ISSN 1936-5330

kcFED Research Working Papers

Big Data Meets the Turbulent Oil Market

Charles W. Calomiris, Nida Cakir Melek, and
Harry Mamaysky

December 2020; updated November 2022
RWP 20-20
http://doi.org/10.18651/RWP2020-20

This paper supersedes the previous version:
"Predicting the Oil Market"

Federal Reserve Bank of Kansas City

Big Data Meets the Turbulent Oil Market

Charles W. Calomiris, Nida Çakır Melek, and Harry Mamaysky[*]

November 2022

Abstract

This paper introduces novel news-based measures for tracking global energy markets. These measures compress thousands of news articles into a parsimonious set of real-time indicators and are successful in-sample forecasters of oil spot, futures, and energy company stock returns, and of changes in oil volatility, production, and inventories, complementing and extending traditional (non-text) predictors. In out-of-sample tests, text-based measures predict oil futures returns and changes in oil spot prices better than traditional predictors, although the latter are more useful for forecasting changes in oil volatility.

Keywords: Asset Pricing, Commodity Markets, Energy Forecasting, Model Validation
JEL: C52, G10, G12, G14, G17, Q47

Our energy words list, text measures, and code are available at:
github.com/hmamaysky/Energy.

[*] Columbia Business School and NBER, cc374@gsb.columbia.edu, Federal Reserve Bank of Kansas City, nida.cakirmelek@kc.frb.org, and Columbia Business School, hm2646@columbia.edu, respectively. The authors are listed in alphabetical order. We thank Christiane Baumeister, Kateryna Holland (discussant), Evgenia Passari (discussant), Brad Paye (discussant), Geert Bekaert, Huixin Bi, Deepa Datta, Jim Hamilton, Lee Smith, Jim Stock, and participants at the EIA 2019 Annual Workshop on Financial and Physical Energy Market Linkages, the FRB-IMF Workshop on New Techniques and Data in Macro Finance, the 2020 CEBRA Workshop for Commodities and Macroeconomics, the 2021 Conference on Non-traditional Data, Machine Learning and Natural Language Processing in Macroeconomics, the 2021 IAAE Annual Meeting, the 4th Annual JPMCC International Commodities Symposium, the 2022 AFA Annual Meeting, the 2022 CEMA Annual Meeting, the Federal Reserve Board, and the Federal Reserve Bank of Kansas City for helpful comments and suggestions. Daliah Al-Shakhshir, Roya Arab Loodaricheh, Sungil Kim, and Hongyu Wu provided excellent research assistance. Colton Tousey helped with construction of the energy words list. The views expressed herein are solely those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Kansas City, or the Federal Reserve System.

1. Introduction

The world is in the midst of an energy crisis, probably the first of its kind. At a time of unexpectedly strong rebound in demand, the oil market must confront a notable supply shortfall in the aftermath of Russia's invasion of Ukraine, combined with political constraints on increasing production in developed economies. What useful information can forecasting models of market returns and risks provide investors at the current moment of extreme uncertainty?

In this paper, using information contained in big (text) data, we provide a comprehensive examination of predictability in oil markets. With a size of around $3.5 trillion in annual consumption, this vast market is essential for economies to function properly and is crucial for assessing macroeconomics risks. The oil market is also notoriously volatile, making timely information and robust forecasting particularly important. We show that real-time news data complements and extends traditional energy-market predictors, and is powerful for forecasting energy market outcomes in- and out-of-sample. We further show that the economic magnitude of out-of-sample forecastability of oil futures returns using text-based measures is large.

Our interest in news-based measures reflects their ability to quickly adjust to new economic and market developments. News-based measures can capture dimensions of oil markets that traditional variables cannot. In oil markets, where extreme movements are common and where the full spectrum of geopolitical events can impact outcomes, such novel data sets, which compress the information content of thousands of news articles into a parsimonious set of real-time indicators, can improve the ability of practitioners and policymakers to understand and respond to events in a timely manner.

We construct a set of novel natural language processing (NLP) measures derived from the analysis of a corpus of energy-relevant articles from Thomson Reuters. Recent work has shown the usefulness of text measures for forecasting the returns and risks of individual stocks and stock indexes, and we find these techniques have value for oil forecasting. While some commonly used predictors of commodity returns, such as industrial production or economic activity indexes, are observed monthly or quarterly and become available with delays, text measures capture a wide range of energy market developments in real-time. Given the volume of news coverage of the energy sector, application of NLP tools in this space is particularly promising. Indeed, we show that our textual measures – obtained utilizing an energy word list that we manually constructed – can algorithmically identify important historical episodes in energy markets, in a way that traditional energy variables are unable to. Our NLP measures include topic-specific frequency and sentiment derived from energy news, and a measure of the unusualness (or entropy) of oil news. We identify seven distinct energy news topics, which are obtained using a network modularity approach, as in Calomiris and Mamaysky (2019a).

We aim to provide a comprehensive and transparent examination of the empirical performance of financial and physical oil market predictability.[1] We introduce new predictive measures derived from energy news articles, but we also consider a broad range of traditional predictors suggested by academic research. Our focus is on forecasting four- and eight-week ahead oil futures returns, oil spot returns, changes in the realized volatility of oil futures returns, the equity returns of three major oil companies, and changes in U.S. oil inventories and U.S. oil production for the period 1998-2020.[2,3] These dependent variables represent crucial information

[1] While forecasting the real price of crude oil and examining supply and demand forces driving the movements in oil prices are central questions of interest in the oil-macro space, they are not the focus of this paper. The models, the time horizon of the analysis, and the variables considered in the oil-macro literature are generally quite different from ours. For some leading contributions, see Alquist, Kilian and Vigfusson (2013), Baumeister and Kilian (2015), Manescu and van Robays (2016), and Baumeister et al. (2022).

[2] We focus on the oil market partly because natural gas is much more localized while oil trades on a world market.

[3] Energy companies' stock returns have not generally been included in studies of energy market forecasting. However, as forward-looking measures of the prospects of energy companies, we expect they contain important information about returns and risks in the energy market.

for investors, policymakers, and analysts, as they seek to understand the dynamics of oil markets. We construct a fully transparent empirical methodology for considering a comprehensive list of potential forecasting variables and investigating their usefulness both in- and out-of-sample. We also study the stability of estimated coefficients over time for in-sample analysis and the persistence of successful predictors for out-of-sample forecasting.

To control for predictors identified in prior studies, the set of explanatory variables we consider is large, including macroeconomic and financial indicators as well as measures that capture time-varying oil returns risk. One of our contributions is to reproduce many predictors used in the prior literature and analyze all of them in a unified framework. Our forecasting variables include lags of our dependent variables, the VIX (an index of short-term implied volatility of S&P 500 options), the yield on the ten-year Treasury note, the trade-weighted value of the dollar, and S&P 500 returns. We include a global measure of industrial production due to Baumeister and Hamilton (2019). We measure the commodity basis of oil futures using the methodology of Hong and Yogo (2012). Following Asness, Moskowitz, and Pedersen (2013), we construct a price-based measure of relative valuation for oil prices. We include several commodity-specific forecasting variables, including momentum, introduced in Boons and Prado (2019) and in Szymanowska et al. (2014). All these explanatory variables are described in detail in Section 2.

Several features distinguish our empirical approach from past work: we begin with a comprehensive list of forecasting variables and employ formal model selection techniques; our methodology for selecting variables is explicit; we use a bootstrap to adjust R-squareds and standard errors for overlapping observations, our variable selection methodology, and other small sample biases; and we consider out-of-sample validation of our models.[4] Our approach avoids reporting biases that are likely to arise when constructing forecasting models by selectively employing only a subset of potential forecasters, and is transparent about the out-of-sample performance of the forecasting variables.

Borrowing from the machine learning literature, we employ a forward selection model capable of selecting parsimonious time series forecasting specifications from the entire list of potential predictors. The forward selection approach accomplishes this via successively choosing each new variable as the one with the greatest incremental contribution to the model R-squared. Hastie, Tibshirani, and Tibshirani (2017) compare forward selection against two other machine learning approaches: best subset selection and the least-absolute shrinkage and selection operator (lasso) regression. They find that forward selection is competitive with the other two methods. But as we next explain, forward selection is particularly useful in the present context.

Our bootstrap methodology produces reliable measures of standard errors by accounting for overlapping observations, the effects of forward selection, and the Stambaugh (1999) bias in predictive regressions. An important reason for the use of forward selection (as opposed to other machine learning approaches) is that the bootstrap yields a distribution for the coefficient on the n^{th} variable chosen out of many (e.g., the distribution for the first variable chosen differs meaningfully from the distribution of the seventh variable chosen under the null of no predictability), allowing us to adjust standard errors accordingly. This approach is informative about the pitfalls of trying many regressors and choosing the best one or two without explicitly accounting for the selection criterion. In fact, we can exactly quantify the bias this approach entails by combining forward selection with bootstrapped standard errors for the n^{th} chosen

[4] Foster et al. (1997) propose techniques for assessing R-squareds of asset pricing regressions when researchers select the best *k* of *m* regressors to use in a forecasting model. Our approach relies on a bootstrap methodology.

variable. Although this point has been made before, for example in Welch and Goyal (2008), it is particularly salient for energy forecasting given the necessary reliance on time-series data.

After controlling for the above issues, we find evidence of robust in-sample predictability for the full sample period. Industry-relevant news corpora contain information that is not present in traditional industry metrics. That our text variables are successful predictors of energy market outcomes, even after controlling for the multitude of traditional forecasting variables, is a striking finding. However, we take heed of past warnings about overreliance on in-sample results (e.g., Welch and Goyal 2008; Harvey, Liu, and Zhu 2016), and consider robustness from two perspectives. First, in-sample stability across subperiods and then out-of-sample performance.

Subperiods for our in-sample analysis were specified in advance of running any regressions to avoid concerns about data mining. We first identified the subperiod that includes the 2007-2009 financial crisis by using the NBER's business cycle dating. Then, we divided the post-crisis subsample into four periods of equal length. Finally, for the pre-crisis subsample, we used the same post-crisis subperiod length of 2.66 years to define pre-crisis subperiods, where the initial subperiod length is the residual (i.e., it is slightly shorter than other periods). Surprisingly, we find that only a few regressors are chosen in the forward-selection model across multiple subperiods, suggesting a fair amount of model instability.

Then, we consider out-of-sample testing. In a time-series context such as ours, parsimonious models are attractive as panel models are not generally appropriate for oil forecasting due to its globally integrated market. So, we adopt a parsimonious modeling discipline by selecting a small number of potential forecasting variables based on rolling OLS tests of forecasting ability. Our out-of-sample testing methodology compares a forecast that blends the output of rolling lasso models, which employ a small number of our OLS-selected forecasting variables, with the value of a rolling mean forecast. We show that the blended forecast using our text measures meaningfully outperforms the rolling mean benchmark for forecasting oil futures returns and changes in oil spot prices, while traditional measures are better than text-based measures at forecasting changes in oil volatility. Using an analysis similar to Campbell and Thompson (2008), we show that the economic impact of this increased forecasting power is indeed large.

A. Related Literature and Our Contribution

Robust out-of-sample performance by forecasting models in finance is, in general, hard to come by. Welch and Goyal (2008) investigate in-sample and out-of-sample performance of equity premium predictions with variables from earlier academic research. They find that models have predicted poorly over their 30-year sample and argue that the historical average excess stock return is a better forecaster of future excess stock returns than out-of-sample regression-based estimates. This motivates our choice of a rolling mean as the benchmark forecast for our out-of-sample tests. Campbell and Thompson (2008), on the other hand, show that simple restrictions – such as having the theoretically predicted sign – on predictive regressions improve out-of-sample performance of key forecasting variables. We contribute to this literature by systematically and transparently investigating in-sample and out-of-sample performance of novel NLP measures, as well as many traditional forecasting variables from previous studies, in predicting several financial and physical oil market outcomes.

A closely related literature in finance investigates predictability in oil and other commodity markets. Examples include Bessembinder and Chan (1992), De Roon, Nijman and Veld (2000), Hong and Yogo (2012), Gorton et al. (2013), and Yang (2013). These studies provide evidence that returns in commodity futures markets can be predicted using a range of aggregate and commodity-specific financial and macroeconomic variables. These studies are typically based on in-sample analysis of a novel forecasting variable with baseline models that contain six or seven predictors. Our approach not only considers new text measures as well as a wide range of financial and macro variables, including many variables considered in this literature, but we also test predictability in oil markets both in sample and out of sample. In addition, we carefully control for small-sample biases.[5]

Identifying relevant news and how it is associated with changes in market returns and risks is a central topic in asset pricing. Recently, economists have analyzed the language that appears in newspaper articles and other textual sources, and applied this analysis to equity, credit, exchange rate markets and volatility (for example, Tetlock 2007; Tetlock, Saar-Tsechansky, and Macskassy, 2008; Calomiris and Mamaysky 2019a, 2019b; Glasserman and Mamaysky 2019; Mamaysky, Shen and Wu 2021). Recent work has applied textual analysis to oil markets. Loughran et al. (2019) develop a list of oil-specific keywords to measure energy article directionality and show that oil prices overreact to news over a one-day horizon. Brandt and Gao (2019) use measures taken directly from RavenPack, a vendor of news analytics, and find that while geopolitical and macro-related news both impact oil prices contemporaneously, only macro-related news forecasts future oil returns over the subsequent one to three months. Li, Shang, and Wang (2019) show that energy news and financial market variables can predict oil prices using machine learning techniques applied to a very limited number (6,756) of news headlines and a small number of financial market explanatory variables. Datta and Dias (2020) characterize oil news as being supply- or demand-related and use a structural VAR to show that

[5] A related recent paper by Conlon et al. (2022) takes a skeptical view of oil return predictability, highlighting the importance of the choice of sampling methodology for returns measurement.

indexes constructed to reflect this information forecast future oil supply and demand. Because of the VAR setting, the number of control variables they used is limited.[6]

Our paper differs in several ways. We perform our own NLP analysis, rather than using processed text data from a vendor. Our text analytics are transparent and straightforward for others to replicate. Our set of text measures, including entropy, topical frequency and sentiment for seven topics, is extensive. We analyze predictability of multiple energy market outcomes, not just oil spot returns. Given the large number of text and non-text forecasting variables used for this purpose, we focus on model selection and inference in our analysis. We also devote a good deal of attention to model stability and out-of-sample performance. Our work thus complements and extends the evolving literature in this space.

Our paper is also closely related to a burgeoning literature that applies machine learning (ML) techniques to forecast equity and corporate bond returns (Gu, Kelly, and Xiu 2020, Kozak, Nagel, and Santosh 2020, Feng, Giglio, Xiu 2020, Giglio, Liao, and Xiu 2021 for the cross-section of equities; and Bali et al. 2021 for corporate bonds). The main difference between our setting and that of the existing ML papers is that our forecasting exercise involves individual time-series (since we forecast each of our eight dependent variables one at a time) and not panel data (as is the case in all the above papers). This limits the available toolset since techniques like partial least squares or neural networks, that involve many parameters needing to be estimated, are subject to overfitting in the absence of panel data (or of a sufficiently long and stable time series). Our forward selection methodology is ideally suited to a time-series, model selection context, because each step only requires the estimation of a small number of parameters.

[6] Cavallo and Wu (2012) is a related paper that uses a VAR approach applied to an energy news series that is hand-collected, and thus difficult to update and maintain. Plante (2019) shows that the counts of articles mentioning "OPEC" can forecast oil volatility.

The paper proceeds as follows. Section 2 lists our non-text forecasting variables and describes our data sources and methodology for variable constructing. Section 3 presents our text analysis and new NLP measures. Section 4 contains our in-sample analysis and discusses our methodology for correcting standard errors and R-squareds for variable selection bias and for overlapping observations. Section 5 presents our out-of-sample analysis. Section 6 concludes and discusses directions for future work. The code for constructing our text-based variables, as well as our replication of traditional forecasting variables, is publicly available on GitHub.

2. Data and Construction of Variables

We consider a comprehensive set of traditional and recently developed predictors that capture returns and risks in the economy and the oil market, as well as new predictors constructed using Reuters news articles relevant for the energy sector. The raw data used to construct our variables come from Bloomberg, the Bureau of Labor Statistics, the Commodity Futures Trading Commission, the Energy Information Administration (EIA), the Wall Street Journal (our source for the VIX index), and the Federal Reserve Board.

Construction of many of the variables we consider is not straightforward, in part because the variables rely on different data sources, and in part because of timing issues that we discuss below. The time frame of our analysis is from April 1998 – March 2020. We forecast all dependent variables on an eight-week ahead basis, using weekly observations.[7] We also tried forecasting dependent variables four weeks ahead and found the results to be qualitatively similar. The four-week ahead results can be found in the Online Appendix.

[7] One of our dependent variables is realized volatility, which is measured over the prior 30 trading days. We cannot include this variable in our four-week ahead regression because it would overlap with our explanatory variables. We chose eight weeks to ensure no overlap between dependent and forecasting variables.

A. Timing Conventions

Our eight dependent variables are eight-week ahead oil spot and future returns, the stock returns of BP, Royal Dutch Shell, and ExxonMobil, changes in realized oil volatility, and changes in U.S. oil production and in U.S. oil inventories.

Oil spot and futures prices are available at 2:30pm Eastern time (ET) on each trading day and the oil majors' stock prices are available after 4pm ET. Our explanatory variables include lags of the dependent variables, and many financial and macro variables described below. We would like to use observations at the highest possible frequency to take full advantage of the links between information arrival and market reactions. Although oil and other market prices are available daily, oil production and inventory data are available at a weekly frequency in the U.S. We therefore perform our analysis using weekly observations.

U.S. crude oil production and crude oil inventories (including the strategic petroleum reserve) data are released by the EIA usually on Wednesdays at 10:30am ET. For some weeks, typically those involving holidays, releases are delayed by one or two days. This feature of the data drives the timing convention for our empirical analysis in order to ensure that oil inventory and production data do not overlap with our dependent variables. When forecasting changes in inventories and production, which we refer to as the *physical regressions*, our dependent variables become available after 10:30am ET on Wednesdays, and occasionally later. For our *price-based regressions* (futures and spot returns, changes in realized oil volatility, and the oil majors' stock returns), for which the inventory and production data will serve as predictors, we take weekly observations for the dependent variables on Fridays, which prevents overlap of dependent and explanatory variables even in weeks when the EIA data releases are delayed.

Our physical regressions have right-hand side variables measured as of Tuesday of week t, and our price-based regressions have right-hand side variables from either Thursday or Friday of week t, whichever timing ensures no overlap with the Friday 2:30pm ET oil market close. Variables can appear either as independent or dependent variables in either the physical or price-based regressions. Each of these four use cases has its own timing convention (i.e., day of week on which the variable is measured) and its own method for dealing with missing observations, which are detailed in the Online Appendix.

A. Dependent Variables

For obtaining oil returns, we use the U.S. oil benchmark, West Texas Intermediate (WTI). We calculate j-week spot price returns as ln (P_{t+j}/P_t), where $j = 4, 8$ weeks. We use the nearest-to-maturity futures price as the spot price, consistent with most studies of commodity futures, as commodity spot markets are frequently illiquid. Spot price changes, while interesting, do not represent an investable return because they ignore storage and transportation costs. To capture investable oil returns, we measure realized returns from investing each week in the front-month oil futures contract. On weeks that the front month future expires, we measure returns using an investment in the second month oil future (which will become the front month week's end). We obtain oil futures returns by constructing j-week cumulative returns as the product of the past j one-week returns. This measure captures the returns to a feasible investment strategy, and reflects changes in spot prices, realizations of risk premia, and changes in risk premia over time.[8]

Energy company stock returns are calculated as j-week percent changes in stock prices. We consider three large multinational oil and gas companies' stock returns: BP, Royal Dutch

[8] Further details on futures returns calculations are available in the Online Appendix.

Shell, and ExxonMobil. For BP, we use the ADR price from the New York Stock Exchange (NYSE); for ExxonMobil we use its NYSE stock price; and for Royal Dutch Shell we use prices from Euronext. Our measure of oil price volatility is the eight-week change in the trailing 30 trading-day realized volatility of WTI prices from Bloomberg. For our physical forecasting regressions, the variables of interest are eight-week ahead changes in oil production and in oil inventories.

B. Forecasting Variables

Our forecasting variables include lags of the four-week versions of our dependent variables. This makes the predictor set for the eight- and four-week ahead regressions identical so that the distinction between the two specifications only involves changing the dependent variable. There are two exceptions. First, rather than using lagged stock returns of the oil majors over the previous four weeks as forecasting variables, we calculate an average of their returns, which we refer to as *StkIdx*. This energy stock index is a less noisy forecaster than individual company returns, and we were unable to find an existing energy stock index with as long a history as *StkIdx*. Our results are qualitatively similar when using the individual stock returns as forecasting variables instead of *StkIdx*. Second, in addition to using the change in realized oil volatility from week *t-4* to week *t as a forecaster*, we also use the week *t* trailing 30-trading day oil volatility to account for mean reversion of future realized oil volatility.

We include an exhaustive set of forecasting variables that have been used in the literature to predict commodity returns. These predictors include the VIX, the yield on the ten-year Treasury notes, the trade-weighted value of the dollar (*DFX*), and S&P 500 returns. We also use the month-over-month growth rate of world industrial production (WIPI) introduced by Baumeister and Hamilton (2019) as a measure of global economic activity. As is common in commodity forecasting, we use a basis measure given by the annualized ratio of the 3-month to 1-month price for crude oil futures, namely $basis_t = (F3_t/F1_t)^6 - 1$ (raising to the power of 6 converts this to an annualized measure). A positive (negative) basis indicates the curve in contango (backwardation), and all other things being equal buying longer-dated futures will lose (earn) money as they roll down the curve. Following Asness, Moskowitz, and Pedersen (2013), we calculate a "book-to-market" ratio for oil prices, *BE/ME*, defined as the average WTI spot price from 4.5 to 5.5 years ago divided by the recent spot price. Momentum, *Mom*, in month *m* is measured as the past cumulative return on WTI front-month futures from *m-11* to *m-1*, i.e. $Mom_m = 100 * (\prod_{s=m-11}^{m-1}(1 + R_s) - 1)$, which is the standard timing convention. Month *m* momentum is then used as a forecasting variable for future four- or eight-week outcomes that start in month *m+1*.

We also consider Boons and Prado's (2019) basis-momentum predictor, *BasMom*, defined as the difference between momentum in a first- and second-nearby futures investment strategy (i.e., constantly rolling each future to the next maturity prior to expiry), where both are measured as the past 12-month cumulative return. Finally, following Szymanowska et al. (2014), we include: inflation beta, *InflaBeta*; dollar beta, *DolBeta*; hedging pressure, *HedgPres*; open interest, *OpenInt*; and liquidity, *liquidity*. For *InflaBet* and *DolBeta*, we use the coefficients from 60-month rolling regressions of monthly WTI futures returns on unexpected inflation and on changes in the log dollar index, respectively..[9] For *HedgPres*, we use the difference between the number of short and long hedging positions by large traders in the crude oil market divided by the total number of hedging positions. The hedging pressure data are released to the market on Friday of week *t* at 3:30pm ET and reflect positioning as of the end of Tuesday of week *t*. We

[9] We use Bloomberg's dollar spot index (DXY) because, of the dollar series we can access, DXY has the longest history.

use the week *t-1* value of *HedgPres* as our week *t* explanatory variable, to ensure no overlap between our week *t* dependent and explanatory variables. *OpenInt* is the total open interest in the crude oil futures markets, in dollar terms. We use Amihud, Mendelson, and Lauterbach (1997)'s *liquidity* measure defined as the log of the ratio of WTI futures trading volume to its absolute return calculated using the daily value of trading volume and daily return.

We refer to the variables from this section as our *baseline (or non-text)* measures. Table I presents definitions for all variables used in the empirical analysis in detail. Table II reports summary statistics for all variables used as either dependent or forecasting variables in the April 1998 – March 2020 sample. For example, the average eight-week return on oil futures has been 1.35% with a standard deviation of 13.78%. The average eight-week return of oil spot prices has been lower, at 0.64%, with a higher standard deviation of 14.75%. Energy company stocks, on the other hand, have lower average returns (ranging between -0.34% and +0.19%) and are less volatile (ranging between 7.61% and 10.01%). The four-week summary statistics look similar.

C. Risk Premium Measures

In addition to the traditional energy market and macro predictors, we include several measures that are useful for gauging market risk premia. The first of these, *vix_diff*, measures the difference between the VIX index and the last 30-day realized volatility of the S&P 500 index. Many researchers, for example Bekaert and Hoerova (2014), argue that the difference between the VIX index and forecasts of future realized volatility reflects the variance risk premium. Here we assume lagged realized volatility is a reasonable proxy for expected future volatility. Similarly, we include *ovx_diff*, which is the difference between the OVX index of implied volatility of an ETF which owns WTI futures and the last 30-day realized volatility of crude oil prices; *ovx_diff* is a proxy for the volatility risk premium in the oil markets.

In addition, we follow Hansen and Jagannathan (1991) and construct another measure of the risk-premium in energy markets. Letting R be an n-dimensional vector of daily gross returns from a candidate set of securities, the unconditional version of the basic no-arbitrage condition of asset pricing is $1 = E[mR]$ (note 1 is an n-dimensional vector), where *m* is the stochastic discount factor (SDF).[10] Assuming *m* is in the linear span of the security returns implies

$$m_t = 1^{\top} E[RR^{\top}]^{-\top} R_t, \tag{1}$$

where $E[RR^{\top}]$ is the unconditional expectation of the $n \times n$ matrix $R_t R_t^{\top}$ for all *t* in the population. Note that this SDF representation does not assume anything about the underlying factor structure of returns. Furthermore, it is well known that the expected excess return on a security is proportional to the negative of its covariance with the SDF (Cochrane 2005). The conditional version of this relationship can be written as

$$E_t R_{i,t+1}^e = -\frac{cov_t\left(m_{t+1}, R_{i,t+1}^e\right)}{E_t m_{t+1}}, \tag{2}$$

where R^e is the daily excess return on security *i* and the expectations are taken as of day *t*. We estimate $E[RR^{\top}]$ in (1) in windows (see below) of our data using daily returns on the Credit-Suisse WTI futures total return index, the total return of the S&P 500 index, a U.S. Treasury total return index from Bloomberg (which roughly tracks 10-year bonds), the total return from investing in 6-month U.S. T-bills, and the total return of the MSCI World Energy Sector index. Then using the estimated SDF $\widehat{m}_t$, we approximate the week *t* conditional expectation in (2) by calculating the covariance between the excess return of the WTI futures index and $\widehat{m}_t$ over the prior 252 trading days, as well as the 252-day mean of $\widehat{m}_t$. We use these two sample moments in (2) to derive an estimate of the conditional WTI risk premium.

[10] This follows by taking the unconditional expectation of the basic no-arbitrage relationship $1 = E_t[m_{t+1} R_{t+1}]$ and applying the law of iterated expectations.

We use three different estimation methods for $\widehat{m}_t$, which differ in their estimates of $E[RR^\top]^{-\top}$. In all three cases, we use the $\widehat{E}_t R^e_{WTI,t+1}$ estimate from the window ending on day *t* as the time *t* estimate of the WTI risk premium. In the first variant, we use a rolling 756-day window (roughly three years) to estimate $E[RR^\top]^{-\top}$. We refer to this series as *sdf_rolling*. In another variant, we use an expanding window that starts at a minimum of 756 days, and then expands for each successive day in the sample. We refer to the WTI risk premium estimate from this approach as *sdf_growing*. Both the rolling and growing SDF is used in our out-of-sample analysis. In our in-sample analysis, we use the SDF constructed with the full-sample estimate of $E[RR^\top]^{-\top}$, which we label *sdf_fullSample*. All calculations are done in windows that end on Tuesdays for the physical regressions and on Thursdays for the price-based regressions.

3. Text Analytics

In order to construct the text measures to be used to forecast energy market outcomes, we apply a broad range of modern NLP techniques, which represent the current state of the art in the use of text analytics for economic forecasting.[11] Our corpus for NLP analysis includes all 2.07 million articles in Reuters that are energy-relevant from January 1996 to March 2020. The text series used in our analysis start in April 1998 because of the lag needed to calculate entropy, as explained below. An article is *energy-relevant* if it is classified by Reuters as belonging to one of 98 energy topics, the full list of which is in the Online Appendix.[12]

To perform topical analysis, we first constructed an energy words list. In doing so, we identified a variety of sources that provided comprehensive coverage of the energy sector, given

[11] The application of NLP to finance and economics is an active research area, and there are new methodologies being developed that may prove useful in the future, e.g., Ke, Kelly, and Xiu 2019; Garcia, D., X. Hu, and M. Rohrer, 2020; Glasserman et al. 2020. We hope to explore these experimental approaches in future work.

[12] This corpus includes not only energy-specific articles, but also broader macro articles that the Reuters editorial staff determines to be relevant for energy markets.

that there is no oil or energy markets textbook that is widely used by scholars or energy analysts. Our list of sources includes popular press books in oil and energy, such as Yergin (1992), or more technical energy and commodities textbooks, such as Dahl (2004) or Geman (2005), as well as industry glossaries.[13] We combined index lists and glossaries of these sources and chose energy markets related words, two-word phrases (bigrams) and three-word phrases (trigrams). We eliminated words or phrases that seemed too technical or not meaningful from a news coverage point of view. This initial version of the energy words list included 1,931 words and phrases. In the second stage of our manual process, we examined the words and phrases one by one, and selected the ones we believed to be more likely to appear in a news article. This stage yielded a list of 685 words or phrases (tokens) including abbreviations. After dropping tokens that never appeared in our Reuters energy news articles corpus, we were left with a list of 387 words and phrases. We began our textual analysis process with that list.

We then constructed a 387×387 token co-occurrence matrix which measures the cosine similarity between this initial list of tokens. The cosine similarity of tokens *i* and *j* is a number between 0 and 1, given by $\frac{w_i^\top w_j}{\|w_i\|\|w_j\|}$ where w_i is the vector measuring the number of times token *i* appears in each of the documents in our Reuters corpus (the length of w_i equals the number of documents in the corpus), and $\|w\|$ is the Euclidean norm of w. A cosine similarity of 1 means tokens *i* and *j* always appear in documents together, and at the same relative frequency, while a cosine similarity of 0 means tokens *i* and *j* never appear together in any document.

Next, we identify disjoint (i.e. non-overlapping) word groups that maximize the *modularity* of the network represented by the token co-occurrence matrix. These word groups

[13] Thanks to Mine K. Yucel and David Rodziewicz for their source suggestions (our full source list is in the Online Appendix).

represent energy topics in the Reuters news archive. Network modularity, introduced by Newman and Girvan (2004), measures the degree to which members of communities or groups in a network are connected to one another above what would be expected by chance. For example, say there is a group of 20 people, 12 of whom are all connected to one another on social media, and 8 of whom are connected to one another, but none of the group of 12 or the group of 8 are connected to a member of the other group. A partition with a community consisting of the 12 connected people and another community consisting of the 8 connected people would have the highest possible modularity across all possible network configurations, because it is highly unlikely that purely by chance no one in either community would be connected to anyone from the other one. In general, finding the network partition with the highest modularity is an NP-complete problem (Brandes et al. 2006), and therefore solution methods for finding high modularity network configurations are heuristic in nature. The algorithm we use is known as the Louvain method, described in Blondel et al. (2008). It generates a high modularity network partition while endogenously determining the optimal number of communities, has efficient run time, and has been shown to perform well in many different settings.[14] In our application of the Louvain algorithm, we set the diagonal of the co-occurrence matrix to zero, meaning an individual is not considered to be connected to itself. This yields eight word groups, or topics, from the Louvain algorithm. The eighth topic contained only several tokens, so we reallocated these tokens from the eighth topic to the other seven topics to maximize the resultant seven-topic partition's modularity.

[14] The algorithm initially assigns every member of the network to its own community. Starting with an arbitrary individual and cycling through all remaining individuals, the algorithm attempts to move the individual currently under consideration to another community to increase the modularity of the resultant partition. A community left with no individuals is deleted. The algorithm ends when no reallocation increases modularity. The remaining communities endogenously determine the optimal community number. The algorithm works well in practice, but has no optimality guarantee.

Once we identified the initial set of seven topics, we calculated the average co-occurrence of a large set of additional candidate energy related words, bigrams and trigrams, beyond the 387 in our original list. We then identified from the list of candidate energy words those whose maximum topical co-occurrence was very high relative to its average topical co-occurrence. For example, the candidate token *shell*, which was not part of our original 387-token list, had an average cosine similarity with the existing tokens in topic 1 of 0.2076, whereas its average co-occurrence across all seven topics was 0.0374. The resultant difference of 0.1702 was the second highest of all our candidate tokens. We therefore included *shell* in our augmented token list. The intuition behind this procedure is that we want to exclude words that have high co-occurrence with *all* our topical clusters because these tend to be generic words such as *said* or *though*. However, words that have a high co-occurrence with a single topic tend to be energy-relevant words, bi- or trigrams, that are related to the topic in question. Applying this process to a large set of candidates yielded an additional 54 tokens, which we then placed into one of the existing seven topical groups to maximize the network modularity of the new, 441-token network. We refer to these 441 tokens as the *energy words*.

Figure 1 displays word clouds for each of our seven topics. The size of each token in the cloud corresponds to its relative frequency in the corpus. We label the topical categories based on our interpretation of the common semantic link of the words that appear in each of these word clouds. Interestingly, the topics represented by the word clouds have readily interpretable meaning and exhibit sufficient variation over time to be useful in our analysis. We label the topics as follows: company (*Co*), global oil market (*Gom*), environment (*Env*), energy/power generation (*Epg*), crude oil physical (*Bbl*), refining and petrochemicals (*Rpc*), and exploration and production (*Ep*). As a robustness check, we verified that latent Dirichlet allocation (LDA) due to Blei, Ng, and Jordan (2003) – another popular topic-modeling approach – produced similar topics to the Louvain-based ones.[15]

We then classify article *i* into topical category τ by looking at the fraction of the energy words appearing in this article that belong to topic τ, or

$$f_{i,\tau} = \frac{N_{i,\tau}}{\sum_{j=1}^{7} N_{j,\tau}},$$

where $N_{i,\tau}$ is the number of energy words in article *i* that belong to topic τ. Notice the article topic weights sum to one. The sentiment of article *i* is defined using the Loughran and McDonald (2011) sentiment dictionary as

$$s_i = \frac{Pos_i - Neg_i}{Total_i}.$$

Here Pos_i, Neg_i, and $Total_i$ are the number of positive, negative and total words in article *i* after stop words have been removed. Following Calomiris and Mamaysky (2019a), we define an article's topic sentiment as the product of topic frequency and sentiment, or

$$s_{i,\tau} = f_{i,\tau} \times s_i.$$

Given that article frequencies sum to one, topical sentiments $s_{i,\tau}$ sum up to article sentiment s_i.

Unusualness is defined using the *entropy* concept introduced in Glasserman and Mamaysky (2019) and used Calomiris and Mamaysky (2019a).[16] Specifically, we define article *i*'s unusualness as the negative average log probability of all 4-grams appearing in that article, or

$$e_i \equiv - \sum_{\substack{j \in 4-grams \\ in\ the\ article}} p_j \times \log \widehat{m}_j,$$

[15] We used a 7-topic LDA model across multiple LDA trials to compare against our word topics. The number of topics in LDA has to be exogenously specified, whereas the Louvain method endogenously determines the topic number. A summary of this analysis is in the Online Appendix.

[16] Glasserman and Mamaysky (2019a) showed that entropy can be used to measure the novelty of an article, and that higher entropy news flow is more informative for forecasting future market outcomes.

where p_j is the fraction of all 4-grams represented by the jth 4-gram in article i, and $\widehat{m}_j$ is the empirical probability of the fourth word in the 4-gram conditional on the first three, estimated over a training corpus using all articles from months $t-27$ to $t-4$.[17] For the jth 4-gram $w_1w_2w_3w_4$ (the w_k's refer to words or tokens), $\widehat{m}_j$ is the fraction of times w_4 follows the word sequence $w_1w_2w_3$ in the training corpus, or

$$\widehat{m}_j = \frac{\hat{c}(w_1w_2w_3w_4)}{\hat{c}(w_1w_2w_3)},$$

where $\hat{c}(\cdot)$ is the count operator. When a 4-gram has not been seen in the training corpus, we assign to it a probability of 0.1.[18]

We aggregate our article-level news measures to the daily level by taking a word-weighted average of all articles released between 2:30pm ET of the prior business day and 2:30pm ET of the present business day. For Mondays, we count articles from 2:30pm ET to midnight on Friday, in addition to articles from 2:30pm ET on Sunday to 2:30pm ET on Monday. We then take an equal-weighted average of the daily news flow measures (topical frequency, topical sentiment, and entropy) ending on Tuesday or Friday of week *t* for the physical and price-based dependent regressions, respectively. We also calculate the average number of daily articles about energy markets in the Reuters archive in weeks ending at 2:30pm ET on Tuesday or Fridays. This yields 16 distinct text-based series: article count, entropy, the seven topical frequency series (labeled *f[Topic]*), and the seven topical sentiment series (*s[Topic]*). We standardize all text series, except entropy, to have mean zero and unit variance. In our regressions, we use four-week rolling averages of all weekly standardized text series.

[17] We do not use months $t-3, t-2, t-1$ in the entropy calculation, because this allows newly emergent words and phrases to remain unusual, i.e. have high $\widehat{m}_j$'s, for several months after their first appearance.

[18] The method is not sensitive to the choice of 0.1. For the entropy analysis, we tokenize and stem the documents, but do not remove stop words. For more details of this methodology, see Glasserman and Mamaysky (2019) and Calomiris and Mamaysky (2019a).

In addition to these, we calculate four additional measures of aggregate news flow: the first principal components (PCAs) of the seven topical frequency series (*PCAfreq*), of the seven topical sentiment series (*PCAsent*), and of all fourteen series together (*PCAall*) as well as the sum of the seven topical sentiment series in a given week (*sent*). The PCAs are calculated using the four-week averages of the weekly series, where the four-week averages have been normalized to mean zero and unit variance. The aggregate measure *sent* is only used in the out-of-sample analysis. We thus allow for the possibility that the common variation across all text series is the key driver of forecasting performance.

A. Validation: Qualitative Analysis of Energy News

We plot four-week averages of the nineteen text-based series in Figure 2. As is clear from the figure, the text-based measures of news flow in energy markets display a large amount of time variation. To gain further insights into our measures of energy news flow, we explore whether unusual movements in our text measures correspond to important real-world events in energy markets. To identify potentially interesting events, we look at four-week averages of our seven topical sentiment series, and then select the two most negative changes in the four-week average series for each topic. For each of these negative topical sentiment episodes, we then identify a set of candidate articles. Candidate articles are those that have entropy scores equal to or higher than 2, that contain 100 or more words after stop words are removed, and that have a topic allocation above 0.8. i.e., $f_{i,\tau} > 0.8$. These articles typically contain stories about specific energy market developments, and are not news alerts, daily summaries, or statistical tables. We then manually looked through the headlines and connected them to specific energy market episodes. We find that almost all extreme moves in topical sentiment were associated with important events in energy markets, and in this validation exercise we chose to focus on six in particular, each of which belongs to a distinct topic. The end dates of the four-week topical sentiment changes associated with these six episodes are marked with stars in Panels A and B of Figure 2.

While the events were identified based on changes in topical sentiment (Panel B), it is clear from Panel A that all of these events are also associated with large increases in the fraction of total news coverage devoted to that particular topic category. This points to a more general feature of the topical sentiment and frequency series, namely that for each topic the two aggregate series are very negatively correlated (the correlations range from -0.57 to -0.93). Spikes in topical frequency tend to occur at times of negative topical sentiment.

Table III shows the six episodes we identified. For each episode, we show the sentiment, entropy, and headlines of the five most negative sentiment articles. The particular historical episodes associated with sharp drops in topical sentiment, with associated topic category in parentheses, are: the UK fuel protests in September of 2000 (company), the attempted Venezuelan coup in 2002 (global oil markets), the Volkswagen emissions scandal in 2015 (environment), the Enron bankruptcy hearings of 2002 (energy/power generation), Hurricane Katrina in 2005 (crude oil physical), and the BP oil spill in 2010 (exploration & production).

First, it is notable that each event is classified into the appropriate topic. For example, many articles discussing the UK fuel protests focused on their impact on business. Others discuss the reduction in OPEC output, caused by the civil unrest in Venezuela, as affecting global oil markets. Second, these events were identified algorithmically, and not cherry-picked by us. Third, since we assigned names to topics by looking only at the word clouds, the close match of headlines with their associated topic names is a validation of the usefulness of our methodology.

These results indicate that our news-based measures of energy markets capture important aspects of energy news with timeliness and specificity that non-text series cannot match. In

Sections 4 and 5, we exploit the information content of these news series for both in- and out-of-sample forecasting of our eight energy-market outcomes.

4. In-sample Predictability

We address two main questions in our in-sample analysis of predictability in energy markets: How well do our text measures work in the presence of non-text measures that were shown in past work to be powerful market forecasters? How stable are in-sample forecasting results across subperiods? The empirical challenge is to determine which subset of our text and non-text measures is most effective in forecasting energy market outcomes while dealing with the limited degrees of freedom inherent in time series analysis. We employ a forward selection model to choose a parsimonious time series forecasting specification from our broad list of potential forecasting variables. The forward selection approach accomplishes this via successively choosing each new variable as the one with the greatest contribution to the model R-squared, given the variables that have already been chosen.[19] We apply this methodology to all our dependent variables, and develop a reliable inference procedure that accounts for the selection criterion of our variables, as well as finite-sample issues inherent to our dataset.

As already discussed, Hastie, Tibshirani, and Tibshirani (2017) found that the forward selection method is competitive with other machine learning model selection techniques. Forward selection is particularly well suited to our application, because it allows us to determine which of a small subset of chosen variables is the first one selected, which is the second, and so on. For conducting inference for our in-sample analysis we use a bootstrapped distribution which takes into account the order in which a given explanatory variable is chosen. This analysis

[19] We use the `fs()` method from the R package *selectiveInference* to perform this analysis.

emphasizes the distortion introduced into model selection by choosing the best of many variables without explicitly accounting for that selection criterion.

Our 41 forecasting variables include lagged measures of our dependent variables (6), macro and energy market indicators and a variety of risk measures (16), and our new NLP measures including article count, entropy, the seven topical frequency series, the seven topical sentiment series, and the three PCAs (19).[20] Prior to running the in-sample forward selection procedure, we first detrend all dependent and independent variables, to ensure that trend does not contribute to finding spurious forecastability. We then residualize the data by regressing out the four-week version of the lagged dependent variable from both the left- and right-hand sides of the in-sample specification; the lagged dependent variables are measured using the explanatory variable timing conventions described in Section 2. We residualize the three oil major stock returns with the lags of their own returns, not with *StkIdx*. The residualization procedure is innocuous: we residualize only because the lagged dependent variable would otherwise be frequently chosen in the forward selection procedure. The residualization procedure is equivalent to forcing forward selection to always include the four-week version of the lagged dependent variable in all specifications. Our forecast horizon is either four- or eight-weeks ahead. We use forward selection to choose seven variables out of our set of 41, after all data have been detrended and residualized. We chose seven variables because the number of variables considered in past studies examining predictability in commodity markets ranges between one

[20] To be conservative, we use only one lag because we already have numerous forecasting variables. The 22 non-text variables: *FutRet*, *DSpot*, *DOilVol*, *OilVol*, *DInv*, *DProd*, *tnote_10y*, *DFX*, *sp500Ret*, *StkIdx*, *basis*, *WIPI*, *VIX*, *vix_diff*, *BE/ME*, *Mom*, *BasMom*, *DolBeta*, *InflaBeta*, *HedgPres*, *liquidity*, *OpenInt*. The 19 text variables: *artcount*, *entropy*, *sCo*, *fCo*, *sGom*, *fGom*, *sEnv*, *fEnv*, *sEpg*, *fEpg*, *sBbl*, *fBbl*, *sRpc*, *fRpc*, *sEp*, *fEp*, *PCAsent*, *PCAfreq*, *PCAall*. Note that *sdf_fullSample* and *ovx_diff* are excluded from this analysis because they are not available for the full sample. Each of the four-week lagged dependent variables can enter into the forecasting regressions for the other seven dependent variables, but not for itself because of our residualization procedure.

and seven (for example see Table 1 in Baumeister and Kilian (2017)); thus, our model can match the complexity of past models used in the literature.

The model is estimated using weekly observations with either four- or eight-week ahead overlapping observations, which increases the possibility of finding spurious forecasting relationships. It is well-known that overlapping observations will downwardly bias standard errors and upwardly bias R-squareds (see, for example, Hodrick 1992, Kirby 1997, Ang and Bekaert 2007, and Boudoukh et al. 2008). Furthermore, we employ forward selection for choosing a parsimonious set of in-sample regressors, which introduces upward bias in the R-squareds, and downward bias in the standard errors. Finally, our forecasting tests include oil spot or futures returns regressed on lagged independent variables related to oil prices, such as the book-to-market ratio, which raises concerns over the Stambaugh (1999) bias. To control for these sources of finite sample bias, we construct bootstrapped distributions for our t-statistics and R-squareds, a methodological contribution of our paper. We now describe our methodology.

A. Controlling for Overlapping Observations and Sample Selection

We assess the in-sample forecasting power of our model by simulating the data and checking whether the empirical R-squareds and t-statistics are anomalous relative to their simulated counterparts. We first estimate an AR(8) process for the dependent variable. We then simulate a new dependent series based on the AR(8) model. The simulated dependent series is constructed to have the same dependence on contemporaneous values of explanatory variables as observed in the data, hence our simulated null distribution accounts for the Stambaugh (1999) bias.[21] Next, we rerun our in-sample forward-selection and regression models, using all of the

[21] Time $t+h$ residuals of our simulated dependent variable use the same loadings on contemporaneous (time $t+h$) explanatory variables, i.e., future regressors, as the actual AR(8) residuals. R-squareds from regressing actual $t+h$ residuals on $t+h$ dependent variables are low (see Online Appendix), so the contribution of these loadings to the variance of the residual is small. Nevertheless, our simulations control for the empirical level of the Stambaugh bias.

actual 41 forecasting variables, except replacing the dependent variable with the simulated series. By construction, the simulated dependent series replicate the autoregressive properties of the actual dependent variables and have just enough correlation with the actual forecasting variables to account for the Stambaugh bias.[22] In one round of the simulation, we calculate the standard OLS t-statistics for the selected variables, keeping track of the order of selection, i.e. the t-statistic for the first selected variables, for the second selected variable, and so on. We also record the R-squared of this one simulation round. We then repeat this process 1,000 times to build a bootstrapped distribution for the ordered t-statistics, as well as for the model R-squared. This process controls for the selection, overlapping observation properties of our in-sample procedure. More details are in the Online Appendix.

Our *no-predictability* null hypothesis controls for the mechanical autoregressive properties of the dependent variable and allows for potential correlation of AR(8) residuals with future regressors to capture the Stambaugh bias. Even accounting for the latter, the simulated dependent series have very little correlation with contemporaneous regressors. To give a sense of the impact of small-sample biases, Figure 3 shows the bootstrapped R-squared distributions for forecasting eight-week ahead oil futures returns and changes in oil volatility. There is a wide range of R-squareds in our simulated runs, and small-sample biases can lead to very high in-sample R-squareds. When reporting our actual R-squareds in Table IV, we show the percentage of simulated R-squareds that are lower than the actual ones (in the table row labeled "CDF"). Rather than interpreting the outright value of the R-squared, a very high CDF value indicates that there is evidence of in-sample predictive ability even in the face of these biases.

[22] In an alternative version of the bootstrap, we simulate the dependent series to be independent of the forecasting variables (i.e., no adjustment for Stambaugh 1999). This yields very similar results to those that we report.

To understand the impact of small-sample biases on p-values, Figure 4 shows the distribution of the ordered t-statistics for the seven forward selected variables, under the null hypothesis of no predictability for forecasting oil futures returns and changes in oil volatility. The butterfly shaped distributions show the extreme bias that forward selection introduces to standard OLS t-statistics. The first chosen t-statistic (the widest bimodal distribution) shows that the modes for the t-statistic of the first selected variables are close to -6 and +6 respectively. The modes for the seventh selected variable are expectedly smaller in magnitude, at approximately -3 and +3. The figure clearly shows that the bias is more pronounced for the first variable chosen than for the second, is higher for the second relative to the third, and so on. The bootstrap applied to forward selection allows us to quantify these differences, whereas other machine learning techniques that choose the variable subset concurrently rather than sequentially would not reveal this pattern. When one sequentially chooses a subset of the best forecasting variables from a large set, their standard error distribution under the null has the butterfly pattern shown in Figure 4. Not adjusting for this introduces obvious biases.

We adjust for this issue by calculating p-values in our in-sample regressions by comparing the OLS t-statistics in our actual regressions to these distributions. Let $\hat{p}$ be the fraction of simulated t-statistics for a given ordered selected variable (e.g. the second selected variable in a given specification) that are less than the t-statistic for the actual ordered selected predictor. Our bootstrapped p-value is reported as $\min(\hat{p}, 1-\hat{p})$. A p-value less than or equal to 0.025 (0.05) indicates significance at the 5% (10%) level. We don’t present bootstrapped distributions of R-squareds and t-statistics for all dependent variables (they are available from the authors), but Table IV, discussed next, summarizes this information.

B. Results for the Full Sample

Table IV presents the eight-week ahead regression results for our 8 dependent variables using stepwise forward selection that chooses seven variables for each model for the full sample.[23] Only the predictors that were chosen by at least one model are presented in the table. For each dependent variable, we present coefficient estimates of the selected predictors, which are standardized, along with corresponding p-values as described in the last section. Our standardized coefficients report the standard deviation change in the dependent variable due to a one standard deviation change in the forecasting variable.[24]

The standardized coefficients for the selected predictors range between 0.05 and 0.67 in absolute value. For example, a one standard deviation increase in ten-year treasury note yields (*tnote_10y*) over the previous month lowers eight-week ahead BP returns by 0.13 standard deviations. Or, a one standard deviation increase in average *sGom* over the past month – positive sentiment about global oil markets – increases oil futures returns by 0.25 standard deviations over the next eight weeks. These are large economic effects. Moreover, around 50% (26/56) of the selected variables are statistically significant, even after adjusting for overlapping observations or variable selection.[25] In other words, the variables chosen by the forward selection method are generally both economically and statistically significant.

The actual adjusted R-squareds of the forecasting regressions are solid, ranging from 6% to 36%. At the bottom of Table IV, we present means of the bootstrapped adjusted R-squareds for each regression and their corresponding CDFs (the percentage of bootstrapped R-squareds

[23] The four-week horizon results in Online Appendix Table A.VIII are consistent with the eight-week results.

[24] These are $b \times sd(RHS)/sd(LHS)$ where b is the estimated coefficient, $sd(RHS)$ is the standard deviation of the forecasting variable, and $sd(LHS)$ is the standard deviation of the dependent variable, calculated for the set of dates available for each individual forecasting regression.

[25] The counts of variables are presented in Online Appendix Table A.IV.

that are lower than the empirical one). We conclude that overall the empirical R-squareds observed in our models are highly unlikely to have been generated by chance. That is, under our no-predictability null hypothesis, the probability of adjusted R-squareds being greater than or equal to the empirical R-squareds reported is less than or equal to 3.7% for all models except ExxonMobil returns and change in oil production (i.e., for six models out of eight).

Turning to the composition of the selected variables, out of the 56 predictors selected across all the models, 32 of them are text measures (about 57 percent), and of these 16 are statistically significant. Recall that 19 of the 41 candidate explanatory variables are text-based measures. For example, three of the text measures that are chosen statistically significantly at least two times are *PCAall, entropy,* and *sGom.* In fact, half of our dependent variables are forecastable statistically significantly by *entropy*. It turns out that although non-text variables represent 54 percent of all forecasting variables (22/41), they account for only about 43 percent of the selected predictors (24/56). In addition, only 42 percent of the selected non-text variables are statistically significant (10/24), which is lower than the 50 percent of the selected text variables that were significant (16/32). For example, only one of the selected non-text variables is chosen statistically significantly more than once: *Mom* for futures returns, oil spot returns and BP returns. These results suggest not only that our new text measures are selected frequently, but also that they are statistically significant more often than the non-text measures. Therefore, we conclude that our text measures are powerful in-sample predictors for the oil market even after controlling for many traditional forecasting measures. The use of modern NLP techniques produces a new set of economically and statistically significant forecasting variables for energy market outcomes.

We next examine whether the text measures are selected because they are proxies for risk. To address this question, we take the forward selection models considered above and presented in Table IV, and add a subset of our risk measures presented in Section 2.2, namely *VIX*, *vix_diff*, *ovx_diff*, and *sdf_fullSample*, one by one after the seven variables were selected by stepwise forward selection. As *VIX* and *vix_diff* were already included in the list of candidate variables for our forward selection procedure, they are included in this test only if they were not selected in the first place. Because *ovx_diff* (data start in May 2007) and *sdf_fullSample* (data start in January 2000) are not available for the full sample period, we do not include them in our core in-sample analysis and analyze them here instead. Risk measures are natural predictors of returns because time variation in expected returns may reflect forward-looking compensation for risk. Looking at how coefficients on text measures change, we find that overall adding the risk measures does not pull the coefficients on the text measures towards zero (Online Appendix Table A.VI shows these results). In other words, one should not interpret the selected text measures as proxies for some omitted risk factor. Interestingly, *sdf_fullSample* does not play a very important role in this in-sample risk analysis; though we will see that the other SDF variables (*sdf_rolling* and *sdf_growing)* often do play an important role in the out-of-sample analysis in the next section.[26]

To sum up, we find compelling evidence of in-sample predictability after carefully controlling for small-sample biases. While it is tempting to stop here and conclude that energy

[26] We also test the usefulness of our variables using a standard F-test. We test the null hypothesis that the coefficients of text, non-text, and all (text and non-text) variables are jointly zero. These three F-tests are done for our eight dependent variables, resulting in a total of 24 F-tests. The results are not bootstrapped. They are presented in the Online Appendix Table A.V and indicate that we can reject this null hypothesis in 22 out of 24 cases with more than 90% confidence (and usually with 99% confidence). Overall, text measures seem more useful than the larger set of non-text measures, and the combined (all) variables perform similar to the text variables.

market outcomes are forecastable, we explore whether the promising in-sample results are reliable from two perspectives: in-sample stability and out-of-sample performance (Section 5).

C. In-Sample Results by Subperiod

To explore stability across subperiods, we ran forward-selection models for each subperiod separately to see whether predicting variables that are chosen in one subperiod are also chosen in other subperiods. Our nine subperiods were chosen prior to running any analysis, and were thus chosen with no data mining involved. The results for subperiods are summarized in Table V.

For each dependent variable and each selected predictor, we report the number of subperiods when each predictor is chosen, distinguishing the number of subperiods where the estimated coefficient of the predictor is either positive or negative. There are few examples of forecasting variables that are chosen for many of the nine subperiods. In the *DOilVol* regression, *OilVol* (the lagged 30-day realized volatility) appears in all nine subperiods and is consistently negative, reflecting its role in forecasting mean reversion of oil return volatility. *OilVol* is also selected as a forecaster of *FutRet* and *DSpot* in three subperiods. The book-to-market ratio (*BE/ME*) appears as a consistently positive forecaster for *FutRet* and *DSpot* in six and four of the nine subperiods, respectively. It is also chosen as a positive predictor in four subperiods for ExxonMobil and BP returns. *HedgPres* is selected as a positive forecaster of *FutRet* in four subperiods. *fGom* is selected as a negative forecaster of *bpRet* in four subperiods. *fEpg* is selected as a negative forecaster of *DInv* in five subperiods. The *VIX* is selected with a positive sign either three or four times for the three oil companies' returns. With these few exceptions, the other variables are not selected with a consistent sign for more than three subperiods. Indeed, instances of three or more subperiods with a consistent sign (shown in bold face in Table V) do not occur very often. For the most part, variables are selected either for only one or two subperiods out of nine.

The instability of forward-selection modeling across subperiods suggests that, in general, forecasting variables estimated in one subperiod may not be reliable forecasters of the subsequent time periods. Hence, impressive in-sample results may not be helpful in an actual forecasting application if the variables chosen in one period are not useful forecasters in the next period. An in-sample analysis that utilizes the entire data without checking subperiod stability can therefore give false hope. We now analyze whether it is possible to devise a useful out-of-sample process that allows for dynamic variable selection and coefficient estimation.

5. Out-of-sample Predictability

The out-of-sample forecasting problem consists of two parts. First, a subset of forecasting variables must be chosen from a candidate set. Second, the coefficient loadings on this subset must be estimated. Neither step should use forward-looking data. Our approach identifies, on an ex-ante basis, a time-varying set of forecasting variables, and checks their out-of-sample performance; this is a joint test of variable selection and coefficient stability. We augment our set of in-sample forecasting variables from Section 4 with the two SDF-based expected return forecasts (*sdf_rolling, sdf_growing*), as well as with *ovx_diff* and *sent*. We also use lagged returns of the three major oil companies directly, in addition to *StkIdx*. We do not use *sdf_fullSample*, as this is calculated using full-sample information. In our out-of-sample tests, we do not detrend or residualize the dependent variables; rather, we include lagged dependent variables along with a time trend as potential explanatory variables. This leaves 48 forecasting variables, 20 of which are text and 28 of which are non-text.[27]

A. Ability of a Time-Varying Variable Set to Forecast Out-of-sample

As suggested by our in-sample subperiod analysis, it is likely that the usefulness of any forecasting variable changes over time. To systematically allow for this possibility, we run univariate OLS regressions of each dependent variable (eight-week ahead changes or returns) on each forecasting variable in rolling five-year training windows. Within each training window for each dependent variable, we then rank each forecasting variable based on its standalone R-squared. We classify our forecasting variables into two groups: the *text* group contains our text-based measures and the *baseline* group contains all non-text forecasting variables. With the R-squared rank of each of the text and baseline variables in hand, for each dependent variable, we form three parsimonious forecasting models. The first model, the 2-0 model, contains only the two best performing baseline variables. The second model, the 0-2 model, contains only the two best performing text variables. The third model, the 2-2 model, contains all four of the variables that were identified in the 0-2 and 2-0 models. We do not believe it is useful to look beyond two variables from each set because of overfitting concerns in the training window.[28] All models are formed using ex-ante information only.

For each of the three models, we run rolling five-year lasso regressions to estimate rolling coefficients for out-of-sample forecasting of eight-week ahead changes or returns of our

[27] The following summarizes the number of variables in different parts of our analysis: (I) In-sample predictors (41): 22 non-text and 19 text variables; (II) Out-of-sample predictor (48): The 41 in-sample variables plus *sdf_rolling*, *sdf_growing*, *ovx_diff*, *sent*, *trend*, and the three oil major returns, minus *VIX*, which is excluded because *vix_diff* and *ovx_diff* are already present; (III) Out-of-sample predictors with observations in all subperiods (45): The 48 out-of-sample variables minus *sdf_rolling*, *sdf_growing*, and *ovx_diff*, which are missing data in some subperiods.

[28] In unreported results, we also tried the 1-0, 0-1, and 1-1 models (i.e., which have a single baseline and text variable). The results are comparable to, but slightly weaker than, the two-variable models.

dependent variables using the two baseline and/or text variables chosen in the univariate regression step.[29] In all lasso regressions, we include a constant in addition to the forecasting variables chosen using the univariate regression described above. We consider rolling regressions rather than expanding windows to account for possible regime shifts in the data. An expanding window would ultimately settle on a single regime, and not allow for structural breaks in the forecasting relationships. For the rolling univariate OLS and lasso models, we ensure each five-year training window uses data only from inside the window.[30] Using the lasso coefficient estimates in each training window, we then use the independent variables available at the end of the window to make an eight-week ahead forecast. We then march the training window forward by one week, re-select the variables using univariate OLS regressions, re-estimate the lasso model, and make another eight-week ahead forecast.

To measure the out-of-sample efficacy of the lasso model, we use the five-year rolling averages of the left-hand side variable as the *benchmark model* (call it $\overline{r_t}$), a standard practice for out-of-sample forecasting performance tests. We are again careful to make sure that the averages of eight-week changes used the benchmark model do not extend outside of the five-year training window. Note that the benchmark model is not completely "information free," since oil markets, as well as equity markets, are characterized by pronounced time-series momentum (see Moskowitz, Ooi, and Pedersen 2012) and reversals (Asness, Moskowitz, and Pedersen 2013). For example, most commodity trading advisors (CTAs), who trade on trend, also trade in energy

[29] The lassos are estimated using automatic penalty parameter selection with ten-fold cross validation. See Hastie, Tibshirani, and Friedman (2009) for an introduction to lasso methods.

[30] The first right-hand side observation in a five-year training window occurs eight weeks after the window's start to ensure that no forecasting variable, e.g., lagged 30-day *OilVol*, extends outside of the window. The last right-hand side observation in the training window occurs eight weeks prior to the end of the window to ensure that the dependent variable does not extend beyond the five-year training window. For specifications where *PCAsent*, *PCAfreq*, or *PCAall* were chosen in the in-sample model selection, we re-estimate the PCA's using normalized four-week averages of the topical sentiment series in each five-year training window.

markets. So, $\overline{r_t}$ may not be a completely information-free benchmark, especially because the presence of CTAs may induce trends.[31]

Our main approach, therefore, is to determine a weight that should be attached to the lasso model in a blended forecast that combines that model with the benchmark model. That is, we investigate whether the *optimal* out-of-sample forecasting model places a positive weight on the lasso model that includes our forecasting variables, rather than only making use of the benchmark, rolling average model. We say the optimal out-of-sample model is the one that maximizes out-of-sample R-squared. As Campbell and Thompson (2008) show, the R-squared approach allows us to pin down the economic magnitude of using a blended forecast, as opposed to using $\overline{r_t}$ alone, an analysis we turn to in the next section.

Our blended forecast approach combines information from our forecasting models with that from the trend variable as follows

$$\widehat{r_t}^{(w)} \equiv w \times \widehat{r_t} + (1-w) \times \overline{r_t}, \tag{3}$$

where $\widehat{r_t}$ is the lasso model forecast and w is the weight. If the errors in $\widehat{r_t}$ and $\overline{r_t}$ are not perfectly correlated, then this blended forecast may do better than either forecast on its own. We then evaluate the effectiveness of the blended forecast relative to the benchmark model (i.e., $\overline{r_t}$) using the Campbell and Thompson (2008) definition of out-of-sample R-squared:

$$R_{OOS}^2(w) = 1 - \frac{\sum_t \left(r_t - \widehat{r_t}^{(w)}\right)^2}{\sum_t (r_t - \overline{r_t})^2}, \tag{4}$$

where r_t is the future outcome variable that is being forecasted.

Figure 5 shows which forecasting variables are chosen at each point in our sample for the 2-0 (two baseline variable) model for *DOilVol* and for the 0-2 (two text variable) model for

[31] To avoid data mining, we did not vary the five-year rolling window for the lasso or the rolling mean analysis. It is possible that the out-of-sample forecasting power of our model would be higher at a different estimation horizon.

FutRet. Figure A.2 in the Online Appendix shows variables selected by the 2-0 and 0-2 models for all dependent variables.[32] Two features are notable. First, there is persistence in which variables are chosen over time, although that is not surprising given our use of overlapping five-year training windows. Second, despite that persistence, there is a good deal of time series variation in the selected forecasters. For example, as shown in Figure 5, for the 2-0 model that chooses the top two non-text variables to forecast *DOilVol* out of sample, *OilVol* (lagged oil volatility) is the most persistent variable chosen, and *vix_diff, ovx_diff,* and *sdf_rolling* (an SDF-based risk-premium forecast) are also frequently chosen. It is not surprising that these volatility-related or risk-premium measures are frequently selected as volatility forecasting variables. On the other hand, for the 0-2 model that chooses the best two text variables to forecast *FutRet*, early in the sample *sent* and *fGom* are chosen, *fEnv* and *artcount* are chosen in the middle of the sample, and *fCo* and *sCo* are chosen later in the sample. This suggests that allowing for time variation in the forecast variable set may be useful, perhaps because the most salient topics in the news are changing over time.

Figure 6 shows, for each dependent variable, what the R-squared maximizing weight is to attach to the 2-0, 0-2, and 2-2 model forecasts in (3). On the y-axis, each chart shows $R^2_{OOS}(w)$ from (4) where $\widehat{r_t}$ comes from our out-of-sample forecasting models, and the x-axis shows the weight w in (3). If $R^2_{OOS}(w) \leq 0$ for all w then blending the rolling average forecast with our lasso-based one is not useful, as the rolling average by itself is better. But if, for a range of ws, $R^2_{OOS}(w) > 0$, then augmenting the rolling average signal with the information from our lasso signal is helpful.

[32] The 2-2 model simply combines the variables from the 2-0 and 0-2 models.

Panel A in Figure 6 presents results for two-variable models using only non-text forecasting variables (2-0 models). None of the dependent variable forecasts are improved by using 2-0 models in addition to $\overline{r_t}$ except *DOilVol*, which is predicted by lagged oil volatility and other risk-based measures as shown in Figure 5 and discussed above, and *DProd* to a limited extent. For forecasting *DOilVol*, the optimally-blended model produces an R-squared of 8.44%, which is clearly an economically large increase relative to using lagged average *DOilVol*, i.e., $\overline{r_t}$, as the only forecasting model. We note that our large selection of baseline (non-text) variables shows no evidence of out-of-sample forecasting ability for oil spot percent changes, futures returns, or stock price returns of the oil majors.

Panel B in Figure 6 shows the analysis using 0-2 models containing only text variables. We see that *DSpot,* and more importantly the tradeable *FutRet,* are highly forecastable relative to the rolling mean of these variables (interestingly text-based models are not helpful for *DOilVol*). For forecasting *FutRet* and *DSpot*, blended forecasts with positive weights on our text-based lasso models produce the highest R-squareds of 1.34% and 0.59%, respectively. These results also provide further evidence for the argument we highlighted in Section 4 that our text variables do not seem to be capturing risk, given that non-text lasso models are good at forecasting *DOilVol* – the dependent variable that is most reflective of risk – but text-based models are not.

Finally, Panel C of Figure 6 shows that the results for *FutRet, DSpot*, and *DOilVol* are weaker in the four-variable model (2-2 model) than in the respective text only or non-text only versions. Only *bpRet* appears forecastable by the four-variable models using two text and two non-text variables, with a maximal R-squared of 0.42% for the blended forecast. So, our results suggest that non-text variables can forecast oil volatility and oil production (to an extent), while text variables can forecast changes in oil spot and futures prices. Only *bpRet* seems forecastable to some degree by combining text and non-text variables.

Several features of this analysis are noteworthy. First, text and non-text variables have different usefulness depending on which dependent variable one focuses on. Second, the out-of-sample R-squareds are generally stable in the vicinity of the optimal blending weight in (3), suggesting that there is a range of blending weights that lead to superior out-of-sample performance. Finally, the magnitude of the R-squareds in models with positive weightings on our text-based variables turn out to be economically large, which we explore next.

B. Evaluating the Economic Impact of Out-of-Sample Forecastability

We showed that for a range of weights in (3), our text-based forecasting models have out-of-sample predictive power for oil futures and oil spot returns. But, is the degree of out-of-sample forecastability economically meaningful? To investigate this question, we follow an approach suggested by Campbell and Thompson (2008). Consider a mean-variance arbitrageur, e.g., a CTA, who invests in oil futures. Assume that eight-week ahead returns on oil futures are given by

$$r_{t+1} = \mu + x_t^{(i)} + \epsilon_{t+1}^{(i)} \tag{5}$$

where μ is the mean return, $x_t^{(i)}$ is time *t*, mean-zero information available from forecasting model *i* – either the rolling average signal or the blended one from (3) – and $\epsilon_{t+1}^{(i)}$ is a mean-zero residual term that is orthogonal to $x_t^{(i)}$. An arbitrageur with access to $x_t^{(i)}$ and risk aversion γ would optimally invest a fraction $w_t^{(i)}$ of his or her wealth into the oil futures contract (with the rest kept in cash) with

$$w_t^{(i)} = \frac{\mu + x_t^{(i)}}{\gamma \sigma_{\epsilon(i)}^2} \tag{6}$$

where $\sigma^2_{\epsilon(i)}$ is the variance of the noise term in (5).[33] [34]

We now evaluate the unconditional expected gain to the arbitrageur from following the trading strategy in (6). Since the excess payout, $\pi^{(i)}_t$, from a dollar invested in this oil futures strategy is $w^{(i)}_t \times r_{t+1}$, the unconditional expectation of the eight-week payout is

$$E\left[\pi^{(i)}_t\right] = \frac{\mu^2 + \sigma^2_{x(i)}}{\gamma\sigma^2_{\epsilon(i)}}, \tag{7}$$

where $\sigma^2_{x(i)}$ is the variance of the signal $x^{(i)}_t$ in (5). The unconditional Sharpe ratio of investing in oil futures with no information (i.e., treating $x^{(i)}_t$ as a random variable) is

$$S = \frac{\mu}{\left(\sigma^2_{x(i)} + \sigma^2_{\epsilon(i)}\right)^{1/2}}.$$

Note that S does not depend on the variance of the signal $x^{(i)}_t$ as long as the sum in the denominator is constant. The R-squared from regressing r_{t+1} on $x^{(i)}_t$ in (5) is

$$R^2_{(i)} = \frac{\sigma^2_{x(i)}}{\sigma^2_{x(i)} + \sigma^2_{\epsilon(i)}}.$$

Dividing both the numerator and denominator of (7) by $\sigma^2_{x(i)} + \sigma^2_{\epsilon(i)}$ and using the definitions of the unconditional Sharpe ratio and R-squared, we obtain

$$E\left[\pi^{(i)}_t\right] = \frac{1}{\gamma} \times \frac{S^2 + R^2_{(i)}}{1 - R^2_{(i)}}. \tag{8}$$

[33] This follows from the standard result that a mean-variance investor would optimally allocate a fraction $E[r - r_f\,]/(\gamma \times \sigma^2_{r-r_f})$ of his wealth to the risky asset. Given information $x^{(i)}_t$, the expected excess return on oil futures is $\mu + x^{(i)}_t$ and the variance of oil future excess returns is $\sigma^2_{\epsilon(i)}$.

[34] It would be more accurate to write (5) as $r_{t+1} = \mu + x^{(i)}_t + \bar{\epsilon}^{(i)} + \epsilon^{(i)}_{t+1}$ where $\bar{\epsilon}^{(i)}$ adjusts for a potentially non-zero bias in the signal $x^{(i)}_t$. In this case, the optimal oil future allocation would still be given by (6), but the mean of $x^{(i)}_t$ would no longer be zero. However, because the bias in the rolling average signal and the blended signal from (3) are almost identical, ignoring this effect will not impact our analysis of the relative benefit of the blended signal.

In other words, the unconditional expected gain to the oil arbitrageur from trading on the signal $x_t^{(i)}$ is proportional to the sum of the square of the unconditional Sharpe ratio and the R-squared of the regression of future returns on the time t signal.

Consider two mean-variance arbitrageurs, one of whom can trade on signal *c*, which represents the rolling mean benchmark $\overline{r_t}$, and one of whom can trade on *b*, the blended forecast $\widehat{r_t}^{(w)}$ from (3). Using (8), the difference in their expected gains is given by

$$E\left[\pi_t^{(b)} - \pi_t^{(c)}\right] = \frac{1}{\gamma}(1+S^2)\frac{R_{(b)}^2 - R_{(c)}^2}{\left(1-R_{(b)}^2\right)\left(1-R_{(c)}^2\right)}. \qquad (9)$$

This is analogous to equation (13) in Campbell and Thompson (2008), except their $R_{(c)}^2$ is assumed to be zero. To get a sense of the magnitude of this effect, consider an arbitrageur with unit risk aversion, $\gamma = 1$, as in the example in Campbell and Thompson (2008). Table 1 shows that the eight-week mean and volatility of oil future returns are 1.349% and 13.78%, respectively. The unconditional average $w_t^{(i)}$ from (6) therefore equals 0.71, which implies the arbitrageur, on average, is 71% invested in oil futures with 29% held in cash. A more aggressive strategy, for example being, on average, fully invested in oil futures, would require a lower γ and would make the economic impact of our predictability result larger (via the $1/\gamma$ term in (9)).

Assuming both R-squareds in (9) are positive and $\gamma = 1$, the eight-week expected performance gain on a dollar investment from trading on the blended signal *b* relative to the rolling average signal *c* can be bounded below by the difference in out-of-sample R-squareds, i.e., $E\left[\pi_t^{(b)} - \pi_t^{(c)}\right] > R_{(b)}^2 - R_{(c)}^2$. The empirical analogue of this difference in R-squareds is the peak increase in the out-of-sample R-squared of the blended model relative to the rolling mean model, which equals 1.34% for *FutRet,* as shown in Panel B of Figure 6. A rough annualization – multiplying this performance gain by 52/8 – yields an expected gain of 8.71% per year. An arbitrageur who statically allocates to oil futures using (6) without using any signal (i.e., $x_t^{(i)} = 0$) would earn an average annual excess return of $0.71 \times 1.349\% \times 52/8 = 6.23\%$ (fraction of portfolio in oil futures times the average 8-week futures excess return from Table II times the annualization factor). The 8.71% expected gain of trading on our text-based *FutRet* signal in combination with the rolling mean, rather than using the rolling mean alone, is therefore quite large relative to this passive benchmark.

Similar arguments show that the economic gain for *bpRet* from a 0.42% R-squared forecasting improvement from using the blended 2-2 model is also sizable.

6. Conclusion

The global energy market is vital for economic growth, but it is also very volatile, making timely information and robust forecasting crucial. Our paper makes several contributions in this context. We construct novel NLP measures for energy markets and show that they perform at least as well in-sample as traditional forecasting variables in predicting a set of energy market outcomes. We show that the forward selection method is particularly well-suited for in-sample forecasting analysis with many predictors. Our bootstrap methodology allows for statistical inference and adjusted R-squareds in forward selection models with overlapping observations, persistent regressors, and the Stambaugh (1999) bias, all of which are challenging and important problems. Our caution that robust in-sample predictability may not translate to robust out-of-sample predictability certainly is a point that has been made before, but our methodology allows for convincing out-of-sample tests that are not subject to reporting bias. An important factor which causes in-sample models to fail out-of-sample is model instability: today's forecasting variables aren't tomorrow's forecasting variables.

Our out-of-sample tests allow for the model to change over time, and confirm the usefulness of text measures as forecasters of oil futures returns and changes in spot prices. For those two dependent variables, a blended forecast that uses the output of a parsimonious two-text-variable model and a rolling average return outperforms either forecast on its own. The degree of out-of-sample forecastability afforded by text variables is economically large. Interestingly, non-text variables show no evidence of out-of-sample forecasting ability for oil spot price changes or oil future returns. For forecasting changes in volatility and changes in production, however, non-text variables prove superior as out-of-sample forecasters. For *bpRet*, a model that combines text and baseline variables shows evidence of out-of-sample forecasting ability. Importantly, in all forecasting tasks, the selected forecasting variables change over time. In the case of text measures used in returns forecasting, we conjecture that this reflects the changing importance of different types of news over time. In the case of volatility, the variables selected not only are all well-known measures of risk; but, interestingly, the ones that prove most useful for forecasting also vary over time.

We leave several interesting questions for future research. Are our forecasting results indicative of time-varying risk premia in oil markets, slow diffusion of information, or investor overconfidence? Are our NLP measures reflective of sentiment or hard information? Does the predictability we find over four- and eight-week horizons carry over to longer time periods? What does the time variation in selected forecasting variables tell us about the economics of energy markets? We hope that the tools we have developed will allow us and other researchers to make progress on these and related questions in future work.

7. References

Alquist, Ron, Lutz Kilian, and Robert Vigfusson. 2013. “Forecasting the Price of Oil,” in: G. Elliott and A. Timmermann (eds.), Handbook of Economic Forecasting, 2A, Amsterdam: North-Holland, 427-507.

Amihud, Yakov, Haim Mendelson, and Beni Lauterbach, 1997. “Market microstructure and securities values: Evidence from the Tel Aviv Stock Exchange," *Journal of Financial Economics*, 45, 365-390.

Ang, Andrew, and Geert Bekaert. 2007. “Stock Return Predictability: Is it there?” *Review of Financial Studies*, 20, 651-707.

Asness, C., T. Moskowitz, and L. Pedersen. 2013. “Value and momentum everywhere,” *Journal of Finance*, 68 (3), 929-985.

Bali, T., A. Goyal, D. Huang, F. Jiang, and Q. Wen. 2022. “Predicting corporate bond returns: Merton meets machine learning,” working paper.

Baumeister, Christiane, and James Hamilton. 2019. “Structural Interpretation of Vector Autoregressions with Incomplete Identification: Revisiting the Role of Oil Supply and Demand Shocks,” *American Economic Review,* 109, 1873-1910.

Baumeister, Christiane, Dimitris Korobilis, and Thomas K. Lee. 2022. “Energy Markets and Global Economic Conditions,” *Review of Economics and Statistics*, 104(4).

Baumeister, Christiane, and Lutz Kilian. 2015. “Forecasting the Real Price of Oil in a Changing World: A Forecast Combination Approach,” *J. Bus. and Econ. Statistics* 33(3): 338-351.

Baumeister, Christiane, and Lutz Kilian. 2017. “A General Approach to Recovering Market Expectations from Futures Prices with an Application to Crude Oil,” working paper.

Bekaert, G. and M. Hoerova. 2014. “The VIX, the variance premium and stock market volatility,” *Journal of Econometrics*, 183, 181-192.

Bessembinder, H., and K. Chan. 1992. “Time-Varying Risk Premia and Forecastable Returns in Futures Markets,” *Journal of Financial Economics*, 32, 169-193.

Blei, D., A. Ng, and M. Jordan, 2003, “Latent Dirichlet allocation,” *Journal of Machine Learning Research*, 3, 993-1022.

Blondel, V., J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, 2008, “Fast unfolding of communities in large networks,” *Journal of Statistical Mechanics*, 10, 10008.

Boons, M. and M. P. Prado, 2018, “Basis-Momentum,” *Journal of Finance, 74 (1), 239—279.*

Boudoukh, J., M. Richardson, and R. Whitelaw. 2008. "The myth of long-horizon predictability," *Review of Financial Studies*, 21 (4), 1577-1605.

Brandes, U., D. Delling, M. Gaertler, R. Goerke, M. Hoefer, Z. Nikoloski, D. Wagner, 2006, "Maximizing modularity is hard," *arXiv/physics*.

Brandt, M. W. and L. Gao. 2019. "Macro fundamentals or geopolitical events? A textual analysis of news events for crude oil," *Journal of Empirical Finance*, 51, 64-94.

Calomiris, Charles W., and Harry Mamaysky. 2019a. "How News and Its Context Drive Risk and Returns Around the World," *Journal of Financial Economics,* 133, 299-336.

Calomiris, Charles W., and Harry Mamaysky. 2019b. "Monetary Policy and Exchange Rate Returns: Time-Varying Risk Regimes," *NBER Working Paper No. 25714.*

Campbell, J. and S. Thompson, 2008, "Predicting excess returns out of sample: Can anything beat the historical average?" *Review of Financial Studies*, 21 (4), 1509-1531.

Cochrane, J. 2005. Asset Pricing, *Princeton University Press*.

Cavallo, M. and T. Wu. 2011. "Measuring oil-price shocks using market-based information," *International Monetary Fund Working Paper No. 12/19.*

Conlon, T., J. Cotter, E. Eyiah-Donkor, 2022, "The illusion of oil return predictability: The choice of data matters," *Journal of Banking and Finance,* 134.

Dahl, Carol. 2004. "International Energy Markets: Understanding Pricing, Policies, and Profits," *Tulsa: PennWell.*

Datta, Deepa, and Daniel Dias. 2020. "Oil Shocks: A Textual Analysis Approach." Mimeo.

De Roon, F.A., T.E. Nijman, and C. Veld. 2000. "Hedging Pressure Effects in Futures Markets," *Journal of Finance*, 55, 1437-1456.

Fama, E. 1965. "The Behavior of Stock-Market Prices," *Journal of Business*, 38 (1), 34-105.

Feng, G., S. Giglio, and D. Xiu. 2020. "Taming the factor zoo: A test of new factors," *Journal of Finance*, 75 (3), 1327–1370.

Foster, F D., T. Smith, and R.E. Whaley. 1997. "Assessing Goodness-of-Fit of Asset Pricing Models: The Distribution of the Maximal R-Squared," *Journal of Finance*, 52 (2), 591-607.

Garcia, D., X. Hu, and M. Rohrer. 2020. "The colour of finance words," working paper.

Geman, Helyette. 2005. "Commodities and Commodity Derivatives: Modeling and Pricing for Agricultural Metals and Energy," West Sussex: John Wiley & Sons.

Giglio, S., Y. Liao, and D. Xiu, 2021. “Thousands of alpha tests,” *Review of Financial Studies*, 34 (7), 3456–3496.

Glasserman, Paul, and Harry Mamaysky. 2019. “Does Unusual News Forecast Market Stress?” *Journal of Financial and Quantitative Analysis*, 54 (5), 1937-1974.

Glasserman, P., K. Krstovski, P. Laliberte, and H. Mamaysky, 2020, “Choosing news topics to explain stock market returns,” *Proc. of ACM Intl. Conf. on AI in Finance*, *ICAIF-2020*.

Gorton, Gary, Fumio Hayashi, and K. Geert Houwenhorst. 2013. “The Fundamentals of Commodity Futures Returns,” *Review of Finance,* 17, 35-105.

Gu, S., B. Kelly, and D. Xiu. 2020. “Empirical asset pricing via machine learning,” *Review of Financial Studies*, 33, 2223–2273.

Hamilton, James D., and J. Cynthia Wu. 2014. “Risk Premia in Crude Oil Futures Prices,” *Journal of International Money and Finance*, 42, 9-37.

Hansen, L. and R. Jagannathan. 1991. “Implications of security market data for models of dynamic economies,” *Journal of Political Economy*, 99 (2), 225-262.

Harvey, C., Y. Liu, and H. Zhu. 2016. “…and the cross-section of expected returns,” *Review of Financial Studies*, 29 (1), 5-68.

Hastie, T. R. Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning.* Springer.

Hastie, T., R. Tibshirani, and R.J. Tibshirani. 2017. “Extended comparisons of best subset selection, forward stepwise selection, and the lasso,” working paper.

Hong, Harrison, and Motohiro Yogo. 2012. “What Does Futures Market Interest Tell Us About the Macroeconomy and Asset Prices?” *Journal of Financial Economics* 105, 173-490.

Hodrick, R. 1992. “`Dividend yields and expected stock returns: Alternative procedures for inference and measurement,” *Review of Financial Studies*, 5 (3), 357-386.

Ke, Z.T., B. Kelly, and D. Xu, 2019, “Predicting returns with text data,” working paper.

Kirby, C., 1997, “Measuring the predictable variation in stock and bonds returns,” *Review of Financial Studies*, 10 (3), 579—630.

Kozak, S., S. Nagel, and S. Santosh. 2020. “Shrinking the cross-section,” *Journal of Financial Economics*, 135, 271–292.

Li, X., W. Shang, and S. Wang, 2019, “Text-based crude oil price forecasting: A deep learning approach,” *International Journal of Forecasting*, 35, 1548-1560.

Loughran, Tim, Bill McDonald, and Ioannis Pragidis. 2019. “Assimilation of Oil News Into Prices,” *International Review of Financial Analysis* 63, 105-118.

Loughran, T. and B. McDonald. 2011. “When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks,” *Journal of Finance*, 66, 35-65.

Mamaysky, H., Y. Shen, and H. Wu, 2021, “Drivers of credit spreads,” working paper.

Manescu, C., and I. van Robays. 2016. “Forecasting the Brent Oil Price: Addressing Time-Variation in Forecast Performance,” mimeo, ECB.

Moskowitz, T., Y. H. Ooi, and L. Pedersen. 2012. “Time series momentum,” *Journal of Financial Economics*, 104, 228–250.

Newman, M.E.J. and M. Girvan, 2004, “Finding and evaluating community structure in networks,” *Physical Review E*, 69, 026113.

Plante, M. 2019. “OPEC in the news,” *Energy Economics*, 80, 163-172.

Stambaugh, R. F. 1999. “Predictive regressions,” *J. Financial Economics*, 54 (3), 375-421.

Szymanowska, M., F. De Roon, T. Nijman And R. Van Den Goorbergh. 2014. “An Anatomy of Commodity Futures Risk Premia,” *The Journal of Finance*, 69 (1), 453-482.

Tetlock, P. C., 2007. “Giving content to investor sentiment: the role of media in the stock market,” *Journal of Finance,* 62, 1139-1168.

Tetlock, P. C., M. Saar-Tsechansky, and S. Macskassy. 2008. “More than words: quantifying language to measure firms’ fundamentals,” *Journal of Finance,* 63, 1437-1467.

Welch, I. and A. Goyal. 2008. “A comprehensive look at the empirical performance of equity premium prediction,” *Review of Financial Studies*, 21 (4), 1455-1508.

Yang, Fan. 2013. “Investment Shocks and the Commodity Basis Spread,” *JFE*, 110, 164-184.

Table I: Data Definitions Summary

Variable	Definition
Dependent Variables	
FutRet^{8}	WTI front-month futures cumulative weekly returns (in %) from the end of week *t* to the end of week *t*+8
DSpot^{8}	Percent change in the WTI spot price from the end of week *t* to the end of week *t*+8
DOilVol^{8}	Level difference in the rolling 30-day realized volatility of WTI physical futures 1-month nearby contract from the end of week *t* to the end of week *t+8*
xomRet^{8}	Exxon Mobil stock returns (in %) from the end of week *t* to the end of week *t*+8 (trades on NYSE)
bpRet^{8}	British Petrol stock returns from the end of week *t* to the end of week *t+8* (ADR trading on NYSE)
rdsaRet^{8}	Royal Dutch Shell class A stock returns from Monday of week *t+1* to Monday of week *t+9* (trades on Euronext)
DInv^{8}	Percent change in U.S. crude inventories including SPR (EOP, mil. bbl) from the end of week *t* to the end of week *t+8*
DProd^{8}	Average weekly percent change in U.S. crude oil field production (mil. bbl/day) from the end of week *t* to the end of week *t+8*
Non-text Variables	
OilVol	Rolling 30-day realized volatility of WTI physical futures 1-month nearby contract
VIX	CBOE market volatility index
DFX	Percent change in the weekly nominal broad dollar index - goods only (Jan 1997 = 100) relative to 4 weeks ago
tnote_10y	10-year treasury note yield at constant maturity (EOP, % p.a.)
sp500Ret	Standard and Poor's 500 weekly stock returns relative to 4 weeks ago
StkIdx	Average of Exxon Mobil, British Petrol, and Royal Dutch Shell class A stock returns from week *t* to week *t*+8
WIPI	Month-over-month growth rate of Baumeister and Hamilton's (2019) monthly World Industrial Production Index
basis	WTI physical annualized 3-month to 1-month basis (when positive curve is upward sloping, capturing contango)
trend	Weekly linear time trend
vix_diff	The difference between CBOE market volatility index and the 30-day volatility of Standard and Poor's 500 index
ovx_diff	The difference between CBOE crude oil volatility index and the 30-day volatility of WTI crude oil prices
sdf_ fullSample	Risk premium calculated from annual covariance with full-sample stochastic discount factor

BE/ME	Average WTI spot price from month t-67 to t-56 relative to the spot price of month t-1
Mom	WTI front-month futures cumulative monthly returns starting in month t-11 through month t-1
BasMom	The difference between momentum (Mom) for the WTI front-month contract and the momentum (Mom) for the WTI month-after-front-month contract
DolBeta	Coefficient from a 60-month rolling regression of monthly WTI futures returns on changes in logarithm of dollar spot index (DXY).
InflaBeta	Coefficient from a 60-month rolling regression of monthly WTI futures returns on unexpected inflation, measured by the change in one-month CPI inflation (yearly change of CPI)
HedgPres	The difference between the number of short and long hedging contracts by large traders in crude oil market relative to the total number of hedging contracts by large traders in crude oil market
liquidity	Logarithm of WTI futures trading volume (number of contracts) relative to the absolute WTI futures return on that trading day
OpenInt	Logarithm of the product of WTI spot price, quantity of WTI futures contracts outstanding, and WTI futures contract size (Tuesday/Friday)
Text Variables	
PCAsent	The first principal components (PCAs) of the seven topical sentiment series, where PCAs are calculated using the four-week averages of the weekly series.
PCAfreq	The first principal components (PCAs) of the seven topical frequency series, where PCAs are calculated using the four-week averages of the weekly series.
PCAall	The first principal components (PCAs) of all fourteen series together, where PCAs are calculated using the four-week averages of the weekly series.
artcount	Average number of articles in the energy corpus over the past 4 weeks
entropy	Average measure of article unusualness over the past 4 weeks
s[Topic]	Average sentiment over the previous 4 weeks due to Topic. Topic is one of company (Co), global oil market (Gom), environment, (Env), energy/power generation (Epg), crude oil physical (Bbl), refining and petrochemicals (Rpc), or exploration and production (Ep).
f[Topic]	Average frequency of articles over the previous 4 weeks in Topic. Topic is one of company (Co), global oil market (Gom), environment, (Env), energy/power generation (Epg), crude oil physical (Bbl), refining and petrochemicals (Rpc), or exploration and production (Ep).

Table II: Descriptive Statistics

Data are weekly observations from April 1998 to March 2020. The variables labeled *t8* show eight-week changes (the *t8* is suppressed in labels used in other tables). The other non-text series are observed weekly, some as changes and some as levels, and the text variables are four-week averages of weekly observations. The data are observed on Tuesday for non-price series, and on Thursday for price-based series. For each variable, the table shows the mean, standard deviation, median, and the 5th and 95th percentiles. *N* is the number of observations in the sample. Variable definitions are presented in Table I. The text measures, except *entropy,* are standardized to mean zero and unit variance in the regressions but are not standardized here.

Panel A: Non-text Variables

VARIABLES	mean	sd	p5	p50	p95	N
FutRet_t8	1.349	13.78	-22.76	2.512	21.62	1,139
DSpot_t8	0.636	14.75	-24.93	2.734	20.31	1,139
DOilVol_t8	0.164	14.44	-22.72	-0.570	23.69	1,139
xomRet_t8	0.191	7.611	-11.68	0.600	11.52	1,139
bpRet_t8	-0.339	10.01	-15.56	0.407	13.03	1,139
rdsaRet_t8	-0.281	9.368	-14.63	0.614	12.76	1,139
DInv_t8	0.137	1.742	-2.596	0.145	2.966	1,139
DProd_t8	0.383	3.056	-2.888	0.402	4.026	1,136
OilVol	35.90	15.96	17.95	32.71	65.35	1,147
VIX	20.05	8.826	11.21	17.97	35.79	1,146
DFX	0.0544	1.507	-2.269	-0.0220	2.424	1,141
tnote_10y	3.567	1.327	1.700	3.580	5.850	1,147
sp500Ret	0.308	4.707	-7.545	1.007	6.077	1,141
StkIdx	-0.134	6.184	-10.208	0.391	8.372	1,141
WIPI	0.208	0.602	-0.674	0.260	1.004	1,147
basis	0.0717	0.314	-0.265	0.0462	0.434	1,147
trend	574	331.3	58	574	1,090	1,147
vix_diff	3.235	4.566	-4.430	3.620	9.480	1,146
ovx_diff	1.820	8.400	-14.91	3.070	12.66	673
sdf_fullSample	0.0414	0.0305	0.00680	0.0329	0.0947	1,052
BE/ME	0.933	0.538	0.349	0.733	1.979	1,147
Mom	7.737	32.85	-46.23	7.295	66.76	1,147
BasMom	0.188	3.728	-5.595	0.0504	6.067	1,147
DolBeta	-0.959	0.792	-2.126	-1.210	0.104	1,147
InflaBeta	6.919	3.589	0.0464	6.706	13.14	1,147
HedgPres	-0.00797	0.0390	-0.0671	-0.0104	0.0609	1,146
liquidity	15.60	1.473	13.40	15.50	18.21	1,141
OpenInt	22.82	1.232	20.83	23.12	24.19	1,146
OpenInt (bln. $)	13.59	10.92	1.111	11.00	32.16	1,146

Panel B: Text Variables

VARIABLES	mean	sd	p5	p50	p95	N
PCAsent	-0	1.489	-2.185	0.270	2.158	1,144
PCAfreq	0	1.764	-2.174	-0.754	2.926	1,144
PCAall	0	2.423	-2.961	-1.047	3.715	1,144
artcount	332.5	114.4	172.9	353.3	521.5	1,144
entropy	2.150	0.116	1.948	2.170	2.305	1,144
sCo	-0.00119	0.000350	-0.00181	-0.00110	-0.000752	1,144
fCo	0.127	0.0474	0.0751	0.120	0.221	1,144
sGom	-0.00472	0.00180	-0.00803	-0.00434	-0.00239	1,144
fGom	0.346	0.104	0.213	0.334	0.508	1,144
sEnv	-0.000561	0.000328	-0.00115	-0.000551	-0.000149	1,144
fEnv	0.0318	0.0174	0.00824	0.0330	0.0579	1,144
sEpg	-0.00564	0.00137	-0.00784	-0.00551	-0.00352	1,144
fEpg	0.355	0.0543	0.261	0.369	0.431	1,144
sBbl	-0.000430	0.000228	-0.000936	-0.000355	-0.000196	1,144
fBbl	0.0387	0.0160	0.0195	0.0345	0.0680	1,144
sRpc	-0.000341	0.000108	-0.000571	-0.000326	-0.000193	1,144
fRpc	0.0203	0.00446	0.0147	0.0194	0.0288	1,144
sEp	-0.000472	0.000197	-0.000766	-0.000443	-0.000226	1,144
fEp	0.0358	0.0116	0.0211	0.0338	0.0554	1,144

Table III: Sample Sentences

This table shows headlines associated with the topic-specific episodes marked with stars in Panels A and B of Figure 2, which identify extreme values of topic-specific sentiment that coincide with high values of entropy. Each episode is labeled with its respective time frame, which is defined by article dates related to the same episode. Articles for each episode must belong predominantly ($f_{i,\tau} > 0.8$) to the episode's topical category. For each event, the headlines of the five most negative sentiment articles are chosen from the candidate set, which consists of articles with an entropy higher than 2 and with a total number of words higher than 100. The *Sentiment* and *Entropy* columns correspond to the values of sentiment and entropy observed for that article.

Sentiment	Entropy	Date	Headline
			Co: UK fuel protests from 2000-08-23 to 2000-09-20
-0.115	2.298	9/12/2000	UK's Blair to hold urgent talks over fuel crisis
-0.092	2.361	9/12/2000	EU asks Belgium for information on trucks protest
-0.072	2.395	9/13/2000	UPDATE 1-UK business says fuel crisis hurting
-0.069	2.347	9/13/2000	Fuel crisis costs UK firms 250 mln stg a day –LCC
-0.068	2.447	9/19/2000	EU govts to hold crisis talks far from Brussels
			Gom: Failed Venezuelan coup from 2002-03-27 to 2002-04-24
-0.132	2.483	4/12/2002	Venezuela PDVSA staff say oil exports being restored
-0.128	2.476	4/12/2002	Venezuela PDVSA staff say oil exports being restored
-0.111	2.44	4/11/2002	U.S. concerned about Venezuela, urges moderation
-0.102	2.403	4/5/2002	UPDATE 1-Oil protest grips Venezuela, disruptions reported
-0.097	2.504	4/12/2002	IPE Brent lower as Venezuela supply concerns ease
			Env: Volkswagen emissions scandal from 2015-09-16 to 2015-10-14
-0.107	2.347	9/24/2015	Nidera says suffers significant loss from biofuels fraud
-0.09	2.364	9/23/2015	BRIEF-Fitch places Volkswagen AG on Rating Watch Negative
-0.09	2.312	10/2/2015	UPDATE 1-VW faces French inquiry for 'aggravated deception' in emissions scandal
-0.071	2.391	9/20/2015	UPDATE 1-Volkswagen orders investigation into breach of US environment rules
-0.063	2.431	9/21/2015	UPDATE 1-Volkswagen shares plunge on U.S. emissions scandal
			Epg: Post-bankruptcy Enron hearings from 2002-01-16 to 2002-02-13
-0.131	2.372	2/12/2002	Calif senate panel seeks contempt citation vs. Enron
-0.123	2.33	2/6/2002	Enron skips Calif. hearing, may face contempt charges
-0.114	2.333	2/4/2002	UPDATE 1-Global Crossing says panel to probe accounting
-0.108	2.312	2/8/2002	Court seen for Enron bigwigs as Congress probes
-0.095	2.34	1/23/2002	Calif. court orders Enron to save documents
			Bbl: Hurricane Katrina from 2005-08-24 to 2005-09-21
-0.075	2.367	9/12/2005	UPDATE 1-FEMA chief Brown resigns in wake of Katrina
-0.059	2.342	9/12/2005	FEMA revises Brown's bio after exaggeration charges
-0.057	2.268	9/2/2005	Bush signs $10.5 bln spending bill for Katrina
-0.055	2.331	9/13/2005	U.S. lawmaker won't reopen bankruptcy for Katrina
-0.055	2.349	8/31/2005	UPDATE 1-Bush says will take years to recover from Katrina
			Ep: BP oil spill aftermath from 2010-05-05 to 2010-06-02
-0.078	2.12	5/6/2010	UPDATE 1-Pioneer Drilling Q1 loss wider than expected
-0.072	2.488	6/1/2010	UPDTAE 1-Goldman removes Halliburton from conviction buy list
-0.061	2.367	5/27/2010	UPDATE 1-Carrefour, unions reach Belgian restructuring deal
-0.058	2.583	6/1/2010	Transocean, Halliburton credit default swaps surge
-0.057	2.32	5/13/2010	UPDATE 1-Transocean seeks to limit spill liability

Table IV: Stepwise Forward Selection at the Eight-Week Horizon

The table shows the forecasting regression results for all eight dependent variables at the eight-week horizon using stepwise forward selection to choose seven regressors from all the variab described in Table I, except *ovx_diff* (which is only available after 2007) and *sdf_fullSampl*e (the values of which reflect future data). We also exclude three energy company stock retu (*xomRet, bpRet, rdsaRet*) from our regressors and instead include *StkIdx* (which is the average of the three stock returns). All dependent and independent variables are first detrended residualized with respect to the lagged four-week dependent variable. Only predictors that were chosen by at least one model are included in this table. Coefficients are standardized using ratios of the standard deviations of the dependent and predicting variables. Superscripts before coefficients indicate order in forward selection (1=chosen first). The p-values are obtained us Monte Carlo simulations that use an AR(8) process to simulate the LHS variable, as well as forward selection to produce both adjusted R^2 and t-statistic distributions. The p-values refer to minimum of the fraction of simulated t-statistics less than the empirical t-statistic, and 1 minus the fraction of simulated t-statistics less than the empirical t-statistic, where the compariso relative to the distribution of the order in which the variables were chosen. The bootstrap was repeated 1,000 times. The table also reports the mean of simulated adjusted R^2 resulting from same bootstrap, as well as the corresponding CDF percentage, computed as the percent of adjusted R^2 simulations less than the empirical adjusted R^2. Statistically significant results at the 1 level or better are shown in bold.

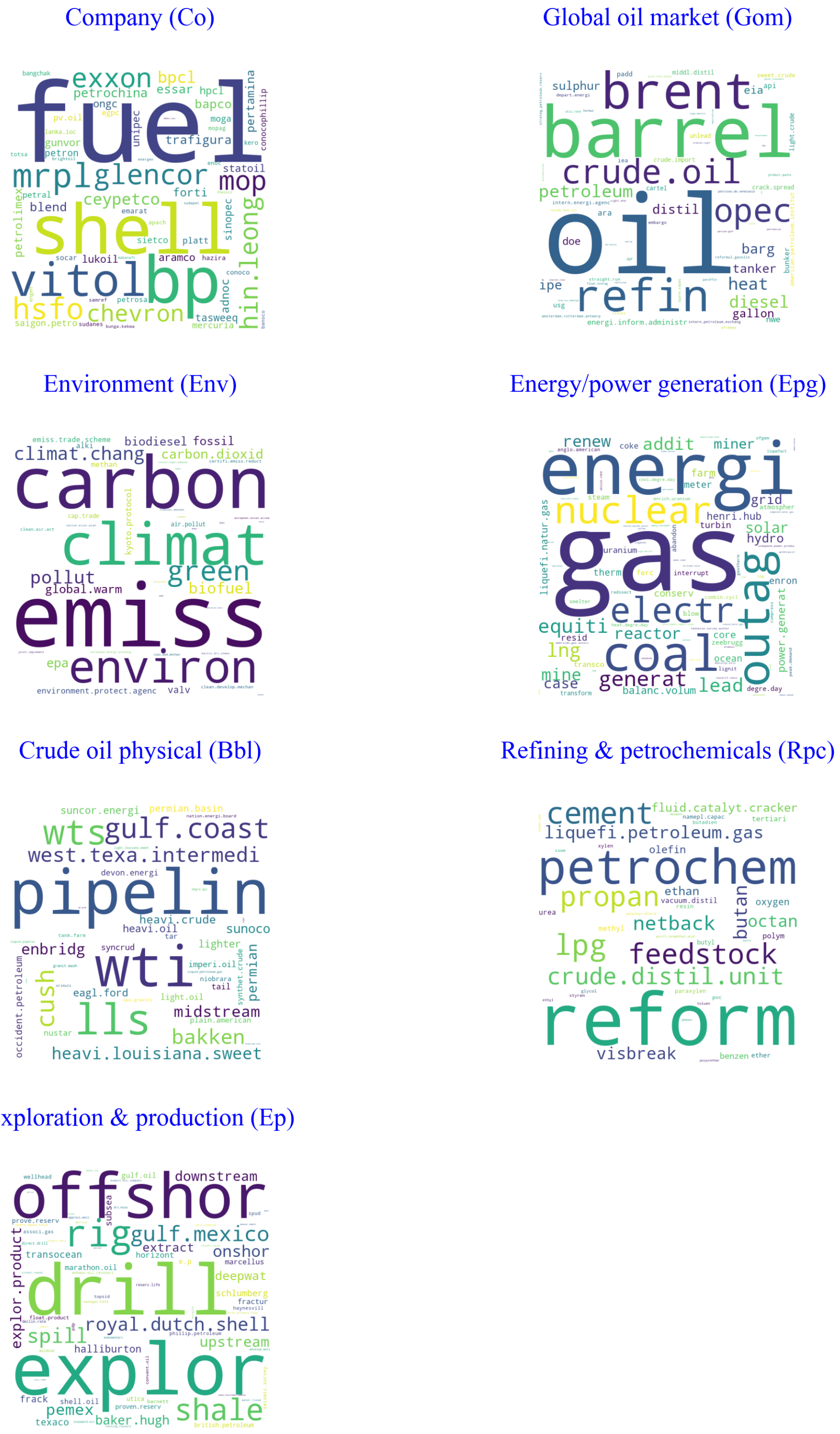
Predictors	*FutRet*		*Dspot*		*DOilVol*		*xomRet*		*bpRet*		*rdsaRet*		*DInv*		*DProd*	
	coef	*pval*	*coef*	*pval*	*coef*	*pval*	*coef*	*pval*	*coef*	*pval*	*coef*	*pval*	*coef*	*pval*	*coef*	*pval*
DSpot					**^{2}-0.24**	0.004									^{3}-0.07	0.462
DInv							^{2}0.11	0.401			^{5}0.11	0.129				
OilVol					**^{1}-0.67**	0.000										
tnote_10y									**^{7}-0.13**	0.003			^{6}-0.11	0.23		
WIPI					**^{6}-0.13**	0.041										
basis	^{2}0.10	0.445	^{1}0.20	0.175					**^{6}-0.14**	0.022						
BE/ME							^{5}0.15	0.113							^{7}-0.12	0.077
InflaBeta											**^{7}-0.13**	0.01			^{4}0.09	0.447
BasMom													^{7}0.10	0.067		
Mom	**^{6}-0.18**	0.002	**^{7}-0.15**	0.007					**^{1}-0.29**	0.004	^{4}-0.12	0.17				
VIX			^{4}-0.05	0.489	**^{3}0.28**	0.001									^{1}0.11	0.438
vix_diff									^{5}-0.14	0.052						
PCAall	**^{5}0.28**	0.003	**^{6}0.24**	0.002												
entropy	**^{3}0.22**	0.017	**^{3}0.27**	0.002	**^{5}-0.28**	0.000	^{6}0.12	0.141	^{3}0.17	0.057	**^{1}0.36**	0.008	^{1}-0.23	0.101		
fCo					**^{7}-0.14**	0.015					^{3}0.16	0.214				
sGom	**^{4}0.25**	0.004	**^{5}0.25**	0.001					**^{4}0.19**	0.019	^{2}0.18	0.084	^{3}-0.09	0.46		
fGom					**^{4}0.26**	0.000									^{2}-0.12	0.376
sEnv									**^{2}0.22**	0.012	^{6}0.13	0.059				
sEpg							^{4}0.11	0.347								
fBbl	**^{7}-0.12**	0.028					^{1}-0.16	0.258							^{6}0.15	0.066
sRpc							^{7}0.09	0.268							^{5}0.06	0.491
fRpc	^{1}0.06	0.484	^{2}0.06	0.483									**^{2}-0.22**	0.008		
sEp							^{3}-0.09	0.396					**^{4}-0.35**	0.016		
fEp													**^{5}-0.34**	0.018		
Observations	1132		1132		1132		1130		1130		1132		1132		1132	
R^2 / R^2 adjusted	0.155 / 0.150		0.176 / 0.171		0.362 / 0.358		0.072 / 0.067		0.148 / 0.142		0.129 / 0.123		0.182 / 0.177		0.067 / 0.061	
Mean of sim. Adj. R2	0.0877		0.0928		0.0896		0.0778		0.0811		0.0798		0.0866		0.0834	
CDF (%)	98.3		99.4		100		31.5		99.2		96.3		100		18.0	

Table V: Measuring Instability of In-Sample Forward Selection Results

This table summarizes subperiod regressions for the 8 dependent variables at the 8-week horizon using stepwise forward selection described in Table IV. There are nine subperiods: 1998-04-01 – 1999-11-30, 1999-12-01 – 2002-07-31, 2002-08-01 – 2005-03-31, 2005-04-01 – 2007-11-30, 2007-12-01 – 2009-06-30, 2009-07-01 – 2012-02-29, 2012-03-01 – 2014-10-31, 2014-11-01 – 2017-06-30, and 2017-07-01 – 2020-03-31, defined as follows: we used NBER recession dating for the period 2007-12-01 to 2009-06-30; then we define post-crisis subperiods of roughly equal (32-month) length. The three pre-crisis subperiods that precede the crisis are also of 32-month length, while the initial (residual) subperiod is 20 months. The table reports pairs of values that represent the number(s) of subperiods a predictor was selected and had positive (left number) and negative (right number) coefficients. The pairs with at least one value greater than or equal to 3 are in brackets, bolded and underlined. The table also reports the average correlation for each dependent variable, computed as the average of correlation coefficients between all possible pairs of the 9 different subperiod regression coefficient vectors. The coefficient estimates of the unselected predictors are regarded as 0.

Predictor	*FutRet*	*DSpot*	*DOilVol*	*xomRet*	*bpRet*	*rdsaRet*	*DInv*	*DProd*
	+,-	+,-	+,-	+,-	+,-	+,-	+,-	+,-
FutRet		1,1	0,1	1,0	[**3**,0]	2,1		2,0
DSpot			0,2			0,1		0,2
DOilVol	1,2	[1,**3**]		2,1	1,0	1,1		1,1
StkIdx		1,0	0,1	1,1	0,1	1,1	0,1	1,1
DInv	1,1	0,1	0,1	[**3**,0]	[**3**,0]	[**3**,0]		2,0
DProd		0,1		0,1			1,0	
OilVol	[**3**,0]	[**3**,0]	[0,**9**]	2,1	1,0	1,0	0,1	2,1
VIX	1,1	1,1	1,1	[**3**,1]	[**4**,1]	[**3**,0]	1,0	[**3**,0]
DFX	2,0	[**3**,0]	1,1	[**3**,1]	1,0	1,0	0,2	0,2
tnote_10y	0,1	0,2	2,1	[0,**3**]	0,2	1,1	2,2	0,2
sp500Ret			1,0	1,0		1,1		
WIPI	2,0	1,0	1,0	[**3**,0]				
basis	[**3**,0]	2,0	[1,**3**]			1,0	1,0	
vix_diff		1,0	1,0	1,0	0,1	0,1	1,0	1,0
BE/ME	[**6**,0]	[**4**,0]	0,1	[**4**,0]	[**4**,1]	[**3**,1]	[**3**,0]	1,1
Mom	0,2	0,1	1,0			[0,**3**]	0,1	[**4**,1]
BasMom	0,1	0,1	1,0	1,0	1,1	1,1	0,2	2,0
DolBeta	0,2	1,1	1,2	1,1	[1,**3**]	0,2	2,0	1,2
InflaBeta	2,2	2,2		1,0	1,0	1,0	[2,**3**]	1,1
HedgPres	[**4**,0]	[**3**,0]	1,0		[**3**,0]	[**3**,0]	[0,**3**]	[1,**3**]
liquidity	0,1		1,0					0,1
OpenInt		0,1						
PCAsent	0,1		1,0		1,2		1,2	1,1
PCAfreq	1,0	1,0	0,1	0,1	1,1		0,2	
PCAall						0,1	1,1	0,1
artcount		0,2	2,0	2,1	1,0	1,0	0,1	0,1
entropy	2,0	2,0	1,2	2,1	2,1	1,0	1,1	1,0
sCo	1,1	1,1	[**3**,1]	0,1	0,1	0,1	1,0	
fCo			0,1	0,1	0,1	1,0	1,1	1,0
sGom	2,0	1,0	0,1	1,1	[2,**3**]	0,1		[**3**,0]
fGom	1,0		1,0		[0,**4**]	0,1	1,1	1,1
sEnv			1,0	1,0		2,0		
fEnv	0,2	0,1	0,1	0,2		2,1	0,2	1,0
sEpg	0,2		[**3**,0]	1,1	1,1	1,1	[1,**3**]	
fEpg	[**3**,0]	[**3**,1]	2,0	1,0		0,1	[1,**5**]	0,1
sBbl	1,0	2,0	1,0	1,1	1,0		0,2	[0,**3**]
fBbl	1,2	1,2	1,0	1,2	0,1	0,2	1,1	2,1
sRpc	[0,**3**]	0,2	1,0	1,1	[0,**3**]	0,1		1,0
fRpc		1,0	0,1	0,1	0,1	2,1	0,1	1,0
sEp	0,1	1,1	0,1			2,1	0,1	0,1
fEp	1,0	1,0	0,1	1,0	2,0	2,0	1,1	1,0
Avg. corr.	0.13	0.07	0.42	0.02	0.06	0.02	0.02	0.02

Figure 1. Word cloud plots for energy topics. This figure shows the word clouds of the energy topics extracted from the energy corpus using the Louvain clustering algorithm. Larger font indicates words that occur more frequently in a given cluster.



Company (Co)

Global oil market (Gom)

Environment (Env)

Energy/power generation (Epg)

Crude oil physical (Bbl)

Refining & petrochemicals (Rpc)

Exploration & production (Ep)

Figure 2. NLP measures over time. This figure shows time series plots of all text series. The series start in April 1998 and end in March 2020. We show 4-week averages of topical frequencies in Panel A, 4-week averages of topical sentiments in Panel B, and 4-week averages of article counts, unusualness (entropy) and the first principal components of normalized 4-week average textual measures in Panel C. The stars in Panels A and B mark the events detailed in Table III. The stars are positioned on the ending date of the time-period associated with the Table III episodes.

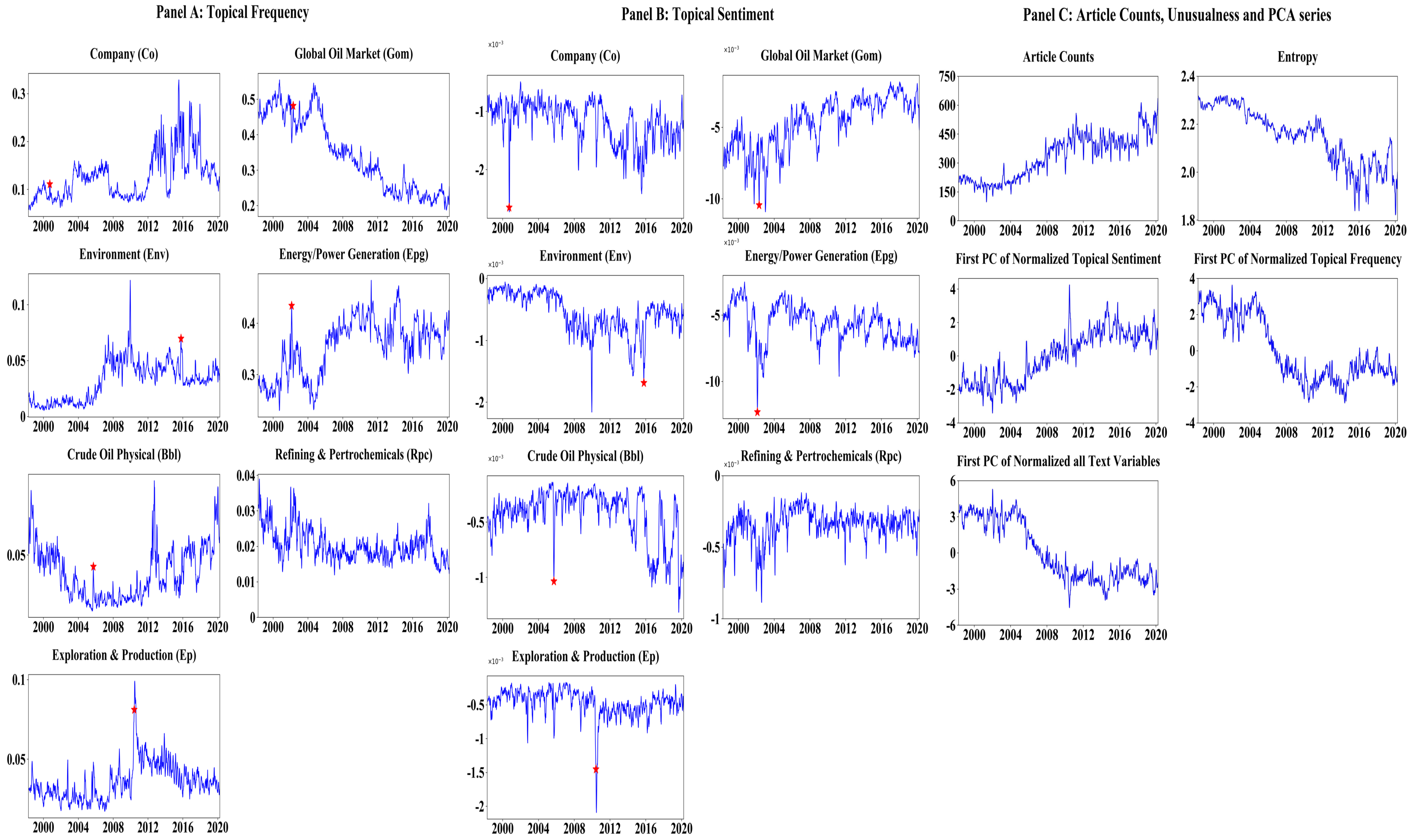
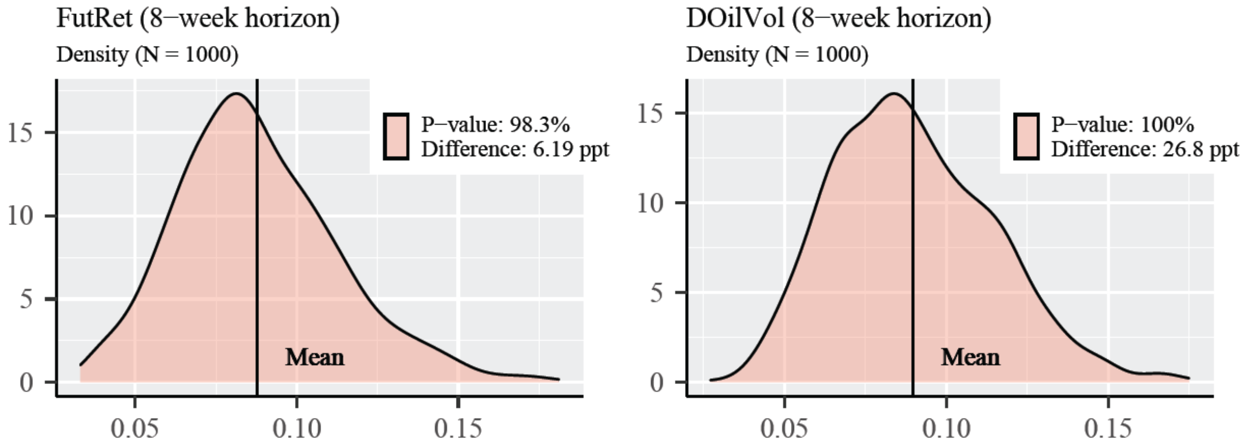


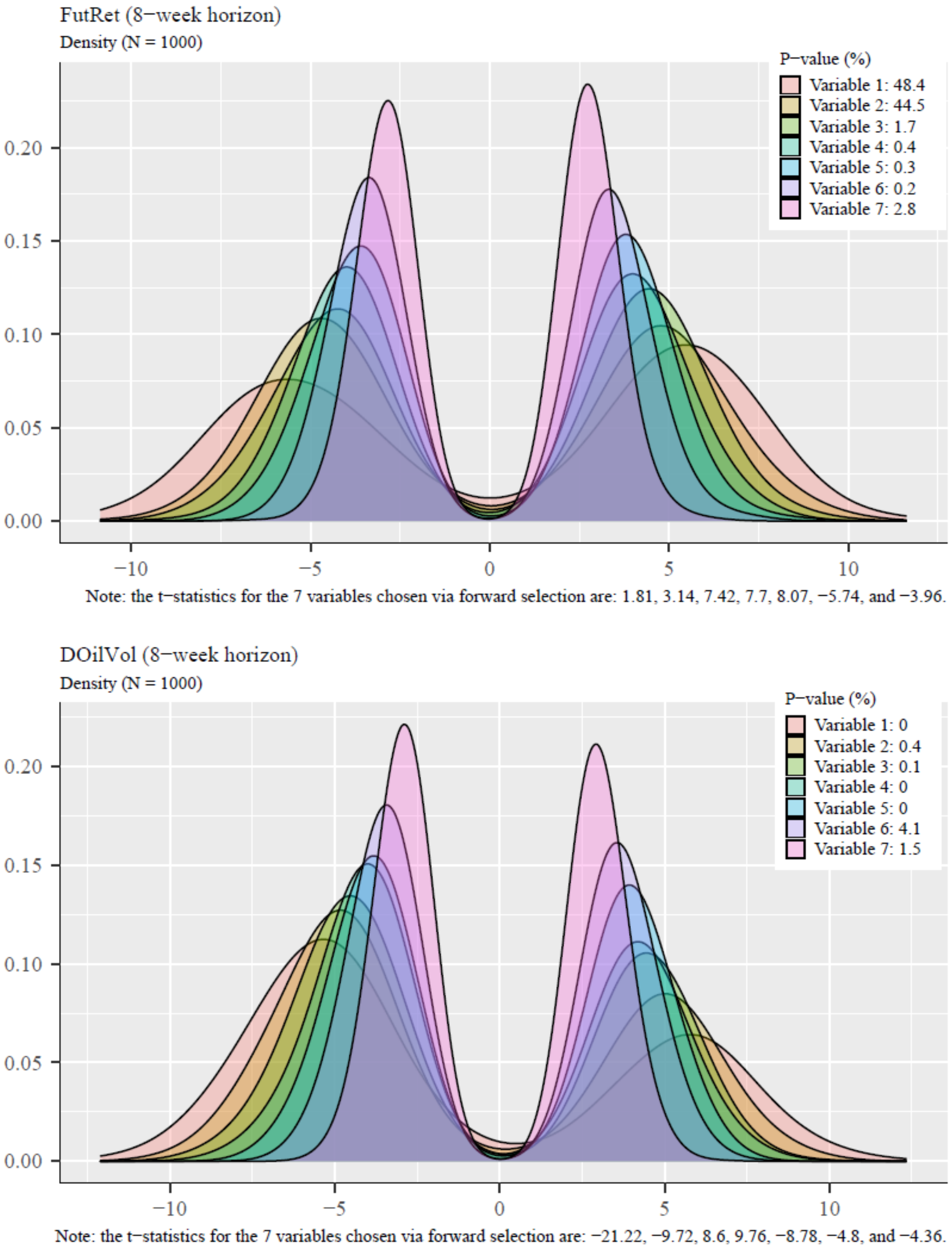
Figure 3. Monte Carlo simulations of adjusted R^2 for the eight-week oil futures returns (FutRet) and eight-week difference in oil volatility (DOilVol) models. We use forward selection to choose 7 variables as in-sample predictors of each dependent variable after the dependent variables have been defined as the residuals of regressions that include a time trend and the lagged value of that dependent variable. In forward selection models, therefore, the lagged dependent variable is not included in the list of selected candidates. The forward selection process includes all variables listed in Table I as candidate variables. In the bootstrapping simulations, raw values of RHS variables are used, while each LHS variable is simulated using an AR8 process. The figure shows the adjusted R-squared density function, and the p-value reported in the upper right corner of each figure measures the percent of the simulated adjusted R-squareds that are less than the empirical adjusted R-squared. The difference shown in the legend refers to the difference between the empirical adjusted R^2 and the mean of the adjusted R^2 simulations. The word "baseline" in the figure refers to the empirical foreword selection model. The Appendix presents a detailed overview of the bootstrapping process.



Note: the adjusted R−squared value in the baseline model is 14.96%.

Note: the adjusted R−squared value in the baseline model is 35.76%.

Figure 4. Monte Carlo simulations of t-statistics for the 7 variables chosen via forward selection for the eight-week oil futures returns (FutRet) and eight-week difference in oil volatility (DOilVol) models. Following the same bootstrap process outlined in the Appendix and used to produce Figure 3, we report here the density functions of the t-statistics for the simulated regression results. The p-value is computed as the minimum of the percent of simulated t-statistics less than the empirical t-statistic, and 1 minus the percent of simulated t-statistics less than the empirical t-statistic. In computing the p-values, we preserve the order of variables chosen in the empirical and bootstrap processes and compare the t-statistics in that order. The empirical t-statistics of the variables chosen via forward selection, in the order in which they were chosen, are listed in the notes below the figures. The p-values presented in Table IV are derived from this process for all variables.


FutRet (8−week horizon)

Density (N = 1000)

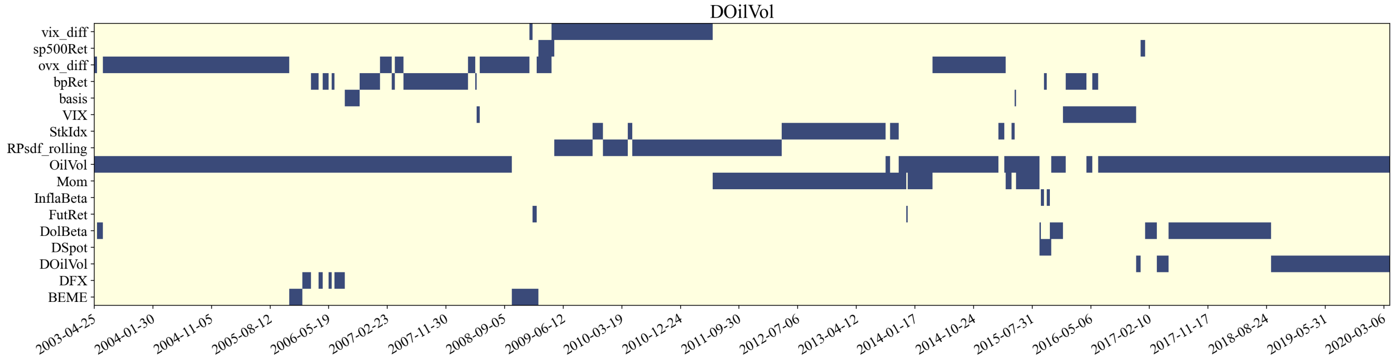
Note: the t−statistics for the 7 variables chosen via forward selection are: 1.81, 3.14, 7.42, 7.7, 8.07, −5.74, and −3.96.

DOilVol (8−week horizon)

Density (N = 1000)

Note: the t−statistics for the 7 variables chosen via forward selection are: −21.22, −9.72, 8.6, 9.76, −8.78, −4.8, and −4.36.

Figure 5. selected non-text and text variables in out-of-sample two-variable models. Selected non-text and text variables in the out-of-sample R^2 based 5-year lasso updating model. Top panel shows the time variation of the non-text variables composing the selected prediction model for *DOilVol*, and the bottom panel shows the time variation of the text variables composing the selected prediction model for *FutRet*. The y-axis lists the variables that at least enter the prediction model once during the whole selection process; the x-axis denotes the time of each forecast point. A blue block indicates the corresponding variable is selected in that prediction window.



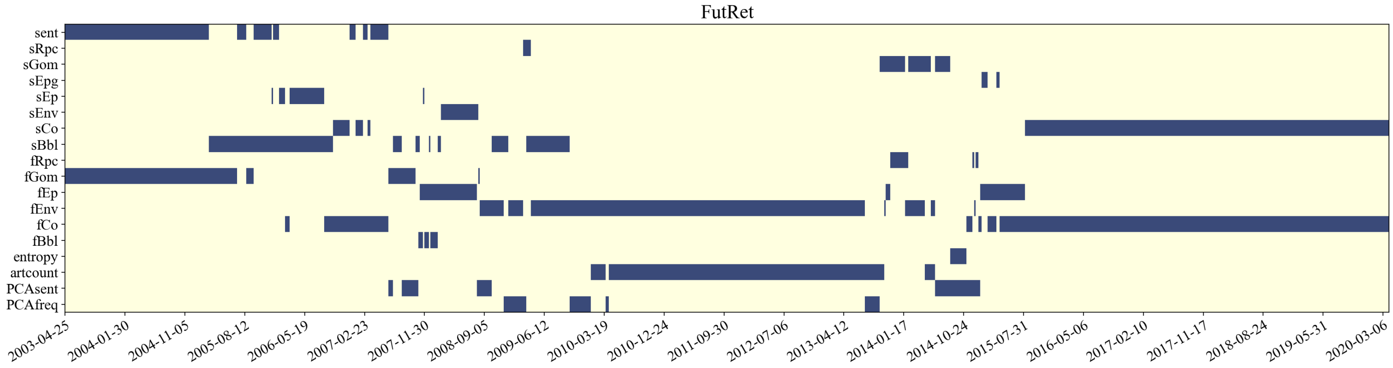
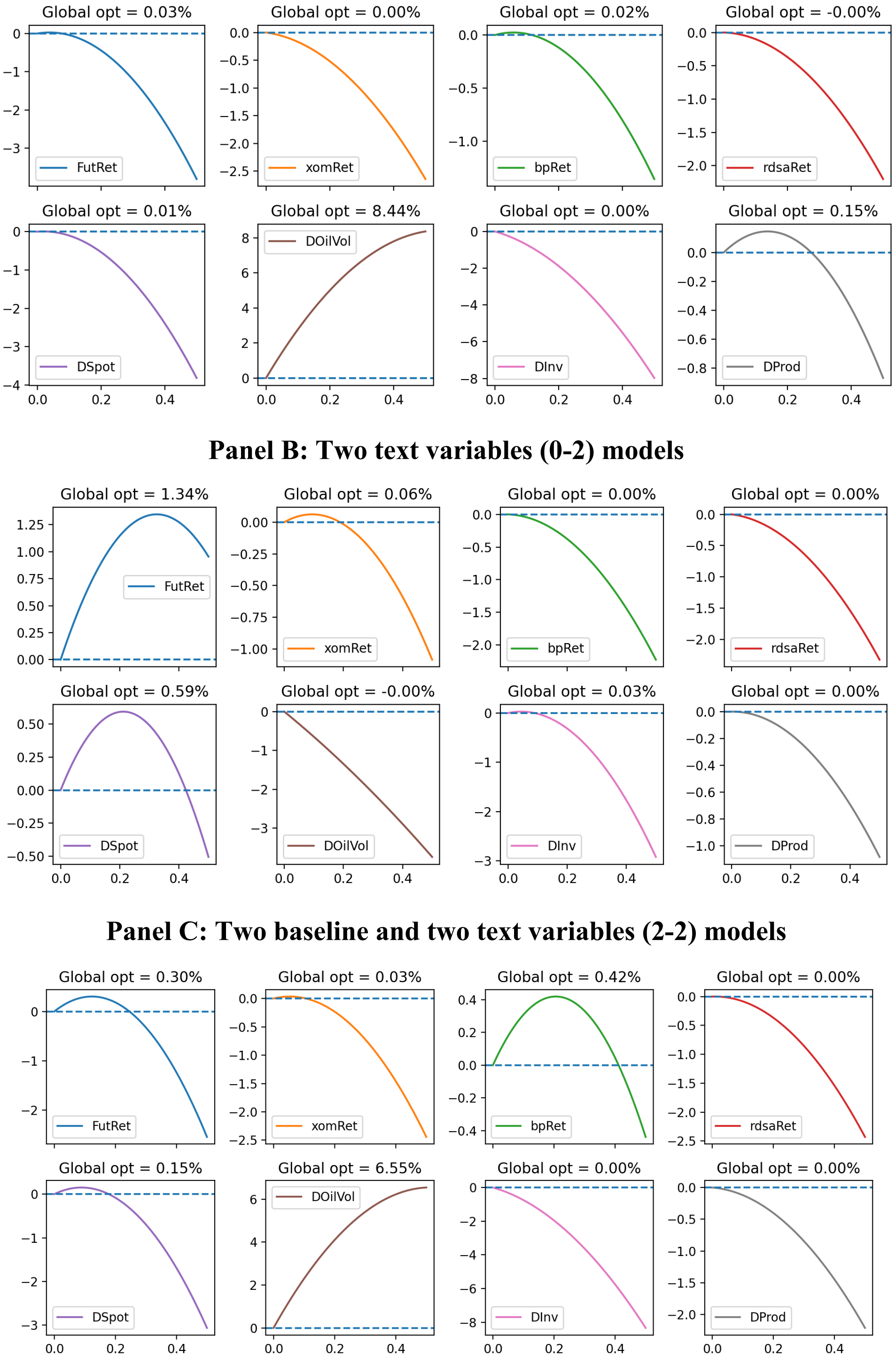


Figure 6: Out-of-sample R-squared comparisons for two-variable models. Panel A shows the $R^2_{OOS}(w)$ measure from equation (4) as a function of the blending weight w for two-variable baseline models for each of our eight dependent variables. Panel B shows the same analysis for two-variable text models. Panel C shows the same analysis for models including two baseline and two text variables. The detailed methodology is explained in Section 5. In all cases, the x-axis shows the weight w from (3).

Panel A: Two baseline variables (2-0) models



Panel B: Two text variables (0-2) models

Panel C: Two baseline and two text variables (2-2) models