Forecasting Industrial Production with Linear, Nonlinear, and Structural Change Models*

Boriss Siliverstovs[§]
DIW Berlin

Dick van Dijk[‡]
Econometric Institute
Erasmus University Rotterdam

Econometric Institute Report EI 2003-16

May 2003

Abstract

We compare the forecasting performance of linear autoregressive models, autoregressive models with structural breaks, self-exciting threshold autoregressive models, and Markov switching autoregressive models in terms of point, interval, and density forecasts for h-month growth rates of industrial production of the G7 countries, for the period January 1960-December 2000. The results of point forecast evaluation tests support the established notion in the forecasting literature on the favorable performance of the linear AR model. By contrast, the Markov switching models render more accurate interval and density forecasts than the other models, including the linear AR model. This encouraging finding supports the idea that non-linear models may outperform linear competitors in terms of describing the uncertainty around future realizations of a time series.

Keywords: Interval forecasts, Density forecasts, Forecast evaluation tests, Nonlinearity, Structural change.

JEL Classification Codes: C22, C53.

*This research was initiated while both authors were visiting the Department of Economics, University of California at San Diego. The hospitality and stimulating research environment which were provided there are gratefully acknowledged. The second author acknowledges financial support from the Netherlands Organization for Scientific Research (N.W.O.). We thank Philip Hans Franses, Niels Haldrup, Svend Hylleberg, Hans Christian Kongsted, and Allan Würtz as well as the participants of the seminars at the Department of Economics, University of Aarhus, German Institute for Economic Research (DIW Berlin), ERC/METU VI Economics conference in Ankara, and the 13th $(EC)^2$ meeting in Bologna for helpful comments and discussions. Also we are grateful to Pierre Perron, Hans-Martin Krolzig, and David Harvey for making their code for estimation of the models with multiple structural breaks, Markov switching autoregressive models, and for the multiple forecast encompassing test, respectively, available. The usual disclaimer applies.

[§]DIW Berlin, German Institute for Economic Research, Königin-Luise-Str. 5, 14195, Berlin, Germany, email: bsiliverstovs@diw.de, (corresponding author)

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

1 Introduction

The amount of evidence for the presence of permanent structural change and business cycle asymmetries in macroeconomic time series such as output, inflation and unemployment is ever increasing; see Stock and Watson (1996) and Marcellino (2002a), among many others. As these features cannot be adequately captured with conventional linear autoregressive models with constant parameters (AR), it comes as no surprise that the use of nonlinear time series models is becoming more and more widespread. The most popular alternative models can roughly be divided in two groups. On the one hand, linear models which allow for infrequent permanent structural changes in the parameters have been used extensively; see Culver and Papell (1997), Bai, Lumsdaine and Stock (1998), and Papell, Murray and Ghiblawi (2000) for recent examples. On the other hand, nonlinear models that allow for some sort of regime-switching have been applied to describe the different dynamic behavior of macroeconomic time series during recessions and expansions, including Markov switching autoregressive (MS-AR) models, self-exciting threshold autoregressive (SETAR) models, and smooth transition autoregressive (STAR) models, see Granger (2001) for a recent survey.

The existence of such a wide variety of alternative models naturally poses the question how different models should be compared and evaluated in empirical research. One possibility is to consider the relative out-of-sample forecasting performance of competing models. Most studies that have taken this route have concentrated on comparing different models in terms of *point* forecasts; see Clements and Krolzig (1998), Stock and Watson (1999) and Marcellino (2002b) for recent examples. As summarized in De Gooijer and Kumar (1992) and Ramsey (1996), a general finding in the forecasting literature is that non-linear models do not render more accurate point forecasts than their linear competitors. The evidence from more recent studies such as the ones cited above does not appear to be at odds with this conclusion. Recently, the idea has arisen that non-linear models may be more suitable for describing the uncertainty around point forecasts, as suggested by Pesaran and Potter (1997), for example. This has led to a small but growing number of empirical studies comparing the forecasting performance of linear and non-linear models in terms of *interval* forecasts and *density* forecasts; see Clements and Smith (2000, 2001), Boero and Marrocu (2002) and Clements *et al.* (2003), among others. From these and other studies it appears that indeed nonlinear models may be superior to linear benchmark models in those respects.

Most previous studies have been limited, not only in the sense that they focus on evaluation of the models in the terms of point or interval or density forecasts only, but also because they usually involve only a pairwise comparison of alternative models, e.g.

AR vs. SETAR, or SETAR vs. MS-AR, etc. In addition, they tend to exclusively focus on US data; notable exceptions include Byers and Peel (1995) and Marcellino (2002b). What appears to be missing in the literature is a thorough evaluation of the forecasting abilities of alternative models in a more general setting, where the performance of linear, structural change, and nonlinear models is compared in terms of point forecasts as well as interval and density forecasts, both for US and other countries' data. In this paper we fill this gap by examining the relative out-of-sample forecasting performance of linear AR models, structural change models, SETAR models, and two MS-AR models for h-month growth rates of industrial production from the G7 countries in terms of point, interval, and density forecasts.

Our most important conclusions can be summarized as follows. The results of point forecast evaluation tests support the established notion in the forecasting literature that the linear AR model is a robust forecasting device, which is rarely beaten by nonlinear models. By contrast, we find that the Markov switching models render more accurate interval and density forecasts than the other models, including the linear AR model. This rather encouraging finding supports the idea that nonlinear models may outperform their linear competitors in terms of describing the uncertainty around future realizations of a time series.

The remainder of the paper is structured as follows. In Section 2, we discuss the industrial production data. The five different models under evaluation are described in Section 3. In Section 4 we provide details on the recursive procedure that is used to specify the models, to estimate the model parameters and to obtain the out-of-sample forecasts. In Section 5, we discuss the evaluation criteria that are used to compare the models in terms of point, interval and density forecasts. Section 6 contains the empirical results and, finally, Section 7 concludes.

2 Data

We examine seasonally adjusted monthly industrial production (IP) series for the G7 countries Canada (CA), Germany (DE), France (FR), Italy (IT), Japan (JP), United Kingdom (UK), and United States (US). The data are taken from the OECD *Main Economic Indicators*. The sample period runs from January 1960 to December 2000 (492 observations), except for Canadian industrial production, which starts in January 1961. The data are transformed to monthly growth rates by taking first differences of logarithms.

The series are adjusted for outliers, which we identified as those observations falling

outside the ±3 standard deviation band around the mean growth rate, computed over the complete sample period, following Simpson, Osborn and Sensier (2001), among others. The relevant observations are adjusted by linear interpolation of the original IP series in levels. The number of observations that are identified as aberrant is fairly small.* In particular, after January 1976, which marks the beginning of the forecasting period, we adjust only 1 observation for Germany and 2 for the UK. We arguably used information from the full sample for the outlier identification and, hence, this procedure would not have been feasible in "real-time" forecasting. However, the alternative approach to identify aberrant observations recursively as the in-sample period expands gives identical results.

3 Models

In this section we briefly describe the five different univariate time series models under scrutiny: a linear autoregressive model, an autoregressive model with multiple structural changes, a self-exciting threshold autoregressive model, and two variants of the Markov switching autoregressive model. For more elaborate discussions of the nonlinear models, we refer to Krolzig (1997), Tong (1990) and Franses and van Dijk (2000).

3.1 Linear autoregressive model

A linear autoregressive (AR) model for the monthly IP growth rate series is given by

$$\Delta y_t = \phi_0 + \phi_1 \Delta y_{t-1} + \ldots + \phi_p \Delta y_{t-p} + \varepsilon_t, \qquad t = 1, \ldots, T, \tag{1}$$

where y_t denotes the log-level of the time series of interest, Δ denotes the first differencing operator, defined by $\Delta_k y_t \equiv y_t - y_{t-k}$ for all $k \neq 0$ and $\Delta \equiv \Delta_1$, $\varepsilon_t \sim \mathsf{iid}(0, \sigma^2)$ is a disturbance term, and T denotes the sample size.

In the recursive specification procedure to be discussed in the next section, we use the Bayesian Information Criterion (BIC) to determine the lag order p in (1),

$$\mathrm{BIC}(p) = T \ln \hat{\sigma}^2(p) + (p+1) \ln T, \tag{2}$$

with the minimum and maximum orders set equal to $p_{\min} = 0$ and $p_{\max} = 12$, respectively.

*The number of identified outliers is equal to 0 for Canada, Japan, and the US, 2 for France and Italy, 3 for Germany, and 6 for the UK.

3.2 Autoregressive model with multiple structural changes

An autoregressive model with m structural changes (MSC) reads

$$\Delta y_t \begin{cases} \phi_{1,0} + \phi_{1,1}\Delta y_{t-1} + \ldots + \phi_{1,p}\Delta y_{t-p} + \sigma_1\eta_t, & \text{if } 1 < t \leq \tau_1, \\ \phi_{2,0} + \phi_{2,1}\Delta y_{t-1} + \ldots + \phi_{2,p}\Delta y_{t-p} + \sigma_2\eta_t, & \text{if } \tau_1 < t \leq \tau_2, \\ \vdots & \\ \phi_{m,0} + \phi_{m,1}\Delta y_{t-1} + \ldots + \phi_{m,p}\Delta y_{t-p} + \sigma_m\eta_t, & \text{if } \tau_{m-1} < t \leq T, \end{cases} \tag{3}$$

where $\tau_1, \ldots, \tau_{m-1}$ denote the $m-1$ break points, and $\eta_t \sim \mathsf{iid}(0,1)$. We use the procedures developed by Bai and Perron (1998) to specify an adequate model as follows. First, we select the lag order p by minimizing the BIC for the linear AR model (1) as above. Second, we apply the sequential test procedure of Bai and Perron (1998) to determine the number of structural breaks. This means that we test for one break at a time, i.e. conditional on finding a first break we test for the presence of a second break and so on, until no further breaks are found or the pre-specified maximum number of breaks is reached. In the sequential break testing procedure, we restrict the maximum number of breaks to 5, require the smallest segment to contain at least 15% of the available observations and use a 5% significance level throughout. Once the number of breaks is determined we apply the repartition procedure discussed in Bai and Perron (1998) to fine-tune the break dates. The out-of-sample forecasts are obtained from either the model estimated for the last segment, if any breaks are detected, or from the linear AR model estimated for the complete sample, if no breaks are detected.

3.3 Self-exciting threshold autoregressive model

A two-regime self-exciting threshold autoregressive model is given by

$$\Delta y_t \begin{cases} \phi_{1,0} + \phi_{1,1}\Delta y_{t-1} + \ldots + \phi_{1,p}\Delta y_{t-p} + \sigma_1\eta_t, & \text{if } \Delta_d y_{t-1} \leq r, \\ \phi_{2,0} + \phi_{2,1}\Delta y_{t-1} + \ldots + \phi_{2,p}\Delta y_{t-p} + \sigma_2\eta_t, & \text{if } \Delta_d y_{t-1} > r, \end{cases} \tag{4}$$

where r is the threshold value and $\eta_t \sim \mathsf{iid}(0,1)$. Note that we allow for regime-dependent heteroskedasticity in the shocks $\varepsilon_t \equiv \sigma_i\eta_t$ for $i = 1, 2$. For the transition variable that governs changes in regime we use the first lag of the d-period growth rate $\Delta_d y_{t-1}$, following Hansen (1997) and Koop and Potter (1999), among others.

The autoregressive order p and the length of the difference d in (4) are determined jointly by minimizing the BIC

$$\text{BIC}(p,d) = T_1 \ln(\hat{\sigma}_1^2(p,d)) + T_2 \ln(\hat{\sigma}_2^2(p,d)) + (p+1)\ln T_1 + (p+1)\ln T_2, \tag{5}$$

where $T_1 = \sum_{t=1}^{T} \mathtt{I}[\Delta_d y_{t-1} \leq r]$, with $\mathtt{I}[A]$ the indicator function for the event A, equals the number of observations in the "lower" regime, and $T_2 = T - T_1$, $\hat{\sigma}_1^2(p,d) =$

$\sum_{t=1}^{T} \mathtt{I}[\Delta_d y_{t-1} \leq r]\hat{\varepsilon}_t^2(p,d)$ is the estimate of the residual variance in the lower regime and $\hat{\sigma}_2^2(p,d)$ is defined similarly. The notation $\hat{\varepsilon}_t(p,d)$ is used to highlight the fact that the residuals and corresponding variances are computed conditional on fixed values of p and d. The BIC(p,d) criterion is minimized by varying p from $p_{\min} = 0$ to $p_{\max} = 12$, and d from $d_{\min} = 1$ to $d_{\max} = 12$.

3.4 Markov switching autoregressive models

We employ two variants of the Markov switching autoregressive model. Using notation from Krolzig (1997), the first model is referred to as MSIAH, which means that we allow for regime-dependent intercepts (I), autoregressive parameters (A), and variances (H), similar to the SETAR model above. The second Markov switching model is the popular model introduced by Hamilton (1989), denoted as MSMH, which allows for regime-specific means (M) and variances (H) with the autoregressive parameters kept constant.

The MSIAH model is given by

$$\Delta y_t = \phi_{s_t,0} + \phi_{s_t,1}\Delta y_{t-1} + \ldots + \phi_{s_t,p}\Delta y_{t-p} + \sigma_{s_t}\eta_t, \tag{6}$$

where $\eta_t \sim\sim \mathsf{iid}\,\mathsf{N}(0,1)$ and s_t is an unobserved two-state first-order Markov process with transition probabilities $\mathsf{Pr}[s_t = j | s_{t-1} = i] = p_{ij}$, $i,j = 1,2$. The autoregressive order p is selected by minimizing the BIC over the range of $p_{\min} = 0$ to $p_{\max} = 12$.

The MSMH model reads

$$\Delta y_t - \mu_{s_t} = \phi_1(\Delta y_{t-1} - \mu_{s_{t-1}}) + \ldots + \phi_p(\Delta y_{t-p} - \mu_{s_{t-p}}) + \sigma_{s_t}\eta_t. \tag{7}$$

One of the key features of this model is that the conditional density of Δy_t depends not only on the current regime s_t, but also on the regimes that prevailed during the most recent p periods $s_{t-1}, \ldots, s_{t-p}$. This is in sharp contrast to the MSIAH model as given in (6), where the conditional density of Δy_t depends on the current regime only. The fact that we have to keep track of the last p regimes implies that using high autoregressive orders in MSMH-type models is not advisable due to induced parameter inflation and local optima of the likelihood function. Therefore, the autoregressive order is determined by minimizing the BIC, allowing p to vary between 0 and 6.

Finally, estimation of both Markov switching models is carried out by means of the Expectation-Maximization (EM) algorithm together with the smoothing filter of Kim (1994), see Hamilton (1994, Chapter 22) for details.

4 Recursive specification, estimation, and forecasting

Specification, estimation, and forecasting are done recursively on an expanding window of observations, starting with 1960.1-1975.12 and ending with 1960.1-2000.11. As we allow for a maximum lag order of 12 in all models (except MSMH), the first window corresponds with an effective sample size of $R = 179$ observations (one observation is lost by taking first differences). For each window and model, we select the autoregressive order (and identify the number of structural changes for the MSC model), estimate the model parameters and compute point, interval, and density forecasts of the h-month growth rates $\{\Delta_h y_t = y_t - y_{t-h}\}_{t=R+h}^{R+P}$ for $h = 1, \ldots, 12$, where $P = 300$. This procedure gives us $P_h = P - (h-1)$ forecasts for the h-month growth rate, $h = 1, \ldots, 12$.

Under the assumption of normally distributed error terms there exist analytic expressions for the point, interval, and density forecasts for the linear AR model. However, this generally is not true if the assumption of Gaussianity is relaxed and/or if nonlinear models are considered. We follow the usual practice and obtain the forecasts of interests by means of Monte Carlo simulation. More precisely, for each model estimated for a given window containing, say, T observations, we simulate 100.000 future paths of the monthly growth rate series up to 12 months ahead, where we draw innovations for each of the models as follows. For the linear AR model we obtain the innovations by drawing with replacement from the vector of the residuals $\{\hat{\varepsilon}_t\}_{t=1}^{T}$. For the MSC model, we obtain the innovations in two steps: first, we standardize the residuals for the different segments using the segment-specific standard deviations $\hat{\sigma}_j$, $j = 1, \ldots, m$, where m is the number of detected structural breaks plus one; second, we draw the innovations from the pooled vector of standardized residuals $\{\hat{\eta}_t\}_{t=1}^{T}$ multiplied by the estimated standard deviation of the last segment, $\hat{\sigma}_m$. We follow a similar procedure in case of the SETAR model, where we first standardize the residuals using the regime-specific standard deviations, and secondly, we obtain the regime-specific innovations by resampling from this vector of standardized residuals multiplied by the relevant standard deviations. For the Markov switching models we obtain innovations by drawing them directly from the standard normal density and multiplying these by the estimated regime-specific standard deviations.

Finally, we should remark that both in constructing and evaluating the out-of-sample forecasts, we treat the point estimates of the model parameters as the true population values. This assumption can be relaxed, as discussed in van Dijk, Teräsvirta and Franses (2002) and McCracken and West (2002).

5 Forecast evaluation

In this section we discuss the evaluation criteria used to compare the predictive performance of the time series models in terms of point forecasts, interval forecasts, and density forecasts.

5.1 Point forecasts

Let $\{\Delta_h \hat{y}^{(i)}_{t|t-h}\}_{t=R+h}^{R+P}$ denote the sequence of forecasts of the h-month growth rate $\Delta_h y_t$ of length $P_h = P - (h-1)$, obtained from model M_i. The corresponding forecast error is denoted $e^{(i)}_{t|t-h} = \Delta_h y_t - \Delta_h \hat{y}^{(i)}_{t|t-h}$. For evaluation of point forecasts, we consider the popular Mean Squared Forecast Error, MSFE$=\frac{1}{P_h}\sum_{t=R+h}^{R+P}\left(\Delta_h y_t - \Delta_h \hat{y}^{(i)}_{t|t-h}\right)^2$, and the Mean Absolute Forecast Error, MAFE$=\frac{1}{P_h}\sum_{t=R+h}^{R+P}\left|\Delta_h y_t - \Delta_h \hat{y}^{(i)}_{t|t-h}\right|$. To assess the statistical significance of differences in these measures for two competing models M_i and M_j we use the test of equal forecast accuracy developed by Diebold and Mariano (1995). Let $g(e^{(i)}_{t|t-h})$ denote the loss associated with the forecast of the h-month growth rate $\Delta_h y_t$ from M_i. The null hypothesis of equal forecast accuracy for models M_i and M_j is given by $\mathsf{E}\left[g(e^{(i)}_{t|t-h})\right] = \mathsf{E}\left[g(e^{(j)}_{t|t-h})\right]$. Put differently, defining the loss differential as $d_t \equiv g(e^{(i)}_{t|t-h}) - g(e^{(j)}_{t|t-h})$, equal forecast accuracy implies $\mathsf{E}[d_t] = 0$. Given a covariance-stationary sequence of loss differentials $\{d_t\}_{t=R+h}^{R+P}$ of length P_h, the Diebold-Mariano (DM) statistic for testing the null hypothesis of equal forecast accuracy is given by

$$\text{DM} = \frac{\overline{d}}{\sqrt{\widehat{V\left(\overline{d}\right)}}} \xrightarrow{d} N\left(0,1\right), \tag{8}$$

where $\overline{d}$ is the sample mean loss differential $\overline{d} = \frac{1}{P_h}\sum_{t=R+h}^{R+P} d_t$, and where $\widehat{V(\overline{d})}$ is a consistent estimate of the asymptotic variance of $\overline{d}$. The latter usually is computed as an unweighted sum of the sample autocovariances up to order $h-1$, that is,

$$\widehat{V(\overline{d})} = \frac{1}{P_h}\left(\hat{\gamma}_0 + 2\sum_{k=0}^{h-1}\hat{\gamma}_k\right), \tag{9}$$

where $\hat{\gamma}_k = \frac{1}{P_h}\sum_{t=R+h+k}^{R+P}\left(d_t - \overline{d}\right)\left(d_{t-k} - \overline{d}\right)$, assuming that the forecast errors for the h-month growth rate exhibit serial dependence up to order $h-1$. We apply the modifications suggested by Harvey, Leybourne and Newbold (1997) to correct the tendency of the Diebold-Mariano statistic to be oversized in small- and medium-size samples and to account for the possibility that forecast errors are fat-tailed. In particular, the

original Diebold-Mariano statistic (8) is multiplied by the correction factor

$$\text{CF} = \left(\frac{P_h + 1 - 2h + h(h-1)/P_h}{P_h} \right), \tag{10}$$

and the resulting statistic MDM=CF $\times$ DM is compared with critical values from the Student's t distribution with $(P_h - 1)$ degrees of freedom.

In addition, we compare point forecasts from different models by means of forecast encompassing tests. M_i is said to forecast encompass the competing model M_j if the forecasts from M_j contain no useful information on top of that contained in the forecasts from M_i. Essentially, a test for forecast encompassing can be based on the composite forecast $\Delta_h \hat{y}^{(c)}_{t|t-h}$, constructed as a linear combination of the forecasts from M_i and M_j,

$$\Delta_h \hat{y}^{(c)}_{t|t-h} = \alpha \Delta_h \hat{y}^{(j)}_{t|t-h} + (1-\alpha) \Delta_h \hat{y}^{(i)}_{t|t-h}, \tag{11}$$

where the coefficient α denotes the optimal weight attached to M_j's forecast. In this context, M_i forecast encompasses M_j if $\alpha = 0$. We use the test put forward by Harvey, Leybourne and Newbold (1998). These authors showed that a test for forecast encompassing can be carried out conveniently within the testing framework of Diebold and Mariano (1995) by computing (8) with $d_t = e^{(i)}_{t|t-h}(e^{(i)}_{t|t-h} - e^{(j)}_{t|t-h})$. The test is modified by incorporating the amendments suggested for the Diebold-Mariano test of equal forecast accuracy discussed above.

Finally, we consider the forecast encompassing test for the case of $K \geq 2$ competing forecasts developed in Harvey and Newbold (2000). The corresponding composite forecast can be written as follows

$$\Delta_h \hat{y}^{(c)}_{t|t-h} = \alpha_1 \Delta_h \hat{y}^{(1)}_{t|t-h} + \ldots + \alpha_{K-1} \Delta_h \hat{y}^{(K-1)}_{t|t-h} + (1 - \alpha_1 - ... - \alpha_{K-1}) \Delta_h \hat{y}^{(K)}_{t|t-h}. \tag{12}$$

In this case, the null hypothesis is that model M_K forecast encompasses all other $K-1$ models, which can be expressed as $\alpha_1 = \ldots = \alpha_{K-1} = 0$ in (12). Defining the $(K-1 \times 1)$ vector D_t with elements $d_{jt} = e^{(K)}_{t|t-h}(e^{(K)}_{t|t-h} - e^{(j)}_{t|t-h})$ for $j = 1, ..., K-1$, the suggested statistic for testing of zero mean of this vector is

$$\text{MS} = \frac{\overline{D} \widehat{V}^{-1} \overline{D}}{(K-1)(P_h - 1)(P_h - K + 1)}, \tag{13}$$

where the $(K-1) \times 1$ vector $\overline{D}$ consists of the sample means of d_{jt} for $j = 1, \ldots, K-1$ and $\widehat{V}$ is the sample covariance matrix constructed using the Newey-West estimator with Bartlett kernel in order to account for the possible serial dependence in the forecast errors.[†] We follow the suggestions of Harvey and Newbold (2000) to multiply the test

[†]Using a rectangular kernel with truncation lag $h-1$ to estimate V does not guarantee that the computed sample covariance matrix is positive-definite in finite samples

statistic MS by the correction factor CF given in (10) and obtain critical values from the F_{K-1,P_h-K+1} distribution.

5.2 Interval forecasts

Christoffersen (1998) argues that a good interval forecast should possess two essential properties. First, its empirical coverage should be close to the nominal coverage probability. Second, in the presence of (conditional) heteroskedasticity, the interval should be narrow in tranquil periods and wide in volatile periods. Put differently, the incidence of observations falling inside or outside the interval forecast should be spread out evenly over the sample and not come in clusters. To assess these two properties for the interval forecasts obtained from the various models, we apply the Pearson-type χ^2 tests developed in Wallis (2002). While these are asymptotically equivalent to the likelihood ratio tests put forward in Christoffersen (1998), the advantage of the Pearson-type tests is that they allow calculation of exact p-values when the number of forecasts is limited. This is relevant for our multiple-month growth rate forecasts, as explained below.

First consider the evaluation of interval forecasts for one-month growth rates. Let $L_{t|t-1}(q)$ and $U_{t|t-1}(q)$ denote the lower and upper limits of the interval forecast of Δy_t made at time $t-1$, for a given nominal coverage probability q. Define the sequence of indicator functions $\{I_{t|t-1}\}_{t=R+1}^{R+P}$ of length $P_1 = P$, where $I_{t|t-1}$ takes the value 1 when the realization Δy_t lies inside the forecast interval and 0 otherwise. The procedure of Wallis (2002) consists of three tests: a test of correct unconditional coverage, a test of independence, and a test of correct conditional coverage. All three tests have the common form

$$X^2 = \sum (O - E)^2 / E, \tag{14}$$

measuring the discrepancy between the observed outcome (O) and the expected outcome (E) under the appropriate null hypothesis. The test of correct unconditional coverage compares the sample proportion of times that the interval forecast includes the realization Δy_t, denoted as $\hat{\pi}$, with the nominal coverage probability q, where $\hat{\pi}$ is computed as

$$\hat{\pi} = \frac{n_1}{n_0 + n_1},$$

where $n_1 = \sum_{t=R+1}^{R+P} I_{t|t-1}$ and $n_0 = P - n_1$. Under the null hypothesis of correct unconditional coverage, the expected frequencies of observing the actual value Δy_t inside and outside the interval forecast are equal to $m_1 = qP$ and $m_0 = (1-q)P$,

respectively. The test statistic for unconditional coverage is given by

$$X^2 = \frac{P(\hat{\pi} - q)^2}{q(1-q)} = \sum_{i=1}^{2} \frac{(n_i - m_i)^2}{m_i}. \tag{15}$$

In the test of correct unconditional coverage only the overall number of interval "hits" matters. The test fails to capture potential temporal dependence in the binary indicator sequence $\{I_{t|t-1}\}_{t=R+1}^{R+P}$, which reveals itself in the fact that ones and zeros tend to cluster instead of being spread out evenly across the sample. In order to detect such dynamics, a test for independence may be performed, which is based on the number of transitions n_{ij} of $I_{t|t-1}$ from "state" i to "state" j, for $i, j = 0, 1$. The observed contingency matrix is

$$\widehat{\Pi}^O = \begin{bmatrix} n_{00} & n_{01} \\ n_{10} & n_{11} \end{bmatrix}. \tag{16}$$

In this case, the expected frequencies are equal to $m_{ij} P(\frac{n_{i+}}{P})(\frac{n_{+j}}{P})$, with $n_{i+} = n_{i0} + n_{i1}$ and $n_{+j} = n_{0j} + n_{1j}$, $i, j = 0, 1$, being the relevant row and column sums in $\widehat{\Pi}^O$, respectively. The resulting test statistic is given by

$$X^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(n_{ij} - m_{ij})^2}{m_{ij}}. \tag{17}$$

Finally, the test for correct conditional coverage combines the tests of correct unconditional coverage and of independence. The test statistics is computed as in (17), where the observed transition counts are taken from (16), while the expected frequencies m_{ij} are now obtained from

$$\widehat{\Pi}_{cc}^E = \begin{bmatrix} (1-q)(n_{00} + n_{01}) & q(n_{00} + n_{01}) \\ (1-q)(n_{10} + n_{11}) & q(n_{10} + n_{11}) \end{bmatrix}.$$

The test statistics of correct unconditional coverage and of independence have asymptotic χ^2 distributions with 1 degree of freedom, while the test of correct conditional coverage follows an asymptotic χ^2 distribution with 2 degrees of freedom. In small samples the use of exact p-values might be desirable, as discussed in Wallis (2002).

Extending the above tests for evaluation of interval forecasts for the h-month growth rate with $h > 1$ is not straightforward, as the corresponding indicator sequence $\{I_{t|t-h}\}_{t=R+h}^{R+P}$ is expected to exhibit serial dependence up to order $(h-1)$ under the null. Here we adopt the approach advocated by Diebold, Gunther and Tay (1998) in the context of evaluation density forecasts, by partitioning $\{I_{t|t-h}\}_{t=R+h}^{R+P}$, into h sub-groups of independent observations $(I_{R+h|R}, I_{R+2h|R+h}, \ldots)$, $(I_{R+1+h|R+1}, I_{R+1+2h|R+1+h}, \ldots)$, $\ldots$, $(I_{R+(h-1)+h|R+(h-1)}, I_{R+(h-1)+2h|R+(h-1)+h}, \ldots)$. We then perform the interval evaluation tests on each of these sub-groups, using a significance level of α/h, and reject the

null hypothesis of correct calibration of the interval forecasts if the null hypothesis is rejected for at least one the h tests performed. This procedure ensures that the overall size of the test procedure is bounded by the selected significance level α. Note that the number of forecasts in the sub-groups quickly becomes smaller as h increases, from 150 for $h = 2$ to to 25 for $h = 12$, demonstrating the relevance of the small sample argument in favor of the Pearson-type tests discussed above.

5.3 Density forecasts

A density forecast is an estimate of the probability distribution of the future realization of the time series being forecast. Density forecasts thus provide a complete description of the uncertainty associated with future predictions. In that respect, density forecasts can be considered as a further elaboration upon interval forecasts, which specify only a single confidence interval for a given coverage probability for the variable of interest.

Evaluation of density forecasts is considered in Diebold *et al.* (1998). First, consider the evaluation of density forecasts of the 1-month growth rate Δy_t. Let $p_{t|t-1}(\cdot)$ be the 1-step ahead density forecast made ad time $t-1$ by a particular model and $f_t(\cdot)$ be the true predictive density of Δy_t. Diebold *et al.* (1998) suggest to compare $p_{t|t-1}(\cdot)$ with $f_t(\cdot)$ via the probability integral transform (PIT) $z_{t|t-1}$ of the actual realization of Δy_t with respect to the density forecast $p_{t|t-1}(\cdot)$:

$$z_{t|t-1} = \int_{-\infty}^{\Delta y_t} p_{t|t-1}(u)\, du \equiv P_{t|t-1}(\Delta y_t), \tag{18}$$

where $P_{t|t-1}(\cdot)$ denotes the CDF corresponding with $p_{t|t-1}(\cdot)$ Under the null hypothesis of equivalence of the density forecast and the true predictive density, i.e., $p_{t|t-1}(\cdot) = f_t(\cdot)$, the sequence of probability integral transforms $\left\{z_{t|t-1}\right\}_{t=R+1}^{R+P}$ is independently uniformly distributed on the [0,1]-interval, $\mathsf{iid}\,\mathsf{U}[0,1]$. We assess the uniformity of the sequence $\left\{z_{t|t-1}\right\}_{t=R+1}^{R+P}$ using the Kolmogorov-Smirnov (KS) test statistic, which compares the maximum deviation of the empirical cumulative density function from the theoretical one, where the appropriate critical value is calculated by means of the analytical formula provided in Miller (1956).

The KS test is based on the assumption of independence. While this may not be an unreasonable assumption in case of density forecasts for the 1-month growth rate, it is likely to fail when we consider h-month growth rate forecasts. Again, we follow the suggestion of Diebold *et al.* (1998) and partition the sequence $\left\{z_{t|t-h}\right\}_{t=R+h}^{R+P}$ into the h subgroups $(z_{R+h|R}, z_{R+2h|R+h}, \ldots)$, $(z_{R+1+h|R+1}, z_{R+1+2h|R+1+h}, \ldots)$, $\ldots$, $(z_{R+(h-1)+h|R+(h-1)}, z_{R+(h-1)+2h|R+(h-1)+h}, \ldots)$, for which the assumption of independence can be maintained. We perform a KS-test at significance level α/h on each of the partitioned

groups and reject the null hypothesis of correct calibration of the density forecasts if the null hypothesis is rejected at least once for the h tests performed.

We examine the maintained assumption of independence by testing for autocorrelation in the partitioned sequences of $\left\{z_{t|t-h}\right\}_{t=R+h}^{R+P}$ using the Ljung-Box test with the significance level α/h with $h = 1 - 12$ for power transformations of the demeaned sequences $(z_{t|t-h} - \overline{z}_{t|t-h})^k$ for $k = 1, \ldots, 4$.

Berkowitz (2001) suggests to apply the inverse normal cumulative density function transformation to the sequence of $\left\{z_{t|t-1}\right\}_{t=R+1}^{R+P}$. Given that under the null hypothesis the sequence $\left\{z_{t|t-1}\right\}_{t=R+1}^{R+P}$ is $\mathsf{iidU}\left[0,1\right]$, the sequence of transformed PIT's $\left\{z_{t|t-1}^{*}\right\}_{t=R+1}^{R+P}$ is $\mathsf{iidN}\left(0,1\right)$. Berkowitz (2001) suggests that normality tests may possess better power properties than that of the KS test. In this paper we test for normality of $\left\{z_{t|t-h}^{*}\right\}_{t=R+h}^{R+P}$ using the test statistic suggested in Doornik and Hansen (1994) (DH). For multiple-step ahead density forecasts, we test the normality assumption after partitioning the corresponding sequences as discussed above.

6 Empirical Results

In this section we discuss the empirical results of the forecasting exercise. First, however, some remarks concerning the in-sample estimates of the various models are in order. For all models, the selected autoregressive order p is small, usually equal to 1 or 2 and never exceeding 4. The sequential break tests of Bai and Perron (1998) generally indicate the presence of a single structural change in the IP growth rates. Only for Japan and the US two breaks are detected when the estimation window extends to the 1990s. The estimated break dates differ widely, ranging from 1969 for the US to 1984 for Germany and the UK. The regimes in the Markov switching models are fairly persistent. The staying probability for the high-growth "expansionary" regime typically is larger than 0.95, while for the low-growth "recessionary" regime it lies between 0.8 and 0.9, confirming the established notion that recessions are shorter in duration than expansions. Finally, the length of the difference d used for the threshold variable $\Delta_d y_{t-1}$ in the SETAR models is fairly small for most countries and estimation windows - between 3 and 6 months. The threshold estimate generally is quite close to 0, indicating that the implied regimes closely correspond with recessions and expansions.

6.1 Evaluation of point forecasts

Table 1 reports the ratios of the MSFE for a given horizon h relative to the MSFE of the AR model with constant parameters, which is taken as the benchmark model.[‡] The rank reported in parentheses is such that the model with the smallest MSFE is assigned rank 1, the second smallest rank 2, etc. Table 2 reports the average rank across forecast horizons $h = 1 - 6$, $h = 7 - 12$, and $h = 1 - 12$.

Several conclusions emerge from these tables. First, note that the ratios reported in Table 1 generally are close to 1, indicating that the MSFE's for the different models do not differ dramatically. Second, the linear AR model has the lowest MSFE on average, see the final column of Table 2. For the shorter forecast horizons $h = 1 - 6$ the MSMH model is second-best, whereas for longer forecast horizons $h = 7 - 12$ the SETAR model comes second, although it should be noted that the average ranks of the different nonlinear models are rather close. Third, the performance of nonlinear models varies substantially across series and across horizons. For example, the SETAR model performs relatively poorly for France, the UK and the US, whereas it is the best-performing model for Canada and Japan. Similarly, for Germany the SETAR model ranks second-lowest for shorter forecast horizons but is superior to the other models for horizons $h = 7 - 12$. A similar observation can be made for the MSC model, which shows good performance for Italy and especially the UK, but renders by far the most inaccurate point forecasts for Canada and the US.

Table 3 provide pairwise model comparisons in terms of MSFE summarized across series and across models, respectively. The entries in the upper panel indicate the number of horizons for which a given row model i has produced smaller MSFE than column model j, where the numbers in parentheses indicate the number of times the difference between the two MSFE values was found to be statistically significant at the conventional 5% level according to the Diebold-Mariano test.

As expected, the linear AR model has the highest score, followed by the MSMH model for shorter forecast horizons and by the SETAR model for longer forecast horizons. According to the upper panel, most of the rejections of the null hypothesis of equal forecast accuracy occurred when tested against the MSIAH model. The entries of the lower panel indicate the number of times a row model rendered a smaller MSFE value than all other models for a given time series. As seen, the rejection rate is not uniform across the different time series. For Canada, Germany, Italy, Japan and the US none or rather few rejections occurred, whereas the opposite is the case for France

[‡]Results for the MAFE criterion are qualitatively similar and not shown here to save space. Detailed results are available upon request.

and the UK.

Tables 4 and 5 summarize the results of the pairwise and multiple forecast encompassing tests, respectively. The entries in the upper panel of Table 4 indicate the number of times the null hypothesis that row model i forecast encompasses column model j is rejected. The bottom panel of this table shows the number of rejections in the pairwise forecast encompassing test that occurred for row model i for a particular time series. The results of the pairwise forecast encompassing test reinforce the earlier conclusions: the AR model dominates the other models both for shorter- and longer forecast horizons, in the sense that it tends to forecast encompass other models and at the same time it is not encompassed by other models. The MSMH model seems to be second-best at the shorter forecast horizons whereas the SETAR model takes the second place at the longer forecast horizons. As seen from Table 5, the null hypothesis that the AR model forecast encompasses all other models simultaneously is rarely rejected. However, the same is true for the MSMH and MSIAH models and for the SETAR model at longer forecast horizons.

In sum, the results of the point forecast evaluation seem to be in line with previous findings in the literature, in that the linear AR model seems to be a robust forecasting device. It provides point forecasts that are often found to be superior to those of the structural change and nonlinear models, across different forecast horizons and across different countries. At the same time, the performance of non-linear models like MSMH and SETAR seems to depend heavily on the forecast horizon as well as the particular time series in question.

6.2 Evaluation of interval forecasts

Table 6 summarizes the outcomes of the interval forecast evaluation tests for nominal coverage probabilities of 50%, 75% and 90%. Because the number of forecasts is quite large for one-step ahead forecasts ($P_1 = 300$) we applied asymptotic critical values for $h = 1$ and calculated exact p-values only for $h = 2 - 12$. Table entries indicate the number of rejections of the respective null hypotheses of correct unconditional coverage, independence and correct conditional coverage at the 5% significance level. The ranking of the models radically changes compared to that based on the point forecast evaluation discussed above. The MSMH and MSIAH models offer the best performance almost uniformly across the three tests and nominal coverage levels considered. The tests of correct unconditional coverage indicate that the Markov switching models have empirical coverage rates that are closest to the nominal ones. Also when testing of independence the MSMH and MSIAH models offer favorable performance compared with

the competing models. It is then not surprising that the tests of correct conditional coverage suggest that the MSMH and MSIAH models outperform the other models. Also note that the linear AR model offers rather disappointing performance, which is either worst or next to worst.

6.3 Evaluation of density forecasts

Tables 7 and 8 provide results of the density forecast evaluation using the Kolmogorov-Smirnov and Doornik-Hansen tests, applied to the PIT and its inverse normal cumulative density function transformation. Table entries indicate the number of rejections of the null hypothesis of correct forecast density calibration at the 5% significance level. The results from the DH test are largely in line with the interval forecast evaluation tests, in the sense that the MSMH and MSIAH models are coined as the best models in terms of forecast densities. By contrast, the results of the KS test suggest that the MSC model provides the most accurate density forecasts.

In order to decide which test results to rely on, consider Table 9. The entries in this table denote the number of times that the actual $h-$period growth rate is completely outside the empirical forecast densities of each models, implying that the PIT is equal to 0 or 1. [§] The results depicted in Table 9 indicate that this number is rather large for the MSC model for Germany, Japan and the US. At the same time, for the MSMH and MSIAH models the actual observation falls outside the empirical density only once. For this reason, in forecast density evaluation we favor the performance of the Markow switching models over that of the MSC model.

As a final step in evaluating the density forecasts, we compute the Ljung-Box test for first-order autocorrelation in the partitioned PIT sequences. The results of this test are shown in Table 10. The incidence of the rejections of the null hypothesis of no autocorrelation using the 5% significance level varies from one time series to another. The number of rejections for MSMH and MSIAH is comparable with those for other models for odd values of $k = 1, 3$ and somewhat lower for even values of $k = 2, 4$.

Our conclusions based on the results in these tables can be summarized as follows. The evidence of the KS test in favor of the MSC model is undermined by the rather large number of actual observations of the $h-$period growth rates that fall outside of the empirical forecast densities. The DH tests indicate that the nonlinear Markov switching models produce superior density forecasts compared to the other models, including the

[§]Clearly, in this case application of the inverse normal cumulative density function transformation yields $z^*_{t|t-h} = \pm\infty$, which invalidates the use of the DH normality test. In order to make the DH test operational again we substituted the PIT with values close to 0 and 1.

linear AR model. This is confirmed by the LB test for first-order autocorrelation, which demonstrates that the MSMH and MSIAH models perform no worse or even somewhat better than the other models.

7 Conclusions

We have investigated the relative forecasting performance of linear autoregressive (AR) models, models with structural changes (MSC), self-exciting threshold autoregressive (SETAR) models, and Markov switching autoregressive models (MSIAH and MSMH) for $h-$month growth rates of industrial production of the G-7 countries for the period 1960.1 - 2000.12. We have evaluated and compared point, interval, and density forecasts generated from these models.

Our findings can be summarized as follows. The linear AR model seems to be quite a robust forecasting device, at least when it comes to point forecasting. The AR point forecasts were mostly found to be superior to those from the other models considered according to MSFE. This finding holds both at the shorter- and longer horizons as well as for different time series. The MSMH model is the best performing nonlinear model at shorter forecast horizons and the SETAR model at longer horizons. At the same time the performance of the SETAR model depends crucially on the time series in question, i.e. for some series it is ranked as the best model while for others it shows the worst performance. The remaining MSC and MSIAH models do not perform particularly well across all horizons and series considered. These results are confirmed by the Diebold-Mariano test of equal forecast accuracy and tests of forecast encompassing.

Both Markov switching models produce superior interval forecasts when compared to those from the other models, including the linear AR model. This finding holds irrespective of the nominal coverage probability considered. Finally, on the basis of density forecast evaluation tests the MSMH model followed by the MSIAH model are the best performers as well. This is a rather encouraging finding and supports the idea that nonlinear models may perform better than linear competitors in terms of describing the uncertainty around future realizations of a time series.

References

Bai, J., R.L. Lumsdaine, and J.H. Stock (1998), Testing for and dating common breaks in multivariate time series, *Review of Economic Studies* **65**, 395–432.

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

Berkowitz, J. (2001), Testing density forecasts, with applications to risk management, *Journal of Business and Economic Statistics* **19**, 465–474.

Boero, G. and E. Marrocu (2002), The performance of non-linear exchange rate models: A forecasting comparison, *Journal of Forecasting* **21**, 513–542.

Byers, J.D. and D.A. Peel (1995), Forecasting industrial production using non-linear methods, *Journal of Forecasting* **14**, 325–336.

Christoffersen, P. (1998), Evaluating interval forecasts, *International Economic Review* **39**, 841–862.

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 (2000), Evaluating the forecast densities of linear and non-linear models: Application to output growth and unemployment, *Journal of Forecasting* **19**, 255–276.

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.

Culver, S.E. and D.H. Papell (1997), Is there a unit root in the inflation rate? Evidence from sequential break and panel data methods, *Journal of Applied Econometrics* **12**, 435–444.

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., T. Gunther, and A. Tay (1998), Evaluating density forecasts with applications to financial risk management, *International Economic Review* **39**, 863–883.

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

Doornik, J.A. and H. Hansen (1994), A practical test for univariate and multivariate normality, Discussion Paper, Nuffield College, Oxford.

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

Granger, C.W.J. (2001), Overview of nonlinear macroeconometric empirical models, *Macroeconomic Dynamics* **5**, 466–481.

Hamilton, J.D. (1989), A new approach to the econometric analysis of nonstationary time series and the business cycle, *Econometrica* **57**, 357–384.

Hamilton, J.D. (1994), *Time Series Analysis*, Princeton: Princeton University Press.

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 (1997), Testing the equality of prediction mean squared errors, *International Journal of Forecasting* **13**, 281–291.

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

Harvey, D.I. and P. Newbold (2000), Tests for multiple forecast encompassing, *Journal of Applied Econometrics* **15**, 471–482.

Kim, C.-J. (1994), Dynamic linear models with Markov-Switching, *Journal of Econometrics* **60**, 1–22.

Koop, G. and S.M. Potter (1999), Dynamic asymmetries in US unemployment, *Journal of Business and Economic Statistics* **17**, 298–313.

Krolzig, H.-M. (1997), *Markov-Switching Vector Autoregressions: Modelling, Statistical Inference, and Application to Business Cycle Analysis*, Lecure Notes in Economic and Mathematical Systems, Vol. 454, Springer-Verlag.

Marcellino, M. (2002a), Instability and nonlinearity in the EMU, CEPR Working Paper No. 3312.

Marcellino, M. (2002b), Forecasting EMU macroeconomic variables, IGIER Working Paper No. 216.

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, pp. 299–321.

Miller, L.H. (1956), Table of percentage points of Kolmogorov statistics, *Journal of the American Statistical Association* **51**, 111–121.

Papell, D.H., C.J. Murray and H. Ghiblawi (2000), The structure of unemployment, *Review of Economics and Statistics* **82**, 309-315.

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.

Simpson, P.W., D.R. Osborn, and M. Sensier (2001), Forecasting UK industrial production over the business cycle, *Journal of Forecasting* **20**, 405–424.

Stock, J.H. and M.W. Watson (1996), Evidence on structural instability in macroeconomic time series relations, *Journal of Business and Economic Statistics* **14**, 11–30.

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.

Tong, H. (1990), *Non-Linear Time Series: A Dynamic System Approach*, Oxford: Oxford University Press.

Van Dijk, D., T. Teräsvirta and P.H. Franses (2002), Smooth transition autoregressive models - A survey of recent developments, *Econometric Reviews* **21**, 1–47.

Wallis, K.F. (2002), Chi-squired tests of interval and density forecasts, and the Bank of England's fan charts, *International Journal of Forecasting*, to appear.

Table 1: Point Forecast Evaluation: MSFE Ratio and Rank

		Forecast horizon h											
		1	2	3	4	5	6	7	8	9	10	11	12
Canada	AR	1.00(3)	1.00(3)	1.00(2)	1.00(2)	1.00(2)	1.00(2)	1.00(2)	1.00(2)	1.00(2)	1.00(2)	1.00(2)	1.00(2)
	MSC	1.03(5)	1.04(5)	1.03(5)	1.06(5)	1.07(5)	1.07(5)	1.09(5)	1.10(5)	1.12(5)	1.14(5)	1.15(5)	1.17(5)
	SETAR	1.00(4)	0.98(1)	0.98(1)	0.96(1)	0.95(1)	0.95(1)	0.95(1)	0.97(1)	0.97(1)	0.97(1)	0.96(1)	0.96(1)
	MSIAH	0.98(1)	1.00(2)	1.01(3)	1.03(3)	1.02(4)	1.03(4)	1.03(4)	1.04(4)	1.06(4)	1.07(4)	1.08(4)	1.08(4)
	MSMH	0.99(2)	1.01(4)	1.03(4)	1.03(4)	1.01(3)	1.02(3)	1.02(3)	1.03(3)	1.05(3)	1.06(3)	1.07(3)	1.07(3)
Germany	AR	1.00(3)	1.00(1)	1.00(2)	1.00(2)	1.00(2)	1.00(3)	1.00(2)	1.00(2)	1.00(2)	1.00(2)	1.00(2)	1.00(2)
	MSC	1.01(4)	1.02(4)	1.01(4)	1.02(3)	1.02(5)	1.03(5)	1.02(4)	1.03(4)	1.03(4)	1.04(4)	1.04(4)	1.05(4)
	SETAR	1.02(5)	1.03(5)	1.03(5)	1.03(5)	1.01(3)	0.99(1)	1.00(1)	0.99(1)	0.99(1)	0.99(1)	1.00(1)	0.99(1)
	MSIAH	0.99(2)	1.01(3)	0.99(1)	0.99(1)	0.99(1)	1.00(2)	1.01(3)	1.02(3)	1.02(3)	1.02(3)	1.03(3)	1.03(3)
	MSMH	0.99(1)	1.01(2)	1.01(3)	1.02(4)	1.02(4)	1.02(4)	1.03(5)	1.03(5)	1.04(5)	1.05(5)	1.05(5)	1.06(5)
France	AR	1.00(1)	1.00(1)	1.00(1)	1.00(1)	1.00(2)	1.00(2)	1.00(2)	1.00(1)	1.00(1)	1.00(1)	1.00(1)	1.00(1)
	MSC	1.01(3)	1.00(2)	1.00(2)	1.01(3)	1.01(3)	1.01(3)	1.01(3)	1.01(3)	1.01(3)	1.01(3)	1.01(3)	1.01(3)
	SETAR	1.09(4)	1.14(5)	1.11(5)	1.09(5)	1.06(4)	1.04(5)	1.04(4)	1.05(5)	1.05(4)	1.05(4)	1.05(4)	1.04(4)
	MSIAH	1.11(5)	1.03(4)	1.08(4)	1.06(4)	1.07(5)	1.04(4)	1.08(5)	1.05(4)	1.09(5)	1.05(5)	1.07(5)	1.06(5)
	MSMH	1.00(2)	1.01(3)	1.00(3)	1.00(2)	1.00(1)	1.00(1)	1.00(1)	1.00(2)	1.01(2)	1.00(2)	1.00(2)	1.00(2)
Italy	AR	1.00(1)	1.00(2)	1.00(3)	1.00(4)	1.00(4)	1.00(4)	1.00(3)	1.00(4)	1.00(4)	1.00(4)	1.00(4)	1.00(4)
	MSC	1.00(3)	0.98(1)	0.96(1)	0.99(2)	0.99(2)	0.99(3)	1.00(4)	0.98(3)	0.97(2)	0.96(2)	0.94(2)	0.93(2)
	SETAR	1.03(5)	1.01(4)	0.99(2)	0.98(1)	0.98(1)	0.97(1)	0.97(1)	0.95(1)	0.95(1)	0.94(1)	0.93(1)	0.93(1)
	MSIAH	1.02(4)	1.02(5)	1.02(5)	1.02(5)	1.02(5)	1.02(5)	1.01(5)	1.01(5)	1.01(5)	1.01(5)	1.01(5)	1.01(5)
	MSMH	1.00(2)	1.00(3)	1.01(4)	1.00(3)	0.99(3)	0.98(2)	0.98(2)	0.98(2)	0.98(3)	0.98(3)	0.98(3)	0.98(3)
Japan	AR	1.00(2)	1.00(2)	1.00(1)	1.00(1)	1.00(1)	1.00(1)	1.00(1)	1.00(1)	1.00(1)	1.00(1)	1.00(1)	1.00(1)
	MSC	1.00(1)	1.00(1)	1.00(2)	1.00(2)	1.00(2)	1.00(2)	1.00(2)	1.00(2)	1.00(2)	1.00(2)	1.00(2)	1.00(2)
	SETAR	1.04(5)	1.03(5)	1.04(5)	1.04(5)	1.04(5)	1.04(5)	1.04(5)	1.04(5)	1.04(5)	1.03(5)	1.03(5)	1.03(5)
	MSIAH	1.03(4)	1.02(4)	1.03(4)	1.02(4)	1.03(4)	1.03(4)	1.03(4)	1.03(4)	1.01(3)	1.00(3)	1.01(3)	1.00(3)
	MSMH	1.00(3)	1.00(3)	1.02(3)	1.01(3)	1.01(3)	1.02(3)	1.02(3)	1.02(3)	1.02(4)	1.01(4)	1.02(4)	1.02(4)
UK	AR	1.00(3)	1.00(2)	1.00(2)	1.00(2)	1.00(2)	1.00(2)	1.00(2)	1.00(2)	1.00(2)	1.00(2)	1.00(2)	1.00(2)
	MSC	1.00(4)	1.00(1)	1.00(1)	1.00(1)	1.00(1)	1.00(1)	1.00(1)	1.00(1)	1.00(1)	1.00(1)	1.00(1)	1.00(1)
	SETAR	1.03(5)	1.03(5)	1.05(5)	1.05(5)	1.08(5)	1.07(5)	1.06(5)	1.05(5)	1.04(5)	1.04(5)	1.04(5)	1.04(5)
	MSIAH	1.00(2)	1.02(4)	1.02(4)	1.02(4)	1.02(4)	1.02(4)	1.02(4)	1.02(4)	1.02(4)	1.02(4)	1.02(4)	1.02(4)
	MSMH	0.99(1)	1.01(3)	1.01(3)	1.01(3)	1.01(3)	1.01(3)	1.01(3)	1.01(3)	1.01(3)	1.01(3)	1.02(3)	1.02(3)
US	AR	1.00(2)	1.00(2)	1.00(1)	1.00(1)	1.00(1)	1.00(1)	1.00(1)	1.00(1)	1.00(1)	1.00(1)	1.00(1)	1.00(1)
	MSC	1.07(5)	1.10(5)	1.12(5)	1.13(5)	1.15(5)	1.16(5)	1.17(5)	1.17(5)	1.19(5)	1.21(5)	1.22(5)	1.24(5)
	SETAR	1.03(4)	1.04(3)	1.04(3)	1.03(3)	1.04(3)	1.03(3)	1.03(3)	1.03(3)	1.03(3)	1.04(3)	1.05(3)	1.05(3)
	MSIAH	1.03(3)	1.06(4)	1.07(4)	1.08(4)	1.08(4)	1.08(4)	1.07(4)	1.07(4)	1.08(4)	1.07(4)	1.07(4)	1.08(4)
	MSMH	0.98(1)	1.00(1)	1.02(2)	1.02(2)	1.02(2)	1.02(2)	1.02(2)	1.02(2)	1.02(2)	1.02(2)	1.03(2)	1.03(2)

Notes: Table entries are the ratio of the MSFE for BP, SETAR, MSIAH, and MSMH models to those of the benchmark AR model. Numbers in parentheses indicate the assigned rank, where 1 corresponds to the model with smallest MSFE, 2 to the second smallest, etc.

Table 2: Point Forecast Evaluation: MSFE Average Rank

	CA	DE	FR	JP	IT	UK	US	Average
$\underline{h = 1 - 6}$								
AR	2.33	2.17	1.33	3.00	1.33	2.17	1.33	1.95
MSC	5.00	4.17	2.67	2.00	1.67	1.50	5.00	3.14
SETAR	1.50	4.00	4.67	2.33	5.00	5.00	3.17	3.67
MSIAH	2.83	1.67	4.33	4.83	4.00	3.67	3.83	3.60
MSMH	3.33	3.00	2.00	2.83	3.00	2.67	1.67	2.64
$\underline{h = 7 - 12}$								
AR	2.00	2.00	1.17	3.83	1.00	2.00	1.00	1.86
MSC	5.00	4.00	3.00	2.50	2.00	1.00	5.00	3.21
SETAR	1.00	1.00	4.17	1.00	5.00	5.00	3.00	2.88
MSIAH	4.00	3.00	4.83	5.00	3.33	4.00	4.00	4.02
MSMH	3.00	5.00	1.83	2.67	3.67	3.00	2.00	3.02
$\underline{h = 1 - 12}$								
AR	2.17	2.08	1.25	3.42	1.17	2.08	1.17	1.90
MSC	5.00	4.08	2.83	2.25	1.83	1.25	5.00	3.18
SETAR	1.25	2.50	4.42	1.67	5.00	5.00	3.08	3.27
MSIAH	3.42	2.33	4.58	4.92	3.67	3.83	3.92	3.81
MSMH	3.17	4.00	1.92	2.75	3.33	2.83	1.83	2.83

Notes: Table entries indicate the average rank assigned to a given model for a certain time series over the specified forecast horizons according to the MSFE. For further details see Table 1.

Table 3: Point Forecast Evaluation: Pairwise DM Test with Quadratic Loss

	AR	BP	SETAR	MSIAH	MSMH	Total
$h = 1-6$						
AR	0(0)	30(2)	32(4)	34(7)	32(0)	128(13)
MSC	12(0)	0(0)	25(3)	23(5)	18(0)	78(8)
SETAR	10(0)	17(0)	0(0)	18(0)	11(0)	56(0)
MSIAH	8(0)	19(0)	24(1)	0(0)	8(0)	59(1)
MSMH	10(0)	24(0)	31(3)	34(5)	0(0)	99(8)
$h = 7-12$						
AR	0(0)	31(0)	24(0)	42(7)	35(0)	132(7)
MSC	11(1)	0(0)	18(0)	24(6)	22(0)	75(7)
SETAR	18(0)	24(0)	0(0)	29(2)	18(0)	89(2)
MSIAH	0(0)	18(0)	13(0)	0(0)	10(0)	41(0)
MSMH	7(0)	20(0)	24(0)	32(5)	0(0)	83(5)
$h = 1-12$						
AR	0(0)	61(2)	56(4)	76(14)	67(0)	260(20)
MSC	23(1)	0(0)	43(3)	47(11)	40(0)	153(15)
SETAR	28(0)	41(0)	0(0)	47(2)	29(0)	145(2)
MSIAH	8(0)	37(0)	37(1)	0(0)	18(0)	100(1)
MSMH	17(0)	44(0)	55(3)	66(10)	0(0)	182(13)

	CA	DE	FR	JP	IT	UK	US	Total
$h = 1-6$								
AR	16(0)	17(0)	22(6)	12(0)	22(1)	17(2)	22(4)	128(13)
MSC	0(0)	5(0)	14(3)	18(0)	20(1)	21(4)	0(0)	78(8)
SETAR	21(0)	6(0)	2(0)	16(0)	0(0)	0(0)	11(0)	56(0)
MSIAH	13(0)	20(0)	4(1)	1(0)	6(0)	8(0)	7(0)	59(1)
MSMH	10(0)	12(0)	18(6)	13(0)	12(1)	14(1)	20(0)	99(8)
$h = 7-12$								
AR	18(0)	18(0)	23(3)	7(0)	24(0)	18(4)	24(0)	132(7)
MSC	0(0)	6(0)	12(0)	15(0)	18(0)	24(7)	0(0)	75(7)
SETAR	24(2)	24(0)	5(0)	24(0)	0(0)	0(0)	12(0)	89(2)
MSIAH	6(0)	12(0)	1(0)	0(0)	10(0)	6(0)	6(0)	41(0)
MSMH	12(0)	0(0)	19(2)	14(0)	8(0)	12(3)	18(0)	83(5)
$h = 1-12$								
AR	34(0)	35(0)	45(9)	19(0)	46(1)	35(6)	46(4)	260(20)
MSC	0(0)	11(0)	26(3)	33(0)	38(1)	45(11)	0(0)	153(15)
SETAR	45(2)	30(0)	7(0)	40(0)	0(0)	0(0)	23(0)	145(2)
MSIAH	19(0)	32(0)	5(1)	1(0)	16(0)	14(0)	13(0)	100(1)
MSMH	22(0)	12(0)	37(8)	27(0)	20(1)	26(4)	38(0)	182(13)

Notes: Table entries in the upper panel are the number of times that row model i has a smaller MSFE value than column model j, summarized across all countries. The number of rejections of equal forecast accuracy in terms of MSFE of row model i and column model j by the Diebold-Mariano test at the 5% significance level are shown in parentheses. Table entries in the lower panel indicate the number of times that row model i produced a smaller MSFE value of than all other models for a particular country. Numbers in parentheses are rejection rates of Diebold-Mariano test for equal forecast accuracy of row model i and all other models for a particular time series at the 5% significance level.

Table 4: Point Forecast Evaluation: Pairwise Forecast Encompassing Test

	AR	BP	SETAR	MSIAH	MSMH	Total
$\underline{h = 1-6}$						
AR	–	8	5	3	3	19
MSC	18	–	20	20	16	74
SETAR	28	24	–	16	23	91
MSIAH	28	21	12	–	26	87
MSMH	4	11	9	5	–	29
$\underline{h = 7-12}$						
AR	–	11	6	0	0	17
MSC	9	–	10	4	8	31
SETAR	8	2	–	1	7	18
MSIAH	19	12	12	–	20	63
MSMH	6	8	12	1	–	27
$\underline{h = 1-12}$						
AR	–	19	11	3	3	36
MSC	27	–	30	24	24	105
SETAR	36	26	–	17	30	109
MSIAH	47	33	24	–	46	150
MSMH	10	19	21	6	–	56

	CA	DE	FR	JP	IT	UK	US	Total
$\underline{h = 1-6}$								
AR	4	5	0	8	0	2	0	19
MSC	14	13	1	18	0	4	24	74
SETAR	5	13	20	7	14	20	12	91
MSIAH	6	7	18	21	6	14	15	87
MSMH	11	3	0	9	0	6	0	29
$\underline{h = 7-12}$								
AR	0	0	0	12	0	5	0	17
MSC	2	3	0	6	0	0	20	31
SETAR	0	0	3	0	0	6	9	18
MSIAH	8	1	12	18	0	17	7	63
MSMH	8	4	0	12	1	2	0	27
$\underline{h = 1-12}$								
AR	4	5	0	20	0	7	0	36
MSC	16	16	1	24	0	4	44	105
SETAR	5	13	23	7	14	26	21	109
MSIAH	14	8	30	39	6	31	22	150
MSMH	19	7	0	21	1	8	0	56

Notes: Table entries in the upper panel are the number of rejections of the null hypothesis that row model i forecast encompasses column model j by the pairwise encompassing test at the 5% significance level, summarized across countries. Entries in the bottom panel are the number of rejections of the null hypothesis that row model i forecast encompasses all other models in a pairwise contest at the 5% significance level.

Table 5: Point Forecast Evaluation: Multiple Forecast Encompassing Test

	CA	DE	FR	JP	IT	UK	US	Total
$\underline{h = 1 - 6}$								
AR	2	1	0	0	1	0	0	4
MSC	5	1	6	0	1	0	2	15
SETAR	2	1	3	1	3	0	0	10
MSIAH	0	0	2	1	1	1	0	5
MSMH	3	0	1	0	0	0	0	4
$\underline{h = 7 - 12}$								
AR	0	0	0	0	0	1	0	1
MSC	0	0	6	0	0	1	0	7
SETAR	0	1	0	0	0	2	0	3
MSIAH	0	1	0	0	0	3	0	4
MSMH	0	1	0	0	0	2	0	3
$\underline{h = 1 - 12}$								
AR	2	1	0	0	1	1	0	5
MSC	5	1	12	0	1	1	2	22
SETAR	2	2	3	1	3	2	0	13
MSIAH	0	1	2	1	1	4	0	9
MSMH	3	1	1	0	0	2	0	7

Notes: Table entries are the number of rejections of the null hypothesis that row model i forecast encompasses all other models simultaneously at the 5% significance level, according to the multiple forecast encompassing test.

Table 6: Interval Forecast Evaluation: Tests of Correct Unconditional Coverage, Independence and Correct Conditional Coverage

	Correct Unconditional Coverage								Independence								Correct Conditional Coverage							
	CA	DE	FR	JP	IT	UK	US	Total	CA	DE	FR	JP	IT	UK	US	Total	CA	DE	FR	JP	IT	UK	US	Total
Nominal Coverage 50%																								
AR	1	4	3	5	5	10	7	35	1	2	0	1	0	2	2	8	3	3	2	5	6	10	10	39
BP	5	1	4	0	5	7	7	29	2	1	0	1	0	0	2	6	7	2	2	1	6	5	7	30
SETAR	1	2	2	7	4	5	4	25	1	1	0	1	0	3	2	8	2	2	0	7	3	5	6	25
MSIAH	0	4	2	1	0	2	4	13	1	3	0	1	0	1	6	12	1	6	2	1	0	3	9	22
MSMH	2	1	2	4	0	2	3	14	1	2	0	1	0	1	4	9	1	2	2	4	0	1	10	20
Nominal Coverage 75%																								
AR	0	2	5	1	6	7	6	27	4	5	1	2	0	3	3	18	3	5	4	2	4	4	6	28
BP	0	0	5	0	6	5	5	21	2	2	1	1	0	3	7	16	1	2	5	1	4	4	7	24
SETAR	2	1	3	4	5	3	4	22	3	0	0	3	0	4	1	11	0	0	4	3	3	4	2	16
MSIAH	2	2	3	2	2	1	5	17	0	1	0	1	0	2	3	7	0	3	3	1	0	1	5	13
MSMH	0	1	4	4	2	1	4	16	2	4	0	2	0	1	1	10	0	3	2	4	0	1	3	13
Nominal Coverage 90%																								
AR	0	0	5	0	2	0	3	10	1	1	0	2	0	1	1	6	2	1	1	1	1	1	3	10
BP	0	0	5	5	2	0	1	13	2	3	0	2	0	1	2	10	3	5	2	6	1	1	3	21
SETAR	0	0	4	0	3	0	0	7	0	1	1	1	0	1	0	4	0	1	1	1	1	1	0	5
MSIAH	0	1	1	2	2	0	0	6	1	0	1	1	0	1	0	4	1	0	0	1	0	1	0	3
MSMH	0	0	1	0	0	0	0	1	1	1	0	2	0	1	0	5	1	1	0	1	0	1	0	4

Notes: Table entries are the number of rejections of the tests of correct unconditional coverage (left panel), independence (middle), and correct conditional coverage (right), summarized for all $h = 1, 2, \ldots, 12$ using the 5% significance level.

Table 7: Density Forecast Evaluation: Kolmogorov-Smirnov test for IID Uniformity of PIT

									h	
	CA	DE	FR	JP	IT	UK	US	Total	1-6	7-12
AR	1	0	6	9	5	1	1	23	18	5
MSC	0	0	0	0	5	0	0	5	5	0
SETAR	1	0	11	5	5	0	0	22	13	9
MSIAH	3	1	6	11	6	0	3	30	23	7
MSMH	0	0	5	5	8	0	0	18	13	5

Notes: Table entries are the number of rejections of the null hypothesis that the probability integral transform is uniformly distributed by the Kolmogorov-Smirnov test at the 5% significance level, summarized across forecast horizons (left) or across countries (right).

Table 8: Density Forecast Evaluation: Doornik-Hansen test for normality of inverse normal CDF of PIT

									h	
	CA	DE	FR	JP	IT	UK	US	Total	1-6	7-12
AR	1	4	0	3	4	10	4	26	15	11
MSC	6	5	0	1	4	8	11	35	14	21
SETAR	7	1	0	2	3	11	0	24	12	12
MSIAH	7	1	0	1	2	4	3	18	9	9
MSMH	3	1	0	1	1	3	1	10	4	6

Notes: Table entries are the number of rejections of the null hypothesis that the inverse normal cumulative density function transformation of the probability integral transform is normally distributed by the Doornik-Hansen test at the 5% significance level, summarized across forecast horizons (left) or across countries (right).

Table 9: Density Forecast Evaluation: Incidence of PIT being equal to 0 or 1

									h	
	CA	DE	FR	JP	IT	UK	US	Total	1-6	7-12
AR	2	0	0	2	2	3	0	9	9	0
MSC	2	22	0	11	2	3	25	65	30	35
SETAR	0	1	1	2	2	6	1	13	13	0
MSIAH	0	0	0	0	0	1	0	1	1	0
MSMH	0	0	0	0	0	1	0	1	1	0

Notes: Table entries are the number of times the actual values of the $h-$period growth rates fall outside the empirical density forecasts, summarized across forecast horizons (left) or across countries (right).

Table 10: Density Forecast Evaluation: Ljung-Box test for first-order autocorrelation in PIT

									h	
	CA	DE	FR	JP	IT	UK	US	Total	1-6	7-12
$\underline{k=1}$										
AR	0	0	1	4	1	1	0	7	7	0
MSC	0	0	1	2	1	1	4	9	7	2
SETAR	0	1	0	4	1	0	0	6	6	0
MSIAH	2	1	1	4	0	0	1	9	9	0
MSMH	1	2	1	4	0	0	0	8	8	0
$\underline{k=2}$										
AR	1	2	1	2	1	7	6	20	18	2
MSC	1	5	1	1	1	4	9	22	15	7
SETAR	5	0	0	2	0	8	1	16	13	3
MSIAH	4	1	0	1	0	1	6	13	8	5
MSMH	1	1	0	2	0	1	4	9	6	3
$\underline{k=3}$										
AR	1	1	1	3	1	1	1	9	9	0
MSC	1	2	0	0	1	1	7	12	5	7
SETAR	3	2	1	3	0	1	0	10	9	1
MSIAH	4	1	2	3	0	1	0	11	10	1
MSMH	1	2	1	4	0	1	0	9	9	0
$\underline{k=4}$										
AR	3	2	1	3	0	8	5	22	17	5
MSC	4	5	1	1	0	5	11	27	15	12
SETAR	6	1	1	2	0	7	2	19	15	4
MSIAH	6	1	0	2	0	2	3	14	10	4
MSMH	3	2	0	2	0	2	3	12	9	3

Notes: Table entries are the number of rejections of the null hypothesis of no first-order autocorrelation in power transformations of the demeaned PIT $(z_{t|t-h} - \overline{z}_{t|t-h})^k$ with $k = 1, 2, 3, 4$ by the Ljung-Box test at the 5% significance level, summarized across forecast horizons (left) or across countries (right).