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

Francesco Ravazzolo[*]
Senior Researcher, Research Department
Norges Bank
Researcher 2
BI Norwegian Business School
Philip Rothman[†]
Professor of Economics
East Carolina University[‡]

April 1, 2012

[*] 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 two anonymous referees, Christiane Baumeister, Hilde Bjørnland, Efrem Castelnuovo, Todd Clark, Vincent Labhard, Lutz Kilian, Mike McCracken, Ken West, and seminar participants at the ECB, EUI, ICEEE 2011, SNDE 2011 conference, the Norges Bank conference on 'Recent Developments in the Econometrics of Macroeconomics and Finance,' the 6^{th} Eurostat Colloquium on 'Modern Tools for Business Cycle Analysis: the Lessons from Global Economic Crisis,' 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 updated nominal index of bulk dry cargo ocean shipping freight rates. The views expressed in this paper are our own and do not necessarily reflect those of Norges Bank.

Abstract

We study the real-time predictive content of crude oil prices for US real GDP growth through a pseudo out-of-sample (OOS) forecasting exercise. Comparing our benchmark model "without oil" against alternatives "with oil," we strongly reject the null hypothesis of no OOS population-level predictability from oil prices to GDP at the longer forecast horizon we consider. This examination of the global OOS relative performance of the models we consider is robust to use of ex-post revised data. But when we focus on the forecasting models' local relative performance, we observe strong differences across use of real-time and ex-post revised data.

JEL Classification: C22, C53, E32, E37

Keywords: oil, GDP, real-time data, predictability, forecasting, recessions

1 Introduction

The goal of this paper is to investigate the predictive relationship between oil prices and US GDP by way of a pseudo real-time out-of-sample (OOS) forecasting exercise. More specifically, we study whether inclusion of oil prices in autoregressive benchmark models helps improve real-time OOS forecasts of real GDP growth rates. We do so conditional on the extensive literature which has explored the relationship between these variables following the seminal paper of Hamilton (1983).[1] A key quantitative question running through this primarily in-sample (IS) literature is whether oil prices have predictive power for GDP.

Bachmeier, Li, and Liu (2008) were among the first to consider this problem within an OOS framework. Using both parametric and nonparametric methods, they strongly conclude that oil prices do not have predictive content for GDP. Their models are estimated with data from the early 1960s and, in some cases, from the mid 1950s, and the oil price measure they employ is the West Texas Intermediate (WTI) spot price.

However, Alquist, Kilian, and Vigfusson (2011) suggest caution against estimation of predictive regressions with pre-1973 oil prices and use of the WTI data. First, since the pre-1973 nominal WTI price was adjusted only at discrete intervals, standard time series techniques are not applicable; this feature of the nominal data also implies problems for use of the associated real WTI price for this period. It follows that it is inappropriate to combine pre-1973 and post-1973 WTI data. Second, they emphasize that the WTI price may not be an accurate measure of the price faced by oil refiners between 1974 and the ending of price controls for the WTI price, since the WTI price was regulated up to the mid-1980s and the import share of oil used in the U.S. increased sharply after 1973. Accordingly, they argue in favor of using data on the refiners' acquisition cost (RAC) of crude oil provided by the U.S. Energy Information Agency (EIA).[2]

Alquist, Kilian, and Vigfusson (2011) find that inclusion of crude oil prices in linear VARs leads to only small improvements in forecasting cumulative real GDP growth. When they allow the predictive relationship between oil prices and real GDP growth to be nonlinear, as in Kilian and Vigfusson (2011), there are larger improvements in forecasting cumulative

[1] Important work in this literature includes, among others, Hamilton (1996), Hooker(1996), Bernanke, Gertler, and Watson (1997), Barsky and Kilian (2002), Hamilton (2003), Barsky and Kilian (2004), Hamilton and Herrera (2004), Baumeister and Peersman (2012), Kilian (2008), Edelstein and Kilian (2009), Hamilton (2009), Kilian (2009), and Kilian (2010).

[2] Three RAC series are available through the U.S. EIA: the RAC for domestically produced oil; the RAC for imported oil; and a composite measure, which is a weighted average of the RACs for domestic and imported crude oil. These monthly data series each begin in January of 1974.

real GDP growth for some specifications, e.g., reductions of Mean Squared Prediction Error (MSPE) over the linear AR(4) benchmark of up to 12%. But they are skeptical of the forecast gains provided by these nonlinear models since they imply the 2007-2009 financial crisis played no role in the real GDP declines of 2008-2009, and because they often generate false positive signals of recession conditional on significant oil price increases.

The OOS real GDP forecast comparisons in Bachmeier, Li, and Liu (2008) and Alquist, Kilian, and Vigfusson (2011) are not conducted in real time, i.e., they are done with use of ex-post revised, not real-time, data. This is of concern since the RAC and real GDP data, as well as other data frequently used in such studies, are revised over time, such that use of ex-post revised versions of these time series assumes the forecaster's information set contains data that would, in fact, be unavailable when constructing the forecasts.[3] Another general ex-post revised versus real-time data issue is that some predictors may be available only with a delay. Due to these concerns, use of ex-post revised data may give a misleading impression of the relative real-time OOS forecasting performance of the alternative models considered. Accordingly, the main empirical issue we examine in this paper is the extent to which imposition of real-time data constraints affects the OOS predictive content of crude oil prices for U.S. real GDP growth rates; we do so using the RAC composite series as our nominal crude oil price measure and by estimating predictive regressions with post-1973 data.

Carlton (2011) carries out an arguably less comprehensive 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 evidence of predictability from oil prices to GDP growth. She presents an interesting interpretation of this apparent predictive content of oil prices for real GDP, by arguing that they may help shorten the "recognition lag" about the state of the business cycle and thereby help improve the efficacy of counter-cyclical stabilization policies. However, her results are subject to the Alquist, Kilian, and Vigfusson (2011) critique mentioned above, since she combines pre- and post-1973 WTI and PPI crude oil price data to estimate her models; she does not use any RAC data in her analysis.

Our main results are as follows.We find very strong statistically significant OOS predictability from oil prices to GDP at the longer forecast horizon we consider, but not at

[3] Baumeister and Kilian (2011) report that the RAC for imported crude oil is revised an average of 1.21 times. The frequency of real GDP revisions is much higher; following the Advance release, there are revisions through the Second and Third releases as well as the One-Year and (roughly five-year) Comprehensive revisions.

the shorter one; the economic significance of the forecast improvements appears to be small. Further examination suggests that the longer horizon results may be due some of 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. These results are similar across use of ex-post revised and real-time data. But when we examine the time path of the models' relative OOS performance, we find that imposition of real-time data constraints does indeed affect the statistical significance of the predictive content of oil prices for GDP.

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

2 Forecasting GDP with Oil Prices

We use data for US real 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 past issues of the EIA's *Petroleum Marketing Monthly* (*PMM*) available in electronic form, we constructed vintages of real-time data for the composite RAC; we use the value of the composite RAC in the third month of the quarter as the quarterly value.[4] 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 Reserve of Saint Louis. We deflate the nominal index of bulk dry cargo ocean shipping freight series of Kilian (2009) by the CPI and then detrend to compute a measure of real global activity for each IS period; the nominal shipping index is available in real time and is not subject to revisions.

We generate $h-$step ahead real-time OOS forecasts, for $h = 1$ and $h = 4$, of quarterly US real GDP growth rates. Our $h = 1$ forecast is a "nowcast" of the quarter $t + 1$ real GDP growth rate using real-time data vintage $t + 1$. This vintage contains the first release of real GDP for quarter t and the first releases of the CPI and nominal composite RAC for the third month of quarter t; the nominal value of the composite RAC is deflated by the CPI to compute the value of the real composite RAC for quarter t. Since this value of the nominal composite RAC is typically released at the beginning of the third month of quarter $t + 1$, our nowcasting exercise mimics a forecast of the quarter $t + 1$ real GDP growth rate being

[4] The date of the first issue of the *PMM* available in this form is 1998M1. Issues of the *PMM* include RAC data for at most three years, so that we backcasted by approximating pre-1995M1 data with ex-post revised data; a similar approach is used by Baumeister and Kilian (2011). When we assembled this data set, the most recent data available were for 2011M1.

generated near the start of the third month of quarter $t + 1$.

The real-time OOS forecast errors are computed with the actual data realization of real GDP given by the first release value (from vintage $t + 2$ in the $h = 1$ case and from vintage $t + 5$ in the $h = 4$ case). For all the models we use direct forecasting for the $4-$step ahead forecasts, such that we do not employ multi-equation systems to produce these forecasts; in contrast, both Bachmeier, Li, and Liu (2008) and Alquist, Kilian, and Vigfusson (2011) use two-equation regressions to generate multi-step-ahead forecasts.

2.1 Predictive Regressions

A standard benchmark to forecast real GDP growth at horizon h is an autoregressive model of order p.

$$\Delta y_{t+h} = \alpha + \sum_{i=0}^{p-1} \beta_i \Delta y_{t-i} + \sigma \epsilon_{t+h}, \tag{1}$$

where $\Delta y_t = \log GDP_t - \log GDP_{t-1}$, GDP_t = real GDP for observation t, and $\epsilon_{t+h} \sim WN$(0,1). In the oil and the macroeconomy literature, the lag order p is often set equal to 4 with quarterly data; see, for example, Hamilton (2003). We follow this practice.[5] The model is estimated and point forecasts are produced via a sequence of recursive windows. The first recursive window IS period is 1975Q1-1989Q4; as per the discussion above, the model is estimated using the 1990Q1 real-time data vintage. For $h = 1$ ($h = 4$), the last IS period is 1975Q1-2009Q3 (1975Q1-2008Q4).

Next we extend the AR(4) benchmark with an oil price measure:

$$\Delta y_{t+h} = \alpha + \sum_{i=0}^{3} \beta_i \Delta y_{t-i} + \sum_{i=0}^{3} \delta_i oil_{t-i} + \sigma \epsilon_{t+h}, \tag{2}$$

where $\epsilon_{t+h} \sim WN$(0,1) and oil_t is the oil price measure at time t. Alquist, Kilian, and Vigfusson (2011) present an extensive discussion on whether one should focus on the predictive content of nominal or real oil prices for real GDP; for completeness, we include both in our study. We use three measures of oil_t: the nominal composite RAC growth rate; the real composite RAC growth rate; and the Net Oil Prince Increase (NOPI) indicator proposed by Hamilton (1996), $oil_t = \max[(\ln(p_t) - \max[\ln(p_{t-1}), .., \ln(p_{t-4})]), 0]$, where p_t is the nominal

[5] We obtain similar results, not reported here, when we identify p according to the AIC.

composite RAC.[6] This leads to three alternatives to the AR(4) benchmark: ADL(4,4)^{nrac}, ADL(4,4)^{rrac}, and ADL(4,4)^{nopi}, where the superscripts '*nrac*,' '*rrac*,' and '*nopi*' indicate, respectively, that the autoregressive distributed lag alternative model includes 4 lags of the nominal composite RAC growth rate, the real composite RAC 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 sensitive to an omitted variable in models (1) and (2). To examine this question, we also consider the following benchmark model:

$$\Delta y_{t+h} = \alpha + \sum_{i=0}^{3} \beta_i \Delta y_{t-i} + \sum_{i=0}^{3} \delta_i z_{t-i} + \sigma \epsilon_{t+h}, \tag{3}$$

where $\epsilon_{t+h} \sim WN(0,1)$ and z_t is a non-oil-price macro variable. As an alternative to these benchmarks, we add an oil price measure:

$$\Delta y_{t+h} = \alpha + \sum_{i=0}^{3} \beta_i \Delta y_{t-i} + \sum_{i=0}^{3} \delta_i z_{t-i} + \sum_{i=0}^{3} \gamma_i oil_{t-i} + \sigma \epsilon_{t+h}, \tag{4}$$

where $\epsilon_{t+h} \sim WN(0,1)$.

The set of potential macro variables z_t to include in forecast comparisons between models (3) and (4) is very large. To guide our choices, we draw upon the literatures which have identified variables as leading indicators of the U.S. business cycle and those variables which may have predictive content for oil prices; see, for example, Estrella and Hardouvelis (1991), Hooker (1996), Stock and Watson (1999), Wright (2006), and Kilian (2009). Initially, the full set of z_t variables we consider are growth rates of the import price deflator, personal consumption expenditures deflator, the global activity measure 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. For each z_t, we then generate forecasts over the OOS periods described below using model (3) and compare the MSPE to that from OOS forecasts from model (1). Adding z_t to the AR(4) benchmark leads to a lower OOS MSPE in only two cases: $z_t =$ the

[6] Hamilton (1996) computes the NOPI using the WTI oil price. For reasons discussed in the Introduction above, we use the composite RAC instead of the WTI price. Noting that oil price increases in 1999 had only recovered from the decreases of the preceding two years, Hamilton (2003) incorporates a 3-year horizon in computing the NOPI measure; in subsequent work, for example, Hamilton (2009, 2011), 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.

growth rate of the import price deflator and z_t = Kilian's (2009) global activity measure. Accordingly, these are the variables we use as z_t in examining the relative OOS performance of models (3) and (4).

2.2 Forecast Evaluation

In comparing OOS forecasts from nested models below, we examine MSPEs of the benchmark and nesting model and carry out tests of OOS population-level predictability. These tests effectively are equivalent to tests of the null hypothesis that the extra parameters in the nesting model are jointly equal to zero. This is in contrast to testing for finite-sample predictability, which focuses on testing the null hypothesis of equal OOS Mean Squared Prediction Errors (MSPEs). The finite-sample predictability null will always be rejected less frequently than the population-level predictability null.[7] We use population-level predictability tests since neither of the available tests for finite-sample predictability, Giacomini and White (2006) and Clark and McCracken (2009a), is appropriate for the case we face; the former is designed for rolling estimation windows, whereas we use recursive estimation windows; the latter does not allow for multiple regressors.

We test for OOS population-level predictability via the Clark and West (2007) (CW) and Hubrich and West (2010) (HW) tests. The first is based on an MSPE adjustment to account for noise induced in the OOS forecasts by way of estimation of parameters with zero population means under the null hypothesis that the benchmark model is the true DGP. The second provides a data snooping check when running the CW test against with a small set of nesting alternatives to the benchmark; we use the "max MSPE-adj $t-$statistic" variant of the HW test.[8] We also use the Giacomini and Rossi (2010) (GR) Fluctuation test to examine the local, as opposed to global, forecasting performance over the OOS period. In our application, this amounts to examination of fixed centered (roughly) 10-year moving windows of a transformation of CW statistics.

Clark and McCracken (2009b) emphasize complications that arise when comparing real-time OOS forecasts due to different degrees of data revision across forecast origins. One approach to deal with such complications is to employ Koenig, Dolmas, and Piger's (2003)

[7] Authoritative sources on the distinction between population-level and finite-sample predictability include Inoue and Kilian (2004), Alquist, Kilian, and Vigfusson (2011), Clark and McCracken (2011a) and Clark and McCracken (2011b).

[8] We do so since we found that the other variant of the HW test, which computes a χ^2 statistic, 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.

(KDP) “strategy 1” for estimation of the predictive regressions: first-release data are used for the left-side variables; at each point in the sample, the latest available data at that date are used for right-side variables. Clark and McCracken (2011b) note that, under this estimation approach, predictability tests developed for the case of non-revised data, such as the CW and the HW tests, can be applied. Accordingly, we follow KDP's strategy 1 for estimation of our models.

2.3 In-Sample Evidence of Predictive Content

Results from Inoue and Kilian (2004) imply that IS predictability is a necessary condition for OOS predictability, such that, using the same models, it would be surprising to find OOS population-level predictability from crude oil prices to US GDP in the absence of IS predictive content. Accordingly, in Figure 1 we present such IS evidence on the predictability of oil prices for US GDP via a sequence of recursive estimation windows of post-1973 data, for which the benchmark model is given by (1) and the nesting models are given by (2). For every estimation window considered, the benchmark model generates a higher value of the Akaike Information Criterion (AIC). Since the AIC is lowered when the benchmark model is augmented by oil prices for every estimation window, there is uniformly strong evidence of IS predictive content from oil prices to GDP for the US.

3 Out-of-Sample Results

We report OOS predictability results for the 1990Q1 to 2009Q4 period as well as for a set of two subsamples, 1995Q1-2009Q4 and 2000Q1-2009Q4. Consideration of these subsamples provides some information about whether the predictive content of crude oil prices for US real GDP has changed over time.

3.1 Global Performance and Subsamples

Table 1 presents results for OOS tests of population-level predictability using both ex-post revised and real-time data at the $h = 1$ and $h = 4$ horizons for the AR(4) benchmarks; we remind the reader that the real-time forecasts at $h = 1$ are nowcasts. For each benchmark model, the MSPE is reported, whereas for the alternatives to the benchmark the ratio of the model's MSPE to the benchmark MSPE is reported. At $h = 1$, addition of an oil price measure to the AR(4) benchmark generates a reduction in MSPE in three out of nine cases for the real-time forecasts; none of these MSPE decreases is obtained when the alternative includes the NOPI measure. In each of these three cases, the CW and HW $p-$values are greater than 10%. Using ex-post revised data, none of the nesting models produces a lower MSPE at $h = 1$, and all CW and HW $p-$values are above 10%.

The real-time ADL(4,4)^{nopi} results at the $h = 4$ forecast horizon mirror those at $h = 1$ by way of MSPE ratios and both the CW and HW tests, all MSPE ratios are greater than one and the CW and HW nulls are never rejected at conventional significance levels. The $h = 4$ ex-post revised ADL(4,4)^{nopi} results are also quite similar to the analogous $h = 1$ results; there is one exception by way of the CW $p-$value being below 0.10 for the full 1990-2009 sample. However, the ADL(4,4)^{nrac} and ADL(4,4)^{rrac} results at the $h = 4$ forecast strongly differ from those at $h = 1$. For the real-time forecasts, the MSPE ratios are less than one in four out of nine cases, and the $p-$values for the CW and HW tests are all below 0.10. At $h = 4$ the results for these alternatives are very similar when using ex-post revised data.

Table 2 presents results for OOS predictability tests in which the benchmark and alternative models are given by, respectively, equations (3) and (4). As explained in Section 2.1 above, we consider $z_t =$ the growth rate of the import price deflator and $z_t =$ Kilian's (2009) global activity measure. To help economize on space and focus on the case for which we observe strong rejections of the CW and HW nulls in Table 1, Table 2 gives results only for the $h = 4$ forecast step. But before discussing these, we note that, at $h = 1$, use of the growth rate of the import price deflator in the benchmark strongly increases the real-time predictive power of the nominal composite RAC relative to the (1) and (2) comparisons as follows: the MSPE ratios are less than one and the CW $p-$ values are below 0.10 for all OOS periods. These results do not carry over, however, to use of ex-post revised data.

The top panel of Table 2 shows the $h = 4$ results when the ADL(4,4) benchmark includes the growth rate of the import price deflator. The CW null hypothesis is rejected at conventional significance levels for the ADL(4,4,4)^{rrac} alternative for all OOS periods with use of both real-time and ex-post revised data. This shows that the ADL(4,4)^{rrac} CW results in Table 1 are not due to omission of the import price deflator from the AR(4) benchmark. In contrast, the CW test $p-$values for the ADL(4,4,4)^{nrac} alternative for the real-time forecasts are all above 0.10, suggesting that omission of the import price deflator from the AR(4) benchmark may be a factor behind rejections of the CW null with the ADL(4,4)^{rrac} alternative in Table 1. The bottom panel of Table 2 shows the $h = 4$ results when the ADL(4,4) benchmark includes Kilian's (2009) global activity measure. For the real-time forecasts, the predictive content of the nominal and real composite RAC is substantially lower than that found in Table 1 at $h = 4$, i.e., the CW test $p-$value is below 0.10 in only one out of six cases; it is below 0.10 in two of six cases with the use of ex-post revised data. Thus, the $h = 4$

CW rejections with the ADL(4,4)^{rrac} and ADL(4,4)^{nrac} alternatives in Table 1 are consistent with Kilian's (2009) global activity measure being omitted from the AR(4) benchmark; this is especially so for the 1995-2009 and 2000-2009 subsamples.

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 has population-level predictive content for real GDP growth conditional on including an oil price measure in the benchmark. Adding Kilian's (2009) global activity measure leads to low CW and HW test $p-$values against both the ADL(4,4)^{nrac}, ADL(4,4)^{rrac}, and ADL(4,4)^{nopi} benchmarks. On the other hand, adding the import price deflator does not provide evidence of OOS predictability for real GDP growth.

3.2 Local Performance

The results in Tables 1 and 2 do not provide much evidence that the OOS predictive content of oil prices for real GDP growth varies across the subsamples considered. The GR Fluctuation test provides a more formal framework for addressing this question. The test is motivated by the idea that if the OOS performance of the two models is time-varying, and averaging this movement over the OOS period will result in a loss of information. In Figure 2, we provide time series plots for the Fluctuation test at $h = 4$ at the 10% significance level using centered rolling windows of CW test statistics (for testing model (1) against (2)). If the value of the Fluctuation test statistic is greater than the critical value at observation t, the null hypothesis that the benchmark model is the true model for the roughly ten year window centered at t is rejected.

In contrast to what we see in Tables 1 and 2, the Fluctuation test results differ rather strongly across the use of ex-post revised and real-time data. We focus initially on the ex-post revised data results. First, as we move through the OOS period, the predictive content of the NOPI measure for real GDP growth increases nearly monotonically. For all windows centered at 1999Q2 and later, the null is rejected is at the 10% significance level. The predictive content of the nominal composite RAC for real GDP growth is nearly identical to that of the real composite RAC. The null is rejected for windows centered at 2001Q1 through 2003Q1; it is not rejected for windows that include that latter quarters of the Great Recession.

With the real-time data, the predictive content of the NOPI measure for real GDP growth decreases nearly monotonically. For all windows centered before 1998, the null is rejected; it is not rejected for later windows. The nominal and real composite RAC once again have very similar predictive content predictive content. For windows centered in the middle 1990s and the early 2000s, the Fluctuation test null is rejected; the predictive content falls as the windows get close to the dates of the recent financial crisis.

3.3 Out-of-Sample Predictive Content of Real GDP for Oil Prices

In their critique of the IS oil prices and the macroeconomy literature, Barsky and Kilian (2002) argue that it is important to note that there may very well be feedback from real 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 population-level OOS predictive content evidence from real GDP growth to oil prices. We do not detail these results here, but note our main finding that real GDP growth generally has only weak OOS predictive content for the crude oil price measures we consider. These results may reflect, as emphasized by Alquist, Kilian, and Vigfusson (2011), that our model is misspecified. For example, it neglects real GDP movements in the rest of the world, such that U.S. real GDP may not be a good proxy for world real GDP.

4 Conclusions

Does the imposition of real-time data constraints affect the predictive content of crude oil prices for U.S. real GDP growth? The answer to this question depends on the particular measure of forecasting performance. More specifically, it depends upon whether the benchmark and nesting models are compared on the basis of global or local relative forecasting results.

When focusing on the global (or average) relative performance, we do not find strong differences between use of real-time and ex-post revised data. At the one-step-ahead forecast horizon, there is practically no predictive content of oil prices for real GDP growth, and at the four-step-ahead forecast horizon, oil prices have statistically significant predictive content for real GDP growth. It is doubtful, however, that these forecast improvements are economically significant. For example, the largest MSPE reduction we observe at this forecast horizon across the full OOS period using real-time data is 1%.

But when focusing on the entire time path of the models' relative OOS performance, the real-time and ex-post revised data results differ considerably. For example, the predictive content of an oil-price censored predictor of real GDP growth which has received a great deal of attention in the literature displays completely opposite behavior (monotonically increasing in one case, and monotonically decreasing in another) across use of ex-post revised and real-time data. On the whole, with both types of data our local examination suggests considerable time variation in the OOS predictive relationship between oil prices and real GDP growth.

We explore whether our statistically strong evidence on the predictive content of oil prices for real GDP growth at the four-step-ahead forecast horizon is sensitive to an omitted variable. Our analysis suggests that these findings may reflect omission of Kilian's (2009) global economic activity measure from our bivariate models.

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 (2010), and Mishkin (2009). Our results suggests it might be useful for this literature to consider Kilian's (2009) measure of global economic activity as a candidate variable for global factors.

Our analysis is agnostic about whether the oil price movements which have predictive content for real GDP are due to demand shocks, supply shocks, or both. We believe it would be informative to determine which type of shocks drive the oil price predictability we uncover by applying, for example, Kilian's (2009) framework to produce estimates of such shocks for the problem we study.

References

Alquist, Ron, Lutz Kilian, and Robert J. Vigfusson. (2011) "Forecasting the Price of Oil." In *Handbook of Economic Forecasting*, edited by Graham Elliott and Allan Timmermann, forthcoming. Amsterdam: North Holland.

Bachmeier, Lance, Qi Li, and Dandan Liu. (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, Robert, and Lutz Kilian. (2002) "Do We Really Know that Oil Caused the Great Stagflation? A Monetary Alternative." In *NBER Macroeconomics Annual 2001*, edited by Ben S. Bernanke and Kenneth Rogoff. Cambridge, MA: MIT Press.

Barsky, Robert B., and Lutz Kilian. (2004) "Oil and the Macroeconomy Since the 1970s." *Journal of Economic Perspectives*, 18:4, 115–134.

Baumeister, Christiane and Lutz Kilian. (2011) "Real-Time Forecasts of the Real Price of Oil." Working Paper 2011-16, Bank of Canada, forthcoming in *Journal of Business and Economic Statistics.*

Baumeister, Christiane, and Gert Peersman. (2012) "Time-Varying Effects of Oil Supply Shocks on the U.S. Economy." Working Paper 2012-2.

Bernanke, Ben S., Mark Gertler, and Mark W. Watson. (1997) "Systematic Monetary Policy and the Effects of Oil Price Shocks." *Brookings Papers on Economic Activity* 1, 91–142.

Borio, Claudio E.V., and Andrew Filardo. (2007) "Globalisation and Inflation: New Cross-Country Evidence on the Global Determinants of Domestic Inflation." BIS Working Paper 227, Bank for International Settlements.

Carlton, Amelie B. (2010) "Oil Prices and Real-Time Output Growth. Working paper, University of Houston.

Clark, Todd E., and Michael W. McCracken. (2009a) "Nested Forecast Model Comparisons: a New Approach to Testing Equal Accuracy." Working Paper 2009-050, Federal Reserve Bank of St. Louis, revised January 2011.

Clark, Todd E., and Michael W. McCracken. (2009b) "Tests of Equal Predictive Ability with Real-Time Data." *Journal of Business and Economic Statistics*," 27:4, 441–454.

Clark, Todd E., and Michael W. McCracken. (2011a) "Testing for Unconditional Predictive

Ability." In *Oxford Handbook on Economic Forecasting*, edited by Michael P. Clements and David F. Hendry, pp. 415-440.Oxford: Oxford University Press.

Clark, Todd E., and Michael W. McCracken. (2011) "Advances in Forecast Evaluation." Working Papers 2011-025, Federal Reserve Bank of St. Louis.

Clark, Todd E., and Kenneth D. West. (2007) "Approximately Normal Tests for Equal Predictive Accuracy in Nested Models." *Journal of Econometrics* 138:1, 291–311.

Edelstein, Paul and Lutz Kilian. (2009) "How Sensitive are Consumer Expenditures to Retail Energy Prices?" *Journal of Monetary Economics* 56:6, 766–779.

Estrella, Arturo, and Gikas A. Hardouvelis. (1991) "The Term Structure as a Predictor of Real Economic Activity." *Journal of Finance* 46:2, 555–576.

Giacomini, Raffaella and Barbara Rossi. (2010) "Forecast Comparisons in Unstable Environments." *Journal of Applied Econometrics* 25:4, 595–620.

Giacomini, R., and Halbert White. (2006) "Tests of Conditional Predictive Ability." *Econometrica* 74:6, 1545–1578.

Hamilton, James D. (1983) "Oil and the Macroeconomy Since World War II." *Journal of Political Economy* 91:2, 228–248.

Hamilton, James D. (1996) "This is What Happened to the Oil Price-Macroeconomy Relationship," *Journal of Monetary Economics* 38:2,225–230.

Hamilton, James D. (2003) "What is an Oil Shock?", *Journal of Econometrics* 113:2, 363–398.

Hamilton, James.D. (2009) "Causes and Consequences of the Oil Shock of 2007-08," *Brookings Papers on Economic Activity* Spring, 215–259.

Hamilton, James D. (2011) "Nonlinearities and the Macroeconomic Effects of Oil Prices," *Macroeconomic Dynamics*, 15:S3, 354–378.

Hamilton, James D., and Ana María Herrera. (2004) "Oil Shocks and Aggregate Macroeconomic Behavior: The Role of Monetary Policy, *Journal of Money, Credit, and Banking* 36:2, 265–286.

Hooker, Mark. (1996) "What Happened to the Oil Price-Macroeconomy Relationship?", *Journal of Monetary Economics* 38:2, 195–213.

Hubrich, Kirstin, and Kenneth D. West. (2010) "Forecast Evaluation of Small Nested Model Sets," *Journal of Applied Econometrics* 25:4, 574–594.

Ihrig, Jane, Steven B. Kamin, Deborah Lindner, and Jaime Marquez. (2010) "Some Simple Tests of the Globalization and Inflation Hypothesis," *International Finance* 13:3, 343–375.

Inoue, Atsushi, and Lutz Kilian. (2004) "In-Sample or Out-of-Sample Tests of Predictability: Which One Should We Use?", *Econometric Reviews* 23:4, 371–402.

Kilian, Lutz. (2008) "The Economic Effects of Energy Price Shocks," *Journal of Economic Literature* 46:4, 871–909.

Kilian, Lutz. (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, Lutz. (2010) "Oil Price Shocks, Monetary Policy and Stagflation." In *Inflation in an Era of Relative Price Shocks*, edited by Renée Fry, Callum Jones, and Christopher Kent. RBA Annual Conference Volume, Reserve Bank of Australia.

Kilian, Lutz, and Robert J. Vigfusson. (2011) "Nonlinearities in the Oil Price-Output Relationship," *Macroeconomic Dynamics* 15:S3, 337–363.

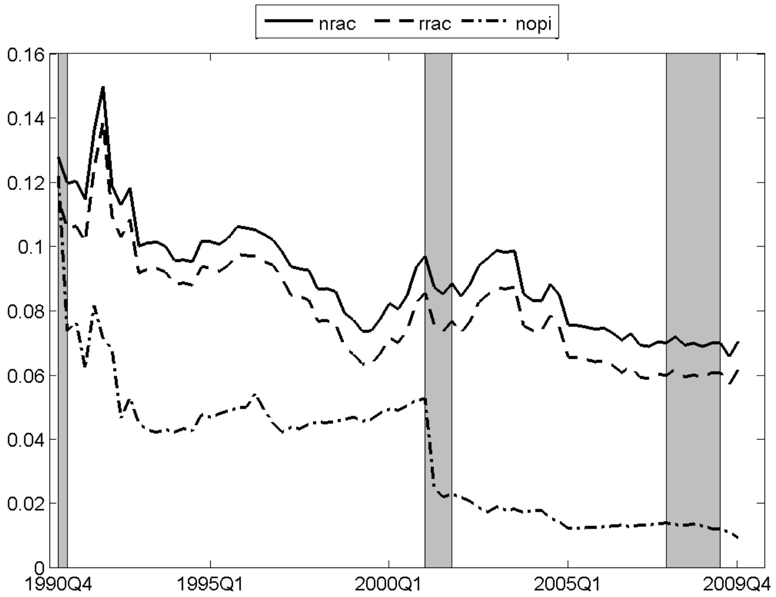
Koenig, Evan F., Sheila Dolmas, and Jeremy Piger. (2003) "The Use and Abuse of Real-Time Data in Economic Forecasting," *The Review of Economics and Statistics* 85:3, 618–628.

Mishkin, Frederic S. (2009) "Globalization, Macroeconomic Performance, and Monetary Policy," *Journal of Money, Credit and Banking* 41:S1, 187–196.

Stock, James H., and Mark W. Watson. (1999) "Forecasting Inflation," *Journal of Monetary Economics* 44:2, 293–335.

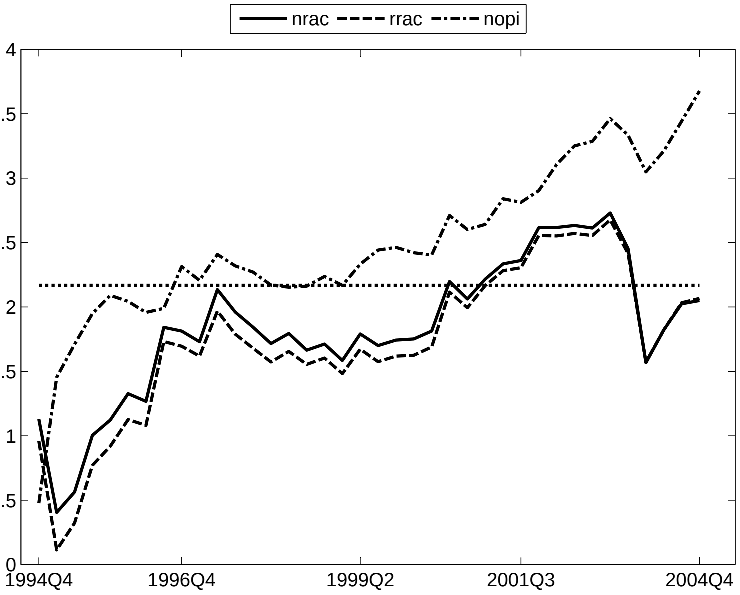
Wright, Jonathan H. (2006) "The Yield Curve and Predicting Recessions," *Finance and Economics Discussion Series 2006-07*, Board of Governors of the Federal Reserve System.

Figure 1: AIC Differences Across Estimation Windows

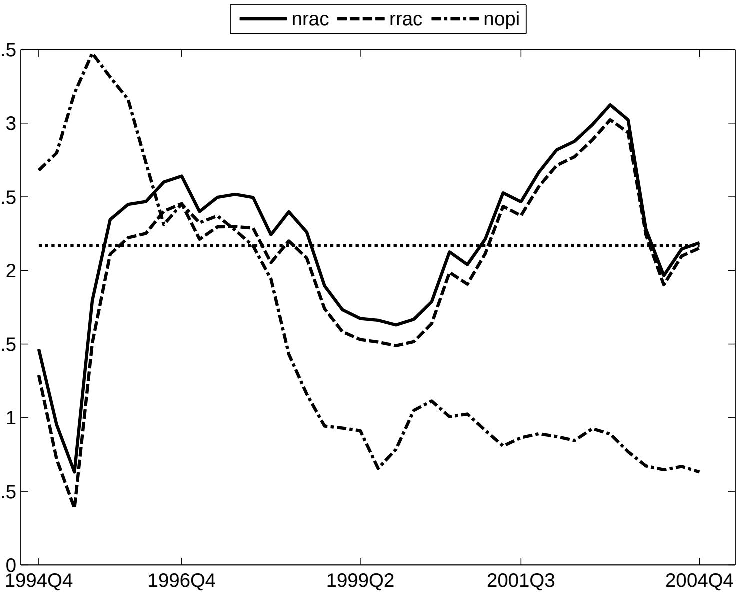


Notes: The graph shows differences in AIC (AIC(benchmark) - AIC(alternative)) for the benchmark model without oil prices and alternative models with an oil price measure included across recursive estimation windows of real-time data; if the benchmark model generates the better fit, then the AIC differences are negative. The labels "nrac," "rrac," and "nopi" indicate that that the ADL alternatives to the benchmark were formed by adding four lags of, respectively, the growth rate of the nominal composite refiners' acquisition cost (RAC) of crude oil, the growth rate of the real composite RAC of crude oil, and the "net oil price increase" (NOPI) measure introduced by Hamilton (1996). The dependent variable in each regression is the growth rate of real GDP for observation $t + 4$, i.e., these regressions are used to generate 4-step-ahead direct out-of-sample forecasts; the dates on the horizontal axis show observation predicted with each regression.

Figure 2: Giacomini and Rossi (2010) Fluctuation Test for Equal Out-of-Sample Predictability at $h = 4$



Ex-Post Revised Data



Real Time

Notes: Giacomini and Rossi (2010) Fluctuation test results centered at time t, based on sequences of Clark and West (2007) (CW) test statistics (for testing model (1) against model (2)), with $\mu = 0.5 = m/P$, where $m =$ the size of the rolling window of CW statistics and $P =$ the number of OOS observations, for the OOS period 1990Q1-2009Q4, such that the length of each window of CW statistics is 38 quarters, i.e., approximately 10 years. Fluctuation test critical value at the 10% significance level shown by dotted horizontal line; if the Fluctuation test statistic exceeds the critical value, the null that the benchmark model is the true model is rejected for the particular window. Benchmark model is an AR(4). The labels "nrac," "rrac," and "nopi" indicate that that the ADL alternatives to the benchmark were formed by adding four lags of, respectively, the growth rate of the nominal RAC of crude oil, the growth rate of the real composite RAC of crude oil, and the "net oil price increase" (NOPI) measure introduced by Hamilton (1996).

Table 1: Tests of Equal Out-of-Sample Population-Level Predictability for Quarterly US GDP Growth Rates with AR(4) Benchmarks

	Ex-Post Revised			Real Time		
	1990-2009	1995-2009	2000-2009	1990-2009	1995-2009	2000-2009
			Forecast horizon $h = 1$			
AR(4) (bench)	0.340	0.360	0.437	0.312	0.338	0.386
vs. ADL(4,4)^{nrac}	1.165	1.088	1.072	1.204	1.022	0.983
	(0.164)	(0.234)	(0.273)	(0.365)	(0.194)	(0.133)
vs. ADL(4,4)^{rrac}	1.163	1.077	1.059	1.190	0.994	0.950
	(0.167)	(0.223)	(0.258)	(0.339)	(0.153)	(0.111)
vs. ADL(4,4)^{nopi}	1.355	1.348	1.216	1.291	1.134	1.067
	(0.193)	(0.424)	(0.338)	(0.773)	(0.772)	(0.659)
HW: vs. 3 models	(0.241)	(0.321)	(0.362)	(0.510)	(0.356)	(0.285)
			Forecast horizon $h = 4$			
AR(4) (bench)	0.471	0.494	0.620	0.325	0.331	0.395
vs. ADL(4,4)^{nrac}	0.967	0.923	0.909	1.069	0.959	0.931
	(**0.007**)	(**0.020**)	(**0.030**)	(**0.004**)	(**0.019**)	(**0.020**)
vs. ADL(4,4)^{rrac}	0.974	0.895	0.873	1.078	0.974	0.938
	(**0.007**)	(**0.019**)	(**0.024**)	(**0.006**)	(**0.026**)	(**0.021**)
vs. ADL(4,4)^{nopi}	1.318	1.090	1.011	1.211	1.158	1.207
	(0.120)	(0.108)	(**0.059**)	(0.337)	(0.771)	(0.882)
HW: vs. 3 models	(**0.013**)	(**0.031**)	(**0.038**)	(**0.009**)	(**0.040**)	(**0.042**)

Notes: Table reports results for out-of-sample tests of equal population-level predictability 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 recursive windows of data; the first in-sample window is 1974Q1-1989Q4. The panel labeled "Ex-Post Revised Data" reports results using the latest vintage of data for both estimation and forecasting. The panel labeled "Real Time" reports results using vintages of real-time data via "strategy 1" of Koenig, Dolmas, and Piger (2003), with OOS forecast errors computed using the first available real-time vintages of data. For the AR(4) benchmark models, MSPEs reported; for alternatives to the benchmark, the ratio of the alternative model's MSPE to the benchmark's MSPE reported. In parentheses under the MSPE ratios are reported $p-$values for the Clark and West (2007) test for equal population-level predictability for nested models. The superscripts "*nrac*," "*rrac*," and "*nopi*" indicate that that the ADL alternatives to the benchmark were formed by adding four lags of, respectively, the growth rate of the nominal composite RAC of crude oil, the growth rate of the real composite RAC of crude oil, 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 equal population-level predictability for a small set of alternative nesting models.

Table 2: Tests of Equal Out-of-Sample Population-Level Predictability for Quarterly US GDP Growth Rates with ADL(4,4) Benchmarks at $h = 4$

	Ex-Post Revised			Real Time		
	1990-2009	1995-2009	2000-2009	1990-2009	1995-2009	2000-2009
			Using Import Price Deflator			
ADL(4,4) (bench)	0.464	0.471	0.603	0.325	0.318	0.367
vs. ADL(4,4,4)^{nrac}	1.010	0.965	0.916	1.219	1.116	1.073
	(0.110)	(**0.073**)	(**0.038**)	(0.876)	(0.925)	(0.701)
vs. ADL(4,4,4)^{rrac}	0.998	0.954	0.904	1.062	1.042	0.992
	(**0.095**)	(**0.057**)	(**0.027**)	(**0.053**)	(**0.070**)	(**0.050**)
vs. ADL(4,4,4)^{nopi}	1.597	1.283	1.121	1.200	1.077	1.106
	(0.320)	(0.379)	(0.162)	(**0.088**)	(**0.089**)	(0.202)
HW: vs. 3 models	(0.188)	(0.102)	(**0.052**)	(0.136)	(0.174)	(0.131)
			Using Index of Global Real Activity			
ADL(4,4) (bench)	0.422	0.445	0.527	0.302	0.312	0.359
vs. ADL(4,4,4)^{nrac}	1.014	1.008	1.017	1.033	1.015	1.019
	(**0.079**)	(0.257)	(0.339)	(0.190)	(0.394)	(0.475)
vs. ADL(4,4,4)^{rrac}	1.014	1.006	1.007	1.094	1.035	0.994
	(**0.092**)	(0.266)	(0.306)	(**0.037**)	(0.276)	(0.139)
vs. ADL(4,4,4)^{nopi}	1.343	1.116	1.068	1.363	1.161	1.123
	(0.259)	(0.334)	(0.231)	(0.393)	(0.898)	(0.781)
HW: vs. 3 models	(0.159)	(0.357)	(0.328)	(**0.085**)	(0.507)	(0.305)

Notes: See notes to Table 1. In the top panel, the benchmark model includes four lags of the growth rates of U.S. real GDP and the import price deflator. In the bottom panel, the benchmark model includes four lags of the growth rate of U.S. real GDP and the index of global real activity of Kilian (2009). The superscripts "*nrac*," "*rrac*," and "*nopi*" indicate that that the ADL alternatives to the benchmark were formed by adding four lags of, respectively, the growth rate of the nominal composite RAC of crude oil, the growth rate of the real composite RAC of crude oil, and the "net oil price increase" (NOPI) measure introduced by Hamilton (1996).