Received: 14 August 2020 | Revised: 12 May 2021 | Accepted: 13 May 2021

DOI: 10.1002/fut.22225

RESEARCH ARTICLE

THE JOURNAL OF FUTURES MARKETS WILEY

Oil price analysts' forecasts

Isabel Figuerola-Ferretti[1] | **Alejandro Rodríguez**[1] | **Eduardo Schwartz**[2]

[1]ICADE, Universidad Pontificia Comillas, Madrid, Spain

[2]Beedie School, Simon Fraser University and NBER, Vancouver, British Columbia, Canada

Correspondence
Isabel Figuerola-Ferretti, ICADE, Universidad Pontificia Comillas, c/Alberto Aguilera 23, 28015 - Madrid, Spain.
Email: ifiguerola@comillas.edu

Funding information
Internal Funding from ICADE, Universidad Pontificia Comillas, Grant/Award Number: PP2018_03; MINECO, Grant/Award Number: PID2019-104960GB-I00

Abstract

Crude oil analysts provide forecasts on future spot prices, which are collected by Bloomberg. We exploit this survey to compare analysts' forecasting ability to futures contracts and also among analysts themselves. We address the problems arising from unstructured forecast data and use the Mean-Squared Prediction Error (MSPE) relative to the no-change forecast and the Diebold and Mariano test. The applied approach represents a substantial improvement compared with the standard MSPE methodology as it corrects for volatility and maturity effects on forecasting performance measures. Finally, we establish that futures prices supersede analyst forecasts and elaborate a performance-based ranking of analyst firms.

KEYWORDS

analyst, crude oil futures, forecast

JEL CLASSIFICATION

C53, G12, G13, G17, Q02, Q35

1 | INTRODUCTION

This paper evaluates the forecasting performance of oil futures prices against the price forecasts provided by investment analysts. While accurate forecasts of oil prices are essential to producer firms and consumers in the optimal allocation process, futures prices have commonly been used in the literature to measure the expectation of futures spot prices (see Alquist & Kilian, 2010; Alquist et al., 2013). Futures markets provide centralized trading where information about supply and demand conditions for a commodity is efficiently incorporated. Traded prices generate an informational consensus by aggregating information among agents. The role of futures prices in forecasting realized spot prices is also relevant to policy-makers. For instance, Svensson (2005) underlines the importance of forecasting oil prices as a first step towards finding the optimal instrument rate plan at the European Central Bank. In his influential speech, Bernanke (2004) discusses the relevance of oil futures prices in reflecting traders' price expectations for future delivery. His remarks also underline that medium- and long-term forecasts of oil prices become significant economic activity predictors. Alquist et al. (2013) highlight the relevance of providing accurate predictions of the crude oil spot price and perform a systematic evaluation of the predictive accuracy of forecasting methodologies based on oil futures.

The forward pricing role of the futures market, also known as the prediction hypothesis, can be framed in the seminal work of Working (1948), who analyzes the role of futures prices in pricing cash market transactions within a model for storable commodities. In a later paper, Black (1976) argues that the futures price is the expected spot price under certain conditions (see also Peck, 1985). A related line of literature addresses the informational role of commodity futures prices in reflecting fundamental conditions. Garbade and Silber (1983) develop a model of simultaneous price dynamics in which they establish that the price discovery role of commodity futures markets arises due to the existence of a higher number of participants. Such findings are confirmed and extended to the metals case in the price discovery framework with endogenous convenience yields introduced by Figuerola-Ferretti and Gonzalo (2010).[1] The role of the marginal convenience yield in explaining the spot and future basis is analyzed in the crude oil market literature under the theoretical framework introduced by Alquist and Kilian (2010). They derive a multiperiod two-country general equilibrium model of spot and futures prices. They show that changes in the predictive accuracy of futures prices may be driven by increased variability in the spot futures basis, which mainly arises due to uncertainty about future supply shortfalls. Coppola (2008) analyzes the forecasting ability of crude oil futures markets within a vector error correction model (VECM) framework highlighting the role of futures prices in forecasting realized spot prices.

The forecasting role of futures prices has also been addressed in commodity pricing models that fit the futures term structure (Gibson & Schwartz, 1990; Schwartz, 1997; Schwartz & Smith, 2000). These models assume that futures prices are the expected spot prices under the risk-neutral probability distribution and provide the true physical distribution of spot prices. In a recent publication, Cortazar et al. (2019) propose a term structure model that supersedes the previous literature in the accuracy of the forecast delivered for expected spot prices. The key element of their contribution lies in the combination of information in futures prices and survey price expectations. The latter is used to provide explicit information on the current risk premium.

This paper follows Cortazar et al. (2019) in that it exploits the Bloomberg crude oil survey data. It is also related to the work of Chang et al. (2009), who use crude oil supply forecast data from Bloomberg to analyze the impact of analyst forecasts on crude oil prices. The price forecasting literature in energy markets has applied different survey data sources (Alquist et al., 2013; Anderson et al., 2011; Baumeister et al., 2014; Sanders et al., 2009; Wiser et al., 2016).[2] While the consensus is that survey data contribute to forecasting future spot prices, the evidence for the crude oil market is not decisive. Alquist and Kilian (2010) use a source of monthly forecasts from the UK-based firm *Consensus Economics Inc.* and conclude that survey forecasts are inferior to the no-change forecast as well as the futures-based forecasts. Alquist et al. (2013) assess the forecasting ability of different sources of crude oil data surveys, including the Energy Information Administration (EIA), The Economist Intelligence Unit, and Consensus Economics. They analyze the accuracy of these data in predicting future spot prices by comparing forecasting results against the predictive performance of the no-change or current spot benchmark. Their main conclusion is that survey data represent at best only moderate improvements in forecast accuracy when compared with the no-change forecast.

Analyst data from the Bloomberg database are unstructured in the sense that it does not deliver a panel of homogeneous observations. Analysts do not provide new forecasts for every maturity on a daily basis. This fact complicates the comparison of predictive performance with respect to the futures and among analysts. We detect two main problems that arise when using the ordinary Mean-Squared Prediction Error (MSPE) metric. First, we find that higher volatility during the forecast period delivers inferior performance. Second, we see that rankings do not account for the number of forecasts provided by each analyst.

To overcome the first issue, we follow the literature (see Alquist & Kilian, 2010; Alquist et al., 2013) and compare the performance of crude oil survey data with futures prices using the MSPE ratios relative to the no-change forecast. We demonstrate that using the current spot price as the basis for comparison eliminates the negative effect that volatility exerts on forecasting performance. The theoretical relevance of the futures to no-change forecast can be framed within the model introduced in Alquist and Kilian (2010), where changes in predictive accuracy are likely to be driven by increased volatility in the futures and spot basis. We address the second problem using the MSPE to the no-change ratio and testing for equal predictive accuracy using the Diebold and Mariano (1995; DM hereafter), which allows testing the null hypothesis of equal predictive performance.

We quantify the extent to which the futures market is unique in the guidance it provides for producers, distributors, and consumers of oil and other commodities (Black, 1976). While futures prices aggregate information among agents, analysts can be regarded as informed investors. Forecasting from informed sources implies that the potential biases are minimized. Brennan et al. (1993) study the contribution of analyst forecasts to the adjustment of equity prices. They use the number of analysts as a proxy for the number of informed investors following a firm. Using a vector autoregressive (VAR) model, they show that the higher the number of informed investors following a firm, the faster the price converges to equilibrium.

[1]In an earlier paper, Figuerola-Ferretti and Gilbert (2005) analyze the increasing importance of the price discovery role of aluminum futures trading under different pricing systems.
[2]Anderson et al. (2011) show that the survey on gas price consumer prices tracks futures prices during the GFCs.

A parallel line of literature argues that investment analysts exhibit behavioral patterns in their price forecasts (De Bondt & Thaler, 1990) and that their performance is affected by conflicts of interest (Fang & Yasuda, 2009). The question of whether investment analysts provide superior forecasting results when compared with futures prices is therefore not straightforward.

Bloomberg is a benchmark data source for market participants, which provides forecasts that are more consistent with the market consensus view than other competing sources (Chen et al., 2013). The Bloomberg crude oil survey reports price forecasts by investment analysts who cover and therefore know the oil industry and the market but not necessarily trade oil. By choosing the Bloomberg database, we aim to guarantee the reliability, completeness, and accuracy of the data set. As Ljungqvist et al. (2009) argued, the correct choice of a database is crucial to avoid tampering attempts and minimize selection bias and measurement errors. In the context where the academic literature is used for trading purposes, a reliable data set is expected to lead to an optimal allocation of resources.

This paper is organized as follows: in Section 2, we describe the data collection process and the data set. Section 3 presents the methodological framework used for the analysis. Section 4 compares the predictive performance of analysts and futures. Section 5 presents a ranking methodology that compares the forecasting performance among analysts while accounting for differences in volatility, statistical significance, and temporal horizon. Section 6 performs an *Ordinary Least Squares* (OLSs) pooled regression designed to explain the extent to which predictive accuracy is related to several independent variables selected from analyst firms' features. Section 7 discusses the main results in light of the literature, and Section 8 offers concluding remarks.

2 | DATA AND CHALLENGES

The purpose of this paper is to assess the performance of analyst forecasts in predicting West Texas Intermediate (WTI) oil spot prices when compared with futures prices and also among themselves. We use the Bloomberg database to answer this question. Bloomberg's oil forecasts survey is reported on a daily basis, providing descriptive statistics on the consensus and also the breakdown of predictions released by different crude oil analysts. These forecasts of spot prices are performed on a quarterly and yearly horizon basis. Consequently, they are heterogeneous among analysts and have maturities up to 62 months.

We collect data on analyst forecasts for every private financial institution along 167 monthly extractions that span from April 30, 2006 to December 31, 2019. No earlier prediction data are available from this source. The whole sample delivers 5748 forecasts released by 60 firms. Additionally, we have daily WTI spot price series and futures New York Mercantile Exchange (NYMEX) WTI prices ranging from 1 to 67 months to maturity for the same sample period (The Bloomberg code for WTI futures ranges from CL_1 to CL_{67}).

Table 1 displays the data provided by the Bloomberg terminal on January 31, 2014, which includes (1) a list of the different analyst firms in the first column labeled as *Firm*, (2) the exact date on which the price forecast is released under the second column designated as *As Of*, (3) the predictions for the average crude oil price in 2014 under the third column labeled *2014*, and (4) subsequent columns with forecasts for 2015–2018.

Some crucial characteristics of the data in Table 1 should be highlighted. First, there are not predictions available for every horizon by each analyst considered. For instance, there is no forecast by *Citigroup Inc.* for 2017 and 2018. Second, predictions remain constant for a number of periods. For example, a close look at the second column shows that the last forecast of *Societe Generale SA* was performed on November 26, 2013, suggesting that forecasts remained constant for over 2 months. Consequently, the data are not homogeneously structured, which implies that the comparison between forecasts and futures prices is not straightforward. Details on the method used to overcome forecast heterogeneity are provided in Section 3.

Forecasts need to satisfy certain regularity conditions to be reported by the Bloomberg terminal. The contributions to the forecast survey are voluntary, which suggests that the analyst firms that report to Bloomberg are heterogeneous, and thus survey constituents comprise firms with diverse expertise levels. This also implies that the number of forecasters for a specific commodity may vary over time. Figure 1 shows that the number of analysts reporting to Bloomberg has decreased significantly since 2018. Although this might suggest that the interest in the oil market has dropped over the past 2 years, we do not have sufficient evidence to empirically support this claim.

Futures trading at NYMEX ends three business days before the 25th calendar day preceding the contract month. If the 25th is not a business day, then trading terminates 4 days earlier. Bloomberg automatically calculates the rolling of futures contracts. When a contract does not trade on a specific day, the gap is filled with the price and open interest of the last available day, and the corresponding volume is set to zero.

TABLE 1 Excerpt of monthly data extraction from Bloomberg on January 31, 2014

Firm	As Of	2014	2015	2016	2017	2018
ABN AMRO Bank NV	15NOV2013	100	95	90	–	–
BNP Paribas SA	17JAN2014	93	94	–	–	–
Banco Santander SA	12DEC2013	97	92	89	88	85
Capital Economics Ltd.	16AUG2013	90	–	–	–	–
Citigroup Inc.	31JAN2014	92.75	86.25	83	–	–
Commerzbank AG	23JAN2014	100	106	–	–	–
Credit Suisse Group AG	06JAN2014	91.75	87.5	85	85	80
DZ Bank AG	29JAN2014	100.25	–	–	–	–
Danske Bank A/S	17JAN2014	93	–	–	–	–
Fitch Solutions	15JAN2014	101.5	101	–	–	–
Incrementum AG	07JAN2014	92.8	83.75	104.4	–	–
Intesa Sanpaolo SpA	18DEC2013	99.2	100.7	102.1	–	–
KLR Group LLC	23AUG2010	85	85	–	–	–
Lloyds Bank PLC	12SEP2013	98.5	94	97	–	–
Macquarie Group Ltd.	11JAN2013	104	108	–	–	–
Natixis SA	10JAN2014	100	101.5	103	–	–
Nomura	03JAN2014	90	85	85	–	–
Prestige Economics LLC	31DEC2013	98.25	106	–	–	–
Societe Generale SA	26NOV2013	99	101	103	105.5	108
UBS Group AG	15JAN2014	98.5	89	86	86	86
UniCredit SpA	20JAN2014	101	101	–	–	–
Westpac Banking Corp	12DEC2013	87.61	94.71	108.81	96.43	83.33

Note: For example, the fifth row shows that *Citigroup* on January 31, 2014 forecasted that the average spot price in 2015 would be 86.25. This value will be compared against the average realized spot in 2015 and with an average of the quoted NYMEX futures prices for maturities ranging from 12 (CL_{12}) to 23 months to maturity (CL_{23}).

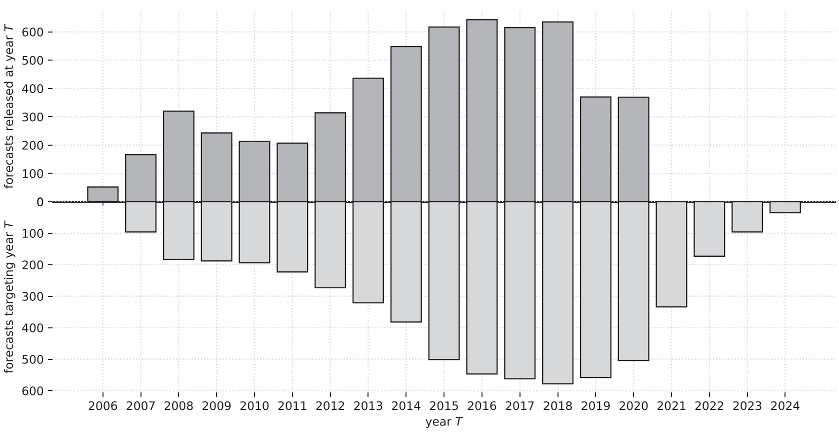

FIGURE 1 For a given year, the top panel shows the number of forecasts released at the time, while the bottom panel shows the number of past forecasts that target it. As the data were retrieved in October 2020 and the number of participants in the survey changed in 2019, charts exhibit an incidental decrease beyond that point

These corrections are more frequent in contracts with longer maturities due to their lower liquidity. Figure 2 shows a sharp decrease in volume and open interest as maturity increases. Moreover, Figure 3 connects the frequency in which futures contracts are traded to the number of forecasts released by analysts. More specifically, the chart measures the percentage of days with futures trading activity for every maturity over the total number of days that the market is open. This can be simultaneously compared with the number of forecasts released with the corresponding

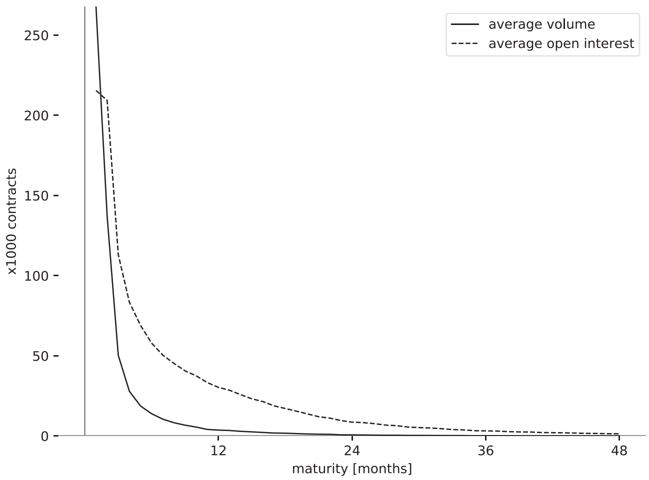

FIGURE 2 Average volume and open interest of the futures contracts in the 2000–2019 period (both inclusive). Maturity up to 48 months

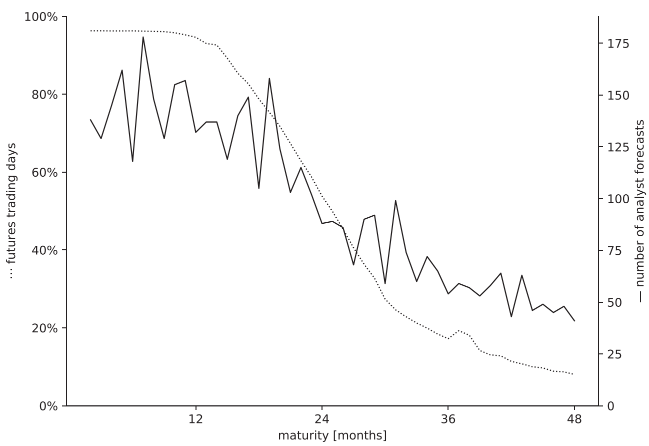

FIGURE 3 The solid line (—) shows the number of forecasts released per each maturity. The dotted line (···) shows the percentage of days that the contracts of that maturity are traded. Short-term forecasts and contracts with lower maturities are more common than long-term ones

temporal horizon. Interestingly, the two lines intersect just beyond the 24 months maturity, where futures are only trading 40% of the time and analysts deliver 75 forecasts.

We select our sample and, following Cortazar et al. (2019) as well as Bianchi and Piana (2017), exclude predictions for the current year. This leaves 3001 usable observations out of the original 5748 and 58 firms out of 60, respectively. Therefore, our sample provides price data that are subsequently used to calculate the mean spot price on a yearly basis up to 2019 (see Equation 1). Note that a significant challenge in this analysis lies in the fact that there is no clear consensus regarding which price the analysts attempt to forecast. We follow the spirit of Bianchi and Piana (2017) and assume that analysts provide point forecasts for the average spot price on the target year.

One question that arises when analyzing Bloomberg analysts' data is to what extent the sample represents a wider population. The criteria behind Bloomberg selection is likely to be supported by business considerations and commercial agreements. In this sense, the sample of analysts' forecasts may not be regarded as statistically random. However, in this paper, we share the view expressed in the work of Chen et al. (2013), where the authors suggest that Bloomberg data are generally consistent with the market consensus. On this basis, we can state that although our sample is not strictly random, it is representative enough to provide statistical inference.

3 | EMPIRICAL FRAMEWORK

In what follows, we present the framework used in our research. Let A_i be the analyst's price forecast by firm i, which is defined as

$$A_i(t, T) = \bar{S}(T) + \varepsilon = \frac{1}{D}\sum_{d=1}^{D} S(T, d) + \varepsilon, \tag{1}$$

$A_i(t, T)$ forecast released on date t by analyst i for year T,
$\bar{S}(T)$ average of spot prices in year T,
D number of trading days in year T,
$S(T, d)$ spot on trading day d of year T,
ϵ error term.

To compare the analyst's forecast of the average spot price with its futures contract counterpart, we construct the following metric based on average futures prices with maturity within the target year (Equation 2).

$$\bar{F}(t, T) = \frac{1}{12} \sum_{m=-6}^{5} F_{M+m}(t), \tag{2}$$

$\bar{F}(t, T)$ average of the 12 futures on release date t that target year T,
M number of months between release date t and July 1 of target year T,
$F_{M+m}(t)$ future price on date t with maturity $M + m$.

The maturities of the corresponding 12 futures contracts are determined by the temporal horizon of the forecast (Figure 4). We follow the approach of Cortazar et al. (2019) and let M be the maturity measured in months to the middle of the target year (July 1) so that the whole year spans from $M - 6$ (January) to $M + 5$ (December). The futures price metric is therefore calculated by averaging futures prices with different maturities. Figure 2 shows how average futures volume and open interest decrease with the time to maturity. This suggests that long-term futures are not as informationally efficient as short-term futures and that there may not be prices available daily for each long-term maturity. By constructing a futures price metric that averages futures prices with different maturities, the problem of low liquidity for long horizons is alleviated. Thus, short- and long-term maturities are used for every observation of the futures series applied to compare analyst forecasts. Cortazar et al. (2019) use both futures prices and analysts' forecasts to calibrate a commodity pricing model and demonstrate that futures data are much more frequent than analysts' forecasts.

The no-change forecast is a naive model that uses the current spot price in a series as a predictor of future values:

$$A^*(t, T) = S(t), \tag{3}$$

$A^*(t, T)$ no-change forecast on release date t for target year T,
$S(t)$ spot on date t.

The no-change forecast has commonly been employed in the literature as a benchmark model (see Alquist et al., 2013). In this paper, we show that its use becomes crucial to overcome problems that arise with heterogeneous data. Equations (1)–(3) are thus central to our analysis.

3.1 | Mean-squared prediction error

There are many available metrics to measure accuracy in univariate time series forecasts (Hyndman & Koehler, 2006). In this paper, we use the MSPE as a standalone metric and in the ratio to the no-change forecasting model. The MSPE metric is computationally straightforward and offers a clear interpretation:

$$\varepsilon_{ij} = A_i(t_j, T_j) - \bar{S}(T_j),$$

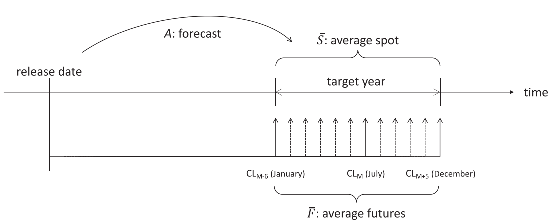

FIGURE 4 An analyst issues at the release date a forecast A which aims to predict the average spot price $\bar{S}$ along the target year. At the same release date, there are 12 futures that mature throughout the same target year and their average is $\bar{F}$

$$MSPE(A_i) = \frac{1}{n}\sum_{j=1}^{n}\varepsilon_{ij}^2, \tag{4}$$

ε_{ij} error term of forecast j by analyst i,
t_j release date of forecast j,
T_j target year of forecast j.

It is important to note that the weight of a forecast error in the MSPE is proportional to the square of the deviation from the true value, and this makes the metric sensitive to outliers. While this feature is often regarded as a drawback of the method, we follow Hyndman and Koehler (2006) and argue that MSPE is an adequate measure in this context due to its computational simplicity and precise interpretation. Besides, such sensitivity to outliers may become an advantage when comparing analysts' performance because it would penalize those whose forecasts experience wild swings.

We analyze the MSPE specified in Equation (4) to compare the predictive performance of the analysts' forecasts against futures contracts at an aggregated level. Next, we apply this MSPE approach to compare the performance of different financial institutions. Our results show that the straight application of MSPE can be misleading because it fails to consider the heterogeneity of the forecasts in terms of underlying uncertainty, statistical significance, and (to a lesser extent) investment horizon. Reported results demonstrate that (a) the quality of the forecast is expected to decrease under abnormal market conditions, (b) the statistical significance of the measure depends on the number of forecasts under consideration, and (c) forecast accuracy is expected to decrease with the forecast horizon.

We shall see that the reported bias of the MSPE metric complicates the task of ranking different firms by their performance as it underrates the performance of those analysts who release a forecast for or during highly turbulent periods. The metric also punishes those firms that provide a large number of forecasts because it does not account for the fact that participants with very few predictions might deliver high accuracy by pure chance.

3.2 | MSPE ratio to no-change forecast

We follow Alquist et al. (2013) and compare the forecasting performance of the different A_i series by presenting MSPE results as ratios relative to the corresponding no-change forecast benchmarks, namely, A_i^*:

$$\delta_{ij} = A^*(t_j, T_j) - \bar{S}(T_j),$$
$$MSPE(A_i^*) = \frac{1}{n}\sum_{j=1}^{n}\delta_{ij}^2, \tag{5}$$

δ_{ij} error term of no-change forecast i,
t_j release date of forecast j,
T_j target year of forecast j.

And the ratio:

$$\phi(A_i) = \frac{\mathrm{MSPE}(A_i)}{\mathrm{MSPE}(A_i^*)}. \tag{6}$$

We shall argue that this allows comparison between analysts' forecasts made under heterogeneous market conditions and different forecast horizons. The ratio $\phi(A_i)$ embeds a series of predictions that take place on equivalent dates for the numerator and denominator. Consequently, it captures the volatility and maturity factors that affect forecast accuracy. This enables improved ranking of the performance among analyst firms, irrespective of the idiosyncrasy of their forecasts.

3.3 | Diebold and Mariano test

In this section, we describe the method used to analyze the statistical significance of the error measure. The Diebold and Mariano (1995) test is the most common choice in the literature for this purpose. It provides a statistically robust methodology to test the null hypothesis of no difference in the accuracy of two competing forecasts. The test is well suited for our analysis as it allows for potentially non-Gaussian, nonzero mean, serially correlated, and contemporaneously correlated forecast errors. The literature has applied the DM test with minor adjustments (Harvey et al., 1997) and bootstrapping techniques (Baumeister & Kilian, 2012; Kilian, 1999; Kilian & Taylor, 2003; Mark, 1995) to tailor it to specific requirements and mitigate some of the limitations that may arise during its application. In this paper, we derive rankings that use DM test results and show that such rankings account for the effects of volatility on forecast accuracy and take statistical significance into consideration.

In what follows, we consider the analysis of two competing forecasts at regular intervals under a predictive horizon of h-steps ahead. We define ε_{it} and ε_{jt} as the resulting forecast errors. While this test supports any arbitrary loss function, we choose a quadratic function of the forecast errors for simplicity. Consequently, if the loss differential d_t is covariance stationary and exhibits short memory, Diebold and Mariano (1995) propose the following statistic S_1:

$$d_t = g(\varepsilon_{it}) - g(\varepsilon_{jt}) = \varepsilon_{it}^2 - \varepsilon_{jt}^2,$$

$$S_1 = \frac{\bar{d}}{\sqrt{Var(\bar{d})}} = \frac{\bar{d}}{\sqrt{\frac{2\pi\hat{f}_d(0)}{n}}} \approx N(0, 1), \tag{7}$$

d_t loss differential at time t,
$g(\varepsilon_t)$ loss function for error ε_t (quadratic in this case),
ε_{it} and ε_{jt} forecast errors of the competing forecasts i and j.

Note that while the last expression allows for forecasting errors and hence loss differentials to be serially correlated (Diebold, 2015), this condition is not imposed. A question that arises when dealing with unstructured data, as is the case in this analysis, is whether the assumption of serial correlation in the data is realistic. Note that Bloomberg's analyst data are heterogeneous in the sense that forecasts take place at arbitrary time intervals and have different temporal horizons.

To illustrate this concern, Figures 5 and 6 describe two hypothetical scenarios. While the former presents the ideal time series of two analysts' forecasts, the latter exhibits the more complex scenario found in this study. Specifically, (1) both competing forecasts may not coincide in time nor temporal horizon (e.g., see the light- and dark-gray cells in the figure, A_i and B_i, respectively), (2) several observations for a single analyst may occur at the same time step (e.g., A_1 and A_2), (3) forecast release dates are irregular and cause gaps in the sample (e.g., $t = 3$), (4) the sample contains observations with different temporal horizons (e.g., A_1, A_2, and A_4), and (5) the sample size may be small due to the limited number of observations available for some of the analysts considered. Specifically, the last points are a cause of concern because Harvey et al. (1997) showed that the DM statistic could deliver oversizing under small samples and large forecast horizons.

The extent to which there is serial correlation in our framework is an empirical question. We therefore select optimal truncation lags for the different analyst series using Andrews' (1991) data-based rule. This allows optimal lag solutions to grow with the sample size and the estimated autoregressive coefficients. Note that when the optimal lag is zero, $2\pi\hat{f}_d(0)$ collapses into the variance of the average loss differential (Ashley, 1998).

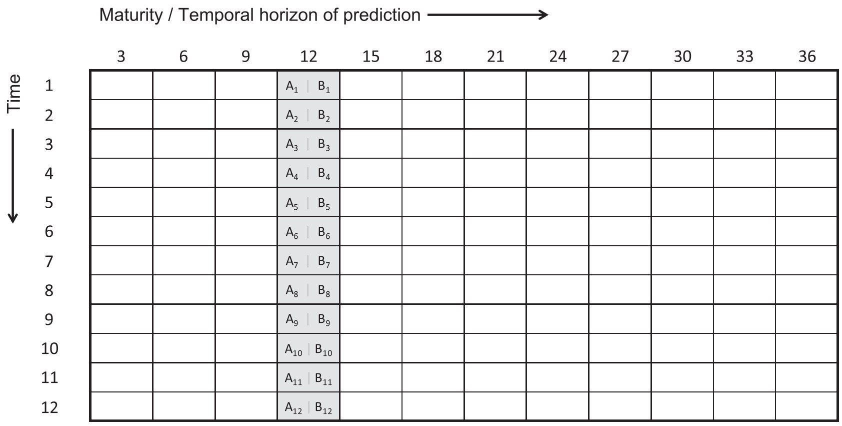

FIGURE 5 Structured data. Competing analysts A and B release forecasts at regular intervals and with the same temporal horizon. Performance comparison is straightforward $A_i \leftrightarrow B_i$

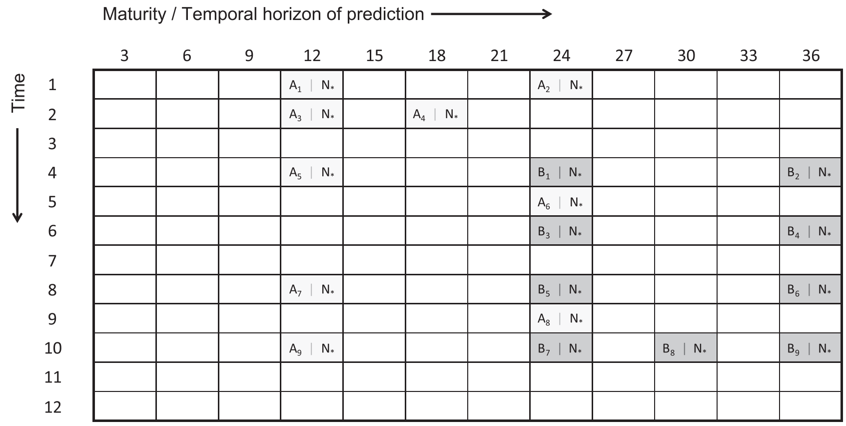

FIGURE 6 Unstructured data. Competing analysts A and B issue forecasts on different dates and with heterogeneous temporal horizons. Comparison in this case can only be performed by combining forecasts with different maturities. Rankings are based on MSPE ratios to their matching no-change counterparts $A_i \leftrightarrow N_*$ and $B_i \leftrightarrow N_*$. MSPE, mean-squared prediction error

The above framework is used to test for the null hypothesis that there is no statistically significant difference in the prediction accuracy of two competing forecasts. We will specifically compare the performance of futures contracts, analyst forecasts, and the no-change forecast benchmark model. When the null is rejected, the underlying method provides the direction of the inequality (i.e., it determines which of the two compared predictions is more accurate).

4 | GLOBAL RESULTS: FUTURES VERSUS ANALYSTS' FORECASTS

This section compares the accuracy of analysts' forecasts and the corresponding futures contracts against the realized spot price for the target year. More specifically, the idea is the following: when an analyst posts a forecast on a release date aiming to predict the average spot price on a given year, the recorded forecast price is compared with (a) the realized spot price and (b) the average price (at the release date) of the 12 futures that mature between the start and the end of the target year. Figure 4 illustrates the set of variables used in the analysis.

In what follows, we compare the analysts' forecasts with the corresponding average futures prices and average target spot prices. We do this for the entire set of 3001 forecasts released by 58 firms over the 2006–2019 sample period:

$$[A_i(t_j, T_j), A_i^*(t_j, T_j)] \leftrightarrow [\bar{F}(t_j, T_j), A_i^*(t_j, T_j)] \leftrightarrow [\bar{S}(T_j)], \tag{8}$$

$A_i(t_j, T_j)$ forecast j by analyst i of the spot price released on t_j targeting year T_j,
$A_i^*(t_j, T_j)$ no-change forecast j matching analyst i of the spot price released on t_j targeting year T_j,
$\bar{F}(t_j, T_j)$ average of the futures valued at t_j that target the 12 months of year T_j,
$\bar{S}(T_j)$ average spot price throughout year T_j.

As a preliminary analysis, first we pool all forecasts for every maturity and compare the performance at the aggregated level between futures prices and analysts' forecasts. Figure 7 shows the time series evolution of the MSPE for futures contracts as well as analysts' forecasts. Both are presented as ratios to the benchmark no-change forecast. The lower panel depicts the difference between both MSPE ratios. We can see that futures prices consistently out-perform both the no-change forecast ($ratio < 1$) and the analysts' forecasts. The documented out-performance increases over our sample period. Moreover, we see that the longer the maturity (temporal horizon) of both futures and analysts' forecasts, the larger the MSPE.

Figure 8 shows that both MSPE ratios to no-change are higher in volatile periods, such as the 2008–2009 GFC or 2014–2016 crude oil price collapse (see Figure 9). In a recent paper, Figuerola-Ferretti et al. (2020) show that crude oil prices exhibit an explosive behavior during such periods which is not consistent with the martingale assumption. It is also interesting to see that the performance diverged in 2011. From that point, the gap widens significantly, which could be explained by a structural change in the series. Possible reasons for this divergence are the financial crisis and the subsequent quantitative easing policies (QE) which contributed to a commodity super-cycle, in which energy corporations increased their CAPEX substantially to take advantage of the low rates (Cervera & Figuerola-Ferretti, 2019). Another possible explanation lies in the Shell revolution (Kilian, 2017), which triggered changes in the supply side and plausibly the increasing number of analysts covering oil (see Table 2).

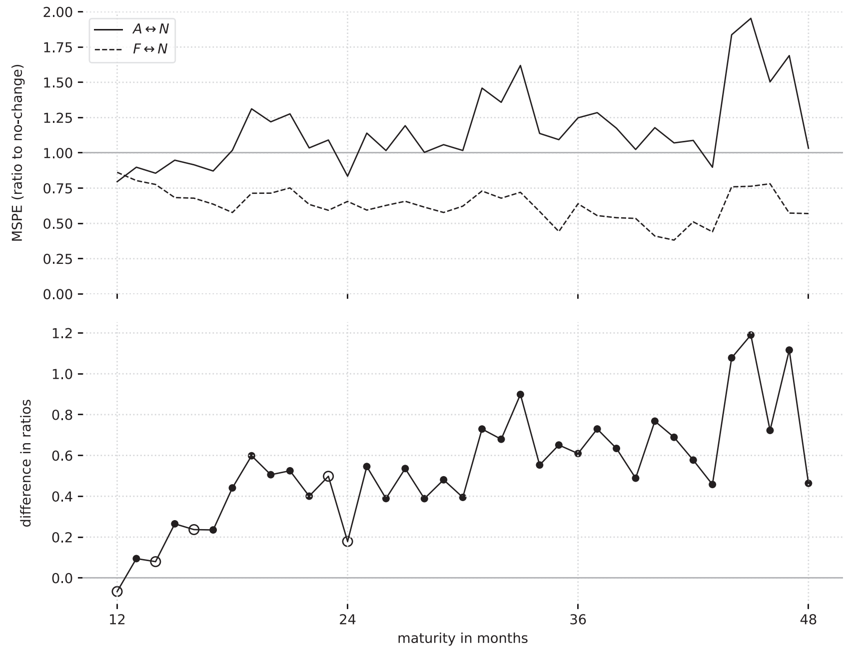

FIGURE 7 The top panel shows the evolution of MSPE ratios of analysts' forecasts (solid line) and matching futures (dashed) as maturity increases. MSPEs are presented as ratios relative to the no-change model. Consequently, values below one are more accurate than the benchmark. The difference is plotted below and increases with maturity. The lower panel observations marked with a solid dot are statistically significant according to the DM test ($\alpha = 0.05$) of equal accuracy between analysts and futures for the given maturity, while those in a hollow circle are not. Despite the notable difference present at longer maturities, certain measures are not significant because of the sharp decline in the number of available long-term forecasts. DM test, Diebold and Mariano test; MSPE, mean-squared prediction error

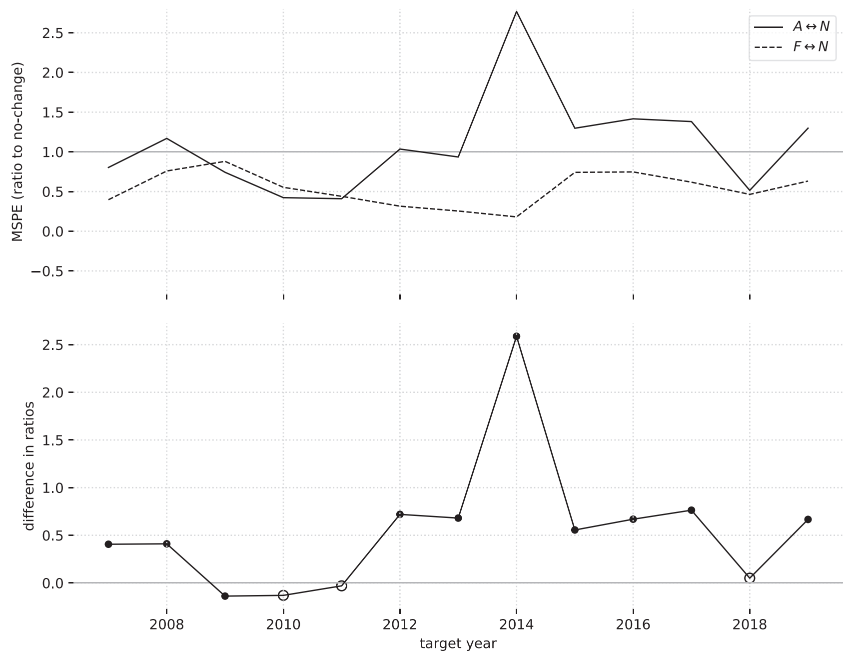

FIGURE 8 The top panel shows the MSPE ratios of analysts' forecasts (solid line) and matching futures (dashed) that target a given year. MSPEs are presented as ratios relative to the no-change model. Consequently, values below one are more accurate than the benchmark. The difference is plotted below and is higher for the newer data points. The observations in the lower panel marked with a solid dot are statistically significant according to the DM test ($\alpha = 0.05$) of equal accuracy between analysts and futures for the given target year, while those in a hollow circle are not. DM test, Diebold and Mariano test; MSPE, mean-squared prediction error

The overall conclusion is that futures contracts outperform analysts' forecasts in accurately estimating the future spot prices of WTI oil at the aggregated level.

5 | ANALYSTS' RANKINGS

5.1 | Forecast heterogeneity

We have shown in Figure 3 that futures prices with maturities of up to 1 year are available on a daily basis. However, contracts for longer maturities exhibit lower trading volumes, and quoted prices may remain unchanged for several days. Despite that, futures price data constitute a structured time series. Conversely, analyst forecast data exhibit considerable heterogeneity because (a) firms do not release estimates at the same points in time and (b) temporal

FIGURE 9 Historical series of WTI oil spot price (top panel) and annualized rolling volatility with a 3-month window (bottom). WTI, West Texas Intermediate

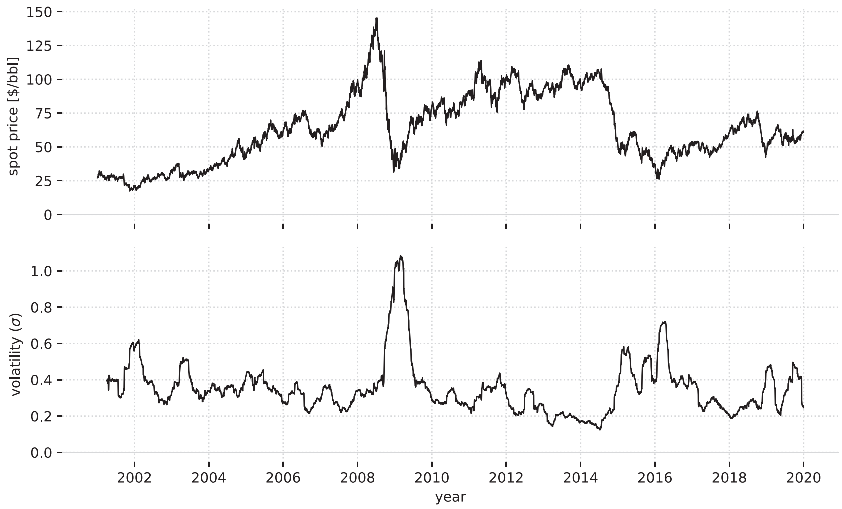

TABLE 2 Number of firms and forecasts per year

	Active firms	Released at	Forecasting
2006	10	52	–
2007	15	166	96
2008	17	320	183
2009	19	243	188
2010	20	213	194
2011	18	207	223
2012	24	314	273
2013	29	436	321
2014	32	548	382
2015	35	617	501
2016	43	643	547
2017	40	615	562
2018	39	635	578
2019	26	370	558
2020	21	369	504
2021	–	–	334
2022	–	–	173
2023	–	–	96
2024	–	–	35

Note: For a given year, the column *active firms* is the number of analyst houses that were active during it. The columns *released at* and *forecasting* show the number of forecasts that were released at that year or aim to predict it, respectively (note: data extracted in October 2020, so 2019 is the last year with full data available).

horizons vary across forecasts (Figure 3). This is important because the higher the forecast horizon, the higher the underlying uncertainty and the more complicated it is to predict future price movements accurately. Besides, those forecasts released for highly volatile target years are likely to exhibit lower performance because of the uncertainty that arises under abnormal market conditions.

Figure 9 plots the time series of daily prices of WTI crude oil (upper panel) and its volatility (lower panel) calculated as the rolling standard deviation over a 3-month window. We can see that the volatility of crude oil WTI prices increases substantially during the 2008–2009 global financial crises (GFCs) and the 2014–2016 crude oil crises. We expect the forecasts released for those target periods to be of lower quality than those posted for calmed years. Figure 3 exhibits the total aggregated number of forecasts per horizon maturity level and shows that the number of forecasts for 1–2-year maturities is higher than for longer maturities.

Table 2 shows that the number of active private financial institutions or analysts that deliver forecasts increases over time. We observe an upward trend in the number of forecasts posted for a given target year which was also documented in Figure 1.

Consequently, while it is feasible to match and compare forecasts with futures prices and current spot prices at an individual level, aggregated comparisons by analyst and time period are not straightforward due to their idiosyncrasy in terms of year of forecast and time horizon.

By comparing any forecast to its corresponding no-change counterpart (see Equation 3), these effects are mitigated because both are affected by matching forecast horizon and market conditions.

5.2 | Forecast error versus volatility and maturity

In this section, we explicitly address the relationship between the forecasting error measured as $|A_i(t, T) - \bar{S}(T)|$ and the volatility of the crude oil price during the forecasting period as well as the temporal horizon of the forecast. This relationship is depicted in Figures 10 and 11, which suggest that there are positive volatility and time to maturity effects on forecasting errors. To quantify these effects, the following linear pooled regression is estimated with all the available observations:

$$|A_i(t, T) - \bar{S}(T)| = \alpha + \beta_1 \cdot volatility + \beta_2 \cdot maturity. \tag{9}$$

Results are reported in Table 3. We can see that both variables are statistically significant and deliver a regression of $R^2 = 0.417$. Given that the forecast error is dependent on maturity and volatility, it becomes essential to control for these effects to guarantee the reliability of the rankings generated.

5.3 | Analyst MSPE versus volatility and maturity

We have seen in the previous section how individual error measures are exposed to volatility and maturity effects. In what follows, we analyze the impact of these variables on aggregated error measures, such as the MSPE. In this case, we use the linear equation:

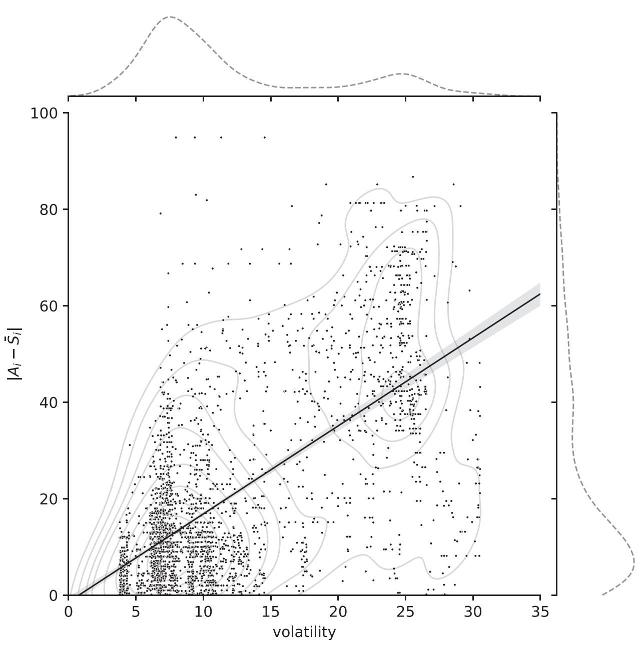

FIGURE 10 Scatter plot of the forecast errors in the form $|A_i - \bar{S}_i|$ versus *volatility*. The marginal plots show the distribution of each variable. As visual aid, the plot includes a regression line with its confidence interval and KDE contour lines. KDE, kernel density estimate

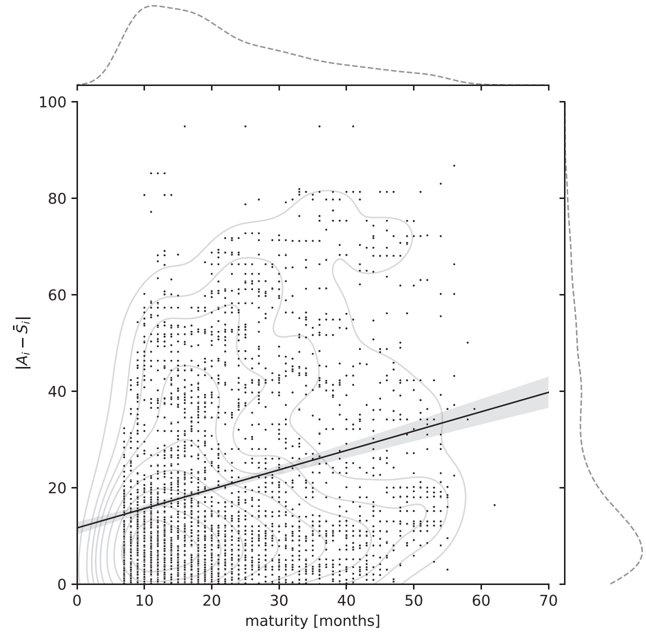

FIGURE 11 Scatter plot of the forecast errors in the form $|A_i - \bar{S}_i|$ versus *maturity*. The marginal plots show the distribution of each variable. As visual aid, the plot includes a regression line with its confidence interval and KDE contour lines. KDE, kernel density estimate

TABLE 3 Pooled OLS regression of all forecasting errors in the form $|A_i(t, T) - \bar{S}(T)|$

Variable	Coefficient	std.err.	*t*-stat.	*p*-Value
Intercept	−2.713	0.663	−4.092	0.000
Volatility	1.778	0.042	42.725	0.000
Maturity	0.084	0.024	3.550	0.000

Note: Explanatory variables volatility and maturity have a positive effect which is statistically significant. $R^2 = 0.417$.

Abbreviation: OLS, ordinary least squares.

TABLE 4 Pooled OLS regression of all analyst errors in the form $MSPE(A_i)$

Variable	Coefficient	std.err.	*t*-stat.	*p*-Value
Intercept	−563.938	269.134	−2.095	0.041
Volatility	2451.414	614.874	3.987	0.000
Maturity	9.078	13.118	0.692	0.492

Note: Only the explanatory variable volatility is statistically significant. $R^2 = 0.355$.

Abbreviations: MSPE, mean-squared prediction error; OLS, ordinary least squares.

$$\mathrm{MSPE}(A_i) = \alpha + \beta_1 \cdot volatility + \beta_2 \cdot maturity \tag{10}$$

Results are reported in Table 4 and illustrated in Figure 12. We can see that, while the maturity effect becomes statistically insignificant, the volatility remains significant at the 5% level, which suggests that MSPE estimates are volatility dependent. $R^2 = 0.355$ in this analysis.

In what follows, we propose the ratio of the MSPE to the no-change as an improved measure of predictive accuracy and estimate the following regression:

$$\phi(A_i) = \frac{\mathrm{MSPE}(A_i)}{\mathrm{MSPE}(A_i^*)} = \alpha + \beta_1 \cdot volatility + \beta_2 \cdot maturity. \tag{11}$$

Results posted in Table 5 and Figure 13 show that volatility and time to maturity are not statistically significant under the revised specification ($R^2 = 0.004$). This finding is a crucial result of the paper as it demonstrates that the no-change weighted measure controls for the volatility and maturity effects delivering an improved measure for ranking purposes.

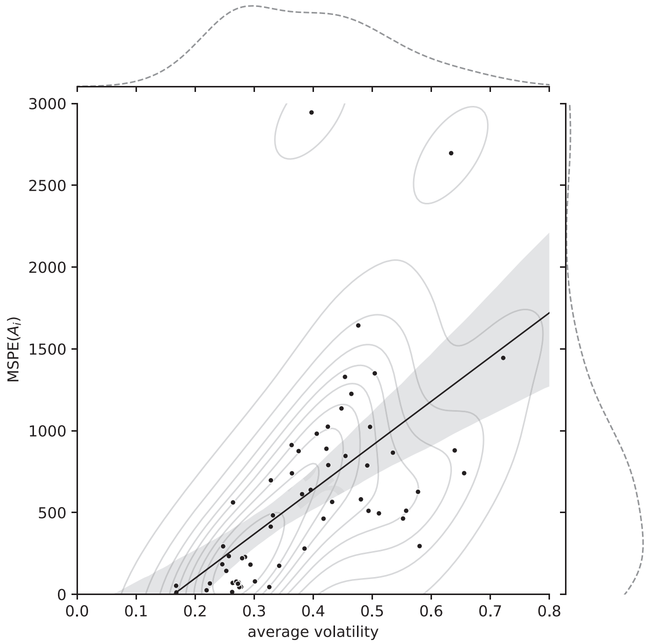

FIGURE 12 Scatter plot of the $MSPE(A_i)$ versus *average volatility*. The marginal plots show the distribution of each variable. As visual aid, the plot includes a regression line with its confidence interval and KDE contour lines. KDE, kernel density estimate; MSPE, mean-squared prediction error

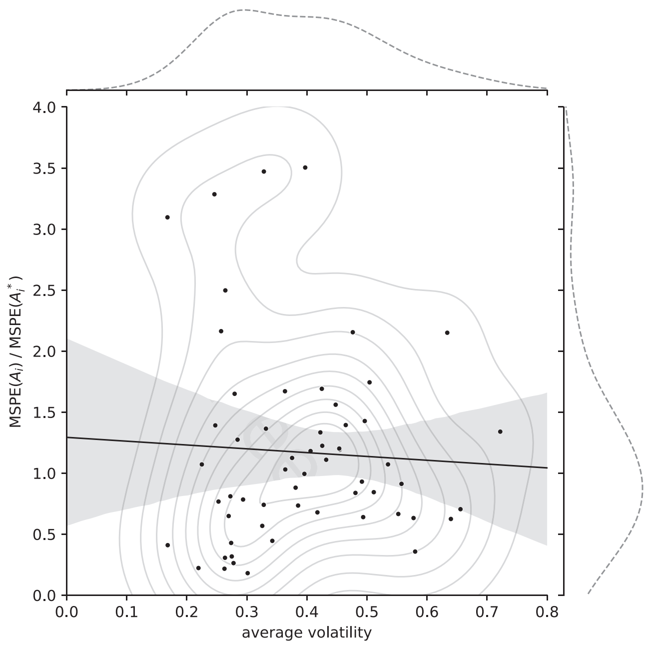

FIGURE 13 Scatter plot of $\frac{MSPE(A_i)}{MSPE(A_i^*)}$ versus *average volatility*. The marginal plots show the distribution of each variable. As visual aid, the plot includes a regression line with its confidence interval and KDE contour lines. KDE, kernel density estimate; MSPE, mean-squared prediction error

TABLE 5 Pooled OLS regression of every analyst MSPE ratio to no-change

Variable	Coefficient	std.err.	*t*-stat.	*p*-Value
Intercept	1.361	0.449	3.035	0.004
Volatility	−0.172	1.025	−0.168	0.867
Maturity	−0.005	0.022	−0.232	0.817

Note: Poor fit ($R^2 = 0.004$) and coefficients are not statistically significant.

Abbreviations: MSPE, mean-squared prediction error; OLS, ordinary least squares.

5.4 | Ranking based on the unweighted MSPE measure

Table 6 shows the analyst firms sorted by forecasting performance as established by the unweighted MSPE metric (see Equation 4). Reported results show that, in line with the results discussed in the previous section, the unweighted MSPE has substantial limitations because it fails to account for (a) the uncertainty between the release date and the target year and (b) the sample size effect.

As a result, the top analyst according to this procedure is a firm with just one forecast (see column *total forecasts* of the first row in Table 6). This reveals the need to account for the sample size effect. Reported volatility figures in column *average volatility* show that it is also evident that allegedly top performers prepare their predictions under lower *average volatility* states than those classified as worse performers. This suggests that better-ranked analysts have released forecasts under quieter market conditions, and the unweighted measure does not properly control for such volatility effect.

While there is no apparent maturity effect in the results reported in column *average maturity*, we see notable maturity biases in the remaining positions. We address the maturity effect on MSPE-based rankings in Figure 14, which demonstrates the need to control for maturity when comparing different analysts' forecasting performance. For instance, if we compare *Australia & New Zealand Banking Group Ltd.* and *Raymond James Financial Inc.*, which hold two consecutive positions (the 29th and 30th, respectively, in Table 6), we can see that they both exhibit almost identical MSPE (510.2 and 510.8, respectively) despite releasing forecasts with disparate temporal horizons.

Moreover, Figure 15 confirms that rankings also require control of the volatility effect. If we compare *Toronto-Dominion Bank/Toronto* and *Credit Suisse Group AG*, which occupy the 38th and 39th positions with almost identical MSPE (739.6 and 740.5, respectively, in Table 6), we see that they released forecasts in very different periods in terms of average price and volatility (see also Figure 9).

5.5 | Ranking by MSPE ratio to no-change

This section uses the MSPE ratio to no-change to build performance rankings. We address in this way the problems that arise with the ordinary MSPE measure and follow the standard practice in the literature (e.g., see Alquist et al., 2013). We also report the DM results on testing the null hypothesis of equal accuracy to the no-change forecast. Table 7 shows the survey constituents ordered by MSPE ratio to no-change. We exclude every analyst that exhibits a ratio greater than one, as that threshold indicates that their performance is worse than the no-change. We also drop from the ranked selection every analyst with a DM p-value greater than $\alpha = 0.05$, thus excluding every individual for which the reported outperformance is not statistically significant when compared with the no-change benchmark. We can see that the resulting ranking changes significantly with respect to the ordinary MSPE measure as the ordering is now robust to the volatility and maturity effects. By applying DM to select the best performers, we take into account statistical significance.

6 | REGRESSION OVER FIRM FEATURES

This section aims to explain the variation in the MSPE ratio to no-change through a cross-section regression.[3] To find explanatory variables for the MSPE ratio (Table 7), several firm characteristics have been considered and regressed against the MSPE ratio using *OLSs*. This is a difficult task because most of the analyst houses are not publicly traded. Ideally, the explanatory variables considered should be publicly available. For instance, they could be reported on the firm web page. They should also be on record for every firm considered so that there is no missing data in the regression. Keeping these constraints under consideration, we selected the following variables:

$$\frac{MSPE\,(A_i)}{MSPE\,(A_i^*)} = \alpha + \beta_1 USA_i + \beta_2 EU_i + \beta_3 UK_i + \beta_4 bank_i + \beta_5 expert_i + \beta_6 team_i + \beta_7 board_i + \epsilon_i, \tag{12}$$

[3]Our approach is related to the work of Roll (1984) who finds a statistically significant relationship between orange juice futures returns and subsequent errors in temperature forecasts.

TABLE 6 Initial ranking ordered by ascending MSPE (most accurate on top)

Rank	Firm	*MSPE* (A_i)	Total forecasts	% Beats nochg	% Beats futures	Average volatility	Average maturity
1	Rabobank International	11.857	1	100	100	0.168	13
2	Reel Kapital Menkul De…	14.121	3	67	67	0.263	24
3	Emirates NBD PJSC	25.003	11	73	64	0.219	18
4	Market Risk Advisory C…	44.016	21	86	71	0.275	23
5	VTB Capital PLC	44.530	2	50	50	0.326	28
6	Bank of Tokyo-Mitsubis…	45.397	5	100	60	0.278	18
7	MPS Capital Services B…	52.467	1	0	0	0.168	12
8	BBVA Research SA	64.830	1	100	0	0.273	18
9	Oxford Economics Ltd.	66.585	7	57	14	0.225	21
10	Coker Palmer Inc.	69.744	13	100	77	0.264	21
11	Bank of Nova Scotia/The	71.928	24	62	50	0.274	24
12	Schneider Electric SE	77.648	8	50	50	0.270	25
13	HSBC Holdings PLC	78.956	5	60	60	0.301	25
14	CIBC	143.343	4	75	50	0.253	15
15	Capital Economics Ltd.	174.413	86	58	49	0.342	26
16	ING Groep NV	181.402	5	60	0	0.294	28
17	Barclays PLC	183.705	11	18	27	0.246	17
18	Guggenheim Securities LLC	220.841	9	44	44	0.280	21
19	HSH Nordbank AG	228.120	13	54	23	0.285	25
20	Norddeutsche Landesban…	233.462	36	44	36	0.257	21
21	BMO Capital Markets Co…	279.646	51	57	41	0.385	33
22	Rising Glory Finance Ltd.	293.417	2	0	0	0.247	21
23	KLR Group LLC	294.879	62	81	48	0.580	37
24	Deutsche Bank AG	413.522	26	65	58	0.328	24
25	Citigroup Inc.	462.373	110	55	44	0.418	23
26	Bank of America Merril…	463.194	49	57	39	0.552	19
27	DZ Bank AG	482.600	23	52	43	0.332	15
28	UniCredit SpA	494.893	64	59	45	0.511	19
29	Australia New Zealan…	510.223	45	67	53	0.494	35
30	Raymond James Financia…	510.824	28	50	39	0.558	18
31	Jefferies LLC	561.691	5	20	20	0.264	23
32	BNP Paribas SA	564.786	81	51	33	0.432	19
33	Wells Fargo Securities…	580.528	97	67	43	0.481	25
34	ABN AMRO Bank NV	612.184	24	54	38	0.381	22
35	Banco Santander SA	626.539	179	70	47	0.578	35
36	Fitch Solutions	638.250	107	50	40	0.396	21
37	Hamburger Sparkasse AG	696.798	2	0	0	0.328	27
38	Toronto-Dominion Bank/…	739.679	32	41	19	0.364	19

TABLE 6 (Continued)

Rank	Firm	$MSPE\ (A_i)$	Total forecasts	% Beats nochg	% Beats futures	Average volatility	Average maturity
39	Credit Suisse Group AG	740.457	144	61	40	0.656	33
40	RBC	787.374	70	53	36	0.492	23
41	Commerzbank AG	789.693	132	37	28	0.426	19
42	Westpac Banking Corp	845.565	190	45	26	0.455	30
43	Societe Generale SA	865.630	239	51	35	0.535	33
44	Danske Bank A/S	875.302	23	43	26	0.375	19
45	UBS Group AG	879.540	37	81	19	0.640	35
46	Lloyds Bank PLC	889.704	30	50	23	0.422	22
47	Landesbank Baden-Wuert...	912.264	38	37	37	0.363	16
48	Natixis SA	981.590	93	41	25	0.406	31
49	Raiffeisen Bank Intern...	1023.281	116	50	43	0.496	27
50	Prestige Economics LLC	1024.474	153	42	35	0.425	23
51	Intesa Sanpaolo SpA	1136.048	115	42	28	0.448	28
52	Itau Unibanco Holding SA	1225.125	185	43	30	0.465	32
53	Macquarie Group Ltd.	1328.861	15	40	7	0.454	25
54	Nomura	1350.488	21	29	33	0.504	33
55	Oversea-Chinese Bankin...	1444.740	2	0	0	0.722	24
56	Standard Chartered Bank	1643.034	38	26	13	0.476	29
57	Sanford C Bernstein...	2695.933	72	19	8	0.634	35
58	Incrementum AG	2944.773	35	29	23	0.397	28

Note: The column *total forecasts* shows the number of forecasts on record, and *% beats nochg* and % beats futures are the percentage of times the forecasts beat the no-change model and the futures, respectively. Columns *average volatility* and average maturity aggregate values for all the forecasts released by a given analyst.

Abbreviations: MSPE, mean-squared prediction error.

USA_i headquarters located in the USA (dummy),
EU_i headquarters located in the EU (dummy),
UK_i headquarters located in the UK (dummy),
$bank_i$ bank subsidiary (dummy),
$expert_i$ expert in commodities (dummy),
$board_i$ discloses board of directors (dummy),
$team_i$ discloses analysts in research team (dummy).

First, *location of headquarters*, which is widely available in the "Contact" or "About Us" sections of corporate web pages. This is split into three dummy variables *USA*, *EU*,[4] and *UK* which are the most frequent geographical areas documented in the data set. Note that we exclude the "*rest of the world*" dummy to prevent multicollinearity. Such a case is covered when all the others are set to zero simultaneously.

Second, we consider the effect of the dichotomy between bank subsidiaries and independent specialized firms. Another dummy variable *bank* is introduced for this purpose.

[4]Liechtenstein, Switzerland, and the United Kingdom are not among the 27 state members of the European Union.

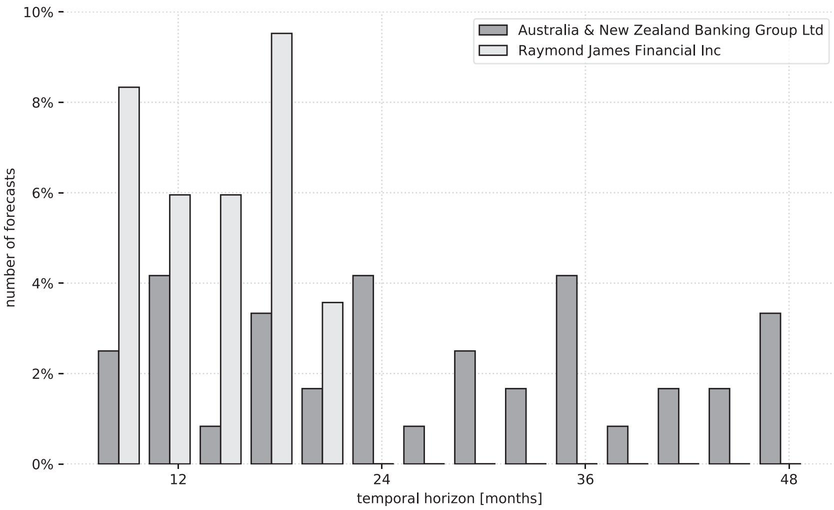

FIGURE 14 Histogram of two competing forecasts by temporal horizon. Comparison between *Australia & New Zealand Banking Group Ltd.* (dark) and *Raymond James Financial Inc.* (light). Both have almost identical MSPE (510.2 and 510.8, respectively) but the latter releases shorter-term forecasts compared to its peer. MSPE, mean-squared prediction error

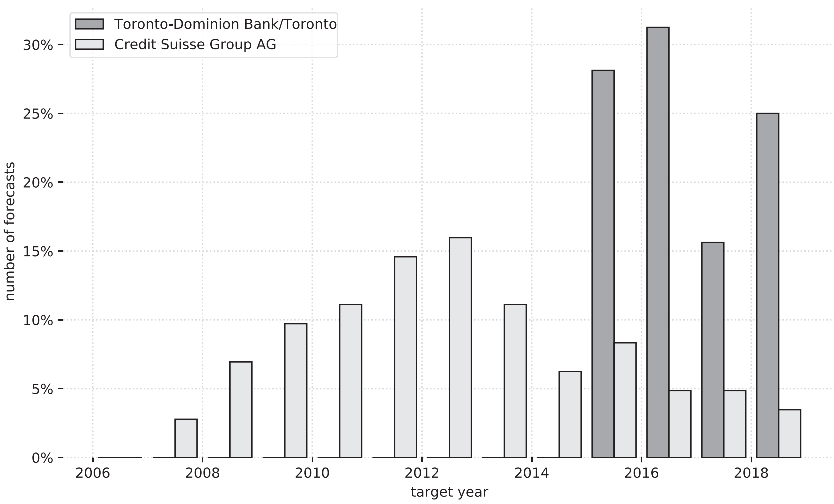

FIGURE 15 Histogram of two competing forecasts by target year. Comparison between *Toronto-Dominion Bank/Toronto* (dark) and *Credit Suisse Group AG* (light). Both have almost identical MSPE (739.6 and 740.5, respectively) but the former has a shorter history compared with its peer. Besides, they released forecasts in different periods in terms of average price and volatility. MSPE, mean-squared prediction error

Third, we address the issue of information transparency by considering (a) a dummy variable *board* to account for firm disclosure of information related to the members of the board of directors and (b) a dummy denoted as *team* that is set to one when the identities of the analysts in the research team are disclosed as well.

Finally, we introduce the variable *expert* that identifies each firm as a generalist or an expert in commodities. This classification rests on the self-reported level of expertise posted on corporate web pages. If a firm explicitly publicizes its coverage of the crude oil market and/or claims certain degree of specialization, then this dummy is set to one. Other interesting features, such as *market capitalization*, *staff size*, or *employee qualifications*, have limited availability and could not be included in the study.

The regression analysis is based on a cross-section with the MSPE ratio to the no-change benchmark for each analyst house as the dependent variable. It does not have a temporal dimension because all forecasts by the same analyst are aggregated into a single figure. Using the MSPE ratio to the no-change, the resulting aggregated data account for volatility, temporal horizon, and sample size effects. OLS regression results are displayed in Table 8. Out of the seven independent variables under consideration (*USA*, *UK*, *EU*, *bank*, *expert*, *team*, and *board*), only *USA* and *bank* are statistically significant at the 5% significance level. Note that the *expert* variable is significant at the 11% level ($p = 0.114$).

Figure 16 exhibits the correlation patterns between the analyzed cross-sectional variables. The disclosure of the *board of directors* takes place in almost every analyst house, possibly due to regulatory requirements. Therefore the discriminatory power of the variable *board* is very faint. The *expert in commodities* and disclosure of *analysts in the research team* are highly correlated but not significant. This suggests that the analyzed firms signal specialization by

TABLE 7 Revised ranking which incorporates the no-change benchmark model and the *p*-values delivered by Diebold and Mariano's test

Rank	Firm	$\frac{MSPE\,(A)}{MSPE\,(A^*)}$	DM *p*-value	Total forecasts	% Beats nochg	% Beats futures	Average volatility	Average maturity
1	Emirates NBD PJSC	0.223	0.008	11	73	64	0.219	18
2	Coker Palmer Inc.	0.307	0.000	13	100	77	0.264	21
3	Market Risk Advisory C...	0.318	0.000	21	86	71	0.275	23
4	KLR Group LLC	0.358	0.005	62	81	48	0.580	37
5	Bank of Nova Scotia/The	0.429	0.003	24	62	50	0.274	24
6	Capital Economics Ltd.	0.446	0.000	86	58	49	0.342	26
7	UBS Group AG	0.625	0.013	37	81	19	0.640	35
8	Banco Santander SA	0.633	0.000	179	70	47	0.578	35
9	Australia New Zealan...	0.640	0.006	45	67	53	0.494	35
10	Bank of America Merril...	0.666	0.018	49	57	39	0.552	19
11	Citigroup Inc.	0.679	0.000	110	55	44	0.418	23
12	Credit Suisse Group AG	0.705	0.003	144	61	40	0.656	33
13	BMO Capital Markets Co...	0.735	0.008	51	57	41	0.385	33
14	UniCredit SpA	0.844	0.047	64	59	45	0.511	19

Note: The ranking is ordered by ascending the MSPE ratio to no-change so that the most accurate forecast is reported in the first row. Analysts with an MSPE ratio >1, DM *p*-value $>\alpha$ (0.05) or less than 10 forecasts on record are dropped from the list. The column *DM p-value* specifies the statistical significance of the test of equal accuracy between the given analyst and the corresponding no-change benchmark. The optimal truncation lag is calculated per analyst according to Andrews' (1991) data-based rule.

Abbreviations: DM test, Diebold and Mariano test; MSPE, mean-squared prediction error.

TABLE 8 Summary of the OLS pooled regression of MSPE ratio to no-change against firm features ($R^2 = 0.25$)

Variable	Coefficient	std.err.	*t*-stat.	*p*-Value
Intercept	2.375	0.523	4.538	0.000
Board	−0.201	0.382	−0.527	0.602
Team	0.171	0.310	0.550	0.587
Expert	−0.552	0.339	−1.629	0.114
Bank	−0.916	0.340	−2.694	0.011
EU	−0.070	0.258	−0.270	0.789
USA	−0.622	0.298	−2.088	0.045
UK	−0.124	0.400	−0.311	0.758

Note: Only analysts with at least 15 valid forecasts on record ($n = 38$). Features are modeled as dummy variables (0 or 1). Being a *bank* subsidiary (rather than an independent firm) and having headquarters in the USA provide a statistically significant advantage in the form of error reduction. *expert* is close to being significant as well ($p = 0.114$).

Abbreviations: MSPE, mean-squared prediction error; OLS, ordinary least squares.

disclosing the characteristics of the analyst team. *USA* and *EU* are also moderately (and negatively) correlated, mainly due to their strong prevalence in the data set: if a firm is not located in the *USA*, then it is very likely to be in the *EU* and vice versa.

The overall conclusion of this section is that firms based in the USA and bank subsidiaries tend to outperform their peers, including allegedly specialized firms. The sign and value of the *bank* and *USA* coefficients provide insight into the direction and magnitude of the effect on the forecast error, suggesting that analysts operating in the US and working for bank subsidiaries release more accurate forecasts.

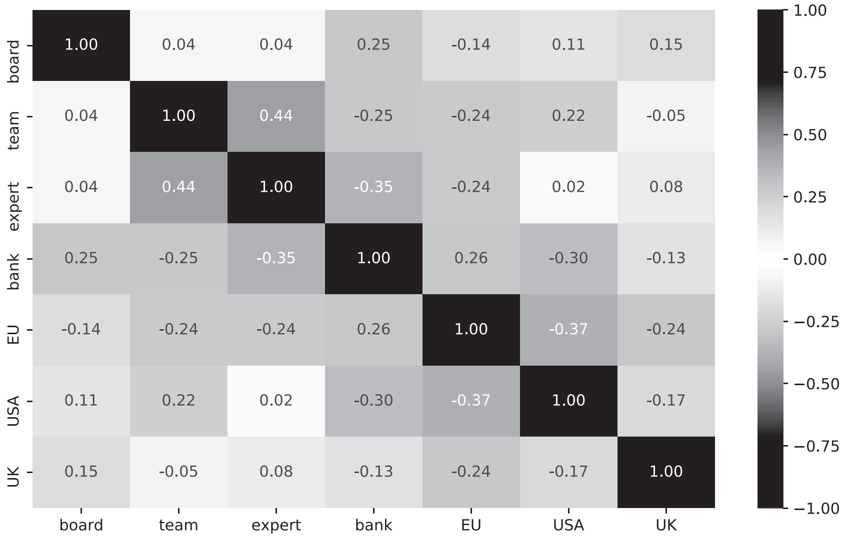

FIGURE 16 Correlation matrix, which shows a strong correlation between *team* ↔ *expert* (0.44) and *EU* ↔ *USA* (−0.37). Diverging colormap in absolute values

7 | DISCUSSION

The main finding of this paper is that analysts' predictions underperform futures in forecasting future spot prices. The interpretation of this result is a challenging task as analysts are, in principle, informed traders that have access to good quality data on crude oil fundamentals. Analysts gather information from multiple sources, including futures markets as well as their peers. However, futures prices are established by matching the supply and demand in the market after multiple participants have taken decisions based on the information available, including the analysts' forecasts. Consequently, there is a tight cross-sectional coupling between both parties (analysts and futures) as well as a temporal feedback loop because their outputs are continuously being fed as new inputs to this complex system.

So, what is the reason that explains the underperformance of the analyst? One of the first researchers to address the forecasting ability of market analysts was Cowles (1933), who noticed the lackluster track record of stock recommendations and predictions released by prominent experts at the time. In his study, he concludes that their performance did not exhibit any special ability and was probably a consequence of pure chance.

A possible explanation of why analysts deliver the weakest predictions is the fact that they are simply human. In their quest to discover a source of rationality in the market, De Bondt and Thaler (1990) considered the segment of the analysts. Contrary to what was originally expected, their work concludes that the changes in the forecasts and other behavioral patterns were too extreme to be accepted as strictly rational. Instead, analysts exhibit human traits, such as overconfidence and self-attribution bias. The former causes them to lean towards private over public information, while the latter is understood as the predisposition to ascribe success to their own skill and failure to external factors, such as misfortune (Daniel et al., 1998; Friesen & Weller, 2006).

As human beings, analysts have flaws, respond to incentives and self-interest. Wolinsky (1993) studied the organization of the markets where sellers happen to be the experts (such as financial markets) and argues that, due to information asymmetry, incentives to opportunistic behavior by the sellers can easily arise.

Therefore, delivering accurate forecasts may not always be a top priority for analysts, as their reputation and success depend on multiple factors. In fact, Hong and Kubik (2003) suggest that investment banks and brokerage houses are not only concerned about forecasting accuracy. Instead, they also reward analysts who optimistically promote stocks because such a tactic attracts underwriting business and trading commissions. Furthermore, Fang and Yasuda (2009) posit that while analyst may be motivated to provide forecast quality for reputation-related reasons, reputation alone is not an effective deterrent against conflict of interests. Consequently, the existence of such conflicts under the principal-agent problem (Ross, 1973) prevents good forecasting performance.

Analysts performance is also motivated by the Wall Street Analyst rankings. A top position on the ladder leads to exceptional compensation and job prospects. However, in practice, forecast accuracy does not weigh much in the selection criteria, and a significant amount of bias is introduced in the eligibility requirements. Voters are additionally subject to lobbying and might not be impartial. Consequently, Emery and Li (2009) disregard such rankings as “popularity contests” and prove that there are no significant differences among the ex post performance of stars and nonstars, reducing the value that such classifications offer to investors.

Croushore (1997) identifies two strategies that analysts follow to maximize their payoff. He suggests that some analysts may shift their forecasts towards the market consensus as a defensive move to mitigate reputational damage if their estimates turn out to be inaccurate. In contrast, other groups release bold forecasts hoping to stand out of the crowd.

Ottaviani and Sørensen (2006b) analyze in depth the idea of strategic dichotomy and develop the “reputational cheap talk” and the contest theories, which constitute the foundation of subsequent research on the topic.

The reputational theory (Ottaviani & Sørensen, 2006b) is developed under the idea that analysts hope to convince the market that they are well informed. Consequently, they are tempted to confirm the original belief of the market by making predictions closer to the prior consensus. They achieve this by disregarding the private information and instead favoring the public view. This strategy is self-defeating as it results in an excessive level of agreement among forecasters, commonly known as “herding.” Other causes of such lack of differentiation arise due to correlated information and incentive structures that encourage imitation (Jegadeesh & Kim, 2010; Scharfstein & Stein, 1990; Trueman, 1994).

On the other hand, the contest theory (Ottaviani & Sørensen, 2006b) describes a winner-take-all competitive environment. In this scenario, analysts exaggerate their predictions away from the expected consensus. This bold move reduces the probability of winning, but on the off-chance of getting it right, their visibility increases and so does the reward as fewer peers are likely to share the same guess.

On a related note, Ottaviani and Sørensen (2006a) suggest that in a competitive environment the experts benefit from having an informed signal and exploit their position by delivering more extreme messages. This allows them to pretend that they have more profound knowledge and forecasting ability than the crowd, leading to the intentional bias of forecasts.

Marinovic et al. (2013) propose a unified, more theoretical framework to model analysts' approaches which combine the strategies described above. They contribute by introducing a unique objective function that aggregates reputational and contest elements.

Note that the corporations considered in the analysis constitute financial firms and banking institutions that are not necessarily specialists in crude oil. Alquist et al. (2013) argue that crude oil analysts that provide forecasts at *Consensus Economics Inc.* are professional macroeconomic forecasters that are not experts in the oil market. This may also be argued for the list of firms considered in this analysis.

Therefore, the overall conclusion is that the literature suggests several frameworks consistent with the finding that analysts do not provide the most accurate prediction of future spot oil prices. Due to principal-agent and behavioral reasons, incentives are not always aligned with the quality of the prediction. In other words, the quality of the forecasts is not the only determinant of an analyst's payoff. Futures may be more precise because they aggregate information and benefit from the “wisdom of the crowds.”

8 | CONCLUSIONS

The main objective of this paper is to assess the forecasting ability of crude oil analysts. As a first approach, we compare the accuracy of analysts' predictions and the corresponding futures contracts in predicting the spot price. In the second stage, we build a ranking of analysts sorted by forecasting performance. The data used in the analysis includes analysts' forecasts, futures contracts, and spot prices are all retrieved from the Bloomberg database.

Forecast quality is evaluated using a standard approach. We follow the current literature and apply the MSPE error measure relative to the no-change forecasting model as a baseline metric. We also apply Diebold and Mariano's method to test for statistical significance. Our analysis demonstrates that the MSPE relative to the no-change facilitates the comparison of forecasts among analysts and across time periods. Our results may be summarized as follows:

First, a set of preliminary regressions demonstrate that the unweighted MSPE provides relative performance measures that suffer from important limitations as they penalize: (i) forecasts that are produced during volatile periods (volatility effect), (ii) firms that perform a large number of forecasts (size effect), and (iii) forecasts that have long-term horizons (maturity effect). Our paper explicitly shows that the MSPE ratio to the no-change model controls for the volatility and maturity effects while applying the Diebold and Mariano methodology allows for testing for statistical significance. We believe that this is an important contribution of the paper.

Second, the ratio of MSPE to the no-change shows that futures prices provide more accurate forecasts of future spot prices than the analysts' predictions at the aggregate level. More than two-thirds of the firms (41 out of 58) exhibit a performance inferior to the alternative based on futures contracts.

Third, a cross-section regression of the relationship between forecasting performance and several firm features shows that firms based in the USA and banks exhibit a statistically significant advantage. The analysts' expertise level turns out to be marginally statistically significant.

We explain our findings in terms of a principal-agent problem in which an analyst firm (the agent) does not necessarily act to maximize the payoff to the investor (the principal). We also argue that there may also be specific behavioral factors that explain our findings. To our knowledge, this is the first paper in the literature that compares the forecasting performance of analysts and futures prices that explicitly addresses the role of the no-change model benchmark in capturing the volatility and maturity factors. Our results contribute to the price discovery and forecasting literature of futures markets.

ACKNOWLEDGMENTS

We are grateful to Carlos Bellón for his comments. Furthermore, we would like to thank Ramón Bermejo for his assistance with the Bloomberg Terminal and Ziwenxi Wang for her work in the data retrieval process. We appreciate the internal financial help from ICADE, Universidad Pontificia Comillas, grant PP2018_03 and the MINECO grant PID2019-104960GB-I00.

DATA AVAILABILITY STATEMENT

The data and source code that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Isabel Figuerola-Ferretti http://orcid.org/0000-0001-7650-9709
Alejandro Rodríguez https://orcid.org/0000-0003-0700-1308
Eduardo Schwartz http://orcid.org/0000-0002-7424-5487

REFERENCES

Alquist, R., & Kilian, L. (2010). What do we learn from the price of crude oil futures? *Journal of Applied Econometrics*, *25*(4), 539–573. https://doi.org/10.1002/jae.1159

Alquist, R., Kilian, L., & Vigfusson, R. J. (2013). Forecasting the price of oil. In G. Elliott & A. Timmermann (Eds.), *Handbook of economic forecasting* (Vol. 2A, pp. 427–507). https://doi.org/10.1016/B978-0-444-53683-9.00008-6

Anderson, S. T., Kellogg, R., Sallee, J. M., & Curtin, R. T. (2011). Forecasting gasoline prices using consumer surveys. *American Economic Review*, *101*(3), 110–114. https://doi.org/10.1257/aer.101.3.110

Andrews, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, *59*(3), 817. https://doi.org/10.2307/2938229

Ashley, R. (1998). A new technique for postsample model selection and validation. *Journal of Economic Dynamics and Control*, *22*(5), 647–665. https://doi.org/10.1016/s0165-1889(97)00085-7

Baumeister, C., & Kilian, L. (2012). Real-time forecasts of the real price of oil. *Journal of Business & Economic Statistics*, *30*(2), 326–336. https://doi.org/10.1080/07350015.2011.648859

Bernanke, B. S. (2004). *FRB speech: Oil and the economy*. Darton College. Retrieved from https://www.federalreserve.gov/boardDocs/speeches/2004/20041021/default.htm

Baumeister, C., Kilian, L., & Lee, T. K. (2014). Are there gains from pooling real-time oil price forecasts? *Energy Economics*, *46*(S1), S33–S43. https://doi.org/10.1016/j.eneco.2014.08.008

Bianchi, D., & Piana, J. (2017). *Expected spot prices and the dynamics of commodity risk premia*. University of Warwick.

Black, F. (1976). The pricing of commodity contracts. *Journal of Financial Economics*, *3*(1–2), 167–179. https://doi.org/10.1016/0304-405X(76)90024-6

Bloomberg Finance, L. P. (2020a). WTI oil spot prices, 05/17/1983 to 11/06/2020. Retrieved 11/07/2020 from Bloomberg Terminal https://www.bloomberg.com

Bloomberg Finance, L. P. (2020b). NYMEX WTI oil futures contracts CL1 through CL104, 07/18/1986 to 11/17/2020. Retrieved 11/18/2020 from Bloomberg Terminal https://www.bloomberg.com

Bloomberg Finance, L. P. (2020c). Crude oil forecasts from the Commodity Prices Forecasts Survey, 01/31/2007 to 11/12/2020. Retrieved 11/11/2020 from Bloomberg Terminal https://www.bloomberg.com

Brennan, M. J., Jegadeesh, N., & Swaminathan, B. (1993). Investment analysis and the adjustment of stock prices to common information. *Review of Financial Studies*, *6*(4), 799–824. https://doi.org/10.1093/rfs/6.4.799

Cervera, I., & Figuerola-Ferretti, I. (2019). *Recent credit risk and bubble behavior in the corporate energy sector*. Retrieved from https://repositorio.comillas.edu/xmlui/handle/11531/26571

Chang, C., Daouk, H., & Wang, A. (2009). Do investors learn about analyst accuracy? A study of the oil futures market. *Journal of Futures Markets*, *29*(5), 414–429. https://doi.org/10.1002/fut.20374 arXiv. https://arxiv.org/abs/arXiv:1406.6943v1

Chen, L. H., Jiang, G. J., & Wang, Q. (2013). Market reaction to information shocks—Does the Bloomberg and Briefing.com survey matter? *Journal of Futures Markets*, *33*(10), 939–964. https://doi.org/10.1002/fut.21564 arXiv. https://arxiv.org/abs/arXiv:1406.6943v1

Coppola, A. (2008). Forecasting oil price movements: Exploiting the information in the futures market. *Journal of Futures Markets*, *28*(1), 34–56. https://doi.org/10.1002/fut.20277 arXiv. https://arxiv.org/abs/arXiv:1406.6943v1

Cortazar, G., Millard, C., Ortega, H., & Schwartz, E. S. (2019). Commodity price forecasts, futures prices, and pricing models. *Management Science*, *65*(9), 4141–4155. https://doi.org/10.1287/mnsc.2018.3035

Cowles, A. (1933). Can stock market forecasters forecast. *Econometrica*, *1*(3), 309. https://doi.org/10.2307/1907042

Croushore, D. (1997). The Livingston survey: Still useful after all these years. *Federal Reserve Bank of Philadelphia Business Review*, *2*(March/April), 1–13.

Daniel, K., Hirshleifer, D., & Subrahmanyam, A. (1998). Investor psychology and security market under- and overreactions. *The Journal of Finance*, *53*(6), 1839–1885. https://doi.org/10.1111/0022-1082.00077

De Bondt, W. F. M., & Thaler, R. H. (1990). Do security analysts overreact? *American Economic Review*, *80*(2), 52. https://doi.org/10.2307/2006542

Diebold, F. X. (2015). Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of Diebold–Mariano tests. *Journal of Business and Economic Statistics*, *33*(1), 1. https://doi.org/10.1080/07350015.2014.983236

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *13*(3), 253–263. https://doi.org/10.2307/1392185

Emery, D. R., & Li, X. (2009). Are the Wall Street Analyst rankings popularity contests. *Journal of Financial and Quantitative Analysis*, *44*(2), 411–437. https://doi.org/10.1017/S0022109009090140

Fang, L., & Yasuda, A. (2009). The effectiveness of reputation as a disciplinary mechanism in sell-side research. *Review of Financial Studies*, *22*(9), 3735–3777. https://doi.org/10.1093/rfs/hhn116

Figuerola-Ferretti, I., & Gilbert, C. L. (2005). Price discovery in the aluminum market. *Journal of Futures Markets*, *25*(10), 967–988. http://dx.doi.org/10.1002/fut.20173

Figuerola-Ferretti, I., & Gonzalo, J. (2010). Modelling and measuring price discovery in commodity markets. *Journal of Econometrics*, *158*(1), 95–107. https://doi.org/10.1016/j.jeconom.2010.03.013

Figuerola-Ferretti, I., McCrorie, J. R., & Paraskevopoulos, I. (2020). Mild explosivity in recent crude oil prices. *Energy Economics*, *87*, 104387. https://doi.org/10.1016/j.eneco.2019.05.002

Friesen, G., & Weller, P. A. (2006). Quantifying cognitive biases in analyst earnings forecasts. *Journal of Financial Markets*, *9*(4), 333–365. https://doi.org/10.1016/j.finmar.2006.07.001

Garbade, K. D., & Silber, W. L. (1983). Price movements and price discovery in futures and cash markets. *The Review of Economics and Statistics*, *65*(2), 289. https://doi.org/10.2307/1924495

Gibson, R., & Schwartz, E. S. (1990). Stochastic convenience yield and the pricing of oil contingent claims. *The Journal of Finance*, *45*(3), 959–976. https://doi.org/10.2307/2328801

Harvey, D., Leybourne, S., & Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, *13*(2), 281–291. https://doi.org/10.1016/S0169-2070(96)00719-4

Hong, H., & Kubik, J. D. (2003). Analyzing the analysts: Career concerns and biased earnings forecasts. *The Journal of Finance*, *58*(1), 313–351. https://doi.org/10.1111/1540-6261.00526

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001

Jegadeesh, N., & Kim, W. (2010). Do analysts herd? An analysis of recommendations and market reactions. *Review of Financial Studies*, *23*(2), 901–937. https://doi.org/10.1093/rfs/hhp093

Kilian, L. (1999). Exchange rates and monetary fundamentals: What do we learn from long-horizon regressions? *Journal of Applied Econometrics*, *14*(5), 491–510. https://doi.org/10.1002/(SICI)1099-1255(199909/10)14:5<491::AID-JAE527>3.0.CO;2-D

Kilian, L. (2017). The impact of the fracking boom on Arab oil producers. *Energy Journal*, *38*(6), 137–160. https://doi.org/10.5547/01956574.38.6.lkil

Kilian, L., & Taylor, M. P. (2003). Why is it so difficult to beat the random walk forecast of exchange rates. *Journal of International Economics*, *60*(1), 85–107. https://doi.org/10.1016/S0022-1996(02)00060-0

Ljungqvist, A., Malloy, C., & Marston, F. (2009). Rewriting history. *The Journal of Finance*, *64*(4), 1935–1960. https://doi.org/10.1111/j.1540-6261.2009.01484.x

Marinovic, I., Ottaviani, M., & Sorensen, P. (2013). Forecasters' objectives and strategies. In *Handbook of economic forecasting* (Vol. 2B, pp. 690–720). Elsevier. https://doi.org/10.1016/B978-0-444-62731-5.00012-9

Mark, N. (1995). Exchange rates and fundamentals: Evidence on long-horizon predictability. *American Economic Review*, *85*(1), 201–218. https://doi.org/10.2307/2118004

Ottaviani, M., & Sørensen, P. N. (2006a). Reputational cheap talk. *The RAND Journal of Economics*, *37*(1), 155–175. https://doi.org/10.1111/j.1756-2171.2006.tb00010.x

Ottaviani, M., & Sørensen, P. N. (2006b). The strategy of professional forecasting. *Journal of Financial Economics*, *81*(2), 441–466. https://doi.org/10.1016/j.jfineco.2005.08.002

Peck, A. E. (1985). The economic role of traditional commodity futures markets. In *Futures markets: Their economic role* (pp. 1–81). American Enterprise Institute for Public Policy Research. Retrieved from http://goo.gl/bsb7J8

Roll, R. (1984). Orange juice and weather. *American Economic Review*, *74*(5), 861–880. https://doi.org/10.2307/549

Ross, S. A. (1973). The economic theory of agency: The principal's problem. *American Economic Review*, *63*(2), 134–139. https://doi.org/10.2307/1817064

Sanders, D. R., Manfredo, M. R., & Boris, K. (2009). Evaluating information in multiple horizon forecasts: The DOE's energy price forecasts. *Energy Economics*, *31*(2), 189–196. https://doi.org/10.2307/2962324

Scharfstein, D., & Stein, J. C. (1990). Herd behavior and investment. *American Economic Review*, *80*(3), 465–479.

Schwartz, E. S. (1997). The stochastic behavior of commodity prices: Implications for valuation and hedging. *The Journal of Finance*, *52*(3), 923–973. https://doi.org/10.2307/2329512

Schwartz, E. S., & Smith, J. E. (2000). Short-term variations and long-term dynamics in commodity prices. *Management Science*, *46*(7), 893–911. https://doi.org/10.1287/mnsc.46.7.893.12034

Svensson, L. E. (2005). Oil prices and ECB monetary policy. *Briefing Paper for the Committee on Economic and Monetary Affairs of the European Parliament*, 1–4.

Trueman, B. (1994). Analyst forecasts and herding behavior. *Review of Financial Studies*, *7*(1), 97–124. https://doi.org/10.1093/rfs/7.1.97

Wiser, R., Jenni, K., Seel, J., Baker, E., Hand, M., Lantz, E., & Smith, A. (2016). Expert elicitation survey on future wind energy costs. *Nature Energy*, *1*(10), 16135. https://doi.org/10.1038/nenergy.2016.135

Wolinsky, A. (1993). Competition in a market for informed experts' services. *The RAND Journal of Economics*, *24*(3), 380. https://doi.org/10.2307/2555964

Working, H. (1948). Theory of the inverse carrying charge in futures markets. *Journal of Farm Economics*, *30*(1), 1–28. https://doi.org/10.2307/1232678

How to cite this article: Figuerola-Ferretti, I., Rodríguez, A., & Schwartz, E. (2021). Oil price analysts' forecasts. *Journal of Futures Markets*, 1–24. https://doi.org/10.1002/fut.22225