2010 | 18

Working Paper

Research Department

Oil and US GDP: A real-time out-of-sample examination

By Francesco Ravazzolo and Philip Rothman

NB NORGES BANK

Working papers fra Norges Bank, fra 1992/1 til 2009/2 kan bestilles over e-post:
servicesenter@norges-bank.no
eller ved henvendelse til:
Norges Bank, Abonnementsservice
Postboks 1179 Sentrum
0107 Oslo
Telefon 22 31 63 83, Telefaks 22 41 31 05

Fra 1999 og fremover er publikasjonene tilgjengelig på www.norges-bank.no

Working papers inneholder forskningsarbeider og utredninger som vanligvis ikke har fått sin endelige form.
Hensikten er blant annet at forfatteren kan motta kommentarer fra kolleger og andre interesserte.
Synspunkter og konklusjoner i arbeidene står for forfatternes regning.

Working papers from Norges Bank, from 1992/1 to 2009/2 can be ordered by e-mail:
servicesenter@norges-bank.no
or from Norges Bank, Subscription service,
P.O.Box. 1179 Sentrum
N-0107Oslo, Norway.
Tel. +47 22 31 63 83, Fax. +47 22 41 31 05

Working papers from 1999 onwards are available on www.norges-bank.no

Norges Bank's working papers present research projects and reports (not usually in their final form)
and are intended inter alia to enable the author to benefit from the comments of colleagues and other interested
parties. Views and conclusions expressed in working papers are the responsibility of the authors alone.

ISSN 1502-8143 (online)
ISBN 978-82-7553-571-7 (online)

Oil and US GDP: A Real-Time Out-of-Sample Examination

Francesco Ravazzolo*
Norges Bank
Philip Rothman†
East Carolina University‡

September 15, 2010

Abstract

We study the real-time Granger-causal relationship between crude oil prices and US GDP growth through a simulated out-of-sample (OOS) forecasting exercise; we also provide strong evidence of in-sample predictability from oil prices to GDP. Comparing our benchmark model "without oil" against alternatives "with oil," we strongly reject the null hypothesis of no OOS predictability from oil prices to GDP via our point forecast comparisons from the mid-1980s through the Great Recession. Further analysis shows that these results may be due to our oil price measures serving as proxies for a recently developed measure of global real economic activity omitted from the alternatives to the benchmark forecasting models in which we only use lags of GDP growth. By way of density forecast OOS comparisons, we find evidence of such oil price predictability for GDP for our full 1970-2009 OOS period. Examination of the density forecasts reveals a massive increase in forecast uncertainty following the 1973 post-Yom Kippur War crude oil price increases.

*Contact: Norges Bank, Bankplassen 2, P.O. Box 1179 Sentrum, 0107 Oslo, Norway, Phone No: +47 22 31 61 72, e-mail: Francesco.ravazzolo@norges-bank.no

†Corresponding author: Brewster A-424, Department of Economics East, Carolina University, Greenville, NC 27858-4353, USA, Phone No: (252) 328-6151, e-mail: rothmanp@ecu.edu

‡We thank Christiane Baumeister, Hilde Bjørnland, Efrem Castelnuovo, Vincent Labhard, Lutz Kilian, Mike McCracken, Ken West, and seminar participants at the Norges Bank conference on 'Recent Developments in the Econometrics of Macroeconomics and Finance' and the University of Oslo conference on 'Empirical Business Cycle Modelling and Policy in the Aftermath of the Financial crisis' for helpful comments. We also thank Lutz Kilian for kindly providing his nominal shipping index series. The views expressed in this paper are our own and do not necessarily reflect those of Norges Bank.

1 Introduction

The goal of this paper is to investigate the predictive relationship between oil prices and US GDP by way of a simulated out-of-sample (OOS) forecasting study. We are motivated to do so by the set of mixed and conflicting results reported by leading scholars in the extensive primarily in-sample (IS) literature which arguably flows from the seminal paper of Hamilton (1983).

Hamilton (1983) shows that large crude oil price increases systematically preceded US recessions from the early post-World War II period to the beginning of the 1980s, such that the positive correlation between oil prices and the US business cycle that was apparent following the post-Yom Kippur War OAPEC embargo, the fall of the Shah of Iran in 1979, and the outbreak of the Iran-Iraq War in 1980 was not a relatively new phenomenon.[1] He also finds that crude oil prices Granger-cause real output over the full 1948-1980 sample period as well as the 1948-1972 and 1973-1980 subsamples. Further, the general failure of the macroeconomic variables considered to Granger-cause oil prices, along with historical and institutional details of the post-World War II oil market studied in Hamilton (1985), leads him to conclude that the crude oil price changes observed in this era were exogenous relative to general business cycle fluctuations. Figure 1 shows a time series plot of a benchmark crude oil price measure and the NBER recession dates from 1955Q1 to 2009Q4.

The data Hamilton (1983) uses end in 1980. With extended data roughly running to the middle of the 1990s, Hooker (1996) establishes that, via the linear time series approach employed by Hamilton (1983), crude oil prices no longer Granger-cause real output. Accordingly, he challenges the then increasing use in the macroeconomics literature of oil prices as instrumental variables at the same time that they appear to play a less important role across the business cycle. In response, Hamilton (1996) demonstrates that use of a nonlinear transformation of oil prices he labels the "net oil price increase" (NOPI), in place of the raw oil price growth rate, produces a Granger-causal relationship from oil prices to output when the more recent data are included.

Subsequent to this exchange between Hooker and Hamilton, several papers document a weakening of the predictive relationship from oil prices to the macroeconomy, including Bernanke, Gertler, and Watson (1997), Edelstein and Kilian (2009), and Blanchard and Galí (2010). Hamilton and Herrera (2004), however, show that the results in Bernanke et al. (1997) are not robust to use of a credible alternative longer lag specification. Also, Hamilton (2009) computes OOS forecasts for several of Edelstein and Kilian's (2009) estimated models over, roughly, the first year of the Great Recession and finds that energy prices explain a large fraction of the forecast errors. In the same paper Hamilton notes that the Blanchard and Galí (2010) structural VAR esti-

[1] While a large literature also argues that these oil price movements, generally interpreted as being produced by negative supply shocks, were a fundamental factor in generating the stagflation of the 1970s, an important and growing line of research, for example, Barsky and Kilian (2002), Barsky and Kilian (2004), Baumeister and Peersman (2008), and Kilian (2010), questions the causal role of oil prices for that stagflation and the association of these oil price fluctuations with supply shocks. For an international comparison, see Bjørnland (2000).

mates imply the US 1981-82 recession would have been, counterintuitively, deeper in the absence of the crude oil price shocks that preceded it, and additionally suggest the 1990-91 US recession might have been averted if oil prices had not shot up following Iraq's 1990 invasion of Kuwait.[2] Further, applying the novel random field approach of Hamilton (2001), Hamilton (2003) presents evidence suggesting that the predictive relationship from oil prices to GDP growth continues to be strong, and argues that measures of oil supply disruptions can serve as useful exogenous instruments in instrumental variables regressions.[3]

Inoue and Kilian (2004) examine the question of IS versus OOS testing of predictability, motivated by the finding that positive IS evidence of predictability is often not associated with OOS predictability. They argue that Ashley, Granger, and Schmalensee's (1980) claim that IS inference without OOS verification is likely to be spurious, with an OOS approach inherently involving less overfitting, is not compelling since there is ample opportunity for the researcher to data mine in a simulated OOS study, and because data snooping adjustments can be made to both IS and OOS tests. Under such adjustments, they show that IS tests are likely to have greater power than OOS tests. The OOS point and density forecasting findings we report in this paper are consistent with this result.

In light of Inoue and Kilian's (2004) analysis, using the same models it would be surprising to find strong OOS predictability from crude oil prices to US GDP in the absence of IS predictability. Accordingly, for models we further explain below, in Figure 2 we present such IS evidence on the predictability of oil prices for US GDP via a sequence of rolling estimation windows of post-World War II data. In each graph comparisons are made against a benchmark model with no oil price measure included and alternatives which do include such oil price data. For every estimation window considered, the benchmark model generates a higher value of the Akaike Information Criterion (AIC) and a lower marginal likelihood.

Following many precedents in the literature, the models with which we generate sequences of OOS forecasts are estimated on vintages of real-time data.[4] The importance of using such data, as opposed to revised data, is at least twofold. First, if the models producing the sequence of forecasts in the OOS study were estimated with the most recent vintage available at the time the research is carried out, this would be equivalent to assuming that economic agents have information that is, in fact, unavailable to them when forecasting future economic activity. Second, use of revised data can give a misleading impression of the relative OOS forecasting performance of the alternative

[2]Since the analysis in Blanchard and Riggi (2009) relies on Blanchard and Galí's (2010) structural VAR estimates, these points Hamilton (2009) raises also apply to Blanchard and Riggi (2009).

[3]Using several econometric specifications, though, Kilian (2008) can not reject the null hypothesis that the instruments suggested by Hamilton (2003) are weak in the sense of Cragg and Donald (1993) and Stock and Yogo (2005).

[4]Croushore and Stark (2003) provide a useful discussion of real-time versus revised data.

models considered.[5]

We carry out our OOS predictability analysis with both point and density forecasts. Our key results from the point forecast comparisons are as follows. We find very strong statistically significant predictability from oil prices to GDP from the 1980s through the Great Recession. Further examination suggests that some of these results may be due to the oil price measures we use proxying for variables omitted from the alternatives to the benchmark, such as Kilian's (2009) real global economic activity measure. Our density forecast comparisons establish OOS predictability from oil prices to GDP growth for the full 1970-2009 OOS period.

Bachmeier, Li, and Liu (2008) also study the OOS predictability from oil prices to GDP growth, reaching the strong conclusion that there is no such predictability. We note that they do so, however, with revised data, such that the above caveats arguably apply. In addition, they only consider point forecast comparisons.

Carlton (2010) carries out a considerably smaller OOS predictability exercise for oil prices and US GDP than we do, but she also uses real-time data. Her OOS period is restricted to a subset of the 2000s, and she reports positive point forecast evidence of predictability from oil prices to GDP growth. Density forecasts are not included as part of her OOS analysis.

The paper proceeds as follows. In Section 2 we discuss our forecasting models and evaluation criteria, and present our results in Section 3. We conclude in Section 4.

2 Forecasting GDP with Oil Prices

We generate $h-$step ahead OOS forecasts, for $h = 1$ and $h = 4$, of quarterly US GDP growth rates using real-time vintage j and compute forecast errors with the first release value of the US GDP (from vintage $j + 1$ in the $h = 1$ case and from vintage $j + 4$ in the $h = 4$ case). For all the models we use direct forecasting for the $h-$step ahead forecasts, such that we do not need to employ multi-equation systems to produce our forecasts.

We use data for US GDP, import prices, the consumer price index (CPI), and the personal consumption expenditures deflator from real-time vintages downloaded from the Philadelphia Federal Reserve Bank's real-time database from 1955Q1 to 2009Q4; the first vintage covers 1955Q1-1969Q4, and the last vintage runs from 1955Q1 to 2009Q4. The main crude oil price measure we focus on is the monthly West Texas Intermediate spot oil price, downloaded from Dow Jones, and compute the arithmetic averages across each quarter to produce our quarterly oil price series; we check the robustness of our results with both the Brent and Dubai crude spot oil price series, downloaded from Bloomberg. The interest rate variables we use are the 10-year Treasury Bond, 3-month Treasury Bill, Federal Funds, Aaa, and Baa rates downloaded from the FRED database at the Federal

[5]This is the case, for example, for the OOS time series forecasts Faust and Wright (2009) analyze.

Reserve of Saint Louis. As a measure of global economic activity, we use the nominal shipping series from Kilian (2009) and employ the transformations he uses to compute a real detrended shipping series for each IS period.

2.1 Forecasting Models

A standard benchmark to forecast GDP growth is an autoregressive model of order p.

$$y_t = \alpha + \sum_{i=1}^{p} \beta_i y_{t-i} + \sigma \epsilon_t, \tag{1}$$

where $\epsilon_t \sim N(0,1)$. In the oil and the macroeconomy literature, the lag order p for the estimated models is often set equal to 4; see, for example, Hamilton (2003).[6] We consider this case and also identify p according to the AIC; in the second case we refer to the model as $\text{AR}(p)_{\text{AIC}}$. Bayesian inference is applied with weak informative conjugate priors to restrict regression coefficients to zero.[7] The model is estimated and point and density forecasts are produced via a sequence of 15-year moving windows; the first moving window IS period is 1955Q1-1969Q4. As Swanson (1998) emphasizes, use of a fixed-length moving window approach allows the data generating process to evolve over time. Our decision to adopt this approach is motivated by the evidence of structural instability in US macroeconomic time series reported by Stock and Watson (1996), Sensier and van Dijk (2004), and others.

Next we extend the AR(p) benchmark with an oil price measure:

$$y_t = \alpha + \sum_{i=1}^{p} \beta_i y_{t-i} + \sum_{i=1}^{p} \delta_i oil_{t-i} + \sigma \epsilon_t, \tag{2}$$

where $\epsilon_t \sim N(0,1)$ and oil_t is the oil price measure at time t. We use two such measures: the oil price growth rate, $oil_t = \ln(p_t) - \ln(p_{t-1})$ where p_t is the West Texas Intermediate spot oil price in quarter t; and the NOPI measure proposed by Hamilton (1996), $oil_t = \max[(\ln(p_t) - \max[\ln(p_{t-1}), .., \ln(p_{t-4})]), 0]$.[8] Given our two schemes for the lag length p, this leads to four alternatives to the AR(4) and $\text{AR}(p)_{\text{AIC}}$ benchmarks: $\text{ARX}(4)^o$, $\text{ARX}(4)^n$,

[6] On identifying the lag order for time series models, Cochrane (2005, p. 26) notes, "we tend to throw in a few extra lags just to be sure and leave it at that."

[7] We use a normal inverted gamma prior with means for α and the β_i equal to zero and variances equal to 100. The predictive densities are Student$-t$ distributed, and the means of densities are used as point forecasts. See, for example, Koop (2003) for details.

[8] While Hamilton (1996) computes the NOPI using the four most recent quarterly lags, Hamilton (2003), noting that oil price increases in 1999 had only recovered from the decreases of the preceding two years, incorporates a 3-year horizon. In subsequent work, for example, Hamilton (2009, 2010), he also uses a 3-year horizon. We find that the OOS predictability results we present are robust to use of a 3-year horizon; $p-$values change only in the third decimal place and beyond.

$\text{ARX}(p)^o_{\text{AIC}}$, and $\text{ARX}(p)^n_{\text{AIC}}$, where the superscripts 'o' and 'n' indicate, respectively, that the ARX alternative model includes p lags of the crude oil price growth rate and the NOPI measure.

It is possible that forecast improvement obtained by adding an oil price measure to the AR(p) benchmark, or failure to achieve such forecast improvement, is driven by an omitted variable in models (1) and (2). To examine this question, we also consider the following benchmark model:

$$y_t = \alpha + \sum_{i=1}^{p} \beta_i y_{t-i} + \sum_{i=1}^{p} \delta_i z_{t-i} + \sigma\epsilon_t, \tag{3}$$

where $\epsilon_t \sim N(0,1)$ and z_t is a non-oil price macro variable. We refer to the benchmarks from (3) as $\text{ARX}(4)^z$ and $\text{ARX}(p)^z_{\text{AIC}}$ for each of the macro variables. As an alternative to these benchmarks, we add an oil price measure:

$$y_t = \alpha + \sum_{i=1}^{p} \beta_i y_{t-i} + \sum_{i=1}^{p} \delta_i z_{t-i} + \sum_{i=1}^{p} \gamma_i oil_{t-i} + \sigma\epsilon_t, \tag{4}$$

where $\epsilon_t \sim N(0,1)$. We refer to these alternatives as $\text{ARX}(4)^{z,o}$, $\text{ARX}(4)^{z,n}$, $\text{ARX}(p)^{z,o}_{AIC}$, and $\text{ARX}(p)^{z,n}_{AIC}$.

To determine which macro variables z_t to include in forecast comparisons between models (3) and (4), we first compare point forecasts using the AR(4) and $\text{AR}(p)_{\text{AIC}}$ benchmarks against, respectively, $\text{ARX}(4)^z$ and $\text{ARX}(p)^z_{\text{AIC}}$ alternatives for the following macro variables: growth rates of the import price deflator, personal consumption expenditures deflator, and nominal shipping freight index of Kilian (2009), the linear detrended real shipping freight index of Kilian (2009), the 3-month T-Bill rate, the 3-month T-Bill/fed funds, 10-year T-Bond/three-month T-Bill, and Moody's Baa/Aaa spreads, and a macro "factor" computed as the first principal component of the preceding variables. Consideration of these variables is based upon a large literature, including Estrella and Hardouvelis (1991), Hooker (1996), Stock and Watson (1999), Wright (2006), and others, as well as our need to use real-time data. Using the tests described below, we find evidence of OOS predictability from z_t to GDP growth for only four of these nine macro variables, the growth rates of the import price and personal consumption expenditures deflators, the linear detrended real shipping freight index, and the macro factor.[9] Accordingly, we use these four variables in our OOS comparisons between models (3) and (4).

It may very well be the case that use of fixed-length moving windows with linear models is not sufficiently flexible to capture the structural change in US GDP dynamics over the period we study.[10] In an attempt to allow for greater flexibility, we note that we also reformulated the models

[9] Full details are available upon request.

[10] The models are linear in the parameters, but when oil_t is the NOPI measure, there is a nonlinear relationship between oil prices and GDP growth.

discussed above with time-varying parameters. In particular, we introduced time instability via breaks in model parameters as in Ravazzolo and Vahey (2010), where shifts are determined by an unobserved stochastic process; the model nests conventional two-state Markov-switching models pioneered by Hamilton (1989), but allows for considerably more general behavior. However, we found that the OOS forecasts generated by this nonlinear approach are rather strongly dominated by those obtained with linear models estimated over fixed-length moving windows. Accordingly, we do not include discussion of these time-varying forecasts below.[11]

2.2 Forecast Evaluation

To examine the predictive power of crude oil prices for GDP growth, we use evaluation statistics for point and density forecasts previously proposed in literature. We compare point forecasts in terms of mean square prediction errors (MSPEs), but for the alternatives to the benchmarks we use "adjusted" MSPEs, where the MSPE adjustment is made as per Clark and West (2007) (hereafter CW), for different models and different OOS periods. Under the null hypothesis that the parsimonious benchmark model is the true DGP, use of estimated non-benchmark models (which nest the benchmark) induces noise into OOS forecasts by way of estimation of parameters with zero population means.[12] The CW MSPE adjustment is an attempt to reduce the role of such noise when making OOS forecasting comparisons for nested models. We test the null hypothesis that the nested benchmark model without an oil price measure has the lower MSPE by way of two tests: (1) the CW test, which compares MSPEs between the benchmark and a single alternative; and (2) the Hubrich and West (2010) (hereafter HW) test, which simultaneously compares MSPEs between the benchmark and a small set of alternatives as a check against data snooping.[13]

To implement the CW test, we compute:

$$\widehat{f}_{t+h} = (y_{t+h} - \widehat{y}_{1,t+h})^2 - [(y_{t+h} - \widehat{y}_{2,t+h})^2 - (\widehat{y}_{1,t+h} - \widehat{y}_{2,t+h})^2], \quad t = N, ..., T-h, \tag{5}$$

where y_{t+h} is the realization of the variable of interest at time $t+h$, $\widehat{y}_{i,t+h}$, $i = 1, 2$, are the h-step

[11] For this class of models there is very little evidence of OOS predictability from oil prices to GDP growth. We believe this reflects the time-varying benchmark's ability to compensate for possible misspecification by allowing for robust time variation in the intercept, the autoregressive coefficients, and the variance of the stochastic error term. Full details are available upon request.

[12] As per Inoue and Kilian (2004), the CW test is a test of no predictability in population. For the case in which there are what they call "weak" additional predictors in the nesting model, Clark and McCracken (2009) develop a test of equal OOS predictability. For both the CW and Clark and McCracken (2009) tests, however, insofar as the null and alternative models under consideration are chosen by data-drive model selection procedures, we speculate that a variant of the "impossibility" theorem in Leeb and Pötscher (2005) might imply that the true distribution of the test statistics might be unknowable even by standard simulation methods.

[13] "Small" in this context means that the number of alternative models is significantly lower than the sample size of the estimation window. We note that HW's simulation study shows that their tests have greater power than White's (2000) "reality check" for the cases considered.

ahead point forecasts conditional on the information at time t from model 1 (the parsimonious nested benchmark) and from model 2 (the larger one), N is the last IS observation, and T is the last OOS observation. The CW test for equal MSPE is carried out by regressing $\widehat{f}_{t+h}$ on a constant and running a $t-$test for the null hypothesis that the constant is less than equal to zero; for $h > 1$, a heteroskedastic and autocorrelation consistent (HAC) standard error is used. Failure to reject the null indicates that model 2 reduces to model 1 at the given significance level.

HW provide two tests, a "max MSPE-adj $t-$statistic," for which the maximum is computed across the set of m CW $t-$statistics, where m is the number of alternatives to the benchmark, and a χ^2 variant. We use the max $t-$statistic test for two reasons. First, it has higher power than the χ^2 test. Second, we found that the χ^2 test can provide misleading inference for the following case: when some of the CW $t-$statistics are large and negative (such that there are not rejections of the one-sided null), the χ^2 can be spuriously large. Below, "HW test" refers to the max $t-$statistic test.

To run the HW test with m alternatives to the benchmark, an $m \times m$ matrix $\hat{\Omega}$ is constructed, where the i, j element of $\hat{\Omega}$ is the sample correlation between the CW $t-$statistics for alternatives i and j. Then the distribution of the max $t-$statistic is estimated by taking a large number of draws from an $N(0, \hat{\Omega})$ distribution, in which the maximum of the random vector is stored from each draw; following HW, we take 50,000 draws. The $p-$value of the observed max $t-$statistic is computed from this empirical distribution of maxima.

Density forecasts are compared using a test based on the Kullback-Leibler information criterion (KLIC) distance measure, which focuses on the difference between two log scores, where the log score of a density forecast for OOS observation $t+h$ is computed as the log of the density forecast for that observation. Amisano and Giacomini (2007) (hereafter AG) derive a KLIC test for equal predictive density accuracy for the case of two nested models estimated using fixed size IS rolling windows of data. For each OOS observation $t+h$, define:

$$WLR_{t+h} = w(y^{std}_{t+h})(\ln(g_1(y_{t+h}|I_t)) - \ln(g_2(y_{t+h}|I_t))), \tag{6}$$

where g_1 and g_2 are, respectively, the scores for the benchmark model 1 and the alternative model 2, $w(\cdot)$ is a weighting function, and y^{std}_{t+h} is the realization y_{t+h} standardized using the IS data with which the density forecasts are estimated. The AG test statistic is computed as:

$$t_n = \frac{\overline{WLR}_n}{\widehat{\sigma}_{t+h}/\sqrt{n}}, \tag{7}$$

where $n = T-h-N$, $\overline{WLR}_n = n^{-1}\sum_N^{T-h} WLR_{t+h}$, and $\widehat{\sigma}_{t+h}$ is the square root of a HAC estimator of the asymptotic variance $\sigma_n^2 \text{var}(\sqrt{n}\ \overline{WLR}_n)$. In reporting our results below, we use the "center of distribution" weighting function of AG, which ignores the effects of any possible outliers.[14] Below "AG test" refers to the test computed using the center of distribution weighting function.

3 Results

We report OOS forecasting results for the 1970Q1 to 2009Q4 period as well as for a set of six subsamples, with each starting five years later than the previous one but also ending in 2009Q4, i.e. 1975Q1-2009Q4, 1980Q1-2009Q4, ..., 2000Q1-2009Q4. Through consideration of these subsamples we are able to obtain an assessment about whether the oil price predictability for US GDP has changed over time, and in particular for specific periods such as the oil crises in the 1970's, the reversal of oil prices in the mid 1980s and subsequent relatively low oil price volatility regime through most of the 1990s, and the eventual high oil price volatility period after 2000.

3.1 Point Forecasts

Table 1 presents results for tests of equal OOS forecast accuracy at the $h = 1$ and $h = 4$ horizons for the AR(4) and $\text{AR}(p)_{\text{AIC}}$ benchmarks. For each benchmark model, the MSPE is reported, whereas for the alternatives to the benchmark the ratio of the model's adjusted MSPE to the benchmark MSPE is reported. At $h = 1$, addition of an oil price measure to the AR(4) benchmark in forecasting GDP growth generates a reduction in MSPE in twenty-six out of twenty-eight cases. The MSPE reduction produced with the $\text{ARX}(4)^n$ model is significant at conventional levels for the full 1970-2009 OOS period via both the CW and HW tests. For the 1975-2009 subsample, however, the MSPE ratios are greater than 1 for both the $\text{ARX}(4)^o$ and $\text{ARX}(4)^n$ alternatives. Though the MSPE ratios are less than 1 for these models in the 1980-2009 subsample, the CW and HW $p-$values are above 0.10. For the last four subsamples, 1985-2009 and onward, the CW $p-$values for these models are all less than 0.10; the rejections are stronger for the $\text{ARX}(4)^n$ forecasts and the HW $p-$values are all less than 0.05.

At the $h = 1$ forecast step, when the lag length p is selected by AIC the addition of the crude oil price growth rate to the AR(p) benchmark does not lead to statistically significant reductions in MSPE for any of the OOS periods. But when the NOPI measure is used, a similar pattern of results is obtained, with some exceptions, relative to setting $p = 4$ for all IS windows. The exceptions are as follows. First, the CW and HW $p-$values are considerably higher, both over 0.10, for the full 1970-2009 OOS period. Second, there are very strong rejections via both tests for the 1980-2009 subsample. The similarity is that, for the last four subsamples, the $p-$values for both the CW and HW tests are quite low.

[14] Our OOS density forecast comparisons are not strongly affected with use of the other three weighting functions AG provide.

The ARX(4)^o and ARX(4)^n results at the $h = 4$ forecast horizon for the most part mirror those at $h = 1$ by way of both the CW and HW tests. One difference is that, even though the ARX(4)^n generates a larger MSPE reduction at $h = 4$ relative to $h = 1$ for the full 1970-2009 OOS period, the CW and HW null hypotheses are not rejected at conventional significance levels. The other is that there is a marginally significant MSPE reduction, via the CW test but not the HW test, produced by the ARX(4)^o forecast for the 1980-2009 OOS subsample. When the lag length is selected by the AIC, at $h = 4$ use of either the crude oil price growth rate or NOPI measure generates a higher MSPE relative to the benchmark for the 1970-2009, 1975-2009, and 1980-2009 OOS periods. The CW test $p-$values are very high for the ARX$(p)^o_{\text{AIC}}$ forecasts for the last four OOS subsamples at $h = 4$. In contrast, for each of these last four subsamples, the ARX$(p)^n_{\text{AIC}}$ forecasts lead to rejection of the CW test null at the 10% significance level; however, the HW test $p-$value is below 0.10 only for the 2000-2009 OOS period.

The results in Table 1 suggest considerable time variation in the point forecast predictability from crude oil prices to GDP growth over the OOS periods we consider. First, when the 1970s, and in most cases the early 1980s, are included in the OOS sample, there generally is no strong evidence of such predictability; the $p-$values for both the CW and HW tests are below 0.10 for only one out of twelve cases. Given the high volatility of oil prices in these years, we find these results surprising; we suggest an explanation in discussion of our density forecasts below. Second, from the mid-1980s, with the onset of the Great Moderation, through the Great Recession, there is very strong evidence of such predictability, with the evidence being marginally stronger at the $h = 1$ forecast horizon.

Table 2 presents results for OOS predictability tests in which the benchmark and alternative models are given by, respectively, equations (3) and (4).[15] The purpose of this set of tests is to help us investigate the possibility that the results reported in Table 1 are influenced by omission of a relevant variable from the models used. To help focus the discussion, Table 2 gives results only for the $h = 1$ forecast step. We first consider those cases in which the lag length p is fixed at 4, and believe they provide four results of interest. First, for the last three subsamples, 1990-2009, 1995-2009, and 2000-2009, the HW test $p-$value is greater than 0.10 in all twelve cases, suggesting that the positive oil price predictability results for these subsamples in Table 1 indeed may be due to our oil price variables proxying for some omitted variables. Second, for the 1985-2009 OOS period, the HW $p-$values are below 0.10 in three out of four cases; the exception is when the macro factor is added to the models. Third, though the oil price alternatives generally produce MSPE reductions relative to the benchmark when the macro factor is included, the CW and HW test $p-$values are below 0.10 in only one out of the twenty-one cases across all OOS periods. These

[15] The nominal shipping index series of Kilian (2009), with which we produce the associated real linear detrended series, begins in 1968Q1. Since we use 15-year estimation windows, the first OOS subsample we have available using this series is 1985Q1-2009Q4.

results are consistent with our macro factor being a parsimonious measure of key macroeconomic behavior missing from equations (3) and (4). Fourth, use of the ARX$(4)^i$ and ARX$(4)^c$ benchmarks leads to strong evidence of predictability from oil prices to GDP for both the 1975-2009 and 1980-2009, implying that our failure to find such evidence for these OOS subsamples in Table 1 may stem from omission of the import price deflator and personal consumption expenditures deflator growth rates.

Next we discuss the results in Table 2 for which the lag length p is selected by the AIC. Two key results are as follows. First, for the last four subsamples, the HW test $p-$values are greater than 0.10 when the import price deflator growth rate, personal consumption expenditures deflator growth rate, and linear detrended real shipping index are included in the benchmark. These results are consistent with what we obtain using these variables in the benchmark models when the lag length p is set equal to 4, and similarly suggest that our oil price predictability results in Table 1 may reflect omission of relevant variables. Second, in contrast to the results reported in the first section of Table 2, when the benchmark includes the macro factor the CW and HW test $p-$values are below 0.10 in thirteen out of the twenty-one cases across all OOS periods; all of these rejections at conventional significance levels occur in subsamples beginning in 1980 or later. These results suggest that omission of our macro factor variable does not play a role in generating the positive oil price predictability evidence in Table 1 for the middle to latter part of our OOS period. Further, we note that the low CW test $p-$values generated by the ARX$(p)^{f,o}_{AIC}$ forecasts for the last four subsamples contrast strongly with what we report for the ARX$(p)^{o}_{AIC}$ forecasts in Table 1, suggesting that the latter results may reflect omission of the macro factor from the ARX$(p)^{o}_{AIC}$ model.

As an additional check, we ran predictability tests in which we use models given by equations (2) and (4) as, respectively, the benchmark and alternative models. Such tests examine whether the macro variable z_t OOS Granger-causes GDP growth conditional on including an oil price measure in the benchmark. We are specifically interested in those subsamples for which Table 1 reports strong evidence of oil price predictability for GDP growth, and accordingly do not report a table of full results across all OOS subsamples. The main findings of interest as are follows. Adding the linear detrended real shipping index leads to low CW and HW test $p-$values for the last three subsamples against both the ARX$(4)^o$ and ARX$(4)^n$ benchmarks. On the other hand, use of the shipping index does not lead to significant MSPE reductions against the ARX$^n_{AIC}$ benchmark for any OOS subsample.[16] Further, for the other macro variables, we generally fail to find evidence of OOS predictability from z_t to GDP growth.

[16] Recall Table 1's result that, when the lag length is selected using the AIC, oil price predictability is found only via NOPI measure.

3.2 Density Forecasts

We next turn to discussion of our density forecast evidence on the OOS predictive power of oil prices for GDP. Table 3 reports log scores and AG test $p-$values for the AR(4) and AR$(p)_{\text{AIC}}$ benchmarks at both the $h = 1$ and $h = 4$ forecast horizons for the same OOS periods considered in the point forecast analysis. We note that higher scores indicate better performance; since all of the log scores in Table 3 are negative, values closer to zero indicate higher density forecast accuracy.

The first notable result is that, in all fifty-six cases, adding an oil price alternative to the AR(p) benchmark yields a higher log score. In contrast, in approximately twenty percent of the cases presented in Table 1, adding oil prices leads to a higher MSPE. Accordingly, by such metrics the density forecasts provide stronger evidence of oil price OOS predictability for GDP growth.

At $h = 1$, the log score improvement produced by the ARX$(4)^o$ forecasts is significant at the conventional levels for the last four OOS subsamples via the AG test. This set of results roughly mirrors the ARX$(4)^o$ CW and HW results in Table 1. Adding the NOPI measure to the AR(4) benchmark leads the AG test $p-$values to be below 0.05 for all seven OOS periods. In contrast, for two OOS subsamples, 1975-2009 and 1980-2009, the ARX$(4)^n$ CW and HW $p-$values in Table 1 are above 0.10.

With respect to the OOS subsamples for which adding an oil price measure leads to statistically significant forecast improvement over the AR$(p)_{\text{AIC}}$ benchmark, the AG test results at $h = 1$ exactly match those for the CW and HW tests in Table 1. Adding the crude oil price growth rate never leads to rejection of the null for any OOS period, and adding the NOPI measure leads to rejection at conventional significance levels for the last five subsamples.

At $h = 4$, the AG test has $p-$values above 0.10 for all OOS period for the ARX$(4)^o$ forecasts, such that, at this longer forecast step, the density forecasts provide far less statistically significant predictability from oil prices to GDP growth over the AR(4) benchmark relative to the CW and HW point forecast results in Table 1. On the other hand, via the AG test the ARX$(4)^n$ forecasts generate statistically significant log score increases for five OOS subsamples at $h = 4$, the last four as well as the full 1970-2009 OOS period.

The ARX$(p)^o_{\text{AIC}}$ log score increases at $h = 4$ are statistically significant for all OOS periods. This is a sharp contrast to the point forecast results in Table 1 for this model at the same forecast step. But the ARX$(p)^n_{\text{AIC}}$ log score increases at $h = 4$ are significant at conventional levels for only the last OOS subsample, 2000-2009.

The fan charts presented in Figures 3 and 4 allow us to examine the uncertainty associated with our forecasts. The left and right panels focus on, respectively, the full 1970-2009 and 2000-2009 OOS periods. One key motivation for providing these two sets of graphs for each class of models is that the 1973 post-Yom Kippur War crude oil price increases manifest themselves in the form of what can arguably be called an "explosion" of forecast uncertainty, such that there appears to be practically no forecast uncertainty afterwards. This is most pronounced for the ARX$(p)^o_{\text{AIC}}$, ARX(4)^o, and ARX(4)^n forecasts, but even for the ARX$(p)^n_{\mathrm{AIC}}$, the width of the fan chart is roughly twice that of the AR$(p)_{\mathrm{AIC}}$ benchmark in the early-to-mid 1970s. The right panel sets of fan charts clearly show, however, that there is considerable forecast uncertainty outside of this period. The 2000-2009 OOS subsample is also of interest, all else equal, since it was also a period of high oil price volatility. In the post-Lehman Brothers collapse period, there is a substantial increase in forecast uncertainty, and the increase is larger for the forecasts produced by adding an oil price measure to the benchmark. But our fan charts demonstrate that there is no similar explosion of forecast uncertainty associated with these oil price movements.

Borrowing from Blinder and Rudd (2009), perhaps the late 1973 oil price shocks indeed were *sui generis*, but in the sense of the subsequent massive increase in forecast uncertainty we document. In light of this finding, we speculate that it is a primary factor behind our general failure to find evidence of point forecast predictability from oil prices to GDP growth when the 1970s and early 1980s are included in the OOS period.

The 2000-2009 fan charts also provide graphical insight on the forecasting benefit of including crude oil prices, especially the NOPI measure, in a forecasting model of GDP growth during the depths of the Great Recession. For both the AR(4) and AR$(p)_{\mathrm{AIC}}$ benchmarks, the movement of actual GDP growth to the trough in 2008Q4 is considerably below the 5% percentiles of the density forecasts, whereas such behavior is not observed for the ARX(4)^n and ARX$(p)^o_{\mathrm{AIC}}$ models.

3.3 Additional Robustness Checks

In their critique of the IS oil prices and the macroeconomy literature, Barsky and Kilian (2002) argue that it is important to note may very well be feedback from GDP growth to crude oil prices. To help address this question for the OOS concerns of our paper, using the approaches described above we examined the evidence on the OOS predictability from GDP growth to oil prices. We do not detail these results here, but note our main finding that GDP growth is generally not Granger-causal for either the growth rate of crude oil prices or the NOPI measure across all of the OOS periods we consider.

Our discussion above focuses on results obtained using oil price measures computed from the West Texas Intermediate spot oil price. We also ran through our procedures using data on the Brent and Dubai spot oil price series, and generally obtained strongly similar results. Given the very high correlation between the growth rates of these series, this is not surprising from a statistical perspective. On the basis of standard arbitrage-based arguments, this is not unexpected on economic grounds.

Given the standard arguments in favor of doing so, we carry out our OOS forecasting exercise with use of real-time data. However, it is useful to know to what extent the results we obtain are affected by this decision. Using the last vintage of data, our point forecast results are roughly similar to what we find with real-time data, and our density forecasting results imply stronger evidence of predictability from oil prices to GDP; at the $h = 1$ forecast horizon, the $p-$values for the AG test are below 0.10 for every OOS subsample we consider against both the AR(4) and $\text{AR}(p)_{\text{AIC}}$ benchmarks.

In their study, Blanchard and Riggi (2009) do not include data past the end of 2007, arguing that inclusion of later data would bias their results in favor of oil prices since it is clear that non-oil price factors were the dominant causes of the sharp output drop in the Great Recession.[17] When we end our simulated OOS forecasts in 2007Q4 we find that part of our point forecast predictability evidence weakens considerably, while the density forecasts at $h = 1$ continue to imply that the NOPI measure is Granger-causal for GDP.[18]

4 Conclusions

We provide several useful results for the literature on the post-World War II question of the Granger-causal relationship between crude oil prices and US GDP growth. First, we show that quite strong evidence can be generated in favor of IS predictability from oil prices to GDP over the past forty years using standard model selection criteria and vintages of real-time data.

Our primary contribution is to examine the extent to which there is OOS forecasting evidence in favor of such predictability using real-time data. Via point forecasts, our key finding from bivariate models of the relationship between GDP growth and crude oil prices is that there is very strong evidence in favor of OOS oil price Granger causality for GDP from the mid-1980s through the end of the Great Recession; further analysis suggests that these findings may reflect omission of Kilian's (2009) real global economic activity measure from our bivariate model.

The density forecasts produce evidence of OOS predictability from oil prices to GDP growth when the 1970s and early 1980s are included in the OOS period. When data from the Great Recession are excluded, the short-horizon density forecasts using the NOPI measure generally dominate the non-oil price benchmark via the statistical test we employ, while the point forecast results show much less predictability from oil prices to GDP. Our density forecasts also show that our oil price alternative models generate a massive bout of forecast uncertainty following the late 1973 crude oil price increases; at no other point in the OOS period is there similar behavior in forecast uncertainty. Accordingly, we believe our analysis demonstrates the usefulness of not restricting attention to the first moment of the estimated probability distribution of future values of GDP in OOS comparisons of models with and without oil prices.

[17]Hamilton (2009) argues that, absent the large oil price increases which preceded it, the 2007Q4-2008Q3 period would not have been included in the most recent recession by the NBER Business Cycle Dating Committee; Blinder (2009) agrees with this claim.

[18]Ending the OOS forecasts in 2008Q4 restores much of the point forecast predictability evidence we obtain when extending the OOS forecasts out to 2009Q4.

We attempt to account for a possibly evolving data generating process through our use of linear in parameters models estimated with fixed-length moving windows. While this may be insufficient in capturing the nature of time variation of our model parameters, this approach produces more accurate OOS forecasts relative to use of models which allow stochastic shifts in model parameters across observations.

In the published discussion of Hamilton (2009), one participant suggests that the IS results presented in that paper may reflect overfitting and thus may overestimate the effect of oil prices on GDP. Among the standard checks against such a claim is carrying out an OOS investigation of the underlying relationship. Accordingly, we believe our results suggest that Hamilton's (2009) findings do not stem from overfitting.

The impulse response functions of Blanchard and Galí (2010) show that the causal relationship between oil prices and GDP weakened markedly during the 2000s. We find, however, strong predictability from oil prices to GDP during this period. Similarly, Nakov and Pescatori (2010) find that smaller oil price shocks and a "reduced share of oil in GDP" were both important factors behind the Great Moderation of 1984-2007, for which we also find that oil prices are Granger-causal for GDP. In future work it would be interesting to see if these results can be reconciled.

Recently there has been a debate about the extent to which, as a result of globalization, international factors have become more important than domestic factors in the data generating process for inflation and the transmission mechanism of monetary policy; see, for example, Borio and Filardo (2007), Ihrig, Kamin, Lindner, and Marquez (2007), and Mishkin (2009). Our results suggests it might be useful for this literature to consider crude oil prices and Kilian's (2009) real global economic activity index as candidate variables for global factors.

Our analysis is agnostic about whether the oil price movements which OOS Granger-cause GDP are due to demand shocks, supply shocks, or both, and we believe it would be informative to determine which type of shocks drive the oil price predictability we uncover. We note two issues of concern with applying, for example, Kilian's (2009) framework to produce estimates of such shocks for the problem we study. First, data availability on world crude oil production would reduce considerably the length of the OOS period. Second, Hamilton (2009) notes that, in several periods for which Kilian's (2009) procedure identifies shocks driven by a large precautionary demand for oil, actual oil inventories in the U.S. decreased. That said, we think it would be fruitful to explore this question in future work.

References

Amisano, G., Giacomini, R., 2007. Comparing density forecasts via weighted likelihood ratio tests. Journal of Business & Economic Statistics 25 (2), 177–190.

Ashley, R., Granger, C. W. J., Schmalensee, R., 1980. Advertising and aggregate consumption: An analysis of causality. Econometrica 48 (5), 1149–1167.

Bachmeier, L., Li, Q., Liu, D., 2008. Should oil prices receive so much attention? An evaluation of the predictive power of oil prices for the u.s. economy. Economic Inquiry 46 (4), 528–539.

Barsky, R. B., Kilian, L., 2002. B.S. Bernanke and K. Rogoff (eds.), NBER Macroeconomics Annual 2001. MIT Press, Cambridge, MA, Ch. Do we really know that oil caused the Great Stagflation? A monetary alternative.

Barsky, R. B., Kilian, L., 2004. Oil and the macroeconomy since the 1970s. Journal of Economic Perspectives 18 (4), 115–134.

Baumeister, C., Peersman, G., 2008. Time-varying effects of oil supply shocks on the us economy. Working Papers of Faculty of Economics and Business Administration, Ghent University, Belgium 08/515, Ghent University, Faculty of Economics and Business Administration.

Bernanke, B. S., Gertler, M., Watson, M. W., 1997. Systematic monetary policy and the effects of oil price shocks. Brookings Papers on Economic Activity (1), 91–142.

Bjørnland, H. C., 2000. The dynamic effects of aggregate demand, supply and oil price shocks–a comparative study. Manchester School 68 (5), 578–607.

Blanchard, O. J., Galí, J., 2010. J. Galí and M. Gertler (eds.), International Dimensions of Monetary Policy. University of Chicago Press, Chicago, IL, Ch. The Macroeconomic Effects of Oil Price Shocks: Why are the 2000s so Different from the 1970s?

Blanchard, O. J., Riggi, M., 2009. Why are the 2000s so different from the 1970s? A structural interpretation of changes in the macroeconomic effects of oil prices. NBER Working Papers 15467, National Bureau of Economic Research, Inc.

Blinder, A. S., 2009. Comment on causes and consequences of the oil shock of 2007-08 by james d. hamilton. Brookings Papers on Economic Activity (Spring), 262–267.

Blinder, A. S., Rudd, J. B., 2009. The supply-shock explanation of the Great Stagflation revisited. Working paper, Princeton University.

Borio, C. E. V., Filardo, A., May 2007. Globalisation and inflation: New cross-country evidence on the global determinants of domestic inflation. BIS Working Papers 227, Bank for International Settlements.

Carlton, A. B., 2010. Oil prices and real-time output growth. Working paper, University of Houston.

Chauvet, M., Piger, J., January 2008. A comparison of the real-time performance of business cycle dating methods. Journal of Business & Economic Statistics 26, 42–49.

Clark, T., West, K., 2007. Approximately normal tests for equal predictive accuracy in nested models. Journal of Econometrics 138 (1), 291–311.

Clark, T. E., McCracken, M. W., 2009. Nested forecast model comparisons: a new approach to testing equal accuracy. Working Paper 2009-050, Federal Reserve Bank of St. Louis.

Cochrane, J. H., 2005. Time Series for Macroeconomics and Finance. Chicago, IL.

Cragg, J., Donald, S. G., 1993. Testing identifiability and specification in instrumental variable models. Econometric Theory 9 (2), 222–240.

Croushore, D., Stark, T., 2003. A real-time data set for macroeconomists: Does the data vintage matter? The Review of Economics and Statistics 85 (3), 605–617.

Edelstein, P., Kilian, L., 2009. How sensitive are consumer expenditures to retail energy prices? Journal of Monetary Economics 56 (6), 766–779.

Estrella, A., Hardouvelis, G. A., 1991. The term structure as a predictor of real economic activity. Journal of Finance 46 (2), 555–576.

Faust, J., Wright, J. H., 2009. Comparing greenbook and reduced form forecasts using a large realtime dataset. Journal of Business & Economic Statistics 27 (4), 468–479.

Hamilton, J. D., 1983. Oil and the macroeconomy since World War II. Journal of Political Economy 91 (2), 228–248.

Hamilton, J. D., 1985. Historical causes of postwar oil shocks and recessions. Energy Journal 6 (1), 97–116.

Hamilton, J. D., 1989. A new approach to the economic analysis of nonstationary time series and the business cycle. Econometrica 57 (2), 357–384.

Hamilton, J. D., 1996. This is what happened to the oil price-macroeconomy relationship. Journal of Monetary Economics 38 (2), 225–230.

Hamilton, J. D., 2001. A parametric approach to flexible nonlinear inference. Econometrica 69, 537–573.

Hamilton, J. D., 2003. What is an oil shock? Journal of Econometrics 113 (2), 363–398.

Hamilton, J. D., 2009. Causes and consequences of the oil shock of 2007-08. Brookings Papers on Economic Activity (Spring), 215–259.

Hamilton, J. D., Jul. 2010. Nonlinearities and the macroeconomic effects of oil prices. NBER Working Papers 16186, National Bureau of Economic Research, Inc.

Hamilton, J. D., Herrera, A. M., 2004. Oil shocks and aggregate macroeconomic behavior: The role of monetary policy. Journal of Money, Credit, and Banking 36 (2), 265–286.

Hooker, M., 1996. What happened to the oil price-macroeconomy relationship? Journal of Monetary Economics 38 (2), 195–213.

Hubrich, K., West, K. D., 2010. Forecast evaluation of small nested model sets. Journal of Applied Econometrics 25 (4), 574–594.

Ihrig, J., Kamin, S. B., Lindner, D., Marquez, J., 2007. Some simple tests of the globalization and inflation hypothesis. Tech. rep.

Inoue, A., Kilian, L., 2004. In-sample or out-of-sample tests of predictability: Which one should we use? Econometric Reviews 23 (4), 371–402.

Kilian, L., 2008. The economic effects of energy price shocks. Journal of Economic Literature 46 (4), 871–909.

Kilian, L., 2009. Not all oil price shocks are alike: Disentangling demand and supply shocks in the crude oil market. American Economic Review 99 (3), 1053–1069.

Kilian, L., 2010. Oil price shocks, monetary policy and stagflation. In: Fry, R., Jones, C., Kent, C. (Eds.), Inflation in an Era of Relative Price Shocks. RBA Annual Conference Volume. Reserve Bank of Australia.

Koop, G., 2003. Bayesian Econometrics. Wiley.

Leeb, H., Pötscher, B. M., 2005. Model selection and inference: Facts and fiction. Econometric Theory 21 (1), 21–59.

Mishkin, F. S., 2009. Globalization, macroeconomic performance, and monetary policy. Journal of Money, Credit and Banking 41 (s1), 187–196.

Nakov, A., Pescatori, A., 2010. Oil and the Great Moderation. Economic Journal 120 (543), 131–156.

Ravazzolo, F., Vahey, S., 2010. Forecast densities for economic aggregates from disaggregate ensembles. Working Paper 2010/02, Norges Bank.

Sensier, M., van Dijk, D., 2004. Testing for volatility changes in u.s. macroeconomic time series. The Review of Economics and Statistics 86 (3), 833–839.

Stock, J. H., Watson, M. W., 1996. Evidence on structural instability in macroeconomic time series relations. Journal of Business & Economic Statistics 14 (1), 11–30.

Stock, J. H., Watson, M. W., 1999. Forecasting inflation. Journal of Monetary Economics 44 (2), 293–335.

Stock, J. H., Yogo, M., 2005. Donald W. K. Andrews and James H. Stock (eds.), Essays in Honor of Thomas Rothenberg. Cambridge University Press, Cambridge and New York, Ch. Testing for Weak Instruments in Linear IV Regression.

Swanson, N. R., 1998. Money and output viewed through a rolling window. Journal of Monetary Economics 41 (3), 455–474.

White, H., 2000. A reality check for data snooping. Econometrica 68 (5), 1097–1126.

Wright, J. H., 2006. The yield curve and predicting recessions. Finance and Economics Discussion Series 2006-07, Board of Governors of the Federal Reserve System (U.S.).

Table 1: Tests of Equal Out-of-Sample Point Forecast Accuracy for Quarterly US GDP Growth Rates with AR Benchmarks

	1970-2009	1975-2009	1980-2009	1985-2009	1990-2009	1995-2009	2000-2009
				Forecast horizon h=1			
AR(4) (bench)	0.623	0.574	0.440	0.251	0.290	0.316	0.389
vs. ARX(4)^o	0.388	1.101	0.869	0.725	0.734	0.684	0.591
	(0.105)	(0.624)	(0.224)	(**0.051**)	(**0.060**)	(**0.057**)	(**0.044**)
vs. ARX(4)^n	0.657	1.010	0.846	0.688	0.632	0.560	0.449
	(**0.065**)	(0.536)	(0.109)	(**0.026**)	(**0.015**)	(**0.013**)	(**0.010**)
HW: vs. 2 models	(**0.099**)	(0.772)	(0.144)	(**0.038**)	(**0.021**)	(**0.018**)	(**0.014**)
AR(p)$_{\text{AIC}}$ (bench)	0.576	0.495	0.418	0.258	0.294	0.321	0.395
vs. ARX(p)$^o_{\text{AIC}}$	0.927	0.984	0.986	1.017	0.987	1.009	0.995
	(0.127)	(0.298)	(0.366)	(0.606)	(0.421)	(0.549)	(0.479)
vs. ARX(p)$^n_{\text{AIC}}$	0.886	0.984	0.897	0.814	0.797	0.779	0.719
	(0.139)	(0.442)	(**0.017**)	(**0.006**)	(**0.005**)	(**0.011**)	(**0.006**)
HW: vs. 2 models	(0.229)	(0.508)	(**0.034**)	(**0.011**)	(**0.011**)	(**0.021**)	(**0.013**)
				Forecast horizon h=4			
AR(4) (bench)	0.806	0.684	0.512	0.297	0.349	0.358	0.456
vs. ARX(4)^o	0.634	0.540	0.897	0.799	0.827	0.896	0.861
	(0.377)	(0.382)	(**0.097**)	(**0.009**)	(**0.009**)	(**0.081**)	(**0.046**)
vs. ARX(4)^n	0.592	0.411	1.011	0.830	0.862	0.829	0.809
	(0.155)	(0.130)	(0.552)	(**0.030**)	(**0.047**)	(**0.049**)	(**0.056**)
HW: vs. 2 models	(0.208)	(0.182)	(0.138)	(**0.015**)	(**0.018**)	(**0.085**)	(**0.079**)
AR(p)$_{\text{AIC}}$ (bench)	0.782	0.644	0.526	0.288	0.339	0.343	0.433
vs. ARX(p)$^o_{\text{AIC}}$	1.052	1.068	1.006	1.000	0.994	0.995	0.989
	(0.824)	(0.814)	(0.723)	(0.489)	(0.337)	(0.396)	(0.290)
vs. ARX(p)$^n_{\text{AIC}}$	1.079	1.064	1.016	0.870	0.865	0.822	0.751
	(0.740)	(0.654)	(0.629)	(**0.056**)	(**0.059**)	(**0.058**)	(**0.027**)
HW: vs. 2 models	(0.982)	(0.953)	(0.851)	(0.114)	(0.116)	(0.115)	(**0.057**)

Notes: Table reports results for tests of equal out-of-sample point forecast accuracy for models of US GDP growth over various out-of-sample periods for two forecasting horizons, $h = 1$ and $h = 4$ steps ahead. The models were estimated using moving windows of real-time data; the first in-sample window is 1955Q1-1969Q4. For benchmark models, MSPEs reported; for alternatives to the benchmark, the ratio of the alternative model's adjusted MSPE to the benchmark's MSPE reported, where the adjusted MSPE was computed as per Clark and West (2007). In parentheses under the MSPE ratios are reported $p-$values for the Clark and West (2007) test for equal forecast accuracy for nested models. "AR(4)" and "ARX(4)" indicate that the lag length p was fixed at 4 for all estimation windows, and the subscript AIC indicates that the lag length was selected using the Akaike Information Criterion. The superscripts "o" and "n" indicate, respectively, that the ARX alternative model includes p lags of the crude oil price growth rate and the "net oil price increase" (NOPI) measure introduced by Hamilton (1996). The row labeled "HW" reports $p-$values for the "max $t-$statistic" variant of the Hubrich and West (2010) test for forecasting accuracy for a small set of nested models.

Table 2: Tests of Equal Out-of-Sample Point Forecast Accuracy for Quarterly US GDP Growth Rates with ARX Benchmarks at Forecast Horizon $h = 1$

	1970-2009	1975-2009	1980-2009	1985-2009	1990-2009	1995-2009	2000-2009
ARX(4)^{i} (bench)	1.215	1.077	0.687	0.495	0.414	0.376	0.486
ARX(4)^{i,o}	1.200	0.830	0.880	0.835	0.964	0.944	0.922
	(0.828)	(**0.041**)	(**0.067**)	(**0.051**)	(0.287)	(0.246)	(0.200)
ARX(4)^{i,n}	0.942	0.770	0.862	0.941	0.997	1.000	1.005
	(0.340)	(**0.076**)	(**0.014**)	(0.251)	(0.487)	(0.500)	(0.516)
HW: vs. 2 models	(0.479)	(**0.064**)	(**0.025**)	(**0.085**)	(0.426)	(0.377)	(0.314)
ARX(4)^{c} (bench)	1.056	0.996	0.674	0.461	0.352	0.273	0.326
ARX(4)^{c,o}	1.167	0.768	0.952	0.828	0.916	0.872	0.837
	(0.719)	(**0.046**)	(0.279)	(**0.026**)	(0.175)	(0.178)	(0.173)
ARX(4)^{c,n}	1.035	0.923	0.866	0.870	0.884	0.890	0.882
	(0.610)	(0.241)	(**0.018**)	(**0.022**)	(**0.086**)	(0.200)	(0.234)
HW: vs. 2 models	(0.750)	(**0.079**)	(**0.037**)	(**0.042**)	(0.145)	(0.275)	(0.272)
ARX(4)^{s}				0.412	0.319	0.273	0.348
ARX(4)^{s,o}				0.781	0.912	0.834	0.794
				(**0.050**)	(0.210)	(0.129)	(0.113)
ARX(4)^{s,n}				0.861	0.932	0.882	0.868
				(**0.066**)	(0.258)	(0.213)	(0.224)
HW: vs. 2 models				(**0.084**)	(0.277)	(0.179)	(0.164)
ARX(4)^{f}	1.309	1.254	0.947	0.504	0.390	0.325	0.416
ARX(4)^{f,o}	0.948	0.736	0.957	0.992	0.990	1.018	1.024
	(0.413)	(**0.076**)	(0.231)	(0.421)	(0.433)	(0.584)	(0.598)
ARX(4)^{f,n}	0.942	0.905	0.935	0.975	0.921	0.937	0.937
	(0.275)	(0.168)	(0.161)	(0.285)	(0.119)	(0.252)	(0.282)
HW: vs. 2 models	(0.385)	(0.112)	(0.265)	(0.434)	(0.201)	(0.410)	(0.448)

continued on next page

continued from previous page

	1970-2009	1975-2009	1980-2009	1985-2009	1990-2009	1995-2009	2000-2009
$\text{ARX}(p)^{i}_{AIC}$ (bench)	0.970	0.911	0.639	0.430	0.314	0.253	0.314
$\text{ARX}(p)^{i,o}_{AIC}$	1.088	1.018	0.992	0.995	0.982	0.980	0.973
	(0.928)	(0.798)	(0.333)	(0.438)	(0.367)	(0.405)	(0.395)
$\text{ARX}(p)^{i,n}_{AIC}$	1.042	1.055	0.944	0.989	1.037	1.005	1.021
	(0.822)	(0.864)	(**0.043**)	(0.393)	(0.719)	(0.519)	(0.569)
HW: vs. 2 models	(0.965)	(0.871)	(**0.083**)	(0.579)	(0.546)	(0.591)	(0.578)
$\text{ARX}(p)^{c}_{AIC}$ (bench)	1.053	1.002	0.689	0.450	0.336	0.251	0.305
$\text{ARX}(p)^{c,o}_{AIC}$	1.036	1.008	0.975	0.940	0.940	0.914	0.895
	(0.895)	(0.635)	(0.178)	(**0.088**)	(0.106)	(0.141)	(0.142)
$\text{ARX}(p)^{c,n}_{AIC}$	1.007	0.998	0.917	0.971	0.977	0.965	0.960
	(0.565)	(0.481)	(**0.018**)	(0.176)	(0.325)	(0.348)	(0.356)
HW: vs. 2 models	(0.802)	(0.699)	(**0.035**)	(0.171)	(0.201)	(0.266)	(0.269)
$\text{ARX}(p)^{s}_{AIC}$ (bench)				0.422	0.319	0.275	0.355
$\text{ARX}(p)^{s,o}_{AIC}$				0.927	0.949	0.910	0.888
				(0.117)	(0.191)	(0.129)	(0.111)
$\text{ARX}(p)^{s,n}_{AIC}$				0.964	0.965	0.902	0.897
				(0.207)	(0.305)	(0.173)	(0.195)
HW: vs. 2 models				(0.214)	(0.297)	(0.209)	(0.186)
$\text{ARX}(p)^{f}_{AIC}$ (bench)	1.215	1.176	0.880	0.478	0.360	0.259	0.330
$\text{ARX}(p)^{f,o}_{AIC}$	0.963	1.004	0.988	0.959	0.915	0.883	0.868
	(0.241)	(0.611)	(0.186)	(**0.063**)	(**0.014**)	(**0.041**)	(**0.044**)
$\text{ARX}(p)^{f,n}_{AIC}$	0.934	0.974	0.926	0.948	0.910	0.851	0.828
	(0.104)	(0.249)	(**0.018**)	(**0.067**)	(**0.056**)	(**0.073**)	(**0.073**)
HW: vs. 2 models	(0.166)	(0.414)	(**0.037**)	(0.122)	(**0.030**)	(**0.080**)	(**0.090**)

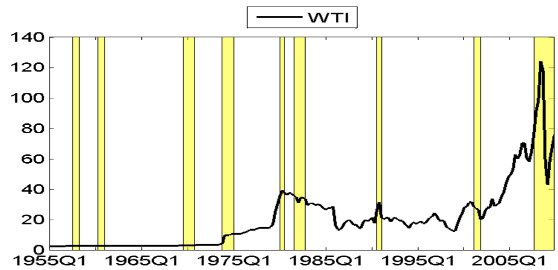
Notes: See notes to Table 1. The superscripts "i," "c," "s," and "f" indicate, respectively, that the ARX model includes p lags of the growth rate of the import price deflator, the growth rate of the personal consumption expenditures deflator, the linear detrended real shipping freight index of Kilian (2009), and a factor series provided by the first principal component for the following set of variables: growth rates of the import price deflator, personal consumption expenditures deflator, and nominal shipping freight index of Kilian (2009), the linear detrended real shipping freight index of Kilian (2009), the 3-month T-Bill rate, and the 3-month T-Bill/fed funds, 10-year T-Bond/three-month T-Bill, and Moody's Baa/Aaa spreads. The superscripts "o" and "n" indicate, respectively, that the ARX alternative model also includes p lags of the crude oil price growth rate and the "net oil price increase" (NOPI) measure introduced by Hamilton (1996).

Table 3: Log Scores for Out-of-Sample Density Forecasts for Quarterly US GDP Growth Rates

	1970-2009	1975-2009	1980-2009	1985-2009	1990-2009	1995-2009	2000-2009
	Forecast horizon h=1						
AR(4) (bench)	-1.184	-1.186	-1.187	-1.171	-1.227	-1.295	-1.376
vs. ARX(4)^o	-1.138	-1.144	-1.146	-1.110	-1.142	-1.187	-1.209
	(0.227)	(0.220)	(0.222)	(**0.040**)	(**0.015**)	(**0.017**)	(**0.012**)
vs. ARX(4)^n	-1.110	-1.120	-1.116	-1.094	-1.123	-1.163	-1.181
	(**0.044**)	(**0.049**)	(**0.035**)	(**0.017**)	(**0.007**)	(**0.008**)	(**0.008**)
AR(p)$_{\text{AIC}}$ (bench)	-1.204	-1.199	-1.205	-1.195	-1.256	-1.332	-1.427
vs. ARX(p)$^o_{\text{AIC}}$	-1.177	-1.172	-1.175	-1.161	-1.211	-1.276	-1.342
	(0.127)	(0.298)	(0.366)	(0.606)	(0.421)	(0.549)	(0.479)
vs. ARX(p)$^n_{\text{AIC}}$	-1.149	-1.155	-1.147	-1.132	-1.177	-1.231	-1.273
	(0.139)	(0.442)	(**0.017**)	(**0.006**)	(**0.005**)	(**0.011**)	(**0.006**)
	Forecast horizon h=4						
AR(4) (bench)	-1.249	-1.207	-1.212	-1.230	-1.303	-1.390	-1.516
vs. ARX(4)^o	-1.198	-1.205	-1.204	-1.207	-1.267	-1.344	-1.438
	(0.306)	(0.379)	(0.338)	(0.250)	(0.184)	(0.172)	(0.124)
vs. ARX(4)^n	-1.175	-1.173	-1.173	-1.166	-1.214	-1.270	-1.336
	(**0.073**)	(0.109)	(0.103)	(**0.042**)	(**0.019**)	(**0.018**)	(**0.019**)
AR(p)$_{\text{AIC}}$ (bench)	-1.266	-1.224	-1.234	-1.245	-1.323	-1.417	-1.570
vs. ARX(p)$^o_{\text{AIC}}$	-1.224	-1.217	-1.225	-1.234	-1.309	-1.400	-1.543
	(**0.023**)	(**0.052**)	(**0.026**)	(**0.027**)	(**0.027**)	(**0.041**)	(**0.023**)
vs. ARX(p)$^n_{\text{AIC}}$	-1.198	-1.192	-1.183	-1.184	-1.245	-1.312	-1.379
	(0.185)	(0.230)	(0.106)	(0.111)	(0.111)	(0.106)	(**0.043**)

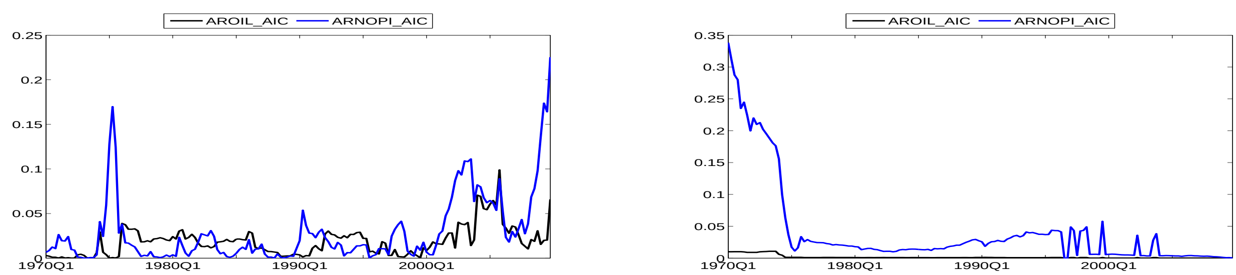
Notes: Table reports the log scores of the out-of-sample quarterly US GDP growth density forecasts over various out-of-sample periods using models described in Section 2 for two forecasting horizons, h=1 and h=4 steps ahead; see notes to Table 1 for explanation of notation used for names of models. In parentheses under the log scores are reported $p-$values for the center of distribution variant of the Amisano and Giacomini (2007) test of equal density predictive accuracy.

Figure 1: Time Series Plot of WTI Crude Oil Price, 1955Q1-2009Q4



Notes: Time series plot of West Texas Intermediate crude oil price, 1955Q1-2009Q4. NBER recession dates are shaded in yellow; end of recession that began in December 2007 determined by the Chauvet and Piger (2008) model.

Figure 2: Model Selection Criteria Across Estimation Windows

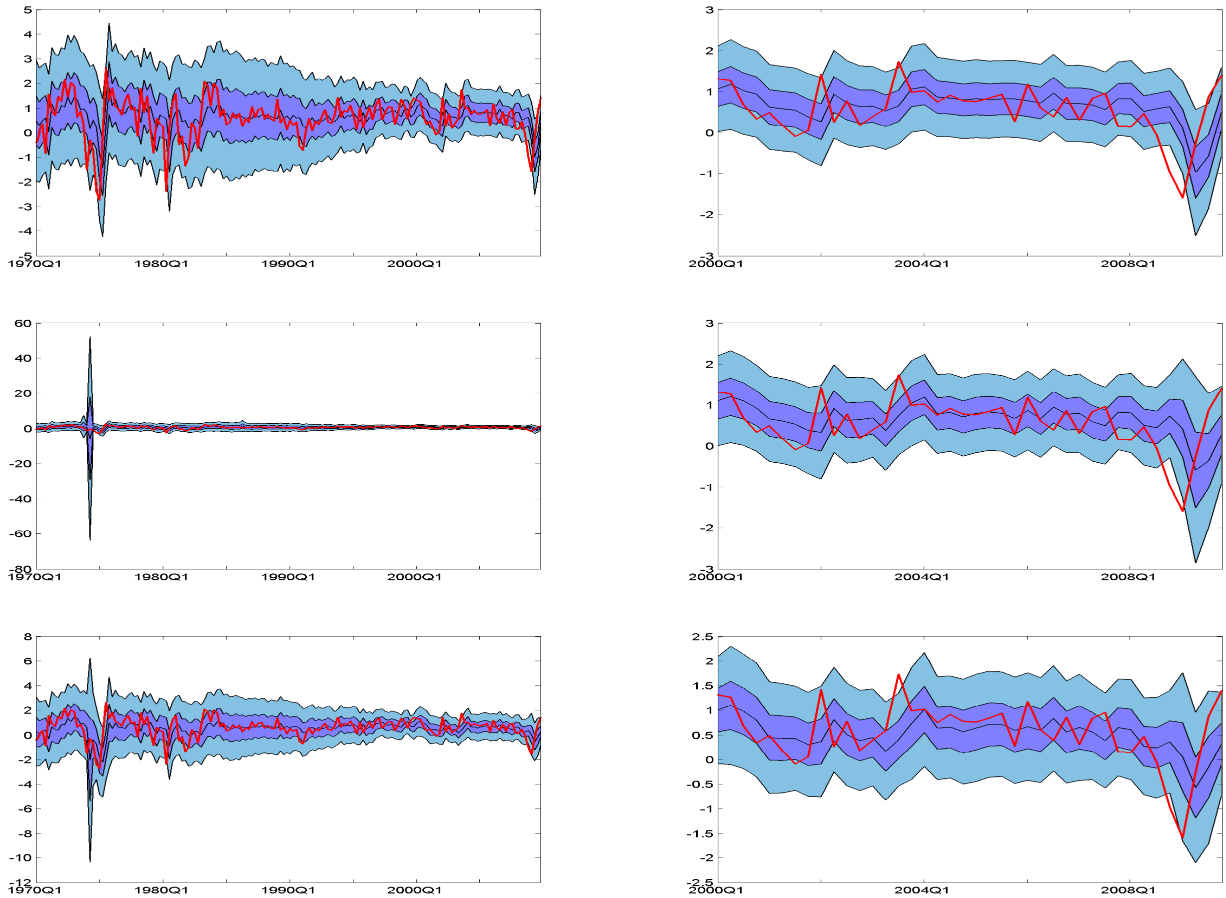


AIC Differences

Bayes Factor

Notes: The graphs show differences in AIC (AIC(benchmark) - AIC(alternative)) and Bayes Factors (Prob(benchmark)/Prob(alternative), where 'Prob' represents marginal likelihood) for the benchmark model without oil prices and alternative models with an oil price measure included across fixed length 15-year moving estimation windows of real-time data; if the benchmark model generates the better fit, then the AIC differences are negative and the Bayes factor is greater than one. The black and blue lines show comparisons between, respectively, the $\mathrm{AR}(p)_{\mathrm{AIC}}$ and $\mathrm{ARX}(p)^{o}_{\mathrm{AIC}}$ models, and $\mathrm{AR}(p)_{\mathrm{AIC}}$ and $\mathrm{ARX}(p)^{n}_{\mathrm{AIC}}$ models; see notes to Table 1 for explanation of notation used for names of models. The first and last in-sample periods are, respectively, 1955Q1-1969Q4 and 1995Q1-2009Q4. The dates on the horizontal axis show the last observation of each estimation window.

Figure 3: Fan Charts, 1-Step Ahead Forecasts, Lag Length Selected by AIC

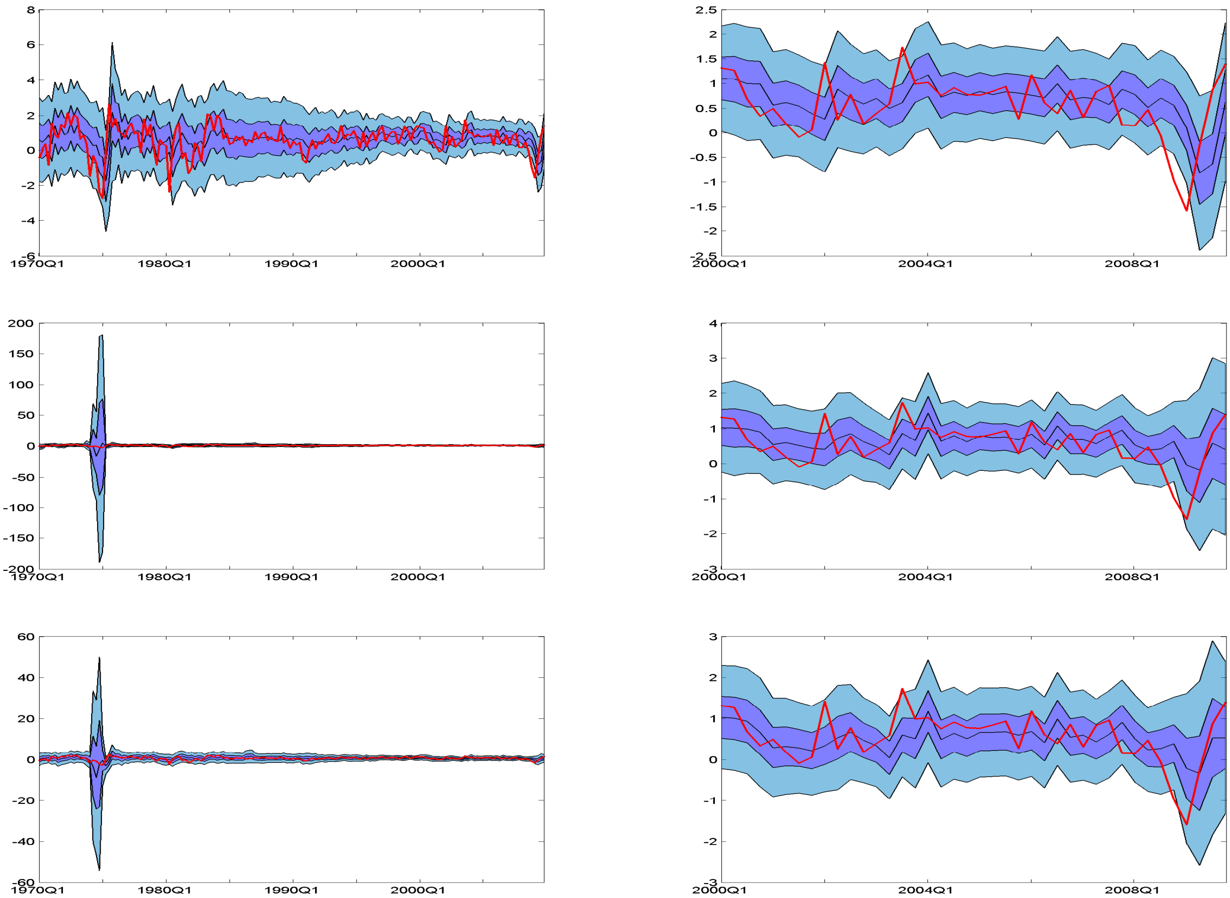


1970 - 2009

2000 - 2009

Notes: For each fan chart, the black solid lines represent the 5%, 25%, 50%, 75%, and 95% percentiles of the corresponding density forecast and the red solid line shows the realized values for real GDP growth, for each out-of-sample observation. In each column, the first, second, and third graphs show the fan charts for the $\text{AR}(p)_{\text{AIC}}$, $\text{ARX}(p)^{o}_{\text{AIC}}$, and $\text{ARX}(p)^{n}_{\text{AIC}}$ models; see notes to Table 1 for explanation of notation used for names of models.

Figure 4: Fan Charts, 1-Step Ahead Forecasts, Lag Length $p = 4$



1970 - 2009

2000 - 2009

Notes: For each fan chart, the black solid lines represent the 5%, 25%, 50%, 75%, and 95% percentiles of the corresponding density forecast and the red solid line shows the realized values for real GDP growth, for each out-of-sample observation. In each column, the first, second, and third graphs show the fan charts for the AR(4), ARX(4)^o, and ARX(4)^n models; see notes to Table 1 for explanation of notation used for names of models.