Oil Shocks: A Textual Analysis Approach*

Deepa D. Datta† Daniel A. Dias‡

February 6, 2022

Abstract

We use textual analysis of oil industry publications to measure oil market developments. By extracting signals from the newly digitized publication archive, we generate historical time series of oil supply and demand indexes. Unlike most official statistics for oil supply and demand that are available only with a publication delay, our indexes can be constructed in real time and contain information on both current and prospective oil market developments. Using a structural VAR model of the oil market, we demonstrate that our supply and demand indexes can help explain oil price movements.

JEL classification codes: Q40, Q41, Q43, C19.

Keywords: Textual analysis; Oil supply and demand; Structural VAR; Oil price shocks.

*We thank Alice Abboud, Mariya Pominova, Ashley Schwanebeck, and Madeleine Yi for excellent research assistance, and Jessica Liu for her contributions to the initial empirical investigations. The views in this paper are solely the responsibility of the authors and should not be interpreted as reflecting the views of the Board of Governors of the Federal Reserve System or of any other person associated with the Federal Reserve System. All errors are our own. Please do not cite or circulate without authors' permission.

†Board of Governors of the Federal Reserve System. Email: deepa.d.datta@frb.gov.

‡Board of Governors of the Federal Reserve System and CEMAPRE. Email: daniel.dias@frb.gov.

1 Introduction

Understanding the relative importance of supply and demand shocks in driving oil market movements is important from both the academic and policy points of view. However, identifying the two sources of variation both accurately and in a timely manner remains difficult. First, while price movements are often driven by news about future market conditions, existing oil supply and demand measures tend to provide information only on current conditions. Second, most direct measures of oil supply and demand conditions are produced with a significant lag, making it impossible to make policy decisions with real-time information at hand. To address this gap, we use specialized oil industry publications whose single purpose is to provide timely daily news. By applying textual analysis techniques to this newly digitized publication archive, we provide a unique and timely source of information about current and prospective oil market developments.

In this paper, we develop a systematic and automated process to extract supply and demand information from specialized industry publications, the *Oil Daily* and the *International Oil Daily*, both published by the Energy Intelligence Group. We categorize words into supply- and demand-related phrases, and based on the frequency of these phrases in the publication archive, we construct a historical time series of supply and demand indexes. Following the oil literature, we use our text-based indexes to identify the underlying economic shocks to supply and demand that drive oil market fluctuations.

Our work improves upon existing approaches by allowing us to provide direct information on current and prospective oil market fluctuations in real time, as the news develops. First, by extracting quantitative signals from daily news articles, we provide information on the sources of oil market fluctuations before data on oil production or measures of economic activity are compiled and published. Importantly, our indexes for supply and demand are constructed directly from the oil industry press, instead of indirectly through signals obtained from external markets, such as those for metals or equities. Second, unlike the typical

measures of monthly oil production and real economic activity, our supply and demand indexes contain both current and prospective information. That is, we don't just have information on flow oil supply and flow oil demand, we also have information on *expected future* oil supply and demand. This prospective information can be especially valuable as oil prices are determined by both current and expected supply and demand conditions. For example, Kanzig (2021) provides strong evidence of the importance of news about future supply for oil price movements and other macroeconomic variables.

Our indexes are highly correlated with existing measures of oil production and global economic activity, both contemporaneously and at long time horizons. They are also strongly correlated with contemporaneous and future movements in oil prices. Importantly, these correlations are in line with economic theory. After constructing our indexes and validating that they are generally consistent with other existing measures of oil production, demand, and oil prices, we use them to estimate a structural model of the oil market, which we then use to decompose episodes of large movements of oil prices into supply- and demand-driven components.

Our paper provides further evidence that important macroeconomic information can be derived from text, and that this information can be used to better understand and measure expected and unexpected movements in macroeconomic variables. Relative to the existing literature using textual analysis to study economic phenomena, we expand the use of text-based analysis to the study of oil market developments. Moreover, and unlike many existing papers using textual analysis, our work moves towards extracting quantitative signals from the text by counting phrases, instead of articles (Baker et al. (2016), Caldara and Iacoviello (2018), Husted et al. (2019)).

In particular, we build on the sentiment literature, which applies textual analysis to study oil market developments. This literature focuses on oil market sentiment in news articles, without probing further the underling supply and demand conditions that may be shaping

that sentiment. In particular, Loughran et al. (2019) focus on how the information in the text is incorporated into prices. Similarly, Calomiris et al. (2020) generate sentiment indexes on a variety of oil market topics, including oil exploration, power generation, and the environment for example, and use these indexes to forecast oil prices. Importantly, neither of these papers apply their text-based measures to better understand the supply- and demand-driven components of oil price movements, which have important macroeconomic implications.

Brandt and Gao (2019) use textual analysis to study how news about macroeconomic fundamentals and about geopolitical events affect oil prices differently. In our approach we don't distinguish between purely macroeconomic news and geopolitical events in the construction of our oil demand indexes because we are less interested in this nuance and also because major geopolitical events are relatively rare.

Among the existing literature applying textual analysis to study the oil market, Cavallo and Wu (2012) is the paper that is closest to ours. These authors combine narrative and quantitative approaches to construct measures of oil price shocks. Their narrative approach involves human auditing of *Oil Daily*, *Oil & Gas Journal*, and the *Monthly Energy Chronology*. Using this approach, they attribute daily changes in oil prices to 22 types of oil-related events, e.g. weather changes, oil field discoveries, political and military actions, and changes in actual or expected inventories. Next, they aggregate select event types to generate exogenous oil shocks series, and show substantial effects of these shocks. While this work is very promising, it depends on human auditing and therefore is extremely difficult to replicate and update. By accomplishing similar analysis using automated processing instead of individual judgment to classify articles, we increase replicability and produce a data set that is easily updated over time. Additionally, whereas the Cavallo and Wu classification only provides an indicator variable for whether each type of event is discussed in a given article, our full text searches for the number of phrases in each article allow us to quantify the relative importance

of various factors within the news narrative and to provide partial attribution of daily price changes based on these counts.

Our paper is obviously also related to the narrative approach to identifying macroeconomic shocks literature - see, for example, Romer and Romer (2004), Romer and Romer (2010), or Cloyne (2013). More generally, there have been recent papers showing that certain developments take time to be reflected in macroeconomic variables, and not accounting for them muddles our view of how certain economic variables develop. For example, Cascaldi-Garcia and Vukoti (2020) show that patent-based news shocks have little effect in the short-run but induce strong permanent effects on total factor productivity after five years. Using our indexes, we find strong evidence that the indexes contain important information about future oil supply, demand, and, consequently, price movements, which, if not taken into account in empirical analysis will give rise to a distorted view on expected and unexpected oil price movements. Furthermore, an important unique strength of our work is that use of textual analysis techniques allows us to easily replicate and extend our news-based indexes.

The rest of the paper is organized as follows: in section 2, we describe the data and how we construct the oil supply and demand indexes; in section 3, we provide empirical evidence on the relationship between our oil price indexes and movements in oil production, economic activity, and oil prices; in section 4, we use our supply and demand indexes to construct a structural model of the oil market that can be used to learn about oil price movements; in section 5 we provide some concluding remarks; and, finally, in the appendices we provide additional information on how we constructed our oil supply and demand indexes and some additional empirical results.

2 Measuring Oil Supply and Demand Developments

A key contribution of this paper is to use word and phrase counts in the newly digitized oil-related news archive to construct indexes reflecting oil supply and demand developments. In this section, we first describe the data source and methodology used to construct the indexes and then provide some summary statistics of the indexes themselves.

2.1 The *Oil Daily* and *International Oil Daily*

Our corpus includes articles published in two daily publications, the *Oil Daily* (OD) and the *International Oil Daily* (IOD), from the well-regarded Energy Intelligence Group (EIG). These publications, which specialize in the oil market, provide us with deeper and broader coverage than would be available from the general financial press, including the *Wall Street Journal*, *Financial Times*, Bloomberg, or Thomson Reuters. Indeed, EIG is so well regarded for its high-quality coverage of oil market developments that its oil market statistics are used by OPEC as an official secondary source, e.g. for monitoring adherence to production quotas. Additionally, the EIG publications contain unique information about the oil market that cannot be easily extracted from oil prices or existing measures of oil market developments. Consequently, although EIG's published oil market statistics are of high quality, we focus on extracting additional information from the article text, or the narratives, and do not use EIG's published data to create the supply and demand indexes.

Our corpus comprises the newly digitized articles from the OD and IOD. Although the OD has been published since 1951, our electronic archive for the OD begins only in 1992 and adds the IOD once it debuts in 2002. As shown in Table 1, aside from the earlier availability of the OD, the two publications are relatively similar along several dimensions. Each publication has about 20 articles per daily issue averaging 300 words per article. As shown in Figure 1, the IOD tends to have a few more articles than the OD each day. Importantly, both

publications have articles about current and prospective oil market developments, and each has a daily article summarizing the previous day's oil price movements as well. The key difference between the two publications is that since the debut of the IOD, the OD has been more focused on the U.S. oil market, while the IOD has been weighted towards the international oil market.

Before implementing our index construction methodology, we clean the data in a few ways. First, because the oil market is global in nature, there is some overlap between the OD and the IOD, which results in some articles being included in both publications. We filter out these duplicates from our data set. Second, we also filter out articles that are not related to oil markets, including articles that focus on natural gas or gasoline. The dropped articles also include announcements of personnel changes at major oil companies, blurbs that very briefly report statistical releases without additional commentary, and brief corrections to earlier articles. Further details on data processing are included in Appendix B.

To give an idea how the cleaned data compare with the original data, we provide in Table 1 the summary statistics for the cleaned data. The first column of the table summarizes the filtering results for the early part of the sample, when only the OD was being published. For this sample, once we drop articles that are unrelated to oil market developments, about 62 percent of the articles are retained and the average number of articles per day drops from 24 to 15. In the later sample, represented by the next two columns, when we drop both unrelated articles and duplicate articles between the OD and IOD, the average number of articles per day is reduced by about half, from 36 to 19 articles per day. The last column summarizes the full sample results. Consistent with the exclusion of many brief articles, overall, our cleaned sample has a slightly higher average word count of about 350 words per article. Finally, we can see in Figure 1 that the number of articles per issue declines somewhat in the later years of the sample. In the later period, each daily issue tends to have fewer, longer articles, and a total word count that is little changed from the earlier period.

Notably, we also find (as discussed below) that the frequency of supply and demand words remains stable over time.

2.2 Index Construction

Our indexes are constructed to capture the fundamental movements in oil markets: increases and decreases in supply and demand. To accomplish this, we generate four broad vocabulary lists of supply, demand, increase, and decrease words. Next, we count the number of times supply and demand words appear in proximity to increase and decrease words. These counts are used to generate our directional supply and demand indexes.

Appendix C provides the four vocabulary lists as well as details on how they are constructed. In sum, we have 49 supply words, 34 demand words, 243 increase words, and 298 decrease words. In Table 2, we list the five most frequently occurring words and their frequencies from each vocabulary list. Notably, there is a large concentration of supply and demand words in just a few expressions; the top five words from our supply and demand word lists account for about 60 and 70 percent of all the supply and demand words, respectively. Similarly, the top five words from our increase and decrease words account for about 40 and 30 percent of all the increase and decrease words, respectively, that are identified in the text of the articles. Importantly, the five most frequently used supply, demand, increase, and decrease words all seem reasonable and are expected to reflect changes in oil supply and demand conditions.

In Table 3, we provide an example from our article corpus to illustrate this process. This article from the midst of the Great Recession discusses how weak oil demand conditions are weighing on prices despite possible OPEC supply cuts in response. It also discusses potential supply cuts that may boost prices imminently or at some time in the future, illustrating the forward-looking nature of our supply and demand indexes. Using our methodology, this example yields a count of 4 supply words (capitalized and underlined) and a count of 3 demand words (capitalized and double underlined). The algorithm next looks for the increase

and decrease words only in sentences with supply and demand words, and it yields 3 supply decrease and 2 demand decrease counts (highlighted in orange) and 0 supply increase and demand increase counts. These counts are well in line with the gist of the article, which is to discuss current weak demand and prospective supply cuts. Notably, the algorithm does not pick up directional cues about prices (e.g. discussion of the decline in oil or stock prices), but rather selects only the directional information on supply and demand.

Of course, our methodology may capture a fair amount of noise in addition to the information we are able to process. However, the results we obtain for this example article illustrate how our methodology might capture the main signals from the articles. In the next section, we provide statistical evidence that our methodology does provide useful information.

2.3 Index Properties

Having described the data we use and how we construct our oil supply and demand indexes, we now show the results of our textual analysis algorithms. In Figure 2, the top panel shows the raw counts of supply and demand words at the monthly frequency. These raw counts increase sharply in 2002, when the sample expands to include IOD articles. The bottom panel shows the monthly average of the individual article ratios of supply and demand words to total words. The ratios in the lower panel do not jump noticeably in 2002. There is significant time variation in all four of the series, and there are approximately twice as many supply words as demand words in the articles. On average, the shares of supply and demand words per article hover around 1.1 and 0.5 percent, respectively.

The word counts give some indication of the relative importance of supply relative to demand for the reporting on oil market developments. Ultimately, however, our goal is to create separate indexes for supply and demand developments that can be associated with movements in oil prices. To construct these, we count supply increase, supply decrease, demand increase, and demand decrease phrases for each article. To create each monthly directional index, as

shown in Figure 3, we take the ratio of the monthly sum of e.g. supply increase phrases to the monthly sum of total words in the articles.

This figure shows that, in general, there are more supply than demand phrases, which may reflect that our oil industry publication has more specialized reporting on oil supply issues than oil demand. We also find more increase phrases than decrease phrases for both supply and demand, though this ordering tends to flip briefly in recession periods (shaded in gray), when we see spikes in both supply decrease and demand decrease phrases. The generally higher frequency of increase phrases may be due to the fact that both oil supply and demand grew considerably over our sample period, or it may reflect higher salience of supply and demand increases in the news, relative to supply and demand decreases.

We also construct net supply and net demand ratios, where net supply is the ratio of supply increase minus supply decrease phrases divided by the total word count, and net demand is the ratio of demand increase minus demand decrease phrases divided by the total word count. Before using our indexes in statistical models, we transform them by standardization relative to a reference period. That is, for each of our four directional indexes and for the net indexes as well, we subtract the mean of the reference period of 1995 to 2004, and divide by the standard deviation of the reference period as well.[1] The standardized versions of the net supply and demand indexes are shown in Figure 4. Notably, both measures fall sharply during U.S. recessions.

[1]Instead of using the whole sample period to standardize our indexes, we opted for a fixed and smaller part of the sample so that the historical indexes do not get revised when the raw data are extended. We chose the 1995 to 2004 period because it is a relatively quiet period in the oil market. Importantly, the choice of reference period has no material implications for the results. It only changes the scale of our indexes, while preserving all other relevant statistical properties of the indexes.

3 Evaluating Our Oil Supply and Demand Indexes

In this section we show how our oil supply and demand indexes vary with existing measures of oil supply, real economic activity, and oil prices. First, we relate our measures to contemporaneous and 12-month ahead global oil production and industrial production. Next, we relate our measures to monthly and weekly oil prices for both spot and farther-dated futures contracts.

For oil supply, we use the log of world total oil production, as reported by the International Energy Agency. For oil demand, we use the log of global industrial production, which Hamilton (2019) proposes as a measure of real economic activity (REA). Lastly, for oil prices, we use West Texas Intermediate crude oil futures prices, measured in logs. We define the spot price using the front-month futures contract, and define the futures price using the 12-month ahead contract.

3.1 Empirical Approach

Our first set of regressions relate our monthly indexes to the *contemporaneous* changes in oil production, REA, and oil spot and futures prices. As explanatory variables, we include either the directional indexes (equation (1)) or the net supply and demand indexes (equation (2)). These regressions help determine whether the indexes provide information beyond that contained in our controls (equation (3)), which include one year of past changes in oil production, REA, and oil prices ($L = 12$).

$$\begin{aligned} y_t - y_{t-1} = \beta_0 &+ \beta_{SI} SupplyIncrease_t + \beta_{SD} SupplyDecrease_t \\ &+ \beta_{DI} DemandIncrease_t + \beta_{DD} DemandDecrease_t \\ &+ Controls(1, L) + \epsilon_t \end{aligned} \tag{1}$$

$$
\begin{aligned}
y_t - y_{t-1} = \beta_0 + \beta_S NetSupply_t + \beta_D NetDemand_t \\
+ Controls(1, L) + \epsilon_t
\end{aligned}
\tag{2}
$$

$$
\begin{aligned}
Controls(j, l) = \sum_{i=j}^{l} \rho_i^{Prod} \Delta ln(OilProd_{t-i}) + \sum_{i=j}^{l} \rho_i^{REA} \Delta ln(REA_{t-i}) \\
+ \sum_{i=j}^{l} \rho_i^{Spot} \Delta ln(Spot_{t-i}) + \sum_{i=j}^{l} \rho_i^{Futures} \Delta ln(Futures_{t-i})
\end{aligned}
\tag{3}
$$

In the next set of regressions, we examine whether our indexes contain any additional information about *future* movements of the four outcome variables, beyond that contained in our controls. Here, the controls include current and past values of oil production, REA, and oil spot and futures prices. For all of our outcome variables, we estimate these equations at the monthly frequency, and set both h and L equal to 12.

$$
\begin{aligned}
y_{t+h} - y_t = & \beta_0 + \beta_{SI} SupplyIncrease_t + \beta_{SD} SupplyDecrease_t \\
& + \beta_{DI} DemandIncrease_t + \beta_{DD} DemandDecrease_t \\
& + Controls(0, L) + \epsilon_t
\end{aligned}
\tag{4}
$$

$$
\begin{aligned}
y_{t+h} - y_t = & \beta_0 + \beta_S NetSupply_t + \beta_D NetDemand_t \\
& + Controls(0, L) + \epsilon_t
\end{aligned}
\tag{5}
$$

Finally, for prices, we also estimate the above regressions at the weekly frequency and evaluate both contemporaneous results and those at the 12-, 26-, and 52-week horizons. For these regressions, we exclude the controls for oil production and REA, since these data are available only at the monthly frequency, and we set $L = 52$ to use lagged information from the previous year.

For all of the above models, we also report results from a test of the hypothesis that the coefficients on the supply increase and supply decrease variables are symmetric, i.e. $H_0 : \beta_{SI} + \beta_{SD} = 0$. We also test symmetry for the demand variables, $H_0 : \beta_{DI} + \beta_{DD} = 0$. Finally,

we test the two restrictions jointly $H_0 : \beta_{SI} + \beta_{SD} = 0 \,\&\, \beta_{DI} + \beta_{DD} = 0$. The results of these tests help us understand if there is any benefit in using the directional indexes separately, or whether we should just use the corresponding net indexes.

Table 5 reports the monthly regressions, and Table 6 reports the weekly regressions. Note that, for ease of comparison across tables, we have annualized all outcome variables before using them in the regressions. In the next subsections we discuss the results separately for oil production and oil demand and for oil prices.

3.2 Our Indexes and Other Measures of Oil Supply and Demand

Table 5 shows that our indexes clearly capture directional information about both contemporaneous and future movements in supply and demand. Overall, we find coefficients indicating that oil production rises with the supply increase index and declines with the supply decrease index, and real economic activity rises with the demand increase index and declines with the demand decrease index.

For example, the first column of the table shows that a one standard deviation increase in the supply increase index is associated with an increase in oil production at about a 3.3 percentage points annual rate, and a one standard deviation increase in the supply decrease index is associated with a decrease in oil production that month, at a 2.8 percentage points annual rate. In the next column we show that our net supply index is also strongly positively associated with contemporaneous oil production. These relationships also hold qualitatively when relating the indexes to the 12-month-ahead change in oil production, though as shown in the lower panel of the table, we find that the coefficients are smaller in magnitude. These results are consistent with the notion that the indexes contain information about both current and prospective developments in oil production.

We also find a strong and statistically significant contemporaneous relationship with real economic activity. The third column in the top panel shows that a one standard deviation increase in the demand increase index is associated with a 1.7 percentage points higher growth rate in industrial production, and a one standard deviation increase in the demand decrease index is associated with a 1.4 percentage points lower growth rate in industrial production. These results also hold for the net demand index, with the effect of a one standard deviation increase in the net demand increase index increasing industrial production growth rate by 0.8 percentage point. Furthermore, we also find that a one standard deviation increase in the supply decrease index is associated with a negative 1.3 percentage point decline in the growth rate of industrial production. The bottom panel of the table shows that in the forward-looking regressions, the estimated relationships between the indexes and real economic activity remain fairly strong, with much less attenuation of the coefficient magnitudes than in the oil production regressions.

Turning to the tests of symmetry for the increase and decrease indexes, the p-values reported at the bottom of the top panel of the table indicate that for the contemporaneous regression, we do not reject the hypothesis of symmetry between SI and SD for the oil production regressions (p=0.261), nor do we reject symmetry between DI and DD for the real economic activity regressions (p=0.374). This suggests a that our indexes do have quantitatively reasonable associations, such that the supply increase and decrease indexes are associated with increasing and decreasing production of a similar magnitude, and the same applies between our demand indexes and REA. That said, we find that when relating the supply indexes to REA, or the demand indexes to oil production, or when considering the forward-looking effects, we tend to reject the hypothesis of symmetry. We conclude that an unrestricted model including the four directional indexes may provide additional information in further work.

Below each regression, Table 5 reports the R^2 for each regression along with the "Indexes R^2 share", which is an estimate of the contribution of the indexes to the goodness-of-fit as measured by the R^2.[2] The R^2 values range from about 0.2 to 0.4, and withe the exception of the forward-looking regression for REA and the net indexes, the indexes R^2 shares of these generally range from 12 to 30 percent.

Overall, the various results from Table 5 show that the indexes we construct contain information about the supply and demand developments in the oil market, both contemporaneously and in the future, with the expected relationships between the supply and demand indexes and oil production and real economic activity. These results affirm the validity of the indexes constructed from textual analysis.

3.3 Our Indexes and Oil Prices

Having documented the close relationship between our indexes and oil production and real economic activity, we now turn to establish the relationship between our indexes of supply and demand and oil spot and 12-month futures prices. We first consider regressions at the monthly frequency before moving on to consider the weekly frequency. As with the oil production and economic activity regressions, we consider the 1-month contemporaneous change in the outcome variables, as well as the 12-month-ahead change. For the weekly regressions, we look at the contemporaneous change, as well as the 12-, 26-, and 52- week ahead price changes. As before, we have annualized all outcome variables before using them in the regressions for ease of comparison. Table 5 reports the monthly regressions, and Table 6 reports the weekly regressions.

[2] To estimate the effect of the indexes on the R^2 of each equation, we use the Shapley-Owen's R^2 decomposition proposed by Huettner et al. (2012). This approach considers all possible combinations of variables or group of variables in the regression to estimate the contribution of each variable or groups of variables to the total R^2 of the equation.

In Table 5, the results in the fifth column of the table show that contemporaneous oil spot prices rise with the supply decrease index, and fall with the supply increase index. In particular, a one standard deviation increase in the supply increase index is associated with a contemporaneous decline in the spot oil price of nearly 15 percent at an annual rate. Similarly, a one standard deviation increase in the supply decrease index is associated with an oil spot price increase of about 7 percent at an annual rate. The coefficient on the supply increase index is statistically significant but that of the supply decrease index is not. In the next column we show that our net supply index is also strongly negatively associated with contemporaneous oil spot prices. A one standard deviation increase in the net supply index is associated with a decrease in oil prices that month, at a 10 percent annual rate.

The demand indexes also show a strong relationship with the contemporaneous oil spot price changes. For example, a one standard deviation increase in the demand increase is associated with a contemporaneous increase in spot prices of 31 percent at an annual rate. For the demand decrease index, we estimate that a one standard deviation increase in the index is associated with a decrease in the spot price of oil at a 17 percent annual rate. The results for the net supply and net demand indexes are similar to those observed for the four directional indexes, with oil prices increasing by 23 percent at annual rate in response to a one standard deviation of the net demand index. All coefficients except for the one on the supply decrease index are statistically significant.

Turning to the results for the 12-month futures price, we find similar results with slightly attenuated magnitudes, in line with the lower volatility observed for 12-month futures prices. Most coefficients are statistically significant, and once again, both supply- and demand-related coefficients have the expected signs.

These relationships also hold when relating the indexes to the 12-month-ahead change in oil spot and futures prices, as shown in the last 4 columns of the lower panel of the table. In general, we find that higher supply decrease and demand increase indexes are associated with

spot and futures prices rising over the next 12 months. Similarly, prices fall in the 12 months following an increase in the supply increase and demand decrease indexes. For the demand indexes, we find that the 12-month-ahead coefficients are somewhat attenuated at around 6 to 14 percent as compared to 15 to 30 percent for the coefficients in the contemporaneous regression. By contrast, the estimated relationships between the supply indexes and 12-month-ahead spot and futures prices remain large in magnitude. A one standard deviation increase in the net supply index is associated with a 10 percent (a.r.) decline in the 12-month-ahead spot price, and a nearly 8 percent decline in the futures price over the following 12 months.

Once again, the indexes contribute significantly to the goodness-of-fit of the regressions, especially in the case of the 12-month-ahead regressions. For the contemporaneous regressions, the indexes contribution to the R^2 varies between 11 and 22 percent. For the 12-month-ahead the contribution of the indexes to the R^2 of the regressions is estimated to be around 40 percent. This is an astonishing result, which suggests that the information contained in our indexes is very informative about oil price movements. The very large increase in the R^2 of the 12-month-ahead regression due to the indexes indicates that oil price movements are more predictable using textual analysis than what previous studies using only oil prices, production, and real economic activity might have found.

Table 6 provides the results for the regressions at the weekly frequency. Here, we consider the contemporaneous change in the spot or futures price in the first two columns. Subsequent columns examine the relationship between the indexes and the price change observed over the following 12, 26, and 52 weeks. In general, we find strong associations between the indexes and the contemporaneous spot and futures price movements. For example, a one standard deviation increase in the net supply index is associated with a 15 percent decline in spot prices and a 7 percent decline in 12-month futures prices. When looking at results for the following weeks, we find weaker results at the 12- and 26-week horizon, and some-

what stronger associations again at the 52-week horizon. Notably, for the 52-week-ahead regressions, the contribution of the indexes to the R^2 is even higher than in the case of the monthly regressions.

Turning to the tests of symmetry for the increase and decrease indexes, the p-values reported at the bottom of the tables indicate we tend to not reject the hypothesis of symmetry between SI and SD, and between DI and DD for the contemporaneous price regressions. However, we do tend to reject the hypotheses for symmetry for the forward-looking regressions. These results suggest that it is informative to include all four directional indexes separately, instead of using the net indexes only.

Overall, the results in this section show that our indexes contain important information about oil market developments. Notably, in comparison to the coefficients of our indexes on oil production and REA, the coefficients for the oil price regressions are much larger. In the next section, we use these indexes to better understand oil price movements.

4 Using the Supply and Demand Indexes to Learn About Oil Price Movements

In this section we propose a structural model of the oil market that utilizes the high information content of the oil supply and demand indexes. Next, we present estimates of the short-run elasticities, the impulse-response functions, and the forecast error variance decomposition of the SVAR model introduced. And last, we use the SVAR model to decompose oil price movements into demand and supply driven at monthly and weekly frequencies.[3]

[3]Note that, we could use the directional or net supply and demand indexes to estimate a daily SVAR model because all indexes can be constructed at a daily frequency.

4.1 A Structural Model of the Oil Market using Supply and Demand Indexes

The workhorse model of the oil market has three main components: 1) information on supply; 2) information on demand; and 3) information on the price of oil. These three components are often represented by just three variables, including monthly global oil production, a measure of global real economic activity, and an oil price, as in Kilian (2009) and subsequent research. In other cases, these three components are represented by more than three variables. For example, Kilian and Murphy (2014) use oil inventories as an additional source of information on oil supply and demand conditions, while Caldara et al. (2019) use metals prices and separate measures of real economic activity for advanced and emerging economies to better model the demand side of the oil market.

Our approach follows a similar structure, in that we use the supply increase and supply decrease indexes as the sources of information for oil supply conditions and the demand increase and demand decrease indexes as the sources of information for oil demand conditions. Notably, we use the individual directional indexes in this model because they have higher informational content than the net supply and net demand indexes and also because we showed in the previous section that the coefficients for supply and demand are asymmetric. As in the rest of the literature, we combine the measures of supply and demand with observed oil prices for a more complete picture of the oil market.

One key advantage of using our indexes to study the oil market is that, unlike the typical measures of monthly oil production and real economic activity, our supply and demand measures contain both current and prospective information. That is, we don't just have information on flow oil supply and flow oil demand, we also have information on *expected future* oil supply and demand. This prospective information can be especially valuable as oil prices are determined by both current and expected supply and demand conditions. In this regard, Kanzig (2021) provides strong evidence of the importance of news about future

supply for oil price movements and other macroeconomic variables. While our supply and demand indexes cannot be interpreted as exogenous as is the case of the OPEC news shocks in Kanzig (2021), they contain information about expected developments in the oil market that are not only about supply or just related to OPEC oil production decisions.

We present a model with just three variables as an example to help guide our discussion, and abstracting from the model dynamics, we have that each variable is a function of three structural shocks:

$$\begin{cases} OilProd_t = h_1(\epsilon_t^S, \epsilon_t^D, \epsilon_t^{PD}) \\ REA_t = h_2(\epsilon_t^S, \epsilon_t^D, \epsilon_t^{PD}) \\ P_t = h_3(\epsilon_t^S, \epsilon_t^D, \epsilon_t^{PD}) \end{cases} \tag{6}$$

In this model, the variables for monthly global oil production ($OilProd_t$), global real economic activity (REA_t), and the price of oil (P_t) are each affected by the same three structural shocks: a flow supply shock (ϵ_t^S), a flow demand shock (ϵ_t^D), and precautionary demand shock (ϵ_t^{PD}).

In this setting, the precautionary demand shock corresponds to the oil price movements that cannot be explained by shocks to *current* or *flow* supply or demand. In particular, many developments have little bearing on current oil supply or oil demand for immediate use, but they can still affect oil prices immediately via expectations of future oil market conditions.

For example, oil prices often move in response to new economic projections published by organizations including the International Monetary Fund or the International Energy Agency. These price movements reflect a change in expected future oil demand instead of a change in flow oil demand. Similarly, oil prices often move sharply after OPEC announcements about changes in future supply targets (as explored in Kanzig (2021)). Again, these oil price movements reflect changing expected future supply conditions, and not changes in the current flow supply or flow demand for oil. In these two examples, as long as oil prices are moving in response to developments not captured by our measures of supply and global

real economic activity, the oil model described in equation (6) would attribute the price movements to precautionary demand shocks. Notably, price movements could be driven by changes in expected future supply or expected future demand. Yet, the model in equation (6) would bundle them together into the precautionary demand shock.

To fully decompose oil price changes into supply- and demand-driven components, we must unpack the oil-specific demand shock. The existing literature shows this decomposition is important because GDP and inflation are affected differently by supply-driven and demand-driven oil price changes, and a better understanding of the decomposition can help determine the optimal monetary policy response.

Our model incorporates 5 structural shocks: positive supply developments shock, $\epsilon_t^{S^+}$, negative supply developments shock, $\epsilon_t^{S^-}$, positive demand developments shock, $\epsilon_t^{D^+}$, negative demand developments shock, $\epsilon_t^{D^-}$, and a non-fundamentals price shock, ϵ_t^{NFP}. Separating the positive and negative shocks instead of using a measure that nets them out has several advantages. In particular, the positive and negative supply shocks may have different dynamic effects, and the additional degrees of freedom can allow for better identification of these effects. For example, suppose there is news of both an increase in OPEC production targets and a pipeline disruption on the same day. While that day's net price effect may be zero, the dynamic effects of these two shocks may be very different than that of a day with no shocks. More generally, it is useful to think of the oil market as a combination of many players and participants, which are constantly being hit by both oil supply and demand shocks.

The fifth shock, which we call the non-fundamentals price shock, captures oil price movements that are not explained by our supply and demand indexes. If our indexes capture all relevant oil supply and demand developments, then any price movements that are not explained by the indexes must be due to non-fundamental factors, such as risk aversion.

In the model, we assume that, contemporaneously, our indexes only interact with each other through the price of oil. However, the price of oil responds to all four indexes contemporaneously. Abstracting from the dynamic effects, the model we propose can be written as follows:

$$\begin{cases} SI_t = \alpha_P^{SI} P_t + \epsilon_t^{SI} \\ SD_t = \alpha_P^{SD} P_t + \epsilon_t^{SD} \\ DI_t = \alpha_P^{DI} P_t + \epsilon_t^{DI} \\ DD_t = \alpha_P^{DD} P_t + \epsilon_t^{DD} \\ P_t = \gamma_{SI}^P SI_t + \gamma_{SD}^P SD_t + \gamma_{DI}^P DI_t + \gamma_{DD}^P DD_t + \epsilon_t^{Resid} \end{cases} \tag{7}$$

$$\begin{cases} S_t = \alpha_P^S P_t + \epsilon_t^S \\ D_t = \alpha_P^D P_t + \epsilon_t^D \\ P_t = \gamma_S^P S_t + \gamma_D^P D_t + \epsilon_t^{Resid} \end{cases} \tag{8}$$

$$\begin{cases} NS_t = SI_t - SD_t = (\alpha_P^{SI} - \alpha_P^{SD}) P_t + (\epsilon_t^{S^+} - \epsilon_t^{S^-}) \\ ND_t = DI_t - DD_t = (\alpha_P^{DI} - \alpha_P^{DD}) P_t + (\epsilon_t^{D^+} - \epsilon_t^{D^-}) \\ P_t = \gamma_S^P (SI_t - SD_t) + \gamma_D^P (DI_t - DD_t) + \epsilon_t^{NFP} \end{cases} \tag{9}$$

Assume $\gamma_S^P = -\gamma_D^P$, use sign restriction

Note that if we assume the supply increase and supply decrease indexes have a symmetric effect on oil prices ($\gamma_{SI} = -\gamma_{SD} = \gamma_S$), and the demand increase and demand decrease indexes also have a symmetric effect on oil prices ($\gamma_{DI} = -\gamma_{DD} = \gamma_D$), we can rearrange the

5-variable model from equation (8) into a 3-variable one:

$$\begin{cases} SI_t - SD_t = (\alpha_P^{SI} - \alpha_P^{SD})P_t + (\epsilon_t^{S^+} - \epsilon_t^{S^-}) \\ DI_t - DD_t = (\alpha_P^{DI} - \alpha_P^{DD})P_t + (\epsilon_t^{D^+} - \epsilon_t^{D^-}) \\ P_t = \gamma_S^P(SI_t - SD_t) + \gamma_D^P(DI_t - DD_t) + \epsilon_t^{NFP} \end{cases} \tag{10}$$

While this model looks more like the one in equation (6), a key difference is that in this model, the supply and demand variables contain information about current and prospective supply and demand conditions, respectively, rather than just incorporating data on flow-supply and flow-demand. However, the results of the tests of symmetry shown in Tables 5 and 6 do not support the symmetry restrictions required to obtain this nested model. Additionally, whereas the assumptions we make to identify the structural shocks are sufficient in the 5-variable model, this 3-variable model would require additional identification assumptions. Estimation of this model is therefore left for later work.

4.2 Reduced Form Oil Price Shocks

In this section, we examine the reduced form shocks that are obtained using the oil price equation in models similar to those presented in the previous section. In particular, we estimate the following models and then compare the residuals that are obtained from each one of them:

$$Baseline \quad : \quad \Delta P_t \quad = \quad \alpha \; + \; \sum_{i=1}^{L}(\rho_i^P \Delta P_{t-i} \; + \; \rho_i^S \Delta S_{t-i} \; + \; \rho_i^{REA} \Delta REA_{t-i}) \; + \; \omega_t^P \tag{11}$$

$$Augmented : \Delta P_t = \alpha + \sum_{i=1}^{L} (\rho_i^P \Delta P_{t-i} + \rho_i^S \Delta S_{t-i} + \rho_i^{REA} \Delta REA_{t-i} + \rho_i^{SI} SI_{t-i} + \rho_i^{SD} SD_{t-i} + \rho_i^{DI} DI_{t-i} + \rho_i^{DD} DD_{t-i}) + \omega_t^P \quad (12)$$

$$Indexes : \Delta P_t = \alpha + \sum_{i=1}^{L} (\rho_i^P \Delta P_{t-i} + \rho_i^{SI} SI_{t-i} + \rho_i^{SD} SD_{t-i} + \rho_i^{DI} DI_{t-i} + \rho_i^{DD} DD_{t-i}) + \omega_t^P \quad (13)$$

The first equation corresponds to the price equation in a standard 3-variable VAR model of the oil market (as in equation (6)); the second equation adds our indexes to the first equation; and the third equation includes only our indexes, which corresponds to the price equation of the 5-variable VAR model of equation (8). We also consider versions of equations (12) and (13) in which we use the net supply and net demand indexes discussed previously. To compare across models, we estimate equations (11), (12), and (13) and compute the residuals. Next, for each regression, we calculate the standard deviation of the residuals as the sum of squared residuals divided by the number of observations minus one, and the root mean square error (RMSE) as the sum of squared residuals divided by the degrees of freedom of the regression, to adjust for the different numbers of right hand side variables across specifications. To compare across models, we compute the deviation of the results for the Augmented and Indexes models (both the directional and the net indexes) relative to those of the Baseline model. The results of these exercises are shown in Table 7.

In looking at the results, we first note that most of the entries in this table are negative, which suggests that in most cases the inclusion of the indexes reduces the average size of the residuals relative to the Baseline model. Focusing on the first two columns of the table, we note that adding the directional or net indexes to the baseline will always reduce the standard error, because adding more variables to a regression model always improves the fit. That said, when looking at the RMSE of these models, which adjusts for the added

variables, we still generally find improvement in model fit, indicating that our directional and net indexes provide additional explanatory information for price movements relative to the baseline model.

Turning to the last two columns of the table, we now compare the baseline model to those using only the indexes. Here, we find that the directional indexes tend to improve fit as measured by the Standard Error, while the net index results are more mixed. When looking at the RMSE measure, we find that we have a slight deterioration in fit of the models when using fewer lags, but an improvement when using more lags. Overall, we conclude that using the indexes may have slightly worse fit in some instances, perhaps due to the imprecision of these measures relative to data on oil production and real economic activity. However, the indexes do provide informational gain, especially when using the four directional indexes and when using net indexes with greater lags. We posit that these gains are likely related to the strength of the indexes in capturing information about prospective oil supply and prospective economic activity.

The results in Table 7 provide additional evidence that the oil market indexes we construct in this paper have relevant oil market information, which goes beyond the information contained in measures of flow oil supply and current economic activity.

4.3 Structural VAR Results

Following the discussion of the SVAR model in the previous sub-section, we now discuss the estimation results of the short-run dynamics in the model, forecast error-variance decomposition of the structural shocks, the impulse response functions, and historical decomposition of oil price movement during certain economic episodes.

4.3.1 Short-Run Dynamics

In Table 8 we show the estimates of the short-run relationships between each of the five variables included in the monthly (Panel A) and weekly (Panel B) models, as outlined in equation (8). For both models, the estimated relationships are in line with economic theory. Additionally, though the coefficient magnitudes vary across the monthly and weekly models, the results are qualitatively very similar: The supply increase (column 1) and the demand decrease (column 4) indexes respond positively to oil price increases, the supply decrease (column 2) and the demand increase (column 3) indexes respond negatively. In both models, the oil price (column 5) responds negatively to the supply increase and demand decrease indexes, and responds positively to the supply decrease and demand increase indexes. With the exception of the response of the demand increase index in the monthly model, all estimates are statistically significant.

4.3.2 Oil Price Forecast Error Variance Decomposition

Table 9 shows for the monthly and weekly models the forecast error variance decomposition of oil prices in the short run ($t = 1$) and in the long run ($t = \infty$) in response to the five structural shocks in the model. The table also reports the sum of the contributions of the two directional supply shocks (S), and the sum of the two directional demand shocks (D).

For both the monthly and weekly models, we find that the supply shocks are most important by a small margin. In the weekly model, for example, supply shocks account for about 45 percent of oil price fluctuations in the short and long run. The main difference between the monthly and weekly models is that demand shocks have a smaller contribution in the weekly model, accounting for only 15 percent of oil price fluctuations in the short run and close to 25 percent in the long run. Focusing on the directional contributions, in the monthly model,

supply increase and supply decrease have about the same importance, but demand decrease shocks are more important than demand increase shocks, especially in the long run.

The importance of non-fundamentals price shocks for oil price fluctuations is a bit higher in the short run than in the long run, which is consistent with the idea that movements driven by non-fundamentals often unwind and become less important over time. The contribution of non-fundamentals remains substantial, however, generally ranging from around 30 to 40 percent.

The relative importance of supply and demand shocks for oil price fluctuations is still an open question in the literature (see, for example, Kilian (2009), Caldara et al. (2019), and Baumeister and Hamilton (2019)). Our results suggest that the frequency and time horizon of the oil price fluctuations may also play an important role in this discussion. In particular, we may find different relative importance of supply and demand shocks when using weekly or monthly frequency data and when considering near and medium term horizons.

4.3.3 Structural Impulse Response Functions

Figure 5 shows the impulse response function of oil prices to each of the five structural shocks based on the monthly (panel a) and weekly (panel b) models.[4] Despite some small differences between the two sets of impulse response functions, the response of oil prices to supply and demand shocks is very similar in both models. Consistent with our earlier results, the price of oil declines in response to either a supply increase or a demand decrease shock, and rises in response to either a supply decrease or a demand increase shock.

With respect to the non-fundamentals price shock, the price of oil price first increases, then converges to near zero for both the monthly and weekly models. The fact that the the non-fundamentals price shock has a non-permanent effect on the price of oil is, in our

[4]Figures D1 and D2 in the Appendix show all the structural impulse response functions associated with the monthly and weekly SVAR models, respectively.

view, consistent with our interpretation of this shock. After controlling for the most relevant supply and demand developments, oil prices may move in response to some non-fundamentals developments, but these movements are expected to be short-lived.

All in all, the impulse response functions based on the SVAR model presented in equation (8) accord with standard economic theory. These results provide further evidence in favor of the oil supply and demand indexes that we construct.

4.3.4 Historical Decomposition of Oil Price Movements

Using our structural model of the oil market represented by equation (8), we can separate oil price movements into their supply- and demand-driven components. This decomposition is particularly useful for guiding monetary policy responses, as monetary authorities are more likely to take a signal and adjust the monetary policy stance in reaction to rising oil prices when the oil price increases are demand-driven rather than supply-driven. The rationale is that the demand pressures that are pushing oil prices higher are also likely to lead to broader inflation as well. In contrast, if oil prices are rising due to supply-side developments, there is much less concern about overall inflation, as other prices don't have the same reasons to move up.

To demonstrate the value of the historical decomposition of oil price changes obtained from our model, we next analyze three episodes with very large oil price changes: 1) the 2007-2009 Great Recession, 2) the 2014 oil price collapse, and 3) the 2020 oil price collapse and recovery from the COVID-19 pandemic.

The 2007-2009 Great Recession

As the Great Recession was among the most severe global recessions, it is perhaps unsurprising that it was also a time of an major oil price collapse. On net, oil prices declined over 25 percent, from about $95 per barrel in November 2007, just before the start of the recession,

to \$70 per barrel at the end of the recession in June 2009. Even more astounding is that this net price decline masks even greater oil price volatility. Prices rose sharply through the start of the recession, reaching over \$130 per barrel in June 2008. Over the following months, oil prices collapsed by over 70% to just under \$40 per barrel in February 2009.

Following Burbidge and Harrison (1985), we use the SVAR model to decompose the the observed oil price change into the contribution of each of the five structural shocks. The main results can be seen in the top panel of Figure 6. Additionally, to summarize the net effects of the positive and negative shocks, the lower panel combines the contributions of the directional shocks into the supply and demand components.

Notably, as shown by our impulse response analysis, shocks at time t can have an effect long afterwards. Therefore, to better understand any specific historical episode, it is useful to consider the continuing effects of the shocks that occurred prior to the start of the episode. For example, in our analysis of the June 2008 to December 2008 oil price collapse as shown in Figure 6, we decompose the total price change into the "expected" effects of shocks that occurred through June 2008, and then the "unexpected" shocks that occurred in July 2008 and after.

According to the results of the monthly and the weekly models, the collapse of oil prices between June and December 2008 was mostly due to demand and non-fundamentals price shocks. About three quarters of the price collapse was unexpected, with mostly demand shocks driving the prices down after June 2008. For the expected price declines after June 2008 shown in the second bar of the top panel, we find that oil prices were already being affected by some negative demand shocks and some non-fundamentals price shocks as well.

To further investigate these results, we show in Figure 7 the price decomposition for the period November 2007 to June 2008. We find that oil prices increased about 40%, mostly driven by non-fundamentals price shocks, and, to a less extent by supply and demand shocks. In particular, we find a large role for non-fundamentals price shocks pushing prices higher

before July 2008. The impulse response function of oil prices shows that, in response to a non-fundamentals price shock, oil prices increase at first and decline later (see Figure 5), with the long-term effect on prices being basically neutral. This is exactly what we see happening in this episode: between November 2007 and June 2008, oil prices increased mostly due non-fundamental reasons (that is, the supply and demand developments did not justify this large increase in oil prices), and this price increase reversed and contributed to later price declines between June and December 2008.

The 2014 Oil Price Collapse

Starting in mid-2014, oil prices plunged by more than 70 percent, from just over $100 per barrel in May 2014 to $30 per barrel in February 2016. In Figure 8 we show the historical decomposition of this oil price change based on the monthly and weekly SVAR models.

As before, the top panel shows the results for the five structural shocks while the lower panel shows the results after combining the positive and negative shocks to supply and demand. Our model shows that supply shocks, in particular supply increase shocks, were the main driver of the price decline, with a significant but slightly smaller role for non-fundamental price shocks as well.

Separating between expected and unexpected shocks shows that about half of the oil price decline observed after May 2014 was the result of shocks that occurred prior to the start of the episode. In the middle bar for each set of 3, it can be seen that supply shocks that occurred through May 2014 were still reducing prices. The middle bars also show that the non-fundamentals price shocks that hit the oil market through May 2014 were contributing to the oil price declines later on. Figure 9 takes a closer look at these, showing the historical decomposition of oil price changes in the year leading up to the oil price collapse of 2014. Interestingly, it shows there were unexpected positive non-fundamentals price shocks that helped maintain the oil price higher through May 2014, consistent with the negative effects driven by the unwinding non-fundamentals shocks in the months to follow. Overall, we find

that though oil prices were mostly unchanged between April 2013 and May 2014, there were some non-fundamentals price shocks occurring during this period that were pushing oil prices up, which lowered prices after June 2014.

To summarize, the mid-2014 oil price collapse was due to a combination of supply and non-fundamentals price shocks. Most most non-fundamentals price shocks occurred on or before June 2014 while supply shocks occurred before, on, and after June 2014.

The 2020 Pandemic Oil Price Collapse and Recovery

As a result of worldwide lockdowns, which were adopted by countries in response to the COVID-19 pandemic, oil prices collapsed at the start of 2020. According to the results in Figure 10, the massive decline in oil prices that occurred from December 2019 to April 2020 was mostly the result of demand decrease shocks, which are consistent with the sharp decline in global economic activity resulting from the lockdowns. Contributions from non-fundamentals also play a role, likely reflecting negative risk sentiment. On the supply front, the lack of agreement between OPEC and Russia on how to adjust production in light of the depressed demand and oil price also contributed to the decline in oil prices. Notably, Figure 10 shows that the vast majority of the oil price collapse was driven by unexpected shocks, consistent with the idea that the pandemic and its effects were unexpected as of December 2019.

Next, we turn to Figure 11, which shows the decomposition of oil prices corresponding to the oil price recovery that happened from April 2020 through December 2020. The results indicate that some of the recovery was just the undoing of part of the shocks that drove prices down in the earlier part of the year – this result can be seen in the middle bars of Figure 10, which shows the expected variation in oil prices given previous shocks. Some of the recovery was also driven by unexpected positive demand shocks, consistent with the unexpected speed of the global economic recovery. Finally, we also see some unexpected positive contributions from supply shocks, consistent with restraint by OPEC in not increasing supplies too rapidly.

5 Concluding Remarks

In this paper, we construct several indexes tracking developments in the oil market based on textual analysis of two highly regarded daily energy-related publications. We show that our indexes, which separate developments that are expected to increase oil supply or demand from those that are expected to decrease oil supply or demand, are correlated in the expected way with observed oil production changes, changes in real economic activity conditions, and oil price changes. We also show that our indexes contain information regarding current and prospective oil market developments, a feature that is unique relative to most of the existing measures oil market developments. Moreover, our findings suggest that important information about the economy can be obtained from specialized industry publications. Indeed, we found that our indexes had significant predictive power of oil prices changes at longer horizons. By not accounting for this information the economist may consider certain developments in the oil market, or in other markets, as unexpected while these developments could be predicted given past developments. Other key characteristics of our indexes are that they can be updated on a daily basis and that these measures can be constructed for different time frequencies.

In an application, we use our indexes to estimate monthly and weekly structural VAR models of the oil market. For both models, the estimated relationships between all the variables included in the model and the structural shocks are in line with economic theory. When using these structural models to decompose large oil price movements around the 2008-09 Great Recession and the 2014 oil price collapse, we find that a large portion of these oil price movements was expected prior to the start of the episodes. By contrast, for the the 2020 pandemic-related oil price collapse and recovery episode we find that the vast majority of the shocks that drove prices in 2020 occurred in 2020. For the three episodes, the resulting historical decomposition of oil price movements during these episodes are plausibly in line with what one would expect to happen to oil around price around an economic recession,

after a large increase in oil production, and after a massive decline and subsequent recovery of economic activity as observed in 2020 due to the COVID-19 shock.

In future work we plan to use our oil supply and demand indexes to estimate oil production and real economic activity in real time. We also would like to use our oil supply and demand indexes to forecast oil prices and compare those projections against other alternative forecasts, such as those based on oil futures.

References

Baker, S. R., Bloom, N., and Davis, S. J. (2016). Measuring economic policy uncertainty. *The Quarterly Journal of Economics*, 131(4):1593–1636.

Baumeister, C. and Hamilton, J. D. (2019). Structural interpretation of vector autoregressions with incomplete identification: Revisiting the role of oil supply and demand shocks. *American Economic Review*, 109(5):1873–1910.

Brandt, M. W. and Gao, L. (2019). Macro fundamentals or geopolitical events? a textual analysis of news events for crude oil. *Journal of Empirical Finance*, 51:64 – 94.

Burbidge, J. and Harrison, A. (1985). An historical decomposition of the great depression to determine the role of money. *Journal of Monetary Economics*, 16(1):45–54.

Caldara, D., Cavallo, M., and Iacoviello, M. (2019). Oil price elasticities and oil price fluctuations. *Journal of Monetary Economics*, 103:1–20.

Caldara, D. and Iacoviello, M. (2018). *Measuring geopolitical risk*. International Finance Discussion Papers 1222.

Calomiris, C. W., Cakir Melek, N., and Mamaysky, H. (2020). Mining for oil forecasts. *Federal Reserve Bank of Kansas City Research Working Paper No. 20-20*.

Cascaldi-Garcia, D. and Vukoti, M. (2020). Patent-based news shocks. *The Review of Economics and Statistics*, 0(ja):1–45.

Cavallo, M. and Wu, T. (2012). *Measuring oil-price shocks using market-based information*. International Monetary Fund Working Paper No. 12/19.

Cloyne, J. (2013). Discretionary tax changes and the macroeconomy: new narrative evidence from the united kingdom. *American Economic Review*, 103(4):1507–28.

Hamilton, J. D. (2019). Measuring global economic activity. *Journal of Applied Econometrics*.

Huettner, F., Sunder, M., et al. (2012). Axiomatic arguments for decomposing goodness of fit according to shapley and owen values. *Electronic Journal of Statistics*, 6:1239–1250.

Husted, L., Rogers, J., and Sun, B. (2019). Monetary policy uncertainty. *Journal of Monetary Economics*.

Kanzig, D. R. (2021). The macroeconomic effects of oil supply news: Evidence from opec announcements. *American Economic Review*, 111(4):1092–1125.

Kilian, L. (2009). Not all oil price shocks are alike: Disentangling demand and supply shocks in the crude oil market. *American Economic Review*, 99(3):1053–69.

Kilian, L. and Murphy, D. P. (2014). The role of inventories and speculative trading in the global market for crude oil. *Journal of Applied Econometrics*, 29(3):454–478.

Loughran, T., McDonald, B., and Pragidis, I. (2019). Assimilation of oil news into prices. *International Review of Financial Analysis*, 63:105–118.

Romer, C. D. and Romer, D. H. (2004). A new measure of monetary shocks: Derivation and implications. *American Economic Review*, 94(4):1055–1084.

Romer, C. D. and Romer, D. H. (2010). The macroeconomic effects of tax changes: estimates based on a new measure of fiscal shocks. *American Economic Review*, 100(3):763–801.

A Figures and Tables

Figure 1: Monthly average of articles per issue in the "Oil Daily" and "International Oil Daily", before and after filtering out duplicates and articles unrelated to the oil market.

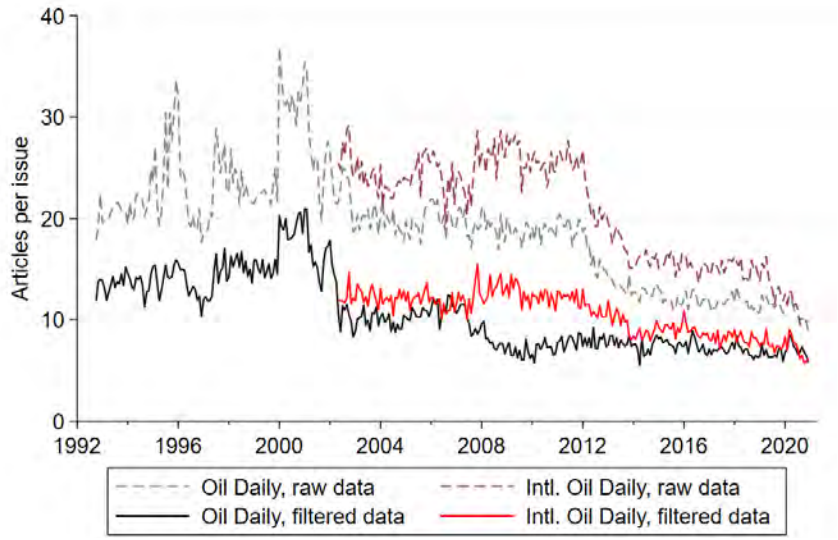


Figure 2: Supply and demand counts per month

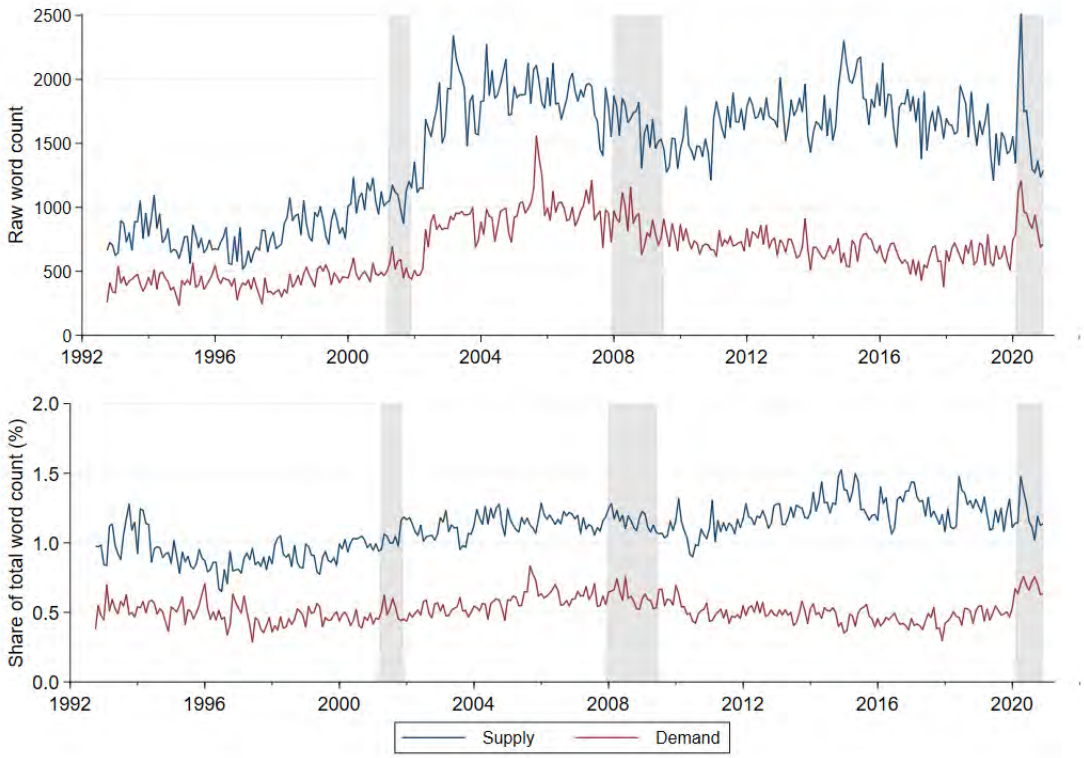


Figure 3: Share of phrase counts per month

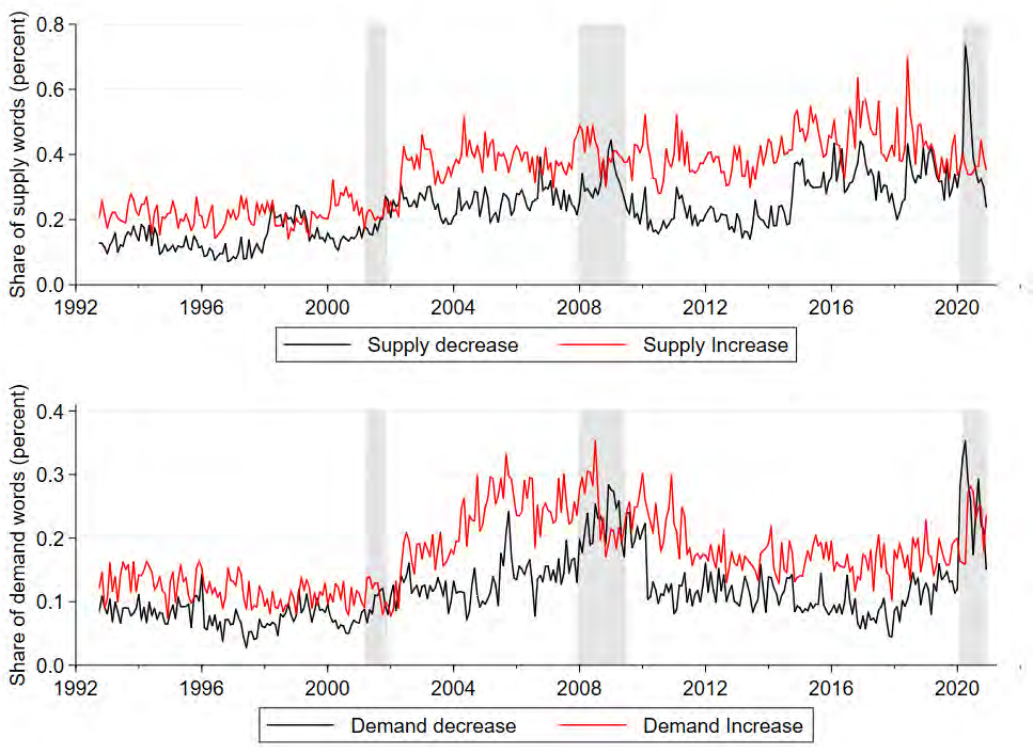


Figure 4: Net supply and net demand word counts per month in the "Oil Daily" and "International Oil Daily" publications combined - standardized.

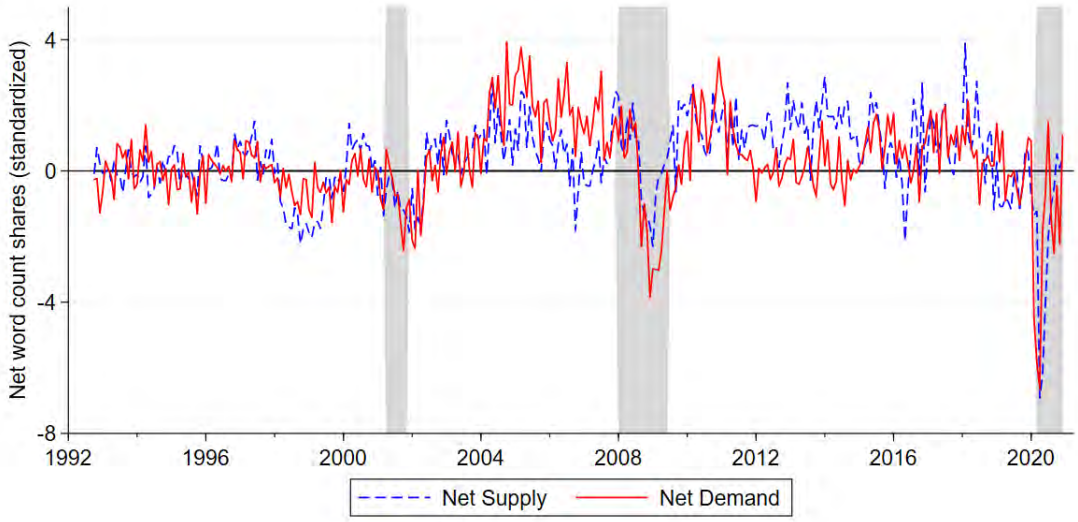
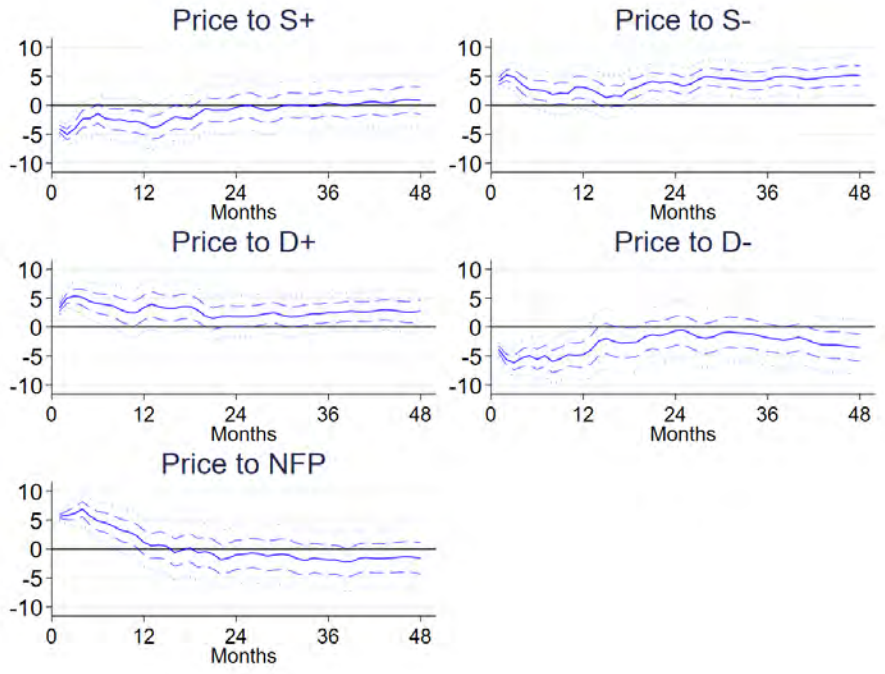


Figure 5: Oil price impulse response to supply demand, and non-fundamentals price shocks - monthly and weekly models

(a) Monthly model



(b) Weekly model

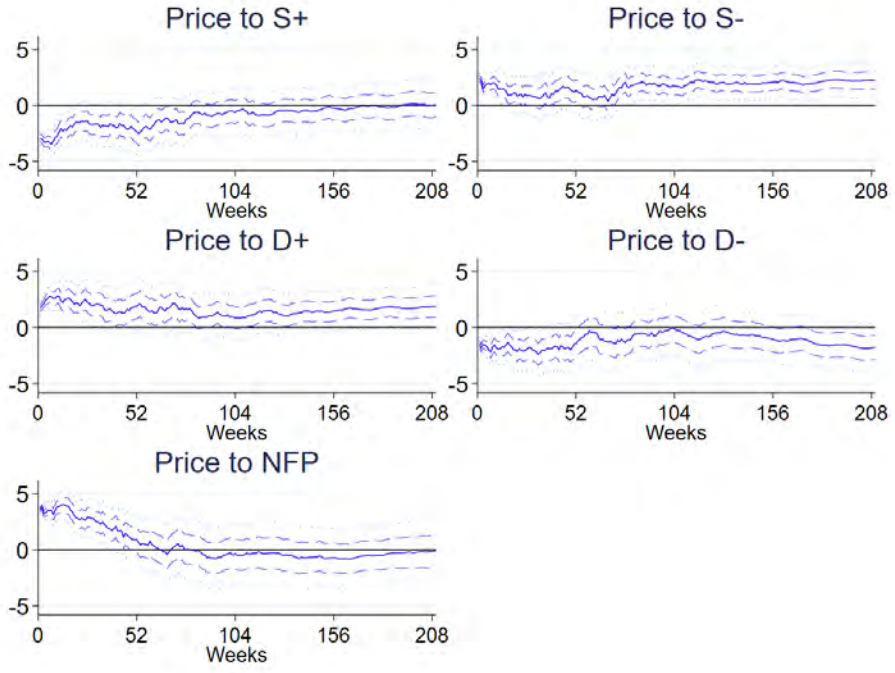
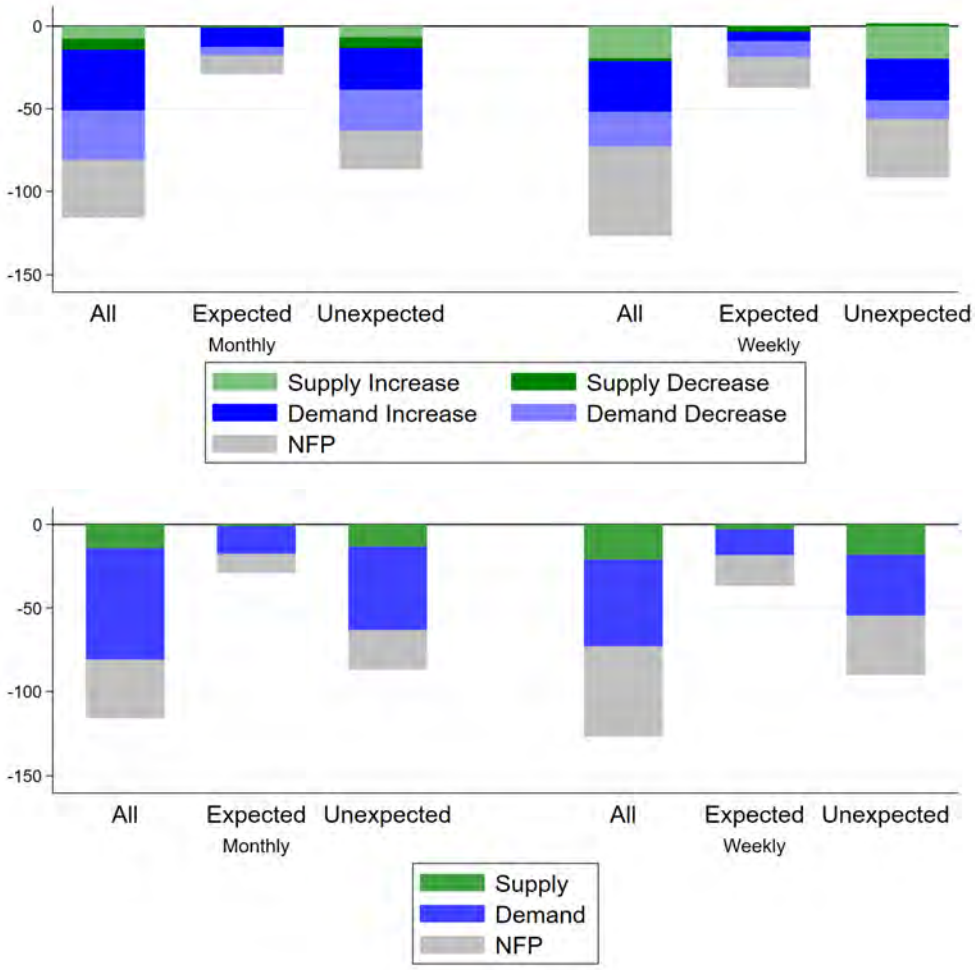
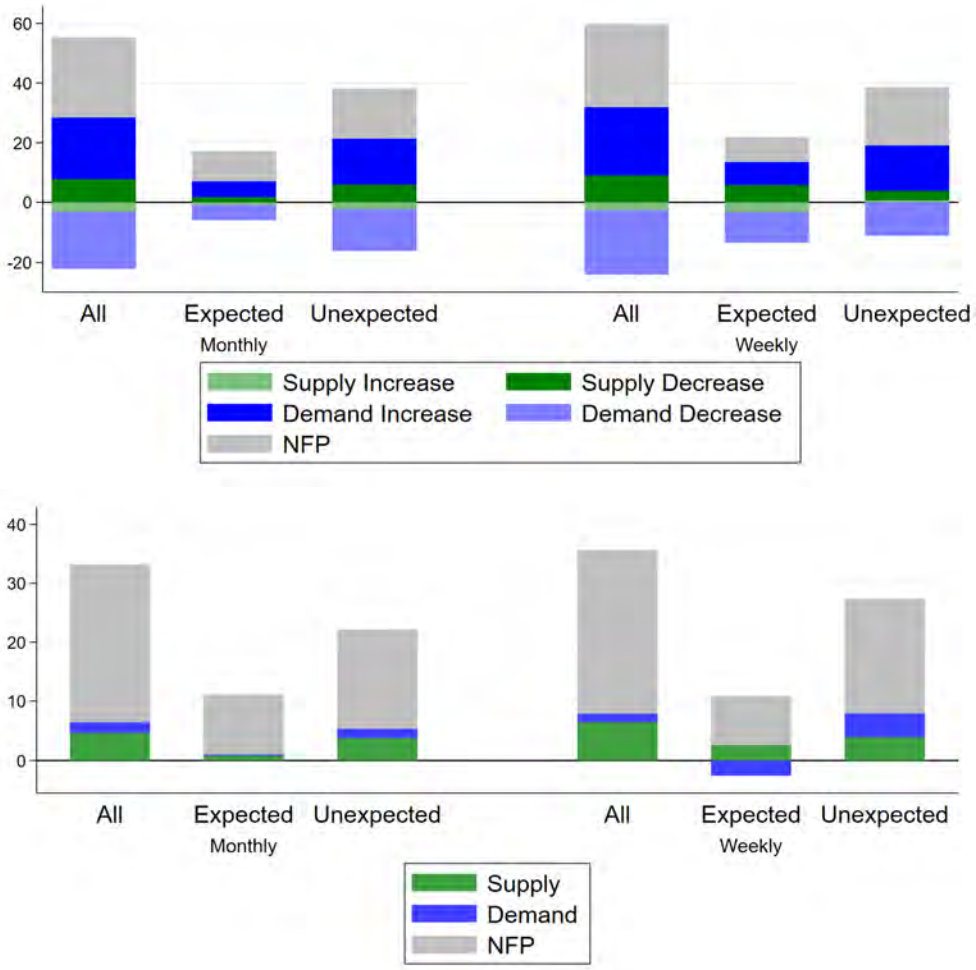


Figure 6: Oil Price Change Decomposition during the 2007-09 Great Recession - June 2008 to December 2008



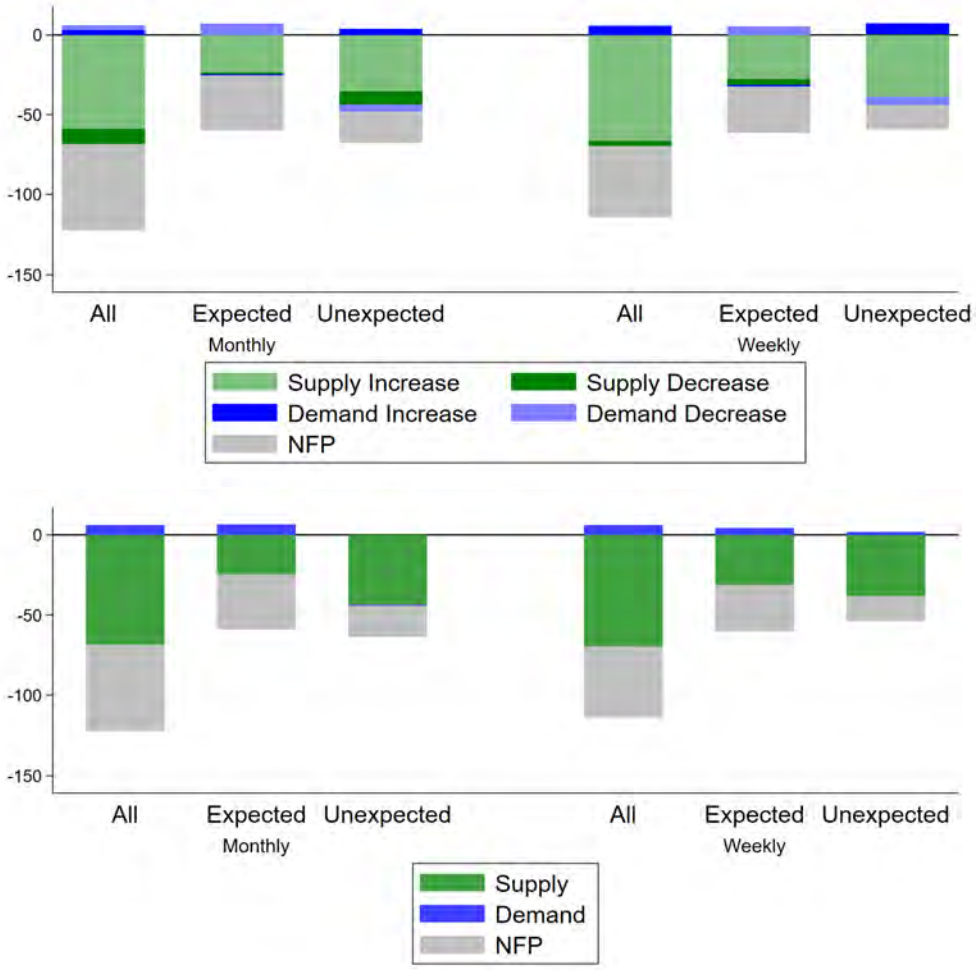
Note: Each bar corresponds to the sum of all the structural shocks causing oil prices to change between June and December 2008. The results shown correspond to log-price changes, as an approximation for percent changes. For very large changes, as in our episodes of interest, this approximation can lead to log-price changes larger than 100 percent.

Figure 7: Oil Price Change Decomposition during the 2007-09 Great Recession - November 2007 to June 2008



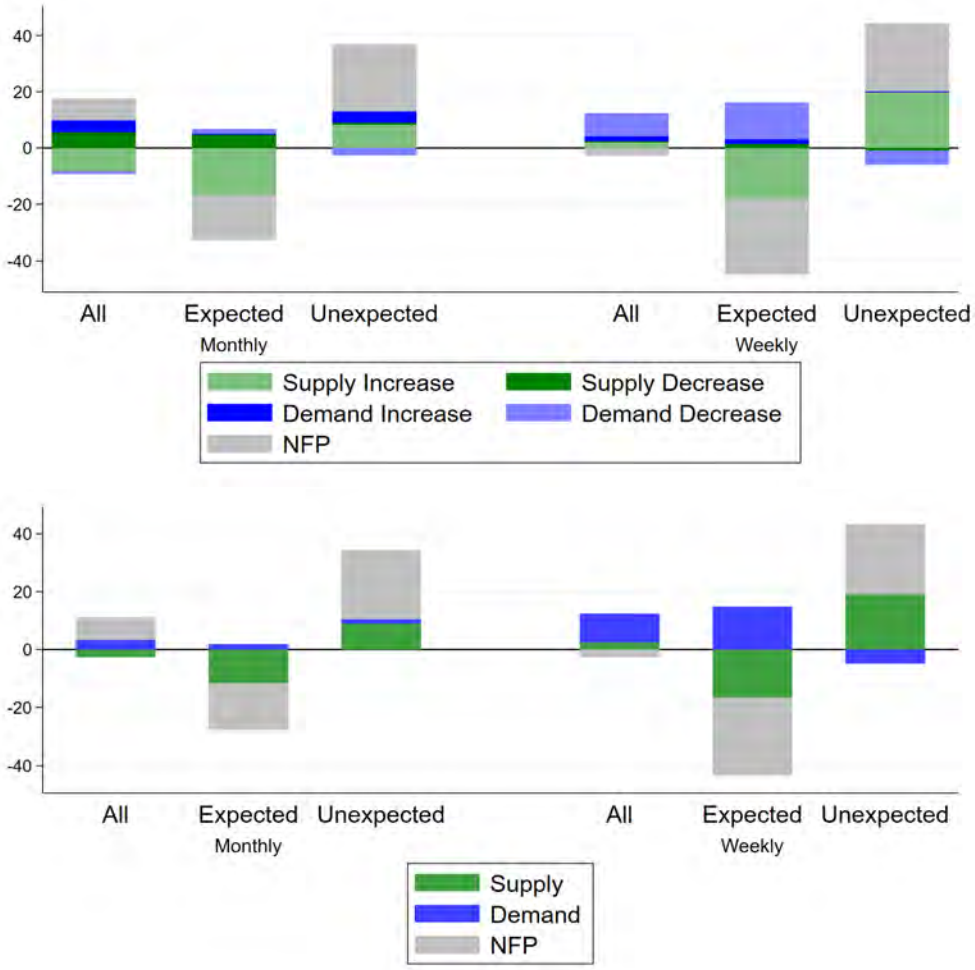
Note: Each bar corresponds to the sum of all the structural shocks causing oil prices to change between November 2007 and June 2008. The results shown correspond to log-price changes, as an approximation for percent changes. For very large changes, as in our episodes of interest, this approximation can lead to log-price changes larger than 100 percent.

Figure 8: Oil Price Change Decomposition of the 2014 Oil Price Collapse



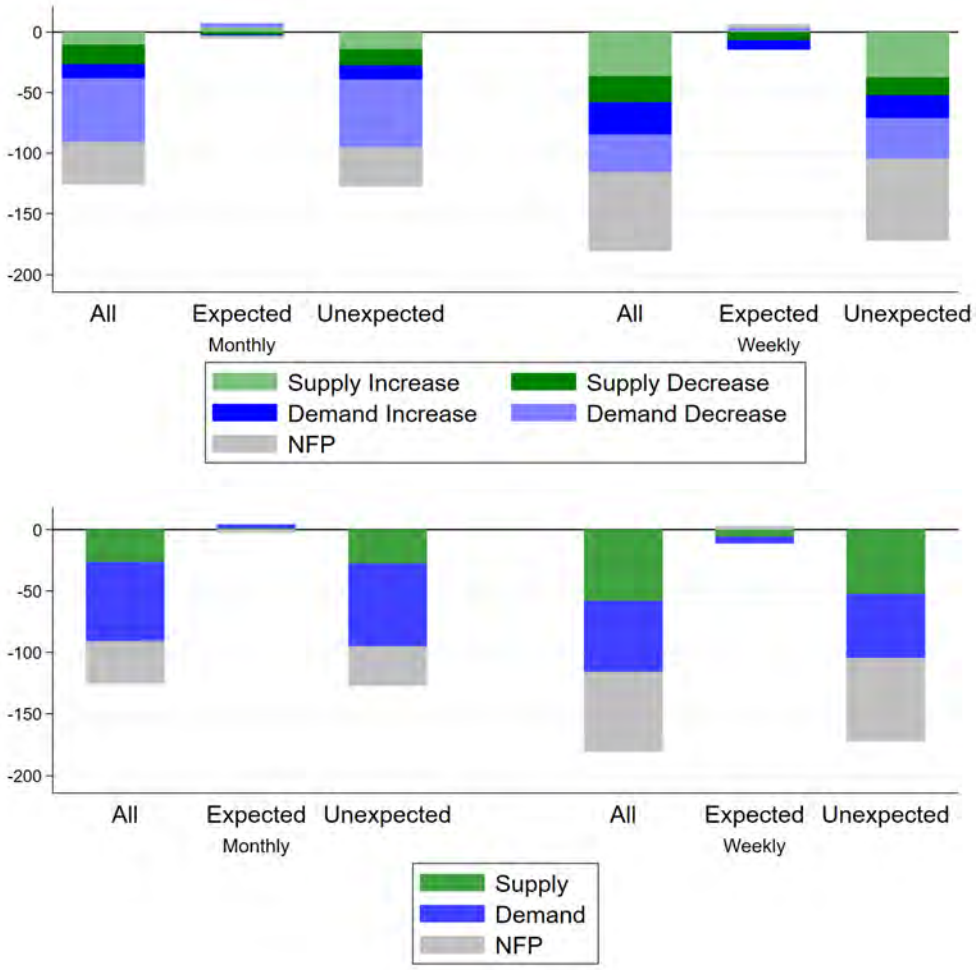
Note: Each bar corresponds to the sum of all the structural shocks causing oil prices to change between May 2014 and February 2016. The results shown correspond to log-price changes, as an approximation for percent changes. For very large changes, as in our episodes of interest, this approximation can lead to log-price changes larger than 100 percent.

Figure 9: Oil Price Change Decomposition One Year Before the 2014 Oil Price Collapse



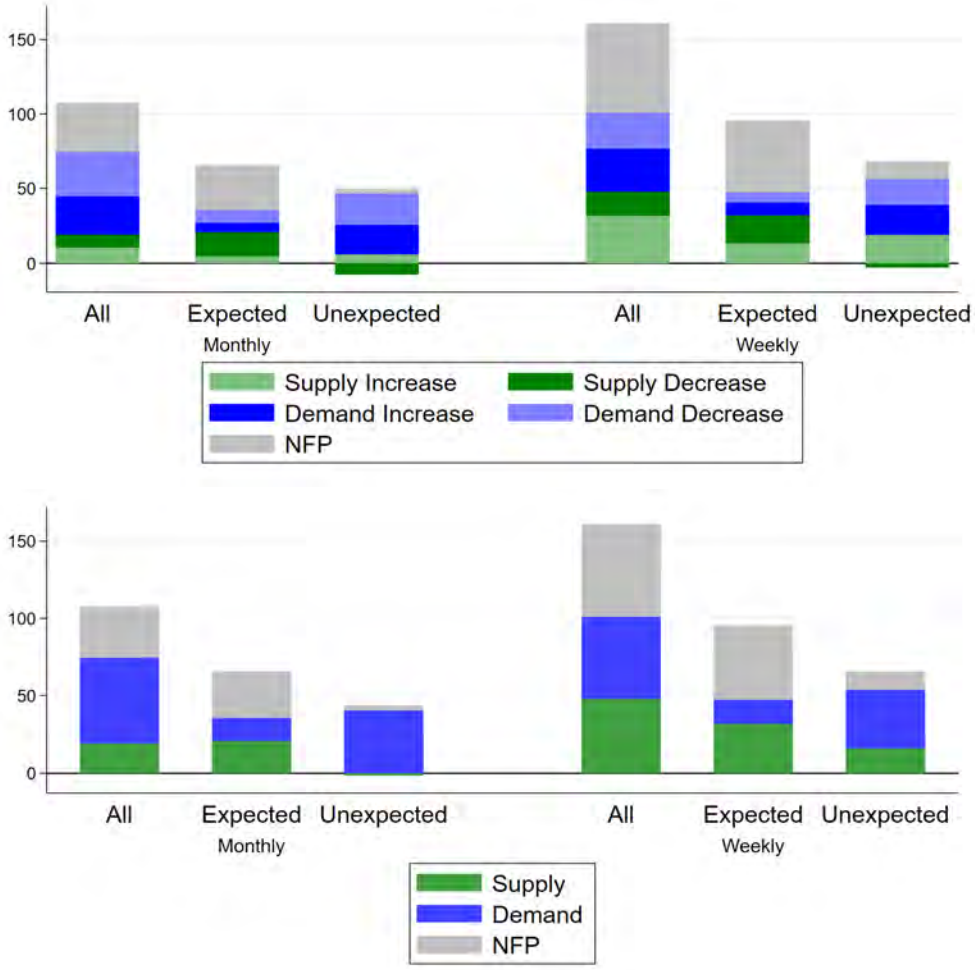
Note: Each bar corresponds to the sum of all the structural shocks causing oil prices to change between April 2013 and May 2014. The results shown correspond to log-price changes, as an approximation for percent changes. For very large changes, as in our episodes of interest, this approximation can lead to log-price changes larger than 100 percent.

Figure 10: Oil Price Change Decomposition of Oil Price Collapse At the Onset of the COVID-19 Pandemic



Note: Each bar corresponds to the sum of all the structural shocks causing oil prices to change between December 2019 and April 2020. The results shown correspond to log-price changes, as an approximation for percent changes. For very large changes, as in our episodes of interest, this approximation can lead to log-price changes larger than 100 percent.

Figure 11: Oil Price Change Decomposition of the Oil Price Recovery During COVID-19 Pandemic



Note: Each bar corresponds to the sum of all the structural shocks causing oil prices to change between April 2020 and December 2020. The results shown correspond to log-price changes, as an approximation for percent changes. For very large changes, as in our episodes of interest, this approximation can lead to log-price changes larger than 100 percent.

Table 1: Summary statistics of the "Oil Daily" and "International Oil Daily" publications

	Early Sample	Late Sample		Full Sample
	Oil Daily	OD	IOD	OD and IOD
Sample start date	Oct. 1992	Apr. 2002	Apr. 2002	Oct. 1992
Sample end date	Mar. 2002	Dec. 2020	Dec. 2020	Dec. 2020
Number of issues	2,335	4,783	4,767	11,885
Before article filtering				
Words per article	240	336	324	307
Articles per day	24	16	20	32
Words per day	5,764	5,418	6,604	9,955
Articles per year	5,613	3,854	4,863	9,219
Number of articles	56,134	77,072	97,269	230,475
After article filtering				
Percent of articles retained	62.3	55.1	47.6	53.7
Words per article	293	375	356	345
Articles per day	15	9	10	17
Words per day	4,392	3,331	3,461	5,997
Articles per year	3,180	2,124	2,316	4,126
Number of articles	34,977	42,478	46,326	123,781

Table 2: Top 5 most frequently used supply, demand, increase, and decrease words

Supply			Demand		
word	**freq.**	**share**	**word**	**freq.**	**share**
production	145,781	30.40%	demand	53,342	23.92%
output	50,062	10.44%	refinery	41,961	18.82%
supply	29,040	6.06%	refining	24,492	10.98%
reserves	28,503	5.94%	imports	21,312	9.56%
exports	28,041	5.85%	refineries	18,447	8.27%
Increase			**Decrease**		
word	**freq.**	**share**	**word**	**freq.**	**share**
more	98,593	11.90%	down	44,239	8.46%
up	90,302	10.90%	under	40,111	7.67%
over	71,397	8.62%	lower	28,510	5.45%
high	35,225	4.25%	low	22,782	4.36%
increase	35,180	4.25%	fell	21,948	4.20%

Table 3: Example of an Oil Daily article

"Crude Prices Fall on Opec Doubts, Demand Worries, Strong Dollar"
Oil Daily, 10/22/2008

"Oil prices gave back Monday's gains, falling by more than $3 on Tuesday, as traders expressed doubts that an Opec PRODUCTION cut will be enough to shore up prices. Saudi Arabia, Opec's largest PRODUCER, has not publicly commented on the group's upcoming meeting. The kingdom is the only Opec member that can make a substantial cut. While some of the other countries are calling for an OUTPUT cut, their statements do not carry as much weight as the Saudis. The Opec pre-meeting bounce will not hold, said a broker. It's another case of BUY the rumor and sell the fact. Opec Secretary-General Abdullah al-Badri said Tuesday that, There has been no decision to cut PRODUCTION so far. Nevertheless, analysts expect Opec to cut because of the sharp drop in both prices and DEMAND. The real question is how much will the group cut and will its members follow through and implement an agreement. In New York, Nymex light, sweet crude for November, in its last day as prompt month, fell by $3.36 to $70.89 per barrel. In London on ICE Futures, Brent crude lost $2.31 at $69.72/bbl. Besides uncertainty about Opec action, early declines in the US stock market also weighed on prices. Another factor which helped to drive down prices was a rally in the dollar, analysts said. But bearishness regarding DEMAND has overwhelmed everything else lately, even talk of Opec taking as much as 2 million barrels per day off the market. Matt Piotrowski, Washington"

Table 4: Monthly regressions: Oil production, real economic activity, and prices

$$ln(y_t) - ln(y_{t-1})$$

	Oil production		REA		WTI spot		WTI 12-month	
Supply increase	3.32		0.45		-14.75		-11.13	
	(0.81)***		(0.43)		(6.27)**		(5.22)**	
Supply decrease	-2.75		-1.25		6.58		3.91	
	(0.69)***		(0.41)***		(4.94)		(4.08)	
Demand increase	-1.19		1.65		30.73		26.37	
	(0.73)		(0.45)***		(7.65)***		(6.55)***	
Demand decrease	0.31		-1.38		-16.55		-14.31	
	(0.59)		(0.44)***		(5.32)***		(3.99)***	
Net supply		2.50		0.36		-10.19		-7.51
		(0.61)***		(0.32)		(4.77)**		(3.91)*
Net demand		-1.11		0.84		22.55		19.02
		(0.58)*		(0.36)**		(5.85)***		(4.85)***
$H_0 : \beta_{SI} + \beta_{SD} = 0$	0.261		0.002		0.051		0.013	
$H_0 : \beta_{DI} + \beta_{DD} = 0$	0.021		0.374		0.001		0.002	
$H_0 : \beta_{SI} + \beta_{SD} = 0$ & $\beta_{DI} + \beta_{DD} = 0$	0.066		0.007		0.003		0.005	
R^2	.272	.266	.423	.391	.314	.311	.312	.306
Indexes R^2 share	.146	.126	.245	.137	.119	.109	.219	.2
N	323	323	323	323	323	323	323	323

$$ln(y_{t+12}) - ln(y_t)$$

	Oil production		REA		WTI spot		WTI 12-month	
Supply increase	0.57		-0.80		-14.39		-10.52	
	(0.29)**		(0.38)**		(5.50)***		(3.90)***	
Supply decrease	-0.97		-0.40		8.57		5.66	
	(0.41)**		(0.51)		(5.31)		(3.79)	
Demand increase	0.08		1.36		13.58		11.25	
	(0.22)		(0.37)***		(3.91)***		(2.86)***	
Demand decrease	-0.07		-0.95		-7.50		-5.58	
	(0.17)		(0.45)**		(3.47)**		(2.50)**	
Net supply		0.49		-0.55		-10.69		-7.50
		(0.26)*		(0.37)		(4.32)**		(3.02)**
Net demand		-0.28		0.48		9.67		7.79
		(0.21)		(0.30)		(2.75)***		(2.11)***
$H_0 : \beta_{SI} + \beta_{SD} = 0$	0.073		0.001		0.147		0.055	
$H_0 : \beta_{DI} + \beta_{DD} = 0$	0.978		0.392		0.035		0.003	
$H_0 : \beta_{SI} + \beta_{SD} = 0$ & $\beta_{DI} + \beta_{DD} = 0$	0.184		0.003		0.095		0.008	
R^2	.241	.19	.322	.232	.267	.262	.215	.204
Indexes R^2 share	.317	.125	.265	.077	.372	.361	.432	.399
N	323	323	323	323	323	323	323	323

Note: *, **, and *** denote statistical significance at 10%, 5%, and 1%. Heteroskedasticity and autocorrelation corrected standard errors in parentheses.

Table 5: Monthly regressions: Oil production, real economic activity, and prices

$$ln(y_t) - ln(y_{t-1})$$

	Oil production		REA		WTI spot		WTI 12-month	
Supply increase	3.44		0.47		-15.36		-11.59	
	(0.84)***		(0.45)		(6.53)**		(5.45)**	
Supply decrease	-2.83		-1.28		6.94		4.22	
	(0.71)***		(0.44)***		(5.14)		(4.25)	
Demand increase	-1.22		1.72		31.99		27.25	
	(0.74)*		(0.44)***		(7.65)***		(6.64)***	
Demand decrease	0.34		-1.44		-17.29		-14.80	
	(0.60)		(0.45)***		(5.42)***		(4.08)***	
Net supply		2.51		0.35		-10.37		-7.66
		(0.61)***		(0.33)		(4.86)**		(4.01)*
Net demand		-1.14		0.86		23.32		19.55
		(0.57)**		(0.36)**		(5.83)***		(4.87)***
R^2	.271	.266	.424	.391	.315	.312	.313	.308
Indexes R^2 share	.146	.126	.246	.137	.12	.111	.219	.202
N	323	323	323	323	323	323	323	323
$\beta_{SI} + \beta_{SD} = 0$	0.245		0.001		0.047		0.012	
$\beta_{DI} + \beta_{DD} = 0$	0.024		0.357		0.001		0.001	
$\beta_{SI} + \beta_{SD} = 0$ & $\beta_{DI} + \beta_{DD} = 0$	0.076		0.004		0.003		0.004	

$$ln(y_{t+12}) - ln(y_t)$$

	Oil production		REA		WTI spot		WTI 12-month	
Supply increase	0.59		-0.80		-14.64		-10.77	
	(0.30)**		(0.40)**		(5.72)**		(4.05)***	
Supply decrease	-0.98		-0.42		8.75		5.83	
	(0.42)**		(0.53)		(5.47)		(3.90)	
Demand increase	0.07		1.41		14.04		11.75	
	(0.23)		(0.37)***		(3.96)***		(2.92)***	
Demand decrease	-0.06		-0.98		-7.65		-5.72	
	(0.18)		(0.47)**		(3.54)**		(2.56)**	
Net supply		0.49		-0.55		-10.54		-7.43
		(0.26)*		(0.37)		(4.35)**		(3.03)**
Net demand		-0.28		0.50		9.99		8.14
		(0.21)		(0.31)		(2.80)***		(2.17)***
R^2	.238	.189	.321	.232	.266	.261	.215	.205
Indexes R^2 share	.308	.124	.26	.074	.366	.353	.431	.396
N	323	323	323	323	323	323	323	323
$\beta_{SI} + \beta_{SD} = 0$	0.083		0.001		0.146		0.053	
$\beta_{DI} + \beta_{DD} = 0$	0.937		0.385		0.033		0.002	
$\beta_{SI} + \beta_{SD} = 0$ & $\beta_{DI} + \beta_{DD} = 0$	0.207		0.002		0.091		0.007	

Note: *, **, and *** denote statistical significance at 10%, 5%, and 1%. Heteroskedasticity and autocorrelation corrected standard errors in parentheses.

Table 6: Weekly regressions: Spot and futures prices

WTI Spot

	$ln(y_t) - ln(y_{t-1})$		$ln(y_{t+12}) - ln(y_t)$		$ln(y_{t+26}) - ln(y_t)$		$ln(y_{t+52}) - ln(y_t)$	
Supply increase	-17.72		-1.07		-5.60		-7.68	
	(8.15)**		(5.42)		(4.27)		(3.53)**	
Supply decrease	15.49		-3.12		1.14		4.51	
	(6.99)**		(4.51)		(3.54)		(3.43)	
Demand increase	29.70		5.28		4.46		3.79	
	(8.77)***		(3.48)		(2.71)*		(1.84)**	
Demand decrease	-17.71		-1.45		-2.35		-1.47	
	(7.58)**		(4.82)		(4.09)		(2.35)	
Net supply		-14.97		0.15		-4.14		-6.11
		(6.52)**		(4.43)		(3.50)		(2.97)**
Net demand		24.31		3.02		2.98		2.56
		(7.24)***		(3.52)		(2.69)		(1.58)
R^2	.124	.124	.136	.133	.137	.133	.112	.109
Indexes R^2 share	.076	.074	.025	.008	.1	.078	.383	.379
N	1355	1355	1355	1355	1355	1355	1355	1355
$\beta_{SI} + \beta_{SD} = 0$	0.744		0.026		0.000		0.000	
$\beta_{DI} + \beta_{DD} = 0$	0.117		0.068		0.169		0.012	
$\beta_{SI} + \beta_{SD} = 0$ & $\beta_{DI} + \beta_{DD} = 0$	0.269		0.055		0.002		0.001	

WTI 12-Month Futures

	$ln(y_t) - ln(y_{t-1})$		$ln(y_{t+12}) - ln(y_t)$		$ln(y_{t+26}) - ln(y_t)$		$ln(y_{t+52}) - ln(y_t)$	
Supply increase	-8.30		0.76		-2.96		-4.56	
	(6.15)		(3.26)		(2.64)		(2.25)**	
Supply decrease	7.53		-3.42		-0.74		2.25	
	(5.06)		(2.92)		(2.38)		(2.19)	
Demand increase	25.10		4.37		3.91		3.12	
	(7.08)***		(2.53)*		(1.96)**		(1.28)**	
Demand decrease	-16.28		-1.56		-1.93		-1.26	
	(5.76)***		(3.72)		(3.28)		(1.98)	
Net supply		-7.14		1.38		-1.89		-3.52
		(4.95)		(2.68)		(2.19)		(1.90)*
Net demand		20.93		2.70		2.53		2.10
		(5.83)***		(2.62)		(2.10)		(1.26)*
R^2	.114	.114	.098	.094	.099	.09	.072	.068
Indexes R^2 share	.146	.144	.07	.039	.129	.06	.387	.359
N	1355	1355	1354	1354	1354	1354	1355	1355
$\beta_{SI} + \beta_{SD} = 0$	0.863		0.041		0.000		0.000	
$\beta_{DI} + \beta_{DD} = 0$	0.085		0.078		0.086		0.007	
$\beta_{SI} + \beta_{SD} = 0$ & $\beta_{DI} + \beta_{DD} = 0$	0.175		0.077		0.000		0.000	

Note: *, **, and *** denote statistical significance at 10%, 5%, and 1%. Heteroskedasticity and autocorrelation corrected standard errors in parentheses.

Table 7: Percent deviations from baseline model's residuals standard errors and regression root mean squared error.

	Baseline + Directional Indexes	Baseline + Net Indexes	Directional Indexes Only	Net Indexes Only
Lag=12				
Standard Error	-9.2	-6.2	-1.9	0.8
RMSE	-0.6	-2.0	2.4	0.8
Lag=18				
Standard Error	-13.7	-8.6	-3.9	1.3
RMSE	1.0	-1.7	3.4	1.3
Lag=24				
Standard Error	-25.0	-14.6	-13.2	-1.5
RMSE	-3.4	-4.6	-3.1	-1.5

Note: Entries in this table are percent deviations from the standard error of the regression residuals and the root mean squared error (RMSE) of each of the models denoted in each column (equations (12) and (13)) relative to those of the baseline model (equation (11)).

Table 8: Structural VAR short-run dynamics parameters estimates.

	SI_t (1)	SD_t (2)	DI_t (3)	DD_t (4)	$\Delta ln(Price_t)$ (5)
Panel A: Monthly model					
SI_t	-	-	-	-	-14.500
					(2.655)***
SD_t	-	-	-	-	14.311
					(2.7)***
DI_t	-	-	-	-	11.883
					(1.709)***
DD_t	-	-	-	-	-8.323
					(1.13)***
$\Delta ln(Price_t)$	0.034	-0.041	-0.007	0.018	-
	(0.007)***	(0.007)***	(0.006)	(0.011)*	
Sample			1994M04 - 2020M12		
Number observations			321		
Panel B: Weekly model					
SI_t	-	-	-	-	-7.575
					(0.864)***
SD_t	-	-	-	-	6.119
					(0.608)***
DI_t	-	-	-	-	4.183
					(0.523)***
DD_t	-	-	-	-	-2.943
					(0.438)***
$\Delta ln(Price_t)$	0.089	-0.059	-0.038	0.033	-
	(0.012)***	(0.008)***	(0.009)***	(0.011)***	
Sample			4/06/1994 - 12/23/2020		
Number observations			1395		

Note: *, **, and *** denote statistical significance at 10%, 5%, and 1%. Analytical standard errors in parentheses.

Table 9: Oil Price Forecast Error Variance Decomposition.

Monthly model							
	S^+	S^-	S	D^+	D^-	D	NFP
$t = 1$	18.4	18.9	37.2	12.0	16.3	28.2	34.5
$t = \infty$	19.3	18.7	38.0	13.2	19.7	33.0	29.1
Weekly model							
	S^+	S^-	S	D^+	D^-	D	NFP
$t = 1$	27.7	18.0	45.7	8.7	6.1	14.8	39.5
$t = \infty$	24.3	19.4	43.7	11.7	11.5	23.2	33.1

Note: Entries in the table are the contribution of each shock, or combination of shocks, to the forecast error variance of oil price in the short run, $t = 1$, and in the long run, $t = \infty$. Columns S and D are the sum of the contribution of shocks S^+ and S^- and D^+ and D^-, respectively.

B Data Cleaning

In this Appendix we explain how we clean the data. In particular, we provide more information on the removal of duplicate articles, articles not related to oil market developments, correction of issues related to numbers and punctuation, and adjustments for other idiosyncratic word usage.

First, we seek to exclude duplicate articles. Because the oil market is global in nature, there is some overlap between the *Oil Daily* (OD) and the *International Oil Daily* (IOD). While some articles are included in both publications, at times there are slightly different versions of the articles on the same day or within a few days. To identify these duplicates, we identify for each OD and IOD article the best match from the opposite publication withing a three-day window. Next, we consider these pairs to be duplicates if they have a cosine similarity above a particular threshold, where the threshold is a bit lower for same-day matches and a bit higher for articles that are published within three days of each other. For each pair of articles, we then keep the longer of the two articles for analysis within our main specifications, and retain the duplicate articles for use in our robustness checks.

Second, we exclude articles that are primarily on topics unrelated to the oil industry, including those related to natural gas and gasoline. Though natural gas supply and demand developments are less closely related to oil markets, we recognize that articles on gasoline may be informative for oil supply and demand. However, we note that while an increase in gasoline *supply* may be indicative of an increase in oil *demand*, our text searches would not adjust to this nuance. Since our algorithm would count gasoline supply developments as oil supply developments, inclusion of these articles would introduce undesirable noise into the analysis. Consequently, we aim to exclude gasoline-focused articles as well. With this in mind, we remove these articles as follows:

1. For each article, we count the number of times the *oil words* "crude", "oil", and "petroleum" appear in the body of the text.

2. Next, we count the number of times the *gas words* "gasoline", "gas", and "LNG" appear in the body of the text. Notably, the word "gas" captures both articles about natural gas and articles about gasoline.

3. We remove from the corpus the articles in which the number of *gas words* is greater than or equal to twice the number of *oil words*. Notably, this step also eliminates all articles for which the count of *oil words* is 0 because in these cases the number of *gas words* will always be greater than or equal to the number of *oil words*.

We also correct some problems related to numbers and punctuation that stem from the digitization of the article archive. We begin by excluding tables of data from the raw text, as our aim is to rely only on textual analysis of words and expressions and not on numerical information provided in the publications. Next, we use the *Sentence Tokenizer* function (part of the *NLTK* suite of libraries for Python) to identify sentences within the article. This is an important step, as our directional word counts rely on word proximity within sentences. Among other things, the tokenizer looks for periods, question marks, and exclamation points to note the ends of sentences. We also adjust the punctuation and numerical characters in the text for better application of the sentence tokenizer by taking the following steps:

1. Foreign currency symbols preceding prices often are replaced by a question mark in the raw data. We replace the pattern "?n", where n could be any numerical digit, with the single character "z".

2. The tokenizer often generates mistakes due to improper identification of decimals as sentence-ending periods. To eliminate this problem, we replace all numerical characters with the character "z". We remove decimal numbers from the text by replacing the pattern "z.z" with "zz" each time it appears. This also allows the tokenizer to better

recognize, for example, the end of a sentence that concludes with a year or other number (e.g., "...2017.").

3. Occasionally, a period at the end of a sentence is (erroneously) not followed by a space. To more clearly delineate sentence endings, we insert a space after all periods.

4. More generally, we find many punctuation marks that are erroneously not followed by a space. In addition to periods, we insert a space following all of the following punctuation characters: commas, double quotes, single quotes/apostrophes, question marks, exclamation points, hyphens, parentheses, brackets, forward and backward slashes, and percent symbol. Since inserting a space after a comma in large numbers would alter our word count, before inserting this spaces, we remove commas from numbers by replacing the pattern "z,z" with "zz" each time it appears.

5. Our vocabulary searches are streamlined by the removal of the possessive forms of words. In particular, we remove the string " 's ", which often appears at the end of singular possessive words and the contraction "it's". This ensures that, e.g. the possessive form of the word "producer's" gets counted along with the word "producer". Similarly, we delete apostrophes at the end of plural possessive words. This ensures that, e.g. the plural possessive word "producers' " gets counted along with the word "producers".

6. We replace the common expression "b\d" with its unabbreviated form, "barrels per day".

Lastly, we also correct some issues related to idiosyncratic word usage. In general, the insertion of the character "z" helps retain the word count in the structure of the original sentence, so that the proximity across words is unchanged after the replacements.

1. The word "shale" is on our supply vocabulary list. However, the common phrase "shale boom" would likely result in supply increase counts which would not be very

informative. To reduce the noise arising from this idiosyncratic phrase, we replace the phrase "shale boom" with "shale z". The retention of the word "shale" maintains the supply count while eliminating the likely noisy supply increase count.

2. To reduce false counts of supply, we replace the phrase "shale gas" with "natural gas." This step has no implication for our deletion of gas-related articles, because the deletion is based on the relative count of *oil words* and *gas words*, and this replacement does not replace the word "gas", which is used in the count.

3. To reduce false counts of decrease words, we replace the phrases "tight oil" and "tight crude" with the phrases "shale oil" and "shale crude", respectively. This issue becomes particularly relevant after 2014, when shale oil started its rise into a major source of oil supply.

4. To reduce false counts of supply, we replace the phrase "industrial production" with the string "indprod" and include "indprod" in our demand vocabulary list.

5. To reduce false increase and decrease counts, we replace the phrase "more or less" with the string "z about z".

6. To reduce false increase counts, we replace the phrases "extra heavy" and "extra light" with "z heavy" and "z light".

7. Because the increase word *up* is frequently used within expressions meaning something other than increase, we replace the word "up" with "z" when it appears as one of the phrases listed in Table B1

Table B1: "Up" Phrases Replaced in Text

back(s/ed) up	fill(s/ed) up	look(s/ed) up	size(s/ed) up	tidy up
blown up	follow(s/ed) up	meet(s) up	speak(s) up	tie(s/d) up
break(s) up	gave up	mess(es) up	spoke up	tighten(s/ed) up
broke up	gear(s/ed) up	messed up	stand(s) up	tore up
brush(es/ed) up	give(s) up	met up	start(s/ed) up	true up
buck(s/ed) up	hang(s) up	mix(es) up	stir(s) up	up front
call(s/ed) up	held up	mixed up	stirred up	wake(s) up
came up with	hole(s/d) up	pipe(s) up	stood up	warm(s/ed) up
clean(s/ed) up	hook(s/ed) up	piped up	suck(s/ed) up	wash(es) up
clear(s/ed) up	hung up	put(s) up	sum(s) up	washed up
come(s) up with	leg up	screw(s/ed) up	summed up	whats up
conjure(s/ed) up	light(s/ed) up	set(s) up	tear(s/ed) up	woke up
cover(s/ed) up	line(s) up	shake(s/d) up	threw up	wrap(s) up
draw(s) up	lined up	shook up	throw(s/ed) up	wrapped up
dress(es/ed) up	lighted up	show(s/ed) up	thumbs up	write(s) up
drew up	lit up	shut(s) up	tidied up	wrote up
fed up	lock(s/ed) up	sign(s/ed) up	tidies up	

C Lists of Words and Phrases

In this Appendix we provide the list of regular expressions used to generate the lists of supply, demand, increase, and decrease words and explain how we used these to obtain the list of words we used in the construction of the indexes.

C.1 Supply and Demand Words and Phrases

To generate our list of supply and demand words, we start with the list of regular expressions reported in Table C1. A search using regular expressions will match the many forms of a single root word. That is, searching for the string *suppl* will yield *supply*, *supplied*, *supplies*, etc. The drawback, however, is that our search will also yield words which are are unrelated, like *supplement*. To generate our final vocabulary list, we search the text of the Oil Daily and International Oil Daily articles for all the matches to our list of regular expressions.

This procedure is applied to the text that has been cleaned as described in Appendix B. Our list of just 11 supply and 9 demand regular expressions results in about 1200 supply and 500 demand matches, including both unrelated words and many obvious typographical errors.[5] We focus on the most frequent words, including the roughly 100 supply and 50 demand words which appear in the corpus of 230,000 articles at least 50 times. Of these, we include in our final supply vocabulary list– shown in Table C2 –the words which are informative for supply and demand. We exclude both unrelated words (like *right(s)* for supply, which matches the regular expression *rig*, and *consummate* for demand, which matches the regular expression *consum*) and words that human auditing have revealed to be misleading (like *product(s)* for supply and *refiner(s)* and for demand). These criteria yielded 42 supply and 34 demand words.

Some of the above vocabulary words, such as *glut* and *outage* are particularly informative not just for supply, but for our primary interest in the directional concepts of supply increase and supply decrease. These words are included not just in the raw counts for supply words, but also add to the counts for supply increase and supply decrease. Lastly, since directional supply words are particularly informative, we augment our vocabulary list to include these less frequently occurring related words, as long as they appear at least 5 times in the article corpus. This augments our supply vocabulary list with 11 additional words, to a total of 53 words. Although a similar concept could be applied to demand words to augment our vocabulary list with words such as *underutilized*, human auditing indicated that such expansion would not necessarily result in improved indexes.

[5] As noted in Appendix B, we replace the phrase *industrial production* with *indprod*, and include this expression as a demand word.

Table C1: Regular expressions

Supply		**Demand**	
discover	*reserv*	*buy*	*indprod*
export	*rig*	*consum*	*refin*
glut	*shale*	*demand*	*throughput*
outage	*suppl*	*deplet*	*utiliz*
output	*surplus*	*import*	
produc			

Table C2: Final vocabulary list

Supply			
discoveries	output	producers	suppliers
discovery	overproduce(+)	produces	supplies
export	overproduced(+)	producing	supply
exported	overproducer(+)	production	supplying
exporter	overproducers(+)	reserve	surplus(+)
exporters	overproducing(+)	reserves	surpluses(+)
exporting	overproduction(+)	reservoir	underproducing(-)
exports	oversupplied(+)	reservoirs	underproduction(-)
glut(+)	oversupplies(+)	rig	undersupplied(-)
gluts(+)	oversupply(+)	rigs	undersupply(-)
glutted(+)	oversupplying(+)	shale	undersupplying(-)
nonproducing(-)	produce	shales	
outage(-)	produced	supplied	
outages(-)	producer	supplier	
Demand			
buy	consumes	importation	refinery
buyer	consuming	imported	refinerys
buyers	consumption	importer	refines
buying	demand	importers	refining
buys	deplete	importing	throughput
consume	depleted	imports	throughputs
consumed	depleting	indprod	utilization
consumer	depletion	refine	
consumers	import	refineries	

Note: (+) and (-) denote supply words which add to the counts for supply increase and supply decrease.

C.2 Increase and Decrease Words

The process to generate the list of increase and decrease words used in our analysis is identical to that used to construct the list of supply and demand words. In this case, we search the text using the list of 83 increase and 100 decrease regular expressions shown in Table C3, which results in around 6000 increase and 4000 decrease matches. We again focus on the most frequent words, including the roughly 850 increase and 700 decrease words which appear at least 40 times. After excluding those words that were not related to increase and decrease, we ended up with 248 increase and 309 decrease words, listed in Tables C4 and C5.

Table C3: Regular expressions

Increase				
above	*buoyant*	*increas*	*push*	*strengthen*
accelerat	*burgeon*	*inflat*	*quicken*	*strong*
accret	*climb*	*intensif*	*rais*	*surg*
accru	*cumulat*	*jump*	*rall*	*surmount*
accumulat	*elevat*	*leap*	*reach*	*surpass*
add	*enlarg*	*lift*	*recover*	*swell*
advanc	*escalat*	*more*	*redoubl*	*takeoff*
amass	*expand*	*mount*	*reviv*	*tookoff*
amplif	*extra*	*multipl*	*ris*	*up*
ascend	*firmed*	*mushroom*	*rocket*	*wax*
augment	*gain*	*over*	*rose*	*zoom*
bloom	*garner*	*perk*	*shoot*	
blossom	*grew*	*pop*	*shot*	
bolster	*grow*	*positive*	*skyrocket*	
boom	*height*	*propagat*	*soar*	
boost	*high*	*propel*	*spik*	
buil	*hik*	*propulsion*	*spring*	
bullish	*improve*	*prosper*	*spurt*	
Decrease				
abat	*depress*	*fall*	*retrench*	*stunt*
attenuat	*destruct*	*falter*	*sank*	*subdu*
bearish	*deteriorat*	*fell*	*sap*	*subside*
below	*diminish*	*flag*	*short*	*subsiding*
collaps	*diminution*	*halt*	*shrank*	*sunk*
constrict	*dip*	*less*	*shrink*	*suppress*
contract	*disrupt*	*loosen*	*shrunk*	*suspend*
cool	*dive*	*los*	*shut*	*tank*
crash	*diving*	*low*	*sink*	*taper*
crimp	*dove*	*maintenance*	*slack*	*tepid*
crush	*down*	*negative*	*slash*	*tight*
curtail	*dried*	*offline*	*slid*	*trickl*
cut	*droop*	*plummet*	*slip*	*trim*
damp	*drop*	*plung*	*slow*	*tumbl*
decay	*dry*	*pull*	*sluggish*	*under*
declin	*dwindl*	*quell*	*slump*	*wane*
decreas	*ebb*	*reced*	*small*	*waning*
demise	*evaporat*	*reduc*	*soft*	*weak*
dent	*fade*	*restrict*	*squeeze*	*wither*
deplet	*fading*	*retreat*	*standstill*	*worsen*

Table C4: Increase vocabulary list

above	boomed	flourish	intensifies	popped	regained	strengthened
accelerate	booming	gain	intensify	popping	regaining	strengthening
accelerated	booms	gained	intensifying	positive	resurgence	strengthens
accelerates	boost	gaining	jump	propel	resurgent	strong
accelerating	boosted	gains	jumped	propelled	revival	stronger
acceleration	booster	gallup	jumping	propelling	revive	strongest
accrue	boosting	grew	jumps	prosper	revived	strongly
accrued	boosts	grow	leap	prosperity	reviving	surge
accruing	build	growing	leaped	push	rise	surged
accumulate	building	grown	leaping	pushed	risen	surges
accumulated	builds	grows	leapt	pushes	rises	surging
accumulating	buildup	growth	lift	pushing	rising	surpass
accumulation	built	heighten	lifted	raise	rocket	surpassed
accumulations	bullish	heightened	lifting	raised	rocketed	surpasses
add	bullishness	heightening	lifts	raises	rocketing	surpassing
added	buoyant	high	more	raising	rockets	swell
adding	burgeoning	higher	mount	rallied	rose	swelled
addition	climb	highest	mounted	rallies	shoot	swelling
additional	climbed	highs	mounting	rally	shooting	up
additions	climbing	hike	mounts	rallying	shot	upbeat
adds	climbs	hiked	over	reach	skyrocket	uplift
advance	elevate	hikes	overcame	reached	skyrocketed	upped
advanced	elevated	hiking	overcome	reaches	skyrocketing	upping
advances	enlarge	improve	overcoming	reaching	soar	upsized
advancing	enlarged	improved	overrun	rebuild	soared	upsurge
amassed	escalate	improvement	overruns	rebuilding	soaring	upswing
amassing	escalated	improvements	overshoot	rebuilt	soars	uptick
amplified	escalates	improves	overshooting	recoup	spike	uptrend
amplify	escalating	increase	overshot	recouped	spiked	upturn
augment	escalation	increased	overtake	recouping	spikes	upward
augmented	expand	increases	overtaken	recover	spiking	upwardly
bolster	expanded	increasing	overtaking	recovered	spring	upwards
bolstered	expanding	inflate	overtook	recovering	springing	
bolstering	expands	inflated	perked	recovers	springs	
bolsters	extra	inflating	pickup	recovery	spurt	
boom	firmed	intensified	pop	regain	strengthen	

Table C5: Decrease vocabulary list

abate	demise	downgraded	lessened	restricts	slowdowns	tepid
abated	dent	downgrades	lessening	retreat	slowed	tight
abatement	dented	downgrading	lesser	retreated	slower	tighten
abating	denting	downs	loosen	retreating	slowest	tightened
bearish	deplete	downsize	loosened	retrenchment	slowing	tightening
bearishness	depleting	downsized	loosening	sank	slows	tightens
below	depletion	downsizing	lose	shortage	sluggish	tighter
collapse	depress	downtrend	loses	shortages	sluggishness	tightest
collapsed	depressed	downturn	losing	shorten	slump	tightly
collapses	depressing	downturns	loss	shortened	slumped	tightness
collapsing	depression	downward	losses	shortening	slumping	trickle
contracted	destruction	downwards	lost	shortfall	small	trickling
contraction	destructive	drawdown	low	shortfalls	smaller	trim
crash	deteriorate	drawdowns	lower	shrank	smallest	trimmed
crashed	deteriorated	drop	lowered	shrink	soft	trimming
crashes	deteriorates	dropped	lowering	shrinkage	soften	tumble
crashing	deteriorating	dropping	lowers	shrinking	softened	tumbled
crimp	deterioration	drops	lowest	shrinks	softening	tumbling
crimped	diminish	dwindle	lows	shrunk	softer	under
crimping	diminished	dwindled	maintenance	shut	softness	underperform
crush	diminishes	dwindling	meltdown	shutdown	squeeze	underperformance
crushed	diminishing	ebb	negative	shutdowns	squeezed	underperformed
crushing	dip	ebbed	negatively	shutter	squeezes	underperforming
curtail	dipped	evaporate	nosedive	shuts	standstill	wane
curtailed	dipping	evaporated	plummet	shuttered	stumble	waned
curtailing	dips	fade	plummeted	shutting	stumbled	waning
curtailment	disappear	faded	plummeting	sink	stumbling	weak
curtailments	disappearance	fading	plunge	sinking	stunt	weaken
cut	disappeared	fall	plunged	sinks	stunted	weakened
cutback	disappearing	fallen	plunging	slack	subdued	weakening
cutbacks	disappoint	falling	pullback	slackening	subside	weakens
cuts	disappointed	falls	quell	slash	subsided	weaker
cutting	disappointing	falter	recede	slashed	subsides	weakest
dampen	disrupt	faltered	receded	slashing	sunk	weakness
dampened	disrupted	faltering	receding	slid	sunken	weaknesses
dampening	disrupting	fell	reduce	slide	suppress	worsen
damper	disruption	flagged	reduced	slides	suppressed	worsens
decline	disruptions	flagging	reduces	sliding	suspend	worsened
declined	disruptive	halt	reducing	slip	suspended	worsening
declines	dive	halted	reduction	slippage	suspending	
declining	dived	halting	reductions	slipped	tanked	
decrease	diving	halts	restrict	slipping	tanking	
decreased	down	landslide	restricted	slips	taper	
decreases	downbeat	less	restricting	slow	tapered	
decreasing	downgrade	lessen	restrictive	slowdown	tapering	