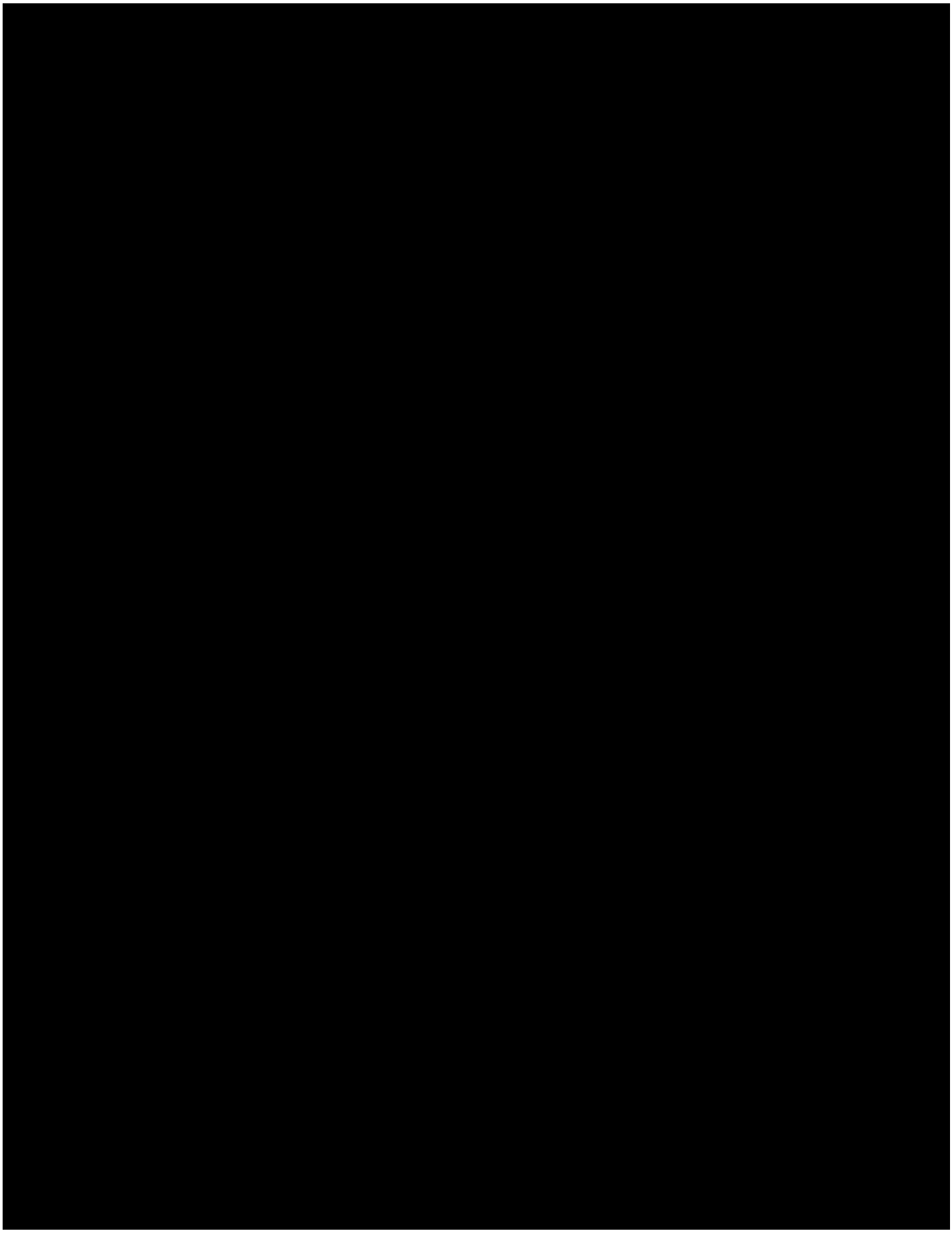


 WP/15/251

IMF Working Paper

Research Department

Forecasting the Nominal Brent Oil Price with VARs—One Model Fits All?

Benjamin Beckers and Samya Beidas-Strom[1]

Authorized for distribution by Thomas Helbling

November 2015

Abstract

We carry out an *ex post* assessment of popular models used to forecast oil prices and propose a host of alternative VAR models based on traditional global macroeconomic and oil market aggregates. While the exact specification of VAR models for nominal oil price prediction is still open to debate, the bias and underprediction in futures and random walk forecasts are larger across all horizons in relation to a large set of VAR specifications. The VAR forecasts generally have the smallest average forecast errors and the highest accuracy, with most specifications outperforming futures and random walk forecasts for horizons up to two years. This calls for caution in reliance on futures or the random walk for forecasting, particularly for near term predictions. Despite the overall strength of VAR models, we highlight some performance instability, with small alterations in specifications, subsamples or lag lengths providing widely different forecasts at times. Combining futures, random walk and VAR models for forecasting have merit for medium term horizons.

JEL Classification Numbers: C53, Q43, C32

Keywords: forecasting, oil, VARs

Authors' E-Mail Addresses: bbeckers@diw.de; sbeidasstrom@imf.org

[1] We thank Christiane Baumeister for generously sharing the Baumeister and Kilian (2012) code, Daniel Rivera Greenwood for excellent research assistance, and Jörg Decressin, Hamid Faruqee, Thomas Helbling, Andrea Pescatori, Emil Stavrev, other colleagues at the IMF Research Department and participants at the George Washington University's Seminar for Forecasting for helpful comments and advice.

Contents **Page**

Tables

Figures

I. MOTIVATION

Large unexpected oil price movements pose a significant macroeconomic risk for both oil-importing and exporting countries. Over the last decade oil prices have tripled from around $40 in 2003 to $140 in 2008, only to drop below $50 in late 2014, with these price movements largely unexpected. The price crash in the aftermath of the Lehman bankruptcy, the subsequent upswing between 2009 and mid-2014, and the sharp fall since then reveal that oil price fluctuations remain prominent, adding to (and potentially feeding back from) the general economic uncertainty and weak recovery.

Since oil prices are notoriously difficult to predict, practitioners have long relied on futures. However, recently developed economic vector auto-regression (VAR) models in the class of Kilian (2009) have been used for short to medium term real oil price forecasting. In particular, Baumeister and Kilian (2012) and Alquist, Kilian and Vigfussion (2013) have shown that these real oil price forecasts provide more accurate predictions of the future path of real oil prices relative to futures or other models.

In this paper, we build on this seminal body of work, by attempting to forecast the *nominal* price of international oil benchmarks (Brent, WTI)—rather than *real* oil prices or the *real* U.S. refiners' acquisition cost (RAC),[2] and by this we fill a gap in the literature to the best of our knowledge. Nominal oil prices are of direct interest to the IMF's *World Economic Outlook* forecasts. We rely on well-established macroeconomic and oil market variables such as global industrial production, oil supply and oil inventories to capture demand and supply as well as speculative (precautionary) demand effects on oil prices as noted by Kilian (2009) and Kilian and Murphy (2014). In addition, for the purpose of providing a nominal oil price forecast, we introduce CPI inflation into our VAR model. Finally, building on Chen *et al.* (2010) and Grisse (2010), we augment the VAR model with the U.S. dollar exchange rate, trade weighted vis-à-vis the largest oil consumers to account for the impact of the exchange rate channel on oil demand.

Four further issues are of particular interest in the literature. First, how should we measure oil demand? We experiment with different specifications; namely global and regional industrial production indices and the Kilian index of real economic activity (REA) based on dry cargo bulk freight rates (Kilian, 2009). Further, we differentiate between advanced and emerging market oil demand, following Aastveit *et al.* (2014) who show that emerging market economies' demand has been more than twice as important as developed economies' demand in driving oil prices over the past two decades. Second, how to handle

[2] U.S. refiners' acquisition costs (RAC) is the average price paid by U.S. refiners for imported crude oil and includes transportation and other fees. See http://www.eia.gov/dnav/pet/PET_PRI_RAC2_DCU_NUS_M.htm

the rise in unconventional oil production from North America? We break down global supply into its major regional producers (e.g., OPEC vs. non-OPEC). Third, how to handle the upward drift in oil prices between 2003 and 2008 and their subsequent decline? Since it is difficult to account for all additional driving factors beyond those variables in our VAR, we assess both a trended oil price model (consistent with Hotelling's 1931 positive intercept) and a detrended model (with a zero intercept in the real oil price equation of the VAR). Fourth, how far back are oil market fundamentals relevant for near- to medium-term predictions? For this, we assess the medium-term memory of the variables in the system and experiment with varying lag structures to capture oil market dynamics over various windows up to 24 months. For all, we evaluate the forecasting performance of our estimates against several other prominent forecasting models from the literature (e.g., futures prices, the no-change prediction and univariate autoregressive models).

The discussion above indicates that the exact specification of a VAR model for the purpose of oil price prediction is still open to debate—whether for real or nominal prices. Hence, as well as experimenting with additional variables, we highlight the instability of rival forecasting models over time. That is, while the standard parsimonious VAR model of oil demand, supply and inventory demand performs best over the full sample, we show that there is merit in other specifications, as well as in the futures forecast during subsample intervals. These differences may be fruitfully combined by forecast averaging.

Evaluating forecasting performance over the past two decades, our paper's main findings are as follows:

- Across the whole sample, the bias of futures forecasts has been larger than that of the VAR across all horizons. The direction of the bias indicates a severe underprediction of the futures forecast. For forecast horizons between five to eight months, even a random walk (with or without drift) outperformed the futures curve in terms of lower bias, lower absolute mean square error, and higher accuracy. For horizons up to 12 months, an AR(6) outperformed both the random walk and futures.

- However, for horizons beyond 12 months up to 24 months, futures forecasts were more accurate than those of the random walk or autoregressive models. In contrast, most of our VAR model specifications outperformed all competitors for both short- and medium-term horizons in terms of accuracy.[3] In addition, we find that removing the trend in oil prices improved the forecast performance for the short-term even further. Beyond the one-year

[3] Throughout our paper, near- or medium-term refers to horizons up to 24 months. Short-term refers to horizons between 1 to 12 months. The long-term, i.e. forecasts in excess of 24 months, are not examined in our paper.

forecast horizon, VAR models including the exchange rate, interest rates, and decomposed oil supply by region provided the most accurate forecasts among all competitor models.

- Forecasting performance varies across time. The random walk forecast performed better during periods of stable oil prices, but the VAR has performed best since 2008. Indeed since the collapse in oil prices in 2008, the VAR has had superior forecasting performance. Measures of activity based on the Baltic Dry Index included valuable information for forecasting oil prices prior to 2008, but thereafter performed poorly.

- Combination forecasts for horizons under 18 months performed poorly. However, for horizons beyond 18 months, we found merit in a combination forecast, particularly one with inverse forecast accuracy weights computed over 24 months.

- Despite the overall strength of VAR models, performance suffered from instability over the full sample, with small alterations in specifications, subsamples, and lag lengths providing widely different forecasts at times. Therefore, we conclude that predicting oil prices on a long sample of data with structural breaks remains difficult.

The rest of this paper is structured as follows. Section 2 provides a brief overview of the existing literature and Section 3 describes the baseline forecasting models, relating them to the main near- to medium-term oil price forecasting models in the literature. Section 4 describes the data used. Section 5 then introduces our evaluation approach, measuring forecast unbiasedness and accuracy and discussing key results. Section 6 offers robustness checks and Section 7 proposes a forecast combination. Section 8 illustrates *ex ante* VAR predictions during the oil price collapse of 2014. Section 9 concludes.

II. RELATION TO THE LITERATURE

After a period of relatively stable oil prices, academic interest in the topic of forecasting oil prices over the near to medium term picked up.[4] The futures forecast has so far been the predominant focus in practice and the academic literature, with its informational content widely reported, as in Reeve and Vigfusson (2011), Reichsfeld and Roache (2011) and Chinn and Coibion (2013). While all three papers investigate the predictive ability of futures (for a range of commodities, showing that energy futures generally perform better in forecasting future spot prices than non-energy commodity futures), different aspects are highlighted.

[4] It is important to note that our paper relates to this class of forecasting models, i.e. short to medium term, defined as those horizons from 1 to 24 months only. For longer horizons, structural models are more common. See Benes *et al.* (2015) for one example of these longer horizon models.

Next to this strand, there is some evidence that recently developed economic vector auto-regression (VAR) models by Kilian (2009) for the determination of the real price of oil could provide more accurate forecasts. Economic theory suggests that a number of global economic aggregates such as oil supply and demand (or related variables), and forward looking variables (such as changes in global crude oil inventories) could contain information about future oil prices. For real oil prices, Alquist *et al.* (2013) first find that some proxies for global oil demand, namely global industrial production (GIP) and the index of global real economic activity (REA) developed by Kilian (2009), feature significant predictive ability, while US GDP does not. Using a factor VAR, Aastveit *et al.* (2014) find that demand from emerging economies, most notably from Asian countries, is more than twice as important as demand from developed countries (as proxied by industrial production in each region) when accounting for the fluctuations in the real U.S. refiners' acquisition cost (RAC) and in oil production.

Second, Alquist *et al.* (2013) demonstrate that a simple VAR model with global oil supply, Kilian's REA index and crude oil inventories outperforms the futures forecast and other models for short forecast horizons (up to 9 months). This result also holds in real-time, as shown by Baumeister and Kilian (2014). These VAR forecasts are found to be robust to various changes in model specification and estimation methods, including Bayesian estimation (Baumeister and Kilian, 2012). Yet, to the best of our knowledge, none of the papers in the literature compare the forecasting performance of *nominal* oil prices against the futures-based forecast, and hence this is the gap our paper fills.

Finally, before turning to our research approach, we augment the baseline model with two variables that may help explain oil price movements. First, to obtain a nominal oil price forecast, we need to forecast inflation. This is directly in line with Alquist *et al.* (2013) who find that monetary variables such as inflation, money growth rates and other nominal commodity prices influence nominal oil price movements. Second, we augment the model with an index to capture the transmission of exchange rate movements on oil demand from oil-importing countries. This is motivated by Chen *et al.* (2010) who find some evidence of predictability when using U.S. dollar exchange rates for a broad index of commodity prices, since global oil demand and supply are influenced by the relative exchange rates of oil importers and exporters in relation to the U.S. dollar. Indeed, Grisse (2010) finds persistent negative correlation between oil prices and the U.S. dollar in recent years.

III. BASELINE FORECASTING MODELS

We now introduce the empirical specification of our baseline VAR out-of-sample forecasting model and discuss the selection of variables. We employ the standard reduced-form VAR model with monthly seasonal dummies, which takes the form:

$$y_{t+1} = c + \delta_m \mathrm{D}_{m,t+1} + A_1 y_t + \cdots + A_p y_{t-p+1} + u_{t+1} \quad (1)$$

where y_t, c and δ_m are $Kx1$ vectors of observables, constants, and monthly dummy parameters, respectively, and $A_i, i = 1, \ldots p$ are KxK coefficient matrices. $\mathrm{D}_{m,t}$ is the monthly dummy indicator which takes the value of 1 if the forecast period t is month m, with $m = \{January, \ldots, November\}$.[5]

The lag length $p = 6$ is obtained by the Akaike information criterion (AIC). The reduced-form residuals u_t are assumed to be $iid\ N(0, \Sigma_u)$, where Σ_u is the variance-covariance matrix of (potentially) correlated residuals. The absence of serial correlation in the residual vectors is important for forecasting. We do not find any evidence of remaining serial correlation for the VAR model with $p = 6$ lags. Henceforth, the VAR model with 6 lags is chosen, VAR(6), but (in Section VI) we evaluate the forecasting performance for all $p \in \{6, \ldots, 24\}$. The model is estimated by multivariate least-squares.[6]

The baseline model, $y_t = [\Delta log\ prod_t, \Delta log\ ip_t, \Delta log\ rpo_t, \Delta crinv_t, \Delta logCPI_t]'$, (Table 1, model A(i)), refers to the vector of endogenous variables, as follows: $\Delta log\ prod_t$, is the log-difference in global crude oil production; $\Delta log\ ip_t$ is the log-difference of the global industrial production index, $\Delta log\ rpo_t$ is the log-differenced real price of oil; and $\Delta crinv_t$, is the level change in OECD inventories. OECD inventories are included as a proxy for global inventories and hence capture speculative (precautionary) demand for oil.[7] By specifying the real price of oil in log-differences, we assure stationarity of the model and remove higher order persistence,[8] and hence longer lag lengths are unnecessary, even harmful since they could reduce estimation efficiency (see Section VI). The real price of oil is computed by dividing the monthly average *nominal* spot price of oil by U.S. CPI inflation. The forecast change in U.S. CPI is used to obtain nominal oil price forecasts.

---

[5] No dummy is needed for December as the constant already captures that month, with other months' constants adjusted relative to December's.

[6] Since the forecasting methodology is well-known, we refer the reader to standard time series textbooks such as Hamilton (1994) or Lütkepohl (2007) for further details on estimation and recursive forecasting.

[7] Kilian and Murphy (2014) first introduced inventories into oil VAR models using U.S. oil inventories to extrapolate global inventories data. Kilian and Li (2014) obtain proprietary data to estimate OECD and non-OECD inventories, including oil in transit (i.e. floating storage and at sea). See www.IEA.org for more details.

[8] When following the literature and specifying the real price of oil in logs, the largest estimated eigenvalue is 0.996, questioning stationarity of the model for our sample period. As a robustness check we run the model with the log-specification and find that the forecasting performance deteriorates.

Table 1. Overview of Forecasting Models

Benchmark 1	Random Walk	$P_{t+1} = P_t + u_{t+1}$
Benchmark 2	Random walk with drift *	$p_{t+1} = v + p_t + u_{t+1}$
Benchmark 3	Futures-based forecast	$P_{t+j} = F_{t+j} + u_{t+j}$
Benchmark 4	ARMA(1,1)	$p_{t+1} = c + \phi p_t + u_{t+1} + \theta u_t$
Benchmark 5	AR(q)	$p_{t+1} = \mu + \phi_1 p_t + \cdots + \phi_p p_{t-p+1} + u_{t+1}$
Benchmark 6	MA(q)	$p_{t+1} = \mu + u_{t+1} + \theta_1 u_t + \cdots + q u_{t-q+1}$

VAR models	**Transformations**		**Variables**	**Lag Lengths**			
A(i)	Log-diff	trend	Global oil supply, global IP, OECD inventories, CPI	6	12	18	24
A(ii)		detrended					
B(i)	Log-diff	trend	A + exchange rate index	6	12	18	24
B(ii)		detrended					
C(i)	Log-diff	trend	A + 3Y interest rate (in differences)	6	12	18	24
C(ii)		detrended					
D(i)	Log-diff	trend	C + 10-3Y interest rate spread	6	12	18	24
D(ii)		detrended					
E(i)	Log-diff	trend	B, with IP broken into advanced, emerging	6	12	18	24
E(ii)		detrended					
F(i)	Log-diff	trend	B, with oil supply broken into OPEC, non-OPEC	6	12	18	24
F(ii)		detrended					
G(i)	Log-diff	trend	B, with oil supply broken into three: OPEC, North, ROW	6	12	18	24
G(ii)		detrended					
H(i)	Log-diff	trend	B, with Kilian's REA, no global IP	6	12	18	24
H(ii)		detrended					

* The drift parameter for the random walk was estimated as the average percentage change in the oil price over the last 12 years—identical to our training period for the out of sample forecast. Small letters indicate logs, caps indicate levels.

In addition to the above variables, motivated by Chen *et al.* (2010), we construct an exchange rate index of the U.S. dollar against the currencies of major oil consumers (Table 1, model B(i)). With this, we capture additional demand-side effects from exchange rate movements of large oil importing countries. Specifically, we weight the exchange rates of these countries by the country's relative share of oil consumption:

$$EXR_t = \frac{1}{N}\sum_{i=1}^{N} \frac{Cons_{i,t}}{\sum_{i=1}^{N} Cons_{i,t}} EXR_{i,t}. \quad (2)$$

For completeness we also explore the forecasting performance with the addition of the following variables:

- Interest rates (U.S. 3-month and 10-year treasury bond yields at constant maturity) (Table 1, model C(i)) since this can be motivated by the Hotelling (1931) model—whereby the nominal price of a non-renewable resource should increase by the nominal interest rate if marginal extraction costs are zero.

- The spread between long- and short-term interest rates (Table 1, model D(i)), i.e., the yield curve, as an indicator for the expected future state of the business cycle and thus future oil demand.

- Disaggregated oil supply (production) and demand (proxied by industrial production indices) of key producers or regions (Table 1, models E(i)-G(i)). Model E follows Aastveit *et al.* (2014), differentiating between developed and emerging market industrial production.

- Kilian's REA, as a measure or proxy for oil demand (Table 1, model H(i)).

These are the main explanatory variables of our reduced-form VAR model.

An important issue is how to handle the presence of trends in our VAR models. While there are several alternatives, the approach so far allows for a time varying trend given that this could be important to capture the run up in oil prices between 2003 and 2008, and is consistent with the intercept term in the spirit of Hotelling (1931).[9] In addition, we introduce a detrended model without removing any seasonality, which would protect against upward bias in post-2008 forecasts (based on estimation after the oil price run up). Indeed, Kilian (2009) argues that the real oil price is stationary and thus not trending.[10] For this, we determine the steady state of the VAR system and deduct this steady-state change of the real price of oil from the system's dynamics (Table 1, models A(ii)-H(ii)).[11]

The steady state of the VAR system is defined as a state in which the endogenous variables y_t do not change over time (in differences, i.e., we have a constant change). This implies that $y_t = y_{t-1} = \cdots = y_{t-p} = \bar{y}$. Inserting this into (1) provides an expression of the VAR system's steady state:

[9] Hotelling (1931) assumes a constant discount/interest rate, however, these are rarely constant over time.

[10] For this reason (i.e., the stationarity of real oil prices), Baumeister and Kilian (2011) take log real oil prices rather than log-differences. However, we ran the models shown in Table 1 with log real oil prices but these performed worse than those with log differences and hence are excluded from the tables for parsimony. See also footnote 8.

[11] In other words, we set the steady-state change in the equation of real oil prices to zero.

$$\bar{y} = \left(I - A_1 - \cdots - A_p\right)^{-1} * c. \quad (3)$$

Kilian (2009) argues that real oil prices are stationary and non-trending (thus ordered last in the vector y_t), thus we only subtract the estimated steady-state value of real prices from the system while allowing all other variables to be trending—by setting the other steady-state values equal to zero:

$$\bar{y}(1, \ldots, K-1) = 0.$$

The forecasts of the detrended models are then obtained by subtracting the steady-state value of real oil prices for each iterative prediction, as in:

$$\widehat{y_{t+1}} - \bar{y} = c + \delta_{t+1} + A_1 y_t + \cdots + A_p y_{t-p+1} - \bar{y}. \quad (4)$$

Setting this constant equal to zero safeguards against an increase in the future real oil price forecast that is primarily due to its long-run trend, on the one hand; and allows us to evaluate how far our results are driven by this behavior and not by the joint dynamics of the system, on the other hand.

IV. DATA

We use monthly data from 1985M01:2014M12. Data on crude oil production, OECD inventories and oil consumption are from the International Energy Agency's (IEA) Monthly Oil Data Service. Oil prices for Brent and WTI crude, U.S. CPI, global and regional industrial production, exchange rates of major oil consumers vis-à-vis the U.S. dollar and U.S. bond yields are obtained from Haver Analytics. Finally, Kilian's index for real economic activity (REA) is taken directly from http://www-personal.umich.edu/~lkilian/.[12] Our sample starts in 1985 due to lack of data for oil consumption and inventories prior to this date. The first forecast is carried out in 1997M01 and thereafter the model is re-estimated each month based on a recursively expanding sample.

We do not adjust for seasonality in order to capture deterministic changes in nominal oil prices. All data has been revised to the latest availability. Real-time data sets were not available for the full sample time period. While this could be considered a limitation of our analysis in this paper, Baumeister and Kilian (2014) show that *ex post* data revisions matter

[12] See Beidas-Strom and Pescatori (2014) and The Economist (2015) for a discussion of the performance of this index since the onset of global financial crisis during a period of overcapacity in the bulk shipping sector.

most for forecasting real refiners' acquisition cost (RAC) but not for non-revised prices of WTI crude oil. This suggests the importance of acknowledging the real-time dimension of data in RAC forecasts is primarily driven by revisions to the RAC price series—and not due to revisions of demand and supply variables. Since we focus on forecasting Brent (and WTI) crude oil prices, revisions of nominal prices of RAC are thus not relevant here. Hence, we would expect only negligible changes to our forecast evaluation based on revised data for Brent and WTI using VAR forecasts with real-time data.

V. FORECAST EVALUATION

In this section, we compare the models' out-of-sample forecasting performance with regard to their unbiasedness and forecasting accuracy. For unbiasedness, we follow the standard approach of Mincer and Zarnowitz (1969) and regress realizations of oil prices h periods from the forecast period t, y_{t+h}, on the prediction $\hat{y}_{t+h}$ and a constant:

$$y_{t+h} = \alpha + \beta \hat{y}_{t+h} + v_{t+h}. \tag{5}$$

If the forecasts are indeed unbiased, we would expect $\alpha = 0$ and $\beta = 1$. Evidence against this joint hypothesis indicates a significant forecast bias. In addition to testing for unbiasedness of our forecast, the accuracy of the prediction is of critical importance. As Reichsfeld and Roache (2011) discuss, the measure of accuracy, however, should generally vary depending on the forecaster's loss function. For oil prices, oil importers are naturally more concerned about upward price risk, while exporters are more vulnerable to downward surprises. To get around these asymmetries, we follow the standard practice of comparing the forecasting accuracy on the basis of the symmetric root mean squared forecast error (RMSE) for the h-period forecast:

$$RMSE_h = \sqrt{\frac{1}{(T-h)} \sum_{t=1}^{(T-h)} (\hat{y}_{t+h} - y_{t+h})^2}. \tag{6}$$

We then compare all forecasts against the accuracy of the random walk (RW) prediction and test the null hypothesis of equal RMSE's of a forecast (from our VAR models and the futures forecast) and the random walk forecast by means of the Diebold-Mariano (1995) test. The asymptotically normally distributed test statistic $DM_h \sim N(0,1)$ can be obtained as:

$$DM_h = \frac{\bar{d}}{\sqrt{\hat{\omega}_d/(T-h)}} \tag{7}$$

where $\bar{d} = \frac{1}{(T-h)} \sum_{t=1}^{T-h} [(\hat{u}_{t+h}^{B})^2 - (\hat{u}_{t+h}^{A})^2]$ and where $\hat{u}_{t+h}^{A}$, $\hat{u}_{t+h}^{B}$ are the h-period ahead oil price forecast errors of the alternative model and the benchmark, respectively, and $\hat{\omega}_d$ is the long run covariance matrix of $\bar{d}$, taking serial correlation into account. For ARMA models that nest the random walk, we do not report the DM statistic.

In order to shed light on when the VAR model forecasts differ from the RW and the futures forecast, we compute 2-year rolling averages of the RMSE for all horizons. A statistical test that attempts to capture time-dependent patterns in different forecasting performances has been proposed by Giacomini and Rossi (2010) for fixing or rolling estimation schemes. However, since we rely on recursive estimation, this test is not applicable. The forecasting performance of our VAR model deteriorates with decreasing length of the estimation window, hence switching to rolling estimation seems unfit and makes a time-dependent statistical comparison hard to justify. We therefore solely compare the forecast performance of the VAR model with the benchmarks for different subsamples. To illustrate general patterns of forecast instability, we focus here on different forecast performances before and after the onset of the global financial crisis in 2008M06 based on the RMSE measure and the Diebold-Mariano test statistic.

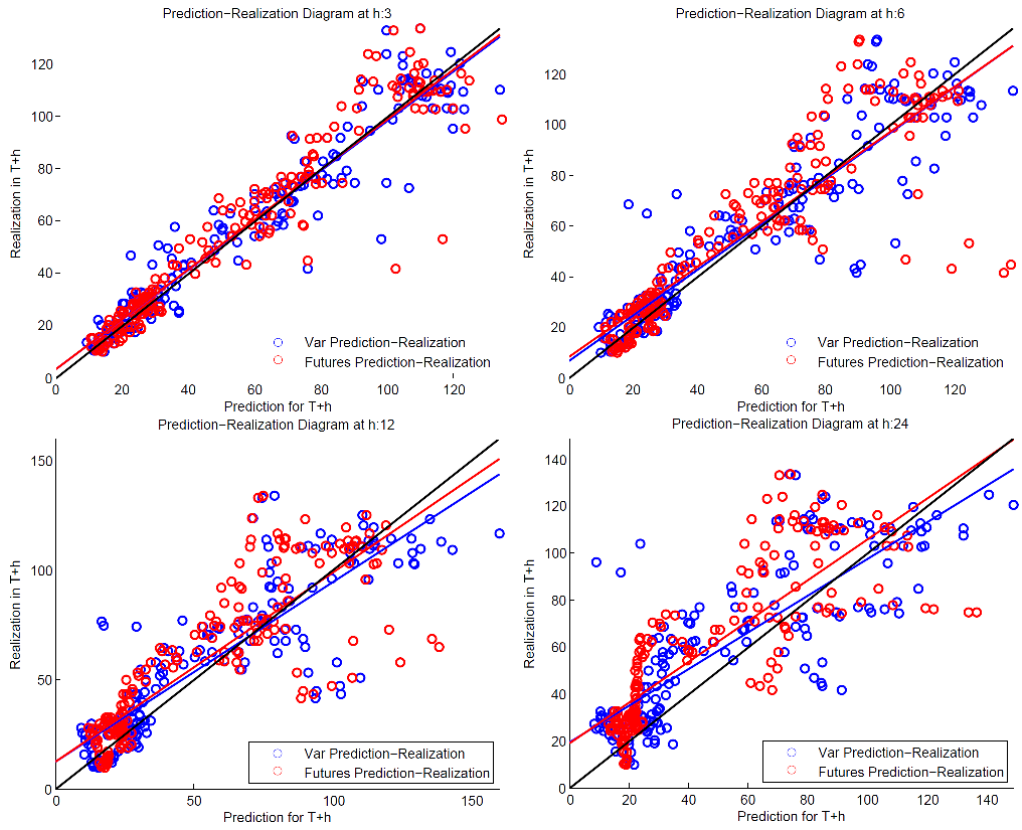
Unbiasedness

We begin with an assessment of the out-of-sample forecasting performance by taking a first pass at forecast unbiasedness.[13] Figure 1 shows the predictions and corresponding realizations for our baseline VAR, model A(i), and the futures forecast for different horizons $h = \{3, 6, 12, 24\}$ and the corresponding best-fit lines from the Mincer- Zarnowitz (MZ) regression from top left to bottom right. The 45 degree line in black gives the ideal forecast with a perfect fit.

Three points emerge from the figure. First, while the bias (indicated by a non-zero intercept and slope of the best-fit-line diverging from unity) seems to be small for horizons up to six months, it becomes more substantial thereafter. Second, there is a general tendency of forecasts to underpredict the oil price, as evidenced by most points lying above the 45 degree line. For the futures forecast, this is more pronounced with a strong underprediction present even at $h = 6$. Third, the futures forecast have the largest forecast errors at each horizon. For all horizons but $h = 3$, there is also some visual evidence that the VAR forecast errors are less dispersed along the 45 degree line, indicating higher accuracy.

[13] In this paper we report the findings for Brent spot price forecasts only, but the results hold equally for WTI.

Figure 1. Prediction-Realization Diagrams for VAR and Futures Forecasts



To test these findings for statistical significance, Table 2 reports the coefficients of the MZ regression along with the joint test of $\alpha = 0$ and $\beta = 1$ for the RW benchmark, the futures baseline forecast and the most parsimonious VAR, model A(i).[14] The first two columns in each panel display the estimated coefficients of the MZ regression along with the F-statistic of joint significant deviations from the null hypothesis in the third column of each panel. Values indicated by *, **, and *** are statistically significant at the 5, 1, and 0.1 percent, respectively. The last two columns of each panel report the average forecast error and its standard deviation.

Three points emerge from this table. First, the α-coefficients of the MZ regression are similar, having the same sign, for the random walk, futures, and the VAR forecasts. However, the futures forecast α-bias is generally the largest, indicating the largest under prediction. Second, the slope coefficient, β, diverges strongly from 1 for horizons longer than three months for the VAR and the random walk, but is strongly biased for the futures forecast

[14] Note that the choice of VAR model A(i) is illustrative since all specifications were checked and conformed to the reported findings.

Table 2. Mincer-Zarnowitz Test for Unbiasedness

Forecast horizon	*Random Walk*					*Futures*					*VAR*				
	α	β	F-test	Mean	Standard deviation	α	β	F-test	Mean	Standard deviation	α	β	F-test	Mean	Standard deviation
1	1.038	0.984	1.078	0.102	5.188	6.725***	0.854***	61.173***	-1.932	7.435	0.920	0.980	1.958	-0.203	4.868
2	2.657	0.960	1.534	0.220	8.526	7.669***	0.842***	25.083***	-1.790	9.185	2.472	0.951*	3.047*	-0.405	7.626
3	4.609	0.931	1.881	0.392	11.304	9.236***	0.820***	14.425***	-1.651	11.531	4.040*	0.923**	3.488*	-0.469	9.587
4	6.543*	0.902	2.112	0.648	13.460	10.840***	0.798***	10.423***	-1.385	13.388	5.427*	0.899***	3.598*	-0.523	11.089
5	8.426*	0.875*	2.345	0.954	15.203	12.387***	0.778***	8.562***	-1.030	14.869	6.835*	0.874**	3.831*	-0.582	12.337
6	10.007*	0.853*	2.477	1.299	16.489	13.726**	0.762***	7.503***	-0.603	15.978	8.346*	0.847**	4.193*	-0.709	13.530
7	11.206*	0.839*	2.493	1.669	17.307	14.783**	0.751***	6.746**	-0.127	16.684	9.766*	0.823**	4.488*	-0.762	14.487
8	12.171*	0.828*	2.475	2.072	17.865	15.644**	0.743***	6.203***	0.376	17.148	11.169*	0.800**	4.838**	-0.779	15.401
9	12.940*	0.821	2.444	2.488	18.242	16.310**	0.738***	5.778**	0.886	17.414	12.471**	0.779**	5.205**	-0.768	16.171
10	13.618**	0.815	2.439	2.893	18.523	16.841**	0.736**	5.453**	1.387	17.537	13.650**	0.761***	5.595**	-0.696	16.790
11	14.240*	0.810	2.462	3.296	18.772	17.276**	0.736**	5.205**	1.886	17.592	14.516**	0.748***	5.886**	-0.619	17.123
12	15.005*	0.803	2.560	3.701	19.082	17.753**	0.735**	5.085**	2.393	17.662	15.164**	0.739***	6.0676**	-0.495	17.260
15	16.912*	0.788	2.918	4.945	19.574	19.018**	0.737**	5.005**	3.986	17.608	16.057**	0.735***	5.882**	0.145	17.009
18	17.485*	0.802	3.235*	6.288	19.084	20.223**	0.740**	5.237**	5.600	17.355	16.326**	0.744**	5.375**	0.924	16.021
21	18.530*	0.805	3.803*	7.725	18.659	21.786***	0.736*	5.716**	7.216	17.293	19.200**	0.707***	6.314**	1.674	16.819
0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

from horizon one. This bias generally increases the longer the forecast horizon, as can be expected. Third, the VAR model features the smallest absolute mean forecast errors for horizons of 4-5 months and for horizons beyond 9 months, with futures coming in first for horizons of 6-8 months (Table 2, third column). Fourth, the VAR model features the smallest standard deviation of the MZ-regression residuals (Table 2, fourth column); hence the parameters in the regression equation can be estimated with the highest accuracy.[15] This finding explains why the bias coefficients of the VAR model are significant more often than those of the random walk forecasts.

Finding 1. Relative to the VAR forecast, the bias in futures forecasts is larger across all horizons and underprediction is more pronounced. Significant biases have been found for all forecasting models with horizons longer than six months. The VAR forecasts have the smallest absolute average forecast errors for horizons longer than three months and feature the lowest dispersion for all but the 24 month forecast horizon. For horizons beyond 5 months, futures perform better than the random walk, in terms of absolute mean forecast errors.

Accuracy

For a deeper discussion of forecast accuracy, Table 3 reports the RMSE of all models introduced in Table 1 relative to the RMSE of the random walk. Values less than one indicate superiority of the forecast model compared to the random walk, and * indicate when the rejection of the null hypothesis of equal predictive ability of the candidate model relative to the random walk using the Diebold-Mariano test is significant at or below the 5 percent level. Values shown in bold indicate the best forecasting model for a particular horizon. Given the popularity of futures for forecasting the oil price, Table 3 also shows the test for equal forecast accuracy of the VAR(6) models relative to futures, indicated by ǂ for the rejection of the null hypothesis of equal predictive ability.

Six points emerge from Table 3:

i. Oil futures generally outperform the random walk forecast for horizons greater than 11 months. This suggests that futures do have predictive content for such horizons.

[15] These findings generally hold for all models shown A(i) to H(i). The futures forecast generally features the largest bias for all horizons, while the random walk and VAR bias are not insubstantial either. Model G(i) features a stronger bias than the random walk for short horizons up to six months. Model H(i) is strongly biased for medium term predictions beyond 18 months, with the bias exceeding that of futures. Detrending reduces the short-term forecast bias, yet induces a larger medium-term bias.

ii. For forecast horizons from one to eight months, the random walk with drift performs (significantly) better than futures. In addition, accounting for inflation by means of a drift component does not improve the forecast performance of the random walk for any horizon. In other words, the random walk without a drift performs better.

iii. For all ARMA-type models, only the AR model with 6 lags, AR(6) is found to outperform the random walk (for all horizons) and futures (for horizons up to 12 months).

iv. In contrast, all VAR(6) specifications outperform the random walk and futures forecasts over the entire forecast horizon, with very few exceptions.[16] Gains relative to the random walk are largely significant up to the 12 month horizon, whereas gains to the futures forecasts are significant only up to 9 month horizon. All models barring H (with the REA measure of oil demand) outperform the futures forecast for horizons up to 9 months. Beyond that horizon, the detrended models largely perform worse, yet still manage to beat futures forecasts for horizons up to 12 months.

v. Relative to the random walk (and futures, given the first result in (i) above), models A and B—the baseline four-variable model augmented with inflation and our exchange rate index—generally perform best for horizons up to 6 months. Here the detrended versions, A(ii) and B(ii), provide the largest gains and outperform their trending counterparts. For horizons longer than a year, the trending models clearly perform better. Some of the forecast gains of *real* oil price trending VARs have been shown before—namely, by Baumeister and Kilian (2014) and Alquist *et al.* (2013). However, to the best of our knowledge, the performance of *nominal* oil price *detrended* models has not been studied before.

vi. Finally, our trending VAR(6) models also beat the random walk and futures forecasts for horizons of one to two years, with the exception of model H (for these outer horizons). This suggests that global industrial production outperforms the REA index as a measure of global oil demand.[17] We find that for longer horizons up to 24 months, models augmented with interest rates as a measure of the business cycle, C(i), as well as those decomposing oil supply by region, G(i), provide the most accurate forecasts. To the best of our knowledge these findings have not been shown before.

[16] While for horizons between 7- 21 the results are not statistically significant at the 5 percent level, this is due to the large variance of the forecast errors. In these instances, the VAR does neither better nor worse than the random walk.

[17] See footnote 10.

Finding 2. Futures provide more accurate forecasts than the random walk or autoregressive models only for horizons beyond 12 months, while most VARs outperform all competitors for both shorter and medium-term horizons. We find that removing the trend in oil prices improves the VAR forecasting performance for the short term even further. Beyond the one-year forecast horizon, VAR models including the exchange rate index, interest rates, and decomposed oil supply by region provide the most accurate forecasts among all competitor models.

Given our findings so far, it makes sense for the remainder of this paper to proceed with robustness checks in relation to the VAR model forecasts alone.

VI. ROBUSTNESS

Structural breaks

An important issue to check for is the presence of structural breaks in the sample, particularly at time of increased emerging market demand for oil, the onset of the global financial crisis and the unconventional oil boom in North America thereafter. A common strategy to deal with this issue would be to rely on a threshold VAR (T-VAR) model and modeling state-dependence explicitly, raising the difficult question of how to forecast the future state of the economy. Another, simpler approach is to rely on a rolling estimation of the VAR.

The notion that rolling VAR estimation for forecasting protects the forecaster against future structural changes has been contested (Baumeister and Kilian, 2012). Nonetheless, it is useful to investigate whether rolling VAR forecasts deliver an improvement in the RMSE relative to the recursive VAR forecasts. Table 4's last rows show our finding that the recursively estimated VAR performs best regardless of the length of the rolling window.[18] This holds true even without controlling for different lengths of the observation period, and implies that there are large gains in forecasting accuracy at the end of the sample, e.g., during the global financial crisis.

A related question to the presence of structural breaks is: When does our VAR(6) forecast perform best? Figure 1 and Table 3 already revealed that the predictions of our VAR model were more accurate than the futures forecast over the whole sample. For more insights on the issue of forecast stability, we next plot two-year rolling averages of the RMSE's for the best performing (i.e., the recursively estimated) VAR A(i), the random walk and futures forecasts.

[18] This is illustrated for model A(i). The same holds true for all ARMA-type models.

Table 3. Root-Mean-Square-Errors (relative to the Random Walk)

		Forecast horizons in months											
Model		**1**	**2**	**3**	**4**	**5**	**6**	**9**	**12**	**15**	**18**	**21**	**24**
RW		5.054	8.435	11.272	13.586	15.454	16.897	19.079	20.274	21.062	20.994	21.081	21.772
RW w/ drift		1.003	1.007	1.010	1.013	1.016	1.019	1.026	1.031	1.034	1.032	1.029	1.028
AR(6)		0.948	0.956	0.969	0.976	0.979	0.984	0.993	0.987	0.987	0.986	0.991	0.995
MA(3)		0.960	0.977	0.995	1.006	1.013	1.022	1.035	1.036	1.046	1.042	1.040	1.039
ARMA(1,1)		0.959	0.975	0.992	1.004	1.011	1.020	1.033	1.034	1.043	1.040	1.037	1.036
Futures		1.807	1.273	1.154	1.099	1.070	1.052	1.024	0.982	0.957	0.968	0.987	0.995
VAR(6) 1/													
A	**trend**	0.931‡	0.894‡	0.840‡	0.820	0.805	0.820	0.913	0.948	0.907	0.866	0.941	0.979
	detrend	**0.927‡**	0.883‡	**0.836‡**	**0.818‡**	0.797‡	0.806	0.894	0.962	0.987	1.057	1.299	1.535
B	**trend**	0.937‡	0.890‡	0.845‡	0.827	0.808	0.823	0.924	0.961	0.917	0.875	0.961	0.996
	detrend	0.932‡	**0.877‡**	0.841‡	0.824‡	**0.793‡**	**0.800**	0.899	0.979	1.024	1.141	1.453*	1.754*
C	**trend**	0.944‡	0.913‡	0.877‡	0.879	0.871	0.892	0.954	0.944	**0.895**	**0.863**	0.931	0.956
	detrend	0.939‡	0.900‡	0.871‡	0.879	0.867	0.882	0.962	1.054	1.175	1.406*	1.833*	2.256*
D	**trend**	0.972‡	0.949‡	0.901‡	0.894	0.880	0.898	0.956	0.947	0.898	0.865	0.928	0.947
	detrend	0.968‡	0.936‡	0.897‡	0.896	0.878	0.891	0.975	1.085	1.235	1.505*	1.974*	2.439*
E	**trend**	0.959‡	0.923‡	0.879‡	0.861	0.841	0.846	0.935	0.956	0.912	0.876	0.943	0.958
	detrend	0.956‡	0.907‡	0.868‡	0.855	0.829	0.835	0.975	1.105	1.223	1.448*	1.856*	2.242*
F	**trend**	0.951‡	0.912‡	0.868‡	0.846	0.822	0.833	0.912	0.959	0.928	0.888	0.973	1.013
	detrend	0.946‡	0.901‡	0.867‡	0.846‡	0.813	0.819	0.899	0.985	1.030	1.127	1.416	1.703
G	**trend**	1.071‡	1.087	1.050	1.015	0.966	0.926	**0.891**	**0.917**	0.927	0.884	**0.904**	**0.933**
	detrend	1.066‡	1.074	1.053	1.029	0.984	0.949	0.960	1.106	1.293‡*	1.533‡*	1.939‡*	2.412‡*
H	**trend**	1.009‡	0.997‡	0.972‡	0.978	0.970	0.977	1.031	1.080	1.090	1.117	1.198	1.265
	detrend	1.001‡	0.994‡	0.998	1.033	1.040	1.084	1.287‡*	1.516‡*	1.783‡*	2.210‡*	2.851‡*	3.509‡*

1/ All VARs shown have six lags. Models with longer lags generally perform worse, in part likley due to estimation inefficiency.

2/ Cells in bold indicate the model with the lowest RMSE for each column/forecast horizon.

3/ * indicates rejection of the null hypothesis of equal predictive ability of the candidate model relative to the random walk model using the Diebold-Mariano test at or less than the 5 percent significance level.

4/ ‡ indicates rejection of the null hypothesis of equal predictive ability of the candidate model relative to futures using the Diebold-Mariano test at or less than the 5 percent significance level.

Table 4. Root-Mean-Square-Errors for Rolling and Recursive Estimations

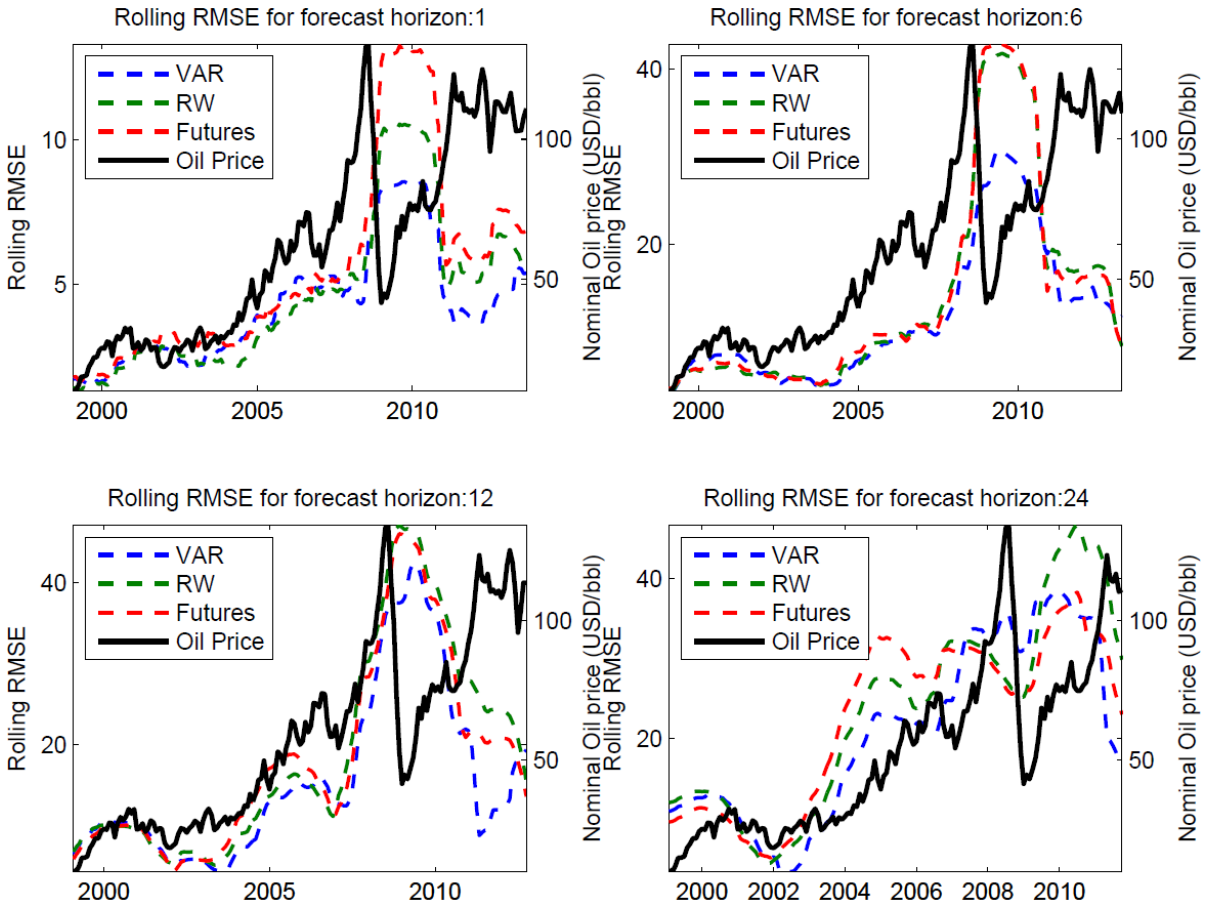
Forecast horizons in months

Start of evaluation period	Window size	1	2	3	4	5	6	7	8	9	12
					Rolling estimation						
1993M1	8 years	1.367	1.360	1.411	1.372	1.348	1.406	1.509	1.628	1.748	2.229
1995M1	10 years	1.340	1.277	1.226	1.179	1.113	1.135	1.214	1.304	1.421	1.663
1997M1	12 years	1.190	1.145	1.105	1.072	1.044	1.074	1.145	1.221	1.291	1.391
1999M1	14 years	1.135	1.124	1.071	1.019	0.994	1.010	1.069	1.132	1.184	1.216
2001M1	16 years	1.058	1.068	1.022	0.971	0.957	0.970	1.017	1.060	1.088	1.111
2003M1	18 years	1.022	1.015	0.955	0.894	0.871	0.887	0.934	0.982	1.013	1.039
2005M1	20 years	0.979	0.946	0.891	0.871	0.875	0.902	0.940	0.981	1.017	1.034
					Recursive estimation						
1997M1		**0.939**	**0.895**	**0.851**	**0.830**	**0.811**	**0.819**	**0.840**	**0.872**	**0.908**	**0.941**

Start of evaluation period	Window size	13	14	15	16	17	18	19	20	21	24
					Rolling estimation						
1993M1	8 years	2.261	2.343	2.516	2.854	3.294	3.747	4.241	5.108	6.135	12.546
1995M1	10 years	1.645	1.608	1.574	1.570	1.584	1.615	1.648	1.695	1.745	1.849
1997M1	12 years	1.372	1.341	1.314	1.305	1.310	1.330	1.360	1.409	1.465	1.533
1999M1	14 years	1.188	1.155	1.127	1.107	1.100	1.102	1.120	1.160	1.217	1.293
2001M1	16 years	1.085	1.052	1.025	1.004	0.992	0.979	0.988	1.020	1.074	1.143
2003M1	18 years	1.014	0.983	0.956	0.939	0.927	0.913	0.922	0.953	1.005	1.091
2005M1	20 years	1.010	0.984	0.965	0.952	0.940	0.928	0.941	0.984	1.041	1.140
					Recursive estimation						
1997M1		**0.919**	**0.903**	**0.884**	**0.868**	**0.852**	**0.842**	**0.849**	**0.879**	**0.921**	**0.990**

Figure 2 shows that while the random walk forecast performs best for short horizons during periods of stable oil prices up to 2005, the VAR performs better than both the random walk and the futures forecasts during periods of high oil prices—specifically since 2006. For forecast horizons shorter than one year, the gains of the VAR model stem mostly from the period following the collapse in oil prices in 2008, while before this period the performance of all three forecasts are very similar. Towards the end of the sample, the performance is ambiguous. For longer horizons up to 24 months, the VAR model performs better than its competitors in almost all periods, and particularly during the run up in oil price.

Figure 2. Rolling Averages of Root Mean Square Errors



Sample splits

The performance gains appear to mostly stem from the period of increasing oil prices between 2003 and 2008 where the futures and random walk forecasts generally unpredicted oil prices, and from periods of high oil volatility following the global financial crisis. A statistical test of significant differences by means of the procedure outlined by Giacomini and Rossi (2010) is not applicable here since we rely on recursive estimation. Therefore, we split our sample at 2008M06 and investigate the relative RMSE before this date when oil prices were stable but running upward, and then after this date when they were volatile culminating in the collapse during the second half of 2014.[19]

[19] We also evaluate the performance of other samples splits of interest. For example, 1997M01-2002M12, 1997M01-2008M06, and 1997M01-2014M06. Results are available upon request.

The results are shown in Tables 5 and 6. Interestingly, we find that the futures forecast performed poorer than the random walk and the simple autoregressive models until the global financial crisis for all forecast horizons (Table 5). Medium-run gains stem solely from the period since the crisis, where it outperforms the random walk and ARMA models for horizons of more than one year (Table 6). Since the absolute level of the forecast error variance is considerably larger in this period, the futures forecast beat the random walk and ARMA models over the full sample too, for those medium-term horizons.

For our VAR model forecasts, we find that most gains to the short-run forecasts stem from the period of the crisis (Table 6), while for longer forecast horizons up to 24 months, the gains are relatively larger during the period up to 2008M06 (Table 5). Prior to the crisis, most VAR models did not beat the random walk model for horizons up to twelve months, but both the futures and random walk forecasts are generally outperformed by the VAR for forecasts of one year ahead up to 24 months. Those medium-term forecasts are largely significantly better than the random walk and futures forecasts—but the shorter ones are not statistically different from the competitor models' predictions (with the exception of the two and four month horizons). During and after the crisis, most VAR models always outperform competitor models, with gains for shorter-term forecasts up to 12 months being particularly large and statistically significant. Again, the detrended models predict the short-run relatively well during this sample split.

Including our exchange rate index improves the forecasting performance in the earlier sample split, model B, when emerging market demand mattered a great deal for the determination of oil prices, while disaggregating global supply and demand provides consistent gains in the more recent period when North American supply may have contributed to lower prices. In addition, VAR models augmented with interest rates, models C and D, perform relatively well in the later sample. It also becomes apparent that the REA index (Kilian 2009) includes valuable information for medium-run forecasts of oil prices up to 2008, whereby model H outperforms other VAR specifications and futures alike. However, from the onset of the crisis, the medium-run predictive ability strongly deteriorates. Moreover, model H was clearly outperformed for short-run forecasts by rival VAR specifications using global industrial production as a proxy for oil demand.[20]

Lag length

Alquist *et al*. (2013) provide evidence that a shorter VAR lag length of $p = 12$ provides more accurate forecasts than $p = 24$ when the real oil price is specified in log

[20] See footnote 10.

levels. Baumeister and Kilian (2015) argue the opposite: longer lag lengths (specifically, $p = 24$) provide more accurate forecasts. We explored longer lags for our results shown so far (Tables 3-6), and found that $p = 6$ generally performed better than VAR specifications with $p > 6$, with the exception of a few models (namely, A, B, and F) when $p = 18$. These few models performed equally well only for forecast horizons between 18 to 24 months and only during the post crisis sample split.[21] Better overall performance of the shorter lag length is likely due to estimation inefficiency at the beginning of the evaluation period.

Since we have eliminated long-run cointegrating relationships—by estimating our VARs in log-differences—increasing the lag length reduces estimation efficiency. However, as shown in Section VIII, when estimating over a longer sample, longer lag lengths may at times provide good forecasts and reduce confidence interval dispersion, narrowing these bands. Before turning to this issue, however, we next explore forecast combinations.

Finding 3. Forecasting performance varies across time, with some performance instability at times. The random walk forecast performed better during periods of stable oil prices, but various VAR specifications have performed best since 2008 across almost all specifications and horizons. Indeed since the collapse in oil prices during 2008, the VAR was found to have superior forecasting performance. Measures of activity based on the Baltic Dry Index included valuable information for forecasting oil prices prior to 2008, but thereafter have performed poorly.

VII. FORECAST COMBINATIONS

We have shown that the VAR model provides better forecasts than both the random walk and futures forecasts across time and for almost all forecast horizons. However, we have also shown that futures or random walk forecast with similar or better accuracy at some horizons and in different subsamples. Thus, there could be merit in a combination forecast (Bates and Granger, 1969; Diebold and Pauly, 1987; and Stock and Watson, 2004) for several reasons. First, even the most accurate forecasting models do not necessarily perform well at all times (Figure 2). Second, we have shown that some forecasting models perform better at short horizons and others at longer horizons. Third, even the forecasting model with the lowest RMSE may potentially improve by incorporating information from other models

[21] Results for Tables 3-7 for VAR lag lengths of 12, 18, and 24 are available upon request.

Table 5. Root-Mean-Square-Errors (relative to the Random Walk): 1997M01 - 2008M06

		Forecast horizons in months											
Model		**1**	**2**	**3**	**4**	**5**	**6**	**9**	**12**	**15**	**18**	**21**	**24**
RW		2.565	3.734	4.424	5.083	5.649	6.363	8.134	9.905	11.599	13.303	15.154	18.254
RW w/ drift		**0.996**	**0.991**	**0.984**	**0.982**	**0.981**	0.982	**0.984**	0.985	0.982	0.966	0.952	0.939
AR(6)		1.037	1.070	1.065	1.074	1.065	1.058	1.050	1.025	1.012	0.997	0.982	0.967
MA(3)		1.031	1.027	1.004	1.010	0.992	0.993	0.998	0.985	0.987	0.974	0.961	0.950
ARMA(1,1)		1.026	1.024	1.001	1.007	0.988	0.989	0.995	**0.984**	0.984	0.972	0.958	0.947
Futures		1.436	1.188	1.173	1.175	1.207	1.218	1.236	1.225	1.220	1.215	1.200	1.160
VAR(6)													
A	**trend**	1.152ǂ	1.202*	1.206*	1.155*	1.053ǂ	1.007ǂ	1.050ǂ	1.043ǂ	0.957ǂ	0.880ǂ*	0.879ǂ*	0.861ǂ*
	detrend	1.153ǂ	1.187*	1.182*	1.121	1.014ǂ	**0.981ǂ**	1.032ǂ	1.074	1.051	1.014	1.075	1.069
B	**trend**	1.193ǂ	1.246*	1.234*	1.170*	1.074ǂ	1.023ǂ	1.050ǂ	1.036ǂ	**0.901ǂ***	**0.795ǂ***	**0.797ǂ***	**0.780ǂ**
	detrend	1.194ǂ	1.235*	1.223*	1.154*	1.065ǂ	1.036ǂ	1.077ǂ	1.110	1.045	0.998ǂ	1.097	1.130
C	**trend**	1.171ǂ	1.217*	1.204*	1.166*	1.064ǂ	1.011ǂ	1.025ǂ	1.012ǂ	0.926ǂ*	0.857ǂ*	0.861ǂ*	0.840ǂ*
	detrend	1.171ǂ	1.207*	1.209*	1.190*	1.094	1.059	1.158	1.372*	1.513ǂ*	1.668ǂ*	1.917ǂ*	2.015ǂ*
D	**trend**	1.193	1.248*	1.231*	1.181*	1.082	1.043ǂ	1.073ǂ	1.042ǂ	0.943ǂ	0.868ǂ*	0.862ǂ*	0.834ǂ*
	detrend	1.191ǂ	1.237*	1.247*	1.222*	1.136	1.125	1.288*	1.545*	1.758*	1.990*	2.316*	2.457*
E	**trend**	1.188ǂ	1.238*	1.257*	1.222*	1.125	1.061ǂ	1.065ǂ	1.049ǂ	0.939ǂ	0.851ǂ*	0.854ǂ*	0.836ǂ*
	detrend	1.194ǂ	1.214*	1.197*	1.139	1.035	0.991ǂ	1.021ǂ	1.085	1.058	1.050	1.177	1.237*
F	**trend**	1.149ǂ	1.235*	1.265*	1.200*	1.096	1.033ǂ	1.040ǂ	1.030ǂ	0.915ǂ*	0.810ǂ*	0.810ǂ*	0.793ǂ*
	detrend	1.149ǂ	1.224*	1.255*	1.182*	1.075	1.029ǂ	1.046ǂ	1.083	1.034	0.969	1.044	1.058
G	**trend**	1.486	1.777*	1.936ǂ*	1.848*	1.679	1.430	1.032ǂ	1.028ǂ	1.002ǂ	0.921ǂ*	0.865ǂ*	0.839ǂ*
	detrend	1.485*	1.758*	1.972*	1.928	1.774	1.553*	1.171	1.249*	1.391*	1.441*	1.512ǂ*	1.544ǂ*
H	**trend**	1.234*	1.260*	1.260*	1.244*	1.182*	1.135*	1.122ǂ*	1.103ǂ*	1.007ǂ	0.909ǂ*	0.887ǂ*	0.844ǂ*
	detrend	1.234*	1.257*	1.338*	1.389*	1.399*	1.434*	1.656ǂ*	2.035ǂ*	2.387ǂ*	2.796ǂ*	3.340ǂ*	3.709ǂ*

1/ All VARs shown have six lags. Models with longer lags generally perform worse, in part likely due to estimation inefficiency.

2/ Cells in bold indicate the model with the lowest RMSE for each column/forecast horizon.

3/ * indicates rejection of the null hypothesis of equal predictive ability of the candidate model relative to the random walk model using the Diebold-Mariano test at or less than the 5 percent significance level.

4/ ǂ indicates rejection of the null hypothesis of equal predictive ability of the candidate model relative to futures using the Diebold-Mariano test at or less than the 5 percent significance level.

Table 6. Root-Mean-Square-Errors (relative to the Random Walk): 2008M07 - 2014M12

		Forecast horizons in months											
Model		**1**	**2**	**3**	**4**	**5**	**6**	**9**	**12**	**15**	**18**	**21**	**24**
RW		7.105	12.122	16.564	20.099	22.994	25.151	27.976	29.131	29.947	29.322	27.933	26.776
RW w/ drift		1.005	1.010	1.014	1.018	1.022	1.025	1.035	1.043	1.052	1.061	1.071	1.100
AR(6)		0.923	0.940	0.963	0.972	0.977	0.984	0.994	0.995	0.990	0.999	1.009	1.033
MA(3)		0.935	0.962	0.989	1.002	1.013	1.025	1.042	1.049	1.060	1.074	1.080	1.112
ARMA(1,1)		0.934	0.961	0.987	1.000	1.012	1.023	1.040	1.048	1.058	1.072	1.077	1.108
Futures		1.829	1.272	1.153	1.102	1.070	1.047	1.004	0.957	0.907	0.885	**0.900**	**0.890**
VAR(6) 1/													
A	**trend**	0.895ǂ*	0.856ǂ*	0.805ǂ*	0.791ǂ*	0.786ǂ*	0.806ǂ*	0.900	0.937	0.900	**0.864**	0.967	1.052ǂ
	detrend	0.890ǂ*	0.846ǂ*	0.804ǂ*	**0.793ǂ***	0.781ǂ*	0.793ǂ*	0.881*	0.948	0.977	1.071ǂ	1.381ǂ	1.780ǂ
B	**trend**	0.894ǂ	0.845ǂ*	0.808ǂ*	0.798ǂ*	0.788ǂ*	0.807ǂ*	0.913	0.952	0.922	0.897	1.021ǂ	1.118ǂ
	detrend	**0.887ǂ**	**0.831ǂ***	**0.804ǂ***	0.796ǂ*	**0.772ǂ***	**0.782ǂ***	0.881*	0.961	1.023	1.179ǂ*	1.577ǂ*	2.068ǂ*
C	**trend**	0.906ǂ	0.877ǂ*	0.847ǂ*	0.856ǂ	0.858ǂ	0.885	0.949	0.936	**0.891***	0.867	0.960	1.027ǂ
	detrend	0.901ǂ	0.863ǂ*	0.840ǂ*	0.853ǂ*	0.850ǂ*	0.870ǂ*	0.942	1.004	1.101ǂ	1.329ǂ*	1.801ǂ*	2.407ǂ*
D	**trend**	0.935ǂ	0.914ǂ	0.871ǂ	0.871ǂ	0.866ǂ	0.888	0.946	0.935	**0.891***	0.865*	0.955	1.017ǂ
	detrend	0.931ǂ	0.900ǂ	0.864ǂ*	0.869ǂ	0.858ǂ	0.873ǂ	0.940	1.005	1.111ǂ	1.346ǂ*	1.823ǂ*	2.431ǂ*
E	**trend**	0.921ǂ	0.884ǂ	0.842ǂ	0.831ǂ	0.820ǂ	0.830ǂ	0.924	0.944	0.909	0.885	0.979ǂ	1.032ǂ
	detrend	0.916ǂ	0.871ǂ*	0.838ǂ*	0.832ǂ*	0.814ǂ*	0.824ǂ*	0.973	1.110ǂ*	1.255ǂ*	1.542ǂ*	2.072ǂ*	2.713ǂ*
F	**trend**	0.918ǂ	0.873ǂ*	0.829ǂ*	0.816ǂ*	0.801ǂ*	0.818ǂ*	0.901	0.951	0.933	0.910	1.033ǂ	1.136ǂ*
	detrend	0.912ǂ	0.861ǂ*	0.829ǂ*	0.818ǂ*	0.793ǂ*	0.803ǂ*	0.885	0.973	1.032	1.168ǂ*	1.544ǂ*	2.022ǂ*
G	**trend**	0.997ǂ	0.988ǂ	0.946ǂ	0.928	0.899	0.879	**0.877**	**0.903**	0.914	0.876	0.921	0.992
	detrend	0.991ǂ	0.976ǂ	0.943ǂ	0.932ǂ	0.906	0.891	0.939	1.086ǂ	1.277ǂ*	1.561ǂ*	2.091ǂ*	2.847ǂ*
H	**trend**	0.972ǂ	0.967ǂ	0.947ǂ	0.958	0.956ǂ	0.967	1.024	1.080	1.108ǂ	1.170ǂ	1.305ǂ*	1.481ǂ*
	detrend	0.963ǂ	0.964ǂ	0.968ǂ	1.004	1.013	1.056	1.246ǂ*	1.430ǂ*	1.646ǂ*	2.026ǂ*	2.634ǂ*	3.376ǂ*

1/ All VARs shown have six lags. Models with longer lags generally perform worse, in part likely due to estimation inefficiency.

2/ Cells in bold indicate the model with the lowest RMSE for each column/forecast horizon.

3/ * indicates rejection of the null hypothesis of equal predictive ability of the candidate model relative to the random walk model using the Diebold-Mariano test at or less than the 5 percent significance level.

4/ ǂ indicates rejection of the null hypothesis of equal predictive ability of the candidate model relative to futures using the Diebold-Mariano test at or less than the 5 percent significance level.

or macroeconomic factors. Finally, employing forecast combinations can partially insure against structural change and model misspecification (Baumeister and Kilian, 2013).

Hence we next present a combination forecast of futures and the baseline VAR specification, model A(i). We retain futures despite their weak performance at the 1 to 6 month horizons since futures are said to contain valuable forward looking information and can reflect market uncertainty—possibly presaging structural or risk premia. In particular, we expect some improvement compared to the VAR forecast for longer horizons up to 24 months where the futures-based forecast outperformed the random walk.

The most important issue for the forecast combination is the question of how to weight each of the forecasts that are being combined. Building on the finding that simple combinations are hard to beat, we experiment with equal (and constant) weights as well as with inverse-RMSE weights for different lengths of rolling windows (Timmermann, 2006). The inverse-RMSE weighted h-period ahead combination forecast $\hat{P}_{t+h|t}$ is hence obtained:

$$\hat{P}_{t+h|t} = \sum_n \omega_{n,h,t} \hat{P}^n_{t+h|t}, \qquad \omega_{n,h,t} = \frac{RMSE^{-1}_{n,h,t}}{\sum_n RMSE^{-1}_{n,h,t}} \tag{8}$$

where $\hat{P}^n_{t+h|t}$ is the h-period ahead forecast of model n (here $n = \{VAR, futures\}$) and $RMSE^{-1}_{n,h,t}$ is the horizon h-specific RMSE of model n of the most recent w forecast errors realized in period t. We experiment with rolling windows of $w = \{6, 12, 18, 24\}$ and find the results to be robust throughout. The equal weights combination forecast is obtained similarly with $\omega_{n,h,t} = 0.5 \; \forall \; n, h, t$.

As shown in Table 7, we find that the combination forecasts all perform worse than the simple VAR model for horizons $h < 18$, which can largely be explained by the poor performance of the futures and/or the random walk forecast for these horizons. For longer horizons, however, the combination forecast irrespective of its specification outperforms the VAR model. The combination design seems to play only a minor role as these results are robust for all weights, although the combination based on inverse RMSE weights computed over 24 months performs best. . These results are robust to longer VAR lag lengths.

Finding 4. Combination forecasts performed poorly for horizons under 18 months. For horizons beyond 18 month, however, there is merit in a combination forecast—particularly for a combination with inverse forecast accuracy weights computed over 24 months.

Table 7. Root- Mean-Square-Errors of Forecast Combinations

	Forecast horizons in months											
Model	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Random Walk	5.835	9.823	13.078	15.681	17.799	19.547	20.876	21.913	22.626	23.067	23.522	24.091
Futures	1.214	1.014	1.019	1.017	1.015	1.009	1.000	0.990	0.982	0.975	0.965	0.954
VAR(6)	**0.885**	**0.872**	**0.821**	**0.816**	**0.796**	**0.791**	**0.797**	**0.814**	**0.843**	**0.871**	**0.879**	**0.869**
Inverse-RMSE (w=24)	1.629	1.165	1.025	0.973	0.942	0.932	0.925	0.924	0.933	0.946	0.944	0.933
Inverse-RMSE (w=12)	1.632	1.162	1.030	0.981	0.952	0.938	0.932	0.928	0.930	0.934	0.926	0.908
Equal weights	1.636	1.160	1.027	0.978	0.949	0.936	0.929	0.925	0.928	0.933	0.924	0.908

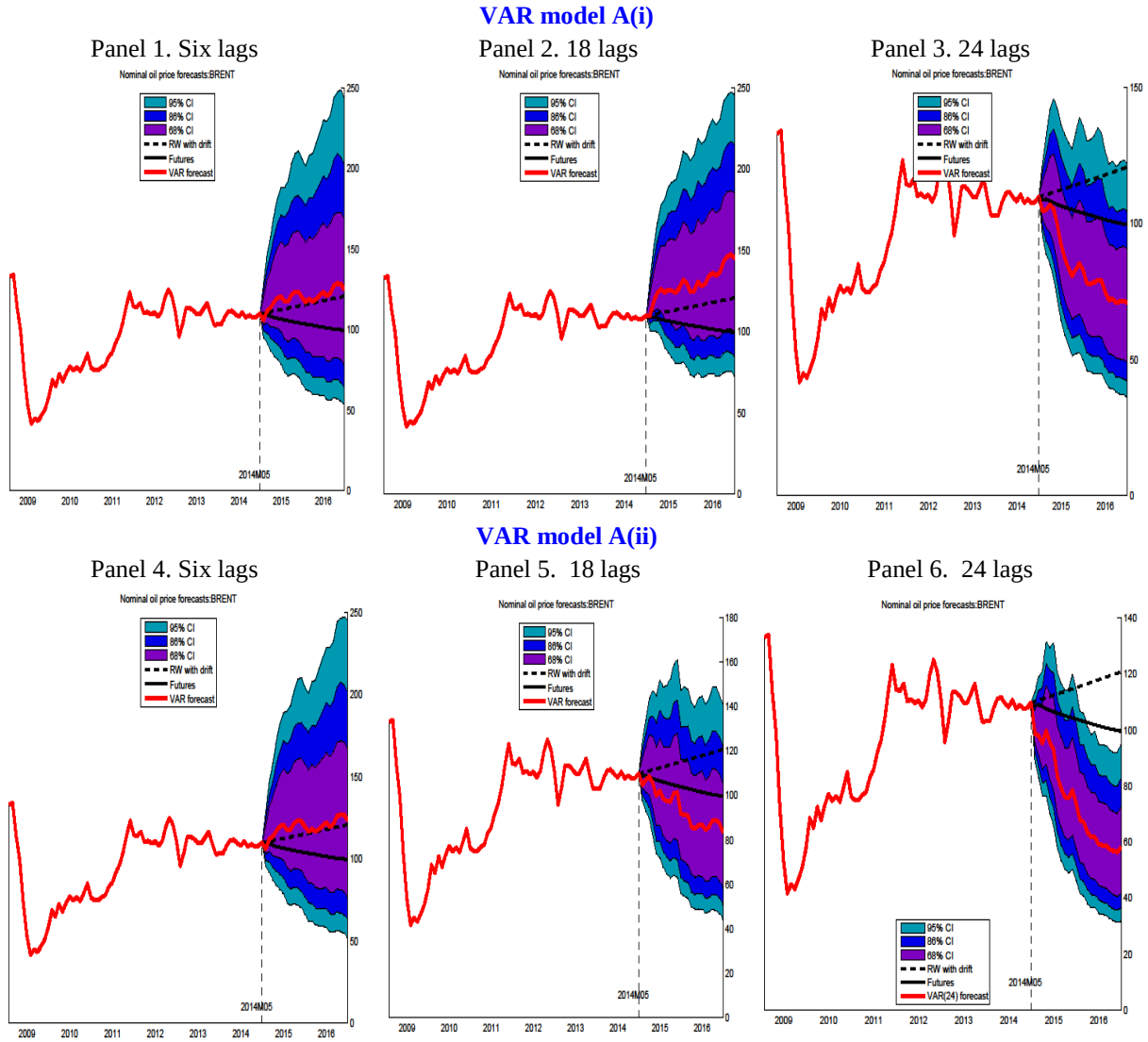
Model	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**
Random Walk	24.717	25.217	25.538	25.853	26.139	26.380	26.606	26.887	27.202	27.617	28.213	28.968
Futures	0.941	0.934	0.929	0.926	0.924	0.924	0.927	0.930	0.934	0.937	0.940	0.940
VAR(6)	**0.856**	**0.851**	**0.834**	**0.816**	**0.798**	**0.788**	0.799	0.808	0.838	0.861	0.872	0.879
Inverse-RMSE (w=24)	0.908	0.879	0.851	0.832	0.817	0.803	**0.798**	**0.795**	**0.804**	**0.813**	**0.813**	**0.806**
Inverse-RMSE (w=12)	0.888	0.874	0.858	0.843	0.827	0.815	0.810	0.807	0.815	0.823	0.822	0.817
Equal weights	0.889	0.875	0.859	0.844	0.828	0.816	0.811	0.808	0.816	0.824	0.822	0.817

The evaluation period runs from 2000M01:2014M12 (1998M01:2014M12) for the inverse RMSE weighted forecast combination with w=24 (w=12) since the data from 1997-2000 (1997-1998) is needed for the initialization of the simple VAR forecasts and the rolling RMSE evaluation.

VIII. THE 2014 OIL PRICE COLLAPSE AND MODEL CHOICE

Despite the overall strength of VAR models, we have highlighted some forecasting performance instability at times. This finding implies that reliance on just one VAR model with the lowest RMSE per horizon is not advisable, despite the fact that our baseline VAR generally has been shown to have the lowest RMSE across all popular models, the full sample, and across horizons up to 24 months. Other VAR specifications that feature small alterations to the selected variables, subsamples or lag length can provide widely different forecasts at times. This wide variation in predication is central to why the oil price collapse in 2014 was largely unexpected—with Brent prices falling from US$ 115 to US$ 55 in late 2014.

Figure 3. Ex Ante Nominal Brent Price Forecast, as of May 2014
(US$)



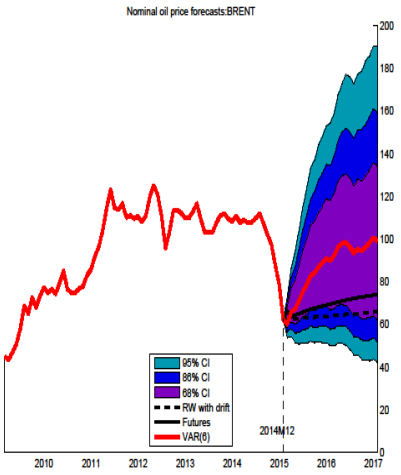
Baumeister and Kilian (2015) show that the oil price collapse was predictable using their VAR with 24 months of lags and real oil prices specified in logs. Our VAR, Model A(i), despite showing superior historical forecast performance over the full sample would not have been able to predict the crash in oil prices; rather it predicts increasing prices (Figure 3, panel 1). The same applies to our detrended model A(ii) with six lags (Figure 3, panel 4). Of the two, only the detrended model with 24 lags would have predicted the full extent of the oil price collapse (Figure 3, panel 6).[22] If anything, this illustrates that best performance over the entire historical sample may not be the right approach for the practitioner from an *ex ante* point of view, as new data and periods arise—highlighting that the best historical forecast model may be off by a considerable margin at times and such a model should be thus viewed with caution.

Next we provide *ex ante* nominal forecasts for Brent oil prices with data as of end December 2014. Figure 4 shows the range of prices that our model A(ii) with various lag lengths predicts. In particular, Brent prices were predicted to range between an average of US$55-95 pb over a two year forecast horizon, depending on the history of oil-market conditions.

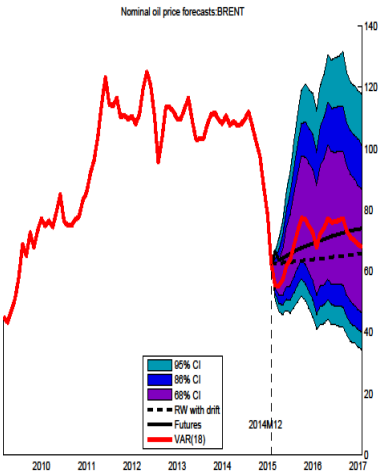
Figure 4. Ex Ante Nominal Brent Price Forecasts, as of December 2014
(US$)

VAR model A(ii)

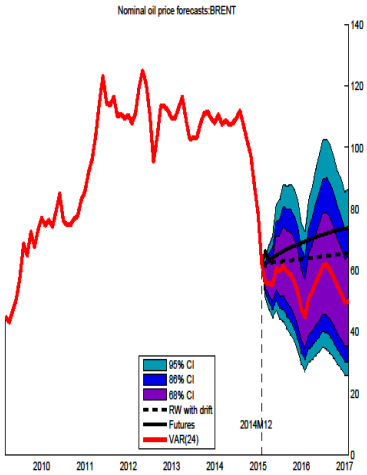
Panel 1. Six lags



Panel 2. 18 lags



Panel 3. 24 lags



[22] For the actual Brent price collapse and its drivers see Box 1.1 of the April 2015 *World Economic Outlook*.

Two points emerge from the figure. First, when considering the model with 24 lags, which predicts that the trend in OECD production would continue as it did two years prior—and thus with global supply predicted to reach 97 mbd over the forecast horizon along with OPEC spare capacity rising to record highs (5 mbd)—Brent prices are forecast to remain low, averaging US$55 pb as demand is expected to be in line with consensus forecasts (Figure 4, panel 3). Second, when OECD production be capped at the slightly lower production which prevailed during the last six months of 2014—such that OPEC spare capacity would then be predicted to fall back to or below 2 mbd—then Brent prices are forecast to rise, averaging US$95 pb (Figure 4, panel 1). Hence the range of forecasts of US$55-95 pb.

Finding 5. Despite the overall strength of VAR models, we highlight performance instability over the full sample, with small alterations in specifications, subsamples and lag lengths providing widely different forecasts at times.

IX. CONCLUSION

Our analysis in this paper employs a set of monthly data from 1985M01 to 2014M12 and allows us to evaluate *ex post* the out-of-sample forecasting performance of several leading benchmark oil price models starting in 1997M01. We compare the different models' predictive abilities over a horizon of approximately 18 years.

We find that the exact specification of a VAR model for oil price prediction is still open to debate. While our standard parsimonious VAR model of oil demand proxied by industrial production, global oil supply and OECD inventory demand performs well over the full sample (despite the presence of structural breaks), we show that there is value in other specifications, as well as in futures forecasts during subsample intervals. However, across our whole sample, we find that the bias of futures forecasts to be larger than that of the VAR, with the direction of the bias indicating a serve underprediction of the futures forecast. Nonetheless, futures do provide more accurate forecasts relative to other simple benchmarks (such as the random walk) for horizons between 12 to 24 months. But futures do not provide more accurate forecasts than our VAR, which outperforms all competitors for all horizons under 24 months.

As expected, forecasting performance varies across time. We find the random walk forecast to have performed better during periods of stable oil prices, but the VAR has performed best since 2008. Despite the overall strength of our VAR models, their performance can suffer from instability over the full sample, with small alterations in specifications, subsamples or lag lengths providing widely different forecasts at times. Therefore, we find merit in combination forecasts for horizons beyond 18 months and

conclude that predicting oil prices on a long sample of data with structural breaks remains difficult.

Finally, a limitation of our class of VAR models is the implied prediction intervals. As the model is only mean-reverting in changes but not in levels, there is no long-run anchor or "equilibrium" oil price in the global oil market. Therefore, a VECM model could overcome this shortcoming and presents a potentially promising area for future research.

REFERENCES

Aastveit, K. A., H. Bjørnland, and L. A. Thorsrud, 2014, "What Drives Oil Prices? Emerging Versus Developed Economies", *Journal of Applied Econometrics*.

Alquist, R., L. Kilian, and R. J. Vigfusson, 2013, "Forecasting the Real Oil Price," forthcoming in the *Handbook of Economic Forecasting*, 2, Amsterdam: North Holland.

Bates, J. M., and C. W. J. Granger, 1969, "The Combination of Forecasts," *Operational Research Society*, 20(4), 451–468.

Baumeister, C. and L. Kilian, 2012, "What Central Bankers Need to Know About Forecasting Oil Prices," *CEPR Discussion Papers* 9118.

_____, 2013, "Forecasting the Real Oil Price in a Changing World: A Forecast Combination Approach," *CEPR Discussion Papers* 9569.

_____, 2014, "Real-Time Forecasts of the Real Price of Oil," *IMF Economic Review*, Palgrave Macmillan, vol. 62(1), pages 119-145, April.

_____, 2015, "Understanding the Decline in the price of Oil since June 2014", forthcoming in the *Journal of the Association of Environmental and Resource Economists*.

Beidas-Strom, S., and A. Pescatori, 2014, "Oil Price Volatility and the Role of Speculation," *IMF Working Paper* 14/218, International Monetary Fund.

Benes, J., M. Chauvet, O. Kamenik, M. Kumhof, D. Laxton, S. Mursula, and J. Selody, 2015, "The future of oil: Geology versus technology," *International Journal of Forecasting*, Elsevier, vol. 31(1), pages 207-221.

Chen, Y-C., K. S. Rogoff, and B. Rossi, 2010, "Can Exchange Rates Forecast Commodity Prices?" *Quarterly Journal of Economics* 125(3), 1145–1194.

Diebold, F. X., and P. Pauly, 1987, "Structural Change and the Combination Forecasts," *Journal of Forecasting*, 6, 21–40.

Diebold, F. X., and R. S. Mariano, 1995, "Comparing Predictive Accuracy," *Journal of Business & Economic Statistics*, 13(3), 134–144.

Giacomini, R., and B. Rossi, 2010, "Forecast Comparisons in Unstable Environments," *Journal of Applied Econometrics*, 25, 595–620.

Grisse, C, 2010. "What Drives the Oil-Dollar Correlation?" Federal Reserve Bank of New York, December 2010. Mimeo.

Hamilton, J. D., 1994, "Time Series Analysis," Princeton University Press.

Hotelling, H., 1931, "The Economics of Exhaustible Resources," *The Journal of Political Economy*, 39(2), 137– 175.

Kilian, L., 2009, "Not all Oil Price Shocks Are Alike: Disentangling Demand and Supply Shocks in the Crude Oil Market," *American Economic Review*, 99, 1053-1069.

Kilian, L. and D. P. Murphy, 2014, "The Role of Inventories and Speculative Trading in the Global Market for Crude Oil." *Journal of Applied Econometrics*, 29: 454–478.

Lütkepohl, H., 2007, "New Introduction to Multiple Time Series Analysis," Springer Publishing Company, Incorporated.

Mincer, J. A., and V. Zarnowitz, 1969, "The Evaluation of Economic Forecasts," NBER Chapters in Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance, 1–46.

Reichsfeld, D. A., and S. K. Roache, 2011, "Do Commodity Futures Help Forecast Spot Prices?" *IMF Working Paper* 11/254. International Monetary Fund.

Stock, J. H., and M. W. Watson, 2004, "Combination Forecasts of Output Growth in a Seven-Country Data Set," *Journal of Forecasting*, 23, 405–430.

The Economist, 2015, "Dry-bulk cargo shipping: Hitting the bottom—Worse is still to come for many bulk carriers", Oct 31st 2015, From the print edition http://www.economist.com/news/business/21677207-worse-still-come-many-bulk-carriers-hitting-bottom

Timmermann, A., 2006, "Forecast Combinations", in the *Handbook of Economic Forecasting*, 1, Elsevier.