Selecting a Nonlinear Time Series Model using Weighted Tests of Equal Forecast Accuracy[*]

Dick van Dijk[†]
Econometric Insitute
Erasmus University Rotterdam

Philip Hans Franses[‡]
Econometric Insitute
Erasmus University Rotterdam

Econometric Institute Report EI 2003-10

March 2003

Abstract

Nonlinear time series models have become fashionable tools to describe and forecast a variety of economic time series. A closer look at reported empirical studies, however, reveals that these models apparently fit well in-sample, but rarely show a substantial improvement in out-of-sample forecasts, at least over linear models. One of the many possible reasons for this finding is that inappropriate model selection criteria and forecast evaluation criteria are used. In this paper we therefore propose a novel criterion, which we believe does more justice to the very nature of nonlinear models. Simulations show that our criterion outperforms currently used criteria, in the sense that the true nonlinear model is more often found to perform better in out-of-sample forecasting than a benchmark linear model. An empirical illustration for US GDP emphasizes its relevance.

Keywords: Nonlinearity, forecasting, forecast evaluation, model selection.
JEL Classification Codes: C22, C52, C53, E32, E37.

[*] This paper has been specifically prepared for presentation at the (EC)^2 conference in Bologna, Italy, December 13-14 2002. Financial support from the Netherlands Organization for Scientific Research (N.W.O.) is gratefully acknowledged.

[†] Econometric Institute, Erasmus University Rotterdam, P.O. Box 1738, NL-3000 DR Rotterdam, The Netherlands, e-mail: djvandijk@few.eur.nl (corresponding author)

[‡] Econometric Institute, Erasmus University Rotterdam, P.O. Box 1738, NL-3000 DR Rotterdam, The Netherlands, e-mail: franses@few.eur.nl

1 Introduction and Motivation

Nonlinear time series models have become fashionable tools to describe and forecast economic time series, including important macro-economic and financial variables such as unemployment, industrial production, and exchange rates. Examples of these models are threshold models, Markov-Switching models and neural networks, see Franses and van Dijk (2000) for a survey. A common feature of these models is that they assume the presence of different regimes, within which the time series under scrutiny can have different means, variances and (auto-)correlation structures. The precise location (or definition) and duration of these regimes are typically unknown, and need to be estimated. Hence, nonlinear time series models can, for example, describe asymmetric business cycle behavior as observed in unemployment, which typically increases rapidly in recessions and decreases only slowly in expansions.

The typical modeling strategy for nonlinear time series models proceeds as follows. One divides the available sample into two parts, with the first part being used for in-sample model specification and estimation and the second part for out-of-sample forecasting and model evaluation. Using the in-sample data, one first specifies a linear model, usually of the autoregressive type. One then considers various diagnostic tests to see if there is any evidence of nonlinearity or, even better, of nonlinearity that can be attributed to a specific alternative model specification. If this is the case, one fits various nonlinear models, and compares these in terms of forecasting performance for the out-of-sample data. This can be done recursively or not, as well as for one-step or multiple steps ahead forecasts. Typically, one also includes a linear model in the out-of-sample forecast competition, perhaps to see if specifying and estimating the nonlinear models was worth the trouble in the first place.

A casual inspection of the available literature on applications of nonlinear models to economic time series reveals that in many cases there is substantial empirical evidence of nonlinearity based on the in-sample data. Interestingly, this evidence rarely seems to translate into out-of-sample forecasting success as, apparently, the linear model frequently wins the forecast competition; see Stock and Watson (1999) for an illustrative example.

Various reasons for this general finding have been put forward and studied in the literature, see Diebold and Nason (1990), de Gooijer and Kumar (1992), Ramsey (1996), and Dacco and Satchell (1999), among others. For example, the nonlinearity detected in the in-sample data may be spurious, in the sense that other features of the time series, such as heteroskedasticity, structural breaks, or outliers, suggest the presence of nonlinearity, see van Dijk, Franses and Lucas (1999), for example. Even though one then might successfully estimate a nonlinear model for such a time series, it is very unlikely that this will result in improved out-of-sample forecasts.

Another possible cause for poor forecasting performance of nonlinear models is that the nonlinearity "does not show up" in the forecast period. In case of regime-switching models it might be that all observations in the forecast period belong to only one of the regimes. Hence, empirical forecasts do not always allow to assess the forecasting quality of the nonlinear models completely.[1]

Finally, it may be that nonlinear models do not improve upon *point* forecasts relative to linear models, but might render superior interval and density forecasts, as suggested in Pesaran and Potter (1997) and Clements and Hendry (1999, p. 285).

All these reasons may be valid, but we believe that an important reason for the apparent poor forecast performance of nonlinear models can be that we use improper criteria to evaluate forecasts from such models. To make this clear, we should go back to the properties of nonlinear models, and how they differ from linear models. Basically, popular nonlinear models such as threshold models allow for a distinct description of observations, depending on whether they get assigned to one or the other regime (in case there are two such regimes). A linear model does not allow for such a distinction. Hence, one would want to compare the two models when matters are really different. For example, a linear time series model will treat recession and expansion observations equally. When one would include all observations in model selection or forecast evaluation with equal weights, it may well be that, on average, the linear model wins, even in situations where the recession observations are much more accurately predicted by the nonlinear model, while other data are predicted relatively poorly.

Therefore, to associate model selection and forecast evaluation criteria with regime-specific observations seems sensible and this has been pursued in Tiao and Tsay (1994), Montgomery *et al.* (1998), and Clements and Smith (1999, 2001), among others. However, the approach pursued in these papers suffers from a major shortcoming, which is that the regimes themselves are defined by the nonlinear model being examined. While we do think that it is sensible to focus on specific observations for model selection and forecast evaluation, to us it seems better not to base the choice of observations of interest on model-specific regimes.

[1] Even in simulation-based evaluations, where the nonlinearity in the forecast period can be "controlled", linear models have been found to be reasonably robust forecasting devices, see Clements and Krolzig (1998) and Clements and Smith (1999).

Of course, there are infinitely many ways to achieve this goal. For example, one could take, say, the NBER turning point dates to split the observations into expansion and recession regimes, but that does not seem to match well with the actual use of models. Indeed, one would perhaps not be interested in models that would have predicted a recession a few months ago, while we just now know that there has been a recession then. Additionally, it could very well be that institutions like the NBER use the same kinds of models when it decides upon the chronology. This would make the model selection criterion fully endogenous.

In many practical forecasting situations, extreme observations of the variable of interest are the most relevant ones, and also the most important ones to forecast accurately. For example, large negative output growth often signals the start of recession, while a large negative stock return often initiates a period of increased volatility and is also of crucial importance for risk management and Value-at-Risk measures. Note that a linear model will most likely miss out on these (unexpectedly) large positive or negative observations, while a nonlinear model could have a specific component to capture such data points. Hence, we propose to choose between nonlinear and linear models by zooming in on the tails of the unconditional distribution of the variable of interest. This is because we believe that one would gain confidence in a model which would have been able to forecast large absolute observations, independent on whether these belong to, say, a recession or not.

For this purpose, we focus on the test of equal predictive accuracy of Diebold and Mariano (1995), which is widely used nowadays for comparing point forecasts of two competing models, and which is commonly regarded as one of the better diagnostic measures. Indeed, Clements *et al.* (2003) find that the Diebold-Mariano statistic is more powerful in discriminating between linear and nonlinear models than techniques based on interval or density forecasts, as developed by Christoffersen (1998), Diebold, Gunther and Tay (1998) and Berkowitz (2001), among others. As we intend to emphasize only particular observations in the forecast evaluation, we modify the standard Diebold-Mariano statistic to allow for different weights being attached to different forecasts.

The remainder of our paper proceeds as follows. In Section 2, we discuss the test statistic of Diebold and Mariano (1995; DM) and introduce a weighted version of this test statistic, which suits our purpose described above. In Section 3, we use extensive Monte Carlo simulations to examine the small sample size and power properties of our new test. We find that it performs exceptionally well, in the sense that if the data are nonlinear, the test finds that the nonlinear model produces the best forecasts. The power of the weighted DM test often is substantially higher than that of the standard test statistic. In Section 4, we illustrate its empirical merits, by examining the forecasting performance of the floor-and-ceiling model for US GDP developed by Pesaran and Potter (1997). In Section 5, we conclude this paper with some remarks and an outline of further research topics.

2 Weighted tests of equal forecast accuracy

Diebold and Mariano (1995) developed a test statistic of the null of equal accuracy of two competing h-step ahead forecasts of a time series variable y_t, denoted as $\widehat{y}_{1,t|t-h}$ and $\widehat{y}_{2,t|t-h}$, respectively, which have been produced for $t = R+h, \ldots, R+P+h-1$, rendering P forecasts in total, where R is the number of observations in the estimation sample (see below for discussion). Specifically, they propose a test of the null of equal forecast accuracy for some arbitrary loss function, $g(e_{i,t|t-h})$, where $e_{i,t|t-h}$ is the corresponding h-step ahead forecast error, that is, $e_{i,t|t-h} = y_t - \widehat{y}_{i,t|t-h}$, $i = 1, 2$. The loss differential is defined as $d_t \equiv g(e_{1,t|t-h}) - g(e_{2,t|t-h})$, so that equal forecast accuracy entails $\mathsf{E}[d_t] = 0$.

Assuming covariance stationarity of the loss differential series, Diebold and Mariano (1995) show that the asymptotic distribution of the sample mean loss differential

$$\overline{d} \equiv \frac{1}{P} \sum_{t=R+h}^{R+P+h-1} d_t, \tag{1}$$

is given by

$$\sqrt{n}\left(\overline{d} - \mu\right) \xrightarrow{D} N\left(0, V\left(\overline{d}\right)\right),$$

where

$$V\left(\overline{d}\right) = \frac{1}{P}\left(\gamma_0 + 2\sum_{i=0}^{h-1} \gamma_i\right),$$

assuming that h-step ahead forecasts exhibit dependence up to order $h-1$. Thus, an asymptotically standard normal statistic for testing the null hypothesis of equal forecast accuracy can be obtained as

$$\mathrm{DM} = \frac{\overline{d}}{\sqrt{\widehat{V}\left(\overline{d}\right)}}, \tag{2}$$

where $\widehat{V}\left(\overline{d}\right)$ is a consistent estimate of $V\left(\overline{d}\right)$, based on the sample autocovariances $\widehat{\gamma}_i$ given by

$$\widehat{\gamma}_i = \frac{1}{P} \sum_{t=R+h+i}^{R+P+h-1} \left(d_t - \overline{d}\right)\left(d_{t-i} - \overline{d}\right).$$

As argued in the introduction, in many practical situations some observations are more important to forecast accurately than others. For example, in empirical macroeconomics, forecasting the start of a recession is of vital importance. Note that this start of a recession would most likely correspond with, for example, large negative observations for output growth or large positive observations for the change in unemployment. Similarly, in financial applications one usually is particularly interested in accurately forecasting extreme negative returns on investments, as evidenced by the enormous interest in Value-at-Risk measures. Hence, when selecting among competing forecasting models, it makes sense to focus on these crucial observations or, put differently, to put more weight on those observations relative to less important ones.

For this purpose, we propose to modify the Diebold-Mariano statistic as given in (2) by using a weighted average loss differential, given by

$$\overline{d}_w = \frac{1}{P} \sum_{t=R+h}^{R+P+h-1} w(\omega_t) d_t, \tag{3}$$

where $\omega_t = \{y_{t-j}, x_{t-j}, j = 0, 1, \ldots\}$ is the information set available at time t, comprising the history of the variable of interest y_t and (possibly) k exogenous variables $x_t = (x_{1t}, \ldots, x_{kt})'$. For most practical situations, this gives the researcher enough flexibility to choose the weight function $w(\cdot)$ in such a way that more weight is placed on the relevant observations. A few examples of sensible weight functions are:

1. $w_{\mathrm{T}}(\omega_t) = 1 - \phi(y_t)/\max(\phi(y_t))$, where $\phi(\cdot)$ is the density function of y_t, which allows to focus on both tails of the distribution of y_t;

2. $w_{\mathrm{LT}}(\omega_t) = 1 - \Phi(y_t)$, where $\Phi(\cdot)$ is the cumulative distribution function of y_t, to focus on the left tail of the distribution of y_t;

3. $w_{\mathrm{RT}}(\omega_t) = \Phi(y_t)$, to focus on the right tail of the distribution of y_t.

See also Giacomini (2002a) for the use of weighted loss differentials in the context of comparison of density forecasts.

The only requirements on the weight function $w(\omega_t)$ are that it is a mapping to the [0,1]-interval, and that it is twice continuously differentiable. Under this assumption, and the usual assumptions regarding forecast errors, it is straightforward to show that the resulting weighted Diebold-Mariano statistic

$$\text{W-DM} = \frac{\overline{d}_w}{\sqrt{\widehat{V}\left(\overline{d}_w\right)}}, \tag{4}$$

has an asymptotic standard normal distribution, where $\widehat{V}\left(\overline{d}_w\right)$ can be estimated in the usual way, using the autocovariances of $w(\omega_t)d_t$.

Recently, a number of modifications and extensions to the original DM statistic in (2) have been considered. These can also be applied to our weighted test statistic (4). First, Harvey, Leybourne and Newbold (1997) proposed a modified version that corrects for the tendency of the DM statistic to be over-sized in small samples, due to bias in the estimated variance $\widehat{V}\left(\overline{d}\right)$. The modified weighted statistic is computed as

$$\text{MW-DM} = \sqrt{\frac{P+1-2h+h(h-1)/P}{P}}\ \text{W-DM}. \qquad (5)$$

Furthermore, Harvey *et al.* (1997) propose to compare the adjusted statistic with critical values obtained from a Student's t-distribution with $P-1$ degrees of freedom, instead of the standard normal distribution.

Second, West (1996,2001) and West and McCracken (1998) have considered modifications of forecast evaluation tests, including tests of equal forecast accuracy, to incorporate parameter uncertainty arising from the fact that typically the competing forecasts $\widehat{y}_{i,t|t-h}$, $i = 1, 2$ are obtained from statistical models, which contain unknown parameters that need to be estimated before the forecasts can be constructed. Essentially, this involves an adjustment of $\widehat{V}\left(\overline{d}\right)$. Note that in the particular case of the DM statistic with quadratic loss function $g(e_{i,t|t-h}) = e^2_{i,t|t-h}$ (on which we focus in the following), it can be shown that parameter uncertainty is irrelevant (at least asymptotically), so that for our present purpose such an adjustment is not necessary, see McCracken and West (2002).

Third, McCracken (2000) and Clark and McCracken (2001) have shown that the asymptotic distribution of the DM-statistic is not standard normal if the models from which the competing forecasts $\widehat{y}_{1,t|t-h}$ and $\widehat{y}_{2,t|t-h}$ are obtained are nested. Intuitively, with nested models, the forecasts are asymptotically the same under the null hypothesis, which leads to a non-standard asymptotic distribution. These asymptotic results have been derived under the assumption that the size of both the estimation sample R and the forecast sample P tend to infinity. More recently, Giacomini (2002b) showed that the asymptotic standard normal distribution of the DM statistic remains valid for nested models when the estimation sample size R remains finite. Essentially this means that parameters in the models are estimated using a rolling window of data, instead of an expanding window. We use this estimation scheme in the Monte Carlo simulations and the empirical application discussed below.

3 Small sample properties

In this section we examine the small sample properties of the proposed weighted Diebold-Mariano statistic by conducting extensive Monte Carlo experiments. We are mainly interested in seeing whether truly nonlinear data would be better picked up by nonlinear models than by linear models in forecast evaluation experiments.

3.1 Monte Carlo design

First, we adopt the experimental design used by Diebold and Mariano (1995), Harvey *et al.* (1997) and Clark (1999), among others. We consider realizations of the bivariate normal forecast error process $\{(e_{1,t}, e_{2,t})\}_{t=R+1}^{R+P}$, with variances k and 1, respectively, with contemporaneous correlation ρ and MA(1) serial correlation with parameter θ. See Diebold and Mariano (1995) for details on the construction of these forecast errors. We set $k = 1$ to investigate the size properties of the tests and $k = 0.9, 0.8, \ldots, 0.5$ to examine the power. We consider forecast sample sizes of $P = 16$, 32, 64, 128 and 256 observations, contemporaneous correlation parameters of $\rho = 0$, 0.5 and 0.9, and MA parameters θ of 0, 0.5 and 0.9.

Second, we generate time series y_t according to an AR(1) model

$$y_t = \phi y_{t-1} + \varepsilon_t, \tag{6}$$

where $\varepsilon_t \sim N(0, 1)$, and ϕ is varied among 0.1,0.3,. . . ,0.9. One-step ahead forecasts $\widehat{y}_{t|t-1}$ for $t = R + 1, \ldots, R + P$ are obtained from an AR(1) model (including an intercept) and a two-regime threshold autoregressive (TAR) model, with an AR(1) model in both regimes and using y_{t-1} as threshold variable, that is

$$y_t = (\phi_{1,0} + \phi_{1,1}y_{t-1})\mathtt{I}[y_{t-1} \le c] + (\phi_{2,0} + \phi_{2,1}y_{t-1})(1 - \mathtt{I}[y_{t-1} \le c]) + \eta_t, \tag{7}$$

where $\mathtt{I}[A]$ is the indicator function for the event A, taking the value 1 if A occurs and 0 otherwise. To estimate the parameters in the AR and TAR models, we use a "rolling" estimation scheme where the parameters are re-estimated for all $t = R, \ldots, R + P - 1$, using a moving window of R observations.[2] The estimation sample size is set equal to $R = 128$ or 256, while the number of forecasts P takes the same set of values as in the previous experiment.

[2] Results for "fixed" estimation (where the parameters are estimated only once, using the observations $t = 1, , \ldots, R$) and "recursive" estimation (where the parameters are re-estimated for all $t = R, \ldots, R + P - 1$, using an expanding window of observations) schemes are qualitatively similar, and are available upon request.

In the third and fourth experiments, the TAR model (7) is taken to be the DGP. We consider models which either have a change in the slope parameter only (by setting $\phi_{1,0} = \phi_{2,0} = 0$, $\phi_{1,1} = -0.7$ and $\phi_{2,1} = 0.3$) or a change in both the slope and the intercept (by setting $\phi_{1,0} = -1.25$, $\phi_{2,0} = 0$, $\phi_{1,1} = -0.7$ and $\phi_{2,1} = 0.3$), and $\eta_t \sim N(0,1)$ throughout. The threshold c is varied in such a way that the "lower" regime ($y_{t-1} \leq c$) contains $20, 30, \ldots, 80\%$ of the observations on average. One-step ahead forecasts are obtained from AR and TAR models, as before.

In all experiments, the competing forecasts are evaluated using weighted and unweighted versions of the Diebold-Mariano test statistic using a quadratic loss function $g(e_{i,t|t-h}) = e^2_{i,t|t-h}$. For the weighted test statistics, we consider the three weight functions discussed in Section 2. The unconditional density function of y_t is estimated using the relevant in-sample observations by means of a standard Nadaraya-Watson kernel estimator with automatic bandwidth selection, while the empirical CDF is used as an estimate of $\Phi(y_t)$.[3] For all tests we incorporate the modification of Harvey *et al.* (1997) as given in (5) and use the Student's t-distribution with P-1 degrees of freedom to obtain the relevant critical values. We report rejection frequencies of the null hypothesis of equal forecast accuracy at a one-sided 5% nominal significance level. Results for other significance levels are available upon request. In the first experiment, the relevant alternative hypothesis is that model 1 forecasts better. In the second (third and fourth) experiment(s), the null is tested against the alternative that the AR (TAR) model performs better than the TAR (AR) model. Finally, we use 5000 replications in all experiments.

3.2 Monte Carlo results

Table 1 displays results from the first experiment, with the first panel referring to size and the second to sixth panels to power. Only results for $\theta = 0$ are shown to save space. Results for other values of θ are qualitatively similar, and are available upon request. All tests are reasonably sized, as the empirical rejection frequencies in the first panel are close to the nominal significance level of 5% throughout. The remaining panels show that for all tests power increases with (i) the number of forecasts P, (ii) with the difference between the forecast error variances $1 - k$, and (iii) with the contemporaneous correlation ρ. Additionally, the unweighted version of the test statistic is more powerful than any of the weighted variants. This was to be expected, given the design of these experiments. The only difference between

[3] For the first experiment, we generate artificial observations y_t, $t = 1, \ldots, R + P$ with $R = 256$ from a standard normal distribution.

the two forecasting models that implicitly are used is a difference in forecast error variance. None of the weighting functions employed here is designed to capture this sort of difference, such that using a weighted test statistic effectively means that less observations are used for the forecast evaluation, inevitably leading to a loss in power.

Table 2 displays results from the second experiment, where the DGP is the AR(1) model (6). It is clear that the power is not extremely high, unless the number of forecasts becomes very large. This should come as no surprise, as the SETAR model nests the linear model, and hence the SETAR parameters will very much look like those of the linear AR model. Power declines as the estimation sample size R increases. The rejection frequencies of the unweighted test statistic and the LT and RT variants, which focus on one of the tails of the distribution of y_t, are approximately the same for all values of ϕ considered, except when $\phi = 0.9$. For the test that puts more weight on both tails of the distribution of y_t (T), power increases with ϕ.

Results from the experiments using the TAR model (7) as DGP are shown in Tables 3 and 4, for the case where only the slope parameter changes and where both the intercept and slope parameter are subject to change, respectively. Table 3 shows that for all tests power increases when the distribution of observations across the two regimes in the DGP is more asymmetric. Also, the power of the MW-DM(T) statistic, which puts more weight on observations in both tails of the distribution of y_t, is considerably higher than the unweighted test statistic. Both the LT and RT statistics perform worse than the unweighted statistic, suggesting that it is not worthwhile to focus on just one of the tails, at least not for these DGPs.[4] Table 4 shows that, in case both the intercepts and slopes are different in the two regimes, power increases as more observations are located in the lower regime. Again, the MW-DM(T) statistic substantially outperforms the unweighted statistic.

4 Forecasting US GDP growth with the floor-and-ceiling model

In this section we apply the weighted Diebold-Mariano tests to forecasts for US GDP generated by linear and nonlinear models. In particular, we consider the floor-

[4] In additional experiments with parameterizations adopted from Hansen (1997) ($\phi_{1,0} = 0$, $\phi_{2,1} = 0$, $\phi_{1,1} \in \{-0.8, -0.6, \dots, 0.8\}$ and $\phi_{2,0} \in \{0, 0.2, \dots, 1.2\}$), we found that the LR and RT test statistics can also be more powerful than the unweighted statistic.

and-ceiling model of Pesaran and Potter (1997), which includes "current-depth-of-recession" and "overheating" variables as additional regressors in a linear autoregressive model for the growth rate of output. To develop the model, recursively define the indicators F_t and C_t for the floor and ceiling regimes, respectively, as

$$F_t = \begin{cases} \mathtt{I}[y_t < r_F] & \text{if } F_{t-1} = 0, \\ \mathtt{I}[CDR_{t-1} + y_t < 0] & \text{if } F_{t-1} = 1, \end{cases} \tag{8}$$

$$C_t = \mathtt{I}[F_t = 0]\mathtt{I}[y_t > r_C]\mathtt{I}[y_{t-1} > r_C], \tag{9}$$

where y_t denotes the annualized quarterly GDP growth rate and the current-depth-of-recession variable is defined as

$$CDR_t = \begin{cases} (y_t - r_F)F_t & \text{if } F_{t-1} = 0, \\ (CDR_{t-1} - y_t)F_t & \text{if } F_{t-1} = 1, \end{cases} \tag{10}$$

and the overheating variable is given by

$$OH_t = C_t(OH_{t-1} + y_t - r_C). \tag{11}$$

The floor-and-ceiling model for output growth then is given by

$$\phi(L)y_t = \phi_0 + \theta_1 CDR_{t-1} + \theta_2 OH_{t-1} + v_t, \tag{12}$$

where $\phi(L) = 1 - \phi_1 L - \ldots - \phi_p L^p$, with the lag operator defined as $L^m y_t = y_{t-m}$ for all m, $\mathsf{E}[v_t|\Omega_{t-1}] = 0$, and the conditional variance of v_t is given by

$$\mathsf{E}[v_t^2|\Omega_{t-1}] \equiv H_t = \sigma_F^2 F_{t-1} + \sigma_{COR}^2 COR_{t-1} + \sigma_C^2 C_{t-1}, \tag{13}$$

where COR_t is the indicator for the corridor regime, defined as

$$COR_t = \mathtt{I}[F_t + C_t = 0].$$

The function of the current-depth-of-recession and overheating variables in (12) is to capture the dampening effects of shocks at business cycle troughs and peaks, respectively; see Pesaran and Potter (1997) for an extensive discussion and motivation of this model.

Following Pesaran and Potter (1997), we set $p = 2$ in (12), and estimate the parameters in the model using a grid search over the floor and ceiling thresholds r_F and r_C. We use quarterly observations on seasonally adjusted real US GDP, from 1953:3-2001:3. The parameter estimates based on the complete sample are given by $\hat{\phi}_0 = 1.52(0.46)$, $\hat{\phi}_1 = 0.35(0.079)$, $\hat{\phi}_2 = 0.21(0.079)$, $\hat{\theta}_1 = -0.45(0.22)$, $\hat{\theta}_2 = -0.041(0.027)$, $\hat{\sigma}_F = 5.03$, $\hat{\sigma}_{COR} = 3.64$, $\hat{\sigma}_C = 2.81$, $\hat{r}_F = -3.51$, $\hat{r}_C = 2.04$, where asymptotic standard errors are given in parentheses. In the effective estimation sample 22, 82 and 93 observations are located in the floor, corridor and ceiling regimes, respectively. Note that the coefficients on the current-depth-of-recession and overheating variables are negative, indicating that indeed they dampen the dynamics during contractions and rapid expansions.

For the forecasting exercise, the parameters in the floor-and-ceiling model are estimated recursively using a moving window of $R = 128$ observations, starting with 1953:3-1985:4 as the initial estimation period. One-step ahead forecasts of y_t are obtained for 1986:1-2001:3 ($P = 63$). Over this forecast period, there are 1, 29 and 33 observations in the floor, corridor and ceiling regimes. The mean squared prediction error (MSPE) of the floor-and-ceiling model is equal to 4.03, compared with an MSPE of 4.15 of a competing linear AR(2) model. The unweighted modified DM statistic is equal to 1.05, corresponding with a p-value of 0.15 when testing against the one-sided alternative that the floor-and-ceiling model renders more accurate forecasts. Hence, we would conclude that the nonlinear model does not have superior forecasting performance, cf. Pesaran and Potter (1997). Regime-specific evaluation of the forecasts does not change this conclusion. The p-values of modified DM statistics for those observations that are classified in the corridor and ceiling regimes are equal to 0.19 and 0.50, respectively.

However, our weighted DM statistics do lead to different outcomes. First, the MW-DM(T) statistic, which puts more weight on observations in both tails of the unconditional distribution of y_t, is equal to 1.67, corresponding with a (one-sided) p-value of 0.050. This suggests that the nonlinear model may be superior to the linear model in forecasting extreme growth rates. Distinguishing between large positive and negative observations of y_t renders even sharper insights: the MW-DM(LT) statistic is equal to 1.97 (0.027), while the MW-DM(RT) statistic is equal to -1.22 (0.89), where p-values are given in parentheses. Hence, the floor-and-ceiling model seems to forecast extreme "recessionary observations" for which the GDP growth rate is small more accurately than the linear model.

5 Concluding remarks

Despite quite abundant evidence for the presence of nonlinearities in macroeconomic and financial variables, nonlinear time series models have not been able to produce consistently superior forecasts compared to linear models. One of the reasons for this poor track record put forward in the literature is the use of inappropriate evaluation criteria. In this paper we build upon this argument by proposing a novel forecast evaluation methodology, where different weights are given to different forecasts. This is based on the idea that in many situations certain observations are more important to predict accurately than others. In particular, extreme observations of the variable of interest (such as output growth or stock returns) often are the most relevant ones and, therefore, we propose to zoom in on the tails of the unconditional distribution of the variable of interest when evaluating forecasts from competing models.

Extensive Monte Carlo experiments showed that the resulting weighted tests of equal forecast accuracy dominate standard versions of the test in terms of power. The empirical application to forecasts of US GDP growth obtained from the floor-and-ceiling model illustrated the empirical relevance of distinguishing between forecasts of different types of observations.

Although in this paper we restricted ourselves to the Diebold-Mariano statistic with quadratic loss function, the idea of attaching different weights to different forecasts can straightforwardly be applied to other forecast evaluation statistics or tests of predictive ability, such as the tests of forecast encompassing discussed in Harvey *et al.* (1998). We leave this for future research. Another topic worth investigating is the impact of parameter uncertainty when forecasts from nonlinear models are involved. We ignored this in the present paper, given the asymptotic irrelevance of this type of uncertainty for the DM statistic with quadratic loss. In finite samples, however, parameter it may still be useful to account for parameter uncertainty.

References

Berkowitz, J. (2001), Testing Density Forecasts with Applications to Risk Management, *Journal of Business & Economic Statistics* **19**, 465–474.

Christoffersen, P.F. (1998), Evaluating Interval Forecasts, *International Economic Review* **39**, 841–862.

Clark, T.E. (1999), Finite-sample Properties of Tests for Equal Forecast Accuracy, *Journal of Forecasting* **18**, 489–504.

Clark, T.E. and M.W. McCracken (2001), Tests of Equal Forecast Accuracy and Encompassing for Nested Models, *Journal of Econometrics* **105**, 85–110.

Clements, M.P. and D.H. Hendry (1999), *Forecasting Non-stationary Economic Time Series*, Cambridge, MA: MIT Press.

Clements, M.P. and H.-M. Krolzig (1998), A Comparison of the Forecast Performance of Markov-Switching and Threshold Autoregressive Models of US GNP, *Econometrics Journal* **1**, C47–C75.

Clements, M.P. and J. Smith (1999), A Monte Carlo Study of the Forecasting Performance of Empirical SETAR Models, *Journal of Applied Econometrics* **14**, 124–141.

Clements, M.P. and J. Smith (2001), Evaluating Forecasts from SETAR Models of Exchange Rates, *Journal of International Money and Finance* **20**, 133–148.

Clements, M.P., P.H. Franses, J. Smith and D. van Dijk (2003), On SETAR Non-Linearity and Forecasting, *Journal of Forecasting*, to appear.

Dacco, R. and S. Satchell (1999), Why do Regime-Switching Models Forecast so Badly?, *Journal of Forecasting* **18**, 1–16.

De Gooijer, J.G. and K. Kumar (1992), Some Recent Developments in Non-Linear Time Series Modelling, Testing and Forecasting, *International Journal of Forecasting* **8**, 135–156.

Diebold, F.X. and R.S. Mariano (1995), Comparing Predictive Accuracy, *Journal of Business & Economic Statistics* **13**, 253–263.

Diebold, F.X. and J.A. Nason (1990), Nonparametric Exchange Rate Prediction, *Journal of International Economics* **28**, 315–332.

Diebold, F.X., T.A. Gunther and A.S. Tay (1998), Evaluating Density Forecasts with Applications to Financial Risk Management, *International Economic Review* **39**, 863–883.

Franses, P.H. and D. van Dijk (2000), *Nonlinear Time Series Models in Empirical Finance*, Cambridge: Cambridge University Press.

Giacomini, R. (2002a), Comparing Density Forecasts via Weighted Likelihood Ratio Tests: Asymptotic and Bootstrap Methods, UCSD Working Paper No. 2002-12, University of California San Diego.

Giacomini, R. (2002b), Tests of Conditional Predictive Ability, mimeo, University of California San Diego.

Harvey, D.I., S.J. Leybourne, and P. Newbold (1997), Testing the Equality of Prediction Mean Squared Errors, *International Journal of Forecasting* **13**, 281–291.

Hansen, B.E. (1997), Inference in TAR Models, *Studies in Nonlinear Dynamics and Econometrics* **2**, 1–14.

Harvey, D.I., S.J. Leybourne, and P. Newbold (1998), Tests for Forecast Encompassing, *Journal of Business & Economic Statistics* **16**, 254–259.

McCracken, M.W. (2000), Robust Out-of-Sample Inference, *Journal of Econometrics*, **99**, 195–223.

McCracken, M.W. and K.D. West (2002), Inference About Predictive Ability, in M.P. Clements and D.F. Hendry (eds.), *A Companion to Economic Forecasting*, Oxford: Basil Blackwell, 299–321.

Montgomery, A.L., V. Zarnowitz, R.S. Tsay and G.C. Tiao (1998), Forecasting the US Unemployment Rate, *Journal of the American Statistical Association* **93**, 478–493.

Pesaran, M.H. and S.M. Potter (1997), A Floor and Ceiling Model of US Output, *Journal of Economic Dynamics and Control* **21**, 661–695.

Ramsey, J.B. (1996), If Nonlinear Models Cannot Forecast, What Use Are They?, *Studies in Nonlinear Dynamics and Econometrics* **1**, 65–86.

Stock, J.H. and M.W. Watson (1999), A Comparison of Linear and Nonlinear Univariate Models for Forecasting Macroeconomic Time Series, in R.F. Engle and H. White (eds.), *Cointegration, Causality, and Forecasting: A Festschrift in Honour of Clive W.J. Granger*, Oxford: Oxford University Press, pp. 1–44.

van Dijk, D., P.H. Franses and A. Lucas (1999), Testing for Smooth Transition Nonlinearity in the Presence of Additive Outliers, *Journal of Business & Economic Statistics* **17**, 217–235.

West, K.D. (1996), Asymptotic Inference about Predictive Ability, *Econometrica* **64**, 1067–1084.

West, K.D. (2001), Tests for Forecast Encompassing when Forecasts Depend on Estimated Regression Parameters, *Journal of Business & Economic Statistics* **19**, 29–33.

West, K.D. and M.W. McCracken (1998), Regression Based Tests of Predictive Ability, *International Economic Review* **39**, 817–840.

Table 1: Size and power of weighted modified Diebold-Mariano test statistics of equal forecast accuracy based on squared loss

		$\rho = 0$					$\rho = 0.5$					$\rho = 0.9$				
Test	P	16	32	64	128	256	16	32	64	128	256	16	32	64	128	256
$k = 1$																
MW-DM		4.9	5.1	5.1	5.2	5.3	5.2	5.1	5.0	5.2	5.5	5.2	4.5	5.2	4.8	5.3
MW-DM(T)		4.3	5.0	5.0	5.0	5.5	4.4	4.6	5.0	4.7	5.1	4.5	4.5	4.8	4.6	4.7
MW-DM(LT)		4.5	4.9	4.9	5.5	5.9	4.8	4.7	5.0	5.4	5.6	4.8	4.7	4.9	5.0	5.4
MW-DM(RT)		4.4	5.2	5.2	5.0	5.2	4.5	5.2	5.1	5.0	5.2	4.5	4.6	4.8	4.8	4.9
$k = 0.9$																
MW-DM		7.5	8.9	11.1	15.1	22.2	7.9	9.5	12.2	17.1	25.5	10.9	15.9	24.1	37.9	60.8
MW-DM(T)		5.5	7.4	9.0	11.6	15.8	6.2	7.7	9.9	12.5	18.0	8.2	11.3	17.7	25.6	40.4
MW-DM(LT)		6.2	7.9	9.8	13.1	18.7	6.6	7.9	10.9	14.9	21.6	9.5	13.4	20.5	31.7	50.9
MW-DM(RT)		6.5	8.2	9.9	12.9	18.9	6.8	9.2	10.9	14.3	21.1	9.3	13.7	20.6	30.9	50.9
$k = 0.8$																
MW-DM		10.9	14.9	22.0	35.2	55.8	11.7	17.2	27.1	41.9	65.7	22.4	38.8	64.1	88.0	99.0
MW-DM(T)		7.7	11.2	15.9	23.8	37.0	8.7	12.3	18.9	27.7	43.6	15.0	25.5	42.5	66.8	90.5
MW-DM(LT)		9.1	12.6	18.6	29.3	46.2	9.5	14.6	22.4	35.1	55.4	18.1	32.0	53.2	79.3	96.7
MW-DM(RT)		8.9	13.3	18.9	29.3	45.9	9.8	14.8	22.3	34.3	55.5	18.4	31.7	53.7	78.8	96.0
$k = 0.7$																
MW-DM		15.3	23.9	40.1	64.2	88.8	18.0	29.5	48.6	73.9	94.6	38.9	68.1	92.9	99.6	100.0
MW-DM(T)		10.6	17.4	27.4	44.0	67.2	12.7	19.4	32.5	52.8	77.2	24.7	46.2	74.1	94.6	99.8
MW-DM(LT)		12.8	20.3	33.1	54.4	80.0	14.8	24.4	40.6	64.2	88.6	31.7	58.5	85.8	98.7	100.0
MW-DM(RT)		13.2	21.1	33.0	54.0	79.1	15.3	24.1	40.1	62.9	87.6	31.1	58.5	85.3	98.3	100.0
$k = 0.6$																
MW-DM		22.1	38.8	63.8	88.1	99.2	26.6	47.2	75.0	94.6	99.9	60.2	90.6	99.6	100.0	100.0
MW-DM(T)		15.1	26.8	43.4	68.2	91.6	17.7	31.1	52.4	77.9	96.4	38.7	70.2	93.8	99.8	100.0
MW-DM(LT)		17.6	31.5	53.1	79.5	97.1	21.7	39.2	64.1	88.7	99.3	48.5	82.0	98.4	100.0	100.0
MW-DM(RT)		19.3	32.2	53.5	79.0	97.3	22.9	38.8	64.0	88.6	98.9	48.9	81.9	98.1	100.0	100.0
$k = 0.5$																
MW-DM		33.4	58.8	85.8	98.5	100.0	39.4	68.6	93.1	99.6	100.0	78.6	98.9	100.0	100.0	100.0
MW-DM(T)		21.0	39.6	64.8	88.8	99.2	25.1	47.0	74.7	94.4	99.9	53.2	86.7	99.2	100.0	100.0
MW-DM(LT)		26.8	48.1	76.6	95.6	99.9	31.6	57.4	85.3	98.5	100.0	66.9	95.1	99.9	100.0	100.0
MW-DM(RT)		26.8	47.7	75.6	95.6	99.8	32.5	57.9	85.5	98.3	100.0	67.1	95.0	100.0	100.0	100.0

Notes: The table reports rejection frequencies of the null of equal forecast accuracy based on a squared loss function, at a one-sided nominal 5% significance level. The DGP is the bivariate normal forecast error process $\{(e_{1,t}, e_{2,t})\}_{t=R+1}^{R+P}$, with variances k and 1, respectively, and with contemporaneous correlation ρ. P denotes the number of forecasts. MW-DM is the standard modified DM statistic. MW-DM(T) is the modified weighted DM statistic as given in (5), with weight function $w_{\mathrm{T}}(\omega_t) = 1 - \phi(y_t)/\max(\phi(y_t))$, where $\phi(\cdot)$ is the density function of y_t. MW-DM(LT) and (RT) are modified weighted DM statistics with weight functions $w_{\mathrm{LT}}(\omega_t) = 1 - \Phi(y_t)$, where $\Phi(\cdot)$ is the cumulative distribution function of y_t, and $w_{\mathrm{RT}}(\omega_t) = \Phi(y_t)$, respectively. All results are based on 5000 replications.

Table 2: Rejection frequencies of weighted (modified) Diebold-Mariano test statistics of equal forecast accuracy based on squared loss

		$R = 128$					$R = 256$				
Test	P	16	32	64	128	256	16	32	64	128	256
$\phi = 0.1$											
MW-DM		7.8	12.5	20.4	34.0	58.2	6.4	9.3	14.3	21.8	36.2
MW-DM(T)		4.6	6.7	9.5	12.9	18.9	4.1	5.6	8.0	9.8	13.7
MW-DM(LT)		8.2	11.4	15.9	24.0	38.9	7.8	10.2	13.0	17.0	25.5
MW-DM(RT)		8.6	12.6	16.9	24.7	40.8	7.1	9.5	12.6	17.1	25.0
$\phi = 0.3$											
MW-DM		7.7	12.5	20.0	33.5	57.1	7.5	11.0	14.9	22.0	35.8
MW-DM(T)		4.6	6.9	9.3	13.1	20.8	4.6	6.3	8.3	10.3	14.7
MW-DM(LT)		8.3	11.5	15.9	24.7	40.2	8.1	10.2	12.6	16.9	25.5
MW-DM(RT)		8.5	12.1	15.7	23.7	40.1	7.6	9.8	11.9	17.1	25.7
$\phi = 0.5$											
MW-DM		9.0	12.7	19.9	33.9	56.7	6.9	9.2	13.7	21.2	36.3
MW-DM(T)		6.0	7.6	10.8	15.7	23.8	4.9	6.8	8.9	11.1	17.1
MW-DM(LT)		8.9	11.3	16.2	25.2	41.0	7.5	10.0	12.5	16.5	25.8
MW-DM(RT)		9.0	12.2	16.8	24.5	40.3	7.5	9.9	12.1	17.6	25.9
$\phi = 0.7$											
MW-DM		8.7	13.4	20.7	34.9	58.5	7.0	10.5	14.7	21.2	35.3
MW-DM(T)		7.6	10.9	14.5	20.4	30.6	6.9	10.0	12.9	16.3	22.4
MW-DM(LT)		9.1	12.2	17.8	26.1	42.9	7.4	10.8	13.8	17.7	26.9
MW-DM(RT)		8.5	12.9	18.1	27.0	42.1	7.5	10.9	13.4	17.9	27.9
$\phi = 0.9$											
MW-DM		8.9	12.4	21.2	37.7	62.8	6.9	9.9	14.1	22.6	37.3
MW-DM(T)		11.7	16.2	23.5	32.1	44.3	9.9	15.3	21.0	27.0	34.1
MW-DM(LT)		9.0	11.6	18.3	28.3	47.0	7.2	10.1	13.5	20.0	30.5
MW-DM(RT)		9.0	11.7	18.1	28.7	46.6	7.1	10.1	14.1	19.5	29.7

Notes: The table reports rejection frequencies of the null hypothesis of equal forecast accuracy based on a squared loss function. The DGP is the AR(1) process $y_t = \phi y_{t-1} + \varepsilon_t$. P one-step ahead forecasts are obtained from AR and TAR models, where the parameters in these models are estimated using a rolling scheme with a moving window of R observations. See Table 1 for a description of the test statistics. The null is tested at a nominal 5% significance level against the one-sided alternative that the AR model renders more accurate forecasts. All results are based on 5000 replications.

Table 3: Rejection frequencies of weighted (modified) Diebold-Mariano test statistics of equal forecast accuracy based on squared loss

		$R = 128$					$R = 256$				
Test	P	16	32	64	128	256	16	32	64	128	256
20-80											
MW-DM		8.8	11.2	14.1	21.7	34.8	11.5	16.3	23.6	39.9	63.8
MW-DM(T)		13.0	22.5	35.5	57.0	81.5	16.2	28.6	45.4	72.0	93.7
MW-DM(LT)		5.1	7.8	10.8	16.1	25.1	6.6	11.8	18.4	30.0	49.5
MW-DM(RT)		5.5	7.6	10.1	13.7	21.4	7.0	10.1	15.1	25.0	42.0
30-70											
MW-DM		7.4	8.4	9.6	12.6	18.7	9.5	12.6	17.4	28.2	44.8
MW-DM(T)		10.6	17.3	25.0	42.4	65.3	13.5	22.0	34.9	58.3	82.9
MW-DM(LT)		4.1	6.1	7.3	9.0	13.3	5.5	9.4	13.1	19.5	32.2
MW-DM(RT)		5.5	6.5	8.0	9.7	12.7	6.6	8.7	12.6	19.2	29.6
40-60											
MW-DM		7.0	7.9	8.8	11.5	16.9	9.2	12.5	16.3	26.1	42.0
MW-DM(T)		10.3	16.6	24.5	40.2	62.4	12.7	21.8	33.4	55.2	80.6
MW-DM(LT)		4.1	6.0	7.1	8.3	11.8	5.2	9.2	12.5	18.8	29.8
MW-DM(RT)		5.4	6.1	7.4	8.8	11.5	6.1	8.5	12.3	18.1	27.9
50-50											
MW-DM		7.8	9.9	11.8	17.3	28.5	10.3	15.1	20.6	33.9	56.4
MW-DM(T)		12.2	19.6	31.0	50.9	75.4	14.4	25.7	41.8	66.3	90.3
MW-DM(LT)		4.0	6.4	8.5	11.2	18.7	5.1	9.5	14.0	23.4	39.8
MW-DM(RT)		6.3	8.1	9.3	12.3	18.5	6.9	11.0	15.2	23.6	37.6
60-40											
MW-DM		10.3	13.9	19.0	30.2	51.5	12.7	19.4	30.0	48.9	77.6
MW-DM(T)		15.8	27.2	43.8	68.0	90.3	17.5	32.4	53.8	80.7	97.4
MW-DM(LT)		4.3	8.2	12.2	19.9	34.7	5.6	10.8	19.1	34.8	59.5
MW-DM(RT)		7.9	10.9	14.1	21.6	34.1	9.6	14.1	21.1	34.1	54.9
70-30											
MW-DM		14.5	20.3	30.4	50.8	78.7	17.6	27.3	43.1	69.2	93.7
MW-DM(T)		21.7	38.6	60.1	83.9	97.6	25.0	45.4	71.0	92.8	99.7
MW-DM(LT)		6.6	11.6	19.2	34.8	60.2	7.7	15.1	28.2	49.8	79.9
MW-DM(RT)		10.6	15.3	22.4	35.4	58.3	12.2	18.4	29.7	48.9	75.7
80-20											
MW-DM		17.1	27.1	42.8	67.5	92.5	20.1	33.6	53.9	80.1	98.1
MW-DM(T)		26.7	49.0	73.6	93.0	99.6	30.4	55.1	80.8	96.8	100.0
MW-DM(LT)		9.8	18.3	31.0	50.5	77.8	11.7	23.6	41.3	66.3	91.8
MW-DM(RT)		9.9	18.2	29.0	48.8	78.2	11.6	21.2	35.4	58.9	86.9

Notes: The table reports rejection frequencies of the null hypothesis of equal forecast accuracy based on a squared loss function. The DGP is the TAR process given in (7), with a change in the slope parameter only ($\phi_{1,0} = \phi_{2,0} = 0$, $\phi_{1,1} = -0.7$ and $\phi_{2,1} = 0.3$). Headings "20-80", etc., indicate the average percentage of observations in the lower-upper regimes of the DGP. P one-step ahead forecasts are obtained from AR and TAR models, where the parameters in these models are estimated using a rolling scheme with a moving window of R observations. See Table 1 for a description of the test statistics. The null is tested at a nominal 5% significance level against the one-sided alternative that the TAR model renders more accurate forecasts. All results are based on 5000 replications.

Table 4: Rejection frequencies of weighted (modified) Diebold-Mariano test statistics of equal forecast accuracy based on squared loss

		$R = 128$					$R = 256$				
Test	P	16	32	64	128	256	16	32	64	128	256
20-80											
MW-DM		3.0	3.2	2.7	2.6	2.0	4.3	5.5	6.4	8.4	11.0
MW-DM(T)		4.9	7.2	9.9	13.4	19.7	6.3	10.3	15.4	26.2	40.6
MW-DM(LT)		2.1	2.2	2.2	1.7	1.4	2.7	4.2	5.0	6.6	7.9
MW-DM(RT)		4.0	3.9	2.5	2.1	2.2	5.6	5.6	4.8	5.8	7.9
30-70											
MW-DM		5.2	6.9	8.7	10.8	15.5	7.0	11.1	15.7	25.2	40.9
MW-DM(T)		8.0	14.4	23.4	36.4	54.8	10.0	20.5	33.8	54.5	77.4
MW-DM(LT)		2.1	3.4	4.8	5.9	8.3	2.1	5.0	9.1	15.6	24.8
MW-DM(RT)		6.3	7.0	7.6	8.2	12.1	8.6	9.8	12.0	17.6	27.9
40-60											
MW-DM		9.7	13.0	19.4	31.5	50.8	12.7	18.2	29.4	49.4	77.4
MW-DM(T)		14.6	27.2	45.3	68.6	88.6	17.4	33.9	57.0	81.4	96.7
MW-DM(LT)		3.2	5.7	9.5	16.4	31.6	3.3	7.3	15.6	29.6	55.9
MW-DM(RT)		9.3	11.7	15.2	23.0	36.9	12.1	16.2	21.8	34.8	57.7
50-50											
MW-DM		14.5	22.2	36.1	59.4	86.6	17.7	27.6	47.1	73.8	95.7
MW-DM(T)		22.8	42.0	67.0	89.4	98.8	26.9	49.0	76.5	95.3	99.9
MW-DM(LT)		4.0	9.2	18.9	36.5	65.7	4.2	11.3	25.9	51.1	83.8
MW-DM(RT)		12.2	17.9	26.2	42.7	70.4	15.9	21.7	33.9	55.7	83.4
60-40											
MW-DM		20.8	34.3	54.6	83.2	98.5	23.8	39.0	64.4	89.2	99.6
MW-DM(T)		32.7	58.0	83.3	97.5	100.0	37.1	63.6	88.8	98.9	100.0
MW-DM(LT)		5.7	15.4	31.5	60.7	90.6	6.3	18.4	39.3	72.8	96.3
MW-DM(RT)		16.5	25.5	40.3	64.5	91.0	18.8	28.7	46.2	72.3	95.2
70-30											
MW-DM		26.0	42.7	69.2	92.7	99.9	29.3	49.2	76.0	96.0	99.9
MW-DM(T)		41.2	67.8	90.6	99.4	100.0	45.0	73.5	94.1	99.8	100.0
MW-DM(LT)		9.6	22.1	45.6	76.4	96.9	11.1	27.2	54.5	86.0	98.9
MW-DM(RT)		18.4	30.9	50.3	78.8	97.7	20.9	34.0	55.2	82.2	98.4
80-20											
MW-DM		29.8	48.7	74.9	95.5	99.9	32.6	54.4	81.3	97.5	100.0
MW-DM(T)		44.8	73.0	93.1	99.5	100.0	48.2	78.2	96.0	99.9	100.0
MW-DM(LT)		15.0	32.0	54.7	81.8	97.9	17.3	37.8	64.3	90.7	99.5
MW-DM(RT)		18.5	33.0	55.6	84.4	99.2	19.8	35.6	59.7	86.3	99.5

Notes: The table reports rejection frequencies of the null hypothesis of equal forecast accuracy based on a squared loss function. The DGP is the TAR process given in (7), with a change in the intercept and slope parameter ($\phi_{1,0} = -1.25$, $\phi_{2,0} = 0$, $\phi_{1,1} = -0.7$ and $\phi_{2,1} = 0.3$). Headings "20-80", etc., indicate the average percentage of observations in the lower-upper regimes of the DGP. P one-step ahead forecasts are obtained from AR and TAR models, where the parameters in these models are estimated using a rolling scheme with a moving window of R observations. See Table 1 for a description of the test statistics. The null is tested at a nominal 5% significance level against the one-sided alternative that the TAR model renders more accurate forecasts. All results are based on 5000 replications.