

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

Francesco Ravazzolo*

Norges Bank

Philip Rothman†

East Carolina University‡

June 9, 2010

Abstract

We study the real-time Granger-causal relationship between crude oil prices and US GDP growth through an out-of-sample (OOS) forecasting exercise; we do so after providing strong evidence of in-sample (IS) predictability from oil prices to GDP. Comparing our benchmark model “without oil” against those “with oil” by way of both point and density forecasts, we find strong evidence in favor of OOS predictability from oil prices to GDP via our point forecast comparisons when we adjust our MSPEs to account for noise introduced under the null hypothesis that the parsimonious benchmark is the true data generating process. These results are consistent with well-known IS results covering part of our OOS period, and also suggest that, in the 1990s and 2000s, oil prices have had greater predictive content for GDP than in the mid to late 1980s. By way of density forecast OOS comparisons, while we do not find statistically significant evidence of such predictability from oil prices to GDP for the full 1970-2008 OOS period, our results qualitatively also suggest substantial time variation in this relationship; predictability from 1970 to 1985, and increasing predictability near the onset of the Great Recession.

*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

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

‡The views expressed in this paper are our own and do not necessarily reflect those of Norges Bank.

1 Introduction

Blinder and Rudd (2008) emphasize that, in apparently helping to produce large macroeconomic effects in the form of high inflation and a deep recession, the 1973 post-Yom Kippur War OPEC oil price increases had a sense of being “something new, if not indeed something *sui generis*, at the time.” In a seminal paper, however, Hamilton (1983) shows that a strong case can be made for the hypothesis that negative oil price shocks systematically preceded recessions from the early post-World War II period to the beginning of the 1980s.

He 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.

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 questions 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 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 relationship between oil prices and the macroeconomy, including Bernanke, Gertler, and Watson (1997), Blanchard and Galí (2008), and Herrera and Pesavento (2009). However, Hamilton and Herrera (2004) show that the results in Bernanke et al. (1997) are not robust, while Hamilton (2009) points out that the Blanchard and Galí (2008) estimates imply, counterintuitively, that the US 1981-82 recession would have been deeper in the absence of the crude oil price shocks that preceded it. Further, applying the novel random field approach of Hamilton (2001), Hamilton (2003) presents evidence suggesting that the causal 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.¹

All of the literature referenced above is based on in-sample (IS) analysis. The goal of this paper is to explore this relationship by way of an out-of-sample (OOS) forecasting study. Our interest in doing so is not motivated by concern that IS inference without OOS verification is likely to be

¹Using 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).

spurious, as Ashley, Granger, and Schmalensee (1980) warn, such that an OOS approach inherently involves less overfitting and is necessarily the correct one to adopt. Rather, we view the results we obtain as a natural complement to the set of mixed and conflicting results reported by leading scholars in the literature and refer to the argument of Welch and Goyal (2008) that they provide “useful diagnostic” information about the underlying dynamic relationship.

Welch and Goyal (2008) argue that it is not reasonable to search for evidence of OOS predictability in the absence of IS predictability. Accordingly, for models we further explain below, in Figure 1 we present evidence of such IS predictability from crude oil prices to US GDP using sequence of both expanding and rolling 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), the Bayesian Information Criterion (BIC), and a higher 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.² The importance of using such data, as opposed to revised data, is twofold. First, if all models in the OSS study were estimated with the most recent vintage available at the time the research is carried, 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 models considered.³

We carry out our OOS predictability analysis with both point and density forecasts. Our key results from the point forecast comparisons are as follows. Via the NOPI measure, we find very strong statistically significant predictability from oil prices to GDP for the longest OOS period we consider, 1970-2008. Further examination suggests that this predictability was strongest during the 1970s and early 1980s, weakened in the 1980s, and increased in strength in the 1990s and 2000s. Our density forecast comparisons suggest moderate OOS predictability from oil prices to GDP growth in the 1970s and early 1980s, but when we consider the full 1970-2008 OOS period, no such statistically significant predictability is detected.

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.

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.

²Croushore and Stark (2003) provide a useful discussion of real-time versus revised data.

³This is the case, for example, for the OOS time series forecasts Faust and Wright (2009) analyze.

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, which implies that we do not need to employ multi-equation systems to produce our forecasts.

We use data for US GDP, import prices, and the GDP deflator from real-time vintages downloaded from the Philadelphia Federal Reserve Bank's real-time database from 1950Q1 to 2008Q4; the first vintage covers 1950Q1-1969Q4, and the last vintage runs from 1950Q1 to 2008Q4.⁴ For a crude oil price measure we obtained data on 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. The interest rate variable we use is the 3-month Treasury Bill rate from the FRED database at Federal Reserve of Saint Louis.⁵

2.1 Forecasting Models

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

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

where $\epsilon_t \sim N(0, 1)$. The Akaike and Bayesian information criteria both identify $p = 4$ across our IS periods. We apply Bayesian inference with weak informative conjugate priors to restrict regression coefficients to zero.⁶ The model is estimated and point and density forecasts are produced via considering two alternative schemes: a sequence of expanding windows, which refer to as AR_e, and a sequence of 15-year moving windows, which we call AR_m. The first expanding window IS period is 1950Q1-1969Q4, while the first moving window IS period is 1955Q1-1969Q4,

The second group of models extends the AR(p) with an oil price measure:

$$y_t = \alpha + \sum_{i=1}^p \beta_i y_{t-i} + \sum_{i=1}^p \gamma_i oil_{t-i} + \sigma \epsilon_t, \quad (2)$$

where $\epsilon_t \sim N(0, 1)$ and oil_t is the oil price measure at time t . We use two alternatives: the oil price growth rate, $oil_t = \ln(p_t) - \ln(p_{t-1})$ where p_t is the West Texas Intermediate spot oil

⁴<http://www.philadelphiafed.org/research-and-data/real-time-center/real-time-data/>

⁵<http://research.stlouisfed.org/fred2/>

⁶We use a normal inverted gamma prior with means for α and β 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. All the linear models in this paper have these properties. See, for example, Koop (2003) for details.

price in quarter t ; and the NOPI measure proposed by Hamilton (1996), $oil_t = \max[(\ln(p_t) - \max[\ln(p_t), \dots, \ln(p_{t-4})]), 0]$. Again, we estimate the models using an expanding window or a 15-year moving window resulting in four different specifications: AROIL_e, AROIL_m, ARNOPI_e, ARNOPI_m respectively.

Hooker (1996) augments model (2) with a set of M macro variables:

$$y_t = \alpha + \sum_{i=1}^p \beta_i y_{t-i} + \sum_{i=1}^p \gamma_i oil_{t-i} + \sum_{j=1}^M \sum_{i=1}^p \delta_{n,i} m_{j,t-i} + \sigma \epsilon_t, \quad (3)$$

where $\epsilon_t \sim N(0, 1)$. He set the lag order p to rather high values, e.g., $p = 12$ for his full IS period. We determine p for each of our IS periods on the basis of both the Akaike and Bayesian information criteria. As in Hooker (1996) we use $M = 3$, specifically the price of imports, the GDP deflator and the interest rate on the 3-month Treasury Bill. As the number of parameters is potentially very high, we select the order of our autoregressive process with regressors by minimizing the Akaike information criteria at each point forecasts are made. Two measures of oil prices, and expanding or 15-years moving windows are used resulting in four specifications: AROILX_e, AROILX_m, ARNOPIX_e and ARNOPIX_m respectively.

The US and world economies have changed over time and modeling with a stable relation or with a simple correction by using 15-year moving windows may be not be sufficiently flexible to capture structural change. We reformulate models in (1) and (2) with time-varying parameters. Time instability is assumed as breaks and parameters are modeled as mixture of normals as in Ravazzolo and Vahey (2010):

$$\begin{aligned} y_t &= \alpha_t + \sum_{i=1}^p \beta_{t,i} y_{t-i} + \sigma_t \epsilon_t \\ \alpha_t &= \alpha_{t-1} + \kappa_{t,1} u_{t,1} \\ \beta_{t,i} &= \beta_{t-1,i} + \kappa_{t,1+i} u_{t,1+i} \\ \ln(\sigma_t^2) &= \ln(\sigma_{t-1}^2) + \kappa_{t,p+2} u_{t,p+2} \end{aligned} \quad (4)$$

where $\epsilon_t \sim N(0, 1)$, $\kappa_{t,s}$, $s = 1, \dots, p + 2$, is a $[0, 1]$ unobserved process with $Pr[\kappa_{t,s} = 1] = \pi_s$, and $u_{t,s} \sim N(0, 1)$; as per the IS estimation results for our AR(p) benchmark, we set $p = 4$.

Hence, the intercept α_t , the autoregressive parameters $\beta_{t,i}$, and the stochastic error standard deviation σ_t remain the same as their previous values for observation $t - 1$ unless $\kappa_{t,s} = 1$. The flexibility of the specification in (4) stems from the fact that the model parameters are allowed to change every time period, but need not. The magnitude of the changes is determined by $u_{t,s}$. We note that the shifts in the individual parameters are not restricted to occur simultaneously, but are allowed to take place at different points in time; if $\kappa_{t,s} = 0$, for $s \neq 1$, $\kappa_{t,1}$ were to follow a first-order Markov process, and $u_{t,1}$ were a scalar, the model (4) would reduce to a conventional two-state Markov-switching model as in Hamilton (1989), in which only the intercept term shifts. See Ravazzolo and Vahey (2010) for estimation details. Giordani and Villani (2009), for example,

show that this type of flexible model can provide accurate OOS forecasts. Similarly, we add oil prices as follows:

$$\begin{aligned}
y_t &= \alpha_t + \sum_{i=1}^p \beta_{t,i} y_{t-i} + \sum_{i=1}^p \gamma_{t,i} \text{Oil}_{t-i} + \sigma_t \epsilon_t \\
\alpha_t &= \alpha_{t-1} + \kappa_{t,1} u_{t,1} \\
\beta_{t,i} &= \beta_{t-1,i} + \kappa_{t,1+i} u_{t,1+i} \\
\gamma_{t,i} &= \gamma_{t-1,i} + \kappa_{t,p+1+i} u_{t,p+1+i} \\
\ln(\sigma_t) &= \ln(\sigma_{t-1}^2) + \kappa_{t,2p+2} u_{t,2p+2}.
\end{aligned} \tag{5}$$

Out two different measures for oil prices are used. The resulting models called: TVPAR, TV-PAROIL, TVPARNOPI respectively. We generate forecasts with these models by way of an expanding window scheme.⁷

2.2 Forecast Evaluation

To shed light on the predictive power of oil, we use a number of evaluation statistics for point and density forecasts previously proposed in literature. We compare point forecasts in terms of mean square prediction errors (MSPEs) and “adjusted” MSPEs (MSPEs-adjusted), where the MSPE adjustment is made as per Clark and West (2007), for different models and different OOS periods.⁸ We test the null hypothesis that the nested benchmark model without an oil price measure has the lower MSPE by way of the McCracken (2007) and Clark and West (2007) tests.

To implement the Clark and West (2007) test, we compute:

$$\hat{f}_{t+h} = (y_{t+h} - \hat{y}_{1,t+h})^2 - [(y_{t+h} - \hat{y}_{2,t+h})^2 - (\hat{y}_{1,t+h} - \hat{y}_{2,t+h})^2], \quad t = N, \dots, T - h, \tag{6}$$

where y_{t+h} is the realization of the variable of interest at time $t + h$, $\hat{y}_{i,t+h}$, $i = 1, 2$ are the h -step 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 Clark and West (2007) test for equal MSPE is carried out by regressing \hat{f}_{t+h} on a constant and running a t -test for the null hypothesis that the constant equals zero. Failure to reject the null indicates that model 2 reduces to model 1 at the given significance level.

Following Welch and Goyal (2008), we also graphically analyze what we call the Cumulative Squared Prediction Error Difference (CSPED):

⁷In the discussion below, we use the term “time-varying” to refer to this group of three models, and “expanding window” to refer to the expanding window implementation of models (1), (2), and (3).

⁸Under 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. Use of MSPE-adjusted is an attempt to reduce the role of such noise when making OOS forecasting comparisons for nested models.

$$CSPED_{t,h} = \sum_{s=N}^t \widehat{f}_{s+h}, \quad (7)$$

where \widehat{f}_{s+h} is given by equation 6.⁹ Increases in $CSPED_{t,h}$ indicate that the alternative to the benchmark predicts better at OOS observation t .

Density forecasts are compared using two different measures: a fit measure, the probability integral transform (*PIT*), and a distance measure, the log score. Use of the *PIT* to evaluate density forecasts was introduced into econometric time series analysis by Diebold, Gunther, and Tay (1998). Let $\{g_i(y_{t+h}|I_t)\}_{t=N+1}^T$ be the density forecast produced by model i conditional on the information set available at time $t = N$. Given the OOS realizations of the process, $\{y_{t+h}\}_{t=N+1}^T$, the *PIT* for each OOS observation is computed as:

$$p_{i,t+h} = \int_{-\infty}^{y_{t+h}} g_i(u|I_t) du, \quad t = N, \dots, T - h, \quad (8)$$

If the h -step ahead density forecasts coincide with the true densities $\{f(y_{t+h}|I_t)\}_{t=N}^{T-h}$, the sequence of the *PITs* is independently and identically distributed *i.i.d* with a uniform distribution, $U(0,1)$.

The goodness-of-fit tests we employ include the Likelihood Ratio (LR) test of Berkowitz (2001), the Anderson-Darling test, and the Pearson (χ^2) test as in Wallis (2003). We use the three degrees of freedom variant of the Berkowitz (2001) test of the null hypothesis that the sequence of transformed *PITs*, $\{z_{t+h}\}_{t=N}^T$, where $z_{i,t+h} = \Phi^{-1}(p_{i,t+h})$, with $\Phi^{-1}(\cdot)$ the inverse of the standard normal distribution function, is *i.i.d* $N(0,1)$ against an AR(1) alternative. The Anderson-Darling (AD) test for uniformity, a modification of the Kolmogorov-Smirnov test, gives more weight to the tails of the forecast density. The Pearson (χ^2) test divides the range of the $p_{i,t+h}$ into eight equiprobable classes and tests for uniformity in the histogram. We also test directly for the independence of the *PITs* using the Ljung-Box (LB) test, based on autocorrelation coefficients up to four.¹⁰ A well-calibrated ensemble should give high probability values for all four of these tests.

Turning to our analysis of relative predictive accuracy, we consider tests 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) 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:

$$WLR_{t+h} = w(y_{t+h}^{std})(\ln(g_1(y_{t+h}|I_t)) - \ln(g_2(y_{t+h}|I_t))), \quad (9)$$

⁹We note that Welch and Goyal (2008) do not make use of the Clark and West (2007) MSPE-adjusted measure.

¹⁰The null of the LB test is that the series being tested is uncorrelated, a weaker condition than independence. But if the null is rejected, the series is clearly not independent.

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_{t+h}^{std} is the realization y_{t+h} standardized using the IS data with which the density forecasts are estimated. The Amisano and Giacomini (2007) test statistic is computed as:

$$t_n = \frac{\overline{WLR}_n}{\widehat{\sigma}_{t+h}/\sqrt{n}}, \quad (10)$$

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 heteroscedastic and autocorrelation (HAC) consistent 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 Amisano and Giacomini (2007), which ignores the effects of any possible outliers.¹¹

Mitchell and Hall (2005) develop a *KLIC* test for nested models for the case in which the density forecasts are estimated using expanding windows of data. They propose the following WALD test:

$$GW_n = n \left(n^{-1} \sum_{t=N}^{T-h} z_{t+h-1} d_{t+h} \right)' \widehat{\Sigma}_{t+h} \left(n^{-1} \sum_{t=N}^{T-h} z_{t+h-1} d_{t+h} \right), \quad (11)$$

where $d_{t+h} = (\ln(g_1(y_{t+h}|I_t)) - \ln(g_2(y_{t+h}|I_t)))$, $z_{t+h-1} = (1, d_{t+h-1})'$, and $\widehat{\Sigma}_{t+h}$ is the HAC estimator for the variance of $(z_{t+h-1} d_{t+h})$. Under the null hypothesis of equal predictive accuracy, $GW_n \sim \chi_2^2$.

Analogous to our use of the CSPED for graphically examining relative MSPEs over time, and following Kascha and Ravazzolo (2010), we define the Cumulative Log Score Difference (*CLSD*):

$$CLSD_{t,h} = - \sum_{s=N}^t d_{s+h}, \quad (12)$$

where d_{t+h} is defined above. If $CLSD_{t,h}$ increases at observation t , this indicates that the alternative to the benchmark has a higher log score.

3 Results

We report OOS forecasting results for the 1970Q1 to 2008Q4 period as well as for nine subsamples: a set of six subsamples, with each starting five years later than the previous one, i.e. 1975Q1-2008Q4, 1980Q1-2008Q4, ..., 2000Q1-2008Q4; and a set of three subsamples focusing on the 1970s through the first half of the 1980s, 1970Q1-1979Q4, 1970Q1-1984Q4, and 1975Q1-1984Q4. Through consideration of these subsamples we are able to obtain an assessment about whether the oil predictability has changed over time, and in particular for specific periods such as the oil crises in

¹¹Our OOS density forecast comparisons are not strongly affected with use of the other three weighting functions Amisano and Giacomini (2007) provide.

the 1970’s, the relatively low oil price volatility regime from the mid 1980s to the 1990s, and the subsequent high oil price volatility period after 2000.

We first summarize the point forecast results found in Tables 1 and 2, which show, respectively, MSPEs and MSPEs-adjusted. The dominant result in Table 1 is that the benchmark “no oil” models produce more accurate point forecasts at both the $h = 1$ and $h = 4$ horizons, irrespective of whether an expanding window, moving window, or time-varying approach is used. The poor relative OOS performance of the models “with oil” is particularly pronounced in the 1970s and first part of the 1980s; in some cases the MSPEs of these models is an order of magnitude larger than in the benchmark case. For those OOS subsamples which exclude these years, the non-benchmark models generally produce higher MSPEs; and in the moving window case, the ARNOPI models generate significant MSPE reductions at the 1% significance level for both $h = 1$ and $h = 4$ in the 1995-2008 OSS period.

As per the analysis of Clark and West (2007), standard MSPE comparisons may be misleading when the alternative model nests the benchmark, and the results Tables 1 and 2 indeed differ considerably. For $h = 1$, three out of the four expanding window alternatives produce significant MSPE-adjusted reductions at the 5% significance level for the 1970-2008 OOS period; the relative performance of the “with oil” alternatives is particularly strong in both the 1970-1980 and 1970-1985 OOS periods. Further, ARNOPI expanding window forecasts dominate at the 1% and 5% significance levels, respectively, for the 1995-2008 and 2000-2008 OOS periods. The OOS predictability pattern from oil prices to GDP suggested by this set of results is that it was particularly strong in the early 1970s, weakened in the 1980s, and increased in strength in the 1990s and 2000s.

A slightly different pattern is suggested by the $h = 1$ moving window results in Table 2. For the relatively long 1970-2008 OOS period, the p -value for the equal MSPE null is less than 0.10 only for the ARNOPI and ARNOPIX models; the same is true for the 1970-1980 and 1970-1985 OOS subsamples. On the other hand, all alternatives to the benchmark generate significant MSPE reductions at the 1% significance level for the 1995-2008 and 2000-2008 subsamples. Further, the ARNOPI $h = 4$ results mirror those of the model’s $h = 1$ performance against the benchmark.

For the class of time-varying models, the null of equal MSPEs is not rejected in almost all cases. We believe this reflects the TVPAR 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.

Figure 2 offers a graphical complement to the results in Table 2, presenting time series plots of the CSPEDs we define above in equation (7) for the 1970-2008 (top panel) and 2000-2008 (lower panel) OOS samples. For the expanding and moving window case, the 1970-2008 results suggest that the forecast dominance of models with oil prices included stems primarily from the period immediately following the jump in oil prices observed in late 1973, with relatively little change afterwards. The 2000-2008 results for these forecast schemes show, however, that scale effects mask

some important later changes, i.e., in several cases the relative predictions from the alternative models steadily improve throughout the 2000s.

We next turn to discussion of our density forecast evidence on the predictive power of oil prices for GDP. The results in Table 3 show that the null hypothesis of the Berkowitz (2001) test that the density forecast is well calibrated is not rejected at the 5% significance level for the the 1970-2008 set of OOS forecasts for only three models at $h=1$, AR_e, AROILX_m and TVPARNOPI, and for only one model at $h = 4$, ARNOPIX_e. In contrast, the p -values for $h = 1$ in the last three columns are generally quite high, implying that the strong set of rejections over the 1970-2008 OOS forecasts stems from behavior in the post-1985 period. The Ljung-Box results in Table 4 suggest that the large number of rejections of the Berkowitz (2001) null we observe in Table 3 are not driven by strong dependence in the *PITs*. Rather, the pattern of rejections in Table 3, as per the Anderson-Darling and Pearson (χ^2) test results in Tables 5 and 6, is likely driven by the density forecast *PITs* not being uniformly distributed for most of the set of OOS forecast samples we consider; the evidence in favor of the uniform null is strongest for the 1970-1980, 1980-1985, and 1975-1985 OOS subsamples. The results in Tables 3-6, then, imply that the estimated densities we use to produce of forecasts are reasonable for the 1970-1985 OOS forecast periods, but generally do not coincide with the true density for the post-1985 OOS subsamples.

Following Mitchell and Wallis (2009), we evaluate relative density forecast performance via log score analysis. These results are presented in Table 7. We repeat that higher scores indicate better performance; since all of the log scores in Table 7 are negative, values closer to zero indicate higher forecast density accuracy. At $h = 1$, there is statistically significant evidence of predictability from oil prices to GDP for AROILX class of models for both the expanding and window schemes in the 1970s and first part of the 1980s. For the same OOS forecast periods, at $h = 4$ several alternatives suggest such predictability in the expanding window case but none do at conventional significance levels in the moving window case. For the post-1985 OOS periods, as well as for the full 1970-2008 OOS sample, the density forecasts when an oil price measure generally do not dominate the benchmark. We also note that our time-varying benchmark appear to be rather flexible in compensating for potential misspecification caused by omitting some relevant explanatory variables.

Figure 3 provides time series plots of our CLSD measure defined in equation (12). The 1970-2008 expanding and moving window cases show that several of the alternative models produce higher log scores in the 1970s and early 1980s. When we focus on the analogous graphs for the 2000-2008 OOS period, we see evidence of relative improvement in several of the non-benchmark alternatives in the last set of observations.

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.

Second, our primary contribution is to examine the extent to which there is OOS forecasting evidence in favor of such predictability. Via point forecasts, our key findings are that: (1) there is very strong evidence in favor of OOS oil price Granger causality for GDP from the 1970s through the mid-1980s; (2) this relationship weakened for roughly the following fifteen years; and (3) there was a strengthening of this relationship in the past decade. Though there are reasonable grounds to question the extent to which our density forecasts are well-calibrated, they do offer evidence that is consistent our point forecasts.

Additional work remains to be done. While we have established OOS Granger-causality from oil prices to US GDP, we should also consider the reverse problem: do GDP growth rates OOS Granger-cause crude oil prices? Barsky and Kilian (2002) argue that there may very well have been feedback from GDP growth to oil prices, and it seems reasonable to address that question in an OOS framework, and Kilian (2009) lists this as one of the two reasons why an exercise of the type we carry out “is not well defined.” The second concern Kilian (2009) raises is due to the issue of disentangling the macroeconomic role of oil demand and supply shocks. We will try to address that concern by examining the robustness of our results to inclusion of the measure he develops on real global economic activity to help us control for both demand and supply shifts in the international oil market.

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.
- Berkowitz, J., 2001. Testing density forecasts, with applications to risk management. *Journal of Business & Economic Statistics* 19 (4), 465–74.
- 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.
- Blanchard, O. J., Galí, J., 2008. 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?
- Blinder, A. S., Rudd, J. B., 2008. The supply-shock explanation of the great stagflation revisited. working paper, Princeton University.
- Clark, T., West, K., 2007. Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics* 138 (1), 291–311.
- 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.
- Diebold, F., Gunther, A., Tay, K., 1998. Evaluating density forecasts with applications to financial risk management. *International Economic Review* 39 (4), 863–883.
- 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.
- Giordani, P., Villani, M., 2009. Forecasting macroeconomic time series with locally adaptive signal extraction. Tech. Rep. forthcoming.

- 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., 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.
- Herrera, A. M., Pesavento, E., 2009. Oil price shocks, systematic monetary policy and the ‘great moderation’. *Macroeconomic Dynamics* 13 (1), 107–137.
- Hooker, M., 1996. What happened to the oil price-macroeconomy relationship? *Journal of Monetary Economics* 38 (2), 195–213.
- Kascha, C., Ravazzolo, F., 2010. Combining inflation density forecasts. *Journal of Forecasting* 29 (1-2), 231–250.
- 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.
- Koop, G., 2003. *Bayesian Econometrics*. Wiley.
- McCracken, M. W., 2007. Asymptotics for out of sample tests of granger causality. *Journal of Econometrics* 140 (2), 719–752.
- Mitchell, J., Hall, S., 2005. Evaluating, comparing and combining density forecasts using the KLIC with an application to the Bank of England and NIESER “fan” charts of inflation. *Oxford Bulletin of Economics and Statistics* 67 (S1).

- Mitchell, J., Wallis, K., Aug. 2009. Evaluating density forecasts: Forecast combinations, model mixtures, calibration and sharpness. NIESR Discussion Papers 320, National Institute of Economic and Social Research.
- Ravazzolo, F., Vahey, S., 2010. Forecast densities for economic aggregates from disaggregate ensembles. Tech. rep., Norges Bank working paper 2010/02.
- 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.
- Wallis, K., 2003. Chi-squared tests of interval and density forecasts, and the Bank of England's fan charts. *International Journal of Forecasting* 19 (2), 165–175.
- Welch, I., Goyal, A., 2008. A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies* 21 (4), 253–303.

Table 1: MSPEs for Quarterly US GDP Out-of-Sample Point Forecasts

	1970-2008	1975-2008	1980-2008	1985-2008	1990-2008	1995-2008	2000-2008	1970-1980	1970-1985	1975-1985
$h = 1$										
AR_e	0.572	0.500	0.396	0.193	0.220	0.223	0.258	1.083	1.178	1.236
AROIL_e	3.262	0.533	0.500	0.264	0.252	0.246	0.266	11.27	8.057	1.177
ARNOPI_e	1.572	0.660	0.441	0.232	0.226	0.225*	0.233**	4.853	3.716	1.685
AROILX_e	4.677	2.218	0.852	0.426	0.384	0.347	0.425	15.77	11.48	6.518
ARNOPIX_e	2.073	1.285	0.957	0.510	0.397	0.342	0.417	5.307	4.573	3.147
AR_m	0.600	0.547	0.404	0.199	0.226	0.231	0.266	1.171	1.242	1.381
AROIL_m	2.566	0.942	0.504	0.245	0.227	0.220***	0.236***	8.547	6.281	2.615
ARNOPI_m	1.142	0.672	0.448	0.227	0.213**	0.206***	0.219***	3.153	2.606	1.741
AROILX_m	7.574	3.125	0.995	0.590	0.447	0.339	0.327**	26.65	18.75	9.206
ARNOPIX_m	2.513	1.980	1.241	0.949	0.658	0.385	0.400*	6.202	5.015	4.453
TVPAR	0.581	0.534	0.393	0.201	0.227	0.238	0.275	1.124	1.187	1.333
TVPAROIL	2.691	1.200	0.806	0.333	0.283	0.253	0.298	8.157	6.463	3.280
TVPARNOPI	0.651	0.639	0.429	0.225	0.231	0.244	0.270**	1.294	1.332	1.633
$h = 4$										
AR_e	0.801	0.671	0.502	0.238	0.275	0.249	0.310	1.738	1.749	1.712
AROIL_e	4.824	5.125	0.623	0.338	0.340	0.267	0.331	17.99	12.38	16.61
ARNOPI_e	0.964	0.788	0.666	0.316	0.327	0.266	0.312*	1.897	2.055	1.922
AROILX_e	6.610	6.847	0.767	0.532	0.441	0.389	0.515	24.93	16.85	22.00
ARNOPIX_e	3.958	3.777	0.754	0.471	0.432	0.404	0.486	14.00	9.830	11.71
AR_m	0.779	0.651	0.467	0.234	0.272	0.254	0.304	1.758	1.697	1.651
AROIL_m	6.903	7.485	0.555	0.297	0.303	0.269	0.307	26.80	18.03	24.74
ARNOPI_m	1.448	1.316	0.586	0.270	0.271*	0.236***	0.278***	4.151	3.431	3.825
AROILX_m	7.291	7.281	1.689	0.983	0.694	0.551	0.670	24.86	17.92	22.40
ARNOPIX_m	6.560	5.918	2.016	1.174	0.778	0.658	0.749	20.81	15.63	17.30
TVPAR	0.868	0.707	0.529	0.267	0.308	0.295	0.353	1.930	1.880	1.765
TVPAROIL	2.805	2.877	0.699	0.488	0.489	0.407	0.522	9.409	6.707	8.610
TVPARNOPI	1.032	0.848	0.633	0.317	0.343	0.324	0.387	2.282	2.236	2.121

Notes: Table reports MSPEs in forecasting quarterly US GDP growth over various out-of-sample periods using the set of models described in Section 2 for two forecasting horizons, $h=1$ and $h=4$ steps ahead. For the expanding window and TVPAR forecasting schemes, the initial IS window is 1950Q1-1969Q4; for the moving window forecasting scheme, the initial IS window is 1955Q1-1969Q4. MSPEs for models that include an oil price measure which are lower than the MSPEs for the respective benchmarks without oil prices are reported in *italics*. ***, **, and * indicate that the null hypothesis of the McCracken (2007) tests for equal accuracy prediction of the parsimonious model without oil prices and the larger models with oil prices is rejected, respectively, at 1%, 5% and 10% significance level.

Table 2: Clark and West (2007) MSPEs-Adjusted for Quarterly US GDP Out-of-Sample Point Forecasts

	1970-2008	1975-2008	1980-2008	1985-2008	1990-2008	1995-2008	2000-2008	1970-1980	1970-1985	1975-1985
$h = 1$										
AR_e	0.572	0.500	0.396	0.193	0.220	0.223	0.258	1.083	1.178	1.236
AROIL_e	<i>-0.283**</i>	<i>0.380</i>	<i>0.363</i>	<i>0.173</i>	<i>0.200</i>	<i>0.198</i>	<i>0.215</i>	<i>-2.157**</i>	<i>-1.013**</i>	<i>0.875</i>
ARNOPI_e	<i>0.230**</i>	<i>0.570</i>	<i>0.371</i>	<i>0.173</i>	<i>0.187*</i>	<i>0.180**</i>	<i>0.180**</i>	<i>-0.179*</i>	<i>0.321*</i>	<i>1.524</i>
AROILX_e	<i>-0.748*</i>	<i>-0.093*</i>	<i>0.350</i>	<i>0.225</i>	<i>0.272</i>	<i>0.257</i>	<i>0.321</i>	<i>-3.992***</i>	<i>-2.304***</i>	<i>-0.856**</i>
ARNOPIX_e	<i>0.009**</i>	<i>0.370</i>	<i>0.356</i>	<i>0.239</i>	<i>0.281</i>	<i>0.257</i>	<i>0.318</i>	<i>-0.997**</i>	<i>-0.359**</i>	<i>0.684</i>
AR_m	0.600	0.547	0.404	0.199	0.226	0.231	0.266	1.171	1.242	1.381
AROIL_m	1.821	0.576	0.440	0.185	<i>0.154***</i>	<i>0.154***</i>	<i>0.155***</i>	5.825	4.439	1.514
ARNOPI_m	<i>-0.240*</i>	<i>0.451**</i>	<i>0.325*</i>	<i>0.136**</i>	<i>0.128***</i>	<i>0.123***</i>	<i>0.117***</i>	<i>-1.876*</i>	<i>-0.840*</i>	<i>1.207</i>
AROILX_m	3.868	1.121	0.605	0.286	<i>0.175</i>	<i>0.110***</i>	<i>0.031***</i>	13.330	9.598	3.127
ARNOPIX_m	<i>-1.522*</i>	<i>-1.642</i>	0.603	0.396	<i>0.180</i>	<i>0.001***</i>	<i>-0.109***</i>	<i>-7.685*</i>	<i>-4.592*</i>	<i>-6.535*</i>
TVPAR	0.581	0.534	0.393	0.201	0.227	0.238	0.275	1.124	1.187	1.333
TVPAROIL	<i>0.525</i>	0.786	0.546	<i>0.192*</i>	<i>0.201</i>	<i>0.180**</i>	<i>0.203*</i>	<i>0.464</i>	<i>1.057</i>	2.211
TVPARNOPI	0.574	0.593	0.399	<i>0.200</i>	<i>0.217</i>	<i>0.232</i>	<i>0.257</i>	<i>1.082</i>	<i>1.174</i>	1.537
$h = 4$										
AR_e	0.801	0.671	0.502	0.238	0.275	0.249	0.310	1.738	1.749	1.712
AROIL_e	<i>-0.012</i>	<i>-0.208</i>	<i>0.465</i>	<i>0.206</i>	<i>0.246</i>	<i>0.225</i>	<i>0.286</i>	<i>-1.507</i>	<i>-0.378</i>	<i>-1.201</i>
ARNOPI_e	<i>0.439</i>	<i>0.225</i>	0.520	<i>0.192*</i>	<i>0.237</i>	<i>0.213</i>	<i>0.250</i>	<i>0.185</i>	<i>0.855</i>	<i>0.302</i>
AROILX_e	1.045	0.908	<i>0.197**</i>	0.278	0.288	0.255	0.329	3.702	2.337	2.419
ARNOPIX_e	0.872	0.756	<i>0.201**</i>	0.249	<i>0.262</i>	<i>0.226</i>	<i>0.270</i>	2.973	1.921	1.972
AR_m	0.779	0.651	0.467	0.234	0.272	0.254	0.304	1.758	1.697	1.651
AROIL_m	4.554	4.861	0.478	0.242	<i>0.240</i>	<i>0.218**</i>	<i>0.251**</i>	17.33	11.82	15.95
ARNOPI_m	<i>-3.989*</i>	<i>-4.021*</i>	<i>0.465</i>	<i>0.191*</i>	<i>0.184***</i>	<i>0.165***</i>	<i>0.189***</i>	<i>-15.47*</i>	<i>-9.420*</i>	<i>-14.13*</i>
AROILX_m	1.868	1.419	0.825	0.436	0.276	<i>0.241</i>	<i>0.284</i>	5.138	4.279	3.776
ARNOPIX_m	<i>-1.185</i>	<i>-2.184*</i>	0.939	0.437	0.371	0.320	0.367	<i>-7.841</i>	<i>-3.915</i>	<i>-8.473*</i>
TVPAR	0.868	0.707	0.529	0.267	0.308	0.295	0.353	1.930	1.880	1.765
TVPAROIL	<i>0.141</i>	<i>-0.058</i>	<i>0.441*</i>	<i>0.255</i>	<i>0.294</i>	<i>0.281</i>	0.408	<i>-0.797</i>	<i>-0.050</i>	<i>-0.811</i>
TVPARNOPI	0.951	0.761	0.579	0.281	0.324	0.314	0.377	2.117	2.079	1.911

Notes: Table reports the MSPEs-adjusted proposed by Clark and West (2007) in forecasting US GDP growth over various out-of-sample periods using the set of models described in Section 2 for two forecasting horizons, $h=1$ and $h=4$ steps ahead; MSPEs for the parsimonious benchmark models are not adjusted. For the expanding window and TVPAR forecasting schemes, the initial IS window is 1950Q1-1969Q4; for the moving window forecasting scheme, the initial IS window is 1955Q1-1969Q4. MSPEs-adjusted for models that include an oil price measure which are lower than the MSPE for the respective benchmarks without oil are reported in *italics*. ***, **, and * indicate that the null of the Clark and West (2007) test for equal accuracy prediction of the parsimonious model without oil prices and the larger models with oil prices is rejected, respectively, at 1%, 5% and 10% significance level.

Table 3: p -Values for Berkowitz (2001) Out-of-Sample Density Forecast LR Test

	1970-2008	1975-2008	1980-2008	1985-2008	1990-2008	1995-2008	2000-2008	1970-1980	1970-1985	1975-1985
$h = 1$										
AR_e	0.292	0.707	0.034	0.000	0.000	0.000	0.000	0.292	0.292	0.707
AROIL_e	0.034	0.000	0.000	0.000	0.000	0.000	0.000	0.060	0.301	0.619
ARNOPI_e	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.472	0.619	0.878
AROILX_e	0.000	0.081	0.000	0.000	0.000	0.000	0.003	0.453	0.472	0.971
ARNOPIX_e	0.035	0.000	0.000	0.000	0.000	0.000	0.001	0.151	0.266	0.473
AR_m	0.000	0.000	0.007	0.000	0.000	0.002	0.000	0.343	0.971	0.664
AROIL_m	0.000	0.009	0.000	0.000	0.000	0.001	0.000	0.289	0.151	0.692
ARNOPI_m	0.000	0.166	0.000	0.000	0.000	0.000	0.000	0.104	0.309	0.767
AROILX_m	0.959	0.000	0.056	0.000	0.000	0.000	0.006	0.055	0.664	0.120
ARNOPIX_m	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.448	0.289	0.977
TVPAR	0.000	0.569	0.000	0.000	0.000	0.000	0.000	0.251	0.447	0.391
TVPAROIL	0.000	0.000	0.125	0.001	0.000	0.000	0.000	0.020	0.767	0.441
TVPARNOPI	0.344	0.000	0.000	0.000	0.000	0.000	0.000	0.271	0.055	0.410
$h = 4$										
AR_e	0.001	0.019	0.000	0.000	0.000	0.000	0.000	0.001	0.001	0.019
AROIL_e	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.802	0.867	0.088
ARNOPI_e	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.045	0.088	0.004
AROILX_e	0.000	0.023	0.001	0.000	0.000	0.000	0.000	0.002	0.045	0.054
ARNOPIX_e	0.510	0.001	0.000	0.000	0.000	0.000	0.001	0.927	0.000	0.214
AR_m	0.001	0.000	0.000	0.000	0.000	0.001	0.000	0.015	0.054	0.005
AROIL_m	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.927	0.020
ARNOPI_m	0.000	0.003	0.000	0.000	0.000	0.000	0.000	0.075	0.001	0.003
AROILX_m	0.014	0.000	0.000	0.000	0.000	0.000	0.006	0.001	0.005	0.001
ARNOPIX_m	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.001	0.053
TVPAR	0.000	0.046	0.000	0.000	0.000	0.000	0.000	0.119	0.011	0.072
TVPAROIL	0.000	0.000	0.002	0.000	0.000	0.000	0.000	0.000	0.003	0.005
TVPARNOPI	0.046	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.001	0.001

Notes: Table reports the p -values of the Berkowitz (2001) likelihood ratio test of a zero mean, unit variance, and independence of $PITs$ of the out-of-sample density forecasts of quarterly US GDP growth over various samples for the set of models described in Section 2 for two forecasting horizons, $h=1$ and $h=4$ steps ahead. For the expanding window and TVPAR forecasting schemes, the initial IS window is 1950Q1-1969Q4; for the moving window forecasting scheme, the initial IS window is 1955Q1-1969Q4.

Table 4: p -Values for Ljung-Box Out-of-Sample Density Forecast Test

	1970-2008	1975-2008	1980-2008	1985-2008	1990-2008	1995-2008	2000-2008	1970-1980	1975-1985
$h = 1$									
AR_e	0.998	0.984	0.991	0.873	0.828	0.942	0.943	0.998	0.984
AROIL_e	0.991	0.873	0.828	0.942	0.955	0.861	0.951	0.993	0.967
ARNOPI_e	0.828	0.942	0.972	0.951	0.925	0.961	0.896	1.000	0.995
AROILX_e	0.898	0.978	0.905	0.925	0.863	0.884	0.962	1.000	0.987
ARNOPIX_e	0.998	0.905	0.925	0.959	0.896	0.923	0.901	0.994	0.976
AR_m	0.905	0.925	0.970	0.952	0.848	0.864	0.921	1.000	0.995
AROIL_m	0.925	0.949	0.983	0.884	0.947	0.998	0.927	0.998	0.984
ARNOPI_m	0.947	0.992	0.896	0.996	0.889	0.943	0.988	0.720	0.891
AROILX_m	0.981	0.952	0.993	0.947	0.938	0.927	0.958	0.936	0.981
ARNOPIX_m	0.983	0.884	0.708	0.840	0.734	0.996	0.989	1.000	0.987
TVPAR	0.896	0.749	0.947	0.920	0.943	0.844	0.246	0.995	0.847
TVPAROIL	0.910	0.708	0.849	0.998	0.771	0.988	0.916	0.996	0.944
TVPARNOPI	0.960	0.947	0.961	0.980	0.830	0.988	0.987	0.990	0.997
$h = 4$									
AR_e	0.390	0.473	0.415	0.206	0.304	0.812	0.985	0.390	0.473
AROIL_e	0.415	0.206	0.304	0.812	0.838	0.586	0.860	0.834	0.809
ARNOPI_e	0.304	0.812	0.955	0.957	0.939	0.678	0.434	0.463	0.332
AROILX_e	0.616	0.859	0.839	0.939	0.552	0.637	0.910	0.580	0.778
ARNOPIX_e	0.796	0.839	0.939	0.849	0.927	0.795	0.434	0.826	0.918
AR_m	0.839	0.939	0.611	0.576	0.089	0.574	0.719	0.186	0.182
AROIL_m	0.939	0.858	0.785	0.734	0.795	0.680	0.914	0.389	0.473
ARNOPI_m	0.341	0.400	0.927	0.282	0.298	0.709	0.325	0.854	0.688
AROILX_m	0.652	0.576	0.424	0.795	0.971	0.837	0.501	0.333	0.152
ARNOPIX_m	0.785	0.629	0.282	0.910	0.042	0.546	0.236	0.580	0.778
TVPAR	0.927	0.735	0.795	0.874	0.517	0.831	0.859	0.446	0.822
TVPAROIL	0.206	0.282	0.696	0.752	0.899	0.759	0.170	0.200	0.688
TVPARNOPI	0.735	0.795	0.811	0.373	0.357	0.836	0.916	0.252	0.410

Notes: Table reports the p -values for the Ljung-Box test that the PII_t s of the out-of-sample quarterly US GDP growth density forecasts are serially uncorrelated over various samples using the set of models described in Section 2 for two forecasting horizons, $h=1$ and $h=4$ steps ahead. For the expanding window and TVPAR forecasting schemes, the initial IS window is 1950Q1-1969Q4; for the moving window forecasting scheme, the initial IS window is 1955Q1-1969Q4.

Table 5: p -Values for Anderson-Darling Out-of-Sample Density Forecast Test

	1970-2008	1975-2008	1980-2008	1985-2008	1990-2008	1995-2008	2000-2008	1970-1980	1975-1985	
$h = 1$										
AR_e	0.375	0.542	0.040	0.000	0.002	0.002	0.004	0.379	0.381	0.524
AROIL_e	0.043	0.000	0.001	0.002	0.001	0.007	0.005	0.034	0.181	0.912
ARNOPI_e	0.002	0.002	0.013	0.001	0.007	0.006	0.010	0.334	0.915	0.864
AROILX_e	0.003	0.212	0.012	0.006	0.004	0.003	0.041	0.481	0.327	0.875
ARNOPIX_e	0.013	0.010	0.005	0.004	0.002	0.038	0.018	0.191	0.307	0.874
AR_m	0.010	0.006	0.028	0.002	0.003	0.037	0.008	0.413	0.879	0.760
AROIL_m	0.006	0.037	0.035	0.003	0.036	0.029	0.004	0.364	0.190	0.538
ARNOPI_m	0.016	0.291	0.003	0.001	0.018	0.014	0.003	0.189	0.795	0.986
AROILX_m	0.671	0.003	0.047	0.039	0.032	0.008	0.014	0.263	0.766	0.047
ARNOPIX_m	0.030	0.003	0.004	0.007	0.005	0.005	0.009	0.479	0.372	0.859
TVPAR	0.003	0.305	0.035	0.009	0.013	0.006	0.045	0.235	0.435	0.421
TVPAROIL	0.014	0.003	0.171	0.032	0.009	0.010	0.021	0.121	0.987	0.738
TVPARNOPI	0.169	0.037	0.005	0.003	0.001	0.008	0.004	0.502	0.262	0.559
$h = 4$										
AR_e	0.152	0.320	0.030	0.001	0.004	0.003	0.001	0.157	0.154	0.320
AROIL_e	0.030	0.001	0.004	0.003	0.001	0.008	0.008	0.570	0.611	0.484
ARNOPI_e	0.006	0.003	0.005	0.001	0.002	0.008	0.018	0.095	0.481	0.550
AROILX_e	0.009	0.065	0.066	0.001	0.016	0.013	0.051	0.102	0.097	0.280
ARNOPIX_e	0.195	0.065	0.002	0.007	0.005	0.047	0.058	0.774	0.159	0.635
AR_m	0.064	0.002	0.145	0.018	0.008	0.107	0.001	0.447	0.280	0.558
AROIL_m	0.002	0.014	0.040	0.010	0.044	0.076	0.013	0.150	0.775	0.317
ARNOPI_m	0.054	0.460	0.005	0.002	0.069	0.092	0.050	0.190	0.788	0.041
AROILX_m	0.081	0.017	0.050	0.044	0.019	0.003	0.014	0.060	0.548	0.191
ARNOPIX_m	0.038	0.014	0.002	0.010	0.061	0.011	0.000	0.100	0.150	0.279
TVPAR	0.005	0.251	0.047	0.010	0.229	0.014	0.005	0.445	0.045	0.791
TVPAROIL	0.075	0.002	0.124	0.066	0.003	0.065	0.001	0.000	0.036	0.120
TVPARNOPI	0.249	0.042	0.030	0.017	0.001	0.049	0.003	0.608	0.056	0.458

Notes: Table reports the (simulated) small sample (equal to the number of observations in each out-of-sample period) p -values for the Anderson-Darling test that the $PITs$ of the out-of-sample quarterly US GDP growth density forecasts are uniformly distributed over various samples using the set of models described in Section 2 for two forecasting horizons, $h=1$ and $h=4$ steps ahead. For the expanding window and TVPAR forecasting schemes, the initial IS window is 1950Q1-1969Q4; for the moving window forecasting scheme, the initial IS window is 1955Q1-1969Q4.

Table 6: p -Values of χ^2 Out-of-Sample Density Forecast Test

	1970-2008	1975-2008	1980-2008	1985-2008	1990-2008	1995-2008	2000-2008	1970-1980	1975-1985	
$h = 1$										
AR_e	0.022	0.059	0.002	0.000	0.000	0.000	0.000	0.022	0.022	0.059
AROIL_e	0.002	0.000	0.000	0.000	0.000	0.001	0.000	0.034	0.016	0.408
ARNOPI_e	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.540	0.408	0.684
AROILX_e	0.000	0.014	0.029	0.000	0.000	0.000	0.034	0.540	0.540	0.934
ARNOPIX_e	0.025	0.029	0.000	0.000	0.000	0.007	0.000	0.267	0.494	0.239
AR_m	0.029	0.000	0.000	0.000	0.000	0.005	0.000	0.167	0.934	0.101
AROIL_m	0.000	0.000	0.004	0.000	0.016	0.001	0.001	0.022	0.267	0.059
ARNOPI_m	0.000	0.636	0.000	0.000	0.003	0.003	0.000	0.267	0.077	0.866
AROILX_m	0.540	0.000	0.002	0.016	0.019	0.000	0.003	0.101	0.101	0.333
ARNOPIX_m	0.004	0.000	0.000	0.001	0.000	0.000	0.000	0.540	0.022	0.934
TVPAR	0.000	0.684	0.016	0.001	0.003	0.000	0.067	0.088	0.022	0.333
TVPAROIL	0.000	0.000	0.299	0.001	0.000	0.002	0.001	0.212	0.866	0.494
TVPARNOPI	0.540	0.016	0.001	0.000	0.000	0.001	0.000	0.825	0.101	0.333
$h = 4$										
AR_e	0.369	0.148	0.000	0.000	0.000	0.000	0.000	0.369	0.369	0.148
AROIL_e	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.189	0.299	0.189
ARNOPI_e	0.000	0.000	0.003	0.001	0.000	0.000	0.002	0.333	0.189	0.540
AROILX_e	0.001	0.077	0.051	0.000	0.002	0.000	0.006	0.333	0.333	0.239
ARNOPIX_e	0.059	0.051	0.000	0.000	0.000	0.039	0.008	0.540	0.115	0.494
AR_m	0.051	0.000	0.212	0.002	0.000	0.115	0.000	0.369	0.239	0.825
AROIL_m	0.000	0.000	0.012	0.000	0.039	0.009	0.002	0.369	0.540	0.148
ARNOPI_m	0.009	0.733	0.000	0.000	0.025	0.034	0.014	0.540	0.960	0.034
AROILX_m	0.408	0.002	0.001	0.039	0.009	0.000	0.067	0.101	0.825	0.333
ARNOPIX_m	0.012	0.000	0.000	0.003	0.025	0.006	0.000	0.333	0.369	0.239
TVPAR	0.000	0.189	0.039	0.003	0.167	0.000	0.004	0.684	0.067	0.780
TVPAROIL	0.008	0.000	0.189	0.006	0.000	0.014	0.000	0.029	0.034	0.267
TVPARNOPI	0.189	0.039	0.006	0.006	0.001	0.059	0.000	0.333	0.101	0.494

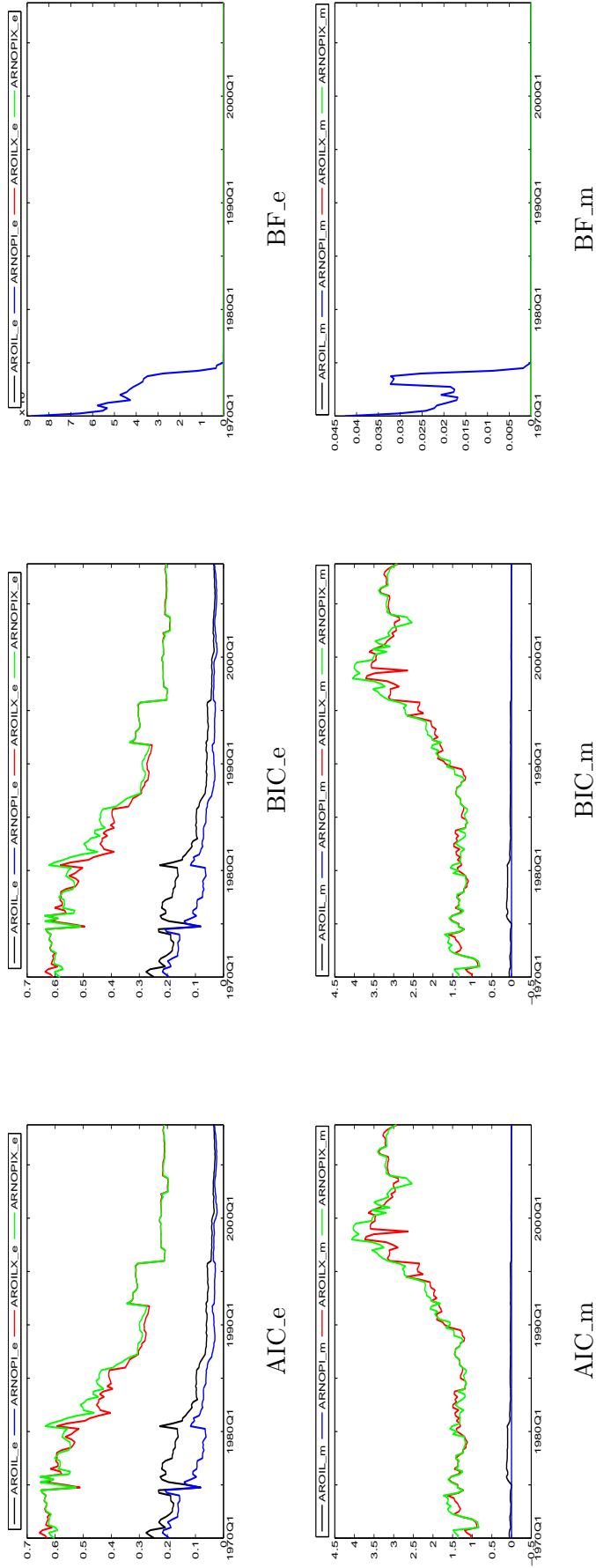
Notes: Table reports the p -values for the Pearson chi-square test that the PIT s of the out-of-sample quarterly US GDP growth density forecasts are uniformly distributed over various samples using the set of models described in Section 2 for two forecasting horizons, $h=1$ and $h=4$ steps ahead. For the expanding window and TVPAR forecasting schemes, the initial IS window is 1950Q1-1969Q4; for the moving window forecasting scheme, the initial IS window is 1955Q1-1969Q4.

Table 7: Log Scores for Out-of-Sample Density Forecasts

	1970-2008	1975-2008	1980-2008	1985-2008	1990-2008	1995-2008	2000-2008	1970-1980	1970-1985	1975-1985
$h = 1$										
AR.e	-1.123	-1.100	-1.071	-1.001	-1.012	-1.014	-1.030	-1.273	-1.318	-1.338
AROIL.e	-1.149	-1.110	-1.101	-1.021	-1.022	-1.021	-1.030	-1.289	-1.353	-1.324**
ARNOPI.e	-1.131	-1.110	-1.075	-1.009	-1.011	-1.012	-1.016	-1.292	-1.325	-1.352
AROILX.e	-1.166	-1.149	-1.138	-1.073	-1.069	-1.056	-1.083	-1.245***	-1.314***	-1.333*
ARNOPIX.e	-1.168	-1.171	-1.160	-1.095	-1.073	-1.052	-1.077	-1.190	-1.285	-1.354
AR.m	-1.168	-1.153	-1.121	-1.063	-1.090	-1.113	-1.145	-1.302	-1.335	-1.370
AROIL.m	-1.176	-1.152	-1.141	-1.062	-1.074	-1.093	-1.106	-1.278	-1.357	-1.368
ARNOPI.m	-1.153	-1.148	-1.112	-1.052	-1.064	-1.080	-1.091	-1.272	-1.315	-1.377
AROILX.m	-1.190	-1.173	-1.162	-1.123	-1.108	-1.077	-1.042	-1.273	-1.299***	-1.293**
ARNOPIX.m	-1.209	-1.223	-1.220	-1.202	-1.164	-1.094	-1.083	-1.177	-1.219	-1.273
TVPAR	-1.150	-1.097	-1.017	-0.886	-0.890	-0.897	-0.917	-1.535	-1.572	-1.603
TVPAROIL	-1.757	-1.729	-1.721	-1.646	-1.627	-1.604	-1.613	-1.862	-1.935	-1.926
TVPARNOPI	-1.244	-1.208	-1.126	-1.001	-0.983	-0.986	-1.000	-1.585	-1.631	-1.702
$h = 4$										
AR.e	-1.189	-1.149	-1.100	-1.016	-1.031	-1.021	-1.047	-1.466	-1.481	-1.468
AROIL.e	-1.198	-1.163	-1.139	-1.046	-1.051	-1.027	-1.053	-1.381**	-1.454***	-1.446**
ARNOPI.e	-1.186**	-1.143	-1.146	-1.039	-1.047	-1.026	-1.044	-1.312*	-1.434*	-1.393
AROILX.e	-1.212	-1.160	-1.132	-1.106	-1.086	-1.068	-1.114	-1.463	-1.391***	-1.292***
ARNOPIX.e	-1.231	-1.163	-1.120	-1.088	-1.083	-1.074	-1.102	-1.581	-1.472	-1.344***
AR.m	-1.224	-1.182	-1.134	-1.076	-1.108	-1.124	-1.167	-1.506	-1.474	-1.437
AROIL.m	-1.215	-1.189	-1.161	-1.090	-1.113	-1.127	-1.161	-1.387	-1.426	-1.426
ARNOPI.m	-1.216	-1.175	-1.156	-1.072	-1.090	-1.095	-1.125	-1.405	-1.458	-1.424
AROILX.m	-1.274	-1.251	-1.249	-1.232	-1.195	-1.181	-1.218	-1.355	-1.346	-1.296
ARNOPIX.m	-1.315	-1.272	-1.280	-1.247	-1.208	-1.211	-1.234	-1.424	-1.430	-1.332
TVPAR	-1.276	-1.207	-1.115	-0.980	-0.990	-0.963	-1.011	-1.779	-1.774	-1.750
TVPAROIL	-1.681	-1.674	-1.575	-1.535	-1.522	-1.507	-1.521	-2.013	-1.927	-2.007
TVPARNOPI	-1.424	-1.354	-1.264	-1.118	-1.097	-1.071	-1.131	-1.925	-1.939	-1.919

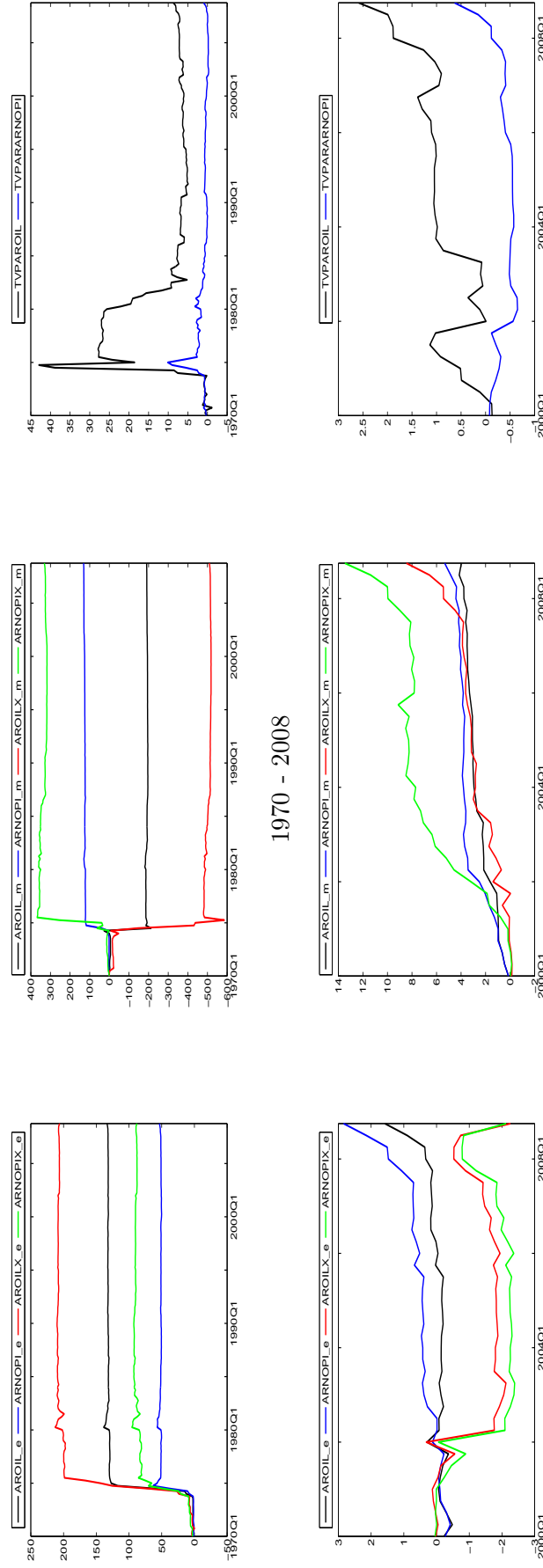
Notes: Table reports the log scores of the out-of-sample quarterly US GDP growth density forecasts over various samples using the set of models described in Section 2 for two forecasting horizons, $h=1$ and $h=4$ steps ahead. For the expanding window and TVPAR forecasting schemes, the initial IS window is 1950Q1-1969Q4; for the moving window forecasting scheme, the initial IS window is 1955Q1-1969Q4. Log scores for models that include an oil price measure which are higher than the log scores for the respective benchmarks without oil are reported in *italics*. ***, **, * and * indicate that the null of the log score tests of equal density predictive accuracy relative to the benchmarks without oil is rejected, respectively, at 1%, 5% and 10% significance level; the Mitchell and Hall (2005) and Amisano and Giacomini (2007) tests are used, respectively, for the expanding window and moving window cases.

Figure 1: Model Selection Criteria Across Estimation Windows



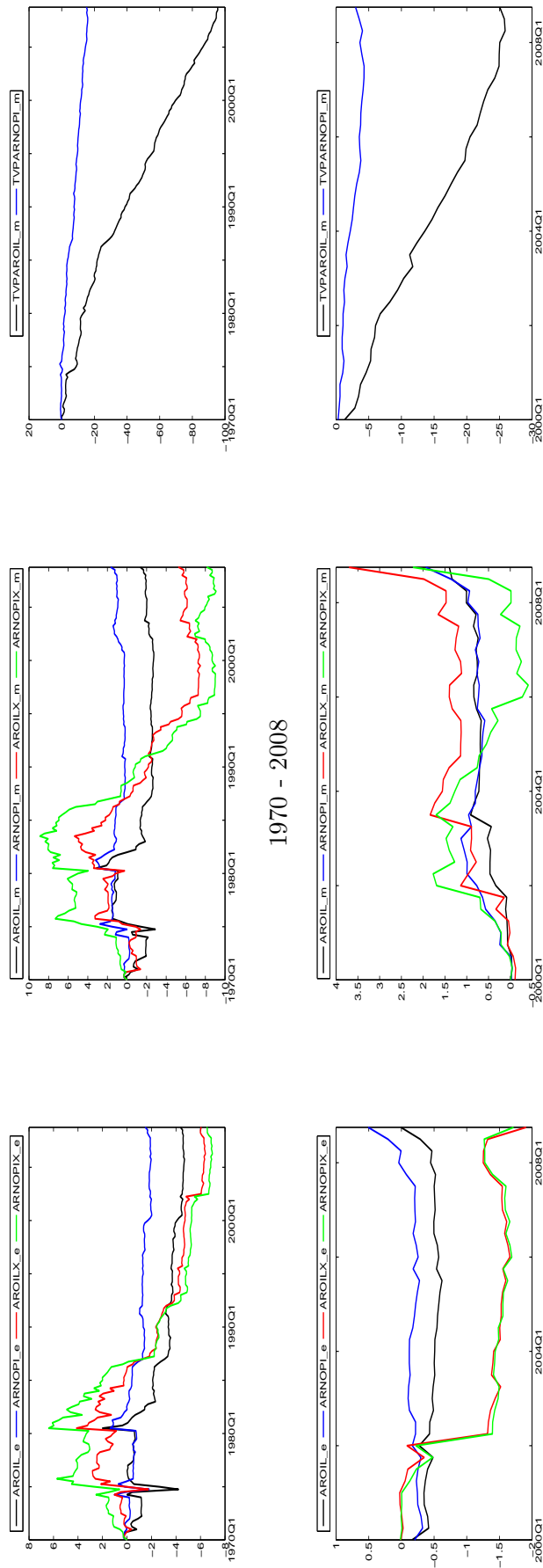
Note: The graphs show differences in AIC differences (AIC(benchmark) - AIC(alternative)), differences in BIC (BIC(benchmark) - BIC(alternative)), Bayes Factors (Prob(benchmark)/Prob(alternative)), where 'Prob' represents marginal likelihood for the benchmark model without oil prices and alternative models with an oil price measure included across expanding estimation windows (top panels) and fixed length 15-year moving estimation windows of IS real-time data (bottom panels); if the benchmark model generates the better fit, then the AIC and BIC differences are positive and the Bayes factor is less than one. For the expanding windows, the first and last IS periods are, respectively, 1950Q1-1969Q4 and 1950Q1-2008Q4; for the moving windows, the first and last IS periods are, respectively, 1955Q1-1969Q4 and 1994Q1-2008Q4.

Figure 2: Cumulative Square Prediction Error Difference, 1-Step Ahead Forecasts



Note: The graphs show the Cumulative Square Prediction Error Difference (CSPED), relative to the benchmark model, for models with oil prices for OOS forecasting samples 1970-2008 (top panel) and 2000-2008 (bottom panel) for the 1-step ahead horizon. Results are shown for the expanding window, moving window, and time-varying estimation schemes in the left, middle, and right panels, respectively.

Figure 3: Cumulative Log Score Difference, 1-step Ahead Forecasts



Note: The graphs show the Cumulative Log Score Difference (CLSD), relative to the benchmark model, for models with oil prices for the OOS forecasting samples 1970-2008 (top panel) and 2000-2008 (bottom panel) for the 1-step ahead horizon. Results are shown for the expanding window, moving window, and time-varying estimation schemes in the left, middle, and right panels, respectively.