

MIT Open Access Articles

Nonlinear and nonparametric modeling approaches for probabilistic forecasting of the US gross national product

The MIT Faculty has made this article openly available. **Please share** how this access benefits you. Your story matters.

Citation: Arora, Siddharth, Max A. Little, and Patrick E. McSharry. "Nonlinear and nonparametric modeling approaches for probabilistic forecasting of the US gross national product." *Studies in Nonlinear Dynamics and Econometrics* 17, no. 4 (January 1, 2013).

As Published: <http://dx.doi.org/10.1515/snde-2012-0029>

Persistent URL: <http://hdl.handle.net/1721.1/87639>

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

Terms of Use: Article is made available in accordance with the publisher's policy and may be subject to US copyright law. Please refer to the publisher's site for terms of use.



Siddharth Arora*, Max A. Little and Patrick E. McSharry

Nonlinear and nonparametric modeling approaches for probabilistic forecasting of the US gross national product

Abstract: Numerous time series models are available for forecasting economic output. Autoregressive models were initially applied to US gross national product (GNP), and have been extended to nonlinear structures, such as the self-exciting threshold autoregressive (SETAR) and Markov-switching autoregressive (MS-AR) models. We show that while these nonlinear models fit the in-sample data better than linear models, the out-of-sample forecast performance of extremely simple methods, such as the unconditional mean is competitive compared with both previously published linear and nonlinear models for GNP time series. Motivated by Occam's razor and the forecasting competitiveness of the unconditional mean, we seek parsimonious models which are based on simple assumptions, incorporate few parameters, and can generate accurate forecasts. We propose nonlinear and nonparametric models based on nearest neighbor and kernel regression by forming a state-space of time delayed GNP observations. The rationale of the proposed methodology lies in identification of local regions of state-space known as nearest neighbor sub-spaces, whereby we utilize future trajectories of the neighboring states and weight them using a double kernel estimator to generate density forecasts. The models proposed in this work incorporate only one or two parameters, and the model estimation framework relies on optimizing the performance of in-sample density forecasts. The proposed modeling approach does not require prior assumptions regarding the form of the error distribution, and can provide transition between regimes of growth and contraction. We compare the forecasts from proposed models with classical time series models for GNP and a range of benchmarks, and validate our results on two post-war GNP time series using different performance scores, such as the root mean squared error (RMSE), mean absolute error (MAE), and the continuous ranked probability score (CRPS). We find that proposed models consistently outperformed previously published models for point and density forecasts of GNP at varying horizons.

Keywords: US GNP; forecasting; autoregressive models; SETAR; MS-AR.

*Corresponding author: **Siddharth Arora**, Oxford Centre for Industrial and Applied Mathematics, Mathematical Institute, University of Oxford, St Giles, Oxford, OX1 3LB, UK, e-mail: siddharth.arora@maths.ox.ac.uk

Max A. Little: Media Lab, Massachusetts Institute of Technology, Ames Street, Cambridge, MA 01239, USA

Patrick E. McSharry: Smith School of Enterprise and the Environment, Hayes House, 75 George Street, Oxford, OX1 2BQ, UK

1 Introduction

The GNP of a country reflects the total value of all goods and services produced by its nationals. It is a macro-economic measure that provides an overall estimate of the economic health of a country. Accurate estimates of future trends in GNP are important for informed decision-making and government policy-making. An accurate GNP forecast facilitates the smooth functioning of the economy through timely changes in policy, providing improved decision-making and leading to greater stability of financial markets. The dynamics underlying GNP time series is complex, and we only have a few number of quarterly post-war observations, making the task of generating reliable economic forecasts extremely challenging (Orrell and McSharry, 2009).

A diverse range of time series models are available for producing GNP forecasts. Nonlinear regime switching models have become extremely popular in contemporary econometrics (Teräsvirta and Anderson 1992; Potter 1995; Krolzig 2001; Clements et al. 2003; Clements, Franses, and Swanson 2004), providing alternatives to Box and Jenkins (1970) class of linear models. Regime switching models like the self-exciting threshold autoregressive (SETAR) models (Tiao and Tsay 1994; Potter 1995), and Markov-switching autoregressive (MS-AR) models (Hamilton 1989) have been employed for modeling US GNP. These models are structured to be able to characterize regimes of growth and recession in a business cycle. There remains no clear consensus

on whether nonlinear models are better than their linear counterparts from the perspective of out-of-sample forecast performance, even when linearity is rejected statistically (Diebold and Nason 1990; De Gooijer and Kumar 1992). Similar results were reported by Swanson and White (1995, 1997) who compared the out-of-sample point forecast accuracy of multivariate adaptive linear and nonlinear models using different macroeconomic variables, and reported that linear models often outperformed nonlinear models in simulated real time. Stock and Watson (1999) found linear models to be more accurate than nonlinear models for forecasting a range of different macroeconomic time series, while Dueker and Neely (2007) reported that a simple benchmark generated more accurate point forecasts compared to a nonlinear regime switching model.

The lack of concrete evidence in support of nonlinear regime switching models over linear models as superior forecasting tools has led to considerable amounts of research into the structural conditions and performance scores under which one class of models may fare better than the other. In this context, nonlinearity is defined based on the presence of different regimes in the data generating process. Note that the GNP follows an asymmetric cyclical process, giving rise to different regimes (growth and recession), with growth periods being much longer than recessionary periods, see Teräsvirta and Anderson (1992). Teräsvirta (1995) found evidence of nonlinearity in the long annual GNP, and used the exponential smooth transition autoregressive (ESTAR) model to characterize business cycles in the time series. Teräsvirta and Anderson (1992) argued that regime switching models may perform better than the linear models, if there are a sufficient number of observations in each separate regime of the time series. This implies that regime switching models, like SETAR and MS-AR, are in general better able to characterize the regime switching events underlying the time series. As also pointed out by Clements and Krolzig (1998), the tendency of an economy to return quickly to a mean level from a negative shock/recession, is a prominent nonlinear switching behavior. Hence, a time series characterized by a substantial number of observations in the recessionary regime would presumably be better modeled by nonlinear regime switching models, compared to the classical linear models. To translate this property of regime switching models into enhanced forecast accuracy requires that the out-of-sample and in-sample data exhibit some minimum degree of switching nonlinearity. Hansen (1992, 1996b) found weak evidence of rejecting linear models in favour of nonlinear models for quarterly GNP time series. Potter (1995) compared the error variance and forecast accuracy of nonlinear and linear models and reported nonlinearity in US GNP time series, while Brock and Sayers (1988) found little evidence of nonlinearity in real US GNP. Using GNP time series, Clements et al. (2003) focused on measuring the degree of nonlinearity (based on a significance level at which test of null of linearity was rejected for heteroscedasticity-robust and standard Hansen tests) under which nonlinear models may be superior to linear models. Marcellino, Stock, and Watson (2003) compared forecasting accuracy of univariate and multivariate models using macroeconomic data, and reported that pooled univariate AR models were competitive with multivariate models. Öcal (2000) employed threshold autoregressive models for UK macroeconomic time series and reported that they were not considerably superior to their linear counterparts. Morley and Piger (2004) argued that reproducing the features of a business cycle requires a nonlinear approach. Engel, Haugh, and Pagan (2005) demonstrated that regime switching models explain the shape and variability of a business cycle better than linear models, but do so at the expense of a very strong re-bound effect, which makes expansions continue for longer durations than they do in reality. Though parametric regime switching models (SETAR and MS-AR) have shown to be useful for reproducing business cycles and analyzing time series, we aim to investigate if these models can generate more accurate point and density forecasts on the out-of-sample data compared to simpler nonparametric modeling approaches. A review on macroeconomic forecasting has been provided by Stekler (2007).

In this study, we propose a new class of nonparametric, nonlinear models based on nearest neighbor and kernel regression for US GNP time series, whereby nonlinearity is defined based on deviation from the classical Gauss-Markov (G-M) assumptions (linearity, homogeneity and independence). The proposed nonlinear models, like the classical nonlinear regime switching models, deviate from the G-M assumptions of linearity, however, the condition of homogeneity (which can be tested using a heteroscedasticity-robust test) may or may not hold true for our nonlinear models. The proposed models are not only much simpler by comparison with the classical models, but also do not rely on assumptions regarding the presence of separate regimes in the data generating process for the model to generate accurate forecasts.

In this paper, we make a systematic comparison between linear and nonlinear models, and show that even the best known forecasting models for US GNP, whether linear or nonlinear, do not outperform a naïve forecast based on the unconditional mean of the time series at multiple forecast horizons, for both point and density forecasts. This result has not been shown before on US GNP time series to the best of our knowledge. For the choice of benchmarks, most of the classical and recent work on GNP and GDP forecasting employ either the random walk model or a first order autoregressive model, but none compare their forecasts with the unconditional forecast, see Potter (1995), Clements and Krolzig (1998), Ang, Piazzesi, and Wei (2006), Zheng and Rossiter (2006) and Barhoumi et al. (2008). Similar results were reported by Dueker and Neely (2007), who found the Markov switching model does not outperform a naïve, constant-return benchmark in predicting excess foreign exchange returns, using the mean squared error and mean absolute error criterion. Regarding combining forecasts, Clemen and Winkler (1986) based on the findings of Makridakis and Winkler (1983), showed that when combining GNP forecasts from different models, a simple average of forecasts from different models leads to a more accurate final forecast, when compared with some of the more complex combination strategies. Galbraith (2003) compares the forecast accuracy of an AR model with the unconditional mean using time series for real GDP and inflation. A majority of previous literature on the comparison between linear and nonlinear models is focused on evaluating point forecast accuracy. Using data for inflation and GDP, Kapetanios, Labhard, and Price (2007) showed that point forecasts from MS-AR and SETAR are slightly more accurate than AR. In recent studies, Buchmann and Hubrich (Unpublished manuscript; https://editorialexpress.com/cgi-bin/conference/download.cgi?db_name=CEF2011&paper_id=360) compare the out-of-sample density performance of AR, vector AR, and MS-AR using the GDP time series, for horizons ranging from one to four quarters ahead. They use the continuous ranked probability score (CRPS) to evaluate density forecasts, and find that MS-AR offers improvement in forecast accuracy over linear modeling approaches. Using GNP time series, we find that density forecasts from MS-AR are competitive with both AR and SETAR, however, none of these models outperform an unconditional forecast. For any meaningful policy-making activity, it is crucial that the uncertainty associated with the forecasts is also taken into consideration. We extend the scope of our work to quantify the level of uncertainty in GNP forecasts, and evaluate density forecasts of both classical and the newly proposed models in this study.

The estimation of classical models (AR, SETAR and MS-AR) and proposed models uses the GNP training set, while model evaluation is undertaken using the testing set. The time index/quarter that divides the GNP time series into non-overlapping training and testing set is known as the forecast origin (last observation in the training set). The testing period was completely reserved for model evaluation, and there was no passing of information from the testing set to the training set. The methods used to quantify the degree of deviation between the actual observations and forecasts are known as performance scores. In this paper, we employ RMSE and MAE to evaluate point forecasts and CRPS to evaluate density forecasts. Note that point forecast corresponds to a single most likely forecast estimate, while density forecasts provide a continuous probability density function that encompasses point forecast and the uncertainty associated with it.

We propose a set of simple nonlinear models, by forming a state-space from time delayed observations. The novelty of our modeling approach lies in the identification of local regions of state-space known as nearest neighbor sub-spaces, and using weighted random sampling of trajectories of neighboring states (based on a double kernel estimator) to issue point and density forecasts. We show that our model outperforms previously published linear and nonlinear models proposed for forecasting US GNP, including simple benchmarks. Specifically, the contributions of this work lies in, a) systematic comparison of the out-of-sample point and density forecast performance of classical linear and nonlinear models employed previously for GNP along with simple benchmarks, b) proposing a new class of parsimonious, nonlinear and nonparametric models based on nearest neighbor and kernel regression, which can be easily generalized to different forecasting applications, and, c) generating forecasts for the US GNP using the proposed models, which convincingly outperforms some of the most widely published parametric models. The proposed model incorporates less than half of the total parameters used in classical models for GNP, and can provide transition between regimes. To validate our results, we compare the forecast performance of linear and nonlinear models on two different GNP time series belonging to different periods, across a range of forecast horizons. The GNP

time series comprises seasonally adjusted quarterly prices of real US GNP (in billions of dollars). As analyzed in Potter (1995), the shorter time series measures GNP at 1982 dollar prices dating from 1947Q1 to 1990Q4. The longer, more recent time series contains US GNP recordings from 1947Q1 to 2008Q3 at 2000 prices. The first GNP series is downloaded from the online data archive of the Journal of Applied Econometrics,¹ while the second time series was obtained from the Federal Reserve Economic Data II (FRED II) affiliated with the Federal Reserve Bank of St. Louis.² Our results were found to be consistent for both the GNP time series; however, for the purpose of demonstration we present results corresponding to the longer time series in this paper.

The paper is arranged as follows, Section 2 estimates the parametric models for US GNP. Section 3 proposes novel one and two parameter nonlinear and nonparametric models for GNP forecasting. Section 4 provides the results and observations pertaining to the out-of-sample point and density forecast comparison of different models, and Section 5 offers conclusions.

2 Parametric models

This section gives a brief overview of three parametric models that have previously been employed for forecasting GNP. Specifically, we present the AR, SETAR and MS-AR models, and provide their parameter estimates.

2.1 Autoregressive (AR) Models

An autoregressive model of order p is denoted as $AR(p)$ and is represented using the following form:

$$x_t = \alpha_0 + \sum_{i=1}^p \alpha_i x_{t-i} + \epsilon_t, \quad \epsilon_t \sim IID(0, \sigma^2) \quad (1)$$

where x_t is an observation in the time series at time t , α_0 is a constant, α_i denotes AR model parameters for $i=1, 2, \dots, p$, and ϵ_t is an independent and identically distributed (IID) process with mean zero and variance σ^2 . The AR model order p can be chosen based on the Akaike information criterion (AIC), as proposed by Akaike (1973).

We estimate the model parameters using the quarterly growth rate (first differences) of logarithm of the GNP, i.e., $x_t = 100 \times \log(y_t/y_{t-1})$, where y_t is the actual GNP value. This transformation is based on the assumption that the log of GNP quarterly growth rate follows a stationary process. The estimates of AR model parameters [obtained using ordinary least squares (OLS)] for training period 1947Q2–1996Q4 are presented in Table 1.

2.2 Self-exciting threshold autoregressive (SETAR) models

SETAR models have been employed previously for US GNP by Tiao and Tsay (1994) and Potter (1995). The rationale for SETAR modeling of economic time series has been to apply separate models for periods of growth and contraction. This makes SETAR a linear piecewise AR process in the state-space.

A SETAR model composed of N_r regimes, for a time series x_t , is denoted as $SETAR(N_r; p_1, \dots, p_{N_r})$, and can be represented as:-

$$x_t = \alpha_{0j} + \sum_{i=1}^{p_j} \alpha_{ij} x_{t-i} + \epsilon_{tj}, \quad \epsilon_{tj} \sim IID(0, \sigma_j^2) \quad (2)$$

¹ <http://econ.queensu.ca/jae/1995-v10.2/potter/potter.data>.

² <http://research.stlouisfed.org/fred2/series/GNPC96>.

Table 1 Parameter estimates for AR model and corresponding standard errors (in parenthesis).

Parameter	Estimate (standard error)
α_0	0.6499 (0.1132)
α_1	0.3120 (0.0719)
α_2	0.1284 (0.0751)
α_3	-0.0832 (0.0751)
α_4	-0.1259 (0.0716)
σ	0.9624
Total parameters	5

where α_{ij} are the autoregressive coefficients for a given regime index j that obeys $r_{j-1} \leq x_{t-d} < r_j$, while r_j for $j=1, 2, \dots, N_r$ are the thresholds that divide the time series into different regimes, p_j is the model order for regime j , and ϵ_{ij} is an IID process with mean zero and variance σ_j^2 . Note that the model order can be different across different regimes. The values of thresholds are chosen from the in-sample time series, while the choice of AR model coefficients depends on the thresholds r_j and the delay length d . The role of the threshold variable is to force the model to switch between different regimes.

The model parameters are estimated by minimizing the overall residual sum of squares using the in-sample data. The AR model order within a regime is selected using an information criterion as employed for SETAR by Clements and Krolzig (1998). We estimate a bi-regime SETAR model ($N_r=2$) using contiguous time lags for the value of delay length d being two. We find that the optimum estimate of threshold r for shorter time series is -0.0580 [as previously reported by Clements and Krolzig (1998)], while for longer time series we obtain $r=0.2808$. For a given threshold and model order, the AR coefficients are obtained by running a least squares regression separately on each regime. The results for parameter estimates are presented in Table 2.

Tiao and Tsay (1994) adopted two and four regime SETAR models for the US GNP, and reported that the performance of SETAR models compared to AR models improves if forecasts are evaluated separately when conditioned on the regime. Note that if most of the observations in a time series belong to a single regime, then fitting a linear AR model across the whole time series is similar to fitting a SETAR model across regimes. Hansen (1996b) compared the two classes of models and found no substantial evidence for rejecting the AR model over SETAR. Pesaran and Potter (1997) reported that AR models perform better than nonlinear models in forecasting the conditional mean, but not the conditional variance.

Table 2 Parameter estimates for SETAR(2;1,1) model.

Parameter	Estimate (standard error)
Lower regime	
α_{0L}	0.4507 (0.1803)
α_{1L}	0.2806 (0.1506)
σ_L	1.2359
N_L	52
Upper regime	
α_{0U}	0.6166 (0.1081)
α_{1U}	0.3159 (0.0769)
σ_U	0.8952
N_U	145
r	0.2808
d	2
Total parameters	8

Clements and Smith (1999) argued that the forecast performance of SETAR relative to the AR model is highlighted when evaluation is conditioned upon the regime, reflecting the ability of SETAR to perform well in certain regimes, but not necessarily across the whole time series. Clements and Krolzig (1998) compared the AR, SETAR and MS-AR models on two different US GNP time series at forecast horizons ranging from 1 to 16 quarters ahead. They concluded that regime switching models do have an edge over linear models in characterizing the underlying shifts in regime, but are not superior to linear models based on the out-of-sample forecast accuracy.

2.3 Markov-switching autoregressive (MS-AR) models

Hamilton (1989) proposed a two-regime Markov-switching mean model for the US GNP. We denote this model as MS(2)-AR(p), defined as:-

$$x_t = \mu(s_t) + \sum_{i=1}^p \alpha_i (x_{t-i} - \mu(s_{t-i})) + \epsilon_t, \quad \epsilon_t \sim IID(0, \sigma^2) \quad (3)$$

where x_t is an observation in the time series, α_i are the AR coefficients for $i=1, 2, \dots, p$, s_t is a regime variable of the system such that, $s_t=1$ corresponds to growth and $s_t=2$ corresponds to recession. Note that $\mu(s_t)$ switches between the two regimes, such that $\mu(s_t)$ is positive if $s_t=1$, and negative otherwise. The transition between the two regimes depends on a Markov chain process. We estimate the model parameters by maximizing the likelihood of observations using expectation maximization. We find the estimates for Hamilton's MS(2)-AR(p) model for the GNP time series, as shown in Table 3, using the MS VAR class for Ox (Krolzig (1998)). The optimum AR model order p for GNP time series (using only in-sample observations) is selected using AIC.

It should be noted that although the MS-AR model does well in identifying the changes in regimes, the superiority of MS-AR over AR remains doubtful when the out-of-sample forecast performance is assessed. Based on standard likelihood ratio tests, Hansen (1992, 1996a) reported no substantial evidence to reject an

Table 3 Parameter estimates for MS(2) – AR(5) model.

Parameter	Estimate (standard error)
Mean	
μ_L	-1.3208 (0.3888)
μ_H	0.9729 (0.1449)
AR(5) coeff.	
α_1	0.3899 (0.0732)
α_2	0.1998 (0.0742)
α_3	-0.1753 (0.0687)
α_4	-0.1550 (0.0685)
α_5	-0.0231 (0.0677)
σ	0.81226
Transitional probability	
p_{LL}	0.2050
p_{HH}	0.9547
Unconditional probability	
p_L	0.0539
p_H	0.9461
Duration	
d_L	1.26
d_H	22.09
Total parameters	10

AR model over MS-AR. Dueker and Neely (2007) evaluated the out-of-sample forecast performance of the MS-AR model for exchange rate returns, and found that MS-AR does not perform better than a naïve forecast. Similar results were reported earlier by Goodwin (1993), who applied Hamilton's model to eight different economies and found that the MS-AR model only provides a slight improvement in the forecast ability compared to the linear models. Bessec and Bouabdallah (2005) argue that the out-of-sample forecasting accuracy of MS-AR models is low because they misclassify the future regimes. Interestingly, it was shown by Loungani (2001), that forecasters do not fare well when it comes to forecasting recession, as only two recessions were predicted a year in advance, out of the 60 recessions that were observed around the world during the 1990s. A detailed review on economic and financial time series forecasting using nonlinear models has been provided by Clements, Franses, and Swanson (2004).

3 Nonparametric models

In this section, we propose three nonparametric models that can effectively capture any temporal dependence structure in the time series, as measured by, for example, the similarity of previous state vector to the current state vector, in a multi-dimensional state-space. The motivation for investigating this modeling approach is that it can cope up with nonlinearities in the data generating process (DGP), requires the estimation of a few parameters, and can provide transition between regimes. The proposed model design is not based on assumptions regarding the presence of regimes in the DGP, as opposed to regime switching models.

We propose nonparametric models, the rationale of which lies in utilizing similar patterns in GNP to generate forecasts using a data-based modeling approach. Conceptually, the proposed models attempt to generate forecast for a given state, based on the trajectories of similar past states. Though parametric models, like AR, SETAR and MS-AR, have been useful for summarizing the characteristics of GNP based on the interpretation of model parameters (Clements and Krolzig 1998; Teräsvirta and Anderson 1992), their forecast performance over simpler benchmarks has been less than satisfactory overall. Given that we have few GNP training observations, a plausible reason for the poor out-of-sample forecasting performance of parametric models is due to these models over-fitting the in-sample data. Note that MS-AR incorporates ten parameters. This motivates us to investigate if parsimonious models, including one or two parameters, could be employed over more complicated models for out-of-sample forecasting. Furthermore, the regime switching models make strong prior assumptions about the functional form for the DGP, whereby the model presumably identifies the exact time period of regime shifts, and adequately characterizes the behavior of GNP in each separate regime. However, it has been shown by Bessec and Bouabdallah (2005) that regime switching models have a tendency to misclassify future regimes on the out-of-sample data. Also, using regime switching models, the form of error distribution needs to be assumed a priori. In order to avoid making potentially incorrect assumptions about the DGP, we attempt to learn the functional form from the data itself using a nonparametric approach. Using this approach, the error distribution is allowed the flexibility to be of different forms (based on the data) at each forecast horizon. We investigate nonlinear models, in order to tackle nonlinearity in the DGP. Note that if the recent GNP observations belong to a period of recession/contraction, then in order to generate forecasts, this method would find similar historical states in the past that includes observations belonging to a recessionary period, and would utilize their future behavior to issue forecast for the current period. This allows the proposed models to provide a transition between regimes, without making prior assumptions about the presence and number of regimes in the time series.

The modeling framework employed for the proposed models relies on parameter estimates obtained by optimizing the performance of in-sample density forecasts, quantified using CRPS. The advantage of using CRPS for quantifying density forecast performance is that it takes into account both *sharpness* and *calibration*. The first property, sharpness, rewards the model if the forecast density has small spread and is highly peaked around the actual observation. The second property, calibration, rewards the model if the agreement between the forecast density and the actual data density is good, for further details please see – Gneiting,

Balabdaoui, and Raftery (2007). Using CRPS, Gneiting and Thorarinsdottir (2010) discuss the issues regarding the choice of benchmarks in density forecasting, and show that forecasts from a simple no-change model are competitive with expert forecasts. Gneiting and Ranjan (2011a) propose a weighted version of the CRPS for economic applications, whereby the weighting allows more emphasis on the quantiles of interest. Gneiting and Ranjan (2011b) provide details of density forecasting studies in economics based on utilizing prediction pools, and discuss linear and nonlinear aggregation methods by taking into account the coherence, calibration and dispersion of forecasts. The proposed models are simple in that they require estimation of only one and two parameters. Note that it is computationally very expensive for one to estimate parameters based on CRPS for regime switching models that comprise a relatively large number of parameters. Estimating model parameters based on density forecast performance ensures that the estimation framework takes into account the divergence between the true and fitted distribution, and not just a point estimate (even if it maximizes in-sample likelihood). This approach does not require assumptions about the form of error distribution, as opposed to previous nonlinear modeling approaches, see Teräsvirta and Anderson (1992). Also, the proposed model does not require the forecast evaluation to be conditioned on separate regimes of GNP time series to convincingly and consistently outperform classical linear and nonlinear models (AR, SETAR and MS-AR) across varying forecast horizons, as contrasted with the forecast evaluation scheme employed by Tiao and Tsay (1994).

Given a discrete time series x_t , with $t=1, 2, \dots, N$, a m -dimensional delay vector can be constructed as, $\mathbf{x}_t = (x_{t(m-1)}, x_{t-1}, \dots, x_t)$. The dimension m is estimated based on minimization of in-sample density forecast error (as explained in the next subsection).

We construct a simple nonlinear analogue forecasting model for the GNP time series, whereby an analogue is defined as a past state that is similar to the current state. As proposed by Lorenz (1969), one forecasting strategy for nonlinear systems is to search for the previous state (analogue) that is relatively similar to the current state, and issue the trajectory (observation corresponding to next instance in time) of the analogue as a forecast for the current state. This approach is based on the assumption that the dynamics underlying the economic time series follows a pattern, such that the trajectory of the economy from the current state would be quite similar to the trajectories of similar past states, given that a similar economic situation had been witnessed. Similarity between two states is typically quantified by the Euclidean distance between them.

The idea of analogue forecasting was extended by Paparella et al. (1997) through identification of a subset of states in a particular neighborhood (local region of state-space) similar to the current state, and choose a state at random from the selected subset for predicting nonlinear time series. The size of the neighborhood can be defined either in terms of the number of states, radii (distance) from the current state or fraction of total number of states. If S is the total number of available states and f ($f \in [0, 1]$) is the optimum neighborhood fraction size, the total number of states within the neighborhood would be given by $f \times S$. Note that as the neighborhood size increases to include all the available state-vectors in the training set, the issued forecast approaches the unconditional mean of the historical data. On the other hand, if the neighborhood size shrinks so as to include only the closest state to the current state, the trajectory of the analogue is issued as the forecast. Estimating an optimum nearest neighborhood sub-space based on minimization of the in-sample density forecast error, tends to find a trade-off between these two extreme cases by capturing the collective behavior of only the most relevant similar states for forecast generation. The specific approach used to sample from the neighborhood takes into account the behavior of similar past states, allowing us to generate different plausible trajectories the economy could take from the current state. Given an optimum neighborhood size, one can also build a local linear model for nearest states, and use the estimated local model for forecasting (Cleveland, Devlin and Grosse (1988)).

3.1 Fraction-nearest neighbor (f -NN)

The nearest neighbor method is a nonparametric form of regression (Hastie, Tibshirani, and Friedman (2009)), and has proven to be useful in forecasting a range of time series. Fernandez-Rodriguez, Sosvilla-Rivero, and

Andhrada-Felix (1999) used nearest neighbor methods for forecasting the exchange rates of nine currencies, and reported that a nearest neighbor model was able to outperform an autoregressive integrated moving average (ARIMA) model and a random walk benchmark. Similar results were reported by Barkoulas, Baum, and Chakraborty (2003) who adopted nearest neighbor method for forecasting US interest rates. They reported that nearest neighbor method outperforms Box and Jenkins (1970) models and a random walk with drift benchmark, in terms of the out-of-sample forecast accuracy. One major drawback of the nearest neighbor method with a fixed neighborhood size is that the neighborhood does not adapt to the change in data density.

The general idea behind the f -NN method is to identify an optimum neighborhood size using cross-validation, and utilize the collective behavior of the ensemble of nearest neighbor trajectories (arising from the nearest neighbor states). This f -NN method requires the estimation of only one parameter, f , that denotes the optimum fraction of points (nearest neighbors) to be identified for forecasting. Estimating the size of the neighborhood in terms of a fraction has the advantage that the radius at each local neighborhood adapts with the density of data in the state-space. Note that the radius of the neighborhood varies with the location in state space (defined by the current state vector \mathbf{x}_t). If the density of state vectors in the vicinity of \mathbf{x}_t is low, the radius of the local neighborhood will be relatively large in order to include sufficient state vectors. In contrast, if the density is high, the radius will be relatively small. This flexibility allows the model to find the balance between state vectors corresponding to different regimes. To compute a forecast for a current state $\mathbf{x}_t = (x_{t-m+1}, x_{t-m}, \dots, x_t)$, we first calculate the Euclidean distance vector $\{D_i\}_{i=m}^{i=t-m}$ (for $t > m$), where $D_i = \|\mathbf{x}_t - \mathbf{x}_i\|$. We then identify a set of nearest states by defining an integer mapping $\vartheta_i: Z \rightarrow Z$, such that $D_{\vartheta_1} \leq D_{\vartheta_2} \leq \dots \leq D_{\vartheta_k}$, whereby D_{ϑ_k} corresponds to the distance of the k^{th} nearest state. The local neighborhood is denoted by the radius $r(\mathbf{x}_t) = D_{\vartheta_k}$, where $k = \lfloor f \times N_t \rfloor$ such that $N_t = t - 2m + 1$ denotes the total number of previous state vectors from a given time instant t . Note that for out-of-sample forecasting, $t \in [T, N]$, where T is the forecast origin, and N is the length of the time series. We generate a probabilistic forecast by simulating an ensemble of Q scenarios by sampling randomly from the discrete set of k state vectors lying within the neighborhood (as defined by f). The point forecast for a particular horizon (h) is obtained by computing the mean of an ensemble of Q scenarios.

The optimum neighborhood fraction f and the dimension of the state-space m are estimated using cross-validation. This is achieved by minimizing one-step ahead in-sample density forecast error quantified using CRPS. In order to estimate the optimum value of f and m , we first divide the in-sample time series into two parts, one for training and another for validation. The value of f is varied from a minimum, to include only the closest state to the current state, to a maximum, when all states from the training set are included in forecast generation. The value of m is varied from one to ten, in increments of one. For a given set of values of m and f , a series of rolling density predictions is generated using random sampling for the validation set. The value of m and f that minimizes in-sample density error is chosen as the optimum estimate, and is employed for generating out-of-sample forecasts.

Specifically, the in-sample time series denoted by $\mathbf{y}_t = (x_1, x_2, \dots, x_T)$, T being the forecast origin, is divided into two unequal parts (for this work, we divide the in-sample set in ratio of 3:1). The first part of the in-sample set (\mathbf{y}_j) comprises observations ranging from x_1 to x_j , while the second half (\mathbf{y}'_T) includes observations ranging from x_{j+1} up to x_T . Hence, the in-sample data has been divided into a training set (\mathbf{y}_j), and a validation set (\mathbf{y}'_T). Using the observations from the training set, probabilistic forecasts are generated and evaluated for the validation set. In order to generate probabilistic forecasts for the observations in the validation set (say observation x_{j+1}), we first compute a distance vector (based on the Euclidean distance) between the current state vector \mathbf{x}_j and the previous state vectors. The distance vector is then sorted based on the integer mapping ϑ_i (as explained previously). Using the sorted distance vector, we identify a set of closest states to the current state \mathbf{x}_j , that lie within the neighborhood defined by f . Once the closest states have been identified, the future observations of these closest states are randomly sampled and issued as a probabilistic forecast for x_{j+1} . Specifically, if say \mathbf{x}_s is one of the closest states to the current state \mathbf{x}_j , i.e., \mathbf{x}_s lies within the neighborhood (defined by f) centered at \mathbf{x}_j , the next instance of \mathbf{x}_s , i.e., x_{s+1} would be issued as a forecast for x_{j+1} . Random sampling from the future instances of the closest states is employed for generating density forecasts. We then roll forward one step, generate a density forecast for x_{j+2} , and keep rolling forward until we obtain a

density forecast for x_t . Using the probabilistic forecasts and the actual observations from the validation set, we compute the CRPS for a given m and f . The values of m and f that minimize in-sample CRPS are used for generating forecasts on the out-of-sample data.

The general idea behind this method is that, at each horizon h , the previous states are ranked based on their distance to the current state, and only the top $100 \times f\%$ of the closest historical states are used in forecast generation. For the longer GNP time series, we estimated $f=0.13$ and $m=5$. The parameter estimates for the proposed models is presented in Table 4. Note that for $k=1(f=1/N_t)$, the f -NN method corresponds to the local analogue, in which the future trajectory of the most similar state is issued as a forecast, see Lorenz (1969).

The advantage of generating an ensemble of Q scenarios for nonlinear models is that it does not require any assumption regarding the forecast distribution at each horizon. This nonparametric approach provides sufficient flexibility to generate forecast distributions with properties such as skewness, heavy tails and multimodality.

For multistep forecasts (horizon $h>1$), we iterate the procedure for each individual scenario. The trajectory for a given scenario is obtained by employing the forecast of the previous time steps to include an additional state vector. For $h=2$, when estimating \hat{x}_{t+2} , we would use the forecast \hat{x}_{t+1} to update the state vector $\hat{\mathbf{x}}_{t+1} = (x_{t-m+2}, x_{t-m+1}, \dots, x_t, \hat{x}_{t+1})$. In this way, the scenarios for horizon $h>1$ depends on all the previous forecasts. Note that as we update the state vector, the number of states available to generate forecasts increase. The advantage of using f -NN while making multistep ahead forecast, is that the neighborhood size takes into account the additional available state vector. This means that $r(\mathbf{x}_{t+1})$ computed using f -NN, adapts not only to the density of data in state-space, but also to the number of available states at each horizon, unlike the case when the neighborhood size is defined based on the number of states or a fixed radii.

3.2 Kernel regression

Kernel regression can be viewed as an extension of nearest neighbor methods that rely on the assumption that the more similar the past state is to the current state, the more likely it becomes for their trajectories to match. One of the major limitations of the f -NN method is that it is not sensitive to the degree of similarity between state vectors, i.e., the state vectors are weighted equally, irrespective of their different distances from the current state vector. Sensitivity to similarity is crucial for analogue prediction methods because if a previous state is relatively far from the current state, then it is less likely to provide an accurate representation of the current state.

The general idea behind the kernel regression method is to assign relatively higher weights to the closer state vectors. The kernel is a smooth weighting function centered at each sampled state. The weighting scheme incorporates a Gaussian kernel, and the width of the kernel is controlled by a bandwidth parameter. It has been shown by Härdle (1990) that the performance of kernel regression is more sensitive to the choice of bandwidth parameters than the choice of a kernel function. A large kernel bandwidth corresponds to a flat weighting function, such that most states are allotted nearly equal weights, while a small

Table 4 Parameter estimates for the f -NN, Kernel regression and Weighted Random Analogue Prediction (WRAP) Model based on minimization of in-sample density forecast error.

Parameter	f -NN	Kernel regression	WRAP
f	0.13	–	0.13
β	–	0.70	0.12
m	5	5	5
Total Parameters	1	1	2

kernel bandwidth leads to a relatively huge variation in the weights. The estimation of the kernel bandwidth is crucial, as a large bandwidth introduces high bias (low variance), while a small kernel bandwidth introduces high variance (low bias). Estimating an optimum bandwidth based on the trade-off between bias and variance is crucial, because low variance leads to poor generalization of the model on the out-of-sample data, whereas low bias points towards the model under-fitting the data. The weight for any state vector \mathbf{x}_i is computed as $w_i = e^{-\beta(D_i)^2}$ where β is the bandwidth parameter and $D_i = \|\mathbf{x}_i - \mathbf{x}_t\|$. Hence, the weights are highest for states closest to the current state \mathbf{x}_t , and decay off with an increase in distance. Note that the choice of the weighting function can be varied, however, we use the above weighting function as it provides accurate forecasts. To issue a forecast, we first find the distance of the current state vector \mathbf{x}_t with the previous state vectors. Based on the distance $\{D_i\}_{i=m}^{t-m}$, we compute weights w_i , such that $\sum_i w_i = 1$. Using the weights w_i , a simple weighted average of the future trajectory of the past state vectors is issued as a forecast.

The optimum β for the GNP time series was found to be 0.70 (as it was varied in the range [0.01, 1] in increments of 0.01), and was estimated using cross-validation (based on CRPS). Probabilistic forecasts using kernel regression were generated by weighted random sampling from all the previous state vectors (using corresponding weights w_i), while multistep forecasts were generated in the same way as explained in Section 3.1 for f -NN.

We extended the notion of similarity between two state vectors, quantified using the Euclidean distance, to also take into account their particular time of observation. Specifically, the similarity between two states, say current state \mathbf{x}_t and some past state \mathbf{x}_i depends on their Euclidean distance, and the difference between their time of occurrence $\Delta t'_{ti}$, where $\Delta t'_{ti} = t - i$, for $t > i$. The advantage of this method is that it not only depends on how close (in state space) a previous state is to the current state, but also takes into account how recent or old it is with respect to the current state. This approach adopts a double kernel, and is based on the assumption that somewhere in the past, there are states that would provide an estimate about the future state, but the farther back we look; the less relevant such states become for forecasting. Using this method, the weight for a given state vector \mathbf{x}_i is computed as $w_i = e^{-\beta(D_i)^2} \times e^{-\alpha(\Delta t'_{ti})^2}$, where α and β are the bandwidth parameters of the two kernels. The kernel with bandwidth parameter α allots a relatively higher weight to the more recent states, whereas the kernel with bandwidth parameter β assigns higher weight to the states that are more similar to the current state in state-space. The similarity between two states in a state-space is quantified using the Euclidean distance. This method is based on the idea of using a double kernel estimator as proposed by Yu and Jones (1998). Harvey and Oryshchenko (2012) propose a modification of the kernel density estimator to capture changes in different quantiles over time. Recently, Jeon and Taylor (2012) propose a conditional kernel density (CKD) estimator involving double kernel estimation for forecasting wind power. They estimate kernel bandwidths based on minimization of the in-sample density forecast error. However, we find that incorporating an extra bandwidth parameter α for the time of observation does not lead to a significant improvement in the forecast accuracy of the model, and hence the results for this double kernel method are not presented in the paper.

3.3 Weighted random analogue prediction (WRAP)

The WRAP model can be viewed as a hybrid between the f -NN and kernel regression. The general idea behind this method is to estimate an optimum nearest neighborhood sub-space, and assign weights to the state vectors within the neighboring sub-space based on their degree of similarity to the current state vector. The WRAP model is suitable for characterizing time series exhibiting regime shifts, as it depends on exploiting similarity to past sequence of states, specifically the location and trajectory of the past states, and also their time of occurrence. Although WRAP does not explicitly assume the presence of distinct regimes in the DGP, it is capable of providing a transition from one regime to another, assuming that the current state of the DGP has been witnessed before, somewhere in the past. Since WRAP is based on a nonparametric modeling approach; the major disadvantage of this method lies in its sensitivity to

the uncertainty in observations. This model requires the estimation of two parameters, f , that gives an estimate of the size of neighboring sub-space, and a bandwidth parameter β , that controls the width of the kernel employed for weighting state vectors within the sub-space. The estimation of the parameter set $\{f, \beta\}$ is undertaken using cross-validation, by varying the two parameters over a two-dimensional grid, and choosing the parameter set that minimizes the one-step ahead density forecast error. Once the states lying within the neighborhood have been identified for a given f , weights are allotted to the neighboring states so that similar states get higher weight, whereby the rate of decay of weights is controlled by the bandwidth parameter β . Specifically, the parameter f is varied such that the corresponding k lies within the range $[1, J-m]$, where J denotes the time index that divides the in-sample data into a training and validation part, while β is varied in range $[0.01, 1]$ in increments of 0.01. The estimates of the parameter set $\{f, \beta\}$ was found to be $\{0.13, 0.12\}$. Using the WRAP model, we are basically defining a local probability sub-space (using weights w_j) for the discrete set of k state vectors in the neighborhood (defined by parameter f), and sampling from this neighborhood (using a weighted random sampling) to generate forecasts.

To summarize, the appeal of proposed models in terms of model structure lies in the fact that they are based on one or two parameters, are nonparametric, and can cope with nonlinearity in the DGP. In terms of application, the proposed models generate more accurate point and density forecasts compared to the parametric models, can provide transition between regimes, and do not require the forecast evaluation to be conditioned on separate regimes of GNP. Also, though we demonstrate the utility of the proposed methods for forecasting GNP, they can easily be adapted for other time series as the underlying modeling process is data oriented.

We discuss the forecasting ability of different models under the conditions of a structural break. The presence of structural breaks can adversely impact the out-of-sample forecasting accuracy of time series models even if they have a good in-sample fit, Clark and McCracken (2002). A structural break can be defined as a permanent change in the parameter vector of a model, which leads to instability in parameter estimates. The impact of structural break on forecast accuracy depends on the type of break and also the form of model under consideration, Clements and Hendry (1998). It has been argued by Hansen (2001) that ignoring structural breaks can lead to economic relationships go off course, leading to inaccurate forecasts. Clements and Hendry (1998) discuss the use of updating, intercept corrections and differencing to alleviate the impact of a structural change. It has been discussed by Kapetanios and Tzavalis (2005) that a major limitation of regime switching models, like SETAR and MS-AR, is that these models assume the structural parameter changes to be of fixed magnitude, which can lead to poor forecasts. They also mention that models with time varying coefficient assume the structural breaks to be continuous. To overcome these limitations, Kapetanios and Tzavalis (2005) propose a model that allows for structural breaks to be stochastic in nature, both in magnitude and in timing. The nonlinear regime switching models are better suited for forecasting under the presence of structural breaks compared to their linear counterparts. The nonlinear and nonparametric models considered in this work have the flexibility to allow for structural breaks to be of different magnitudes, which makes them more appealing for out-of-sample forecasting compared to the regime switching models. Note that in order to generate forecasts under the conditions of a structural break, the nonparametric models would quantify the behavior of the past economic periods, which had witnessed a similar structural break as the current period. Also, since the proposed models adapt with the density of states in the state space, the models would tend to include adequate number of analogous past states (exhibiting similar structural breaks) in the neighborhood, and utilize their future trajectories to generate forecasts for the current period. This modeling approach allows the nonparametric models to adapt differently to the conditions of structural breaks based on their respective magnitudes. The precondition for the proposed nonparametric models to perform well under the conditions of a structural break for a given period, however, is that a similar break must have already been witnessed before, somewhere in the past.

The aforementioned reasons make the proposed models appealing conceptually, which is reflected in their ability to generate more accurate point and density forecasts than parametric models for the GNP time series, as demonstrated in the next section.

4 Results and observations

In this section, we compare the out-of-sample point and density forecast performance of the classical and proposed models. The point forecast performance is evaluated using the root mean squared error (RMSE) and the mean absolute error (MAE), given by:-

$$RMSE_h = \sqrt{\frac{1}{N-T-h+1} \sum_{i=T}^{N-h} (x_{i+h} - \hat{x}_{i+h|i})^2}, \quad (4)$$

$$MAE_h = \frac{1}{N-T-h+1} \sum_{i=T}^{N-h} |x_{i+h} - \hat{x}_{i+h|i}|, \quad (5)$$

where $RMSE_h$ (MAE_h) is the RMSE (MAE) at horizon h , x_{i+h} is the actual observation, $\hat{x}_{i+h|i}$ is the h -step ahead forecast, T is the forecast origin and N is the length of the time series. The advantage of using MAE over RMSE, is that it is much less sensitive than RMSE to the occasional outlier. Note that the RMSE gives an estimate about the models ability to forecast the mean of the time series, while MAE reflects the models efficacy in capturing the median.

In order to evaluate the density forecast performance, we use CRPS as it quantifies both calibration and sharpness, and is also a proper scoring rule. Note that CRPS can be viewed as the distributional analogue of the mean absolute error, and is defined in terms of predictive cumulative distribution function (CDF) denoted by F as follows,

$$CRPS(F, x) = \int_{-\infty}^{\infty} \{F(y) - \mathbf{1}(y \geq x)\}^2 dy, \quad (6)$$

where x is the actual observation, and $\mathbf{1}$ is an indicator function that equals one for $y \geq x$. The empirical form of CRPS, as given by Gneiting and Raftery (2007) is represented as,

$$CRPS(F, x) = E_F |X - x| - \frac{1}{2} E_F |X - X'|, \quad (7)$$

where X and X' are independent samples drawn from the forecast density function, each having the same distribution F , while E_F is the expectation with respect to the distribution F . We employ the empirical form of CRPS over the analytic form as it is relatively easy to compute. Note that for $X=X'$, i.e., when F reduces to a point forecast, CRPS is the same as the MAE.

A series of forecasts for horizons ranging from one to sixteen quarters ahead is generated using a rolling forecast scheme, as previously used for GNP time series by Clements and Krolzig (1998). In this work, the GNP time series dating from 1947Q2-2008Q3, is divided into 1947Q2-1996Q4 (199 observations) for training and from 1997Q1-2008Q3 (47 observations) for testing. Specifically, we first generate a sequence of 47 one-step ahead forecasts from 1997Q1 up to 2008Q3 by rolling the forecast origin through the out-of-sample data. We then include the one-step ahead forecast for 1997Q1 in the model, and generate a two-step forecast for 1997Q2. We continue rolling forward so as to obtain a sequence of 46 two-step ahead forecasts. The above process is repeated for increasing horizons, until we obtain 31 sixteen-step ahead forecasts. Note that we do not estimate horizon-specific models, but rather estimate a single model by minimizing one-step ahead error using the in-sample data. Having estimated the model, we then iterate the forecasts forward across multiple time periods. Multistep ahead forecasts for the nonlinear models were generated using Monte Carlo simulations, by constructing an ensemble of Q scenarios. In this work, we chose Q equal to ten thousand. Note that the nonparametric models proposed in this work do not rely on assumptions regarding the forecast distribution at any horizon, and different forecast scenarios for the current state are generated by random sampling of the next instances of similar past states. For kernel regression and WRAP, we employ weighted random sampling to generate density forecasts, whereby the weights are derived based on a kernel (as explained in the previous section). Further details on generating multi-step ahead forecasts using nonlinear models can be found in Granger and Teräsvirta (1992). The model parameters for all models were estimated using only the in-sample data and then held fixed, and each model had same amount of information for making

out-of-sample forecasts. The advantage of this scheme is that it includes previous step forecasts in the model for generating forecasts for a given horizon. Using the same forecasting scheme, Clements and Krolzig (1998) reported that AR is a more robust out-of-sample GNP predictor compared to both SETAR and MS-AR. A comparison of iterated and direct forecasting schemes series has been provided by Marcellino, Stock, and Watson (2006). Iterated forecasts are generated using a model estimated based on one-step ahead error, whereby multi-step ahead forecasts are constructed by iterating the forecasts forward for multiple periods. On the other hand, a direct forecast is generated using a horizon-specific model. Using a range of different macroeconomic time series, Marcellino, Stock, and Watson (2006) show that iterated forecasts are more accurate than direct forecasts, whereby the performance of iterated forecasts improves with longer horizons.

For the GNP time series ranging from 1947Q2-1990Q4, as analyzed by Potter (1995), we choose data from 1947Q2-1984Q4 as the training set, and from 1985Q1-1990Q4 as the testing set. Using a rolling forecast origin, we generate a sequence of one to sixteen-step ahead forecasts as explained above, and evaluate the out-of-sample forecast accuracy of the model. Given the limited size of this time series compared to the more recent GNP time series, it is relatively hard for one to draw reliable conclusions about models forecast accuracy using the shorter time series.

In this work, we use the unconditional mean forecast as a benchmark. For point forecasts, we compute the unconditional mean of the in-sample data, and issue it as a forecast for the out-of-sample data. Note that, in order to make a h -step ahead forecast, original data only up to time index $T-h+1$ is used (T being the forecast origin). The actual number of data points employed to compute RMSE/MAE varies with the forecast horizon. The whole out-of-sample data (starting from $T+1$) is employed to compute one-step ahead error, while the two-step ahead error is computed starting from $T+2$ until the last available quarter. The process is repeated until we obtain 16-quarter ahead forecast error. We find that the unconditional forecast performs well for all horizons compared to the classical models, across both the time series. This result points towards the efficacy and robustness of the unconditional method as a benchmark for GNP time series. We also compared the random walk against the unconditional mean benchmark in terms of forecast accuracy. In order to issue h -step ahead forecast using the random walk benchmark (also known as persistence), the current observation is issued as a forecast for the h -step ahead observation. The unconditional benchmark easily outperformed the persistence benchmark (results not presented here). Note that the f -NN method defaults to the unconditional density forecast as f approaches one.

The out-of-sample point forecast performance of different models is presented in Figure 1 and 2, while density forecast performance is presented in Figure 3. When evaluated using RMSE (Figure 1), we find that f -NN, kernel regression, WRAP and SETAR(2;1,1) do extremely well for shorter horizons. The f -NN outperforms

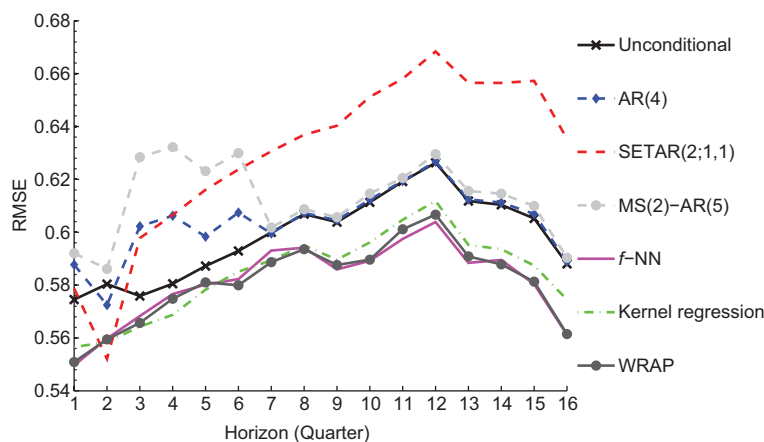


Figure 1 Out-of-sample forecast performance of classical models (AR, SETAR and MS-AR), unconditional method and proposed models (f -NN, kernel regression and WRAP) for the longer US GNP time series (1947Q2-2008Q3) based on the root mean square error (RMSE).

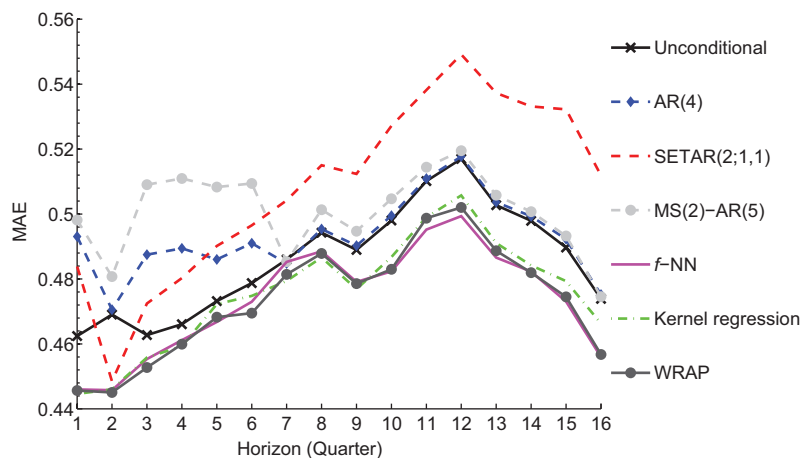


Figure 2 Out-of-sample forecast performance of classical models (AR, SETAR and MS-AR), unconditional method and proposed models (f -NN, kernel regression and WRAP) for the longer US GNP time series (1947Q2-2008Q3) based on the mean absolute error (MAE).

all the other models for most horizons. Given the limited number of samples available for computing RMSE for longer horizons, f -NN was however, not the best model overall. SETAR(2;1,1) performs well only for shorter horizons, whereas the forecast performance of AR(4) and MS(2)-AR(5) is not distinguishably superior to the unconditional forecast. The proposed models (f -NN, kernel regression and WRAP) outperform all other models on all horizons except for horizon $h=2$ when SETAR is slightly better. In the case of MAE (Figure 2), the proposed models (f -NN, kernel regression and WRAP) outperform all other models on all forecast horizons.

From the CRPS results (Figure 3), it is evident that f -NN, kernel regression and WRAP consistently outperform the classical models on all horizons. The superiority of these proposed models is enhanced when quantifying their density forecasts using CRPS. The superior density forecasts of the proposed models could be related to their parameter estimation process which optimizes CRPS. The greater point forecast accuracy of the proposed models relative to the classical models, when evaluated using RMSE and MAE, signifies that the forecasting ability of the proposed models is not restricted by the choice of model estimation framework (CRPS or OLS). To verify this further, we investigated an alternate case whereby the parameter estimation process for nonparametric models involved minimization of in-sample RMSE, as opposed to minimization

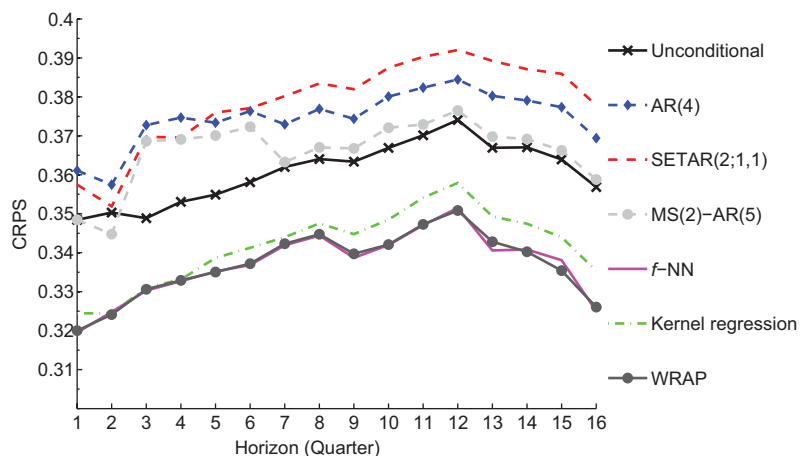


Figure 3 Out-of-sample forecast performance of classical models (AR, SETAR and MS-AR), unconditional method and proposed models (f -NN, kernel regression and WRAP) for the longer US GNP time series (1947Q2-2008Q3) based on the continuous ranked probability score (CRPS).

of in-sample CRPS. This was achieved by generating one-step ahead point forecasts using nonparametric models for a range of different parameters, and choosing the parameters which minimized one-step ahead RMSE on the in-sample validation set. We found that the nonparametric models (estimated using RMSE) outperformed the parametric models and the unconditional benchmark on most forecast horizons, when the out-of-sample point forecast accuracy was quantified using both RMSE and MAE. Also, note that it is not straightforward to estimate more than, say, two parameter values by optimizing CRPS, making the parameter estimation for AR/SETAR/MS-AR models extremely difficult. The superior forecast performance of the proposed models compared to the unconditional forecast across varying horizons, demonstrates the ability of the proposed models for generating both short (1 quarter) and long term (4 years) forecasts. Interestingly, we find that MS(2)-AR(5), which was selected over the fourth order Markov switching model MS(2)-AR(4), (as proposed by Hamilton (1989) for GNP) as the best model using AIC, performed worse than the MS(2)-AR(4) in terms of its forecasting ability. Hence, from the perspective of forecasting, it remains doubtful whether AIC is the right choice for selecting the lag order for the nonlinear models, as also pointed out previously by Clements and Krolzig (1998).

We show the plot of one-step ahead out-of-sample predictive error density obtained using different models in Figure 4. This figure shows that the predictive error density obtained from the nonparametric models is narrower and centered close to zero compared to the parametric models. We also plot one-step ahead out-of-sample forecasts generated using the parametric and nonparametric models along with the corresponding actual observations in Figure 5. It can be seen from this figure that forecasts from the parametric models deviate more from the actual observations compared to the nonparametric models. Figure 6 plots the actual GNP quarterly returns (Δx_t) and corresponding one-step ahead forecasts ($\hat{\Delta x}_{t|t-1}$) obtained from different models versus GNP quarterly returns lagged (Δx_{t-1}). To investigate the suitability of the proposed models in capturing the data density, we generate a sequence of one-step ahead out-of-sample density forecasts from 1997Q1 up to 2008Q3 using the WRAP model and analyze the calibration of the model. In Figure 7, the different shaded regions correspond to the different quantile ranges (centered around the median) of the forecast distribution. Note that there are several instances in time (around 2002, 2007 and 2008) when the GNP recorded a high change with respect to the last quarter. The point forecast corresponding to the particular time of observation fails to capture these sharp changes in the GNP, but it is during these times, that there is a marked increase in the uncertainty associated with the forecasts. It is evident from Figure 7, that the forecast density estimates gives more informed and valid inference about the changes in GNP, compared to point estimates. This underscores the need for generating density forecasts along with point forecasts, in order to communicate forecast uncertainty to policy-makers.

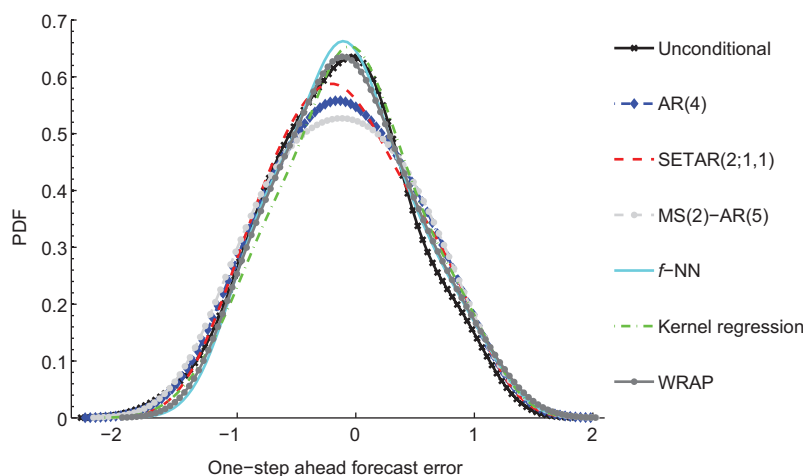


Figure 4 One-step ahead out-of-sample predictive error density obtained using classical models (AR, SETAR and MS-AR), unconditional method and proposed models (f -NN, kernel regression and WRAP).

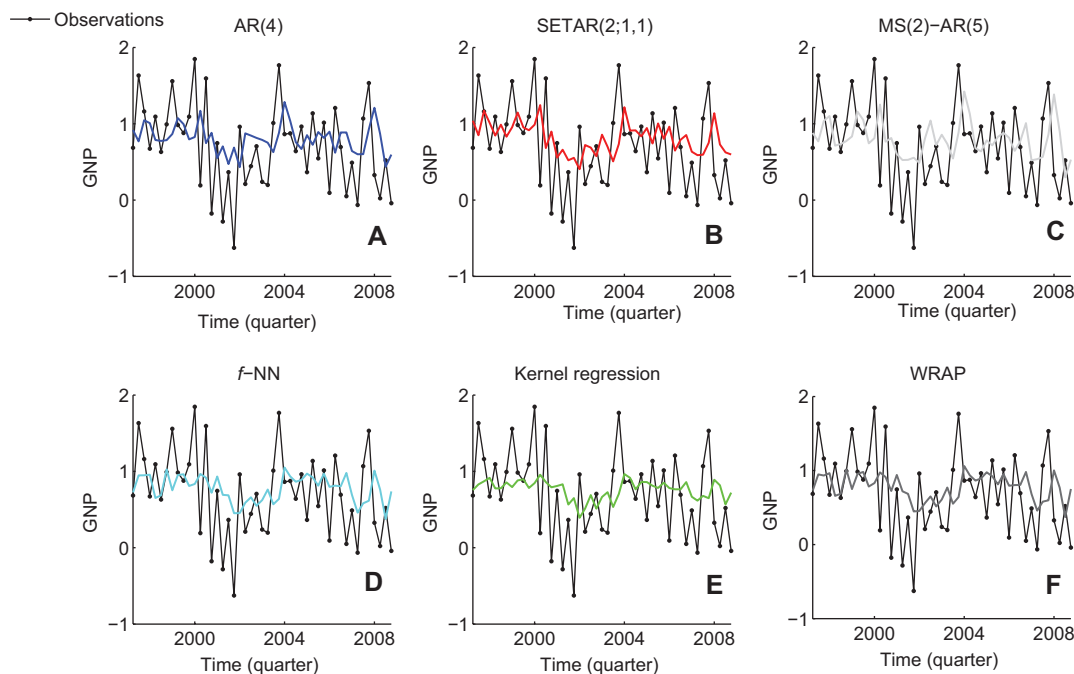


Figure 5 One-step ahead out-of-sample point forecasts of GNP quarterly returns from 1997Q1 up to 2008Q3 generated using (A) AR, (B) SETAR, (C) MS-AR, (D) f -NN, (E) kernel regression, and, (F) WRAP. The colored lines correspond to the forecasts, while the back dotted line indicates the actual observations.

The quality of the density forecasts generated using the WRAP model is further evaluated using a reliability diagram. A reliability diagram measures the degree to which the actual observations lie within the different quantile ranges of the density forecast, also known as coverage. Note that each quantile range of the density forecast is expected to encompass a specific proportion of the actual observations, as quantified by coverage. The difference between the expected and obtained coverage is known as bias. We find that the bias for the WRAP model across different quantiles is small, as shown in Figure 8.

In addition to using the RMSE and MAE performance scores, we employ tests for equal predictive ability and forecast encompassing to evaluate point forecasts from different models. Specifically, we use the McCracken (2007) MSE-F statistic, which is a variant of the Diebold and Mariano (1995) and West (1996) and west statistic, to test for equal predictive ability. In order to test for forecast encompassing, we use the Clark and McCracken (2001) ENC-NEW statistic, which is a variant of the Harvey, Leybourne, and Newbold (1998) statistic. Clark and McCracken (2001, 2004) find these statistics to be more powerful than their original counterparts in extensive Monte Carlo simulations.

The MSE-F statistic for equal predictive ability is based on comparing the difference in forecast accuracy between two models using the mean squared error criterion. Given two linear regression models, whereby one comprises of say k_1 parameters (restrictive model), whereas the other comprises of k_1+k_2 parameters (unrestrictive model), the MSE-F statistic checks if inclusion of an extra k_2 number of parameters is justified, based on whether the unrestricted model leads to a reduction in forecast error over the restricted model. An unrestrictive model (referred to as model 2) can basically be viewed as an extension of the restrictive model (referred to as model 1). If model 2 does not offer improvement in forecast accuracy over model 1, then it can be inferred that model 2 nests models 1, and hence, the inclusion of an extra k_2 parameters cannot be justified. Alternatively, if model 2 generates more accurate forecast than model 1, then model 2 can be deemed correct and the inclusion of extra parameters is reasonable. The MSE-F statistic uses the loss differential, $\hat{d}_{t+h} = (\hat{u}_{1,t+h})^2 - (\hat{u}_{2,t+h})^2$, where $\hat{u}_{1,t+h} = (x_{t+h} - \hat{x}_{1,t+h|t})$ is the h -step ahead out-of-sample forecast error obtained using model 1, while $\hat{u}_{2,t+h} = (x_{t+h} - \hat{x}_{2,t+h|t})$ is the error from model 2. Using $\bar{d} = (N-T-h+1)^{-1} \sum_{t=T}^{N-h} \hat{d}_{t+h}$, the MSE-F statistic is defined as,

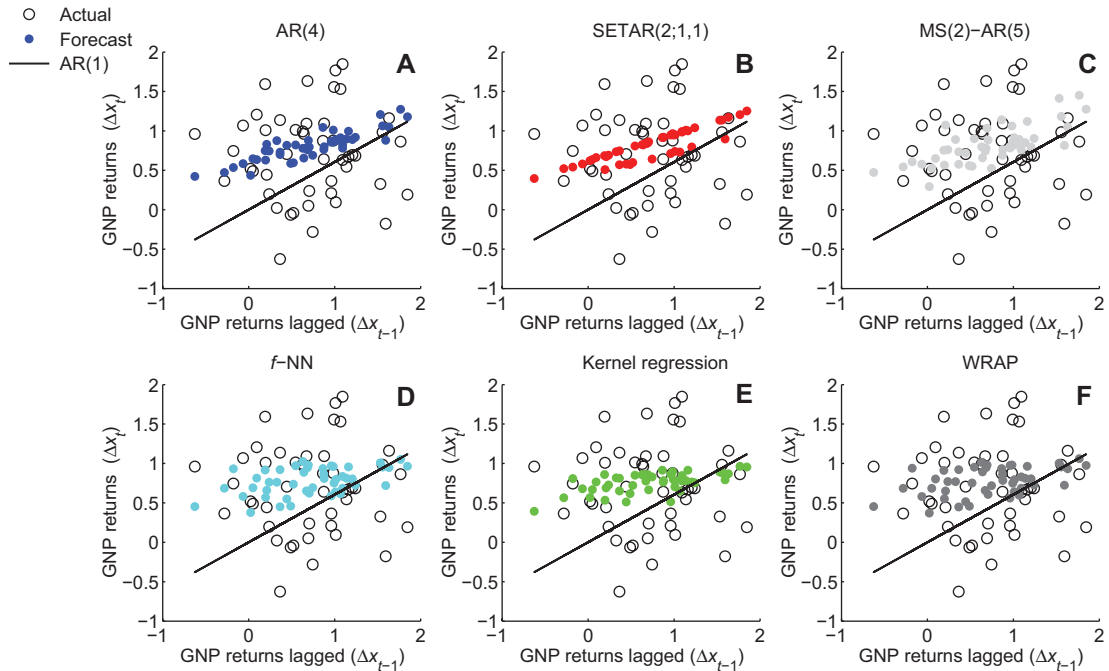


Figure 6 Plot of GNP quarterly returns (Δx_t) versus GNP quarterly returns lagged (Δx_{t-1}) (black circles) using the out-of-sample data, along with the plot of corresponding one-step ahead forecasts ($\Delta \hat{x}_{t|t-1}$) versus (Δx_{t-1}) (colored dots) for (A) AR, (B) SETAR, (C) MS-AR, (D) f -NN, (E) kernel regression, and, (F) WRAP. The black line denotes one-step ahead forecasts obtained from an AR(1) model.

$$MSE-F = (N-T-h+1) \times \frac{\bar{d}}{MSE_2}, \quad (8)$$

where $MSE_2 = (N-T-h+1)^{-1} \sum_{t=T}^{N-h} (x_{t+h} - \hat{x}_{2,t+h|t})^2$ denotes the mean squared error for model 2 on the out-of-sample forecast window. Note that if model 2 offers improvement in forecast accuracy over model 1 overall, then the errors from model 2 would be less than model 1, and the MSE-F statistic would be positive. A significant MSE-F statistic points towards the utility of an unrestricted model over a restricted model. In the context of this work, we are interested in finding the justification for using complicated models comprising large

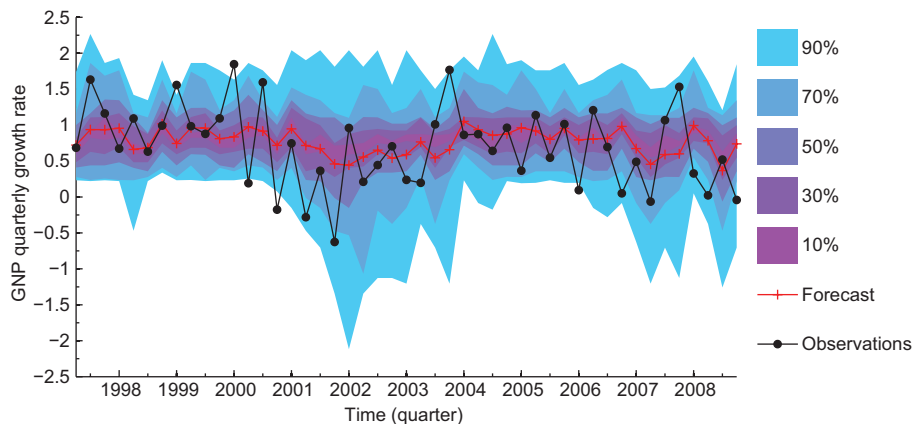


Figure 7 One-step ahead out-of-sample density forecasts of GNP quarterly returns from 1997Q1 up to 2008Q3 using the weighted random analogue prediction (WRAP) model. The shaded regions correspond to the different quantiles of the density forecasts centered on the median. The red line corresponds to the point forecasts provided by the mean of the forecast density and the black line indicates the observations.

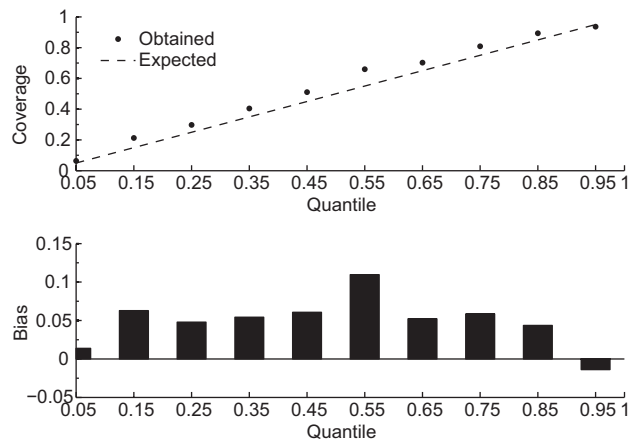


Figure 8 Reliability diagram (coverage and bias) for one-step ahead out-of-sample forecasts from 1997Q1 up to 2008Q3 using the weighted random analogue prediction (WRAP) model. Coverage reflects the degree to which actual observations lie within different quantile ranges of forecast distribution. For a perfect density forecast, the coverage would lie along the diagonal (dashed line), and hence the bias would be zero across all quantiles.

number of parameters, over parsimonious models that employ fewer parameters. Hence, we use the unconditional mean as a restrictive model, and treat the parametric and nonparametric models as the unrestricted models. This allows us to gauge the improvement in forecast accuracy offered using different modeling approaches over the unconditional forecast.

We use the ENC-NEW statistics to test for forecast encompassing, which measures the covariance between $\hat{u}_{1,t+h}$ and $\hat{u}_{1,t+h} - \hat{u}_{2,t+h}$. If model 1 forecasts encompasses model 2, the covariance between the above two terms would be less than or equal to zero (Clark and McCracken 2004). However, if model 2 consists of additional information over model 1, the covariance would be positive. If $\hat{c}_{t+h} = \hat{u}_{1,t+h}(\hat{u}_{1,t+h} - \hat{u}_{2,t+h})$, and $\bar{c} = (N - T - h + 1)^{-1} \sum_{t=T}^{N-h} \hat{c}_{t+h}$, the statistic is defined as,

$$ENC-NEW = (N - T - h + 1) \times \frac{\bar{c}}{MSE_2}. \quad (9)$$

As pointed by Rapach, Wohar, and Rangvid (2005), if model 1 encompasses model 2, the weight attached to the unrestricted model is zero under the optimal composite forecast, which is a convex combination of the restricted and unrestricted model forecasts. Gupta and Modise (2011) employ these statistics to compare the out-of-sample predictability of stock returns using macroeconomic variables. Clark and McCracken (2002) investigate the impact of structural breaks on tests for equal predictive ability and forecast encompassing. They mention that the out-sample tests are useful in revealing the predictive ability of some variables over others, and report that structural breaks can lead to poor out-of-sample forecasts even if models fits the in-sample data well. In order to provide a comprehensive out-of-sample forecast evaluation, we show the MSE-F and ENC-NEW statistics for both parametric and nonparametric models across varying forecast horizons. The statistics are computed using only the out-of-sample data. A detailed overview on forecast encompassing can be found in Clements and Hendry (1998).

The MSE-F statistic is provided Figure 9, while Figure 10 gives the ENC-NEW statistic for different models across all horizons. The parametric and nonparametric models are treated as the unrestricted model, while the unconditional benchmark is the restricted model. It can be seen from these two figures that the parametric models (AR, SETAR and MS-AR) do not provide any additional information over the unconditional benchmark overall, whereas the nonparametric models (f-NN, kernel regression and WRAP) offer improvement in forecast accuracy at all horizons. Forecasts from parametric models are competitive with the unconditional forecast only for the first two horizons, and deteriorate afterwards. Among parametric models, AR is the most consistent predictor, while SETAR is the worst performing model overall.

Interestingly, Clements and Krolzig (1998), and Galbraith (2003) have reported similar results regarding the comparison of the out-of-sample point macroeconomic forecasts obtained from different parametric models. Clements and Krolzig (1998) showed that if the DGP is nonlinear, SETAR and MS-AR are competitive with AR only for one and two-quarter ahead forecasts. They report that for horizons beyond two-quarters, AR is the most accurate model, while MS-AR is worse than AR, but more accurate than SETAR. Galbraith (2003) defines content horizon as the forecast horizon beyond which forecasts obtained using the unconditional mean are no worse than the forecasts obtained from the model under consideration, and report that an AR model is not superior to the unconditional mean beyond two quarters for forecasting US GDP. Among nonparametric models, we find that both f -NN and WRAP are more accurate than kernel regression, and all three models are superior to the unconditional benchmark. Note that these results are in accordance with the out-of-sample point forecast results based on RMSE and MAE reported in Figure 1 and 2, whereby the nonparametric model forecasts are more accurate than the unconditional forecast.

We use the two sided Kolmogorov-Smirnov test, proposed by Kolmogorov (1933) and Smirnov (1939), to further investigate if the forecast errors from different models are the same. Assuming a null hypothesis (H_0) that the forecast accuracy of different models is the same, the Kolmogorov-Smirnov (KS) statistic and the corresponding p -values can be used to test this hypothesis. The KS test quantifies the difference between the two empirical cumulative distribution functions (ECDFs) based on the maximum difference between them. The KS test is a nonparametric statistical test, i.e., it does not assume a specific form of the distributions. The two sided KS statistic (D_n) is defined as:-

$$D_n = \sup_x |F_n(x) - G_n(x)|, \quad (10)$$

where sup (or supremum) is the least upper bound, while F and G are the two ECDFs under consideration. To compute the KS statistic, we first generate a sequence of one-step ahead density forecasts for the out-of-sample GNP data (1997Q1-2008Q3) using all the models and the unconditional mean benchmark. The KS statistic and the corresponding p -value are then computed between different pairs of models and the benchmark, for each quarter on the out-of-sample period. We report the average KS statistic and the p -value in Table 5. Based

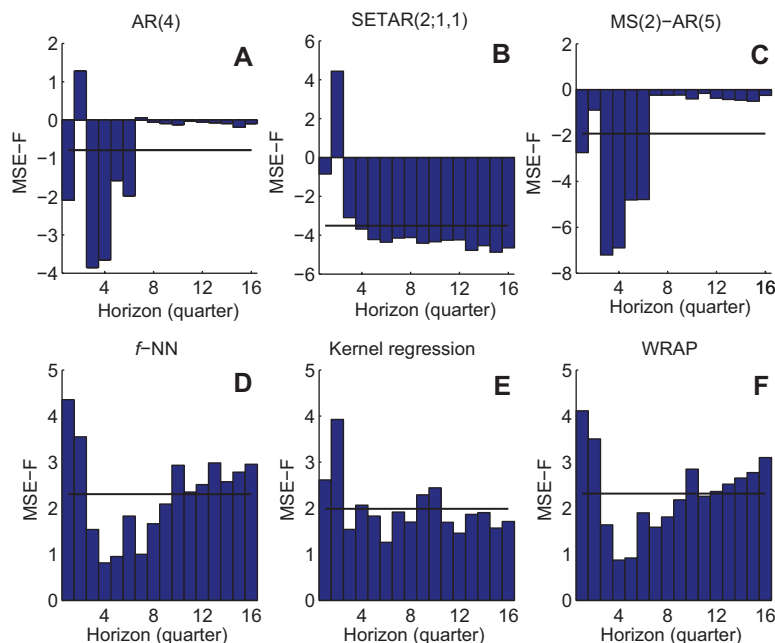


Figure 9 The MSE-F statistic for the test of equal predictive ability reported for all 16 horizons using the unconditional forecast as the restrictive model, and (A) AR, (B) SETAR, (C) MS-AR, (D) f -NN, (E) kernel regression, and, (F) WRAP as the unrestrictive model, computed using the out-of-sample data. The horizontal black line corresponds to the average of MSE-F statistic across all horizons.

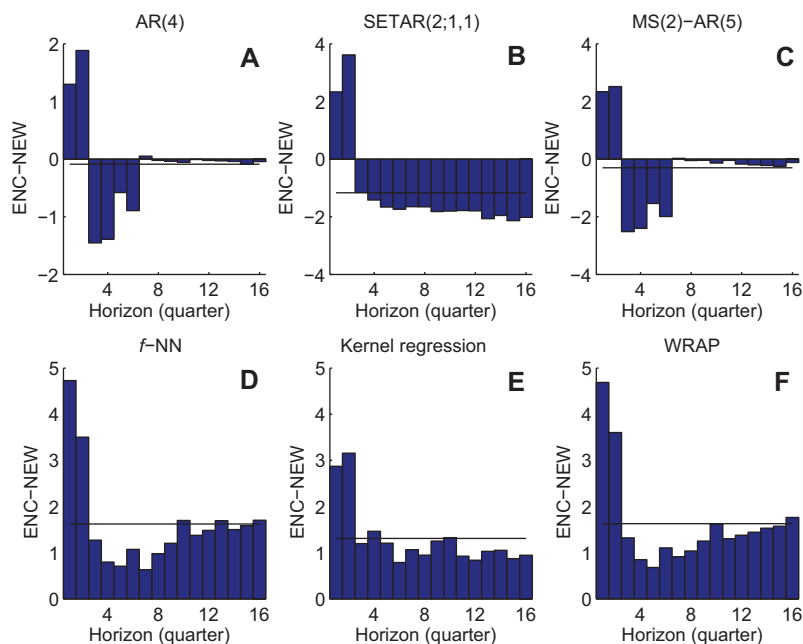


Figure 10 The ENC-NEW statistic for the test for forecast encompassing reported for all 16 horizons using the unconditional forecast as the restrictive model, and (A) AR, (B) SETAR, (C) MS-AR, (D) f -NN, (E) kernel regression, and, (F) WRAP as the unrestrictive model, computed using the out-of-sample data. The horizontal black line corresponds to the average of ENC-NEW statistic across all horizons.

on a 1% significance level, the cases where the null hypothesis is rejected are reported in bold. It is evident from the Table 5, that the forecasts from nonparametric models (f -NN, kernel regression and WRAP) are significantly different from the parametric models (AR, SETAR and MS-AR), as also evident from the model evaluation based on the different performance scores as reported earlier.

To evaluate the out-of-sample density forecasts from different models, specifically the forecast calibration, we use the probability integral transform (PIT), as also employed by Gneiting, Balabdaoui, and Raftery

Table 5 Kolmogorov-Smirnov (KS) statistic for different pairs of models and the benchmark, computed using one-step ahead density forecasts on the out-of-sample GNP data (1997Q1-2008Q3). The null hypothesis (H_0) is that the accuracy of forecasts generated using different models is the same. Cases where H_0 is rejected (1% significance level) are reported in **bold**.

Model	Uncond.	AR	SETAR	MS-AR	f -NN	K. Reg
Uncond.	—	—	—	—	—	—
AR	0.113 (***)	—	—	—	—	—
SETAR	0.125 (***)	0.057 (**)	—	—	—	—
MS-AR	0.126 (***)	0.073 (***)	0.095 (***)	—	—	—
f -NN	0.193 (***)	0.227 (***)	0.228 (***)	0.210 (***)	—	—
K. Reg	0.136 (***)	0.185 (***)	0.188 (***)	0.175 (***)	0.089 (***)	—
WRAP	0.193 (***)	0.226 (***)	0.228 (***)	0.210 (***)	0.0132 (*)	0.087 (***)

*($p \geq 0.01$); **($p < 0.01$); ***($p < 0.001$).

(2007), Little, McSharpy, and Taylor (2009), and Harvey and Oryshchenko (2012). Note that calibration refers to the consistency between forecast density and the actual data density. The PIT is defined as the value that the predictive cumulative distribution function attains at the observation, and is computed as,

$$z_n = \int_{-\infty}^{x_n} p(u) du, \quad (11)$$

where x_n is the actual observation and p is the one-step ahead forecast density function. If the forecasts are perfectly calibrated, i.e., the forecast density matches completely with the true density, then z_n would be independent and identically distributed (IID) with uniform density between 0 and 1. A visual inspection of PIT histogram is used to infer the deviation of PIT values from uniformity. In the context of evaluating density forecasts from econometric models, the utility of PIT has been shown by Gençay and Selçuk (1998), who analyze the PIT histogram along with correlograms of different transformations of the PIT values.

We use stochastic interpolation to calculate z_n , by drawing 10,000 sample values from the out-of-sample forecast distribution in order to construct the empirical cumulative density function, as done in Little, McSharpy, and Taylor (2009). The PIT histogram (using 10 bins) for each of the parametric and nonparametric models is provided in Figure 11. A visual inspection of the histograms suggests that nonparametric models are relatively more uniformly distributed compared to parametric models. Note that a U-shaped PIT histogram points towards the predictive distribution being too narrow (with respect to the actual distribution), whereas an inverted U-shaped (hump shaped) histogram corresponds to an over dispersed distribution, Gneiting, Balabdaoui, and Raftery (2007). However, the PIT histogram is not a skill score that quantifies the relative performance of models, but is a visual aid that can be used in conjunction with different performance scores. We also plot the autocorrelation function (ACF) of PIT values (with 95% confidence intervals) to check for the presence of serial correlations. Note that if the structure in a time series is adequately captured by a model, this would be reflected in corresponding z_n being serially independent. It can be seen from Figure 12 that the autocorrelation in PIT values across different models is not significant across most lags, showing the absence of serial correlations in z_n obtained from both parametric

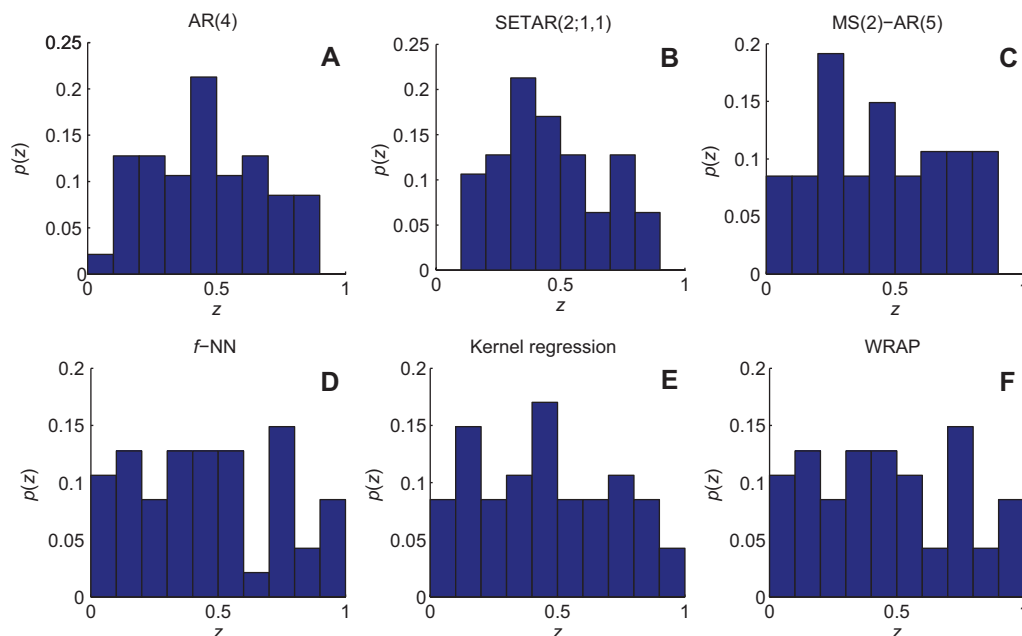


Figure 11 Probability integral transform (PIT) histogram for one-step ahead out-of-sample forecasts obtained using (A) AR, (B) SETAR, (C) MS-AR, (D) f -NN, (E) kernel regression, and (F) WRAP.

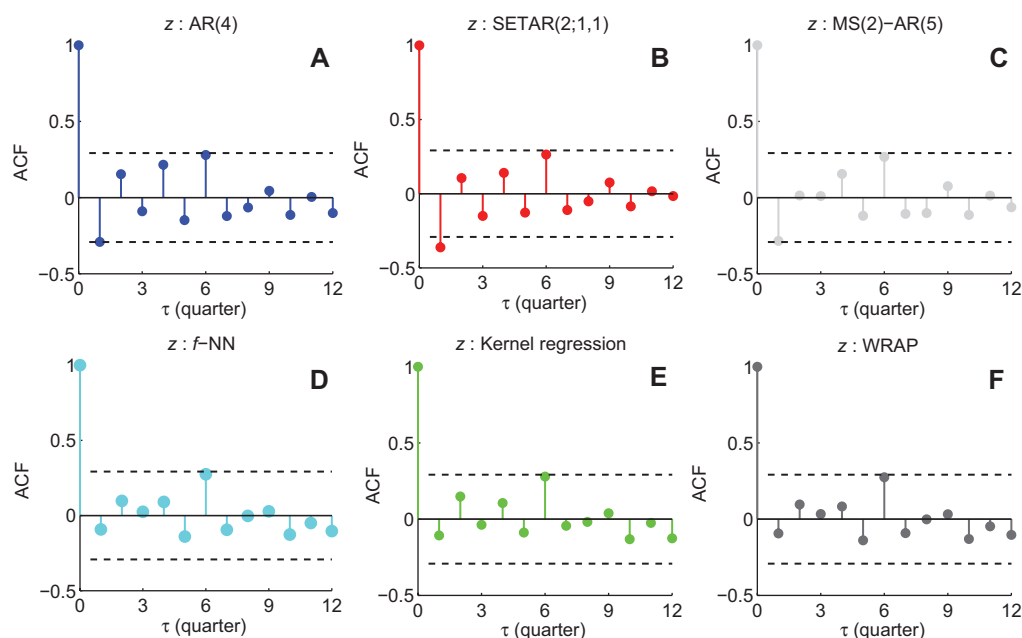


Figure 12 Autocorrelation function (ACF) for z -series (associated with the corresponding histogram shown above) for one-step ahead out-of-sample forecasts obtained using (A) AR, (B) SETAR, (C) MS-AR, (D) f -NN, (E) kernel regression, and (F) WRAP. The horizontal dashed lines correspond to the 95% confidence interval.

and nonparametric models. This result points towards different models capturing the serial dependence present in the GNP time series.

5 Conclusion

In this paper, we proposed three nonlinear models based on the estimation of one and two parameters that outperform previously published models for forecasting the US GNP. Our proposed models are not only parsimonious, but also rely on very few assumptions regarding the presence of regimes in the time series and provide a smooth transition between periods of contraction and expansion. We find that neither the linear AR model, nor the classical regime switching models outperform a naïve point and density forecast based on the unconditional mean. This indicates the importance of using the unconditional forecast as a benchmark instead of persistence or a random walk benchmark as used previously by Clemen and Guerard (1989). Overall, a novel two-parameter model, WRAP, is found to provide the most accurate forecasts across multiple horizons on both GNP time series, when compared with the more complicated SETAR and MS-AR models, and simple benchmarks. Out of the two single parameter models (f -NN and kernel regression), we found that f -NN performs best and is highly competitive with the two-parameter WRAP model. Given the simplicity of f -NN, and its competitive performance, we recommend that future economic forecast evaluators employ it as a benchmark. Despite its simplicity, the proposed models (f -NN, kernel regression and WRAP) outperform the more complex nonlinear models (SETAR(2;1,1) and MS(2)-AR(5) require 8 and 10 parameters respectively) in terms of point and density forecasts.

Given the importance of quantifying uncertainty in forecasts for informed decision-making, it is important to correctly assess candidate models, whether linear or nonlinear, based on their ability to generate probabilistic forecasts. It is worth emphasizing that the linear AR(4) model requires 5 parameters and therefore could be viewed as being more complex and less parsimonious than the simple nonlinear model, f -NN. We recommend that for generating accurate density forecasts, parameter estimation must be undertaken

based on minimization of in-sample density forecast error, ensuring that the risk associated with decision-making based on such forecasts is adequately assessed.

Acknowledgments: We wish to thank the editor and two anonymous reviewers for their useful comments on our work, which has greatly helped to improve the quality of this manuscript. We are also very grateful to James W. Taylor for many insightful suggestions.

References

- Akaike, H. 1973. "Information Theory and an Extension of the Maximum Likelihood Principle." In *Second International Symposium on Information Theory*, edited by B. N. Petrov and F. Csaki, 267–281. Budapest: Akademiai Kiado.
- Ang, A., M. Piazzesi, and M. Wei. 2006. "What Does the Yield Curve Tell us About GDP Growth?" *Journal of Econometrics* 131: 359–403.
- Barhoumi, K., S. Benk, R. Cristadoro, A. D. Reijer, A. Jakaitiene, P. Jelonek, A. Rua, G. Rünstler, K. Ruth, and C. V. Nieuwenhuyze. 2008. "Short-term Forecasting of GDP Using Large Monthly Datasets: A Pseudo Real-time Forecast Evaluation Exercise. European Central Bank." Occasional Paper Series No 84/April 2008.
- Barkoulas, J., C. F. Baum, and A. Chakraborty. 2003. "Nearest-neighbor Forecasts of U.S. Interest Rates." Boston College Working Papers in Economics, Number 313.
- Bessec, M., and O. Bouabdallah. 2005. "What Causes the Forecasting Failure of Markov-switching Models? A Monte Carlo Study." *Studies in Nonlinear Dynamics and Econometrics* 9: Article 6.
- Box, G. E. P., and G. M. Jenkins. 1970. *Time Series Analysis, Forecasting and Control*. San Francisco: Holden-Day.
- Brock, W. A., and C. L. Sayers. 1988. "Is The Business Cycle Characterized by Deterministic Chaos?" *Journal of Monetary Economics* 22: 71–90.
- Clark, T. E., and M. W. McCracken. 2001. "Tests of Equal Forecast Accuracy and Encompassing for Nested Models." *Journal of Econometrics* 105: 85–110.
- Clark, T. E., and M. W. McCracken. 2002. "Forecast-based Model Selection in the Presence of Structural Breaks." Federal Reserve Bank of Kansas City, RWP 02-05.
- Clark, T. E., and M. W. McCracken. 2004. *Evaluating Long-horizon Forecasts*. USA: University of Missouri-Columbia, Manuscript.
- Clemen, R. T., and J. B. Guerard. 1989. "Econometric GNP Forecasts: Incremental Information Relative to Naive Extrapolation." *International Journal of Forecasting* 5: 417–426.
- Clemen, R. T., and R. L. Winkler. 1986. "Combining Economic Forecasts." *Journal of Business & Economic Statistics* 4: 39–46.
- Clements, M. P., and D. F. Hendry. 1998. "Forecasting Economic Processes." *International Journal of Forecasting* 14: 111–131.
- 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: 123–141.
- Clements, M. P., P. H. Franses, J. Smith, and D. van Dijk. 2003. "On SETAR Non-linearity and Forecasting." *Journal of Forecasting* 22: 359–375.
- Clements, M. P., P. H. Franses, and N. R. Swanson. 2004. "Forecasting Economic and Financial Time-Series with Non-linear models." *International Journal of Forecasting* 20: 169–183.
- Cleveland, W. S., S. J. Devlin, and E. Grosse. 1988. "Regression by Local Fitting: Methods, Properties and Computational Algorithms." *Journal of Econometrics* 37: 87–114.
- 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 J. A. Nason. 1990. "Nonparametric Exchange Rate Prediction." *Journal of International Economics* 28: 315–332.
- Diebold, F. X., and R. S. Mariano. 1995. "Comparing Predictive Accuracy." *Journal of Economics and Business Statistics* 13: 253–263.
- Dueker, M., and C. J. Neely. 2007. "Can Markov Switching Models Predict Excess Foreign Exchange Returns?." *Journal of Banking & Finance* 31: 279–296.
- Engel, J., D. Haugh, and A. Pagan. 2005. "Some Methods for Assessing the Need for Non-linear Models in Business Cycle Analysis." *International Journal of Forecasting* 21: 651–662.
- Fernandez-Rodriguez, F., S. Sosvilla-Rivero, and J. Andrada-Felix. 1999. "Exchange-rate Forecasts with Simultaneous Nearest-neighbor Methods: Evidence from the EMS." *International Journal of Forecasting* 15: 383–392.
- Galbraith, J. W. 2003. "Content Horizons for Univariate Time-series Forecasts." *International Journal of Forecasting* 19: 43–55.

- Gençay, R., and F. Selçuk. 1998. "A Visual Goodness-of-Fit Test for Econometric Models." *Studies in Nonlinear Dynamics & Econometrics* 3: 157–167.
- Gneiting, T., and A. E. Raftery. 2007. "Strictly Proper Scoring Rules, Prediction, and Estimation." *Journal of the American Statistical Association* 102: 359–378.
- Gneiting, T., F. Balabdaoui, and A. E. Raftery. 2007. "Probabilistic Forecasts, Calibration and Sharpness." *Journal of the Royal Statistical Society* 69: 243–268.
- Gneiting, T., and T. L. Thorarinsdottir. 2010. "Predicting Inflation: Professional Experts Versus No-Change Forecasts." arXiv:1010.2318v1.
- Gneiting, T., and R. Ranjan. 2011a. "Comparing Density Forecasts Using Threshold and Quantile-Weighted Scoring Rules." *Journal of Business & Economic Statistics* 29: 411–422.
- Gneiting, T., and R. Ranjan. 2011b. "Combining Predictive Distributions." arXiv:1106.1638v1.
- Goodwin, T. H. 1993. "Business-Cycle Analysis with a Markov-Switching Model." *Journal of Business & Economic Statistics* 11: 331–339.
- Granger, C. W. J., and T. Teräsvirta. 1992. *Modelling Nonlinear Economic Relationships*. USA: Oxford University Press.
- Gupta, R., and M. P. Modise. 2011. *Macroeconomic Variables and South African Stock Return Predictability*. University of Pretoria, Working Paper: 2011–07.
- Hamilton, J. D. 1989. "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle." *Econometrica* 57: 357–384.
- Hansen, B. E. 1992. "The Likelihood Ratio Test under Nonstandard Conditions: Testing the Markov Switching Model of GNP." *Journal of Applied Econometrics* 7: S61–S82.
- Hansen, B. E. 1996a. "Erratum: The Likelihood Ratio Test Under Nonstandard Conditions: Testing the Markov Switching Model of GNP." *Journal of Applied Econometrics* 11: 195–198.
- Hansen, B. E. 1996b. "Inference When a Nuisance Parameter is Not Identified under the Null Hypothesis." *Econometrica* 64: 413–430.
- Hansen, B. E. 2001. "The New Econometrics of Structural Change: Dating Breaks in U.S. Labor Productivity." *The Journal of Economic Perspectives* 15: 117–128.
- Härdle, W. 1990. "Applied Nonparametric Regression." *Econometric Society Monographs*. England: Cambridge University Press.
- Harvey, A., and V. Oryshchenko. 2012. "Kernel Density Estimation for Time Series Data." *International Journal of Forecasting* 28: 3–14.
- Harvey, D. I., S. J. Leybourne, and P. Newbold. 1998. "Tests for Forecast Encompassing." *Journal of Business & Economic Statistics* 16: 254–259.
- Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. USA: Springer Series in Statistics.
- Jeon, J., and J. W. Taylor. 2012. "Using Conditional Kernel Density Estimation for Wind Power Density Forecasting." *Journal of the American Statistical Association* 107: 66–79.
- Kapetanios, G., and E. Tzavalis. 2005. "Nonlinear Modelling of Autoregressive Structural Breaks in a US Diffusion Index Dataset." Bank of England, Working Paper no. 537.
- Kapetanios, G., V. Labhard, and S. Price. 2007. "Forecast Combination and the Bank of England's Suite of Statistical Forecasting Models." Bank of England, Working Paper no. 323.
- Kolmogorov, A. N. 1933. "On the Empirical Determination of a Distribution Function." *Giornale dell'Istituto Italiano degli Attuari* 4: 83–91.
- Krolzig, H-M. 1998. "Econometric Modelling of Markov-switching Vector Autoregressions using MS VAR for Ox." Discussion paper, University of Oxford.
- Krolzig, H-M. 2001. "Business Cycle Measurement in the Presence of Structural Change: International Evidence." *International Journal of Forecasting* 17: 349–368.
- Little, M. A., P. E. McSharry, and J. W. Taylor. 2009. "Generalised linear Models for Site-specific Density Forecasting of UK Daily Rainfall." *Monthly Weather Review* 137: 1031–1047.
- Lorenz, E. N. 1969. "Atmospheric Predictability as Revealed by Naturally Occurring Analogues." *Journal of the Atmospheric Sciences* 26: 636–646.
- Loungani, P. 2001. "How Accurate are Private Sector Forecasts? Cross-Country Evidence from Consensus Forecasts of Output Growth." *International Journal of Forecasting* 17: 419–432.
- Makridakis, S., and R. L. Winkler. 1983. "Averages of Forecasts: Some Empirical Results." *Management Science* 29: 987–996.
- Marcellino, M., J. H. Stock, and M. W. Watson. 2003. "Macroeconomic Forecasting in the Euro Area: Country Specific versus Area-wide Information." *European Economic Review* 47: 1–18.
- Marcellino, M., J. H. Stock, and M. W. Watson. 2006. "A Comparison of Direct and Iterated Multistep AR Methods for Forecasting Macroeconomic Time Series." *Journal of Econometrics* 135: 499–526.
- McCracken, M. W. 2007. "Asymptotics for Out-of-sample Tests of Granger Causality." *Journal of Econometrics* 140 (2): 719–752.
- Morley, J., and J. Piger. 2004. "The Importance of Nonlinearity in Reproducing Business Cycle Features." Federal Reserve Bank of St. Louis, Working Paper 2004-032A.

- Öcal, N. 2000. "Nonlinear Models for U.K. Macroeconomic Time Series." *Studies in Nonlinear Dynamics & Econometrics* 4: 123–135.
- Orrell, D., and P. E. McSharry. 2009. "Systems Economics: Overcoming the Pitfalls of Forecasting Models via a Multidisciplinary Approach." *International Journal of Forecasting* 25: 734–743.
- Paparella, F., A. Provenzale, L. A. Smith, C. Taricco and R. Vio. 1997. "Local Random Analogue Prediction of Nonlinear Processes." *Physics Letters A* 235: 233–240.
- 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.
- Potter, S. M. 1995. "A Nonlinear Approach to US GNP." *Journal of Applied Econometrics* 10: 109–125.
- Rapach, D. E., M. E. Wohar, and J. Rangvid. 2005. "Macro Variables and International Stock Return Predictability." *International Journal of Forecasting* 21: 137–166.
- Smirnov, N. V. 1939. "On the Estimation of the Discrepancy Between Empirical Curves of Distribution for Two Independent Samples." *Bulletin of Moscow University* 2: 3–16.
- Stekler, H. O. 2007. "The Future of Macroeconomic Forecasting: Understanding the Forecasting Process." *International Journal of Forecasting* 23: 237–248.
- Stock, J. H., and M. W. Watson. 1999. "A Comparison of Linear and Nonlinear Univariate Models for Forecasting Macroeconomic Time Series." In *Ch. 1 in Cointegration, Causality and Forecasting: A Festschrift for Clive*, edited by W. J. Granger, R. Engle, and H. White, 1–44. Oxford: Oxford University Press.
- Swanson, N. R., and H. White. 1995. "A Model Selection Approach to Assessing the Information in the Term Structure Using Linear Models and Artificial Neural Networks." *Journal of Business and Economic Statistics* 13: 265–275.
- Swanson, N. R., and H. White. 1997. "A Model Selection Approach to Real-time Macroeconomic Forecasting Using Linear Models and Artificial Neural Networks." *Review of Economics and Statistics* 79: 540–550.
- Teräsvirta, T. 1995. "Modelling Nonlinearity in U.S. Gross National Product 1889–1987." *Empirical Economics* 20: 577–597.
- Teräsvirta, T., and H. M. Anderson. 1992. "Characterizing Nonlinearities in Business Cycles Using Smooth Transition Autoregressive Models." *Journal of Applied Econometrics* 7: S119–S136.
- Tiao, G. C., and R. S. Tsay. 1994. "Some Advances in Non-linear and Adaptive Modelling in Time-series." *Journal of Forecasting* 13: 109–131.
- West, K. D. 1996. "Asymptotic Inference about Predictive Ability." *Econometrica* 64: 1067–1084.
- Yu, K., and M. C. Jones. 1998. "Local Linear Quantile Regression." *Journal of the American Statistical Society* 93: 228–237.
- Zheng, I. Y., and J. Rossiter. 2006. "Using Monthly Indicators to Predict Quarterly GDP." Bank of Canada Working, Paper No. 2006–26.