTING PERIOD 1991.7-1997.7



A robust result emerging from the evaluation of interval forecasts revealed gains to the GARCH models, especially at the wider intervals, implying correct coverage of the tails of the distribution. This is an interesting finding, suggesting that the GARCH model can be more useful in practical applications, and especially for risk management. The analysis has also shown that the static interval forecasts from the AR model and the two regime interval forecasts from the SETAR model are clearly not good 'conditional' interval forecasts. The tests also revealed that all models failed to produce forecasts with correct coverage for narrower intervals (.50%), suggesting that some aspects of the underlying data generating process were not adequately captured by the models. However, for practical applications, particularly with financial variables, where attention is typically confined to the tails of the distribution (large losses and gains), less accurate forecast performance in the middle range of the distribution (small changes of the series) may be of minor importance.

These results clearly reflect the fact that the forecasting models are all suboptimal, and it is therefore possible that some do better than others in predicting certain regions of the distribution, but worse in predicting other aspects. From all this we see pointers emerging for future research in various directions. For example, consideration of alternative error distributions might be a promising avenue to follow. It would also be interesting to see whether alternative types of nonlinear models not pursued here (Artificial Neural Networks, STAR,

Markov Switching) or some combination of forecasts from different models might yield better performance.

REFERENCES

BOERO, G. and E. MARROCU (2000), "Modelli nonlineari per i tassi di cambio: un confronto previsivo con dati a diversa frequenza", *Moneta e Credito*, vol. 53, pp. 385-415.

BOERO G. and E. MARROCU (2002), "The performance of nonlinear exchange rate models: a forecast comparison", *Journal of Forecasting*, vol. 21, pp. 513-42.

BOERO G. and E. MARROCU (2004), "The performance of SETAR models: a regime conditional evaluation of point, interval and density forecasts", *International Journal of Forecasting*, vol. 20, pp. 305-20.

BOLLERSLEV, T. (1986), "Generalized autoregressive conditional heteroscedasticity", *Journal of Econometrics*, vol. 31, pp. 307-27.

CHRISTOFFERSEN, P. (1998), "Evaluating interval forecasts", *International Economic Review*, vol. 39, pp. 841-62.

CLARÍDA, R.H., L. SARNO, M.P. TAYLOR and G. VALENTE (2003), "The out-of-sample success of term structure models as exchange rate predictors: a step beyond", *Journal of International Economics*, vol. 60, pp. 61-83.

CLEMENTS, M.P. and J.P. SMITH (1999), "A Monte Carlo study of the forecasting performance of empirical SETAR models", *Journal of Applied Econometrics*, vol. 14, pp. 123-41.

CLEMENTS M.P. and J.P. SMITH (2001), "Evaluating forecasts from SETAR models of exchange rates", *Journal of International Money and Finance*, vol. 20, pp. 133-48.

CLEMENTS, M.P., P.H. FRANSES, J.P. SMITH and D. VAN DIJK (2003), "On SETAR non-linearity and forecasting, *Journal of Forecasting*, vol. 22, pp. 359-75.

DACCO, R. and S. SATCHELL (1999), "Why do regime-switching models forecast so badly?", *Journal of Forecasting*, vol. 18, pp. 1-16.

DIEBOLD, F.X. and R.S. MARIANO (1995), "Comparing predictive accuracy", *Journal of Business and Economic Statistics*, vol. 13, pp. 253-63.

DIEBOLD, F.X. and J.A. NASON (1990), "Nonparametric exchange rate prediction?", *Journal of International Economics*, vol. 28, pp. 315-32.

ENGEL C. (1994), "Can the Markov switching model forecast exchange rate?", *Journal of International Economics*, vol. 36, pp. 151-65.

ENGEL C. and J.D. HAMILTON (1990), "Long swings in the dollar: are they in the data and do markets know it?", *American Economic Review*, vol. 80, pp. 689-713.

ENGLE, R.F. (1982), "Autoregressive conditional heteroscedasticity with estimates of the variance of the United Kingdom inflation", *Econometrica*, vol. 50, pp. 987-1008.

FILDES, R. (1992), "The evaluation of extrapolative forecasting methods", *International Journal of Forecasting*, vol. 8, pp. 81-98.

GRANGER C.W.J and T. TERÄSVIRTA (1993), *Modelling Nonlinear Economic Relationships*, Oxford University Press, Oxford.

HAMILTON, J.D. (1989), "A new approach to the economic analysis of nonstationary time series and the business cycle", *Econometrica*, vol. 57, pp. 357-84.

HSIEH, D.A. (1989), "A nonlinear stochastic rational expectations model of exchange rates", Fuqua School of Business, Duke University, unpublished manuscript.

KRÄGER, H. and P. KUGLER (1993), "Nonlinearities in foreign exchange markets: a different perspective", *Journal of International Money and Finance*, vol. 12, pp. 195-208.

LUUKKONEN, R., P. SAIKKONEN and T. TERÄSVIRTA (1988), "Testing linearity in univariate time series models", *Scandinavian Journal of Statistics*, vol. 15, pp. 161-75.

MARK, N.C. and D. SUL (2001), "Nominal exchange rates and monetary fundamentals: evidence from a small post-Bretton Woods panel", *Journal of International Economics*, vol. 53, pp. 29-52.

MEESE R. and K. ROGOFF (1983), "Empirical exchange rate models of the seventies: do they fit out of sample?", *Journal of International Economics*, vol. 14, pp. 3-24.

MEESE, R.A. and A.K. ROSE (1991), "An empirical assessment of nonlinearities in models of exchange rate determination", *Review of Economic Studies*, vol. 58, pp. 603-19.

NELSON, D.B. (1991), "Conditional heteroskedasticity in asset returns: a new approach", *Econometrica*, vol. 59, pp. 347-70.

PEEL, D.A. and A.E. SPEIGHT (1994), "Testing for nonlinear dependence in inter-war exchange rates", *Weltwirtschaftliches Archiv*, Bd. 130, pp. 391-417.

POTTER, S. (1995), "A nonlinear approach to US GNP", *Journal of Applied Econometrics*, vol. 10, pp. 109-25.

PRIESTLEY, M.B. (1988), *Nonlinear and Non-stationary Time Series Analysis*, Academic Press, London.

RYDEN, T., T. TERÄSVIRTA and S. ASBRINK (1998), "Stylised facts of daily returns series and the hidden Markov model", *Journal of Applied Econometrics*, vol. 13, pp. 217-44.

TASHMAN, L.J. (2000), "Out-of-sample tests of forecasting accuracy: an analysis and review", *International Journal of Forecasting*, vol. 16, pp. 437-50.

TIAO G.C. and R.S. TSAY (1994), "Some advances in non-linear and adaptive modelling in time series", *Journal of Forecasting*, vol. 13, pp. 109-31.

TONG H. (1978), "On a threshold model. In pattern recognition and a signal processing", in C.H. Chen ed., *Pattern Recognition and Signal Processing*, Sijhoff and Noordoff, Amsterdam, pp. 101-41.

TONG, H. (1983), *Threshold Models in Nonlinear Time Series Analysis*, Springer-Verlag, New York.

TONG, H. (1995), *Non-linear Time Series. A Dynamical System Approach*, Clarendon Press, Oxford.

TONG, H. and K.S. LIM (1980), "Thresholds autoregression, limit cycles and cyclical data", *Journal of the Royal Statistical Society* B, vol. 42, pp. 245-92.

ZAKOIAN, J.M. (1994), "Threshold heteroskedastic models", *Journal of Economic Dynamics and Control*, vol. 18, pp. 931-55.