Journal of Forecasting
J. Forecast. **21**, 513–542 (2002)
Published online 12 July 2002 in Wiley InterScience (www.interscience.wiley.com). **DOI**: 10.1002/for.837

The Performance of Non-linear Exchange Rate Models: a Forecasting Comparison

GIANNA BOERO[1]* AND EMANUELA MARROCU[2]
[1] *University of Cagliari, CRENoS and University of Warwick*
[2] *University of Cagliari and CRENoS*

ABSTRACT

In recent years there has been a considerable development in modelling non-linearities and asymmetries in economic and financial variables. The aim of the current paper is to compare the forecasting performance of different models for the returns of three of the most traded exchange rates in terms of the US dollar, namely the French franc (FF/$), the German mark (DM/$) and the Japanese yen (Y/$). The relative performance of non-linear models of the SETAR, STAR and GARCH types is contrasted with their linear counterparts. The results show that if attention is restricted to mean square forecast errors, the performance of the models, when distinguishable, tends to favour the linear models. The forecast performance of the models is evaluated also conditional on the regime at the forecast origin and on density forecasts. This analysis produces more evidence of forecasting gains from non-linear models. Copyright © 2002 John Wiley & Sons, Ltd.

KEY WORDS non-linearity; forecasting accuracy; point forecasts; density forecasts; exchange rates

INTRODUCTION

The problem of exchange rate determination and its predictability is a very controversial issue in the international economics literature. In particular, the empirical specification of non-linear models for exchange rates has been largely motivated by non-linear solutions presented for such variables in a number of theoretical models. We refer, for example, to the literature on target zone models (Krugman, 1991) and to the rational expectations model with central bank stochastic intervention rules (Hsieh, unpublished manuscript, 1989). The development of alternative models and the evaluation of their forecasting abilities have been motivated by the evidence provided in Meese and Rogoff (1983) that the simple random walk model outperformed the complex structural models in forecasting exchange rate variables. Since then, a large number of studies have been carried out, some corroborating the importance of Meese and Rogoff's results, others stressing the relevance of the economic fundamentals in determining exchange rate behaviour and reaffirming the

* Correspondence to: Gianna Boero, Department of Economics, University of Warwick, Coventry CV4 7AL, UK.
E-mail: gianna.boero@warwick.ac.uk

forecasting superiority of the structural models over the random walk, at least for medium-long-term horizons. An exhaustive empirical assessment of non-linearities in the context of structural models can be found in Meese and Rose (1991), while an up-to-date review of the debate on the exchange rate determination can be found in the recent contributions by Dixon (1999), Rogoff (1999), Flood and Rose (1999) and MacDonald (1999).

Several other studies have been conducted in the context of univariate models, exploiting recent developments in non-linear time series econometrics. These studies focus mainly on the dynamic representation of exchange rates and on their short-run predictability, and are theoretically based on the efficient market hypothesis. The main rationale for these models is that if the exchange rate market is characterized by some degree of efficiency, it is plausible to assume that all the relevant information is embodied in the most recent exchange rate returns, so that it becomes unnecessary to include the economic fundamentals in the set of explanatory variables. Among the most commonly applied non-linear models, the GARCH (generalized autoregressive conditional heteroscedastic) and the SETAR (self-exciting threshold autoregressive) models have proved successful in describing the dynamic behaviour of many economic and financial variables; moreover, they offer the advantage of being readily interpretable in economic terms (see, among others, Kräger and Kugler, 1993, Peel and Speight, 1994; Chappell *et al*., 1996). The GARCH models allow one to specify the process governing both the mean and the variance of the series, while the SETAR models represent a stochastic process generated by the alternation of different regimes.

In general, although there have been extensive applications of new techniques to describe the non-linearities and asymmetries which characterize exchange rate dynamics, there are still few studies on the forecasting performance of the different models for historical time series data. Typically, comparisons have been carried out 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, 1997, 1999).

The aim of this paper is to compare the forecasting performance of alternative univariate models for the returns of three of the most traded currencies: the French franc (FF/$), the German mark (DM/$) and the Japanese yen (Y/$). Initially we conduct the analysis using data at different frequencies: monthly and weekly. As has been shown in a number of studies, non-linearities are more evident if a series is observed at high frequencies (daily and weekly), while they tend to disappear for series observed at lower frequencies (monthly and quarterly); this result reflects the effects of temporal aggregation or systematic sampling as documented in Weiss (1984) and Granger and Lee (1999). The use of data at different frequencies also allows us to evaluate such effects from a forecasting perspective.[1]

The evaluation of the forecast performance of the models is conducted according to different criteria. We first evaluate the average performance of the monthly and weekly models on mean square forecast errors (MSFEs), *unconditionally*, over the whole forecast period. Then, we evaluate the weekly models *conditioning* the forecast observations on being in each of the regimes of the SETAR models; this analysis should better exploit potential gains of the SETAR models in specific regimes. Finally, in order to take into account the varying degrees of uncertainty associated with the point forecasts, we compare the models on their ability to produce correct density forecasts. We do this by implementing the methodology recently proposed by Diebold *et al*. (1998) and surveyed by Tay and Wallis (2000). Applications of the methodology can be found in Diebold *et al*. (1999) and Clements and Smith (2000, 2001).

[1] An analysis of these series at a daily frequency was conducted in Boero and Marrocu (2000).

The rest of the paper is organised as follows. In the next section we present a review of recent work on exchange rate determination and forecasting. The methodological issues, the models adopted and the tests performed to detect the presence of non-linearities are described in the third section. The statistical properties of the data are discussed in the fourth section, while the findings from the modelling and forecasting exercises are reported in the fifth section. Finally, we summarize the main results and make concluding remarks.

LITERATURE REVIEW

An extensive empirical literature documents the finding that although exchange rate changes are weakly autocorrelated, there are strong dependencies in the data. In most empirical studies, which refer to the post-Bretton Woods floating exchange-rate regime, it is shown that the nature of the dependency can be adequately represented by the autoregressive conditional heteroskedastic (ARCH) model proposed by Engle (1982), or by its generalization represented by the GARCH model, suggested by Bollerslev (1986). This class of models is 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; however, this kind of dependency can be exploited only to improve interval or density forecasts, but not point forecasts. An improvement in point forecasts can be achieved by the GARCH in Mean (GARCH-M) model, where the conditional variance estimate enters as a regressor in the mean equation of the series. Recently, many authors have also stressed the empirical relevance of non-linearity in mean for the exchange rate returns; we refer, among others, to Meese and Rose (1991), Kräger and Kugler (1993), Peel and Speight (1994), Chappell *et al*. (1996) and Brooks (1997). However, the significant presence of mean non-linearities for the in-sample period only rarely has provided better out-of-sample forecasts compared with 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.

Kräger and Kugler (1993) estimate threshold autoregressive models for the returns of the French franc, Italian lire, Japanese yen, German mark and Swiss franc, all quoted against the US dollar (weekly observations for the period 1980.6–1990.1). Kräger and Kugler find evidence of three different regimes, with the outer regimes exhibiting much higher estimated standard deviations than the inner regime, and argue that this finding is probably due to the central bank interventions aimed at avoiding excessive appreciation (first regime) or depreciation (third regime). The theoretical background of the empirical analysis presented in Kräger and Kugler is the rational expectations monetary model with stochastic intervention rules proposed by Hsieh (1989). Thus, a three-regime autoregressive model is considered a good candidate to approximate Hsieh's model, which, according to Kräger and Kugler, provides a better understanding of the managed floating exchange-rate regime than the target zone model (Krugman, 1991). In Hsieh's model central bank intervention is triggered by large exchange-rate changes, while in Krugman's model the intervention takes place when the level of the variable is in the vicinity of the bounds. In order to evaluate the relative importance of mean and variance non-linearities, GARCH models are also estimated for the variables listed above. Kräger and Kugler conclude that neither the threshold models nor the GARCH models prove successful in describing adequately the non-linearity present in the series.

Peel and Speight (1994) analyse the changes of the British pound exchange rate against the US dollar, the French franc and the Reichsmark for the interwar period (weekly data). Having

found strong evidence of a generic form of non-linearity in all the series, the authors proceed by estimating alternative non-linear models: GARCH, bilinear and threshold autoregressive models.[2] The forecasting performance of the models is evaluated only for a one-step-ahead horizon: the linear-ARCH models exhibit a lower MSFE compared to the bilinear models for all the series, but in the case of the pound/US dollar exchange rate the most accurate forecasts are provided by the threshold models.

The study in Chappell *et al.* (1996) differs from those presented in Kräger and Kugler (1993) and in Peel and Speight (1994) since it is focused on the forecasting performance of non-linear models fitted to the *levels*, rather than the *changes*, of some bilateral ERM exchange rates considered at daily frequency. It is important to stress that if the forecast assessment (for more than one step ahead) is carried out on the basis of criteria such as the MSFE, the choice of data transformations is not neutral as shown by Clements and Hendry (1993, 1995): evaluation in differences is penalising relative to evaluation in levels. The issue of whether to evaluate the forecasting performance for the differences or the levels of the series is distinct from the issue of whether to estimate a model in the differences rather than the levels. According to Chappell *et al.* (1996), the inherent design of the ERM, based on the existence of a band in which the exchange rate is allowed to fluctuate without intervention by the central banks, could be the rationale for the presence of at least one threshold. Thus, the exchange rate follows a random walk process within the band but stationary autoregressive processes in the proximity of the ceiling or the floor such that the whole process exhibits mean-reverting features. In this case the process is globally, but not locally, stationary.[3] In contrast with most of the studies in which it is documented that the forecasting superiority of the non-linear models is often confined to the one-step-ahead horizon, the SETAR models estimated by Chappell *et al.* (1996) yield noticeable gains outperforming the random walk and the linear model at horizons as long as five and ten steps ahead.

Brooks (1996, 1997) analyses the daily British pound/US dollar exchange rate returns for the period 1974.1–1994.7. The main findings are that the non-linear models adopted, namely GARCH, SETAR and bilinear, produce forecasts only marginally more accurate than the ones obtained from a random walk model for all the horizons considered (up to 20 steps ahead). Moreover, on the basis of the Pesaran–Timmermann (1992) test, Brooks shows that the estimated models do not feature any *market timing ability*.

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. These are (1) non-linearities concern the even-ordered conditional moments and therefore are 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, and (4) non-linearities are present but they are captured by the wrong type of non-linear 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. On the basis of an extensive Monte Carlo study, Clements and Smith (2001), using the

[2] The threshold autoregressive models estimated by Peel and Speight (1994) exhibit three regimes with symmetric thresholds for the exchange rate against the dollar and two regimes for those against the French franc and the Reichsmark.

[3] Pippinger and Goering (1993) show that the Dickey–Fuller test for time series which exhibit the same behaviour described above has a very low power leading to a more frequent acceptance of the null of non-stationarity.

SETAR specifications of Kräger and Kugler (1993) discussed above, show that whether the non-linearities present in the data can be exploited to forecast better than a random walk depends both on how forecast accuracy is measured and on the *state of nature*. As suggested in their study, the evaluation of the whole forecast density may reveal gains to the non-linear models that are systematically masked if the comparison is carried out only in terms of MSFE.

Dacco and Satchell (1999) point out the predominance of the *random walk* model in forecasting exchange rates is mostly based on MSFE measures. Therefore, they suggest that the method of evaluation has to be chosen according to the nature of the problem examined. Methods based on the profitability criterion should turn out to be more adequate in the case of financial variables. Tests for the percentage of correct sign predictions, such as that proposed by Pesaran and Timmermann (1992), are expected to be more informative in deciding whether to buy or sell foreign currencies.

METHODOLOGY

In this section we present the models adopted to describe and forecast the exchange rate returns and the tests applied to detect the presence of non-linear features in the analysed series.

The models

The threshold autoregressive models

Threshold autoregressive models were first proposed by Tong (1978, 1983, 1990) and Tong and Lim (1980). The essential idea of this class of non-linear model 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, the model is known as *self-exciting*, hence the acronym SETAR.

Note that the threshold variable y_{t-d} is continuous on $\Re$, so that partitioning the real line defines the number of regimes that the process may follow:

$$-\infty < r_0 < r_1 < \cdots < r_n < r_{n+1} < \infty$$

where the r_j are referred to as *thresholds*. Thus, a SETAR model is piecewise-linear in the space of the threshold variable, rather than in time. If the process is in the jth regime, the pth order linear autoregression is formally defined as:

$$y_t = \phi_0^{(j)} + \phi_1^{(j)} y_{t-1} + \cdots + \phi_p^{(j)} y_{t-p} + \varepsilon_t^{(j)} \quad \text{for } r_{j-1} \le y_{t-d} < r_j \tag{1a}$$

$$\varepsilon_t^{(j)} \sim \text{IID}(0, \sigma^{2(j)})$$

Note that, 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.

A variant of the SETAR model, extensively explored by Teräsvirta and Anderson (1992) and Granger and Teräsvirta (1993), can be obtained if the parameters are allowed to change smoothly over time. The resulting model is called a smooth threshold autoregressive model (STAR) and has the following general expression:

$$y_t = \pi_0 + \pi_1' x_t + (\theta_0 + \theta_1' x_t) F(y_{t-d}) + u_t$$

where the error is assumed to be n.i.d $(0, \sigma^2)$, $x_t = (y_{t-1}, \ldots, y_{t-p})'$, $\pi_1 = (\pi_{11}, \ldots, \pi_{1p})'$ and $\theta_1 = (\theta_{11}, \ldots, \theta_{1p})'$, and $F(.)$ is the transition function.

The most common specifications for the transition function are the logistic and the exponential:

$$F(y_{t-d}) = \{1 + \exp[-\gamma(y_{t-d} - r)]\}^{-1}$$
$$F(y_{t-d}) = 1 - \exp[-\gamma(y_{t-d} - r)^2]$$

In the logistic STAR (LSTAR) model the parameters change monotonically with y_{t-d}. When γ tends to infinity, $F(y_{t-d})$ becomes a Heaviside function which assumes value 0 if the threshold variable is equal or smaller than r and value 1 if it is greater than r; in this case the model becomes a SETAR model. On the other hand, if γ tends to zero, the STAR reduces to a linear AR(p) model. In the exponential STAR (ESTAR) case, the parameters change symmetrically about r with y_{t-d}. When γ tends to either infinity or zero the model becomes linear because, on the boundary, one regime has probability 1 and the other zero.

SETAR model estimation 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*). Alternatively, indicator functions, which implicitly constrain the residual error variance to be constant across regimes, can be employed and the model can be reformulated as follows:

$$\begin{aligned} y_t = {} & (\phi_0^{(1)} + \phi_1^{(1)} y_{t-1} + \cdots + \phi_p^{(1)} y_{t-p})(1 - I[y_{t-d} > r]) \\ & + (\phi_0^{(2)} + \phi_1^{(2)} y_{t-1} + \cdots + \phi_p^{(2)} y_{t-p})(I[y_{t-d} > r]) + \varepsilon_t \end{aligned} \tag{1b}$$

where $I[\mathrm{A}]$ is an indicator function with $I[\mathrm{A}] = 1$ if the event A occurs and $I[\mathrm{A}] = 0$ otherwise.

In the more common case, in which the threshold parameter (r) and the delay parameter (d) are unknown, Tong (1983) suggests an empirical procedure that allows selecting as the 'best' model the one which yields the minimum Akaike Information Criteria (AIC). However, as stressed by Priestley (1988), such a procedure has 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 is 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, the re-estimation of the separate AR models allows the determination of the r parameter, as the one for which AIC(d) attains its minimum value. In stage three the search over d is carried out by repeating both stage 1 and stage 2 for $d = d_1, d_2, \ldots, d_p$. The selected value of d is, again, the value that minimizes AIC(d).

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:

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

where x_t is a vector of $k \times 1$ explanatory variables, which in our study includes only lagged values of y_t, and β is a $k \times 1$ 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 $y_{t-1}, y_{t-2}, \ldots . y_{t-k}$. In particular, the model is based on the assumption that:

$$\varepsilon_t|\Psi_{t-1} \sim \text{N}(0, h_t) \tag{3}$$

where $h_t = \alpha_0 + \alpha_1\varepsilon_{t-1}^2 + \cdots + \alpha_q\varepsilon_{t-q}^2$ with $\alpha_0 > 0$ and $\alpha_i \geq 0$, $i = 1, \ldots, q$, 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 generalization of the ARCH model, which leads to the following specification of the conditional variance:

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

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 $i = 1, \ldots, q$, and $\beta_i \geq 0$ for $i = 1, \ldots, p$.[4] 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.

GARCH in mean

Engle *et al*. (1987) extend the ARCH model by introducing the conditional variance as a regressor in the mean equation of the variable:

$$y_t = x_t'\beta + \delta h_t + \varepsilon_t \quad t = 1, \ldots, T \tag{5}$$

where $\varepsilon_t|\Psi_{t-1} \sim \text{N}(0, h_t)$ and h_t is a (G)ARCH process.

In the (G)ARCH-M model the conditional variance is included in the mean equation according to different functional forms: $\log(h_t)$, $\sqrt{h_t}$ and h_t.

Asymmetric GARCH A relevant extension of the GARCH models is represented by the class of asymmetric models. These allow one to capture possible asymmetries in the conditional variance induced by the sign and the magnitude of past shocks. Most applied specifications are the threshold heteroscedastic model (TARCH) (Glosten *et al*., 1993; Zakoian, 1994) and the Exponential GARCH model (EGARCH; Nelson, 1991).

[4] These are sufficient restrictions for the conditional variance to be positive, but they are not necessary (see Nelson and Cao, 1992).

Linearity tests

Testing for linearity is not a "standard testing situation" (Granger, 1993, p. 237) due to the characteristics of the null and alternative hypotheses. In fact, the specification of the null of linearity involves a very large number of parameters, while the alternative implies a 'huge' number of models, each one with many parameters (Granger, pp. 236–237). Therefore, Granger suggests a testing strategy based on the application of a battery of tests. Each test needs not to be interpreted formally, but rather as an indication of the presence of non-linear features in the data. In order to detect the presence of non-linear components in the returns series of the French franc, the German mark and the Japanese yen, we apply four different linearity tests: the RESET test, the Tsay (1986) test, the S_2 test proposed by Luukkonen *et al*. (1988), and the McLeod and Li (1983) test. All the tests are devised for the null hypothesis of linearity.

The RESET test is applied in its traditional form (Ramsey, 1969) and in the modified version found to be superior by Thursby and Schmidt (1977). In its original form, 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. The modified RESET test requires that all the initial regressors enter linearly and up to a certain power h in the auxiliary regression; Thursby and Schimdt suggest using $h = 4$. The Lagrange multiplier form (Granger and Teräsvirta, 1993) of the test is adopted in this study, thus the test is distributed as a χ^2 with up to $3p$ degrees of freedom for the modified version.

The Tsay (1986) test belongs to the class of tests based on Volterra expansions (Priestley, 1980) and represents a generalization of the Keenan (1985) test. Under the null hypothesis the series is described by a linear AR(p) model, while the auxiliary regression includes squares and cross-products terms at different lag lengths, such as $y_{t-i}y_{t-j}$. Tsay shows that the test is more powerful than Keenan's test. Under the null hypothesis, which imposes the condition that all the coefficients of the non-linear terms are jointly equal to zero, the test is distributed as an F with $p(p+1)/2$ and $(T\text{-}p\text{-}p(p+1)/2\text{-}1)$ degrees of freedom.

While the RESET and the Tsay tests are devised for a generic form of misspecification, the S_2 test, suggested by Luukkonen *et al*. (1988), is formulated for a specific alternative hypothesis, i.e. STAR-type non-linearity. However, the authors show that the S_2 test has reasonable power even when the true model is a SETAR one. The S_2 test follows a χ^2 with $p(p+1)/2 + 2p^2$ degrees of freedom and 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 following model:

$$y_t + \beta_0 + \beta' w_t + \sum_{i=1}^{p}\sum_{j=1}^{p} \xi_{ij} y_{t-i} y_{t-j} + \sum_{i=1}^{p}\sum_{j=1}^{p} \psi_{ij} y_{t-i} y_{t-j}^2 + \sum_{i=1}^{p}\sum_{j=1}^{p} \kappa_{ij} y_{t-i} y_{t-j}^3 = \varepsilon_t \qquad (6)$$

where the vector w_t includes the lags of y_t.

The maximum lag p is usually unknown and is determined from the data according to some model selection criterion (AIC). If the true model is non-linear, it is possible that the maximum lag in the AR(p) model is greater than the one in the non-linear model and this can lower the power of the test compared with the case in which p is known. On the other hand, if p is so low that the linear model has autocorrelated residuals, the test is then biased towards rejection if the true model is linear, because the test also has power against serially correlated errors (Teräsvirta, 1994). If d is assumed to be known, y_{t-d} can be substituted in the above auxiliary regression for y_{t-j}, and the resulting test has a χ^2 distribution with $3p$ degrees of freedom under the null. Note that even

when d is fixed the test statistic requires a large number of degrees of freedom if the lag p in the linear model is high. For this reason, the S_2 test is applied assuming that the delay parameter d is known and takes values in the range [1,6]. If the null is rejected for different values of d, the delay parameter is selected as the one that yields the lowest *probability* value.

To check for the presence of non-linearity in variance we perform the McLeod and Li (1983) test. The test is similar to the test proposed by Ljung and Box (1978); both tests are based on the estimation of the correlation function of the squared residuals obtained from a linear model. Granger and Anderson (1978) point out that even if the residuals from Box–Jenkins (1976) linear models can appear to be non-correlated, the squared residuals are often correlated if the series is non-linear. Therefore, according to Granger and Anderson (1978) the autocorrelation function is a useful tool for identifying and selecting non-linear models. The McLeod and Li test is calculated as follows. The estimated residual series $\hat{u}_t$ is obtained from the *best* AR(p) or ARMA(p,q) model and the autocorrelation function is computed according to the following expression:

$$\hat{r}_u^2(k) = \frac{\sum_{t=k+1}^{T} (\hat{u}_t^2 - \hat{\sigma}^2)(\hat{u}_{t-k}^2 - \hat{\sigma}^2)}{\sum_{t=1}^{T} (\hat{u}_t^2 - \hat{\sigma}^2)^2} \quad \text{where } \hat{\sigma}^2 = \frac{1}{T}\sum \hat{u}_t^2$$

The test is computed as the *portmanteau* $Q^*(m)$ statistics,

$$Q^*(m) = T(T+2)\sum_{i=1}^{m} \frac{\hat{r}_u^2(i)}{(T-i)}$$

which is distributed asymptotically as a χ^2 with m degrees of freedom, if the estimated residuals $\hat{u}_t$ are independent.[5]

PRELIMINARY DATA ANALYSIS

The empirical analysis has been carried out on the exchange rate returns measured in log-differences. The monthly series for the log-levels and the returns for the period 1973.1–1997.7 (294 observations) are depicted in Figure 1. Figure 1 also shows the behaviour of the weekly series over the same period (1281 observations). All the returns series are mean-stationary, while the variance features the typical *volatility clustering* phenomenon with periods of high volatility followed by periods of low volatility, and this is particularly evident for weekly series. If such a feature turns out to be relevant we expect the GARCH models to capture it adequately.

Table I reports the summary of the descriptive statistics for the exchange rate returns. All the series, especially those at weekly frequencies, are characterized by excessive kurtosis and asymmetry. The Jarque–Bera test strongly rejects the normality hypothesis for all the series.

Tables II(a) and II(b) report the probability values for the linearity test performed on the log-differences of the series. For each test the linear model under the null hypothesis has been estimated

[5] Note, however, that there are suggestions that the asymptotic distribution of this test may not be invariant to heteroscedasticity (see Li and Mak, 1994).

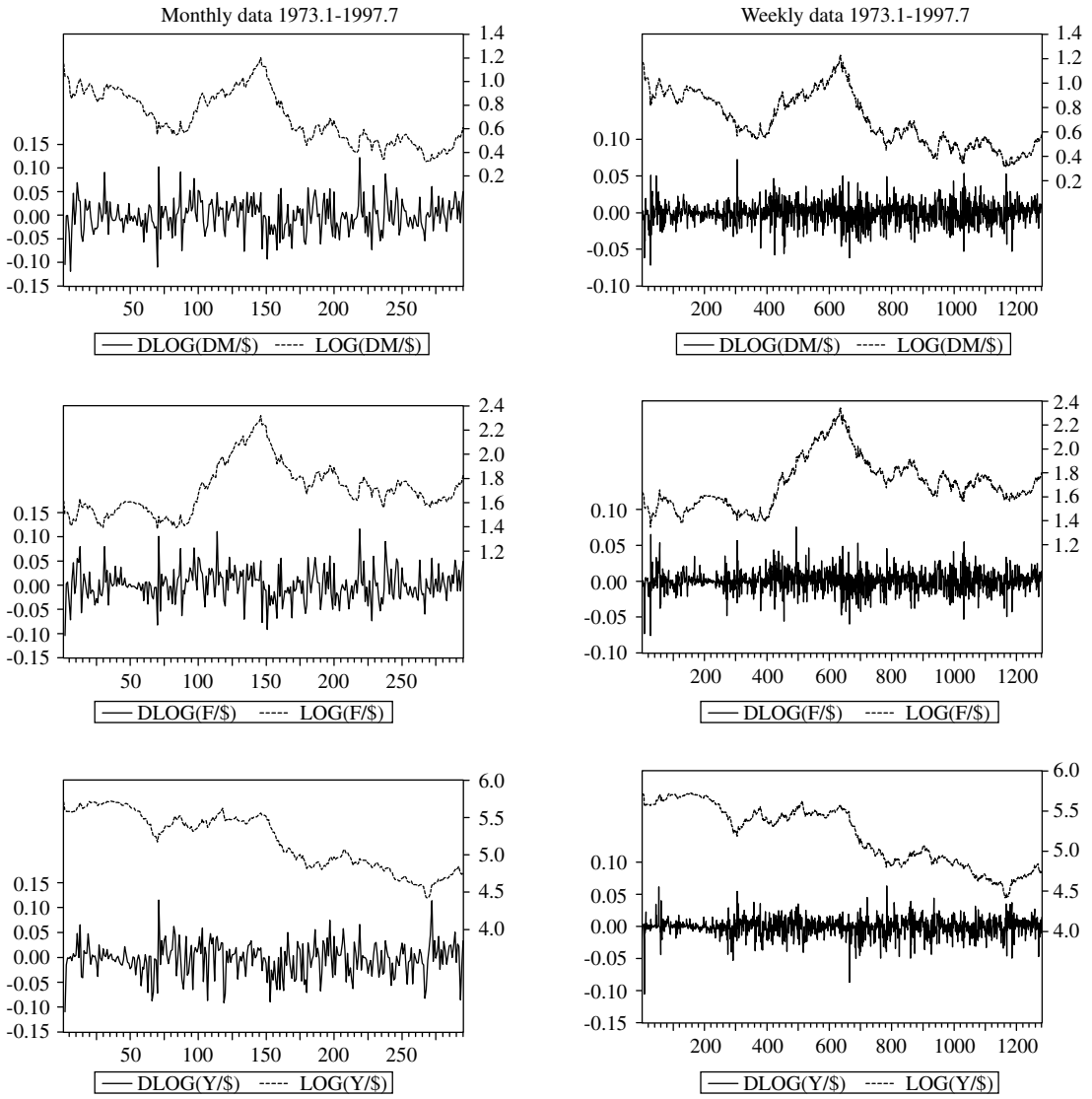

Figure 1. Exchange rate log-levels (dotted lines) and returns (solid lines)

assuming different lag structures ($p = 2, \ldots, 6$). As pointed out by Teräsvirta (1994), if the 'true' model is non-linear it is possible that the maximum lag in the AR(p) model is greater than the lag order in the non-linear model and this can lower the power of the test compared with the case in which p is known. In Tables II(a) and II(b) we report the results only for $p = 4, 5, 6$.

The RESET test, in both its traditional and modified version, has been carried out according to three different specifications. The first includes only the squared terms (fitted values or initial regressors), the second adds the cubic terms, and the third includes terms up to the fourth power. Focusing first on *monthly* series (Table II(a)), the S_2 test, performed under the specific alternative

Table I(a). Descriptive statistics for the exchange rate returns (monthly data)

	F/$	DM/$	Y/$
Mean	0.000700	−0.001851	−0.003179
Median	−0.001087	−0.002323	0.000851
Maximum	0.116373	0.121743	0.115344
Minimum	−0.104154	−0.118541	−0.109147
Std dev.	0.033329	0.034553	0.033336
Skewness	0.210835	0.008767	−0.241786
Kurtosis	3.968137	4.075787	3.954479
Jarque–Bera	13.65990	14.18089	14.02467
Probability	0.001081	0.000833	0.000901
Observations	294	294	294

Table I(b). Descriptive statistics for the exchange rate returns (weekly data)

	F/$	DM/$	Y/$
Mean	−0.000435	0.000147	−0.000740
Median	−0.000118	0.000000	0.000000
Maximum	0.072321	0.075292	0.063120
Minimum	−0.071506	−0.075878	−0.105679
Std dev.	0.015196	0.014831	0.014186
Skewness	−0.166860	−0.069270	−0.702024
Kurtosis	4.932147	5.754028	7.815579
Jarque–Bera	205.2034	405.8562	1342.976
Probability	0.000000	0.000000	0.000000
Observations	1281	1281	1281

Table II(a). Linearity tests—*P*-values (monthly data)

	French franc			German mark			Japanese yen		
p	4	5	6	4	5	6	4	5	6
RESET-2	0.404	0.899	0.271	0.236	0.267	0.660	0.383	0.429	0.217
RESET-3	0.673	0.570	0.095	**0.032**	0.228	0.895	0.311	0.395	0.374
RESET-4	0.687	0.685	0.182	**0.075**	0.372	0.962	0.420	0.520	0.448
Mod.RESET-2	0.764	0.982	0.930	0.718	0.894	0.893	0.645	0.120	0.132
Mod.RESET-3	0.336	0.563	0.465	0.156	0.492	0.152	0.852	0.432	0.485
Mod.RESET-4	**0.085**	0.208	0.228	0.194	0.574	**0.060**	0.796	0.460	0.602
Tsay[a]	0.315	0.428	0.472	0.749	0.819	0.739	0.554	0.334	**0.097**
S_2, $d = 1$	**0.004**	**0.01**	**0.026**	**0.000**	**0.000**	**0.001**	**0.046**	**0.054**	0.121
S_2, $d = 2$	0.508	0.505	0.516	0.517	0.311	**0.059**	0.243	0.299	0.257
S_2, $d = 3$	0.877	0.921	0.837	0.980	0.970	0.990	0.970	0.453	0.557
S_2, $d = 4$	0.312	0.441	0.667	0.190	0.536	0.642	0.645	**0.046**	**0.069**
S_2, $d = 5$	0.144	0.194	0.144	**0.005**	0.216	0.115	0.245	0.154	**0.076**
S_2, $d = 6$	0.192	0.576	0.866	**0.069**	0.781	0.970	0.587	0.713	0.442
$Q^*(12)$	0.883	0.991	0.998	**0.007**	0.186	**0.045**	**0.000**	**0.000**	**0.000**

Table II(b). Linearity tests—*P*-values (weekly data)

	French franc			German mark			Japanese yen		
p	4	5	6	4	5	6	4	5	6
RESET-2	0.906	0.983	0.677	0.113	0.116	0.524	**0.036**	**0.028**	**0.061**
RESET-3	**0.060**	0.344	0.475	**0.000**	**0.000**	**0.023**	**0.025**	**0.013**	**0.006**
RESET-4	0.131	0.520	0.518	**0.000**	**0.000**	**0.058**	**0.055**	**0.011**	**0.003**
Mod.RESET-2	**0.042**	**0.047**	**0.075**	**0.057**	**0.034**	**0.058**	0.166	0.193	0.160
Mod.RESET-3	**0.004**	**0.001**	**0.001**	**0.015**	**0.003**	**0.005**	0.323	0.099	**0.070**
Mod.RESET-4	**0.014**	**0.002**	**0.002**	**0.047**	**0.015**	**0.027**	0.099	**0.011**	**0.005**
Tsay[a]	**0.025**	**0.035**	**0.092**	**0.022**	**0.024**	**0.024**	0.099	**0.024**	**0.000**
S_2, $d = 1$	**0.001**	**0.000**	**0.000**	**0.007**	**0.002**	**0.003**	**0.032**	**0.000**	**0.000**
S_2, $d = 2$	**0.001**	**0.001**	**0.002**	**0.018**	**0.026**	**0.022**	**0.001**	**0.002**	**0.002**
S_2, $d = 3$	0.548	0.764	0.577	**0.000**	**0.001**	**0.002**	0.093	0.169	**0.084**
S_2, $d = 4$	0.342	**0.038**	**0.064**	0.308	**0.015**	**0.023**	0.637	0.388	0.231
S_2, $d = 5$	0.146	**0.082**	**0.052**	0.103	0.110	0.194	**0.000**	**0.000**	**0.000**
S_2, $d = 6$	**0.000**	**0.000**	**0.004**	**0.000**	**0.000**	**0.001**	**0.000**	**0.000**	**0.001**
$Q^*(12)$	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**

p is the autoregressive lag order under the null hypothesis of linearity.
[a] The Tsay test follows an *F*-distribution, while all the other tests are distributed as χ^2.

hypothesis of STAR-type non-linearity, indicates the presence of non-linearities in a number of cases. The test is robust with respect to the autoregressive structure and the lowest *P*-values are found for $d = 1$. The results of the RESET and Tsay tests, devised for a generic alternative hypothesis of non-linearity, lead to a rejection of the null only in a small number of cases. Finally, the McLeod–Li test shows the presence of non-linearity in variance for the Japanese yen and for the German mark. Although the test was carried out for $m = 1, 2, \ldots, 12$, we report results only for $m = 12$, since the statistic did not change appreciably for different values of m.

Turning to *weekly* series, the linearity tests (Table II(b)) lead to the rejection of the null in a higher number of cases compared with the monthly series. There is, therefore, stronger evidence of non-linear components for the high-frequency data. In particular, the RESET test in the Thursby–Schmidt version turns out to be more powerful than the traditional formulation, allowing rejection of the null hypothesis in a larger number of cases with respect to the French franc and the German mark. The Japanese yen series, on the other hand, appears more non-linear if the evidence is based on the classic RESET test.

The results of the S_2 test show rejections in most cases, with only few exceptions depending on the selected delay parameter d. The Tsay test and the McLeod–Li test, for which the results are reported for $m = 12$, leads to rejection of linearity for all the series, regardless of the dynamic structure adopted.

EMPIRICAL RESULTS

Model estimation results

Linear models

The models were estimated over the period 1973.2–1991.6. Examination of the in-sample returns data revealed more significant serial correlation in the weekly than in the monthly series. The

monthly returns of the German mark and Japanese yen showed marginal evidence of serial correlation at lags 2 and 3, respectively. Thus, on the basis of the AIC and SC criteria an AR(2) model was selected for the German mark and an AR(3) for the yen. Both specifications were restricted with intermediate autoregressive coefficients equal to zero and marginally preferred to a random walk with drift specification (lag order equal zero). More distinct evidence of serial correlation emerged in the monthly returns of the French franc, for which a restricted AR(4) specification was clearly selected against the random walk with drift.

For weekly returns, we selected a restricted AR(2) model for the French franc and the Japanese yen, and an unrestricted AR(2) for the German mark.

SETAR and STAR models

With regard to the SETAR models, we estimated specifications with one threshold (two regimes) and two thresholds (three regimes), following the estimation procedure suggested by Tong (1983).[6] The model selection has been 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. Moreover, we considered only models with a maximum lag order $p = 6$. The models selected are reported in Table III.[7] In general, the dynamic structure and the estimated error variance differ across regimes, indicating that both monthly (Table III(a)) and weekly (Table III(b)) series are strongly characterized by non-linearities.

Looking in more detail at the results for monthly data, according to the two-regime specification, the French franc returns are described by an AR(2) process in the first regime and simply by a constant term in the second regime. For the three-regime model, the French series follows an AR(2) process in the middle regime, while it is represented by a constant in the outer regimes of strong appreciation and depreciation, respectively. We also note that for this specification the estimated standard deviation is considerably lower in the inner regime. Noticeable differences across regimes were also found in the case of the German mark and the Japanese yen. Moreover, the dynamics of the three-regime SETAR model for the latter currencies are very similar to those discussed above for the French franc. These results are in line with the theoretical model described in Hsieh (1989) and with the empirical evidence reported by Kräger and Kugler (1993).

Further evidence of this kind of non-linearity is provided by the SETAR specifications selected for the weekly data, and shown in Table III(b). It is also interesting to note that for all SETAR specifications the estimated total standard deviation is always smaller than the one obtained from linear models, which to facilitate the comparison has also been reported in the last columns of Tables III(a) and III(b). The improved goodness-of-fit of the non-linear models with respect to the AR models is driven by the lower residual standard deviation for the upper regime in the SETAR-2 models, and for the middle regime in the SETAR-3 specification.

For weekly frequencies data we have also estimated STAR models, specifically a logistic STAR for the French franc and the German mark, and an exponential STAR for the Japanese yen. The results of this estimation are reported in Table III(c).

[6] All the models have been selected and estimated with Eviews codes; the codes are available from the authors upon request. Note that in carrying out the selection procedure we allow each regime to include at least 10% of the total number of observations included in the estimation sample.

[7] Note that even the simplest SETAR models with an AR(1) process in each regime can generate complex dynamic behaviour. Moreover, it is worth stressing that the constant term plays a relevant role in non-linear models.

Table III(a). SETAR models for the exchange rate returns (monthly data)

SETAR	AIC	Thresholds	d	Total st. dev.	Regime 1			Regime 2			Regime 3			Linear st. dev.
					T	No. coeff.[a]	St. dev.	T	No. coeff.[a]	St. dev.	T	No. coeff.[a]	St. dev.	
French franc														
SETAR-2	−6.8387	−0.0019	4	0.0325	98	3	0.0313	117	1	0.0334				0.0334
SETAR-3	−6.8746	−0.0046; 0.0065	3	0.0330	89	1	0.0342	38	3	0.0199	88	1	0.0361	
German mark														
SETAR-2	−6.7768	−0.0077	1	0.0334	91	2	0.0331	124	3	0.0336				0.0351
SETAR-3	−6.8228	−0.0131; −0.0032	2	0.0340	73	2	0.0389	31	5	0.0181	111	1	0.0339	
Japanese yen														
SETAR-2	−6.8047	0.0030	3	0.0332	115	1	0.0346	100	1	0.0316				0.0334
SETAR-3	−6.8480	−0.0030; 0.0044	1	0.0322	92	1	0.0344	30	5	0.0284	93	1	0.0312	

Table III(b). SETAR model specifications for the exchange rate returns (weekly data)

SETAR	AIC	Thresholds	d	Total st. dev.	Regime 1			Regime 2			Regime 3			Linear st. dev.
					T	No. coeff.[a]	St. dev.	T	No. coeff.[a]	St. dev.	T	No. coeff.[a]	St. dev.	
French franc														
SETAR-2	−8.479	−0.0125	1	0.0145	145	2	0.0185	813	7	0.0137				0.0149
SETAR-3	−8.531	−0.0089; 0.0118	1	0.0145	210	2	0.0172	583	4	0.0120	165	7	0.0181	
German mark														
SETAR-2	−8.4408	−0.0108	1	0.0149	196	1	0.0189	762	4	0.0137				0.0151
SETAR-3	−8.4822	−0.0108; 0.0123	1	0.0149	195	1	0.0189	613	4	0.0125	150	1	0.0174	
Japanese yen														
SETAR-2	−8.5704	0.0072	3	0.0138	736	3	0.0134	222	1	0.0150				0.0143
SETAR-3	−8.6490	−0.0032; 0.0057	1	0.0137	332	1	0.0164	361	3	0.0103	265	1	0.0140	

[a] Number of estimated coefficients including the constant.

Table III(c). STAR model specifications for the exchange rate returns (weekly data)

STAR	Threshold	d	Total st. dev.	γ	AR linear terms[a]	AR terms multiplying $F(y_{t-d})$[a]
French franc	−0.0058	1	0.0144	2.82	2	5
German mark	−0.0027	1	0.0149	6.96	2	3
Japanese yen	−0.0025	1	0.0136	2.40	5	3

[a] Number of estimated coefficients including the constant.

GARCH models

While GARCH components turned out to be significant only for the yen returns when considering the data at monthly frequencies, they were strongly present in all the weekly series, thus capturing the evident volatility clustering illustrated in Figure 1. In order to describe appropriately such components, we identified some alternative models—namely, a simple GARCH(1,1), an EGARCH(1,1) and a TARCH(1,1)—to take into account possible asymmetries in the conditional variance, and a GARCH in mean (GARCH-M(1,1)). The *best* model was selected according to the Akaike (AIC) and Schwarz (SIC) information criteria. A significant variance component was found in the mean equation for the monthly and weekly returns of the Japanese yen and for the weekly returns of the German mark. In these cases, the resulting joint estimated AR models for the mean of the returns were of a lower order (AR(1) processes) than that reported under AR estimation alone. In the case of the weekly returns of the French franc, the model order (AR(2)) continued to hold for jointly estimated GARCH, and a variance component was found only marginally significant in the mean equation. Therefore, as shown in the next section, the mean forecasts for this model, when computed over the whole forecast period, are virtually the same as those obtained from the AR linear model.

The forecasting exercise

The forecasting performance of the models is evaluated in different ways. First, we compute MSFEs for the various models for different steps ahead and compare the relative performance of the models by means of the Diebold and Mariano (1995) test. This exercise is conducted over the forecast period as a whole, and we refer to it as *unconditional* (without conditioning on being in a particular regime). Second, following other authors (Tiao and Tsay, 1994; Clements and Smith, 1999, 2001), we analyse the forecasting performance of the models *conditional* on the regimes of the SETAR models. This evaluation, also conducted on MSFEs, will explore whether the SETAR and STAR models show a better performance for observations falling in specific regimes. Third, we supplement the evaluation on MSFEs by assessing the ability of the models to produce 'correct' density forecasts. For this analysis we employ recently developed techniques discussed in Diebold *et al.* (1998) and surveyed by Tay and Wallis (2000). Previous applications of these techniques have shown some gains of SETAR models over linear counterparts, using Monte Carlo methods (see Clements and Smith, 2000, 2001). It is therefore interesting to see whether these gains hold in our evaluation with historical data. Furthermore, models of conditional variance such as GARCH are mostly useful when the object of the analysis is also to provide some indication of the uncertainty around the mean. Potential gains of the GARCH models over the linear models can therefore be better exploited in a comparative evaluation of density forecasts.

Unconditional point forecasts

The forecasts of the exchange rate returns have been calculated recursively for the monthly and weekly series from one to 24 steps ahead and from one to five steps ahead, respectively. The models were identified and specified only once, over the first estimation periods, 1973.2–1991.6. They were then re-estimated (but not re-specified) by expanding the sample with one observation each time, over the period 1991.7–1997.7, thereby obtaining 50 point forecasts in the case of monthly data and 313 point forecasts with weekly data for each forecasting horizon (h).

We recall that the computation of multi-step-ahead forecasts ($h > 1$) from non-linear models (SETAR) involves the solution of 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[8] method, following the suggestions in Clements and Smith (1997, 1999).

In this comparative exercise the model forecasting ability is firstly assessed by means of the MSFE and in terms of the percentage of correct sign predictions. For returns series, in fact, it may be particularly informative to accompany standard measures of forecast accuracy, such as the MSFE, with an indicator of the number of times the sign is correctly predicted.[9] The results of our forecasting exercise (MSFE and Sign) are reported in Table IV, while in Table V we present the MSFE normalized 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

Table IV(a). Forecasting performance (monthly data)

	Number of steps ahead											
	1		3		6		9		12		24	
	MSFE	Sign	MSFE	Sign	MSFE	Sign	MSFE	Sign	MSFE	Sign	MSFE	Sign
French franc												
Naïve	**10.860**	—	**10.733**	—	10.636	—	**8.990**	—	8.979	—	7.190	—
Linear AR(4)	11.560	0.44	11.089	0.50	10.628	0.34**	9.000	0.34**	9.000	0.34**	7.182	0.42
SETAR-2	12.532	0.44	11.492	**0.56**	**10.563**	0.44	9.303	**0.40**	**8.880**	**0.48**	7.076	0.50
SETAR-3	11.492	**0.54**	11.424	0.46	10.628	**0.54**	9.364	0.34**	**8.880**	0.38	**7.023**	**0.56**
German mark												
Naïve	10.640	—	**10.584**	—	10.510	—	8.976	—	9.003	—	7.575	—
Linear AR(2)	11.022	0.54	10.628	**0.60**	10.498	**0.58**	**8.940**	**0.58**	7.508	**0.60**	7.508	**0.52**
SETAR-2	**10.240**	**0.58**	10.890	0.54	**10.240**	0.56	9.610	0.52	**7.290**	0.58	**7.290**	0.50
SETAR-3	11.560	0.38*	10.758	0.46	10.433	0.52	9.303	**0.58**	**7.290**	0.58	**7.290**	**0.52**
Japanese yen												
Naïve	10.296	—	10.618	—	10.873	—	10.589	—	10.442	—	11.441	—
Linear AR(3)	10.563	0.56	10.890	0.52	10.890	0.54	**10.563**	0.56	10.433	0.56	**11.357**	0.48
GARCH-M	**9.672**	0.62	**10.433**	**0.58**	**10.824**	0.54	10.628	0.56	10.498	0.56	11.424	0.48
SETAR-2	11.022	0.42	11.290	0.44	**10.824**	0.54	10.890	0.56	10.433	0.56	11.560	0.48
SETAR-3	11.424	**0.64****	11.156	0.50	**10.824**	0.54	10.956	0.56	10.433	0.56	11.560	0.48

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

[9] As pointed out by an anonymous referee, recent work by Christoffersen and Diebold (2001) has shown that sign forecasting may be comparatively simple and does not imply any superiority in mean forecasting.

Table IV(b). Forecasting performance (weekly data)

	Number of steps ahead									
	1		2		3		4		5	
	MSFE	Sign	MSFE	Sign	MSFE	Sign	MSFE	Sign	MSFE	Sign
French franc										
Naïve	2.1333	—	**2.1329**	—	2.1262	—	2.1147	—	2.1308	—
Linear AR(2)	2.1461	0.51	2.1470	**0.51**	**2.1260**	**0.51**	**2.1145**	**0.52**	2.1309	0.50
GARCH-M	2.1461	0.51	2.1470	**0.51**	**2.1260**	**0.51**	**2.1145**	**0.52**	2.1309	0.50
SETAR-2	**2.1120**	**0.53**	2.2261	0.47	2.1780	0.45*	2.1619	0.48	2.1316	0.51
SETAR-3	2.2053	0.48	2.2144	0.50	2.1889	0.47	2.1467	0.49	**2.1262**	**0.52**
LSTAR	2.1804	0.48	2.2418	0.49	2.1680	0.48	2.1675	0.47	2.1338	0.49
German mark										
Naïve	**2.3214**	—	2.3210	—	2.3129	—	2.3004	—	2.3215	—
Linear AR(2)	2.3276	0.50	2.3278	0.50	2.3124	0.50	2.2999	0.49	2.3215	0.49
GARCH-M	2.3549	0.49	**2.3116**	0.50	**2.2964**	**0.51**	**2.2880**	0.50	2.3090	0.50
SETAR-2	2.3784	0.50	2.3778	**0.53**	2.3513	0.48	2.3384	0.50	**2.2848**	0.50
SETAR-3	2.4204	0.50	2.4383	0.51	2.3817	0.49	2.3182	0.51	2.2866	0.49
LSTAR	2.3663	0.50	2.4095	**0.53**	2.3422	0.50	2.3464	**0.52**	2.2864	0.50
Japanese yen										
Naïve	1.8917	—	1.8944	—	1.8915	—	1.8985	—	1.8992	—
Linear AR(2)	1.8929	0.51	1.8892	0.53	1.8837	0.49	1.8948	0.49	1.8992	0.49
GARCH-M	1.8980	0.55	1.8903	0.55	1.8785	0.49	**1.8913**	0.49	1.8965	0.49
SETAR-2	1.8643	**0.55**	**1.8777**	0.51	1.8569	**0.55**	1.9306	0.46	1.8686	**0.54**
SETAR-3	1.9324	0.53	1.9384	0.56**	1.8847	0.51	1.9153	0.49	1.8747	0.50
ESTAR	1.9210	0.53	1.9220	**0.58****	**1.8525**	0.53	1.9192	0.48	**1.8617**	0.53

Note: The value of MSFE has been rescaled by multiplying by 10^4.
*, ** indicates statistically different from 0.50 at 10% and 5% respectively.

Table V(a). Normalized MFSE (monthly data)

	Number of steps ahead					
	1	3	6	9	12	24
French franc						
SETAR-2	1.084**	1.036*	0.994	1.034*	0.987	0.985
SETAR-3	0.994	1.030	1.000	1.040**	0.987	0.978
German mark						
SETAR-2	0.929	1.025*	0.975	1.075**	1.000	0.971
SETAR-3	1.049	1.012	0.994	1.041**	0.993	0.971
Japanese yen						
GARCH-M	0.916**	0.958*	0.994	1.006*	1.006	1.006
SETAR-2	1.044*	1.037**	0.994	1.031**	1.000	1.018
SETAR-3	1.082	1.024	0.994**	1.037**	1.000	1.018

whether such a superiority is statistically significant we perform the Diebold–Mariano (DM) test. Values leading to the rejection of the null hypothesis of equality of forecast accuracy are indicated with asterisks in Tables V(a) and V(b). Similarly, sign predictions that are statistically different

Table V(b). Normalized MSFE (weekly data)

	Number of steps ahead				
	1	2	3	4	5
French franc					
GARCH-M	1.000	1.000	1.000	1.000	1.000
SETAR-2	0.984	1.037**	1.024**	1.022*	1.000
SETAR-3	1.028	1.031**	1.030**	1.015	0.998
LSTAR	1.016	1.044**	1.020**	1.025**	1.001
German mark					
GARCH-M	1.012	0.993	0.993	0.995	0.995
SETAR-2	1.022**	1.021**	1.017*	1.017*	0.984**
SETAR-3	1.040**	1.047**	1.030**	1.008	0.985*
LSTAR	1.017	1.035**	1.013	1.020**	0.985**
Japanese yen					
GARCH-M	1.003	1.001	0.997	0.998	0.999
SETAR-2	0.985	0.994	0.986	1.019**	0.984**
SETAR-3	1.021	1.026**	1.001	1.011	0.987
ESTAR	1.015	1.017	0.983*	1.013	0.980**

Notes: The normalized MSFE is calculated as the ratio $MSFE_{NL}/MSFE_{L}$.
*, ** denote significance of the Diebold–Mariano test at 10% and 5%.

from a proportion of 50% (under the assumption of independence between forecasts and actual values) are indicated in Table IV with asterisks. Table IV also reports the MSFE obtained from a *naïve* forecast by assuming that the levels of the exchange rates follow a *random walk* with drift process.

Focusing first on *monthly* data, an interesting result is that the random walk model shows some gains in terms of MSFE over the other models only in a limited number of cases (for one, three, and nine steps ahead for the French franc and for the three steps ahead for the German mark). In fact, the oft-claimed superiority of the random walk, as emphasized in previous studies, is questioned by the more accurate forecasting performance of the non-linear models as appears from the highlighted values reported in Table IV(a). These models also show some gains over the linear AR(p) model in terms of either MSFE or Sign, although there are only few cases when the proportion of correct sign predictions is greater than 50%. Moreover, in the case of the French franc and the German mark, gains of non-linear models appear to be only marginal when assessed by means of the DM test. A close look at Table V(a), in fact, shows a larger number of values greater than one that indicate a significantly better performance of the linear models.

A more supportive picture in favour of the non-linear models emerges for the Japanese yen. In particular, the SETAR-3 model shows the highest percentage of correct sign predictions in the one-step-ahead, and offers significant forecasting gains over the linear model in the six-step-ahead forecasts. Even more substantial gains are obtained in terms of MSFEs with the GARCH-M model in the one- and three-step-ahead forecasts. This latter result reflects a significant contribution of the mean component of the GARCH model in forecasting the conditional mean of this variable.

Turning now to the *weekly* forecasts, in Tables IV(b) and V(b) we report the results for the one-to five-step-ahead forecasts. For this comparison, we have extended the forecasting experiment to include an alternative specification of threshold models, namely the STAR models. As pointed out above, one of the tests presented in Table II(b), the S_2 test, showed strong evidence in favour of

a STAR form of non-linearity. These models, as discussed above, assume a more gradual transition between the different regimes, and are obtained by replacing the indicator function in the SETAR models (equation 1b) by a continuous function, typically logistic or exponential, which changes smoothly from 0 to 1. It seems interesting to examine whether the kind of non-linearity captured by the STAR models provides some forecasting gains over the linear models which cannot be observed with SETAR models. We have estimated a Logistic STAR (LSTAR) model for the French franc and the German mark, while an Exponential STAR (ESTAR) turned out to be appropriate in the case of the Japanese yen.

From Table IV(b) it is interesting to note that in terms of both MSFE and Sign, generally the models exhibit similar values. In particular, for the French franc, we find that the apparent advantages of the SETAR-2 model for the one-step-ahead and of the SETAR-3 model for the five-step-ahead forecasts (Table IV(b)) are not significant in terms of the DM test (Table V(b)). Moreover, for the intermediate steps ahead (two to four) the DM test detects a significantly different performance that favours the linear model against the SETAR and LSTAR models. With regard to the German mark, the linear model turns out to be superior to both specifications of the SETAR models (with two and three regimes) and to the LSTAR model in the majority of cases, with the exception of the five-step-ahead forecasts, when the non-linear models are clearly superior in terms of the DM test. For the Japanese yen the performance of the models is significantly different in five cases out of the twenty considered, with three cases in favour of the non-linear models. In particular the SETAR-2 and the ESTAR model dominate the linear model in the five-step-ahead forecasts, with the ESTAR model showing a significant advantage also in the three-step-ahead forecasts. SETAR-3 and ESTAR also show a percentage of correct sign predictions significantly higher than 50%, for the two-step-ahead forecasts.

Overall, the performance of SETAR and STAR models appears similar. In particular it is interesting to note that the STAR models do not provide more gains over the linear models than those observed in the SETAR specifications. The performance of the GARCH models is in no case distinguishable from that of the linear AR models, indicating that the 'in mean' component of the GARCH-M models contributes only marginally to the point forecasts. However, as discussed in the next sub-section, potential forecasting gains of GARCH models can be better explored by evaluating density forecasts.

More generally, although the linearity tests showed clearer evidence of non-linearity in the weekly returns, from Tables V(a) and V(b) it appears that non-linear models applied to high-frequency data do not offer greater forecasting gains with respect to the same models estimated on monthly data. Various explanations for the inadequate forecasting performance of the non-linear models have been offered. As already discussed, Diebold and Nason (1991) argue that non-linearity may not be pronounced enough over the whole forecast period to guarantee greater forecasting accuracy. Others have shown that forecasting gains of non-linear models may be masked by the evaluation method adopted or may depend on the *state of nature* (Clements and Smith, 2001). To investigate this possibility, in what follows we conduct a forecast exercise by conditioning the forecast observations on the regimes of the SETAR models and examine whether, as suggested by other authors, the SETAR forecasts are superior to the linear forecasts conditional on a specific regime. Tiao and Tsay (1994), for example, have shown that SETAR models produce US GNP forecasts which are superior to those obtained from a linear model, when the forecasts are obtained from the regime with fewer observations (when the economy is recovering from recession). Similarly, Clements and Smith (2001) have shown, by means of Monte Carlo simulations, that a three-regime SETAR model for the yen exchange rate returns records significant gains (over 40%) relative to the random walk

model for the one-step-ahead forecasts conditional on being in the middle regime. As in the study by Tiao and Tsay, the regime for which it was possible to exploit gains was the minority-observations regime (with 15% of the total observations).

Point forecasts conditional on the regimes of the SETAR models

In this exercise, we have divided up the forecast sample by the regimes of the SETAR models to explore the dependence of forecast performance of the non-linear models on the regime at the forecast origin. In Tables VI(a) and VI(b) we report the results for the one-step-ahead point forecasts for the weekly models of the three exchange rate returns. The models under comparison are, as in the previous assessment, the non-linear SETAR and STAR models, the GARCH, and AR models. The tables report the MSFEs for each model, the MSFEs normalized by those for the linear AR model and the *P*-values of the Diebold and Mariano (1995) test of equal forecast accuracy. Focusing on the performance of the SETAR and STAR models relative to that of the AR counterparts, the results reveal significant forecast gains that can be achieved in specific regimes with the SETAR and STAR models, gains that were not noticeable when the models were assessed unconditionally over the entire forecast period. Interestingly, and in line with previous findings, our results across exchange rates show that gains of the non-linear models over the linear AR alternatives occur, in most cases, for the minority-regime observations. More specifically, for the French franc, the SETAR-2 model (Table VI(a)) significantly outperforms the AR model conditional on being in regime 1 (17% observations), while the two models show a similar performance for observations in regime 2. Similar gains are achieved by the LSTAR model and SETAR-3 model (Table VI(b)). The latter, however, is significantly outperformed by the linear model in regimes 2 and 3. For the German mark, the SETAR-2 and LSTAR models significantly outperform the linear AR model

Table VI(a). Point forecasts conditional on the regimes of the SETAR-2 model: one-step-ahead MSFE and normalized MSFE (weekly data)

	MSFE			Normalized MSFE		
	Entire sample	Regime 1	Regime 2	Entire sample	Regime 1	Regime 2
French franc						
no. obs.	313	53	260	313	53	260
Linear AR(2)	2.146	2.867	1.999	—	—	—
GARCH-M	2.146	2.888	1.998	1.000	1.007	0.999
SETAR-2	2.112	2.446*	2.044	0.984	0.853*	1.023
LSTAR	2.180	2.563*	2.000	1.016	0.894*	1.001
German mark						
no. obs.	313	71	242	313	71	242
Linear AR(2)	2.328	2.794	2.229	—	—	—
GARCH-M	2.355	2.725	2.246	1.012	0.975	1.008
SETAR-2	2.378**	2.690**	2.287	1.022**	0.963**	1.026
LSTAR	2.366	2.515**	2.221	1.017	0.900**	0.996
Japanese yen						
no. obs.	313	235	78	313	235	78
Linear AR(2)	1.893	1.921	1.791	—	—	—
GARCH-M	1.898	1.939	1.784	1.003	1.009	0.996
SETAR-2	1.864	1.924	1.685*	0.985	1.002	0.941*
ESTAR	1.921	2.002**	1.711*	1.015	1.042**	0.955*

Table VI(b). Point forecasts conditional on the regimes of the SETAR-3 model: one-step-ahead MSFE and normalized MSFE (weekly data)

	MSFE				Normalized MSFE			
	Entire sample	Regime 1	Regime 2	Regime 3	Entire sample	Regime 1	Regime 2	Regime 3
French franc								
no. obs.	313	74	184	55	313	74	184	55
Linear AR(2)	2.146	2.460	2.008	2.186	—	—	—	—
GARCH-M	2.146	2.480	2.003	2.193	1.000	1.008	0.997	1.004
SETAR-3	2.205	2.166*	2.067**	2.662**	1.028	0.881*	1.029**	1.218**
LSTAR	2.180	2.235*	2.070*	2.478**	1.016	0.908*	1.031*	1.134**
German mark								
no. obs.	313	71	187	55	313	71	187	55
Linear AR(2)	2.328	2.794	2.320	1.917	—	—	—	—
GARCH-M	2.355	2.725**	2.321	1.992	1.012	0.975**	1.000	1.039
SETAR-3	2.420**	2.689**	2.446**	1.982	1.010**	0.963**	1.054**	1.034
LSTAR	2.366	2.515**	2.420**	1.992*	1.017	0.900**	1.043**	1.039*
Japanese yen								
no. obs.	313	116	103	94	313	116	103	94
Linear AR(2)	1.893	1.994	1.816	1.838	—	—	—	—
GARCH-M	1.898	2.059	1.821	1.792	1.003	1.032	1.003	0.975
SETAR-3	1.932	2.093*	1.911*	1.759*	1.021	1.050*	1.052*	0.957*
ESTAR	1.921	2.038	1.914*	1.784	1.015	1.022	1.054*	0.971

Notes: The normalized MSFE is calculated as the ratio $MSFE_{NL}/MSFE_{L}$.
*, ** denotes significance of the Diebold–Mariano test at 10% and 5%.

conditional on being in regime 1 (23% observation), while their performance is indistinguishable from that of the linear model in regime 2. The SETAR-3 model also shows significant gains over the linear model in regime 1 (coinciding with regime 1 in SETAR-2), it is outperformed in regime 2, and the performance of the models is not distinguishable in regime 3. Finally, for the Japanese yen, there is some evidence of gains for the SETAR-2 and ESTAR models conditional on being in regime 2 (25% observations) over the AR model. There is no evidence of the SETAR-2 model outperforming the linear counterpart in regime 1, which is the regime with the largest number of observations, while the ESTAR model is, in this regime, significantly outperformed by the linear AR model. Some gains are also obtained for the SETAR-3 model over the AR model when conditioning on regime 3, which is again the regime with the fewest observations. Conversely, the AR model does significantly better than the SETAR-3 model conditional on being in regimes 1 and 2.

To conclude, the evaluation of the models has shown significant gains of the SETAR and STAR models versus the linear alternative when the forecast origin is conditioned on the regimes. The forecast performance of the non-linear models is in most cases superior to that of the linear counterparts conditional on the regime with fewer observations. On the other hand, when conditioning on the regime(s) with more observations, model performance is either not distinguishable or the SETAR and STAR models are outperformed by the linear models. These results, based on actual data and on a genuine out-of-sample forecast exercise, confirm previous findings by Tiao and Tsay (1994) for the US GNP, and those obtained by Clements and Smith (2001), by means of a Monte Carlo study for the exchange rate returns.

Consideration of the performance of the GARCH models versus the AR models shows, again, that there is not much to choose between these two models when they are evaluated on their ability to produce point forecasts. This confirms the marginal contribution that the 'mean component' of the GARCH models plays in forecasting the conditional mean.

Density forecasts

Previous authors have found that non-linear models of the TAR-type perform well in terms of predicting the overall density, rather than just the first moment (Clements and Smith, 2000). Moreover, GARCH models are useful for providing some indication of the uncertainty around the mean by modelling the conditional variance of the process. Thus, an evaluation based on density forecasts may reveal greater discrimination over the linear models than evaluations based on the first moment. In this section, we apply the density forecasts evaluation methods suggested by Diebold *et al.* (1998) and surveyed by Tay and Wallis (2000). The approach is based on the analysis of the probability integral transforms of the actual realizations of the variables with respect to the forecast densities of the models (see also Dawid, 1984). These are defined as $z_t = F_t(y_t)$, where $F(\cdot)$ is the density forecast distribution function and y is the observed outcome. Thus, z_t is the cumulative density function corresponding to the density $F_t(y_t)$ evaluated at y_t, that is, the forecast probability of observing an outcome no greater than that actually realized. If a sequence of density forecasts correspond to the true density, then Diebold *et al.* (1998) show that the corresponding sequence of probability integral transforms $\{z_t\}_{t=1}^{313}$ is i.i.d. uniform (0,1). The forecast densities are then assessed by testing whether the sequence of probability integral transforms departs from the i.i.d. uniform hypothesis. The evaluation of the density forecasts proceed, as suggested by Diebold *et al.* (1998), by examining the distributional and autocorrelation properties of the z_t series. The distributional properties can be examined by visual inspection of plots of either the histogram or the empirical distribution function of the z_t series, which are visually compared with those of a uniform (0,1). These graphical devices are then supplemented by more formal tests. Confidence intervals are computed for the histograms under the null hypothesis of i.i.d. U(0,1), exploiting the binomial structure bin-by-bin (Diebold *et al.*, 1998). Equivalently, the Kolmogorov–Smirnov test (the maximum absolute difference between the empirical distribution function and the distribution function under the null hypothesis of uniformity) is used on the sample distribution function of the z_t series (see Diebold *et al.*, 1999; Tay and Wallis, 2000). These two approaches address the unconditional uniformity hypothesis. Alternatively, unconditional uniformity can be tested by applying the Pearson's chi-squared goodness-of-fit test (see the recent discussion in Wallis, 2001, with applications to inflation forecasts). In our analysis below, we use both the Kolmogorov–Smirnov test and the Pearson χ^2 test. With respect to the i.i.d. *part of the hypothesis* for the z_t series, Diebold *et al.* (1998) again suggest visual assessment of graphical tools, such as correlograms, combined with the usual confidence intervals. Along similar lines, other authors have used in their applications the LM test for higher order serial correlation (Clements and Smith, 2000). Dependence is assessed not only linearly, but also with regard to higher-order moments (conditional variance, skewness and kurtosis).

In what follows, we assess the one-step-ahead density forecasts of the three exchange rate returns, obtained under the assumption of Gaussian errors, for the AR, GARCH and S(E)TAR models. In Figure 2 we report a selection of plots of the empirical distribution function of the z_t series against the 45° line—the theoretical uniform distribution function. The 95% confidence intervals along side the 45° line are calculated using the critical values of the Kolmogorov–Smirnov test, reported in Lilliefors (1967, Table 1, p. 400), in the presence of estimated parameters. Although not much

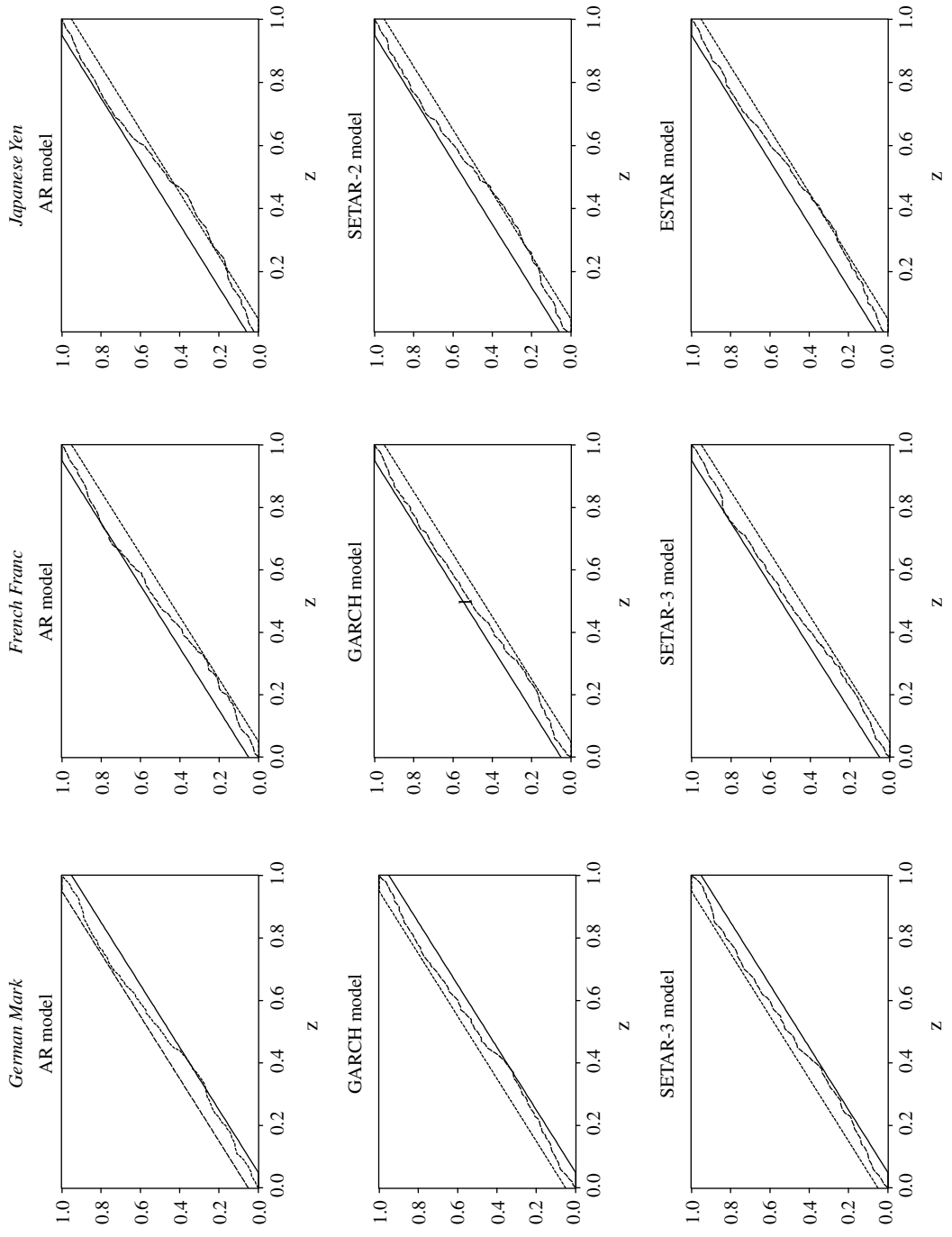

Figure 2. Density Forecasts

is known about parameter estimation errors in the density evaluation literature, the standard critical values of the Kolmogorov–Smirnov test are probably a conservative estimate of the 'correct' critical values when certain parameters of the distribution must be estimated from the sample.[10] Table VII reports the results from the Pearson χ^2 test, and Table VIII the *P*-values of the Ljung–Box *Q*-statistics for serial correlation, computed on the first six sample autocorrelations for $(z-\overline{z})$, $(z-\overline{z})^2$, $(z-\overline{z})^3$ and $(z-\overline{z})^4$.

By examining Figure 2 we see that some models exhibit greater evidence of departure from the uniform null than do others. In the case of the Japanese yen, we found that all models approached the bounds very closely, but the AR showed more evidence of departure from the null hypothesis by crossing the bounds at various ranges of the density. Also, there was clear evidence that the GARCH, SETAR and STAR models did comparatively better than the AR model in the case of the French franc, exhibiting empirical distribution functions which were overall closer to the 45° line. Finally, the evidence for the German mark was somewhat mixed, giving stronger indication of departure for the AR and STAR model.

Next, unconditional uniformity is tested by applying Pearson's chi-squared goodness-of-fit test. The use of complementary techniques may offer further guidance on the nature and strength of the deficiencies of the density forecasts, and be particularly useful in real data applications. We have

Table VII. χ^2 goodness-of-fit test (weekly data)

	$k = 20$	
	χ^2	*P*-value*
French franc		
Linear AR(2)	21.505	0.309
GARCH-M	16.137	0.648
SETAR-2	16.521	0.622
SETAR-3	19.971	0.396
LSTAR	25.083	0.158
German mark		
Linear AR(2)	26.617	0.114
GARCH-M	24.188	0.189
SETAR-2	25.978	0.131
SETAR-3	31.601	**0.035**
LSTAR	37.863	**0.006**
Japanese yen		
Linear AR(2)	39.012	**0.004**
GARCH-M	35.818	**0.011**
SETAR-2	27.128	0.102
SETAR-3	31.728	**0.033**
ESTAR	21.505	0.309

* Asymptotic *P*-values.

[10] We are grateful to an anonymous referee for bringing this point to our attention. The formula reported in Lilliefors (1967) for $T > 30$, level of significance 0.05, is given by $0.886/\sqrt{T}$. As we have $T = 313$ observations, the 95% confidence intervals are then the 45° line ±0.05008. In a previous version of this paper we used standard asymptotic critical values, as in Miller (1956, Eq. 3, page 115), and obtained more conservative results.

Table VIII. *P*-values of the Ljung–box *Q*-statistics for serial correlation (weekly data)

		Moments			
		$(z-\overline{z})$	$(z-\overline{z})^2$	$(z-\overline{z})^3$	$(z-\overline{z})^4$
French franc	Linear AR(2)	0.264	**0.001**	0.858	**0.000**
	GARCH-M	0.465	0.586	0.962	0.665
	SETAR-2	0.084	**0.028**	0.228	**0.001**
	SETAR-3	0.099	**0.003**	0.742	**0.010**
	LSTAR	**0.030**	**0.000**	0.228	**0.000**
German mark	Linear AR(2)	0.078	**0.002**	0.495	**0.004**
	GARCH-M	0.147	0.837	0.821	0.863
	SETAR-2	**0.051**	**0.018**	0.302	**0.021**
	SETAR-3	**0.053**	**0.023**	0.454	0.084
	LSTAR	**0.022**	**0.001**	0.412	**0.000**
Japanese yen	Linear AR(2)	0.178	0.207	**0.054**	**0.033**
	GARCH-M	0.019	0.862	0.116	0.736
	SETAR-2	0.258	0.322	0.178	0.131
	SETAR-3	0.066	0.839	0.149	0.840
	ESTAR	0.419	0.304	0.589	0.070

divided the range of the z_t series into k equiprobable classes and computed the test

$$X^2 = \sum (n_i - n/k)^2/(n/k) = (k/n)\sum n_i^2 - n$$

where k is the number of classes, n_i the observed frequencies, n the number of observations (313 forecasts). This test has a limiting χ^2 distribution with $k-1$ degrees of freedom under the null hypothesis. In Table VII we report the results of the tests computed for $k = 20$.[11] The table shows that for the French franc, all models produce density forecasts with correct coverage which, a part from the AR case, is consistent with the assessment of unconditional uniformity in Figure 2. For the German mark, SETAR-3 and LSTAR density forecasts seem to departure significantly from the correct density, while the AR forecasts appear correct, somewhat in contrast to the indications from the Kolmogorov–Smirnov test. Finally, for the Japanese yen, only the SETAR-2 and ESTAR models give correct density forecasts, whereas there is stronger evidence against uniformity for the AR and GARCH models.

One well-known limitation of these tests is that they rest on an assumption of random sampling and little is known about the impact on the distribution of these tests of departures from independence.[12] This means that when rejection occurs, the tests provide no guidance as to *why* (Diebold *et al*., 1998, p. 869); departure from uniformity may suggest improper distributional assumptions, or poorly captured dynamics, or both these aspects. As we can see from the results of the Ljung–Box *Q*-statistics reported in Table VIII, there are cases of violation of the independence assumption mostly for the second (conditional heteroscedasticity) and fourth (conditional kurtosis) moments. More specifically, in the case of the French franc and German mark, the GARCH model is the

[11] Since the results of the test may depend on the selected value for k, as discussed in Stuart *et al*. (1999), we performed the test for values of k up to 40 but the results did not lead us to substantial different conclusions from those reported above.
[12] For a preliminary study of the size and power of alternative tests see Noceti *et al*. (2000).

only model which shows no evidence of any kind of autocorrelation, while there is some evidence of misspecification in the second and fourth moment for the other models. Thus, combined with the evidence in Figure 2 and Table VII, we can conclude that for these two exchange rates (French franc and German mark) there is strong support for the GARCH model in favour of the hypothesis of correct density forecasts. Similarly, a clear conclusion can be reached in the case of the Japanese yen, but this time in favour of the TAR-type models, which seem to be the only models to produce z_t series consistent with the i.i.d. hypothesis. In the case of the Japanese yen, there is in fact evidence of violation of the independence assumption for the AR model (third- and fourth-ordered moments) and for the GARCH-M model (first-order moment).[13]

Overall, the results presented in this section allow for clearer discrimination among the competing models, providing more evidence supporting the forecasting superiority of both classes of non-linear models than in evaluations based on just the first moment.

CONCLUSIONS

In this study we have compared the forecasting performance of alternative univariate time series models for the returns of three exchange rates quoted against the US dollar: the French franc, the German mark and the Japanese yen. Application of linearity tests to monthly and weekly series has provided evidence of non-linear components at both frequencies although, as expected, the evidence of non-linearities was more marked in the weekly series. Various non-linear models, namely a two-regime SETAR, a three-regime SETAR, and a GARCH-M model, were compared and contrasted with simpler linear alternatives (AR and random walk processes). STAR models were also used with weekly returns.

The SETAR and GARCH models proved successful in describing non-linear features of the data. In particular, the SETAR models have provided strong in-sample evidence for the existence of different regimes, in which the exchange rate returns exhibit quite different dynamics, whereas the GARCH models appeared to capture adequately non-linearities in the second-order conditional moment.

In comparing the forecast performance of different models, we have used different criteria and, within the same criterion, we have adopted alternative procedures. The steps taken in this paper can be used as examples to illustrate the kind of problems and choices faced in applications with actual data. The forecast comparison has been conducted initially on both monthly and weekly data using the MSFE and the percentage of the correct sign predictions, for one-step- and multistep-ahead forecasts. The percentage of correct sign predictions was tested against the null hypothesis of independence between forecasts and actual values, and differences in MSFEs between models were evaluated by means of the Diebold and Mariano test. This analysis was conducted over the entire forecast period and did not show significant forecast gains for the non-linear models over the linear benchmark. In a second exercise, we have assessed the forecast performance of the weekly models by *conditioning* the forecast origin on each of the regimes of the SETAR models. This exercise was carried out for the one-step-ahead forecasts and enabled us to explore further the added value of the non-linear features of the SETAR and STAR models. Finally, we have evaluated the one-step-ahead density forecasts associated with each model by examining the distributional and autocorrelation

[13] As an interesting extension, one could explore whether joint estimation of threshold and GARCH models offer further opportunities of forecasting gains.

properties of the probability integral transforms of the actual exchange rates returns with respect to the forecast densities of the models (Diebold *et al*., 1998). The comparative evaluation of density forecasts has provided stronger evidence supporting the forecasting superiority of both classes of non-linear models, GARCH and TAR, and allowed for clearer discrimination with respect to the linear competing models.

Overall, our results, based on actual data for exchange rate returns and on a *genuine* out-of-sample forecasting exercise, confirm and reinforce recent findings that have shown that the comparative advantages of non-linear models over the linear counterparts depend on both the criteria used to assess forecast accuracy (MSFE, Sign, Density Forecasts) and on the 'state of nature' (Clements and Smith, 2001). These results question the oft-claimed forecasting superiority of the linear models and call for a more articulate analysis of forecast performance and of the stochastic characteristics of the period considered.

Finally, on methodological issues, the procedures adopted in the evaluation of density forecasts are relatively new, and currently there is little experience of their use. Moreover, comparison of alternative approaches to testing forecast densities is at an early stage, and more research is required in this area. In the meantime, as more experience and research develop, we have shown that application of complementary techniques may provide useful indications on the nature and strength of the deficiencies of the forecasts and, more generally, can offer further guidance to 'rank' the models in a comparative exercise.

ACKNOWLEDGEMENTS

We wish to thank the editor and two anonymous referees for helpful comments. We are also indebted to Mike Clements, Jeremy Smith and Ken Wallis for fruitful discussions.

REFERENCES

Boero G, Marrocu E. 2000. Modelli nonlineari per i tassi di cambio: un confronto previsivo con dati a diversa frequenza. *Moneta e Credito* **53**: 385–415.

Bollerslev T. 1986. Generalized autoregressive conditional heteroscedasticity. *Journal of Econometrics* **31**: 307–327.

Box GEP, Jenkins GM. 1976. *Time Series Analysis: Forecasting and Control*, 2nd edn. Holden-day: San Francisco, CA.

Brooks C. 1996. Testing for nonlinearity in daily sterling exchange rates. *Applied Financial Economics* **6**: 307–311.

Brooks C. 1997. Linear and nonlinear (non-) predictability of high-frequency exchange rates. *Journal of Forecasting* **16**: 125–145.

Chan KS, Tong H. 1986. On estimating thresholds in autoregressive models. *Journal of Time Series Analysis* **7**: 179–190.

Chappell D, Padmore J, Mistry P, Ellis C. 1996. A threshold model for the French franc/Deutschemark exchange rate. *Journal of Forecasting* **15**: 155–164.

Christoffersen PF, Diebold FX. 2001. Financial asset returns, market timing, and volatility dynamics. Working Paper, McGill School of Business.

Clements MP, Hendry DF. 1993. On limitations of comparing mean squared forecast errors. *Journal of Forecasting* **12**: 617–637.

Clements MP, Hendry DF. 1995. Forecasting in cointegrated systems. *Journal of Applied Econometrics* **10**: 127–146.

Clements MP, Smith JP. 1997. The performance of alternative forecasting methods for SETAR models. *International Journal of Forecasting* **13**: 463–75.
Clements MP, Smith JP. 1999. A Monte Carlo study of the forecasting performance of empirical SETAR models. *Journal of Applied Econometrics* **14**: 123–41.
Clements MP, Smith JP. 2000. Evaluating the forecast densities of linear and non-linear models: applications to output growth and unemployment. *Journal of Forecasting* **19**: 255–276.
Clements MP, Smith JP. 2001. Evaluating forecasts from SETAR models of exchange rates. *Journal of International Money and Finance* **20**: 133–148.
Dacco R, Satchell S. 1999. Why do regime-switching models forecast so badly? *Journal of Forecasting* **18**: 1–16.
Dawid AP. 1984. Statistical theory: the prequential approach. *Journal of the Royal Statistical Society A* **147**: 278–292.
Diebold FX, Gunther TA, Tay AS. 1998. Evaluating density forecasts with applications to financial risk management. *International Economic Review* **39**: 863–883.
Diebold FX, Mariano RS. 1995. Comparing predictive accuracy. *Journal of Business and Economic Statistics* **13**: 253–263.
Diebold FX, Nason JA. 1990. Nonparametric exchange rate prediction? *Journal of International Economics* **28**: 315–332.
Diebold FX, Tay AS, Wallis KF. 1999. Evaluating density forecasts of inflation: the Survey of Professional Forecasters. In *Cointegration, Causality, and Forecasting: A Festschrift in Honour of Clive W.J. Granger*, Engle RF, White H (eds). Oxford University Press: Oxford.
Dixon H. 1999. Controversy: exchange rates and fundamentals. *The Economic Journal* **109**: F652–F654.
Engle RF. 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of the United Kingdom inflation. *Econometrica* **50**: 987–1008.
Engle RF, Lilien DM, Robins RP. 1987. Estimating time varying risk premia in the term structure: the ARCH-M model. *Econometrica* **55**: 391–407.
Flood RP, Rose AK. 1999. Understanding exchange rate volatility without the contrivance of macroeconomics. *The Economic Journal* **109**: F660–F672.
Glosten LR, Jagannathan R, Runkle D. 1993. On the relation between the expected value and the volatility of the nominal excess return on stocks. *Journal of Finance* **48**: 1779–1802.
Granger CWJ. 1993. Strategies for modelling nonlinear time-series relationships. *The Economic Record* **69**: 233–238.
Granger CWJ, Anderson AP. 1978. *An Introduction to Bilinear Time Series Models*. Vandenhoeck & Ruprecht: Göttingen.
Granger CWJ, Lee T-H. 1999. The effects of aggregation on non-linearity. *Econometric Reviews* **18**: 259–269.
Granger CWJ, Teräsvirta T. 1993. *Modelling Nonlinear Economic Relationships*. Oxford University Press: Oxford.
Hsieh DA. 1989. A nonlinear stochastic rational expectations model of exchange rates. Unpublished manuscript, Fuque School of Business, Duke University.
Keenan DM. 1985. A Tukey nonadditivity-type test for time series non-linearity. *Biometrika* **72**: 39–44.
Kräger H, Kugler P. 1993. Nonlinearities in foreign exchange markets: a different perspective. *Journal of International Money and Finance* **12**: 195–208.
Krugman P. 1991. Target zones and exchange rate dynamics. *Quarterly Journal of Economics* **106**: 669–682.
Li WK, Mak TK. 1994. On the squared residual autocorrelations in non-linear time series with conditional heteroskedasticity. *Journal of Time Series Analysis* **15**: 627–636.
Lilliefors HW. 1967. On the Kolmogorov–Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association* **62**: 399–402.
Ljung GM, Box GEP. 1978. On a measure of lack of fit in time series models. *Biometrika* **65**: 297–303.
Luukkonen R, Saikkonen P, Teräsvirta T. 1988. Testing linearity against smooth transition autoregressive models. *Biometrika* **75**: 491–499.
MacDonald R. 1999. Exchange rate behaviour: are fundamentals important? *The Economic Journal* **109**: F673–F691.
McLeod AI, Li WK. 1983. Diagnostic checking ARMA time series models using squared-residual autocorrelations. *Journal of Time Series Analysis* **4**: 269–273.

Meese R, Rogoff K. 1983. Empirical exchange rate models of the seventies: do they fit out of sample? *Journal of International Economics* **14**: 3–24.
Meese RA, Rose AK. 1991. An empirical assessment of nonlinearities in models of exchange rate determination. *Review of Economic Studies* **58**: 603–619.
Miller LH. 1956. Table of percentage point of Kolmogorov statistics. *Journal of the American Statistical Association* **51**: 111–121.
Nelson DB. 1991. Conditional heteroskedasticity in asset returns: a new approach. *Econometrica* **59**: 347–370.
Nelson DB, Cao CQ. 1992. Inequality constraint in the univariate GARCH(1,1) model. *Journal of Business and Economic Statistics* **10**: 229–235.
Noceti P, Smith JP, Hodges S. 2000. An evaluation of tests of distributional forecasts. Discussion Paper, FORC, University of Warwick, No. 102.
Peel DA, Speight AE. 1994. Testing for nonlinear dependence in inter-war exchange rates. *Weltwirtschaftliches Archiv* **130**: 391–417.
Pesaran MH, Timmermann A. 1992. A simple nonparametric test of predictive performance. *Journal of Business and Economic Statistics* **10**: 461–465.
Pippinger MK, Goering GE. 1993. A note on the empirical power of unit root tests under threshold processes. *Oxford Bulletin of Economics and Statistics* **55**: 473–481.
Priestley MB. 1980. State-dependent models: a general approach to nonlinear time series analysis. *Journal of Time Series Analysis* **1**: 47–71.
Priestley MB. 1988. *Nonlinear and Non-stationary Time Series Analysis*. Academic Press: London.
Ramsey JB. 1969. Tests for specification errors in classical linear least squares regression analysis. *Journal of the Royal Statistical Society B* **31**: 350–371.
Rogoff K. 1999. Monetary models of dollar/yen/euro nominal exchange rates: dead or undead? *The Economic Journal* **109**: F655–F659.
Stuart A, Ord JK, Arnold S. 1999. *Kendall's Advanced Theory of Statistics*, vol. 2 A. Arnold Publishers: London.
Tay AS, Wallis KF. 2000. Density forecasting: a survey. *Journal of Forecasting* **19**: 235–254.
Teräsvirta T. 1994. Specification, estimation and evaluation of smooth transition autoregressive models. *Journal of the American Statistical Association* **89**: 208–218.
Teräsvirta T, Anderson HM. 1992. Characterizing nonlinearities in business cycles using smooth transition autoregressive models. *Journal of Applied Econometrics* **7**: S119–S139.
Thursby JG, Schimdt P. 1977. Some properties of tests for the specification error in a linear regression model. *Journal of the American Statistical Association* **72**: 635–41.
Tiao GC, Tsay RS. 1994. Some advances in non-linear and adaptive modelling in time series. *Journal of Forecasting* **13**: 109–131.
Tong H. 1978. On a threshold model. In *Pattern Recognition and a Signal Processing*, Chen CH (ed.). Sijhoff and Noordoff: Amsterdam.
Tong H. 1983. *Threshold Models in Nonlinear Time Series Analysis*. Springer-Verlag: New York.
Tong H. 1990. *Nonlinear Time Series. A Dynamical System Approach*. Clarendon Press: Oxford.
Tong H, Lim KS. 1980. Thresholds autoregression, limit cycles and cyclical data. *Journal of the Royal Statistical Society B* **42**: 245–292.
Tong H, Moeanaddin R. 1988. On multi-step nonlinear squares prediction. *The Statistician* **37**: 101–110.
Tsay RS. 1986. Nonlinearity tests for time series. *Biometrika* **73**: 461–466.
Wallis KF. 2001. Chi-squared tests of interval and density forecasts, and the Bank of England's fan charts. *International Journal of Forecasting* (forthcoming).
Weiss AA. 1984. Systematic sampling and temporal aggregation in time series models. *Journal of Econometrics* **26**: 271–281.
Zakoian JM. 1994. Threshold heteroskedastic models. *Journal of Economic Dynamics and Control* **18**: 931–955.

Authors' biographies:

Gianna Boero is associate professor of econometrics at the University of Cagliari (Italy), research fellow of CRENoS, and part-time lecturer at the University of Warwick. Her recent research is in the field of time series econometric modelling with applications to financial variables.

Emanuela Marrocu is a lecturer at the University of Cagliari, Italy, and a research fellow of CRENoS. She holds a PhD in Economics from the University of Warwick. Her research is in the field of non-linear econometric modelling and forecasting.

Authors' addresses:
Gianna Boero, University of Warwick (UK), Department of Economics, Gibbet Hill, Coventry CV4 7AL, UK.

Emanuela Marrocu, University of Cagliari, Viale Fra Ignazio, 78, 09123 Cagliari, Italy.

