ISSN 1936-5330

kcFED Research Working Papers

Predicting the Oil Market

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

December 2020; updated September 2021

RWP 20-20

http://doi.org/10.18651/RWP2020-20

This paper supersedes the previous version:

"Mining for Oil Forecasts"

Federal Reserve Bank of Kansas City

Predicting the Oil Market

Charles W. Calomiris, Nida Çakır Melek, and Harry Mamaysky[1]

Current version: September 2021
First version: December 2020

Abstract

We study the performance of many traditional and novel, text-based variables for in-sample and out-of-sample forecasting of oil spot, futures, and energy company stock returns, and changes in oil volatility, production, and inventories. After controlling for small-sample biases, we find evidence of in-sample predictability. Our text measures, derived using energy news articles, hold their own against traditional variables. While we cannot identify ex-ante rules for selecting successful out-of-sample forecasters, an analysis of all possible two-variable models reveals out-of-sample performance above that expected under random variation. Our findings provide new directions for identifying robust forecasting models for oil markets, and beyond.

Keywords: Asset Pricing, Commodity Markets, Energy Forecasting, Model Validation

JEL: C52, G10, G12, G14, G17, Q47

[1] 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. An earlier draft was circulated as "Mining for Oil Forecasts." We thank Christiane Baumeister, Kateryna Holland (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 IAAE Annual Meeting, the 4th Annual JPMCC International Commodities Symposium, and the Federal Reserve Bank of Kansas City for helpful comments and suggestions. Daliah Al-Shakhshir, Sungil Kim, and Hongyu Wu provided excellent research assistance for this study. Roya Arab Loodaricheh helped with the textual analysis and 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

Predictability of stock returns is a long-standing question in finance. Although different techniques, variables, or time periods may yield conflicting results, the consensus is that stock returns are, to some extent, predictable. With the increase in commodity investing and the emergence of commodities as a popular asset class since the early 2000s, interest in examining commodity return predictability has gained momentum. Among commodity markets, the oil market receives particular attention, not only among financial economists, but also among macroeconomists and general news outlets. This reflects its liquidity, its unparalleled importance as an input to production and consumption goods, its regional and geopolitical significance for U.S. states or foreign countries that produce and consume petroleum products in large quantity, as well as its importance in assessing macroeconomic risks.

In this study, we provide a comprehensive examination of the empirical performance of financial and physical oil market predictability. We introduce new predictive measures derived from energy news articles and consider a broad range of predictors suggested by academic research to forecast four- and eight-week ahead oil futures returns, oil spot returns, change in realized volatility of oil prices, the equity returns of oil companies, and changes in U.S. oil inventories and U.S. oil production for the period 1998-2020.[2,3] Our goal is to construct a fully transparent empirical methodology for considering a comprehensive list of potential forecasting variables and investigating their usefulness both in-sample and out-of-sample. We also study the stability of estimated coefficients over time for in-sample analysis and the persistence of our measures for out-of-sample forecasting.

We consider a wide range of explanatory variables – many of which have been used in prior studies – including macroeconomic and financial indicators, as well as various measures that capture time-varying oil

[2] While forecasting the real price of oil is a central question of interest in the oil-macro space, it is outside the scope 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. (2020).

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

returns risk.[4] Our first contribution 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 using the methodology of Hong and Yogo (2012). Following Asness et al. (2013), we construct a measure of value for oil prices. We also 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 2.1.

We construct a set of novel natural language processing (NLP) measures derived from the analysis of a corpus of oil news articles from Thomson Reuters (TR). 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 monthly 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 seems particularly promising. Indeed, we show that our textual measures – calculated utilizing an energy word list that we manually constructed for this paper – can algorithmically identify important historical episodes in energy markets, in a way that traditional energy variables are unable to do. Our NLP measures include topic-specific frequency and sentiment derived from energy news, as well as a measure of the unusualness or "entropy" of oil news (i.e., the frequency of occurrence of unusual word phrases). Topics are obtained from a corpus of TR articles using a network modularity approach, as in Calomiris and Mamaysky (2019a). The second contribution of this paper is the development of a rigorous text analysis methodology that captures important aspects of energy-related news flow, and should prove useful in future research. Indeed, industry-specific news corpora seems to contain information that is not

[4] A range of aggregate and commodity market-specific financial and macroeconomic variables used to predict commodity market outcomes are examined in Baumeister and Kilian 2017. Hamilton and Wu 2014 document significant changes in risk premia in crude oil futures contracts since the early 2000s.

present in traditional industry summary statistics, as our novel text measures hold their own in both in-sample and out-of-sample tests versus traditional predictors of energy market outcomes.

Several features distinguish our empirical approach from the literature: we begin with a comprehensive list of forecasting variables; our methodology for selecting variables is explicit; we use a bootstrap to adjust R-squareds and standard errors for overlapping observations and for our variable selection methodology; and we consider multiple approaches to out-of-sample validation of our models.[5] Therefore, our approach avoids reporting biases that are likely to arise when constructing forecasting models, and is transparent about the out-of-sample performance of the forecasting variables.

For example, a study might show the significance of a particular variable for forecasting, but does that variable prove significant if forced to compete for inclusion with a full range of other candidate variables? How should that variable's standard error be adjusted upward to reflect the fact that it was selected from a list of other variables because it was found to be a useful forecaster? And did the study in question report on all the other variables that were tried but that did not work? Additional reporting biases may arise from selective reporting of out-of-sample tests. For example, one could do an exhaustive search across many possible specifications to identify forecasting models that "work" *both* in-sample and out-of-sample, and only report in-sample results for models that pass this test. But such a search undermines the legitimacy of out-of-sample testing. Can one have confidence in any out-of-sample test that is reported simultaneously with the construction of an in-sample model? What out-of-sample validation technique can one use to provide a convincing validation?

Our approach takes explicit account of a broad set of possible modeling choices, both in our in-sample analysis and our out-of-sample validation. 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

[5] 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.

new variable as the one with the greatest incremental contribution to the model R-squared. Hastie, Tibshirani, and Tibshirani (2017) compare the forward selection methodology against two other machine learning approaches, best subset selection and the lasso regression. They find that forward selection is competitive with the other two methods. 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 both overlapping observations and for the effects of forward selection. 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 nth 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. Hence, this approach is informative about the pitfalls that arise from 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 nth chosen variable. Although this point has been made before, see for example Welch and Goyal 2008, we think it is particularly salient for energy forecasting given the lack of cross-sectional data.

After controlling for all of the above issues, we find evidence of robust in-sample predictability from both our text and non-text measures for the sample period as a whole. This validates past predictability results with regard to many of our forecasting variables, and establishes that our novel text-based measures are at least as powerful as non-text predictors of energy market outcomes. 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 stability.

Subperiods were specified in advance of running any models 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, in general, few regressors are chosen in the forward-selection model across multiple subperiods, suggesting a fair amount of model instability.

Then, we consider various approaches to out-of-sample testing. In a time series context, such as ours (as panel models are not generally appropriate for oil forecasting due to its globally integrated market), parsimonious models are attractive. Nevertheless, even when adopting a parsimonious modeling discipline by selecting only a small number of potential forecasting variables, for most of our dependent variables, it proves difficult to construct a methodology for systematically identifying, in real time, a set of forecasting variables that work well out-of-sample.

However, we identify additional candidates for parsimonious oil forecasting models by resorting to a brute force search over all possible forecasting models. When estimating loadings on pairs of forecasting variables using rolling lasso regressions, we find many variable combinations with successful quasi out-of-sample forecasting performance in subperiods of the data. This is *quasi* out-of-sample, because while the coefficients in the forecasting model are chosen using only ex-ante data, the choice of which pair of forecasting variables to select is not known ex-ante. We analyze whether success in quasi out-of-sample forecasting in the past is an indication of true out-of-sample forecasting success in the future by constructing a distribution for the incidence of successful out-of-sample forecasting runs of a given length under the null of no persistence. We find evidence of this: variables that show quasi out-of-sample forecasting outperformance in one subperiod are more likely to be successful true out-of-sample forecasters in the next one or two subperiods than they would be by chance. We stop short of declaring success in the ex-ante identification of forecasting variables that will prove successful because we do not identify a systematic strategy for choosing future successful pairs of forecasting variables. The evidence is nonetheless encouraging.

A. Related Literature and Our Contribution

Robust out-of-sample performance by forecasting models in finance is, in general, hard to come by. For example, 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. 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 investigating in-sample and out-of-sample performance of novel NLP measures as well as 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. They usually propose new predicting variables or modifications of existing predictors, and examine whether these improve predictability in commodity markets. These studies are typically based on in-sample analysis with baseline models that contain six or seven predictors. Our approach not only considers a wide range of financial and macro variables, including many variables considered in this literature as well as new text measures, but we also test predictability in oil markets both in-sample and out-of-sample. In addition, we carefully control for small-sample biases.

Finally, 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 brought new tools to bear in examining this question, including the analysis of various aspects of language that appear in newspaper articles or other textual sources, which have been applied 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). In this context, our work also relates to recent work using textual analysis in analyzing the oil market (Brandt and Gao 2019, Datta and Dias 2019, Loughran et al. 2019 and Plante 2019). Our approach differs from those studies by constructing and considering a more comprehensive set of NLP measures.

This paper makes several methodological contributions. We construct novel NLP measures for energy markets and examine their usefulness 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, both of which are challenging and potentially 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 – of subjecting in-sample analysis to out-of-sample tests – considers how to perform convincing out-of-sample tests that are not subject to reporting bias. We conjecture that an important factor which causes in-sample models to fail out-of-sample is model instability: today's forecasting variables may not be tomorrow's forecasting variables. To take both reporting bias and model instability into account, we derive a distribution for the number of successful forecasting runs of a certain length under the null of no out-of-sample persistence. We then systematically analyze all possible, parsimonious forecasting models from a given set, and show that these models exhibit out-of-sample forecasting persistence beyond what may be expected purely by chance. This suggests that past winning out-of-sample models may continue to be future winning out-of-sample models. We believe many of these methodological contributions may prove useful for financial market analysis more broadly.

The remainder of the paper proceeds as follows. Section 2 presents the list of non-text forecasting variables we consider, and describes our data sources and our methods for constructing them. Section 3 presents our text analysis and new NLP measures included in the models. Section 4 explains our choice of in-sample modeling structure, which involves a forward selection model using overlapping weekly observations to forecast eight-week ahead energy market outcomes. Section 4 also discusses our methodology for correcting standard errors and R-squareds for variable selection bias and for overlapping observations, and presents our in-sample results, for the sample period as a whole as well as for various subperiods. Section 5 presents our out-of-sample analysis and runs tests. Section 6 concludes. Our code, energy word list, and text measures are available at https://github.com/hmamaysky/Energy.

2. Data and Construction of Variables

We consider a variety 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 TR news articles about the energy sector. The raw data used to construct the variables in our analysis 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 in detail 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. We also tried forecasting dependent variables four weeks ahead, except for realized oil volatility because of timing issues explained below, and found the results to be qualitatively similar. To conserve space, the four-week results are reported in the Online Appendix.

A. Timing Conventions

Our dependent variables – oil spot and futures returns, company stock returns for oil majors, change in realized oil volatility, and changes in oil production and in oil inventories – represent crucial information for investors, policymakers, and analysts, as they seek to understand the dynamics of oil markets. Our explanatory variables include lags of the dependent variables, as well as 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 arriving in the market and market reactions to that information. 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 Eastern time (ET). For some weeks, typically those involving holidays, releases are delayed by one or two days. This feature of the data release process drives the timing convention for our empirical analysis. When forecasting changes in inventories and production, which we refer to as the physical – as opposed to price-based – regressions, our dependent variables become available after 10:30am ET on Wednesdays, and occasionally later. When forecasting price-based series (futures and spot returns, change 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 price-based dependent variables from Friday of week *t* to Friday of week *t+8*.

Oil spot and futures prices are available at 2:30pm ET on each trading day and the oil majors' stock prices are available after 4pm ET. One of the oil majors we consider, Royal Dutch Shell (RDS), trades in Europe and its Friday close occurs in the morning ET. So, for RDS, we calculate its forward return from Monday of week *t+1* to Monday of week *t+9*. We measure the price-based dependent variables on Fridays (or the next Monday for RDS) to make sure there is no overlap with the physical data releases in week *t*. For example, in some weeks physical data releases can be delayed by one or two days, and may come out on either Thursday or Friday at 10:30am ET. Therefore, starting our dependent variable measurements from Friday at 2:30pm ET or later is conservative and ensures no overlap between the dependent price-based series and the physical oil forecasting variables. Our oil volatility measure, *DOilVol*, is the change in the trailing 30 trading-day realized volatility of WTI prices from the Friday of week *t* to the Friday of week *t+8*. Hence, this dependent variable has no overlap with any week *t* explanatory variable. However, we cannot run four-week ahead forecasting regressions with this variable due to overlap with the four-week lagged explanatory variables. Given the data release schedule for oil inventories and production, our physical (Wednesday) and price-based (Friday or Monday) timing convention for dependent variables is the natural choice for any analysis of U.S. oil markets that involves both physical and price-based data.

Since we have two timing conventions for our dependent variables, we need to have two versions of all explanatory variables as well: those that are used for the physical forecasting regressions, and those that are used for the price-based forecasting regressions. For the physical regressions, we use values for variables that are the most recently available ones on the Tuesday of week *t*. This ensures no overlap with the physical dependent variables which are released on Wednesday of week *t* at the earliest. For the price-based regressions, we can use the latest value of a forecasting variable that is not after 2:30pm ET on the Friday of week *t*, otherwise there would be overlap with the forward-looking oil spot and oil futures returns. For example, the week *t* Friday S&P 500 return would overlap with the oil spot and futures dependent variable for that week, because the S&P 500 return is measured as of 4pm ET on Friday. In all such cases, we use the Thursday end of period (EOP) value of the forecasting variable. In cases where the right-hand side variable does not overlap with the Friday dependent variables, such as lagged futures returns, we use the Friday value of the forecasting variable.

To summarize, 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. In our analysis, variables can appear either as independent or dependent variables in either the physical or price-based regressions. We list the day of the week for each dependent and explanatory variable based on its use context in the Online Appendix.

Occasionally, a dependent or explanatory variable observation is not available in a given week *t*, often due to holidays. In this case, we would be unable to calculate either the forward-looking week *t* dependent variable or the backward-looking week *t* forecasting variable. When a week *t* price-based dependent variable is missing its Friday level, we calculate forward returns using either the Monday or Tuesday (if Monday is missing) price point of week *t+1* (for RDS, a missing Monday observation would be replaced with either the Tuesday or Wednesday one). For the price-based dependent variable regressions, when a right-hand side variable is missing its week *t* observation and is measured as of Thursday, we use either the week *t* Wednesday or Tuesday value; if the right-hand side variable is measured as of Friday, we use either the week *t* Thursday or Wednesday value. For the physical regressions, when we are missing a week *t* Tuesday value for an explanatory variable, we use the week *t* Monday or week *t-1* Friday value instead. These rules allow us to avoid missing a dependent or independent variable due to a single missing observation, while avoid overlaps between explanatory and dependent variables.

B. Dependent and Explanatory Variables Involving Energy, Macro, and Financial Data

For obtaining oil returns, we consider the U.S. oil benchmark, West Texas Intermediate (WTI). Our measure of j-week spot price returns is ln (P_{t+j}/P_t), where t is measured in weeks and $j = 4, 8$. We use the nearest-to-maturity futures price as the spot price, consistent with most other studies of commodity futures, as commodity spot markets are frequently illiquid. While modeling spot returns is useful for capturing the dynamics of oil price changes, spot price changes do not represent an investable return because they ignore storage and transportation costs. To capture investable oil price 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 at the end of the week). We construct j-week cumulative returns as the product of the past j one-week returns. This measure captures the returns to a specific investment strategy, and reflects changes in spot prices, the realization of risk premia, and changes in risk premia over time.[6]

In a similar vein, 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 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 the Euronext exchange. 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. Therefore, our eight dependent variables are eight-week changes in oil production and in oil inventories, oil spot and future returns, eight-week changes in oil realized volatility, and the stock returns of BP, Royal Dutch Shell, and ExxonMobil.

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

[6] Further details on calculating futures returns are available in the Online Appendix.

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* (BP and ExxonMobil are the Thursday-to-Thursday returns, and RDS is Friday-to-Friday for the price-based regressions; and all are measured Tuesday-to-Tuesday for the physical regressions). This energy stock index is a less noisy forecaster than individual company returns, and we were unable to find an existing energy stock index that extends as far back 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*, 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 basis indicates the curve in contango, and all other things being equal buying longer-dated futures will lose money as they roll down the curve. A negative basis indicates backwardation, and if the curve remains fixed, buying longer dated futures will earn positive returns as they roll up the curve. Following Asness et al. (2013), we calculate a book-to-market-type 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*.

In addition, we 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 (defined as the month-over-month change in the year-over-year CPI inflation) and on changes in the log dollar index, respectively.[7] 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 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, measured in dollar terms; more precisely it is the log of the WTI price times the contract size times the total open interest at the end of each trading day. We use the measure for *liquidity* proposed in Amihud, Mendelson, and Lauterbach (1997), and 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. The underlying data for *OpenInt* and *liquidity* are obtained from Bloomberg.

We refer to the variables defined in this subsection as our *baseline (or non-text)* measures. Table I presents definitions for all variables used in the empirical analysis. 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.

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

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).[8] Assuming m is in the linear span of the security returns implies

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

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. 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

[8] 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.

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.

We use three different estimation methods for $\widehat{m}_t$, which differ in the time period used to estimate $E[RR^\top]^{-\top}$. In all three cases, we use the $\hat{E}_t R^e_{WTI,t+1}$ 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.[9] Our corpus for NLP analysis includes all 2.07 million articles in Thomson Reuters (TR) that are labeled as being energy-related 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 related* if it is classified by TR as belonging to one of 98 energy topics, the full list of which is in the Online Appendix.

[9] 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.

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 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.[10] 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 TR 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 between 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 TR 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 two 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 represent energy topics in the TR

[10] Our full list of sources are as follows: Dahl (2004), Downey (2009), Geman (2005), Griffin and Steele (1986), Griffin and Teece (1982), Raymond and Leffler (2006), Yergin (1992), Yergin (2011), as well as Deutsche Bank (2013), Devold (2013), the IEA Oil Market Report glossary, and the Platts energy industry glossary, covering common terms and abbreviations from the oil, power, petrochemicals, nuclear, gas, coal and metals markets. Thanks to Mine K. Yucel and David Rodziewicz for their source suggestions.

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 and is 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.[11] 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.

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,

[11] 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.

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 the 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, which will be discussed further in the next subsection. 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.[12]

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

[12] 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.

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

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. 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 the sentiment s_i of each article.

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

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

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 $t-4$. 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. For the jth 4-gram $w_1 w_2 w_3 w_4$ (the w_k's refer to words or tokens), $\widehat{m}_j$ is the fraction of times w_4 follows the word sequence $w_1 w_2 w_3$ in the training corpus, or

$$\widehat{m}_j = \frac{\hat{c}(w_1 w_2 w_3 w_4)}{\hat{c}(w_1 w_2 w_3)},$$

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.[13] Glasserman and Mamaysky (2019) 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.

[13] 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).

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 physical and price-based dependent variables, respectively. Similarly, we calculate the average number of daily articles about energy markets in the TR corpus 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 (labeled *s[Topic]*). We standardize all text-based 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.

In addition to these, we also add three 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*). The PCAs are calculated using the four-week averages of the weekly series, where the four-week averages have been normalized to be mean zero and unit variance.

A. Behavior 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 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 of interest that 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 dependent energy-market variables.

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.[14] 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 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,

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

entropy, the seven topical frequency series, the seven topical sentiment series, and the three PCAs (19).[15] 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*. We residualize because the lagged dependent variable would otherwise frequently be 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.[16]

The model is estimated using weekly observations with either four- or eight-week ahead overlapping observations, which substantially 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 tends to introduce upward bias in the R-squareds, and downward bias in the standard errors as well. To control for both of 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 in more detail.

[15] 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.

[16] The number of variables considered in studies examining return predictability in commodity markets ranges between one and seven (see Table 1 in Baumeister and Kilian 2017). In our out-of-sample analysis, we also consider a two-variable in-sample selection model.

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 are anomalous relative to the simulated R-squareds. We first estimate an AR(8) process for the dependent variable. We then simulate a new dependent series based on the AR(8) model. Next, we rerun our in-sample forward-selection and regression models, using all of the actual 41 forecasting variables, except replacing the lagged dependent variable with the simulated series. By construction the simulated dependent series is independent of all our forecasting variables, except for the lagged dependent series itself which controls for the mechanical autoregressive properties of the dependent variable. 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 both the selection and overlapping observation properties of our in-sample procedure. More details are in the Appendix.

To give a sense for 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. Under the null hypothesis of no predictability, outside the mechanical autoregressive properties of both series, there is a wide range of R-squareds in our simulated runs. In fact, the dual small-sample problems of overlapping observations and variable selection lead to very high in-sample R-squared. 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.

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.[17] Recall the forward selection model is applied to detrended variables, that are then orthogonalized by regressing out the lagged four-week dependent variable. 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

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

standard deviation change in the dependent variable due to a one standard deviation change in the forecasting variable.[18]

The standardized coefficients for the selected predictors range between 0.04 and 0.67 in absolute value. For example, a one standard deviation increase in ten-year treasury note yield (*tnote_10y*) over the previous month lowers eight-week ahead BP returns by 1.3% $(0.13 \times 10.01\%)$. Or, a one standard deviation increase in average $sGom$ over the past month – positive sentiment about global oil markets – increases oil futures returns by 3.9% $(0.28 \times 13.78\%)$ over the next eight weeks. Moreover, around 50% (28/56) of the selected variables are statistically significant, even after adjusting for overlapping observations and variable selection.[19] 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 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 the null hypothesis of no relationship between the dependent and the independent variables (except for the presence of the lagged independent variable), the probability of adjusted R-squareds being greater than or equal to the empirical R-squareds reported is less than 0.7% 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, 30 of them are text measures (about 54 percent), and of these 17 are statistically significant. Recall that 19 of the 41 candidate explanatory variables are text-based measures. For example, four of the text measures that are chosen statistically significantly at least two times are *PCAall, entropy, sGom,* and *sEnv.* In fact, all

[18] 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.

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

dependent variables except ExxonMobil returns and changes in oil production and oil inventories 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 only about 46 percent of the selected predictors (26/56). In addition, only 42 percent of the selected non-text variables are statistically significant (11/26), which is lower than the 57 percent of the selected text variables that were significant (17/30). For example, only two of the selected non-text variables are chosen statistically significantly two times: *tnote_10y* for futures returns and for BP returns, and *Mom* for 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. We conclude, therefore, that our text measures are powerful in-sample predictors for the oil market. 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.[20]

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 market outcomes are robustly forecastable, below 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 choice of subperiods is explained in the introduction, and as we emphasize there, the 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 also 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 positive forecaster of both *FutRet* and *DSpot* in four subperiods. The book-to-market measure (*BE/ME*) appears as a 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

[20] 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 20 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.

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 parsimonious subset of forecasting variables must be chosen from a larger candidate set. Second, the coefficient loadings on this subset must be estimated. Neither of these steps should use forward-looking data. We consider two out-of-sample approaches. Our first 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. The second approach conducts a brute force search over all possible two-variable forecasting models, and checks their out-of-sample performance and stability over subperiods. In this approach, coefficient loadings are estimated using only historical data, but we do not (yet) take a stand on which of all possible pairs of forecasting variables is a good choice. We augment our set of in-sample forecasting variables from Section 3 with the two SDF-based expected return forecasts (*sdf_rolling, sdf_growing*), as well as *ovx_diff* and *sent*, which is the sum of the four-week topical sentiments in a given week and is spanned by our seven topical sentiment series. 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. This leaves 48 forecasting variables, 20 of which are text and 28 of which are non-text.[21]

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 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, we then form three forecasting models. The first contains only the baseline variable with the highest R-squared in the training window; this is the 1-0 model. The second model contains only the text variable with the highest training window R-squared, the 0-1 model. Finally, the third contains both the top baseline and text variables, the 1-1 model. All three models are formed using ex-ante information only.

For each of the three models, we then estimate a lasso regression in the training window. We run rolling five-year lasso regressions with automatic penalty parameter selection using ten-fold cross validation, to estimate rolling coefficients for out-of-sample forecasting of eight-week ahead changes or returns of our dependent variables. In all lasso regressions, in addition to the forecasting variables under consideration, we also include a constant and the four-week version of the lagged dependent variable to control for mechanical autocorrelation properties of the data. 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

[21] The following summarizes the number of variables in different parts of our analysis:

- In-sample predictors (41): 22 non-text and 19 text variables.
- 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.
- 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.

regime, and not allow for structural breaks in the forecasting relationships. We ensure each five-year training window uses data only from inside the window.[22] Using the lasso coefficient estimates in each training window, we then use the dependent 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 and re-estimate the model, and make another eight-week ahead forecast.

We also explore a variation of the above approach in which we choose 2 non-text variables (the 2-0 model), or 2 text variables (the 0-2 model), or the two top non-text and text variables (the 2-2 model), where all variables are ranked by their training window R-squareds within their peer variable set (i.e., text relative to text, and non-text relative to non-text). This variation allows for the possibility that more than one forecasting variable from each set may be useful, and is again completely ex-ante.

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 a baseline model. We are again careful to make sure that the averages of eight-week changes used for what we call the *constant model* do not extend outside of the five-year training window. Our main test is to look at the ratio of the mean-squared-error (MSE) of our lasso models relative to the mean-squared-error of the constant model, i.e.

$$MSE\ ratio = \frac{MSE_{lasso}}{MSE_{constant}}.$$

Note that the out-of-sample R-squared is simply one minus this ratio. Therefore, when this ratio is above one, i.e., when the lasso model generates a higher error variance than the constant model, the out-of-sample R-squared will be negative.

For each of our eight-week ahead dependent variables, Figure A.2 in the Online Appendix shows which forecasting variable is chosen at each point in our sample in the 1-1 model. Two features are notable. First,

[22] 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.

there is some persistence in which variable is chosen over time, though this is not surprising given our use of overlapping five-year training windows. Second, despite this persistence there is a large amount of time series variation in the selected non-text and text variables. This suggests that allowing for time variation in the forecast variable set may be useful.

Unfortunately, as Table VI shows, this approach to out-of-sample forecasting does not yield a model that performs better than the constant. Panel 1 shows the performance of 0-1 and 1-1 models relative to the constant model as measured by the MSE ratio. The MSE ratio is above one in all the 16 cases. When we move to the 0-2 and 2-2 models in Panel 2, the performance is similar, as 15 out of 16 cases have above one MSE ratios. The only dependent variable with some out-of-sample success is *DOilVol* when forecasted with the 2-2 models. Overall, our variable selection methodology does not lead to out-of-sample outperformance relative to the simple constant benchmark.

We also explore the relative value of text variables in forecasting compared to non-text variables. Here, we consider 1-0 and 2-0 models as our benchmark to test the incremental value of our text measures in out-of-sample forecasting relative to using the non-text forecasting variables only. We find that relying on text measures does improve forecasting results significantly. Panel 3 of Table VI shows the ratio of the MSE of models 0-1 (text only) and 1-1 (non-text and text) relative to the model 1-0 (non-text only). The results are now much more positive. In many cases, we find that the MSE ratio of the models that include text measures are lower than the MSE ratios of the non-text models. This is true also for the two text and non-text variable models as shown in Panel 4 of the table. Apparently, adding text-based information to non-text forecasters of energy market outcomes is useful, but it is not sufficiently useful to make forecasting models outperform the forecasts of the constant model in out-of-sample tests.

B. Successful Out-of-sample Forecasting Models Using a Fixed Pair of Forecasters

We have attempted to devise a transparent and systematic methodology that identifies a subset of successful out-of-sample forecasters from a large pool of potential forecasters. Our lack of success relative to the constant model suggests that it may be very difficult to select, in real-time, a model that beats the random walk for out-of-sample forecasting of returns and volatility in the oil market as well as oil production and inventory changes. But it does not mean that there are no combinations of forecasting variables that can beat the constant model; merely, that our technique for variable selection and coefficient updating does not seem to be capable of identifying those combinations. We now ask whether there are *any* combinations of forecasting variables that appear to have successful out-of-sample performance. To address this question, we undertake a brute-force search over all possible forecasting combinations of two variables. In other words, we use data mining to identify combinations of variables that turn out to be "successful" in out-of-sample forecasting. For a given pair of forecasting variables, we estimate a rolling lasso model (to choose time-varying coefficient values for the two fixed variables) in a five-year training window using the approach described in the previous subsection. Using these rolling coefficient estimates, we then make an out-of-sample eight-week ahead forecast, then move the estimation window forward by one week, and repeat the analysis. For a particular pair of fixed forecasting variables, this is fully an out-of-sample analysis, where the coefficients used to forecast the outcome variable do not use any forward-looking information. What makes this approach only a quasi out-of-sample analysis is that we analyze *all* pairs of forecasting variables to find pairs that succeed as out-of-sample forecasters.

There are 48 forecasting variables in our study, and therefore there are $\binom{48}{2} = 1{,}128$ potential models.[23] Table A.IX of the Online Appendix shows that the majority of potential models cannot outperform the constant model in out-of-sample forecasting when evaluated over the entire data set. In light of the finding in Table V that in-sample forecasting variables are not stable over time, the inability of most two-variable models to outperform the constant model over the entire sample is not very surprising. Taking a cue from the in-sample analysis, rather than checking whether a given model outperforms the constant over the entire sample, we instead check whether a given model outperforms the constant model in different subperiods. For our out-of-

[23] To fit 1,128 lasso models, each of which requires cross-validation to estimate the coefficient penalty parameter, requires running the code in parallel on dozens of nodes on Columbia's research computing grid. We did not repeat this analysis for a larger number of forecasting variables because of the enormous computational burden of doing so. There are 48 choose 3 = 17,296 three variable models, and this would take close to a month of runtime on a highly parallel computational grid.

sample analysis, we employ the same subperiods we used for our in-sample analysis. As already mentioned, the subperiods were fixed beforehand and were not adjusted to influence our results. Since the out-of-sample analysis needs an initial five-year window to estimate the first lasso model, we lose the two early subperiods from our in-sample analysis, resulting in seven rather than nine subperiods. So, the first subperiod in which we have out-of-sample results is from 2003-04-25 to 2005-03-31, and there is a total of seven out-of-sample subperiods. Given the seven subperiods, for each dependent variable, each model can outperform the constant model from zero to seven times.

Table A.X on the Online Appendix is analogous to Table A.IX, but shows out-of-sample test performance by subperiod. There are now far more winning forecasting models than in Table A.IX, where success requires a pair of variables to outperform the constant over the *entire* sample. For example, for *FutRet* Table A.IX shows that 26 out of 1128 two pair models outperformed the constant over the whole sample, whereas Table A.X shows that roughly 10 times as many models outperform the constant in out-of-sample *FutRet* forecasting in any given subperiod. The same general pattern applies to all the dependent variables. This suggests instability in out-of-sample forecasting models, implying that the variables that successfully forecast energy market outcomes change over time. The challenge then is how to identify future successful pairs of forecasting variables. One approach is to check whether certain forecasting variables consistently show up in winning forecasting models. In addition, one can check whether a successful forecasting model in one period is likely to remain a successful model in the next period.

To check the consistency of our forecasting variables, we count the number of forecasting models that outperform the constant model in three or more *consecutive* subperiods. We refer to these models as *consistent* models. We choose three consecutive subperiods to gauge out-of-sample performance, because we believe that it would be asking too much of one pair of forecasting variables to generate successful out-of-sample forecasting performance for more than three or four consecutive subperiods out of seven, given the fundamental changes that occurred during our sample period that are related to technological changes (e.g., fracking boom) or to the financial crisis of 2007-2009.

For each dependent variable, Table VII shows the number of consistent models containing a given forecasting variable.[24] The column labeled *Total* shows the total number of consistent models in which a given forecasting variable appears for all dependent variables. The rows are sorted by the *Total* column. For example, for *DOilVol,* 42 pairs including the predictor *DSpot* outperform the constant model in three or more consecutive subperiods. Given that there are only 47 models that contain *DSpot* (since there are 48 forecasting variables), the maximum possible number of consistent models containing *DSpot* is 47. That 42 of these had runs of at least length three suggests a great deal of out-of-sample consistency. Similarly, the forecasting variable *FutRet* appears in 41 consistent models. The other prominent forecasting variables in consistent models for *DOilVoil* are *rdsaRet*, *OilVol*, and *bpRet*.

The top two predictors based on the total number of consistent models in which they appear are lagged values of *DSPot* and *FutRet*, due largely to their out-of-sample forecasting power for *DOilVol*, the dependent variable with the largest number of successful full period out-of-sample forecasting models (Table A.IX in the Online Appendix). The next best predictor is *rdsaRet* which appears prominently for *DOilVol* and *DProd* (apparently stock returns of this particular oil major are very informative for future oil production). Of the next ten top forecasting variables, six are text variables and four (*DInv*, *WIPI*, *sp500Ret*, *liquidity*) are non-text. Text variables are well represented in successful instances of model-subperiod out-of-sample outperformance.

We emphasize that missing from this quasi out-of-sample analysis is a rule for identifying which model will successfully forecast a given dependent variable in a future subperiod. Given that one can always employ data mining to identify a successful fixed model for out-of-sample forecasting, is there any way to validate the "successful" fixed models as reflecting something more than random success?

Figure 5 addresses this challenge by explicitly testing for the non-randomness of successful models based on their persistence across subperiods. For a given dependent variable and subperiod, a forecasting model can be in one of two states: negative (higher MSE than the constant model) and positive (lower MSE than the constant model). Figure 5 shows the transition probabilities for these states for two dependent

[24] This table only has 46 rows because *ovx_diff* and *trend* are not consistent forecasters for any dependent variable.

variables, *FutRet* and *DOilVol*, in each subperiod. For each state transition, we calculate three probabilities: one for all two-variable models, one for two-variable models containing at least one text variable, and one for two-variable models containing no text variables. For example, for *FutRet* the bottom right panel shows the probability that a positive model in subperiod *t-1* remains a positive model in subperiod *t*: roughly 40% of text models (i.e., models that contain at least one text variable) that were successful out-of-sample forecasters for *FutRet* in subperiod (6) remain successful out-of-sample models in subperiod (7). The bottom right panel for *DOilVol* shows that close to 80% of all successful out-of-sample models involving at least one text variable in subperiod (6) remain successful forecasting models in subperiod (7). Interestingly, for both *FutRet* and *DOilVol,* the worst positive to positive transition probability occurs from subperiod 3, which is the Global Financial Crisis, to subperiod 4. As may be expected, models that worked during a very stressed subperiod may no longer work when the economy transitions to calmer economic times.

For both *FutRet* and *DOilVol,* the negative-to-negative transition probability is around 80% in most periods, suggesting that high MSE models tend to remain high MSE models in subsequent subperiods. To conserve space, we do not show the transition probability results for the remaining six dependent variables, but they are available in the Online Appendix and are qualitatively similar to the ones presented here. The takeaways from this analysis are that successful models in subperiod *t-1* tend to remain successful in period *t*, and that there is an even stronger tendency for unsuccessful models in subperiod *t-1* to remain unsuccessful in subperiod *t*. We now present a methodology that allows us to establish a null hypothesis for how much persistence in model success may be expected, and to test whether the observed persistence is in line or in excess of the null hypothesis.

C. Successful Forecasting Runs

In many statistical applications involving successful or unsuccessful trials, it is possible to assess whether the observed sequence of trial successes and failures is consistent with the null hypothesis of independence. In our context, we are particularly interested whether a successful out-of-sample model in one subperiod is more likely to remain a successful out-of-sample model in the next subperiod than would be expected by chance. If so, then a good candidate for a successful future forecasting model would be a model that was successful in out-of-sample forecasting in the recent past. How far back the "recent past" extends depends on the persistence of successful models. Consider a particular forecasting model being tested for out-of-sample forecasting success relative to the constant model for a given dependent variable in each of the seven out-of-sample subperiods. This model-subperiod combination may be characterized as a length-7 string of 0s and 1s, where 1s indicate outperformance by the forecasting model relative to the constant in a given subperiod. For example, assume that *FutRet* is the dependent variable, and the model consisting of the 10-yr Treasury yield (*tnote_10y*) and the frequency of energy and power generation news (*fEpg*) beats the constant model in out-of-sample tests in subperiods 1, 4, 5, and 6. Then, the corresponding string would be "1001110".

Our approach to testing for better-than-random out-of-sample performance focuses on *persistence*. For a given dependent variable, we analyze how frequently different runs of successes happen across adjacent subperiods across all candidate models. In the example case of the string 1001110, there is one run of length 1, and one run of length 3. A *run* is defined as a series of successful trials at the start, middle or end of a given string of length 7. A run at the start of the string is a series of 1s followed by a zero; a run in the middle of a string is a sequence of 1s surrounded by 0s on either side; and a run at the end of the string is a sequence of 1s preceded by a 0. Note that the use of runs in finance goes back to Fama (1965) who analyzes runs of price changes of the same sign; though there may be even earlier applications of which we are not aware.

For a given dependent variable, we have 45 forecasting variables with data in each of the seven subperiods, of which 20 are text variables (see the footnote at the start of Section 5). This means that for each dependent variable, we have $\binom{45}{2} = 990$ trials, each of which is a string of length 7 indicating the success of the given model in each of the seven out-of-sample subperiods. There are $25 \times 20 + \binom{20}{2} = 690$ models which contain at least one text variable. Our particular interest is to see how many of these 990 trials or 690 trials have at least one run of length k for $k = 1,2,3,\ldots,7$. For example, the top panel of Table VIII shows that for *FutRet*, out of the 990 total models, 695 have at least one run of length 1, 181 have at least one run of length 2, 34 have at least one run of length 3, and there is just a single instance of a run of length 4.

To gauge our findings against random success we must ascertain what number of trials for *FutRet* were expected to have runs of length *k* under the null of no persistence. To address this question, we first need to determine the likelihood that a given model will outperform the constant model in out-of-sample forecasting in a given subperiod. There are a total of $990 \times 7 = 6{,}930$ model-subperiod observations, and of these for *FutRet* 21.1% had successful out-of-sample performance of the forecasting model relative to the constant model. This is shown in the q row in the top panel of the table. Given an individual trial success probability of q, we can then determine the probability of observing at least one run of length *k* is a string of seven *independent* trials. Rather than constructing a structural model to assess the time series properties of the dependent and forecasting variables that are consistent with the observed q, we instead take q as given and then ask whether the observed number of runs of length *k* is likely to have occurred under independence, or whether the observed number of runs is indicative of persistence. This is a nonparametric test of persistence that is agnostic to the time series properties of the underlying data.

A run of length 7 has a probability of q^7, and a run of length 6 has a probability of $2q^6(1-q)$, which is the probability of 6 1s at the start of the string, followed by a zero, plus the probability of a zero, followed by 6 1s. A run of length 5 has a probability of $2q^5(1-q) + q^5(1-q)^2$, which is the probability of two corner runs and one middle run. Having a run of length 4 has probability $2q^4(1-q) + 2q^4(1-q)^2$, which is two corner runs and two middle runs. Calculating the probabilities of having at least one run of length 3, 2, or 1 is more complex, because in these cases a string of length 7 can have more than one run, e.g., "1110111" has two runs of length 3. In this case, when we calculate the probability of an ending corner run "…0111", we need to adjust for the probability of not having seen a prior run of length 3 to avoid double counting. These probabilities are calculated recursively, and the formula is shown in the Appendix. We validated our calculations by constructing Monte Carlo simulations that obtain the same probabilities of observing randomly generated runs.

We refer to the probability of observing at least one run of length k in a string of length 7 given a probability of success in a single trial of q as $\Pr[q, k, 7]$.

With this, the distribution of the number of strings of length 7 which have at least one run of length k given either 990 or 690 independent draws is given by the binomial distribution with size of either 990 or 690, and a probability of a successful trial given by $\Pr[q, k, 7]$. The top panel of Table VIII shows the number of models which utilize all forecasting variables, for which we have data in all seven subperiods, that have at least one run of length $k \in \{1,2,3,4,5\}$. There are no runs of length 6 or 7 in our sample. The p-value below each of the run counts gives the probability of observing strictly more than this number of runs of length k under the assumption of independence across the 990 or 690 draws, and across the seven trials for a given draw. These counts and p-values are given for each of the eight dependent variables in our analysis.

For *FutRet*, *DSpot*, *DOilVol*, and *bpRet* the p-values for length-1 runs are high, suggesting that we observe too few runs of length-1 relative to independence. For the other four dependent variables, on the other hand, we observe too many length-1 runs. For example, for *xomRet* the p-value for the 828 models with at least one length-1 run is 0.00, suggesting a far higher number of length-1 runs than is expected to occur by chance. Turning to runs of length 2, we see that for five of the eight dependent variables, *DSpot, DOilVol*, and the three oil major stock prices, the number of length-2 runs is much higher than would be expected under independence. Also, the number of length-3 runs is anomalously high for four of the dependent variables, *DSpot*, *DOilVol*, *DInv*, and *DProd*. However, the number of length-4 and length-5 runs is, in most cases, far lower than would be expected under independence as all p-values are close to 1.

In summary, the evidence consistently points towards greater persistence in out-of-sample performance over two or three subperiods for the majority of our dependent variables. The flipside of this is that there are fewer runs of lengths 4 and 5 than might be expected under independence. We interpret the absence of length-4 and length-5 runs as an indication that long-term structural instability (due possibly to technological changes and the financial crisis) makes it infeasible for forecasting models to do well for more than three consecutive subperiods. Interestingly, *FutRet*'s p-values are all between 0.23 and 0.97, suggesting that out-of-sample forecasting performance for *FutRet* is consistent with a 21.1% probability of successful model outperformance but zero persistence. Therefore, past successful models for *FutRet* are no more likely to outperform the constant than past unsuccessful models.

We also investigate the relative usefulness of text variables as persistent forecasters in our successful fixed models. The bottom panel of Table VIII shows a similar analysis for the 690 models that include at least one text variable. The results are qualitatively quite similar to those in the top panel. The probabilities q of a successful model-subperiod outcome are almost identical for the text models, suggesting that text variables are equally successful out-of-sample forecasters as non-text variables. To sum up, the main takeaways of the out-of-sample analysis of all two variable forecasting models are: (1) there are many cases in which models outperform the constant in some subperiods; (2) such outperformance tends to be more persistent over two or three subperiods than it would be by chance; (3) the runs test we develop has power against the random-predictability null; and (4) text variables are equally successful out-of-sample predictors as non-text variables. These suggest that a deeper analysis of which models are persistent in which subperiods and for which dependent variables is an important area for further study.

6. Conclusion

Predicting financial and physical oil market outcomes is a challenging task, especially given that the globally integrated oil market implies the need to employ time series, rather than country panel, regression analysis. Furthermore, the time period for which data are available is relatively short, and even if it weren't, regime shifts would undermine the usefulness of much of the earlier data. Traditional in-sample estimation approaches suffer from an implicit variable selection bias, although researchers typically do not formally adjust for this. Taking our cue from recent findings that text-based measures are useful for predicting returns and risk in equity markets, we construct novel text-based measures for the energy market. Adding these to a long list of standard and relatively more recent forecasting variables, we formally model the variable selection problem for forecasting energy market outcomes, and control for selection bias in our R-squared evaluation and in our standard errors. We find systematic patterns of in-sample predictability. Many of the successful in-sample forecasting variables are derived from measures based on the text of Thomson Reuters news articles about the energy space.

Despite successful in-sample predictability, we find it difficult to identify a systematic strategy for finding forecasting variables that lead to out-of-sample performance that is better than a rolling mean of the dependent variable. However, an analysis of all possible two variable forecasting models identifies many successful out-of-sample models across subperiods. We present strong statistical evidence of persistence of out-of-sample outperformance by documenting a tendency of successful forecasting models to have outperformance runs of two to three subperiods.

This paper contains several methodological contributions that are useful for energy markets forecasting, and that should prove useful for more broad financial market applications. These include our code and word lists, our NLP methodology, our forward selection methodology, as well as the associated bootstrap procedure that can be used for inference and for tabulating R-squared distributions. The results on both in-sample and out-of-sample model instability also are likely to be salient for other – not only energy – markets. Finally, our runs analysis provides a powerful test of out-of-sample forecastability and should be useful in non-energy settings as well.

What remains unclear is whether a systematic variable selection method exists that will identify successful out-of-sample forecasting models. Although we do not propose a particular forecasting approach based on our current results, our findings could be used to construct a forecasting model now, and then one could test the predictions of that forecasting model in a true out-of-sample analysis, say, two or three years in the future. We regard this as the next step in our analysis. How would one build a model today to make use of our findings? Note that our findings of persistence imply that the set of all fixed models that beat the constant can be thought of as comprised of two types of models: those that randomly beat the constant and those that beat the constant because of true out-of-sample forecasting power. It appears that at least some of the models are of the latter type. A "meta model" to forecast any dependent variable could be constructed as an average of all the successful fixed model forecasts, and that meta model should out-perform the constant. Rather than giving equal weight to all the successful fixed models we identify when constructing the meta model, one could give greater weight to successful models that worked for the past two consecutive subperiods. One could also give greater weight to models that include variables that tend to reappear frequently in successful forecasts of more than one dependent variable. For example, looking at Tables V and VII, and focusing on the dependent variables *FutRet*, *DSpot* and *DOilVol* (which should be closely related structurally), we note that some variables appear more frequently than others in the list of useful forecasting variables for these three dependent variables in the two tables. It would be reasonable to give models that contain those variables greater weight when building a meta model.

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," *The Journal of Finance*, 68 (3), 929-985.

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. 2020. "Energy Markets and Global Economic Conditions," *Review of Economics and Statistics*, forthcoming.

Baumeister, Christiane, and Lutz Kilian. 2015. "Forecasting the Real Price of Oil in a Changing World: A Forecast Combination Approach," *Journal of Business and Economic 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*.

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

Datta, Deepa, and Daniel Dias. 2019. "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.

Deutsche Bank Markets Research. 2013. “Oil & Gas for Beginners: A Guide to the Oil & Gas Industry.”

Devold, Havard. 2013. “Oil and gas production handbook: An introduction to oil and gas production, transport, refining and petrochemical industry,” *ABB Oil & Gas.*

Downey, Morgan. 2009. “Oil 101,” Los Angeles: Wooden Table Press.

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

Foster, F Douglas, Tom Smith, and Robert E. Whaley. 1997. “Assessing Goodness-of-Fit of Asset Pricing Models: The Distribution of the Maximal R-Squared,” Journal of Finance, 52 (2), 591-60.

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.

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,” *Proceedings of the ACM International Conference 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.

Griffin, James M., and Henry B. Steele. 1986. “Energy Economics and Policy,” *Orlando: Academic Press*.

Griffin, James M., and David J. Teece. 1982. “OPEC Behavior and World Oil Prices,” *London: George Allen & Unwin*.

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 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.

International Energy Agency. Oil Market Report Glossary available at https://www.iea.org/oilmarketreport/glossary/

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.

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.

Raymond, Martin, and William L. Leffler. 2006. “Oil and Gas Production in Nontechnical Language,” *Tulsa: PennWell.*

S&P Global Platts Glossary available at https://www.platts.com/glossary

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,” *Journal of Financial Economics*, 110, 164-184.

Yergin, Daniel. 1992. “The Prize: The Epic Quest for Oil, Money & Power,” *New York: Free Press*.

Yergin, Daniel. 2011. “The Quest: Energy, Security, and the Remaking of the Modern World,” *New York: Penguin Group.*

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*
Nontext 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

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

Panel A: Nontext 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 variables 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 returns (*xomRet, bpRet, rdsaRet*) from our regressors and instead include *StkIdx* (which is the average of the three stock returns). Only predictors that were chosen by at least one model are included in this table. Coefficients are standardized using the ratios of the standard errors of the dependent and predicting variables. Superscripts before coefficients indicate order in forward selection (1=chosen first). The p-values are obtained using Monte Carlo simulations that use an AR8 process to simulate the LHS variable, as well as forward selection to produce both adjusted R^2 and t-statistic simulations. The p-values refer to the 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 comparison is relative to 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 the 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 significance 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**	0.00									^{3}-0.07	0.45
DInv							^{2}0.10	0.37			^{5}0.11	0.12				
OilVol					^{1}**-0.67**	0.00										
tnote_10y	^{7}**-0.11**	0.048							^{7}**-0.13**	0.01			^{6}-0.13	0.09		
sp500Ret																
WIPI					^{6}-0.12	0.053										
basis			^{1}0.20	0.13					^{6}**-0.14**	0.02						
BE/ME	^{1}0.19	0.26					^{6}0.12	0.15							^{7}-0.12	0.07
InflaBeta											^{7}**-0.13**	0.01			^{4}0.09	0.42
HedgPres							^{7}0.08	0.21					^{7}**-0.11**	0.03		
Mom	^{6}-0.14	0.06	^{5}**-0.16**	0.047					^{1}**-0.29**	0.00	^{4}-0.13	0.12				
VIX			^{4}-0.04	0.48	^{3}**0.29**	0.01									^{1}0.12	0.39
vix_diff									^{5}**-0.14**	0.03						
PCAall	^{5}**0.25**	0.00	^{7}**0.25**	0.00												
entropy	^{3}**0.23**	0.00	^{3}**0.27**	0.00	^{5}**-0.28**	0.00	^{5}0.14	0.11	^{3}**0.18**	0.03	^{1}**0.36**	0.01	^{1}-0.24	0.052		
fCo					^{7}**-0.14**	0.00					^{3}0.16	0.23				
sGom	^{4}**0.28**	0.00	^{6}**0.25**	0.00					^{4}**0.19**	0.01	^{2}0.17	0.10	^{3}-0.09	0.44		
fGom					^{4}**0.26**	0.00									^{2}-0.12	0.36
sEnv									^{2}**0.22**	0.01	^{6}**0.13**	0.049				
sEpg							^{4}0.14	0.10								
fBbl	^{2}-0.20	0.08					^{1}-0.16	0.28							^{6}0.16	0.06
sRpc															^{5}0.07	0.50
fRpc			^{2}0.05	0.50									^{2}**-0.25**	0.01		
sEp							^{3}-0.09	0.39					^{4}**-0.36**	0.01		
fEp													^{5}**-0.35**	0.01		
Observations	1122		1122		1122		1120		1120		1122		1117		1117	
R^2 / R^2 adjusted	0.172 / 0.167		0.176 / 0.171		0.361 / 0.357		0.074 / 0.068		0.149 / 0.144		0.130 / 0.125		0.186 / 0.181		0.068 / 0.062	
Mean of sim. Adj. R2	0.0794		0.0829		0.0795		0.0669		0.067		0.0677		0.0774		0.0755	
CDF (%)	99.9		99.8		100		55.3		100		99.3		99.9		29.3	

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 (negative) coefficients. The pairs with at least one value greater than or equal to 3 are in brackets, in bold, and are 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		1,0	[**3**,0]	2,1		2,0
DSpot	0,1		0,2			0,1		0,2
DOilVol	[1,**3**]	[1,**4**]		2,1	1,0	1,2	0,1	0,1
StkIdx		1,0	0,1	1,0	0,1	1,1	0,1	1,0
DInv	0,1	1,1	0,1	2,0	[**3**,0]	[**3**,0]		[**3**,0]
DProd		0,1		0,1			1,0	
OilVol	[**4**,0]	[**4**,0]	[0,**9**]	[**3**,1]	1,0	2,0	2,1	2,1
VIX	1,1	1,1	2,1	[**3**,1]	[**4**,1]	[**3**,0]	1,0	[**3**,0]
DFX	2,0	[**3**,0]	2,0	[**3**,1]	1,0	1,0	0,2	0,2
tnote_10y	0,1	0,2	2,2	[0,**3**]	0,2	1,2	2,2	0,2
sp500Ret			1,0			1,0	0,1	
WIPI	2,0		1,0	2,0				0,1
basis	1,0	2,0	[1,**3**]	1,0			1,0	
vix_diff	1,0	1,0	1,1		0,1	0,1		1,0
BE/ME	[**6**,0]	[**4**,0]	1,1	[**4**,0]	[**4**,1]	2,1	2,0	1,1
Mom	0,2	0,1	1,0	0,1		[0,**3**]	0,1	[**4**,1]
BasMom		0,1	1,0	1,0	1,1	2,1	0,2	2,0
DolBeta	0,2	1,1	1,1	1,1	[1,**3**]	0,2	1,0	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]	2,0	0,2	[0,**3]**
liquidity	0,1							0,1
OpenInt		0,1						
PCAsent	0,1		1,1		1,2		0,2	2,1
PCAfreq	1,0		0,1	0,1	1,1		0,2	
PCAall						0,1	1,1	0,1
artcount		0,2	1,0	2,2	1,0	1,0	0,2	0,1
entropy	[**3**,0]	2,0	[2,**3**]	[**3**,1]	2,1	1,0	1,1	1,1
sCo	1,1	1,1	[**3**,1]	0,1	0,1	0,1		
fCo			0,1	1,0	0,1	1,0	1,0	1,0
sGom	2,0	1,0	0,1	2,1	[2,**3**]	0,2	1,0	[**3**,0]
fGom	1,0		1,0		[0,**4**]	0,1	2,1	1,1
sEnv						2,0		
fEnv	0,2	0,1	0,1	0,2		2,1	0,2	1,1
sEpg	0,2	0,1	[**3**,1]	1,1	1,1	1,1	[1,**3**]	
fEpg	**[3**,0]	2,0	1,0	2,0			[0,**5**]	
sBbl	1,0	2,0	1,0	1,1	1,0	0,1	1,2	0,2
fBbl	1,2	1,2	1,0	1,1	1,1	0,2	0,1	2,1
sRpc	[0,**3**]	[0,**3**]	1,0	1,1	[0,**3**]	0,1		1,0
fRpc		1,0	0,1	0,1	0,1	2,0	0,2	1,0
sEp		1,0				2,1	0,1	0,1
fEp	1,0	1,0		1,0	1,0	2,0	1,1	1,0
Avg. corr.	0.16	0.11	0.40	0.02	0.06	0.04	0.03	0.01

Table VI
Out-of-Sample Forecast Accuracy of Parsimonious (1-1 and 2-2) Lasso Models

The table displays the MSE ratio of the out-of-sample 1-1 and 2-2 Lasso updating model against benchmark models. Each week, an n-n (where n=1 or 2) Lasso model selects n text and n non-text predictors separately. In each week, we run univariate regressions for each potential predicting variable with a 5-year lookback to determine which variables to include in our parsimonious Lasso models. We rank the potential text and non-text variables separately for each week by their R^2 in these univariate OLS regressions. Then we choose each week the top n text or non-text variable (s) to form that date's forecast model. Each week the model updates the coefficients of the predictors using a Lasso model loss function using the rolling 5-year lookback window to predict the dependent variable eight weeks ahead. Each predicting variable is defined by the last day's observation in each five-year window. There are two benchmark models: the Constant model and the Non-Text model, with specifications annotated in the table as *0 nontext + 0 text* or *n nontext + 0 text* respectively. Alternative models are the Text model and the Full model, with specifications as *0 nontext + n text* or *n nontext + n text*. The MSE for a model is produced once all the weekly forecasts are calculated. After the MSE calculations, the MSE ratio is determined by dividing the MSE of an alternative model by that of a benchmark model. Boldface indicates superior performance relative to the benchmark.

Panel A: Full Model specification: 1 Nontext + 1 Text				
	(1)		(3)	
Benchmark Model	Constant *0 nontext + 0 text*		Nontext *1 nontext + 0 text*	
Alternative Model	Text *0 nontext + 1 text*	Full *1 nontext + 1 text*	Text *0 nontext + 1 text*	Full *1 nontext + 1 text*
Out-of-Sample R2 Relative to Benchmark with respect to Each Predicted Variable				
FutRet	1.053	1.097	**0.967**	1.008
xomRet	1.032	1.047	**0.989**	1.004
bpRet	1.062	1.059	1.015	1.012
rdsaRet	1.044	1.053	**0.997**	1.006
DSpot	1.050	1.070	**0.981**	**0.999**
DOilVol	1.034	1.007	1.031	1.004
DInv	1.075	1.133	**0.962**	1.014
DProd	1.006	1.028	**0.978**	**0.9995**

Panel B: Full Model specification: 2 Nontext + 2 Text				
	(2)		(4)	
Benchmark Model	Constant *0 base + 0 text*		Nontext *2 base + 0 text*	
Alternative Model	Text *0 nontext + 2 text*	Full *2 nontext + 2 text*	Text *0 nontext + 2 text*	Full *2 nontext + 2 text*
Out-of-Sample R2 Relative to Benchmark with respect to Each Predicted Variable				
FutRet	1.022	1.074	**0.945**	**0.994**
xomRet	1.028	1.054	**0.981**	1.006
bpRet	1.043	1.030	1.011	**0.998**
rdsaRet	1.043	1.049	**0.999**	1.006
DSpot	1.037	1.076	**0.963**	**0.999**
DOilVol	1.043	**0.998**	1.056	1.011
DInv	1.063	1.129	**0.946**	1.005
DProd	1.022	1.054	**0.982**	1.013

Table VII
Frequency of Variables Present in the Consistent Out-of-Sample Forecasting Models

This table counts the number of times a variable appears in a two-variable out-of-sample fixed model with a run of length 3 or greater for a given dependent variable. Given that there are 48 forecasting variables, the maximum number of times a variable can appear in a model with at least a length-3 run is 47. The 3rd subperiod is the 2007-2009 NBER recession. Subperiods 4 to 7 are divided evenly after the 3rd subperiod. The length of the 2nd subperiod matches the ones after the 3rd. The first subperiod is the residual. The seven subperiods are 2003-04-25 – 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 – 2020-01-31. The numeric values of the five variables that show up the most are in bold with brackets. We rank the variables allowing for ties, so there are some columns that have more than five values shown in bold with brackets.

Predictor	*FutRet*	*DSpot*	*DOilVol*	*xomRet*	*bpRet*	*rdsaRet*	*DInv*	*DProd*	*Total*
DSpot	1	3	**[42]**	3	0	0	2	1	52
FutRet	1	2	**[41]**	1	1	0	1	2	49
rdsaRet	0	1	**[27]**	6	3	1	1	**[10]**	49
sCo	4	**[5]**	6	4	**[8]**	**[10]**	7	3	47
fEnv	4	4	4	**[8]**	**[10]**	5	8	3	46
DInv	0	1	4	1	3	**[17]**	**[12]**	3	41
fCo	1	1	9	4	4	3	8	**[11]**	41
sEnv	**[7]**	**[5]**	3	4	**[10]**	2	7	3	41
WIPI	2	**[7]**	7	**[14]**	3	3	3	1	40
sp500Ret	0	1	4	**[22]**	0	**[9]**	2	0	38
entropy	3	**[7]**	7	**[9]**	4	4	0	3	37
fBbl	2	2	4	1	**[7]**	**[8]**	**[10]**	0	34
liquidity	0	0	6	6	3	5	7	6	33
basis	0	0	6	6	**[7]**	3	7	3	32
fEp	1	0	5	**[7]**	2	5	6	6	32
sdf_growing	2	3	13	1	0	1	**[11]**	1	32
HedgPres	3	3	2	2	3	4	2	**[12]**	31
OilVol	**[5]**	4	**[22]**	0	0	0	0	0	31
tnote_10y	**[8]**	**[8]**	1	2	1	0	8	2	30
bpRet	1	1	**[18]**	5	1	0	2	1	29
DProd	1	**[6]**	3	4	2	2	**[11]**	0	29
InflaBeta	0	0	1	2	0	0	**[26]**	0	29
PCAfreq	2	4	5	2	4	0	3	**[8]**	28
sRpc	**[5]**	3	7	0	3	7	3	0	28
fRpc	0	1	4	2	**[7]**	1	7	5	27
StkIdx	1	1	15	5	2	1	1	1	27
artcount	2	1	7	2	6	1	2	5	26
fGom	**[5]**	2	7	3	3	1	3	2	26
BE/ME	0	0	0	6	2	**[10]**	2	5	25
DolBeta	**[5]**	1	3	1	**[7]**	0	7	1	25
OpenInt	1	2	3	4	0	3	6	6	25
sEp	2	2	4	2	1	1	8	4	24
PCAsent	0	1	7	0	1	1	9	4	23
sEpg	0	0	7	4	2	2	7	1	23
DFX	0	0	11	2	0	1	3	2	19
Mom	1	3	3	0	0	0	0	**[12]**	19
sGom	1	0	5	6	1	1	3	2	19
fEpg	2	0	6	2	2	3	1	2	18
sBbl	0	1	4	4	1	4	4	0	18
xomRet	0	0	4	3	4	2	1	2	16
BasMom	0	0	4	4	0	2	3	2	15
sent	0	0	2	4	0	7	1	0	14
PCAall	0	1	5	1	0	1	3	2	13
vix_diff	0	1	3	1	0	3	3	1	12
DOilVol	1	0	6	0	0	0	1	0	8
sdf_rolling	0	0	3	0	0	0	0	0	3

Table VIII
Out-Of-Sample Analysis of Runs

We evaluate every combination of two forecasting variables in each of seven test periods. A *run of length k* is defined as *k* consecutive periods in which a model consisting of a pair of forecasting variables outperforms the constant model out-of-sample (has a positive out-of-sample R-squared). The *k* consecutive periods of outperformance must either be at the start of the seven-period window and be followed by a period of underperformance, or be in the interior of the seven-period window and be surrounded by two periods of underperformance, or be at the end of the seven-period window and be preceded by a period of underperformance. Table columns correspond to one dependent variable. The *Runs* row shows the number of variable pairs that have at least one run of any length. The *Max Runs* row shows the maximum possible number of runs. The *q (All)* row shows the sample probability of any pair of forecasting variables outperforming the constant model in any of the seven test periods. The *# All/Txt* row shows, respectively, the number of all variables and the number of text variables for which data are available in each of the seven subperiods. The *Run[k]* row shows the number of variable pairs with at least one length-*k* run. The *Run[k]-p* row shows the probability that strictly more than *Run[k]* runs would have been observed under the null of independent draws from a binomial distribution of size *Max Runs* with a probability of a successful outcome given by $\Pr[q, k, 7]$, the probability of seeing at least one run of length *k* in a string of length 7 given independent draws with probability of success given by q. The calculation of $\Pr[q, k, 7]$ is explained in the text. The top panel shows the results for all forecasting variables; and the bottom panel shows results for models that contain at least one text variable.

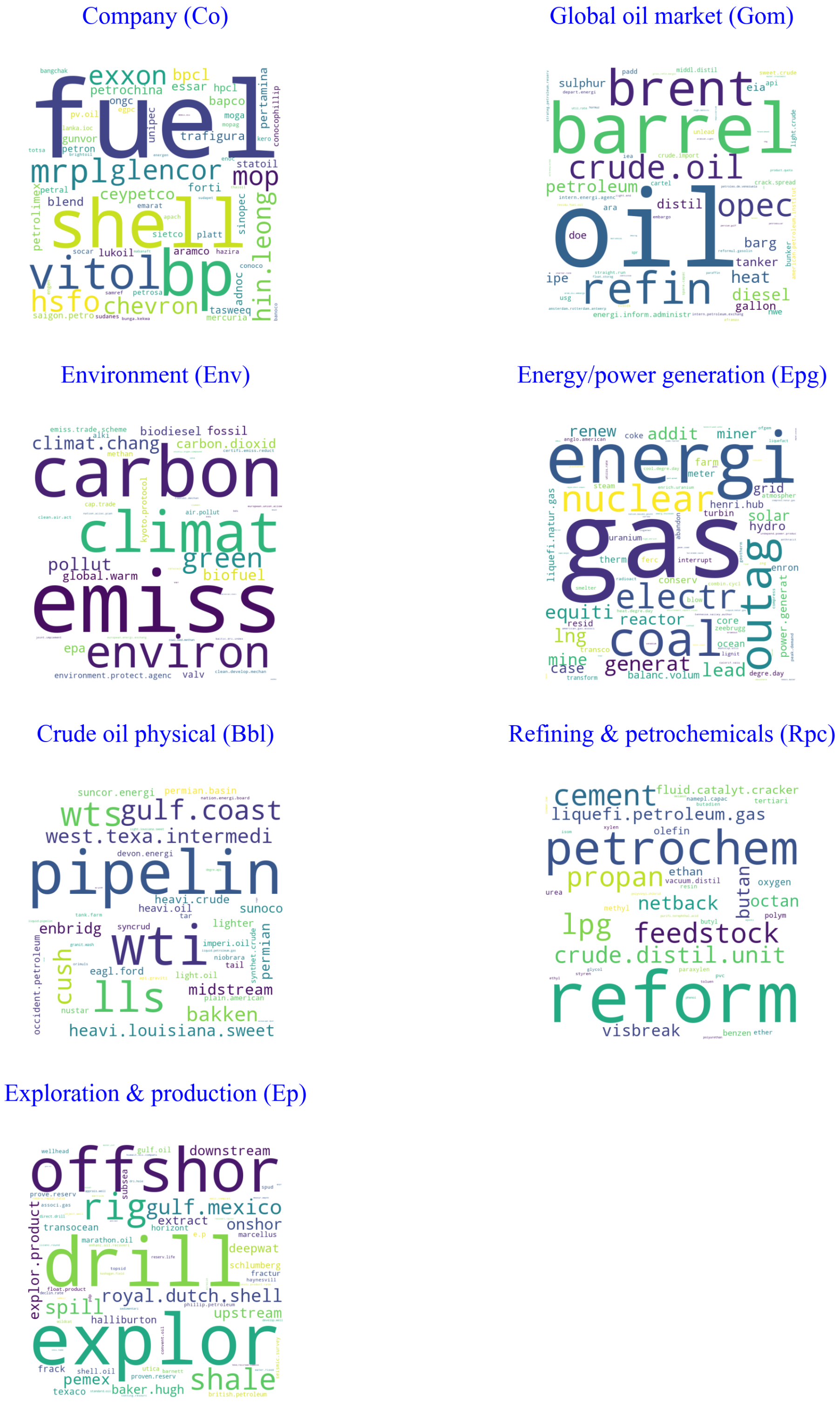
Out-of-sample runs analysis: All models

Dep Var	FutRet	DSpot	DOilVol	xomRet	bpRet	rdsaRet	DInv	DProd
Runs	829	824	943	977	905	906	891	892
Max Runs	990	990	990	990	990	990	990	990
q	0.211	0.214	0.348	0.356	0.286	0.308	0.29	0.261
# All/Txt	45/20	45/20	45/20	45/20	45/20	45/20	45/20	45/20
Run1	695	684	712	828	739	768	760	747
Run1-p	(0.36)	(0.73)	(1.00)	(0.00)	(0.55)	(0.06)	(0.09)	(0.10)
Run2	181	210	401	359	309	346	236	212
Run2-p	(0.23)	(0.00)	(0.00)	(0.07)	(0.00)	(0.00)	(0.99)	(0.95)
Run3	34	41	148	76	55	59	88	64
Run3-p	(0.32)	(0.07)	(0.00)	(1.00)	(0.95)	(1.00)	(0.01)	(0.09)
Run4	1	0	16	6	3	7	8	4
Run4-p	(0.97)	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)	(0.99)	(0.99)
Run5	0	0	0	2	1	0	4	0
Run5-p	(0.60)	(0.62)	(1.00)	(1.00)	(0.88)	(0.99)	(0.35)	(0.91)

Out-of-sample runs analysis: Text models

Dep Var	FutRet	DSpot	DOilVol	xomRet	bpRet	rdsaRet	DInv	DProd
Runs	586	577	657	679	644	638	621	629
Max Runs	690	690	690	690	690	690	690	690
q	0.222	0.22	0.334	0.351	0.291	0.312	0.298	0.269
# All/Txt	45/20	45/20	45/20	45/20	45/20	45/20	45/20	45/20
Run1	484	483	512	572	515	530	531	534
Run1-p	(0.64)	(0.63)	(0.84)	(0.00)	(0.58)	(0.22)	(0.14)	(0.02)
Run2	138	148	267	259	207	254	170	151
Run2-p	(0.20)	(0.02)	(0.00)	(0.01)	(0.05)	(0.00)	(0.97)	(0.95)
Run3	27	28	89	46	50	45	63	44
Run3-p	(0.33)	(0.22)	(0.00)	(1.00)	(0.43)	(0.96)	(0.06)	(0.28)
Run4	0	0	6	5	3	3	4	3
Run4-p	(0.99)	(0.99)	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)	(0.98)
Run5	0	0	0	2	1	0	3	0
Run5-p	(0.55)	(0.54)	(0.99)	(0.95)	(0.76)	(0.98)	(0.37)	(0.86)

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)

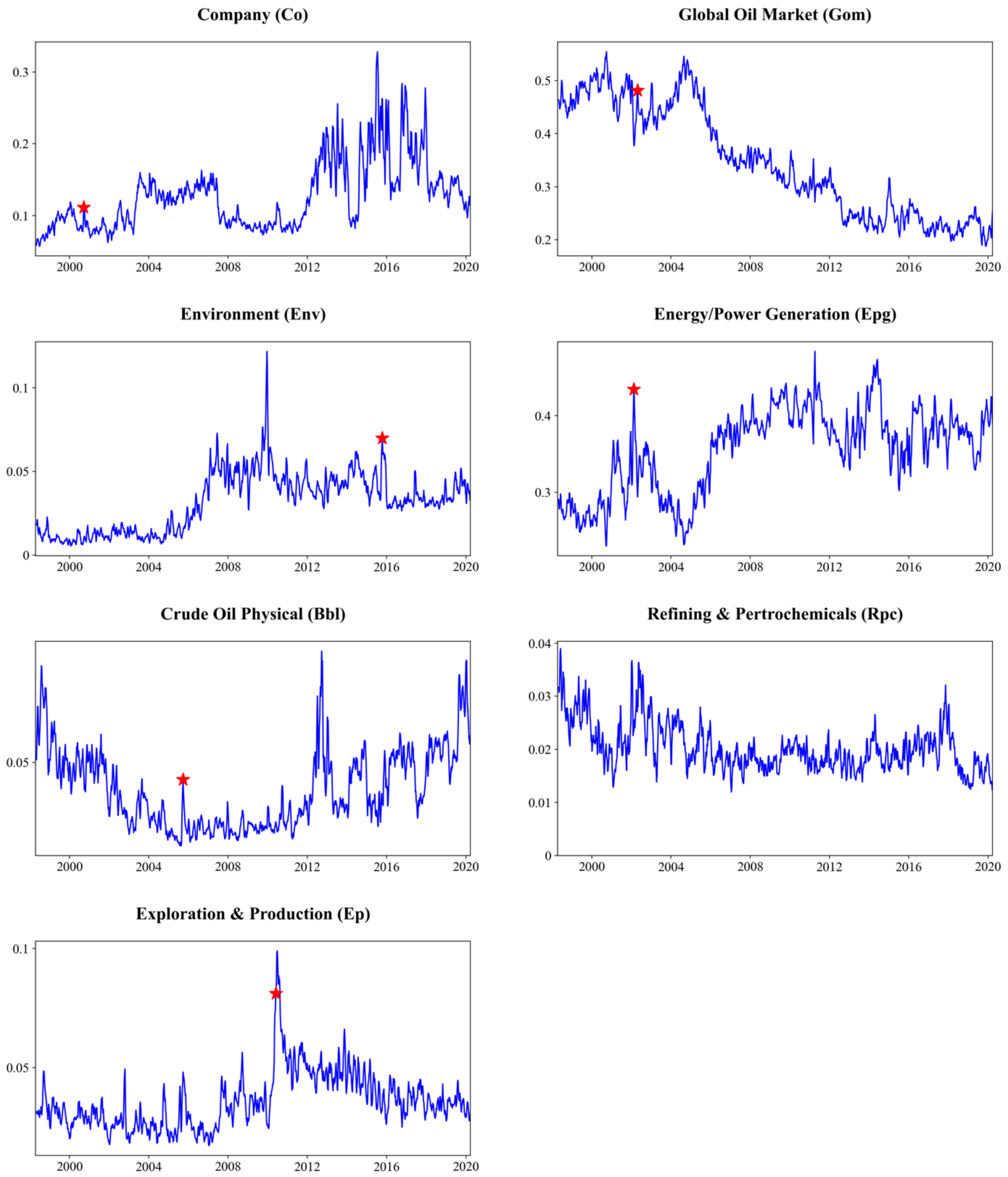
Crude oil physical (Bbl)

Refining & petrochemicals (Rpc)

Exploration & production (Ep)

Figure 2: NLP measures over time. This figure shows the time series plots of all the textual series in this paper. All series start from April 1998 and end in March 2020. We display the 4-week averages of topical frequencies in Panel A, and 4-week averages of topical sentiments in Panel B. 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. In addition, Panel C shows 4-week averages of the article counts, the unusualness (entropy) and the first principal components of normalized 4-week average textual measures.

Panel A: Topical Frequency



Company (Co)

Global Oil Market (Gom)

Environment (Env)

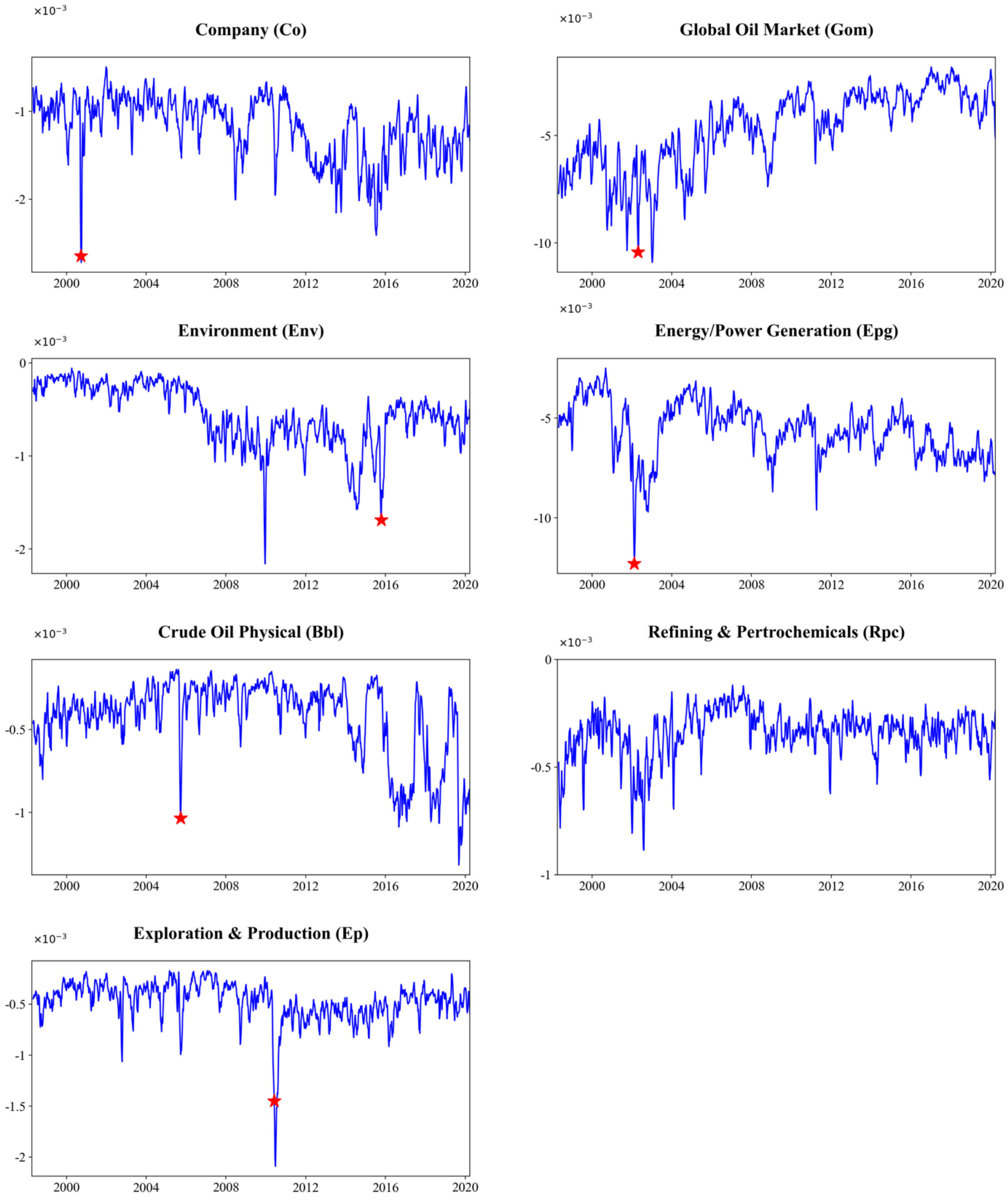
Energy/Power Generation (Epg)

Crude Oil Physical (Bbl)

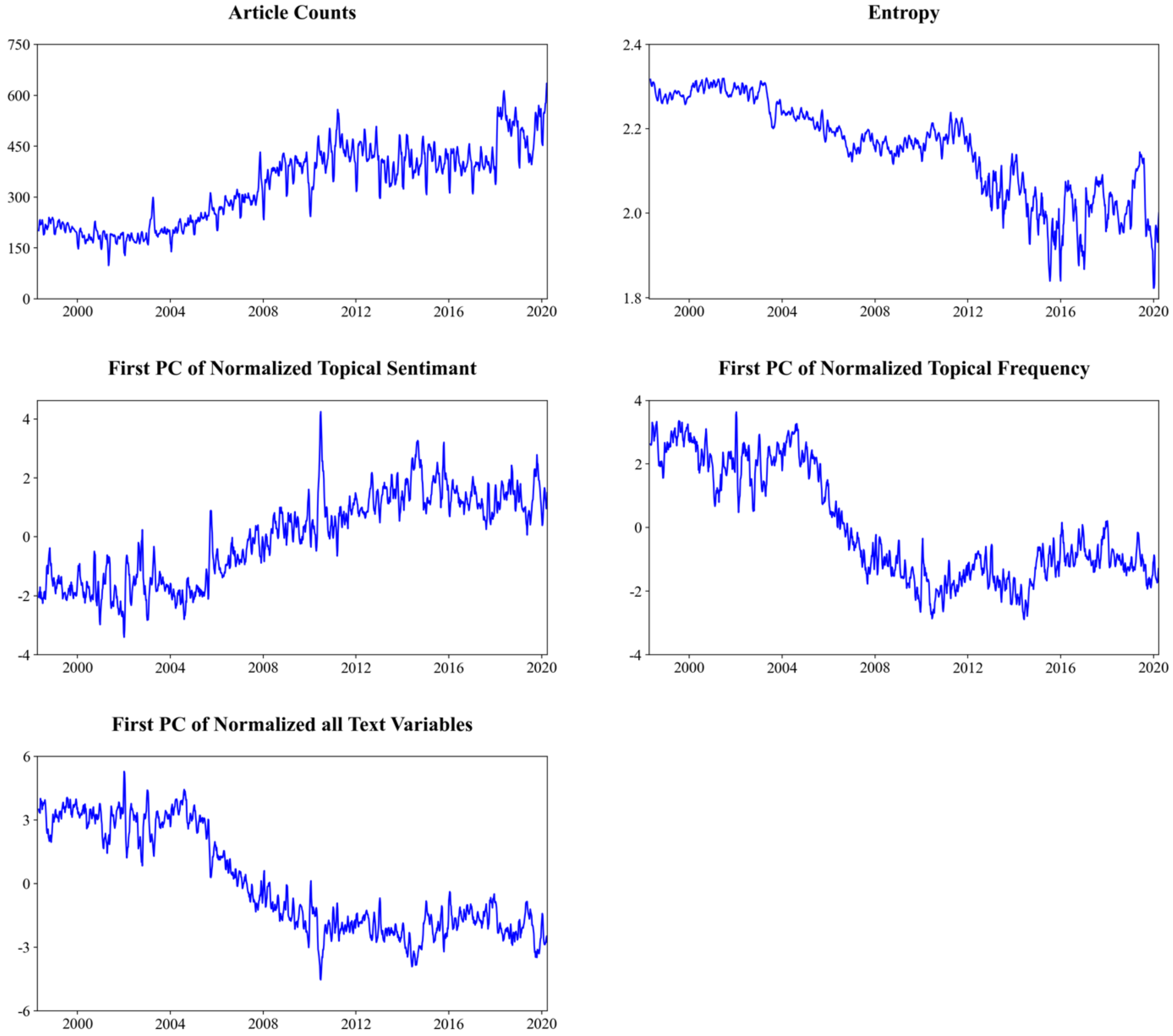
Refining & Pertrochemicals (Rpc)

Exploration & Production (Ep)

Panel B: Topical Sentiment



Panel C: Article Counts, Unusualness and PCA series



Article Counts

Entropy

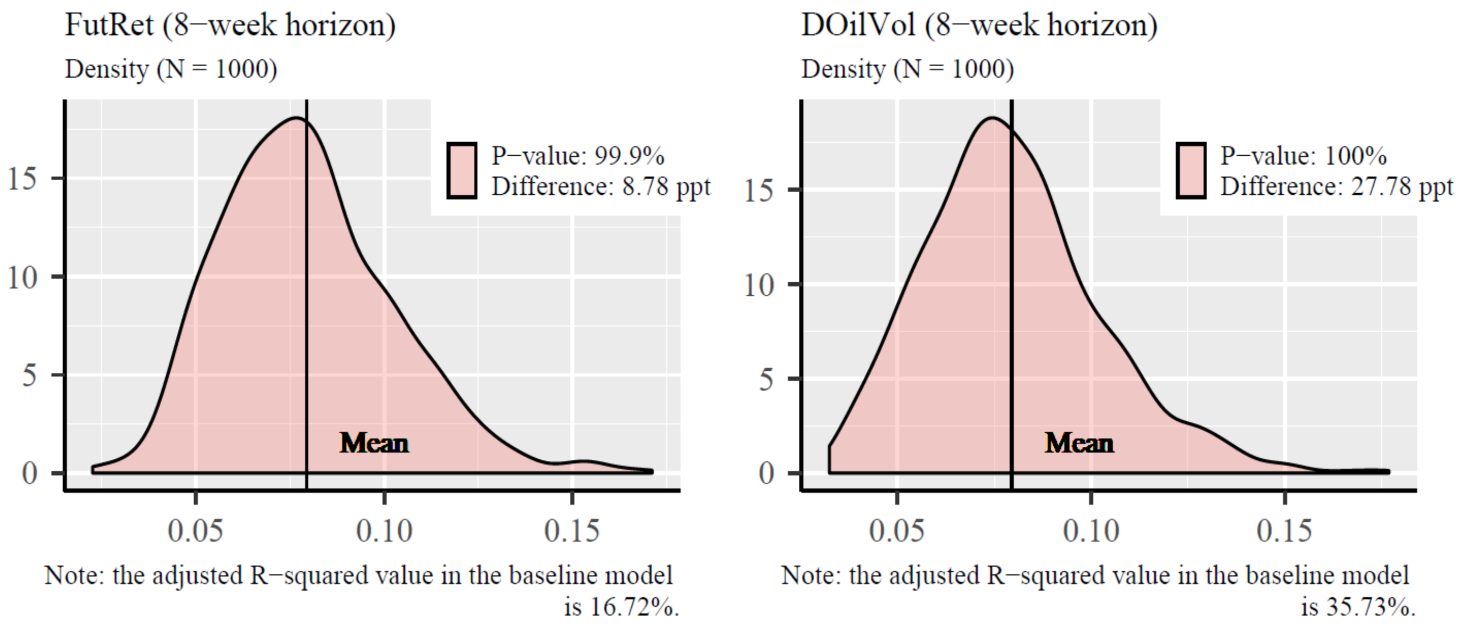
First PC of Normalized Topical Sentimant

First PC of Normalized Topical Frequency

First PC of Normalized all Text Variables

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.


FutRet (8–week horizon)

Density (N = 1000)

Note: the adjusted R–squared value in the baseline model is 16.72%.

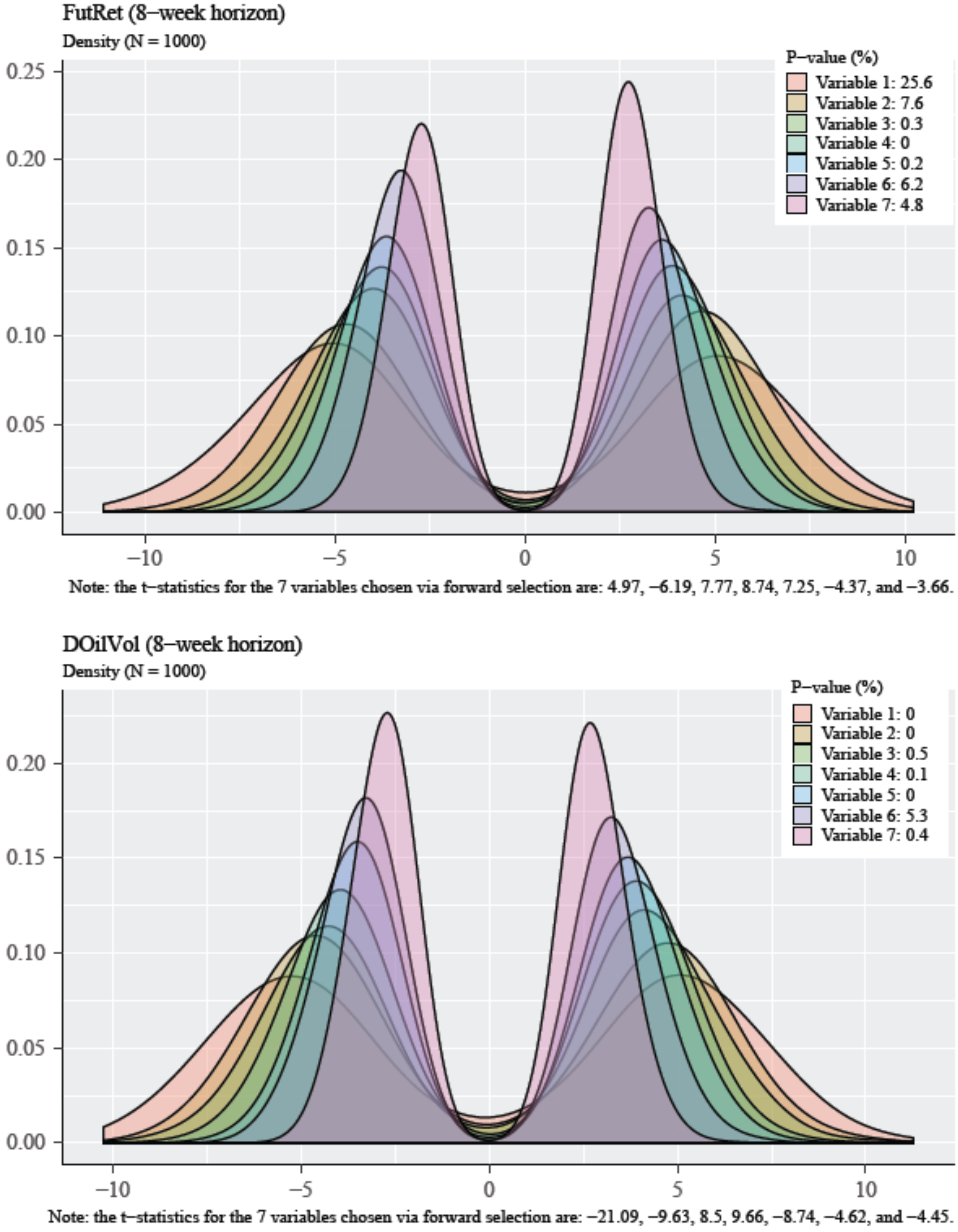
DOilVol (8–week horizon)

Density (N = 1000)

Note: the adjusted R–squared value in the baseline model is 35.73%.

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.

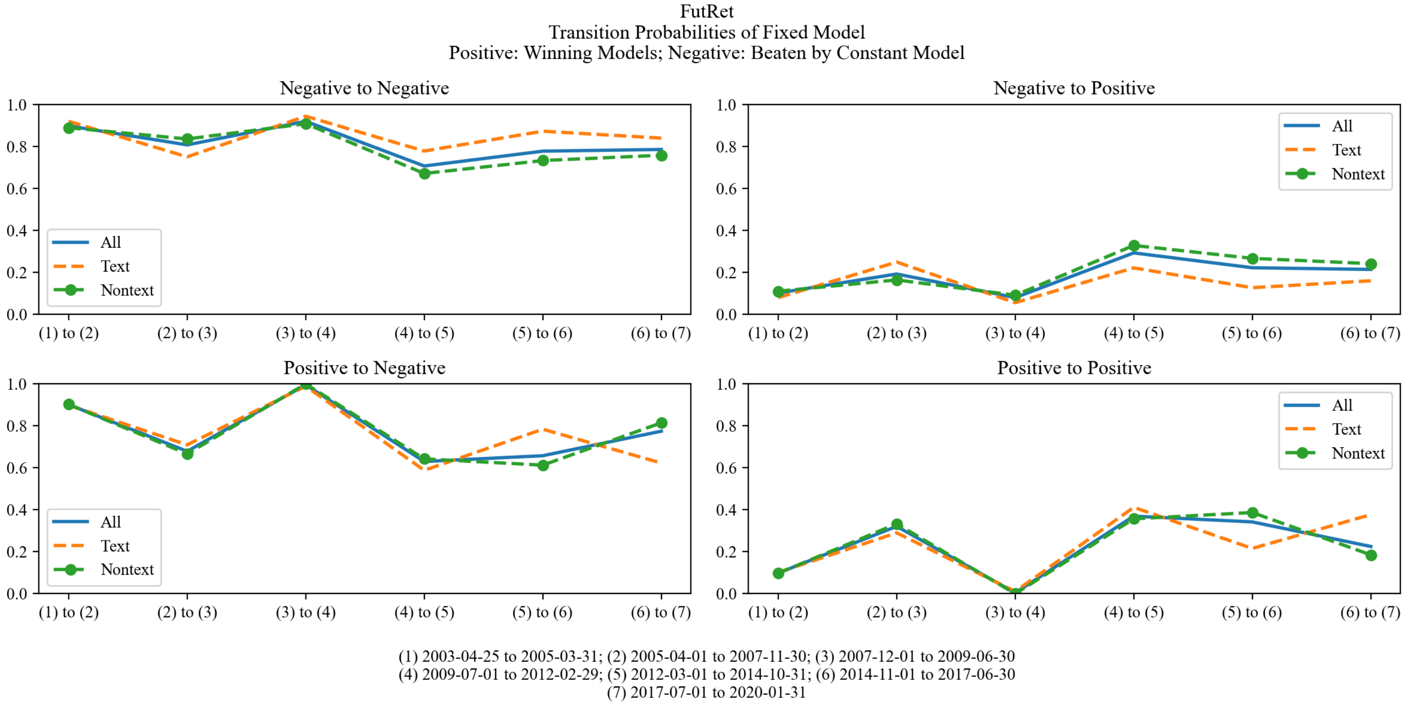
Following the same bootstrap process outlined in the Appendix and used to produce Figure 3, we report here the density functions of the t-statistics for the simulated regression results. The p-value is computed as the minimum of the percent of simulated t-statistics less than the empirical t-statistic, 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.

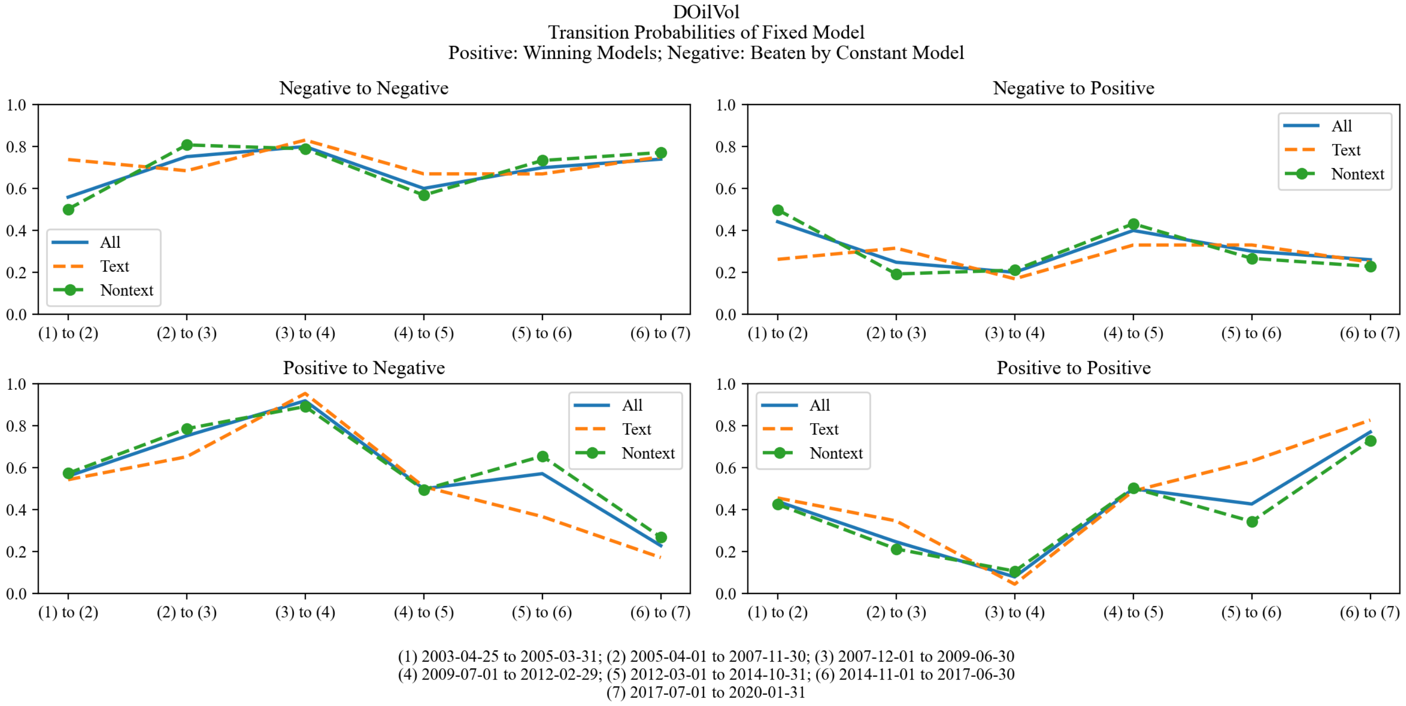


Note: the t–statistics for the 7 variables chosen via forward selection are: 4.97, −6.19, 7.77, 8.74, 7.25, −4.37, and −3.66.

Note: the t–statistics for the 7 variables chosen via forward selection are: −21.09, −9.63, 8.5, 9.66, −8.74, −4.62, and −4.45.

Figure 5: Transition Matrix for Out-of-Sample Fixed Models. This figure shows the transition matrix that captures consistency over time in successful fixed models (i.e., those with MSE ratios less than one). A negative model outcome for a subperiod is defined as one where the MSE value is greater than that of the constant model. A positive model outcome for a subperiod is the opposite. *Negative to Negative* represents the probability of a negative model in period *t* continuing to be beaten by the constant model in period *t+1*; *Negative to Positive* represents the probability of a negative mode in period *t* beats the constant model in period *t+1*; *Positive to Positive* represents the probability of a positive mode in period *t* remains positive in period *t+1*; *Positive to Negative* represents the probability of a positive mode in period *t* turns negative in period *t+1*. There are 7 subperiods in total, the 3rd period is the NBER recession period and periods 4 to 7 are divided evenly in the subsequent samples after the 3rd one. The length of the 2nd period matches the ones after the 3rd and the first period contains data from the beginning of our sample till the start of the 2nd. There are 3 lines in each subplot, referring to *Text* models (models with at least 1 text variable), *Non-text* models (models without any text variable), and *All* models.





Appendix

1. Bootstrapping methodology

We need to control for two deviations from standard assumptions. First, we likely have serial correlation in the residuals of our time-series regressions because of overlapping observations. Second, we use forward selection for choosing a parsimonious set of in-sample regressors. Both of these considerations may introduce upward bias in the R-squareds, and downward bias in the standard errors. To control for both of these sources of finite sample bias, we bootstrap the data and construct bootstrapped distributions for our t-statistics and R-squareds.

We first detrend all dependent and forecasting variables. We then residualize each dependent variable by regressing out its lagged four-week version. Our in-sample analysis assumes the following specification for the detrended and residualized series:

$$y_{t:t+h} = X_t^{(M)}\beta + \epsilon_{t:t+h}, \tag{3}$$

where $X_t^{(M)}, \beta \in R^M$, M is the number of chosen explanatory variables, and the time index t is in weeks. We assume the $X_t^{(M)}$s are chosen from the larger set X_t of $N > M$ variables using forward selection. Under the null we assume that $\beta = 0$ for all M variable subsets. To match the empirical properties of the data, we estimate an $AR(K)$ model for the dependent variable:

$$y_{t:t+h} = b_1 y_{t-1:t-1+h} + \cdots + b_K y_{t-K,t-K+h} + e_{t:t+h}. \tag{4}$$

We run the analysis with $K = h$, i.e. eight lags for eight-week ahead forecasts, and four lags for four-week ahead forecasts. We estimate the above model to get the empirical $\hat{b}_1, \dots, \hat{b}_K$ and the innovation variance $\widehat{var}(e)$; these are the parameters that describe the behavior of the actual data. We then use this to calculate a single run of the simulation as follows:

1. Set $y_{1:h}, \dots, y_{K:K+h} = 0$
2. Draw $e_{K+1:K+1+h}$ from a normal distribution with mean zero and variance $\widehat{var}(e)$.
3. Use the above relationship to generate the next element $y_{K+1:K+1+h}$ in (4).
4. Run the model for 100 steps as a burn-in period, and at step 101 begin collecting data.
5. Collect the y variables until we match the number of empirical observations.
6. Run the forward selection algorithm using the simulated y's and the detrended and residualized X's. This selects a subset $X^{(M)}$ of explanatory variables.
7. Keep track of the adjusted R-squared of this simulation run.
8. Keep track of the standard (no adjustments) OLS t-statistic for each of the variables that are selected by the forward selection algorithm. In every simulated path this will result in M t-statistics, $\{\hat{t}_1, \dots, \hat{t}_M\}$.

Here $\hat{t}_1$ correspond to the t-statistic of the first variable chosen by the forward selection algorithm, $\hat{t}_2$ is the t-statistics of the second chosen variable, and so on. We refer to these as the *ordered t-statistics*.

We repeat this procedure 1,000 times to generate a distribution for the observed R-squareds and the observed t-statistics. The simulated R-squareds and ordered t-statistics adjust for overlapping observations and variable selection under the null hypothesis of no relationship between the dependent and the independent variables.

We evaluate the adjusted R-squared for a given dependent variable via the percentage of simulated adjusted R-squareds that are lower. Since this is a one-sided test, a value of above 95% indicates significance at the 5% level. For p-values of the coefficient estimates in the actual regression, we compare the t-statistic of the *nth* chosen variable in the forward selection method to the *nth* ordered t-statistic distribution under the null hypothesis. We report the outcome of the two-sided test min $(\hat{p}, 1-\hat{p})$ where $\hat{p}$ is the number of simulated t-statistics for the *nth* chosen variable that are less than the t-statistic for the actual *nth* chosen variable. For purposes of this test, all t-statistics are calculated using standard OLS assumptions of independence and homoscedasticity. The simulated t-statistic distributions will reflect all of the OLS biases.

2. Derivation of $\Pr[q,k,n]$

We are interested in finding the probability of having at least one run of 1s of length *k* in a string of length *n* (*n=7* in the paper) where the probability of observing a 1 in a given trial is *q* and all trials are independent. Some boundary cases are easy to establish. The probability of having a run of length *n* is q^n. The probability of having a run of length *n-1* is $2q^{n-1}(1-q)$ which is two times the probability of a corner run, i.e. 1s followed by a zero or 1s preceded by a zero. The probability of having a run of length *n-2* is $2q^{n-2}(1-q)+q^{n-2}(1-q)^2$, which is two corner runs and one interior run, i.e., a series of 1s preceded and followed by a zero. This counting approach extends to cases of runs of length less than *n-2* except that one must be careful to avoid double counting as multiple runs of length *k* can occur in the same length *n* string in order. For example, a string of length 7 can contain two length-3 runs, as in “1110111”. In this case, when calculating the probability of the second run, we must adjust for the fact that strings starting with “1110…” already include strings ending in “…0111”. The following recursive procedure makes these adjustments.

We start with a length *n-k+1* array $p[j]$ for $j=1,\dots,n-k+1$. If $n=k$, then $p[1]=q^k$ and we are done. Otherwise, we proceed as follows

$$\tilde{p}[j]=\begin{cases} q^k(1-q) & if\ j\in\{1,n-k+1\} \\ q^k(1-q)^2 & if\ j\in\{2,3,\dots,n-k+2\} \end{cases}$$

$$p[j]=\begin{cases} \tilde{p}[j] & if\ j<k+2 \\ \left(1-\sum_{l=1}^{j-(k+2)} p[l]-\frac{p[j-(k+1)]}{1-q}\right)\times\tilde{p}[j] & if\ j\geq k+2 \end{cases}$$

We verified this recursive solution by running Monte Carlo simulations to calculate the sample analogues of $\Pr[q, k, n]$.

To understand the intuition for this formula, let's consider the case of $n = 7$ and $k = 3$. There are five possible ways for a run of length 3 to appear in a string of length 7:

1	**2**	**3**	**4**	**5**	**6**	**7**	**j =**
1	1	1	0				**1**
0	1	1	1	0			**2**
	0	1	1	1	0		**3**
		0	1	1	1	0	**4**
			0	1	1	1	**5**

Each of the $n - k + 1$ possible runs is labeled with j, which goes from 1,…,5 as indicated in the algorithm. Runs 1, 2, 3, and 4 (shown in the column labeled j) are mutually exclusive, and correspond to one corner run with probability $q^3(1 - q)$ and three interior runs with total probability of $3q^3(1 - q)^2$. The 5th run (which happens to be a corner run) is already included as a possibility in the *j=1* case, though this 5th run can also happen without the 1st run happening. So, when we calculate the probability of the 5th run, we need to look at the joint event of run 5 happening and run 1 not happening which will avoid double counting. This is why we need the adjustment term in going from $\tilde{p}(j)$ to $p(j)$. The $1/(1 - q)$ term in the last term in the summation in the adjustment formula controls for the fact that both $\tilde{p}[1]$ and $\tilde{p}[5]$ include an overlapping 0.