

RESEARCH ARTICLE



Volatility forecasts embedded in the prices of crude-oil options

Dudley Gilder¹  | Leonidas Tsiaras² 

¹Department of Accounting and Finance,
Cardiff Business School, Cardiff
University, Cardiff, UK

²Department of Economics, Finance and
Entrepreneurship, Aston Business School,
Aston University, Birmingham, UK

Correspondence

Leonidas Tsiaras, Department of
Economics, Finance and
Entrepreneurship, Aston Business School,
Aston University, B4 7ET Birmingham,
UK.

Email: l.tsiaras@aston.ac.uk

Abstract

This paper evaluates the ability of alternative option-implied volatility measures to forecast crude-oil return volatility. We find that a corridor implied volatility measure that aggregates information from a narrow range of option contracts consistently outperforms forecasts obtained by the popular Black–Scholes and model-free volatility expectations, as well as those generated by a realized volatility model. This measure ranks favorably in regression-based tests, delivers the lowest forecast errors under different loss functions, and generates economically significant gains in volatility timing exercises. Our results also show that the Chicago Board Options Exchange's “oil-VIX” index performs poorly, as it routinely produces the least accurate forecasts.

KEYWORDS

option-implied volatility, realized variance, volatility forecasting

JEL CLASSIFICATION

C53; C58; G17

1 | INTRODUCTION

In economic terms, crude-oil is the most important traded commodity. Unsurprisingly, a wide range of economic agents, from individual investors to policy makers, closely monitor its price and routinely attempt to make predictions about the future. Unlike standard financial assets, however, one salient feature of crude-oil prices is that they can experience dramatic shifts for reasons that are largely unrelated to global macroeconomic conditions, such as OPEC policy changes or geopolitical instability in oil-producing regions. It is therefore tempting to expand the information set of standard time-series models, which rely exclusively on the record of historical prices, with measures that have “forward-looking” features by construction.

Traded options appear as a natural candidate for this task. In theory, market efficiency dictates that their prices should instantaneously adjust to any piece of relevant information pertaining to their expiration

horizon.¹ In practice, option-implied measures have indeed shown remarkable forecast performance across a variety of markets and asset classes. As evidenced by the voluminous literature, this is particularly true for volatility forecasting.^{2,3}

However, given a panel of option prices, different measures of expected, risk-neutral, return variation can be explored, so empirical evidence contrasting the various alternatives are needed to proceed. In this paper, we attempt to fill this gap for the case of crude-oil volatility. We do so by comparing the forecast performance, using both statistical and economic criteria, of various popular alternatives, such as the Black–Scholes at-the-money implied volatility (ATMIV), the so-called model-free implied volatility (MFIV), and the Crude Oil Volatility Index (OVX) published by the Chicago Board of Options Exchange (CBOE). More important, we investigate, for the first time in the commodities literature, corridor implied volatility (CIV) measures introduced by Carr and Madan (1998) and Andersen and Bondarenko (2007) and obtain very promising results.

The main strand of the crude-oil volatility forecasting literature focuses on the performance of models belonging either to the GARCH family, introduced by Bollerslev (1986), or those based on realized variance measures, notably the heterogeneous autoregressive (HAR) model of Corsi (2009) and its extensions. Sadorsky (2006) finds that simple GARCH models perform better than more complicated alternatives such as bivariate GARCH, vector autoregression, and state-space models. Kang, Kang, and Yoon (2009) compare various GARCH models with a particular focus on those that exhibit long memory. More recently, Sévi (2014) and Prokopczuk, Symeonidis, and Wese (2016) utilize high frequency returns and compare the original HAR model of Corsi (2009) with its variants that separately model the continuous and discontinuous variance components. Interestingly, both studies find that jump decompositions do not outperform the simple HAR model in terms of out-of-sample forecast accuracy.

Commencing with Latané and Rendleman (1976), there is a large collection of papers examining the information content of option-implied volatilities for the purpose of forecasting, with a handful of papers focusing on crude-oil. Naturally, early studies focused on producing forecasts using information in the ATMIV. Day and Lewis (1993) were the first to examine this measure and found that ATMIV-based forecasts encompass those generated by GARCH and EGARCH models. Szakmary, Ors, Kim, and Davidson (2003) also compared ATMIV with GARCH-based forecasts for 35 different assets, including crude-oil, and found that ATMIV was superior in their forecasting comparisons. Similarly, Martens and Zein (2004) find that option-implied forecasts rank favorably against GARCH alternatives, although the best models are those that combine both option-implied and high-frequency return information. Agnolucci (2009) also favors models that combine GARCH and option-implied forecasts, however, the former appear to be more accurate than the latter in individual forecast comparisons.

While the early literature has examined the information content of Black–Scholes implied volatilities calculated from different strikes and maturities (Beckers, 1981; Chiras & Manaster, 1978; Fung, Lie, & Moreno, 1990; Gemmill, 1986; Trippi, 1977) the consensus is that the simple ATMIV of a contract expiring as close to the forecast horizon appears to provide the most reliable results. More recently, ATMIV forecasts have been compared to model-free volatility expectations that have a number of appealing theoretical properties.⁴ The empirical evidence, however, has produced inconclusive results. Jiang and Tian (2005) study the S&P500 index and find MFIV to be more informative than ATMIV, while the opposite conclusion is reached by Andersen and Bondarenko (2007) for the same underlying asset. Taylor, Yadav, and Zhang (2010) examine individual U.S. stocks and report that ATMIV provides more accurate volatility forecasts than its model-free counterpart. Finally, in their study of three energy markets, Prokopczuk and Simen (2014) find that MFIV is more informative than ATMIV in predicting either crude-oil, heating oil, or natural gas volatility. They also find that a simple adjustment for volatility risk-premia enhances the forecast performance of all option-implied measures.

When the task at hand is to predict future return variation MFIV is not without shortcomings. This is mainly for two reasons. First, some options included in the calculation of this measure (such as deep out-of-the money puts for the case of equities) tend to be very sensitive to volatility risk-premia fluctuations. This can introduce substantial variation in the

¹Melick and Thomas (1997) provide an insightful study of how the option-implied probability distributions of oil prices reflected market commentary during the first Gulf War.

²See Christoffersen, Jacobs, and Chang (2013), for a recent review of literature concerning the forecasting performance of various option-implied measures, including volatility.

³Although, strictly speaking, this paper focuses on forecasting variance, we will use the terms “volatility” and “variance” interchangeably.

⁴As shown in Carr and Madan (1998), Demeterfi, Derman, Kamal, and Zou (1999), and Britten-Jones and Neuberger (2000), MFIV is a, risk-neutral, ex ante expectation of future return variation and, unlike the Black–Scholes implied volatilities, can be directly obtained from observed option price data without assuming a particular option pricing model.

option-implied measures that is largely unrelated to the forecast target, namely integrated variance (IVAR). Second, calculating MFIV requires that market prices of options with extreme strikes are observed. In practice, this means that either some extrapolation scheme must be implemented, or that options beyond a certain strike range should be excluded from the calculation.

The most popular estimates of MFIV measures are the volatility indices produced and published by CBOE, such as the volatility index (VIX) for the case of the S&P500 and the OVX for the case of crude-oil. CBOE's implementation algorithm, which is common for both the VIX and OVX indices, adopts a liquidity-based cut-off point that determines the range of options to be included in the MFIV measure calculation. The choice of this algorithm by CBOE has recently attracted some criticism. Andersen and Bondarenko (2007) were the first to note that the VIX is in fact an *ex ante* measure of corridor integrated variance (CIVAR), rather than IVAR. In a comprehensive empirical study, Andersen, Bondarenko, and Gonzalez-Perez (2015) use high-frequency option data and report that the VIX calculation method introduces systematic biases to the extracted measure, including artificial jumps, which become particularly pronounced during periods of market stress. From a different viewpoint, Griffin and Shams (2018) put forth evidence pointing toward market manipulation of the VIX futures market. In essence, this is facilitated by CBOE's adopted cut-off algorithm, as speculators can temporarily boost the liquidity of deep out-of-the money S&P 500 options, increasing the level of the VIX just before the settlement price for VIX futures is determined. Given that the same methodology is used to calculate both the VIX and OVX indices, all the above raise reasonable concerns regarding the informational efficiency of OVX-based forecasts. In addition, since the popularity of volatility indices has recently become widespread in the finance industry, a comparison between the OVX and other option-implied alternatives appears to be long overdue.

Our work builds on the study of Andersen and Bondarenko (2007) who explore an alternative measure of *ex ante* risk-neutral expectation of volatility, the so-called CIV. Similar to the MFIV, and unlike the Black–Scholes model, this measure aggregates volatility information from several options and does not depend on a particular option pricing model. However, the extracted measure is not a risk-neutral expectation of IVAR but CIVAR, that is return variation accumulated only when the asset price lies within a corridor of two prespecified price levels. The advantage of this approach is that one can select a corridor width that, while containing a wide-range of option prices, excludes those with extreme strikes, avoiding both price extrapolations and liquidity-driven cut-off points that may influence the reliability of the extracted measure. Moreover, since each corridor corresponds to a different range of option data, one can explore CIV measures that may be less sensitive to volatility risk-premia fluctuations.⁵

The contribution of this paper is threefold. First, we examine the forecast performance of CIV measures vis-à-vis a collection of competing alternatives, including HAR, MFIV, OVX, and ATMIV forecasts, for the case of crude-oil. Our paper builds, but significantly expands, on the work done by Prokopczuk and Simen (2014) who compare the performance of MFIV and ATMIV forecasts. Besides considering additional option-implied measures, our study also includes models that utilize high-frequency return information, while forecasts are ranked using both statistical and economic criteria. Second, we evaluate volatility forecasts for the case of the United States Oil Fund (USO), a exchange-traded fund (ETF) that attempts to track the price of West Texas Intermediate light sweet crude-oil. Considering this alternative, yet closely related, target quantity, enables us to further scrutinize our earlier findings. Moreover, to the best of our knowledge, we are first to construct and evaluate option-implied forecasts for this data set. Third, we provide the first empirical evaluation of the OVX index, used in the forecasting study of Haugom, Langeland, Molnár, and Westgaard (2014), against other option-implied alternatives. We do so for both crude-oil forecasts, which are more important in applied practice, as well as for USO volatility forecasts, which enable more reliable methodological comparisons. This is because USO options constitute the basis for the OVX calculation.

Our empirical results provide insights on a number of issues. We find that a particular CIV measure, that uses a relatively narrow range of option prices, consistently ranks favorably against all other competing measures using a variety of statistical and economic criteria. In particular, model forecasts that utilize this measure achieve the highest R^2 in Minzer-Zarnowitz regressions, remain significant in encompassing regression tests, and deliver the most accurate forecasts under both the symmetric and asymmetric loss functions we consider. Moreover, volatility timing exercises show that utilizing this measure results in significant economic gains. The superior performance of this narrow CIV

⁵For the case of the S&P 500, Dotsis and Vlastakis (2016) decompose MFIV into various corridor components and show that, indeed, each CIV measure reflects different risk premia.

measure is evident in our full crude-oil data set (1996–2016), our “postfinancialization” sample⁶ period (2007–2016), as well as, our USO ETF series (2007–2016).

With respect to the performance of the OVX, we find clear evidence that it provides problematic crude-oil volatility forecasts as it is routinely outperformed by other option-based alternatives. When the target quantity is USO volatility, the OVX generates improved, but far from optimal, forecasts. Our complementary analysis of the algorithm used to calculate the OVX indicates that the currently adopted cut-off rule introduces noise which hinders its forecast accuracy. Finally, in contrast to Prokopczuk and Simen (2014), we find that ATMIV is more informative about future crude-oil volatility than the MFIV measure,⁷ although the opposite is true for the case of USO volatility forecasts.

The structure of the paper is as follows. Section 2 discusses various measures of volatility that we use in this study. Section 3 describes the data set and the details of our methodology. The empirical results for our crude-oil data set are presented in Section 4 while those corresponding to the USO ETF are presented in Section 5. Our robustness checks can be found in Section 6. Section 7 concludes.

2 | VOLATILITY MEASURES

In this section, we describe the alternative volatility measures we use to construct forecasts. However, before doing so, we state the assumptions we make about asset price dynamics and outline the relevant theory on which all our volatility measures are based.

2.1 | The dynamics of futures prices

Assume that over the period $t \in [0, T]$ investors can continuously trade in a frictionless and arbitrage-free market. In the filtered probability space $(\Omega, \mathbb{F}, \mathbb{P}; \mathbb{F}_{t \in [0, T]})$, the futures price of a contract expiring at time $0 < T < \mathbb{T}$, denoted as F_t , evolves according to the following general diffusion,

$$\frac{dF_t}{F_t} = \mu_t dt + \sigma_t dW_t, \quad (1)$$

where W_t is a Wiener process. The drift μ_t and volatility σ_t can change across time according to the filtration \mathbb{F}_t .

The constraint imposed on the futures price dynamics is that the stochastic process is a semimartingale without jumps in prices.⁸ It is worth noting that the only restriction imposed on the volatility dynamics is that σ_t is a strictly positive (càdlàg) stochastic process, so volatility can exhibit jumps.

According to these price dynamics, the total variation of logarithmic futures price changes from $t = 0$ to T is given by IVAR, defined as,

$$IVAR_{0,T} = \int_0^T \sigma_t^2 dt. \quad (2)$$

Although total return variation is the forecast target of this paper, we also utilize the concept of CIVAR, that is variance accumulated only when the underlying asset (F_t in our case) lies between two “barrier” price levels B_1 and B_2 .

⁶With the introduction of new securities linked to commodities, as well as with regulatory changes related to trading, the popularity of investing in commodity products has had a significant impact of their price dynamics. See, for example, Henderson, Pearson, and Wang (2014) and Singleton (2014), among others.

⁷The same result is also reported in Andersen and Bondarenko (2007) for the case of the S&P 500, and Taylor et al. (2010), for the case of individual stocks.

⁸Price jumps are excluded from this representation because the option-implied expectations, discussed later in the paper, will be biased when prices are subject to discontinuous movements. Jiang and Tian (2005) and Carr and Wu (2009) note that the bias will not be sizeable for small or moderate jumps, although large jumps could have a significant impact, as argued by Carr, Lee, and Wu (2012).

Defining the indicator function I_t that takes the value of 1 if $B_1 \leq F_t \leq B_2$ and 0 otherwise, CIVAR is given by the following expression,

$$CIVAR_{0,T} = \int_0^T \sigma_t^2 I_t(B_1, B_2) dt. \quad (3)$$

Obviously, when the corridor defined by B_1 and B_2 is sufficiently wide to contain all levels that the futures price can reach with positive probability under \mathbb{P} , CIVAR will be equal to IVAR. In other words, IVAR can be seen as a special case of CIVAR, since for $B_1 = 0$ and $B_2 = \infty$ the definitions of the two measures coincide.

2.2 | Volatility expectations from option prices

Option markets may be informative about future volatility, since observed prices can be utilized to extract forward-looking expectations of the aforementioned volatility measures. In particular, suppose European plain vanilla options, written on an underlying futures contract F_t and expiring at time $t = T$, trade for a continuum of strike prices K . As shown in Carr and Madan (1998), Demeterfi et al. (1999), and Britten-Jones and Neuberger (2000), ex ante risk-neutral expectations of the future IVAR can be obtained by calculating the value of a static position in a portfolio of European options. Specifically, the expected IVAR from time $t = 0$ to time T , under the risk-neutral measure \mathbb{Q} , can be calculated from:

$$MFIV_{0,T} = E^{\mathbb{Q}}[IVAR_{0,T}] = E^{\mathbb{Q}}\left[\int_0^T \sigma_t^2 dt\right] = 2e^{rT} \int_0^\infty \frac{M_{0,T}(K)}{K^2} dK, \quad (4)$$

where $M_{0,T}(K)$ is the price of a European out-of-the money option (i.e., either put or call), with strike price K and maturity T . Since this expectation does not depend on a particular option pricing model (such as the Black–Scholes model, e.g.), it is referred to as model free.

Similarly, as shown in Carr and Madan (1998) and Andersen and Bondarenko (2007), CIV, that is the risk-neutral expectation of future CIVAR, can be obtained by calculating the value of a static position in European options with strikes ranging from B_1 to B_2 ,

$$CIV_{0,T} = E^{\mathbb{Q}}[CIVAR_{0,T}] = E^{\mathbb{Q}}\left[\int_0^T \sigma_t^2 I_t(B_1, B_2) dt\right] = 2e^{rT} \int_{B_1}^{B_2} \frac{M_{0,T}(K)}{K^2} dK. \quad (5)$$

2.3 | CBOE's crude-oil volatility index (OVX)

The last option-based measure we consider is the Crude Oil ETF Volatility Index (OVX), also known as the “Oil VIX.” The OVX, which is produced and disseminated by the CBOE, intends to measure the market's (risk-neutral) expectation of crude-oil price volatility over the next month. It is defined as the square root of MFIV, given in Equation (4). The data underlying the OVX computation are options written on the USO, an ETF that is designed to track the price of West Texas Intermediate light sweet crude-oil.

For the construction of the OVX the CBOE adopts the same methodology as the one employed for the popular S&P 500 VIX index. Notably, CBOE applies a liquidity criterion to determine the range of option contracts included in the calculation of the index. In particular, moving from high (low) strike, out-of-the-money, put (call) options toward those with lower (higher) strikes, once two contracts with consecutive strike prices have zero bid prices a cut-off point is applied and no further contracts are considered. Therefore, both the VIX and the OVX are, in reality, CIV measures, with a corridor width determined by market liquidity.

2.4 | Realized variance measures

While our forecast target, that is IVAR, is inherently latent, accurate ex post IVAR estimates can be obtained using high-frequency price observations. In particular, Barndorff-Nielsen and Shephard (2002), Meddahi (2002), and Andersen,

Bollerslev, Diebold, and Labys (2003) show that summing squared intraday returns leads to an estimator which converges in probability to IVAR and is referred to as *realized variance* (RV). To calculate RV, suppose on day t there are $M + 1$ equally spaced intraday price observations at times t_i , $i = 0, \dots, M$. We will also assume the interval between the observations is $1/M$, that is the length of a day is standardized to unity. If the log price at time t_i is p_{t_i} , then the intraday return between times t_{i-1} and t_i is $r_{t_i} = p_{t_i} - p_{t_{i-1}}$. It is then straightforward to calculate RV on day t as,

$$RV_t = \sum_{i=1}^M r_{t_i}^2.$$

Theoretically, RV becomes more accurate as M increases, that is as more intraday prices are observed over shorter and shorter intervals. However, if prices are observed over very short intervals, RV will be contaminated by microstructure noise, which causes an upward bias (Andersen, Bollerslev, Diebold, & Labys, 2000). A common remedy is to use prices observed over a relatively coarse set of intraday times so that the effects of microstructure noise are mitigated. Typically, prices recorded over 5-min intervals are used, despite transactions occurring at a much higher frequency. Although using a coarse intraday sampling interval solves the microstructure noise problem, it results in information being discarded. To recover some of this information, ensuring our RVs are estimated with as much accuracy as possible, and to continue avoiding microstructure noise by using a coarse sampling interval, we use subsampled RVs (Zhang, Mykland, & Ait-Sahalia, 2005) which are calculated as follows:

$$\widetilde{RV}_t = \frac{1}{\Delta} \left(\frac{M/\Delta}{(M/\Delta) - 1} \sum_{j=2}^{\Delta-1} \sum_{i=1}^{M/\Delta-1} \bar{r}_{\Delta, t_{\Delta(i-1)+j}}^2 + \sum_{k=1}^{M/\Delta} \bar{r}_{\Delta, t_{\Delta(k-1)+1}}^2 \right), \quad (6)$$

where $\bar{r}_{\Delta, t_i} = \sum_{j=1}^{\Delta} r_{t_{i+j-1}} = p_{t_{i-1+\Delta}} - p_{t_{i-1}}$ is the Δ -period intraday return between times t_{i-1} and $t_{i-1+\Delta}$, and M and Δ are chosen such that M/Δ is an integer.

Later in the paper we will require an accurate proxy of CIVAR. For this purpose, we introduce the concept of CRV. With subsampling, CRV is calculated as follows:

$$\widetilde{CRV}(B_1, B_2)_t = \frac{1}{\Delta} \left(\frac{M/\Delta}{(M/\Delta) - 1} \sum_{i=2}^{M-\Delta-1} (\tilde{r}_{\Delta, t_{\Delta(i-1)+j}}(B_1, B_2))^2 + \sum_{k=1}^{M/\Delta} (\tilde{r}_{\Delta, t_{\Delta(k-1)+1}}(B_1, B_2))^2 \right), \quad (7)$$

where $\tilde{r}_{\Delta, t_{\Delta(i-1)+j}}(B_1, B_2) = \max(\min(p_{t_{i-1+\Delta}}, B_2), B_1) - \max(\min(p_{t_{i-1}}, B_2), B_1)$ is the Δ -period intraday corridor return between times t_{i-1} and $t_{i-1+\Delta}$. Therefore, if either $p_{t_{i-1+\Delta}}$ or $p_{t_{i-1}}$ are above B_2 then they are set equal to B_2 , whereas if either are below B_1 they are set equal to B_1 . Again, M and Δ are chosen such that M/Δ is an integer.

3 | METHODOLOGY

3.1 | Data and sample construction

Our main data set consists of options and high-frequency prices for the WTI Light Sweet Crude Oil futures, which currently trade at CME, the world's most liquid commodity derivatives market. Our options and futures data sets start in January 1996 and end in April 2016. Our secondary data set contains options and high-frequency prices of the USO ETF from May 2007 to December 2016.

3.1.1 | Crude-oil option data

Option contracts are written on futures contracts for physical delivery of light sweet crude-oil. In particular, the underlying asset is the futures contract whose delivery date is 3 business days after the expiration of the option. These

TABLE 1 Summary statistics of the option data sets

	Period	Obs	All options			Calls (Puts)		
			Mean	Min	Max	Mean	Min	Max
Full sample	1996–2016	280	61.1	7	213	33.8 (27.3)	3 (3)	184 (94)
Post-Fin sample	2007–2016	108	112.5	54	213	63.2 (49.3)	29 (16)	184 (94)
USO sample	2007–2016	118	38.7	14	123	20.6 (18.1)	7 (7)	64 (63)

Note: We report the number of unique option cross-sections (Obs) used to construct our nonoverlapping, option-implied, monthly volatility forecasts, as well as the mean, minimum, and maximum number of option contracts used in our estimations. In the full and postfinancialization (Post-Fin) samples the underlying asset is the WTI Light Sweet Crude Oil futures contract. In the third sample the underlying asset is the United States Oil (USO) Fund.

contracts have American-style exercise and are settled in cash. Our option data consist of daily settlement prices, which are recorded at 14:30 ET each day.

We make several adjustments to the raw data before we proceed with the estimation of the option-implied volatility (OIV) measures. As the latter require price data on European options rather than American ones for their calculation, we attempt to alleviate this problem in two ways. First, we eliminate all in-the-money options and only keep out-of-the-money options, for which the early exercise premium is significantly lower and, for the case of deep-out-of-the-money options, almost quantitatively negligible. Second, we estimate the early exercise premium of each option using the Barone-Adesi and Whaley (1987) American option pricing formula and subsequently calculate the price of their European-style counterparts.

Finally, to guard against recording errors and other market microstructure effects, we eliminate options with a price less than \$0.01 and filter all call/put prices that violate standard arbitrage bounds. An overview of our final option data sample is provided in Table 1. It is noteworthy that the number of traded option contracts has increased substantially over the last few years.

3.1.2 | High-frequency crude-oil futures data

Our high-frequency data comprise of transaction prices recorded at 30-s intervals. Until June 2006, futures were traded between 09:00 and 14:30 ET using an open outcry system in a trading pit, resulting in 661 price observations per day. Subsequently, they have been traded between 18:00 and 17:00 ET the following day on the electronic GLOBEX platform, resulting in 2,761 price observations per day.

To construct RV we use 5-min intraday returns. This frequency is commonly used in the empirical literature, for example, see Andersen, Bollerslev, Diebold, and Labys (2001), Andersen et al. (2003), and Liu, Patton, and Sheppard (2015) among others, as it is deemed to provide an appropriate trade-off between the objective of incorporating as much information as possible from intraday prices and the necessity to avoid contamination from microstructure noise. Hence, in applying our subsampled RV estimator in Equation (6) we set $M = 660$ or 2,760, depending on the data are from before June 2006 or not, and $\Delta = 10$.

3.1.3 | USO option and high-frequency return data

Options on the USO ETF trade 08:30–15:00 CT on the CBOE and correspond to the delivery of 100 shares of the USO ETF, where the delivery date is three business days following the exercise of the option. These contracts have American-style exercise. Our USO ETF option prices consist of end-of-day bid and ask quotes. Before proceeding with the estimation of our OIV measures we estimate the early exercise premium and apply the same data filter procedures as with our crude-oil option data set.

In analogy to our crude-oil futures data, our high-frequency USO ETF data consists of transaction prices recorded at 30-s intervals. The USO ETF trades on the NYSE Arca between 09:30 and 16:00 ET, resulting in 781 price observations per day. Again, we use 5-min returns to construct RV so that $M = 780$ and $\Delta = 10$ when applying our subsampled RV estimator in Equation (6).

3.2 | Forecast samples

We examine the forecast performance of alternative option-implied measures using three distinct samples. The first, and most important one, corresponds to the full history of our crude-oil option and futures data (January 1996–April 2016). We refer to it as our *full sample*.

Motivated by the literature that finds that the behavior of commodities has experienced significant changes after their “financialization” period, we conduct a separate analysis using data that belong exclusively to that era. This sample, which we refer to as *Post-Fin*, starts in May 2007, when the OVX, our natural benchmark, was first reported by the CBOE, and ends in April 2016.

Finally, we examine option data and high-frequency returns corresponding to the USO ETF. Since the latter underlies CBOE's OVX computation we believe that the analysis of USO volatility forecasts can provide interesting insights. This sample, which we refer to as *USO sample*, spans the period between May 2007 and December 2016. In the interests of clarity, we present the empirical results of the USO sample separately in Section 5.

Our empirical study focuses on monthly variance forecasts. Along these lines, we study options that have approximately 22 trading days to expiration.⁹ Throughout the sample, the day that we collect option prices is never before the maturity date of the previous option chain we studied, that is all our option-based forecasts are nonoverlapping.

3.3 | Construction of implied volatility measures

The computation of the MFIV and CIV measures requires the existence of options trading for a continuum of strike prices, an assumption which is of course not satisfied in practice. To address this, we first estimate a risk-neutral distribution using the prices of observed options for each relevant date, which enables us to subsequently generate option prices for arbitrary strike prices. Our preferred risk-neutral distribution is the flexible Generalized Beta Distribution of the second kind (GB2) which, as discussed in Taylor (2005), has a number of appealing properties.¹⁰

The calculation of the CIV measures also requires a selection of the relevant corridor width, that is the barrier price levels B_1 and B_2 . We consider four CIV measures¹¹ in total, with the barriers determined by evaluating the quantile function of the risk-neutral distribution F_Q . Specifically, defining $B_1 = F_Q^{-1}(p)$ and $B_2 = F_Q^{-1}(1 - p)$, the CIV1 to CIV4 measures are obtained by first setting $p = .45, .35, .25, .10$, respectively, and subsequently evaluating Equation (5).

Figure 1 plots several CIV measures together with the MFIV estimates corresponding to crude-oil over the 1996–2016 period. As expected, all measures exhibit a strong degree of covariation, although narrow CIV measures appear more stable than their wide corridor counterparts. As shown in Table 2, all CIV measures are highly correlated with both MFIV and ATMIV. The autocorrelation patterns for 1, 6, and 12 lags also reveal that CIV measures become less persistent as the width of the corridor widens. As expected, the unconditional moments of the CIV measures approach those of the full MFIV as more options are included in the calculations, while for the case of CIV4 the two measures have almost identical properties.

3.4 | Variance forecasts using the HAR model

There is a considerable literature demonstrating the superiority of forecasts generated from time-series models of RV over those obtained by a random walk model or specification using daily return data (see, e.g., Bollerslev, Patton, &

⁹If for a given month options with an expiration horizon of 22 trading days are not available, we shorten the target horizon by one or (if needed) 2 days and scale the resulting OIV such that its expiration horizon corresponds to 22 days. For example, if the expiration horizon of the OIV was N trading days, we scale the OIV by multiplying it by $22/N$.

¹⁰The GB2 distribution, first proposed by Bookstaber and McDonald (1987), allows for general levels of skewness and kurtosis, while European option prices can be computed in closed-form. We elaborate on our choice and examine its robustness in Section 6.1.1.

¹¹Supplementary analysis for alternative corridor measures is presented in Section 6.1.2.

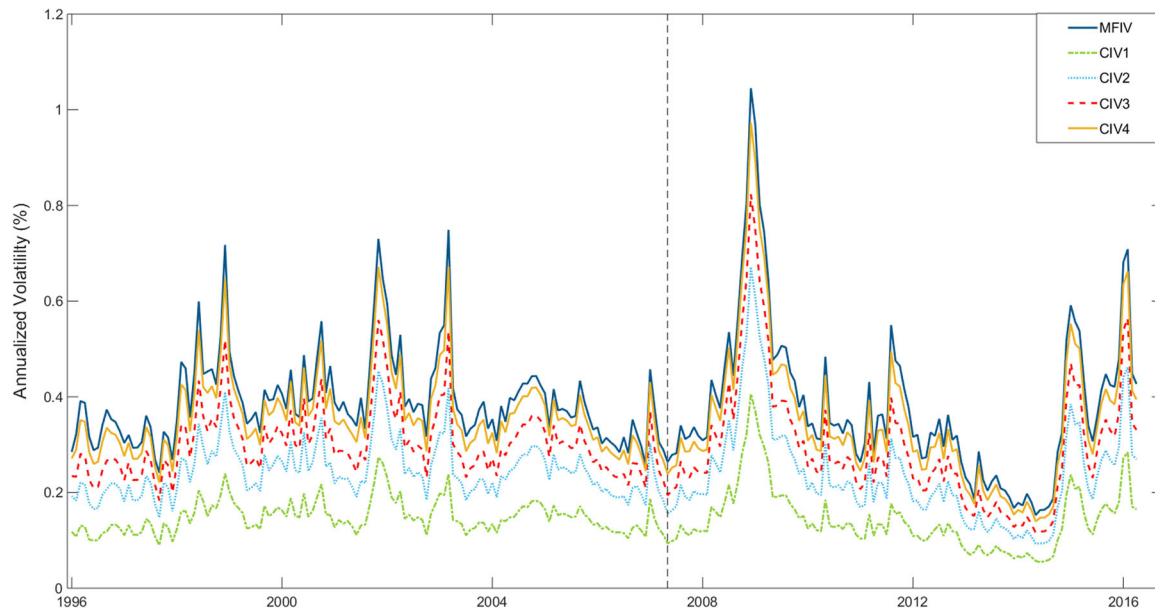


FIGURE 1 Time-series plots of different corridor implied volatility measures for the 1996–2016 period. All expectations represent annualized standard deviations corresponding to a monthly horizon. The dashed vertical line indicates the beginning of the Post-Fin sample. CIV, corridor implied volatility; MFIV, model-free implied volatility [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 2 Descriptive statistics of option-implied volatility measures

	Mean	SD	Skew	Kurt	ρ_1	ρ_6	ρ_{12}	ρ_{MF}	ρ_{BS}
CIV1	0.146	0.051	1.531	7.514	0.856	0.33	0.125	0.981	0.987
CIV2	0.241	0.084	1.564	7.65	0.85	0.319	0.116	0.988	0.993
CIV3	0.298	0.102	1.572	7.68	0.845	0.314	0.114	0.993	0.995
CIV4	0.356	0.12	1.583	7.699	0.837	0.306	0.112	0.999	0.995
MFIV	0.386	0.129	1.591	7.711	0.829	0.301	0.112	1	0.993
ATMIV	0.365	0.124	1.549	7.509	0.836	0.306	0.115	0.993	1

Note: The first four columns show the unconditional sample moments for each measure. Sample autocorrelations for 1, 6, and 12 lags are denoted as ρ_1 , ρ_6 , and ρ_{12} , respectively. The last two columns show the correlation of each measure with the model-free (ρ_{MF}) and Black–Scholes (ρ_{BS}) implied volatilities.

Abbreviations: ATMIV, at-the-money implied volatility; CIV, corridor implied volatility; MFIV, model-free implied volatility.

Quaedvlieg, 2016; Bollerslev, Hood, Huss, & Pedersen, 2018; Corsi, 2009; and Patton & Sheppard, 2015). Along these lines, we use the HAR model of Corsi (2009) to generate RV-based forecasts,

$$RV_{t+1,t+22} = \beta_0 + \beta_1 RV_{t,t} + \beta_2 RV_{t-4,t} + \beta_3 RV_{t-21,t} + \epsilon_{t+22}, \quad (8)$$

where, $RV_{u,v} = \sum_{s=u}^v \widetilde{RV}_s$, that is our realized variances are calculated using the subsampled RV estimates described in Equation (6). While various extensions of the HAR model could be considered, our study relies exclusively on the baseline version. This is because, as shown in the comprehensive studies of Sévi (2014) and Prokopczuk et al. (2016), sophisticated HAR extensions do not outperform the simple HAR benchmark for the case of crude-oil.¹² We estimate the HAR model parameters using a rolling window of 60 monthly observations.

¹²Experimentation with alternative HAR models corroborated this finding for our sample. Since our results replicate those of Sévi (2014) and Prokopczuk et al. (2016) we do not report them to conserve space.

TABLE 3 Mincer–Zarnowitz regression results

Forecast	Panel A: Full sample			Panel B: Post-Fin sample		
	β_0	β_1	R^2	β_0	β_1	R^2
HAR	7.499E–05** (3.749)	0.785** (14.197)	0.700	4.159E–02* (2.055)	0.633** (4.549)	0.368
MFIV	–9.079E–06 (<0.001)	0.776** (11.458)	0.747	3.099E–03 (0.308)	0.877** (9.898)	0.720
ATMIV	–4.743E–06 (<0.001)	0.857** (12.049)	0.757	5.837E–03 (0.607)	0.964** (9.801)	0.728
CIV1	2.746E–06 (0.127)	5.141** (12.278)	0.766	1.197E–02 (1.272)	5.576** (9.695)	0.737
CIV2	1.351E–06 (0.064)	1.907** (12.545)	0.766	9.836E–03 (1.031)	2.109** (9.722)	0.734
CIV3	–1.839E–06 (<0.001)	1.268** (12.410)	0.764	8.059E–03 (0.836)	1.411** (9.783)	0.731
CIV4	–6.678E–06 (<0.001)	0.903** (11.942)	0.756	5.187E–03 (0.527)	1.013** (9.871)	0.725
OVX				–8.312E–04 (<0.001)	0.886** (9.525)	0.725

Note: This table reports parameter estimates and values of R^2 for the Mincer–Zarnowitz regressions in Equation (9). Panel A presents results for the full sample. Panel B presents results for the Post-Fin sample, which includes the OVX forecasts. All t statistics, reported in parentheses, are based on Newey–West standard errors. The superscripts * and ** indicate significance at the 5% and 1% levels, respectively.

Abbreviations: ATMIV, at-the-money implied volatility; CIV, corridor implied volatility; HAR, heterogeneous autoregressive model; MFIV, model-free implied volatility; OVX, Crude Oil Volatility Index.

4 | EMPIRICAL RESULTS

We examine the forecast accuracy of competing forecasts using several techniques. First, we evaluate the information content of each of our forecasts using Mincer–Zarnowitz regressions. Second, we make comparisons between the information content of our forecasts using encompassing regressions. Third, we analyse the prediction errors of our forecasts using statistical loss functions. Lastly, we assess the economic value of our forecasts by implementing a volatility timing exercise.

4.1 | Mincer–Zarnowitz regressions

Our first evaluation procedure assesses the information content of our HAR-based forecasts and the OIVs. This is done by running Mincer–Zarnowitz regressions (Mincer & Zarnowitz, 1969), whereby we regress our variance target, the RV calculated over each out-of-sample monthly forecast horizon, against the competing forecasts. More precisely, for each forecast i , we run the following regression,

$$RV_{t+1,t+22} = \beta_0 + \beta_1 f_{t,t+22}^i + \epsilon_{t+22}, \quad (9)$$

where $f_{t,t+22}^i$ is the forecast from model i using information available up until day t for the variance between days $t + 1$ and $t + 22$. The information content of each forecast is measured by the R^2 of this regression.

In Table 3 we present results for the Mincer–Zarnowitz regressions. In Panel A we report results for the full sample, while in Panel B we summarize results for the Post-Fin sample. The values of R^2 suggest the following. First, OIVs have markedly higher information content compared to forecasts based solely on past RVs. Specifically, HAR forecasts have the lowest R^2 values, while the R^2 for the OIV forecasts are larger by 5–7 percentage points for the full sample and

35–37 percentage points for the Post-Fin sample. Second, CIV1 appears to have the highest information content, since CIV1 forecasts result in the largest R^2 s in both the full and Post-Fin samples. Third, the information content of our CIVs appears to be inversely proportional to the width of the corridor; as we move from CIV4 to CIV1, the R^2 s progressively increase. Lastly, the OVX has a low information content with its R^2 being approximately on par with CIV4, the worse performing CIV, and lower than ATMIV.

It should also be noted that the parameter estimates corresponding to Equation (9) differ substantially between the forecasts. This is as expected and a consequence of the differing levels of bias in the forecasts. If a forecast is unbiased, then we would expect $\beta_0 = 0$ and $\beta_1 = 1$. While all of the β_0 parameters are close to zero, there are large differences between the values of β_1 . Overall, the MFIV, ATMIV, and OVX are upwardly biased, consistent with the presence of a variance risk-premium, while CIV1, CIV2, and CIV3 are downwardly biased, reflecting the fact that they provide risk-neutral expectations for CIVAR rather than IVAR. The observed biases are consistent with the mean values of the OIVs reported in Table 2.

4.2 | Encompassing regressions

Our second evaluation procedure assesses the relative performance of the alternative forecasts by running encompassing regressions. We make comparisons between two forecasts by running the following bivariate regressions,

$$RV_{t+1,t+22} = \beta_0 + \beta_1 f_{t,t+22}^i + \beta_2 f_{t,t+22}^j + \epsilon_{t+22}. \quad (10)$$

We can determine whether the forecast from one model encompasses the other by examining the significance of the individual regression parameters. If the information contained in the forecast from model i is subsumed by the information in the forecast from model j , then we expect β_1 to be insignificant and β_2 to be significant. If this occurs, then we say that the forecast from model i encompasses the forecast from model j and vice versa.

We make the following comparisons using our encompassing regressions: (a) we compare the information content of the HAR model forecasts, which are based on RV alone, against our OIVs to examine whether the forward-looking information in our OIVs is useful vis-à-vis the backward-looking information contained in RV; and (b) we compare the information content of our alternative OIVs to examine which of our OIVs contains the most useful forecasting information.

The results of the encompassing regressions are summarized in Table 4. Results for comparisons between the OIVs and the HAR forecasts are summarized in Panels A and B for the full and Post-Fin samples, respectively. Panel A shows that CIV1, CIV2, and CIV3 encompass the HAR forecasts, at the 1% level. In the remaining encompassing regressions, although the parameters on MFIV, ATMIV, and CIV4 are significant at the 1% level, the parameters on HAR are also significant; albeit only at the 10% level for the encompassing regressions involving ATMIV and CIV4. Thus, although they do not encompass the HAR forecasts, there appears to be incremental information in the MFIV, ATMIV, and CIV4 OIVs which is not captured in the historical time-series of RV. However, in Panel B, all OIVs, including the OVX, encompass the HAR forecasts, at the 1% level.

Results for comparisons between the OIVs are summarized in Panels C and D for the full and Post-Fin samples, respectively. Focusing on the full sample results, it can be seen that MFIV is encompassed by all the other OIVs at the 5% level. Comparisons of ATMIV and CIVs show that ATMIV is encompassed by CIV1, CIV2, and CIV3, at the 5% level. The encompassing regressions involving CIV1, CIV2, CIV3, and CIV4 show that CIV4 is encompassed by CIV1, CIV2, and CIV3, at the 5% level. Therefore, the results show that narrowing the corridor of our CIVs leads to improved information content. These findings are consistent with the arguments of Andersen and Bondarenko (2007) that the presence of options with extreme strikes, which are thinly traded and typically very sensitive to volatility risk premia fluctuations, contaminates the volatility information contained in the MFIV and wide CIV measures.¹³

The results in Panel D are weaker than those in Panel C and likely reflect the smaller size of the Post-Fin sample. Nonetheless, although often significant at the 10% level only, the results show that CIV1 encompasses all other OIVs.

¹³We investigate the role of the variance risk premium in determining the forecast accuracy of our OIVs in more depth in Section 4.6.

TABLE 4 Encompassing regression results

Forecast 1	Forecast 2	Parameters				Parameters			
		β_0	β_1	β_2	adj R^2	β_0	β_1	β_2	adj R^2
		Panel A: HAR vs. option-implied forecasts (Full sample)				Panel B: HAR vs. option-implied forecasts (Post-Fin sample)			
HAR	MFIV	5.360E−06 (0.279)	0.284** (2.141)	0.530*** (4.121)	0.763	4.867E−03 (0.492)	−0.390 (<0.001)	1.210*** (4.595)	0.756
HAR	ATMIV	6.529E−06 (0.342)	0.247* (1.813)	0.621*** (4.205)	0.768	8.965E−03 (0.981)	−0.413 (<0.001)	1.350*** (4.589)	0.769
HAR	CIV1	1.045E−05 (0.541)	0.218 (1.554)	3.898*** (4.138)	0.774	1.807E−02** (2.047)	−0.434 (<0.001)	7.909*** (4.815)	0.783
HAR	CIV2	9.438E−06 (0.490)	0.214 (1.522)	1.453*** (4.180)	0.774	1.487E−02* (1.677)	−0.428 (<0.001)	2.982*** (4.774)	0.778
HAR	CIV3	7.475E−06 (0.387)	0.223 (1.611)	0.953*** (4.210)	0.772	1.224E−02 (1.365)	−0.423 (<0.001)	1.989*** (4.755)	0.774
HAR	CIV4	5.372E−06 (0.278)	0.253* (1.875)	0.648*** (4.199)	0.767	7.958E−03 (0.849)	−0.407 (<0.001)	1.413*** (4.677)	0.765
HAR	OVX					−1.157E−04 (<0.001)	−0.364 (<0.001)	1.195*** (4.598)	0.757
		Panel C: Option-implied forecasts (Full sample)				Panel D: Option-implied forecasts (Post-Fin sample)			
MFIV	ATMIV	6.383E−07 (0.031)	−0.549 (<0.001)	1.457** (2.099)	0.758	1.267E−02 (0.963)	−1.750 (<0.001)	2.875 (1.426)	0.728
MFIV	CIV1	2.848E−06 (0.145)	−0.004 (<0.001)	5.166** (2.546)	0.765	2.317E−02 (1.593)	−0.945 (<0.001)	11.491** (2.032)	0.738
MFIV	CIV2	7.549E−06 (0.393)	−0.279 (<0.001)	2.579** (2.464)	0.766	2.094E−02 (1.506)	−1.229 (<0.001)	5.028* (1.875)	0.735
MFIV	CIV3	8.505E−06 (0.441)	−0.647 (<0.001)	2.309** (2.397)	0.766	2.055E−02 (1.495)	−1.856 (<0.001)	4.368* (1.848)	0.735
MFIV	CIV4	8.481E−06 (0.435)	−2.501 (<0.001)	3.794** (2.265)	0.767	1.967E−02 (1.466)	−5.038 (<0.001)	6.806* (1.798)	0.735
ATMIV	CIV1	2.314E−06 (0.115)	0.031 (0.087)	4.956** (2.019)	0.765	2.623E−02* (1.928)	−2.000 (<0.001)	17.055** (2.040)	0.740
ATMIV	CIV2	5.580E−06 (0.289)	−0.394 (<0.001)	2.775** (1.995)	0.766	2.222E−02* (1.969)	−2.681 (<0.001)	7.948* (1.773)	0.735
ATMIV	CIV3	2.377E−06 (0.120)	−0.734 (<0.001)	2.348** (1.980)	0.764	1.696E−02** (2.002)	−3.449 (<0.001)	6.447 (1.535)	0.731
ATMIV	CIV4	−5.149E−06 (<0.001)	0.726 (0.874)	0.138 (0.159)	0.756	7.344E−03 (0.640)	2.648 (0.708)	−1.773 (<0.001)	0.724
CIV1	CIV2	2.049E−06 (0.097)	3.130 (0.630)	0.747 (0.430)	0.765	2.647E−02* (1.680)	39.351* (1.704)	−12.809 (<0.001)	0.742
CIV1	CIV3	1.901E−06 (0.093)	4.436 (1.348)	0.175 (0.239)	0.765	2.481E−02 (1.632)	21.911* (1.709)	−4.153 (<0.001)	0.739
CIV1	CIV4	2.399E−06 (0.120)	5.018** (2.123)	0.022 (0.061)	0.765	2.386E−02 (1.604)	14.032* (1.865)	−1.552 (<0.001)	0.738
CIV2	CIV3	6.274E−06	4.282	1.582	0.766	2.109E−02	14.244	−8.132	0.734

TABLE 4 (Continued)

Forecast 1	Forecast 2	Parameters				Parameters			
		β_0	β_1	β_2	adj R^2	β_0	β_1	β_2	adj R^2
		(0.325)	(1.332)	(<0.001)		(1.493)	(1.362)	(<0.001)	
CIV2	CIV4	7.217E-06 (0.375)	2.868** (2.045)	−0.460 (<0.001)	0.766	2.130E−02 (1.509)	6.628* (1.661)	−2.185 (<0.001)	0.735
CIV3	CIV4	8.374E-06 (0.434)	3.090** (1.988)	−1.305 (<0.001)	0.766	2.110E−02 (1.506)	6.925 (1.641)	−3.976 (<0.001)	0.736
OVX	MFIV					−9.208E−04 (<0.001)	0.911 (0.768)	−0.025 (<0.001)	0.719
OVX	ATMIV					3.244E−03 (0.286)	0.308 (0.459)	0.631 (0.866)	0.723
OVX	CIV1					1.051E−02 (0.961)	0.089 (0.200)	5.025* (1.676)	0.732
OVX	CIV2					7.877E−03 (0.735)	0.144 (0.286)	1.772 (1.412)	0.728
OVX	CIV3					6.003E−03 (0.565)	0.182 (0.322)	1.125 (1.212)	0.726
OVX	CIV4					1.999E−03 (0.182)	0.433 (0.555)	0.521 (0.587)	0.721

Note: The table reports parameter estimates and values of the adjusted R^2 for the encompassing regressions in Equation (10). Estimates reported in the columns labeled β_1 are for the forecast listed in the column-labeled *Forecast 1*. Estimates reported in the columns labeled β_2 are for the forecast listed in the column-labeled *Forecast 2*. Panel A presents results for encompassing regressions involving comparisons between the HAR forecasts and each of the option-implied variances for the full sample. Panel B presents results for encompassing regressions involving comparisons between the HAR forecasts and each of the option-implied variances for the Post-Fin sample, which includes the OVX forecasts. Panel C presents results for encompassing regressions involving comparisons between the option-implied variances for the full sample. Panel D presents results for encompassing regressions involving comparisons between the option-implied variances for the Post-Fin sample. All t statistics, reported in parentheses, are based on Newey–West standard errors. The superscripts *, ** and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

Abbreviations: ATMIV, at-the-money implied volatility; CIV, corridor implied volatility; HAR, heterogeneous autoregressive model; MFIV, model-free implied volatility; OVX, Crude Oil Volatility Index.

Notably, the OVX does not encompass any of the other OIVs, consistent with our Mincer–Zarnowitz regression results in Section 4.1 which showed the information content of the OVX to be inferior relative to our other OIVs.

In summary, the results from the encompassing regressions support those from the Mincer–Zarnowitz regressions in Section 4.1. The information content of the OIVs appears to be superior relative to those based on RV alone, that is the HAR forecasts. Furthermore, of the OIVs, the CIV1 appears to be informationally superior, since it encompasses our other OIVs in both the full and Post-Fin samples. In contrast, the information content of the OVX is seemingly limited as it does not encompass any other OIV while its respective parameter is always less significant when combined with any CIV.

4.3 | Generating out-of-sample forecasts

The presence of a variance risk premia as well as the fact that corridor implied variances provide risk-neutral forecasts of expected CIVAR, rather than IVAR, generally causes the raw OIV measures to provide biased forecasts of future return variation. Empirically, this can be observed by the parameters of the Mincer–Zarnowitz regressions presented in Table 3. Along these lines, out-of-sample variance forecasts of models utilizing option-implied information are obtained through the following *univariate models*,

$$RV_{t+1,t+22} = \beta_0 + \beta_1 OIV_{t,t+22} + \epsilon_{t+22}, \quad (11)$$

where $OIV_{t,t+22}$ is an option-implied measure (i.e., either MFIV, ATMBS, CIV1, CIV2, CIV3, CIV4, or, where applicable, OVX) calculated on day t and corresponding to a forecast for the variance between days $t + 1$ and $t + 22$. We use BC-MFIV to denote the bias-corrected MFIV forecast, BC-ATMIV to denote the bias-corrected ATMIV forecast, and so on.

Moreover, by comparing the adjusted R^2 s of the encompassing regressions in Table 4 to the R^2 s of the Mincer–Zarnowitz regressions in Table 3, we can see that a linear combination of RVs and OIVs leads to a higher information content than can be attained with any individual forecast. This motivates evaluating the accuracy of forecasts generated by *augmented models* that utilize information from both option prices and historical return data. Consistent with the prior literature (e.g., Busch, Christensen, & Nielsen, 2011), we use the following augmented HAR model to generate forecasts,

$$RV_{t+1,t+22} = \beta_0 + \beta_1 RV_{t,t} + \beta_2 RV_{t-4,t} + \beta_3 RV_{t-21,t} + \beta_4 OIV_{t,t+22} + \epsilon_{t+22}. \quad (12)$$

We use HAR-MFIV to denote the forecasts from an HAR model augmented with MFIV, HAR-ATMIV the forecasts from an HAR model augmented with ATMIV and so on.

In total, we examine out-of-sample forecasts generated by 13 different models for the *full sample* and 15 for the *Post-Fin* and *USO* samples, since the latter two include the BC-OVX and HAR-OVX forecasts. The parameters for all models defined in Equations (8), (11), and (12) are estimated using a rolling window of 60 monthly observations.

4.4 | Forecast evaluation using statistical criteria

Although Mincer–Zarnowitz and encompassing regressions provide insights into the information content of our forecasts, they do not provide much information about their precision. From the perspective of economic agents, quantifying forecast accuracy is of paramount importance. Thus, we now turn to assessing prediction errors by means of statistical loss functions.

4.4.1 | Statistical loss functions

To evaluate the accuracy of our forecasts, we use a symmetric, the mean squared error (MSE), and an asymmetric, the quasi-likelihood (QLIKE), loss function. A lower MSE and/or QLIKE corresponds to smaller prediction errors. These loss functions were chosen because they are commonly employed in the volatility forecasting literature and, as shown in Patton (2011), they are robust to measurement error in the IVAR proxy. More precisely, using the MSE and QLIKE loss functions ensures that the ranking of two forecasts in terms of expected loss is preserved even when the true IVAR is replaced by a conditionally unbiased, but imperfect, proxy.

Denoting f_t^i as the time t variance forecast generated from a reference model i , and $RV_{t,T}$ as the corresponding realization of the target quantity, MSE and QLIKE losses from n total forecasts are defined as:

$$L_{MSE,i} = \frac{1}{n} \sum_{t=1}^n (f_t^i - RV_{t,T})^2, \quad L_{QLIKE,i} = \frac{1}{n} \sum_{t=1}^n \ln(f_t^i) - \frac{f_t^i}{RV_{t,T}}$$

To test for significant differences between the MSE and QLIKE of competing forecasts we use Diebold–Mariano tests (Diebold & Mariano, 1995) with Newey–West (Newey & West, 1987) standard errors.

4.4.2 | Forecast evaluation results

The out-of-sample MSEs and QLIKES of our competing forecasts are presented in Table 5. Panel A reports results for the full sample while Panel B summarizes results for the Post-Fin sample. In each panel we also report the results of Diebold–Mariano tests in which the HAR forecasts are used as a benchmark. Therefore, any significance reported indicates the MSE or QLIKE is significantly different to that of the HAR forecasts. In Panel B, to assess the accuracy of

TABLE 5 Out-of-sample MSE and QLIKE results

Forecast	Panel A: Full sample		Panel B: Post-Fin sample	
	MSE	QLIKE	MSE	QLIKE
HAR	2.721	−1.634	9.760	−1.504
BC-MFIV	2.562	−1.637	4.916*****	−1.504
BC-ATMIV	2.427	−1.639	4.626*****	−1.519*****
BC-CIV1	2.229	−1.636	4.279*****	−1.528*****
BC-CIV2	2.268	−1.638	4.413*****	−1.527*****
BC-CIV3	2.314	−1.639	4.517*****	−1.523*****
BC-CIV4	2.430	−1.638	4.729*****	−1.514****
BC-OVX			5.010*****	−1.445
HAR-MFIV	2.721	−1.642*	6.447*****	−1.519*****
HAR-ATMIV	2.600	−1.644**	5.583*****	−1.526*****
HAR-CIV1	2.393	− 1.646**	4.440****	− 1.533*****
HAR-CIV2	2.460	−1.645**	4.900****	−1.531*****
HAR-CIV3	2.518	−1.644**	5.276*****	−1.528*****
HAR-CIV4	2.629	−1.643*	5.943*****	−1.523*****
HAR-OVX			7.196	−1.494

Note: Panel A reports the values of MSE and QLIKE for the full sample. Panel B reports values of MSE and QLIKE for the Post-Fin sample, which includes the BC-OVX and HAR-OVX forecasts. Boldface is used to highlight minimum values. The results of Diebold–Mariano tests, based on Newey–West standard errors, are also presented. In Panel A the HAR forecasts are the benchmark against which all other forecasts are compared. In Panel B the HAR and HAR-OVX forecasts are used as benchmarks against which all other forecasts are compared. The superscripts *, ** and *** indicate significantly different forecast losses between the reference model and the HAR benchmark at the 10%, 5%, and 1% levels, respectively. Similarly, the superscripts ****, ***** and ***** indicate significantly different forecast losses between the reference model and the HAR-OVX benchmark at the 10%, 5%, and 1% levels, respectively. Abbreviations: ATMIV, at-the-money implied volatility; BC, bias-corrected; CIV, corridor implied volatility; HAR, heterogeneous autoregressive model; MFIV, model-free implied volatility; MSE, mean squared error; OVX, Crude Oil Volatility Index; QLIKE, quasi-likelihood.

forecasts which incorporate information from the OVX, we additionally report results from Diebold–Mariano tests where the HAR-OVX forecasts are treated as the benchmark.¹⁴

From Panel A it can be seen that, overall, the bias-corrected forecasts outperform the corresponding HAR forecasts according to the MSE. The BC-CIV1 forecasts result in the lowest MSE and, of the HAR-based forecasts, the lowest MSE is associated with HAR-CIV1. However, Diebold–Mariano tests show that no forecast results in an MSE which is significantly lower than that generated by the HAR forecasts.

The QLIKE results in Panel A differ slightly to those for the MSE. Under this loss function the HAR-based forecasts outperform the bias-corrected forecasts, with the most accurate forecasts being those of HAR-CIV1. However, in contrast to the MSE results, we find that the QLIKES for HAR-CIV1, HAR-CIV2, HAR-CIV3, and HAR-ATMIV are all significantly lower than for HAR at the 5% level. The QLIKES for HAR-MFIV and HAR-CIV4 are also significantly lower at the 10% level. The fact that we find significant differences when using the QLIKE loss function is most likely associated with the ability of this loss function to more accurately discriminate between competing variance forecasts (Patton & Sheppard, 2009). Overall, the results in Panel A suggest that using information in CIV1 leads to the most accurate variance forecasts.

The results for the Post-Fin sample (Panel B) are analogous to those of the full sample (Panel A) and support our overall conclusion that incorporating information from CIV1 leads to the most accurate forecasts. Importantly, it can also be seen that forecasts based on the OVX perform poorly. When the HAR-OVX forecasts are used as a benchmark, it can be seen that nearly all non-OVX option-implied forecasts produce significantly lower QLIKES and MSEs at the 5%

¹⁴Between the OVX-based forecasts, the BC-OVX forecasts result in a lower MSE, while the HAR-OVX forecasts generate a lower QLIKE. However, to avoid unnecessary complexity, we only use the HAR and HAR-OVX forecasts as benchmarks in our Diebold–Mariano tests reported in Table 5. This choice has no impact on our conclusions.

level.¹⁵ Although the BC-OVX forecasts result in an MSE that is significantly lower than that for the HAR-OVX forecasts, of the bias-corrected forecasts, BC-OVX result in the largest MSE.

In summary, the results show that prediction errors can be minimized when CIV1 is employed. The results are consistent with those from the Mincer–Zarnowitz and encompassing regressions, which showed that the OIVs were informationally superior to forecasts based on RV and that CIV1 had a higher information content relative to other OIVs. Our results also suggest that forecasts based on the OVX perform significantly worse than those based on our alternative OIVs. Further insights regarding the performance of the OVX, including the effect of CBOE's cut-off methodology, are provided by the analysis of our USO sample in Section 5.

4.5 | Forecast evaluation using economic criteria

Although evaluating variance forecasts with MSE and QLIKE is common, they are statistical loss functions. It is not clear how minimizing the MSE and/or QLIKE translates into economic gains for the forecaster. Therefore, to ascertain whether the improved forecast accuracy we observe leads to economic gains, we perform a volatility timing exercise.

We consider an agent whose investment opportunity set consists of the WTI crude-oil futures and a risk-free asset and assume the agent's objective is to maximize the utility of a portfolio consisting of these two assets. We follow the set-up of Bollerslev et al. (2018) and assume Sharpe ratios are constant and that a quadratic utility function provides an accurate approximation to investors' true utility functions. With these assumptions, the following function describes investors' utility per unit wealth (UoW),

$$UoW_t = \frac{U(w_t)}{W_t} = \left(w_t SR \sqrt{E_t(RV_{t+1,t+22})} - \frac{\gamma}{2} w_t^2 E_t(RV_{t+1,t+22}) \right), \quad (13)$$

where W_t is the investors wealth, w_t is the proportion of the investors wealth held in crude-oil futures, SR is the Sharpe ratio and γ is the investors level of risk aversion.¹⁶ It can be shown that constructing a portfolio to maximize expected utility is equivalent to forming a portfolio with a specific volatility target, where the optimal proportion of wealth to invest in crude-oil futures is,

$$w_t^* = \frac{SR/\gamma}{\sqrt{E_t(RV_{t+1,t+22})}}. \quad (14)$$

The ratio in the numerator corresponds to the volatility target and the denominator is the expected volatility. To operationalize the strategy, $E_t(RV_{t+1,t+22})$ in denominator is replaced with a variance forecast. We also follow Bollerslev et al. (2018) and set $SR = 0.4$ and $\gamma = 2$, which they argue are sensible parameters when forecasting variance over a monthly horizon, and results in a volatility target of 20%. Using Equation (14) to substitute for w_t in Equation (13), replacing $E_t(RV_{t+1,t+22})$ with our variance forecasts, and plugging-in our assumed values of SR and γ leads to the following expression for the utility per unit wealth based on forecast $f_{t+1,t+22}$,

$$UoW_t = 8\% \frac{\sqrt{RV_{t+1,t+22}}}{\sqrt{f_{t+1,t+22}}} - 4\% \frac{RV_{t+1,t+22}}{\sqrt{f_{t+1,t+22}}}. \quad (15)$$

Comparisons between models are then made using realized utility,

$$RUoW = \frac{1}{T} \sum_{t=1}^T UoW_t. \quad (16)$$

Note, realized utility is expressed as a percentage return and, given our assumptions, can take a maximum value of 4%.

¹⁵The BC-CIV4 (HAR-CIV1) forecasts generate a QLIKE (MSE) that is significantly lower than the QLIKE (MSE) for the HAR-OVX forecasts at the 10% level only.

¹⁶See Bollerslev et al. (2018) for further details.

TABLE 6 Realized utility results

Forecast	Panel A: Full sample							Panel B: Post-Fin sample						
	Realized utility	Quantiles						Realized utility	Quantiles					
		2.5%	5%	10%	25%	50%	75%		2.5%	5%	10%	25%	50%	75%
HAR	3.807	2.610	3.106	3.518	3.826	3.937	3.977	3.713	2.624	2.756	3.122	3.578	3.890	3.980
BC-MFIV	3.805	2.567	3.060	3.561	3.824	3.941	3.985	3.691	1.670	2.574	2.936	3.666	3.934	3.982
BC-ATMIV	3.811	2.661	3.169	3.527	3.815	3.944	3.988	3.725	1.805	2.613	3.266	3.680	3.940	3.996
BC-CIV1	3.804	2.294	3.174	3.568	3.800	3.937	3.989	3.748	2.037	2.826	3.287	3.705	3.931	3.989
BC-CIV2	3.808	2.528	3.171	3.580	3.818	3.942	3.990	3.744	1.920	2.739	3.258	3.706	3.939	3.996
BC-CIV3	3.809	2.605	3.163	3.563	3.818	3.942	3.989	3.735	1.842	2.686	3.268	3.689	3.939	3.997
BC-CIV4	3.809	2.576	3.121	3.565	3.817	3.943	3.989	3.713	1.695	2.604	3.192	3.663	3.938	3.991
BC-OVX								3.547	0.566	1.433	2.728	3.634	3.886	3.970
HAR-MFIV	3.822	2.684	3.145	3.509	3.815	3.951	3.988	3.735**	2.476	2.920	3.057	3.701	3.935	3.992
HAR-ATMIV	3.825*	2.706	3.124	3.536	3.828	3.953	3.989	3.747***	2.505	2.755	3.185	3.711	3.926	3.994
HAR-CIV1	3.829**	2.702	3.222	3.575	3.828	3.952	3.987	3.765***	2.503	2.789	3.248	3.755	3.927	3.994
HAR-CIV2	3.826**	2.714	3.163	3.570	3.820	3.952	3.988	3.757***	2.413	2.776	3.231	3.701	3.929	3.994
HAR-CIV3	3.825*	2.713	3.128	3.552	3.811	3.953	3.988	3.752***	2.442	2.772	3.201	3.699	3.931	3.994
HAR-CIV4	3.823	2.703	3.083	3.545	3.815	3.954	3.988	3.742***	2.481	2.823	3.123	3.705	3.932	3.994
HAR-OVX								3.696	2.502	2.759	3.076	3.622	3.835	3.984

Note: The realized utility is given in Equation (16) and the per period UoW is given in Equation (15). All values in the table are reported as percentages. In the columns reporting realized utility, boldface is used to highlight maximum values. Panel A reports the realized utility and quantiles of the UoW for forecasts calculated using the full sample. Panel B reports the realized utility and quantiles of the UoW for forecasts calculated using the Post-Fin sample. We also report the results of Diebold–Mariano tests (based on Newey–West standard errors) for differences between the realized utility of the HAR forecasts, which are the benchmark forecasts, and each of the remaining competing forecasts. The superscripts *, ** and *** indicate a significant difference to the benchmark at the 10%, 5%, and 1% levels, respectively.

Abbreviations: ATMIV, at-the-money implied volatility; BC, bias-corrected; CIV, corridor implied volatility; HAR, heterogeneous autoregressive model; MFIV, model-free implied volatility; OVX, Crude Oil Volatility Index; UoW, utility per unit wealth.

4.5.1 | Realized utility results

Panels A and B of Table 6 summarize the realized utility results for our full and Post-Fin samples, respectively. The first column of Panel A reports the realized utility. It can be seen that the HAR-based forecasts outperform the bias-corrected forecasts. In addition, of the HAR-based forecasts, HAR-CIV1 generates the highest realized utility.

In the first column of Panel A we also report the results from Diebold–Mariano tests of whether the realized utility associated with a forecast is significantly different to the realized utility generated by the HAR forecasts. It can be seen that HAR-CIV1, HAR-CIV2, and HAR-CIV3 produce realized utilities that are significantly higher, at the 5% level for HAR-CIV1 and HAR-CIV2 and the 10% level for HAR-CIV3, than the realized utility attained with the HAR forecasts. Therefore, the pattern in forecasting performance observed with the statistical loss functions is retained when we use an economic loss function. The results also confirm that the improvements in forecasting accuracy we observed with the statistical loss functions translate into economic benefits.

The difference between the value of the realized utility for the HAR-CIV1 forecasts relative to the HAR forecasts in Panel A is 2 bp. Although this may appear to be a relatively modest difference, there are two reasons why this represents a material economic improvement. First, as highlighted by Bollerslev et al. (2018), there has been a drive by investment management companies, in particular mutual funds and ETFs, toward lowering fees. The fees now charged by low-cost funds are of the order of 10 of basis points. As argued by Bollerslev et al. (2018), this means that a single digit basis point increase in fees is relatively substantial. The difference in realized utility of 2 bp means that a fund using the HAR-CIV1 forecasts instead of the HAR forecasts will be able to increase its fees by 2 bp and remain equally attractive to investors.

Second, the realized utilities reported are unconditional. Therefore, in any given month the economic benefit of using the HAR-CIV1 over the HAR forecasts could be much larger than 2 bp. To examine this further, we report in columns 2–7 of Panel A in Table 6 2.5%, 5%, 10%, 25%, 50%, and 75% quantiles of the UoW for each set of forecasts. Comparing the HAR-CIV1 to the HAR forecasts, it is clear that there are substantial differences, of 5–10 bp, in the UoW at the 2.5%, 5% and 10% quantiles. The difference between the lower quantiles of the UoW distributions suggests the HAR-CIV1 forecasts tend to outperform the HAR forecasts precisely when it is most difficult to forecast volatility and when an accurate forecast is in greatest demand, for example, when there is a volatility shock which causes UoW to be low.

The results for the Post-Fin sample in Panel B of Table 6 are analogous to those for the full sample in Panel A. Again, the HAR-CIV1 results in the highest realized utility, while, besides the OVX-HAR case, all augmented HAR models generate a realized utility higher than that of the HAR. The differences between the realized utilities of each non-OVX augmented HAR forecast and the HAR forecasts are also on the whole larger than those reported in Panel A of Table 6, being approximately 2–5 bp.

To examine the economic benefit of using our OIVs over the OVX, in Panel B we test for a significant difference between each forecast and the HAR-OVX forecasts.¹⁷ It can be seen that the realized utility of all non-OVX augmented HAR forecasts are significantly higher, typically at the 1% level, than the realized utility associated with the HAR-OVX forecasts. Therefore, these results further support our conclusion that there is value in constructing our OIVs directly from option prices rather than relying on the CBOE's methodology.

Similar to our findings in Panel A, there are potentially large differences between the conditional values of UoW for the Post-Fin sample. For the 25–50% UoW quantiles the difference between the realized utilities associated with the HAR and augmented HAR forecasts is approximately 4–13 bp, while the difference between the HAR-OVX and the other non-OVX augmented HAR forecasts is approximately 9 bp.

Of course, the magnitude of the economic benefits derived from each forecast depends on the assumptions employed. As Bollerslev et al. (2018) highlight, the value of the realized utility is a linear function of the volatility target. Thus, if the volatility target doubles, for example, through a doubling of the Sharpe ratio or a halving of the coefficient of risk aversion, the size of the economic benefits also double. Nevertheless, the framework above provides a sensible approximation, and therefore, the magnitude of the economic benefits presented should be reasonable.

In conclusion, the results demonstrate that the improvements in forecasting accuracy observed when using OIVs translate into economic benefits under reasonable assumptions. They also further corroborate our preference for the CIV1 among our OIVs.

4.5.2 | Realized utility results with transaction costs

To make our analysis more realistic, we also take into consideration transaction costs. Specifically, transaction costs are calculated as being a proportion of turnover,

$$TC = \sum_{t=1}^T c \cdot \text{turnover}_t, \quad (17)$$

where,

$$\text{turnover}_t = |w_t - w_{t-1}|.$$

The precise level of transaction costs is controlled by c . We follow Wang, Liu, Ma, and Wu (2016) and Caldeira, Moura, Nogales, and Santos (2017) and set c to be either 0.033% or 0.15%. It should be noted that transaction costs are low in futures markets (Locke & Venkatesh, 1997) and have decreased markedly over the past few decades. Realized utility net of transaction costs is then given by,

$$\text{NetRUoW} = \text{RUoW} - TC. \quad (18)$$

Panels A and B of Table 7 summarize the net realized utility and transaction costs for the full and Post-Fin samples, respectively. In Panel A (Panel B) we also use Diebold–Mariano tests to formally evaluate whether the net realized

¹⁷We use the HAR-OVX forecasts rather than the BC-OVX forecasts as a benchmark because, of the two, the former generate a higher realized utility.

TABLE 7 Realized utilities with transaction costs

Forecast	Panel A: Full sample				Panel B: Post-Fin sample			
	$c = 0.15\%$		$c = 0.033\%$		$c = 0.15\%$		$c = 0.033\%$	
	Net realized utility	TC	Net realized utility	Transaction costs	Net realized utility	TC	Net realized utility	Transaction costs
HAR	3.793	0.026	3.804	0.0056	3.697	0.0079	3.710	0.0017
BC-MFIV	3.791	0.026	3.802	0.0054	3.674	0.0081	3.687	0.0018
BC-ATMIV	3.797	0.025	3.808	0.0055	3.710	0.0073	3.721	0.0016
BC-CIV1	3.791	0.025	3.801	0.0054	3.734	0.0070*	3.745	0.0015*
BC-CIV2	3.795	0.024	3.805	0.0054	3.729	0.0069*	3.740	0.0015*
BC-CIV3	3.796	0.025	3.807	0.0055	3.720	0.0071	3.732	0.0016
BC-CIV4	3.795	0.025	3.806	0.0056	3.698	0.0076	3.710	0.0017
BC-OVX					3.524	0.0109	3.542	0.0024
HAR-MFIV	3.808	0.026	3.819	0.0057	3.718**	0.0080	3.731**	0.0018
HAR-ATMIV	3.811*	0.026	3.822*	0.0057	3.731***	0.0079	3.744***	0.0017
HAR-CIV1	3.815**	0.026	3.826**	0.0057	3.749***	0.0076	3.761***	0.0017
HAR-CIV2	3.812**	0.026	3.823**	0.0057	3.741***	0.0076	3.754***	0.0017
HAR-CIV3	3.811*	0.026	3.822*	0.0057	3.735***	0.0078	3.748***	0.0017
HAR-CIV4	3.809	0.026	3.820	0.0057	3.725***	0.0079	3.738***	0.0017
HAR-OVX					3.679	0.0080	3.692	0.0018

Note: The net realized utility is given in Equation (18) and transaction costs (TC) are given in Equation (17). Net realized utility and TC are presented for $c = 0.15\%$ and 0.033% . All values in the table are reported as percentages. In the columns reporting net realized utility, boldface is used to highlight maximum values. Panel A reports the net realized utility and TC for forecasts calculated using the full sample. Panel B reports the net realized utility and TC for forecasts calculated using the Post-Fin sample. We also report the results of Diebold–Mariano tests for differences between the net realized utility of the HAR forecasts, which are the benchmark forecasts, and each of the competing alternatives. Analogous Diebold–Mariano tests are conducted for TC, where the TC of the HAR forecasts are the benchmark. Newey–West standard errors were used in all Diebold–Mariano tests. The superscripts *, ** and *** indicate a significant difference to the benchmark at the 10%, 5%, and 1% levels, respectively.

Abbreviations: ATMIV, at-the-money implied volatility; BC, bias-corrected; CIV, corridor implied volatility; HAR, heterogeneous autoregressive model; MFIV, model-free implied volatility; OVX, Crude Oil Volatility Index.

utility and transaction costs associated with each forecast are significantly different to the net realized utility and transaction costs generated by the HAR (HAR-OVX) forecasts.

In both Panels A and B it can be seen that transaction costs, whether $c = 0.015\%$ or 0.0033% , do not differ substantially between the competing forecasts. In Panel A, it can be seen that there are no significant differences in transaction costs, while in Panel B, BC-CIV1 and BC-CIV2 lead to significantly lower transaction costs at the 10% level. Consequently, because none of the forecasts lead to unusually high transaction costs, the ranking of the forecasts in Panels A and B of Table 7 are identical to those in Panels A and B of Table 6. In particular, the HAR-CIV1 forecasts produce the highest net realized utility in both the full and Post-Fin samples. Within the full sample, the HAR-CIV1 and HAR-CIV2 forecasts result in net realized utilities that are significantly higher than the net realized utility of the HAR forecasts at the 5% level. While for the Post-Fin sample, all the non-OVX augmented HAR forecasts result in significantly higher net realized utilities than the HAR-OVX forecasts, typically at the 1% level. Thus, differences in trading volumes are small and do not have a material impact on the relative economic benefits of the forecasts. Therefore, the results also support our conclusions in Section 4.5.1.

4.6 | Forecast performance and variance risk premia

CIV measures discard some information from option panels as contracts with strikes outside the corridors are not included in their calculation. This creates a mismatch between the option-implied quantity (i.e., expected risk-neutral

corridor variance) and the target quantity (expected real-world IVAR). If narrow CIV measures perform well empirically, it must be the case that the benefits from discarding these options outweighs the costs.

There are two reasons that can justify this trade-off. The first is that the discarded options, due to thin or nonexistent trading, merely contaminate option-implied expectations with measurement error. The second is that certain options embed risk premia, in the sense that a higher expected volatility is required to explain their price.

For the purpose of forecasting, the presence of such risk premia may not be a significant concern if they are relatively constant across time, and a sufficiently long history of past data is available to estimate them, so that the bias from the option-implied forecasts can be removed. What is notoriously more challenging is the possibility that they vary substantially over time, causing variations in the option-implied measures that are unrelated with the target quantity, that is expected IVAR. In this section we investigate whether the behavior of risk-premia can help us explain why our narrow corridor measure perform well in practice.

4.6.1 | Variance and corridor risk premia

Thus far we have presented our empirical results without attempting to explain why CIV1 provides the most accurate variance forecasts. In the stock market, the *variance risk premium* (VRP), defined as the difference between real-world and risk-neutral expected values of variance, is known to vary across time (Todorov, 2009) and have an unconditional expected value that is negative (Bollerslev, Tauchen, & Zhou, 2009; Carr & Wu, 2009). Naturally, these properties of the VRP adversely affect the accuracy of volatility forecasts derived from the prices of options.

To gain an insight into whether a similar argument applies to our crude oil data, we analyze the time series properties of risk premia associated with our option-implied measures.

To estimate the VRP we follow Bollerslev et al. (2009) and use¹⁸

$$VRP_t = E^{\mathbb{Q}}[IVAR_{0,T}] - E[IVAR_{0,T}] = MFIV_{t,t+22} - RV_{t-21,t},$$

where the IVAR measures for the real-world (\mathbb{P}) and risk-neutral world (\mathbb{Q}) have been discussed in Section 2.1 and Section 2.2, respectively. Due to our CIVs providing risk-neutral expectations of CIVAR, it is not appropriate to estimate their corresponding risk premia using the equation above. Instead, we use the analogous concept of *corridor variance risk premium* (CVRP), defined as

$$CVRP_t^{(B_1, B_2)} = E^{\mathbb{Q}}[CIVAR_{0,T}^{(B_1, B_2)}] - E[CIVAR_{0,T}^{(B_1, B_2)}] = CIV_{t,t+22}^{(B_1, B_2)} - CRV_{t-21,t}^{(B_1, B_2)},$$

where $CIV_{t,t+22}^{(B_1, B_2)}$ is equal to CIV1, CIV2, CIV3, or CIV4, and $CRV_{t-21,t}^{(B_1, B_2)}$ represents the associated corridor realized variances (CRV) which we denote as CRV1, CRV2, CRV3, and CRV4.¹⁹ Similarly, the barriers of $CVRP^{(B_1, B_2)}$ correspond to those of our four CIV measures, so we denote the respective corridor variance risk premia estimates as $CVRP^1$, $CVRP^2$, $CVRP^3$, and $CVRP^4$.

4.6.2 | Statistical properties of CRV

Since CRVs have not been explored in the extant literature, we first examine their time-series properties. Panel A of Table 8 provides summary statistics for our realized variance measures. As expected, the means and standard deviations of the CRVs are lower than those of the RV. The skewness and kurtosis of CRV1–CRV3 are also markedly lower. Therefore, the time-series properties of our CRVs are analogous to those of the CIVs in that their time variation is lower

¹⁸This estimate, which essentially assumes that monthly RVs follow a random walk, has the advantage that it does not require a model to generate expected variance forecasts. Both Bollerslev et al. (2009) and Feunou, Jahan-Parvar, and Okou (2017) report that they obtain qualitatively similar results when model-based forecasts are used, so they opt for this simpler estimate instead.

¹⁹Both the RV and CRV measures were computed using the subsampling method discussed in Section 2.4.

TABLE 8 Descriptive statistics for corridor realized variances and corridor risk premia

Model	Mean	SD	Skew	Kurt	ρ_1	ρ_6	ρ_{12}
Panel A							
RV	0.111	0.148	6.846	71.679	0.538	0.109	0.008
CRV1	0.015	0.021	4.395	30.497	0.572	0.105	0.052
CRV2	0.043	0.058	4.164	26.661	0.658	0.137	0.033
CRV3	0.068	0.093	5.012	36.690	0.635	0.130	0.029
CRV4	0.100	0.144	7.398	80.654	0.532	0.087	0.025
Panel B							
R1	0.144	0.118	1.030	3.764	0.024	0.015	0.009
R2	0.408	0.250	0.439	2.379	0.003	0.027	0.059
R3	0.632	0.285	-0.262	1.931	-0.013	0.112	0.038
R4	0.917	0.195	-1.909	7.680	-0.003	-0.015	-0.021
Panel C							
VRP	0.038***	0.081	-3.739	49.090	0.233	0.167	0.173
CVRP ¹	0.007***	0.017	-0.350	12.904	0.106	0.013	0.045
CVRP ²	0.014***	0.041	-0.625	13.652	0.134	0.015	0.046
CVRP ³	0.020***	0.057	-0.901	13.194	0.161	0.048	0.103
CVRP ⁴	0.026***	0.084	-6.429	87.237	0.198	0.070	0.137

Note: This table reports the mean, standard deviation, skewness, kurtosis and 1-, 6- and 12-order autocorrelations (ρ_1 , ρ_6 , and ρ_{12} , respectively) for realized variance (RV), corridor realized variances (CRV), realized ratios R (calculated as CRV/RV), as well as variance risk premia (VRP) and corridor variance risk premia (CVRP). Panel A reports the summary statistics for RV and CRV1–CRV4. Panel B reports the summary statistics for R1–R4. Panel C reports the summary statistics VRP and CVRP¹–CVRP⁴.

and have fewer extreme values than their full variance counterparts. Although arguably a little higher for the CRVs, the autocorrelations are of a similar magnitude with those of the RV.

Our CIV measures provide (risk-neutral) forecasts of future CIVAR. Our target quantity in this paper, however, is IVAR. Therefore, a closer inspection at the relationship between two, as measured by our respective proxies (i.e., CRV and RV), can provide some useful insights. For instance, if the relative differences are stable over time, then CIVAR forecasts can be easily scaled up to provide reasonable predictions of IVAR.

Along these lines, we examine the statistical properties of the ratios between the CRV and RV measures (i.e., CRV/RV) for each of our corridors. For CRV1 to CRV4 these ratios are denoted as R1 to R4. The summary statistics are displayed in Panel B of Table 8. As it can be seen therein, although the skewness and kurtosis of R1 are a little larger than those of R2 and R3, R1 has a much lower standard deviation. In terms of persistence, the autocorrelations for all the realized variance ratios are of comparable magnitude. Therefore, the relationship between CRV1 and RV appears to be more stable than for the other wider corridors, so it reasonable to assume that sensible IVAR predictions can be obtained by scaling up the corresponding CIVAR forecasts.

4.6.3 | Statistical properties of corridor risk premia

To understand the behavior of corridor risk premia and examine what impact they might have on our forecast results, we take a closer look at their statistical properties. As it can be seen from Panel C of Table 8, several observations are noteworthy. First, VRP and CVRPs are significantly different to zero in all cases, meaning that all the option-implied measures provide biased forecasts of their respective target quantities. Second, both the mean and standard deviation of the CVRP¹ are substantially lower compared to the rest. Third, the skewness and kurtosis of CVRP¹, CVRP², and CVRP³ are comparatively low, indicating that narrow corridors are associated with fewer extreme CVRPs values. This suggests that for narrow corridor measures the difference between risk-neutral and real-world expectations will not be prone to sudden large deviations.

We continue our discussion by looking at the most important link between the statistical properties of the various CVRPs and the forecast performance of the associated CIV measures. Specifically, we examine if variance and corridor risk premia vary across time by testing for zero autocorrelations in the series $(x - \bar{x})$ where x is a reference VRP or

CVRP estimate and \tilde{x} denotes its sample mean. We consider 1, 6, or 12 lags and the null hypothesis that all autocorrelations up to these lags are jointly zero is evaluated using the Ljung-Box test. This test should uncover evidence of conditional mean dynamics in each series. We find that at the 5% level the constant mean assumption cannot be rejected for any of the CVRPs. On the contrary, at the same significance level, the zero autocorrelation assumption is rejected at both 1 and 6 lags for the case of the VRP.²⁰ As a time-varying risk-premium contaminates option-implied forecasts with variations that are irrelevant to the target quantity (expected IVAR) but rather reflect adjustments in the pricing of risk premia, this empirical finding directly demonstrates the advantage of CIV forecasts over those obtained from the full MFIV measure.

We conclude our empirical evidence on time-varying risk premia by doing a more careful comparison between volatility expectations extracted from options at the tails of the risk-neutral density versus those that have strikes around its mean. To do so, we compare the risk-premia associated exclusively with extreme strike options, defined as $CVRP^{tail} = VRP - CVRP^4$, with those that correspond to the central support of the RND. For the latter, we use the $CVRP^2$ estimate that contains a comparable (risk-neutral) probability mass. Since risk-premia should vary with the state of the economy, we use the economic activity index of Aruoba, Diebold, and Scotti (2009), which we denote as ADS, to capture any changes in the conditional mean of each series. In particular, we estimate the following regression models:

$$CVRP_t^2 = c + \beta ADS_{t-1} + e_t, \quad CVRP_t^{tail} = c^* + \beta^* ADS_{t-1} + e_t^*$$

The results of these regressions are revealing. For the case of the $CVRP^2$ measure, the estimated parameters are $c = 0.028$ (significant at the 1% level) and $\beta = -0.00102$, which has the correct sign (i.e., risk premia are higher in bad economic states), but is insignificant with a p-value of 93%. With respect to the $CVRP^{tail}$ measure, the estimated parameters are $c^* = 0.0133$ (significant at the 1% level) and $\beta^* = -0.00648$ which has the correct sign as well, but is significant at the 5% level using HAC standard errors.²¹ In connection with our forecasting study, this result carries two important messages for option-implied measures that rely on extreme strike options. First, it provides further evidence that the risk-premia associated with extreme strike options are time-varying, which generally impairs their forecast accuracy. Second, it highlights that their risk-neutral forecasts will abruptly drift apart from real-world expectations during times of sudden market turmoil, so that their forecast performance will deteriorate when it is needed the most. So, overall, studying the dynamics of variance and corridor risk premia provides some intuition regarding the good forecasting performance of our narrow CIV measures, as well as strengthens the argument for preferring them in practice.

5 | FURTHER INSIGHTS USING USO OPTIONS

Although the main focus of this paper is forecasting crude-oil volatility, interesting insights can be obtained by studying a closely related quantity, namely the volatility of the USO ETF. First, given that the two underlyings are tightly linked but trade in different markets, offers a unique opportunity to examine if our key findings are mainly driven by trading venue idiosyncrasies such as market liquidity. Furthermore, the fact that the USO data set underlies CBOE's computation of the OVX paves the way for methodological comparisons. For instance, although our main results clearly showed that our CIV measures deliver more accurate crude-oil volatility forecasts than the OVX, directly contrasting the performance of the two methodologies was inappropriate as their implementation was based on different data. Finally, the USO data set enables us to explore alternatives to CBOE's cut-off point rule and evaluate their forecast performance.

The details of the USO data set, that consists of options and associated high-frequency returns of the underlying ETF, can be found in Section 3.1.3. To remain consistent with our main empirical analysis on crude-oil data, we adopt exactly the same methodological approach. Along these lines, our target quantity is the monthly realized variance of the asset underlying the option contracts, that is the USO ETF. All option-implied measures are obtained from option contracts that have approximately 22 trading days to expiration and our forecasts are strictly nonoverlapping.

²⁰The p values corresponding 1 lag (6 lags) are 0.093 (0.75) for the $CVRP^1$, 0.71 (0.56) for $CVRP^2$, 0.73 (0.251) for $CVRP^3$, and 0.16 (0.74) for $CVRP^4$. When we test for zero correlation at the first 12 lags, the p value of the VRP is .07 while all CVRP estimates have p values above .7.

²¹We have also experimented with the ADS_{t-2} variable, since the CVRP measures are constructed using RV_{t-1} , but our findings remain unaltered.

5.1 | Forecasting USO volatility

In this section, we investigate the information content of various volatility forecasts extracted from the prices of our USO options, that is we examine the performance of the MFIV, ATMIV, OVX, and our four CIV measures. Following our empirical exercise for crude-oil volatility, we first conduct regression analysis and subsequently assess each alternative using two statistical loss functions.

5.1.1 | Regression analysis

We begin our analysis by inspecting the results of the various forecast regressions.²² Table 9 shows the (adjusted) R^2 from Mincer–Zarnowitz regressions when each of the competing alternatives is used to predict the monthly variance of the USO fund. As it can be seen therein, the CIV1 forecasts have the highest correlation with the target quantity, followed by CIV2, CIV3, and CIV4. Consistent with our earlier findings for crude-oil volatility, the information content of the HAR forecasts is notably inferior to any of the OIV measures. The OVX forecasts are ranked second to last.

Since combinations of forecasts from both option-implied and realized variance models are often preferable in practice, it is important to assess our OIVs measures with respect to the incremental volatility information they contain, that is, information that is not present in RV-based forecasts. Along these lines, we examine the adjusted R^2 from bivariate specifications where realized variance is regressed on both a reference OIV measure as well as the HAR forecast. As our results show, comparisons in terms of adjusted R^2 reveal that including HAR forecasts is not productive since every univariate model ranks favorably compared to its bivariate counterpart. Notably, the relative ranking between bivariate models remains remarkably similar to that of univariate specifications. Focusing exclusively on bivariate models, the most accurate forecasts are generated by those containing the CIV1 measure, followed by those that utilize CIV2 and CIV3, while the lowest correlation between forecasts and realizations are those corresponding to the OVX and ATMIV measures.

Overall, using the adjusted R^2 to rank all models clearly shows that narrow corridor measures contain the most useful information in term of predicting future USO volatility. In particular, the best results are those corresponding to the univariate models containing CIV1, CIV2, and CIV3 followed by their bivariate counterparts that combine them with HAR forecasts. On the other side, both univariate and bivariate OVX forecasts rank toward the bottom of the list. It is also worth mentioning the MFIV measure delivers less informative forecasts compared to any of the CIV alternatives. The superior performance of the CIV1, the unimpressive ranking of the OVX, as well as the inferior MFIV forecasts vis-à-vis our CIV measures, mirror exactly our earlier findings for the crude-oil data set.

5.1.2 | Statistical loss functions

We complete our empirical exercise by evaluating the accuracy of our models using our two benchmark statistical loss functions, that is the symmetric MSE and the asymmetric QLIKE. Consistent with our forecast setting for the case of crude-oil, we generate monthly variance forecasts from 15 separate forecast models. Specifically, seven specifications are based exclusively on OIV measures, one utilizes only past return data (HAR), and seven are augmented models that combine HAR with OIV forecasts. One-step ahead forecasts are generated by estimating each regression model, described in Sections 3.4 (HAR model) and 4.3 (univariate and augmented models), using a rolling window of 60 monthly observations, so that all forecasts are bias-corrected.

Table 9 displays the results of this forecasting horse race. Under the MSE loss function, the first thing that becomes apparent is that all univariate OIV models outperform every augmented specification. Second, forecasts obtained exclusively by relying on historical return information (i.e., generated by the HAR model) are less sharp than those of any other model. Third, the most accurate forecasts are generated by the univariate CIV1 specification, followed by those corresponding to the univariate CIV2 measure. It is also noteworthy that OVX forecasts deliver very low MSE losses as well, as they are only worst

²²It should be noted that the number of observations in these forecast regressions is lower than those corresponding to the crude-oil Post-Fin sample. This is because the first part of the data had to be reserved for the estimation of the HAR model. This was not a concern for the crude-oil data set since, unlike case of the USO, a long record of historical futures prices is available.

TABLE 9 Summary of forecasting results for the USO volatility

Model	Adjusted R^2 (%)	MSE ($\times 1,000$)	QLIKE ($\times 10$)
HAR	22.684	8.571	-24.846
MFIV	52.680	6.800	-24.532
ATMIV	51.943	6.856	-24.517
CIV1	53.835	6.518	-24.619
CIV2	53.686	6.663	-24.569
CIV3	53.671	6.704	-24.551
CIV4	53.252	6.761	-24.538
OVX	51.790	6.669	-24.596
HAR-MFIV	52.547	7.864	-24.898
HAR-ATMIV	51.742	7.863	-24.888
HAR-CIV1	53.612	7.824	-24.918
HAR-CIV2	53.546	7.822	-24.908
HAR-CIV3	53.577	7.823	-24.906
HAR-CIV4	53.158	7.843	-24.901
HAR-OVX	51.540	7.870	-24.903

Note: This table displays the adjusted R^2 for univariate and bivariate forecast regressions on monthly realized variance. Results for the univariate models are obtained by evaluating Equation (9), while for bivariate models, that is, specifications that combine option-implied and HAR forecasts, they are obtained through Equation (10). The table also displays the MSE and QLIKE losses corresponding to bias-corrected and augmented forecasts, generated through Equations (11) and (12), respectively.

Abbreviations: ATMIV, at-the-money implied volatility; CIV, corridor implied; HAR, heterogeneous autoregressive model; MFIV, model-free implied volatility; MSE, mean squared error; OVX, Crude Oil Volatility Index; QLIKE, quasi-likelihood; USO, United States Oil Fund.

than those of CIV1 and CIV2. The augmented OVX specification continues to perform very poorly though, coming second to last compared to all other models. Finally, for both univariate and augmented specifications, the MFIV measure continues to deliver less accurate volatility forecasts compared to any CIV measure.

Under QLIKE losses, where a stronger penalty is imposed for underpredicting variance, models that combine two sources of information, namely option prices and historical returns, deliver more accurate forecasts than those obtained by univariate specifications. The sharpest forecasts are provided by the HAR-CIV1 model, followed by those of HAR-CIV2 and HAR-CIV3. On the other end, the worst forecasts are obtained by the univariate ATMIV and MFIV models. The HAR model generates lower QLIKE losses than any univariate OIV model but higher than any augmented specification. Finally, out of 15 the models, those that involve the OVX measure rank fourth (univariate) and tenth (augmented) from the top, so the performance of the measure is reasonable overall.

5.1.3 | Crude-oil versus USO volatility forecasts

How much do our key empirical findings change when examine the volatility of the USO ETF instead of that of crude-oil futures? This is answer is very little, if at all. In both cases, regression analysis indicated that forecasts from historical return models are outmatched by option-implied forecasts. Furthermore, narrow corridor implied volatilities outperform either the MFIV or ATMIV measures. Using the adjusted R^2 as a ranking criterion, the CIV1 contains more information about future volatility levels compared to any other measure. With respect to MSE losses, univariate models that utilize OIV measures deliver sharper forecasts compared to augmented specifications. Once again, the most accurate forecasts are generated by narrow CIV measures, notably CIV1, instead of other benchmarks, such as the HAR, ATMIV, MFIV, or OVX forecasts. When the QLIKE loss function is considered, models that combine both option-implied and past realized variances rank favorably compared to their univariate counterparts. The lowest forecast errors are still obtained by models that utilize narrow corridor measures, with the augmented HAR-CIV1 model providing the best results.

All the above are true for both the USO ETF, as well as, the two crude-oil data sets (full and Post-Fin samples). The only notable, yet rather unsurprising, difference in the USO results is that the performance of the OVX index is not as poor as in the crude-oil sample. In particular, while still ranking below any alternative option-implied measure in univariate or bivariate forecast regressions, models that utilize the OVX generally have, in relative terms, lower realized losses than before. Some improvement was obviously expected since for the case of crude-oil forecasts there was a mismatch between the target quantity (crude-oil volatility) and the underlying asset of the options used in the OVX calculation (USO ETF).

5.2 | Alternative cut-off points for the OVX calculation

In general, the usefulness of the OVX index compared to other alternatives depends on the application at hand. For example, if one is interested in studying risk premia associated with extreme price movements, narrow CIV measures will be of little value as they discard options with extreme strikes. With that in mind, and consistent with the spirit of our paper, the alternative OVX indices we construct are only evaluated with respect to their forecast accuracy. In the interests of brevity, we only assess the forecasts using regression analysis.

All our alternative OVX indices are computed by applying the CBOE methodology on same option chains, but each of them differs on how the cut-off point, determining which options will be included in the calculation of the measure, is decided. In particular, we construct indices where the cut-off point is obtained by targeting fixed values of the effective range (ER) statistic of Andersen et al. (2015), defined as

$$ER = \left[\frac{\ln(K_1/F)}{\sigma_{BS} \sqrt{T}}, \frac{\ln(K_N/F)}{\sigma_{BS} \sqrt{T}} \right],$$

where (K_1, K_N) represents the strike range, σ_{BS} is the at-the-money Black–Scholes implied volatility (ATMIV in our notation), and T is the time-to-maturity of the underlying options. We include five OVX alternatives in our comparisons by setting $K_1 = -z$ and $K_N = z$ for $z = 1, 1.5, 2, 2.5, 3$. The corresponding measures are denoted as $OVX_1, OVX_{1.5}, OVX_2, OVX_{2.5}$, and OVX_3 . The measure obtained by applying the CBOE's cut-off algorithm is simply denoted as OVX.

Our proposed methodology for directly using the effective range as a means to determine which options should be included in OVX calculation formula has a number of appealing features. First, it is a simple modification of CBOE's cut-off methodology that is straightforward to implement. Second, besides calculating the ATMIV, no additional assumptions, such as the estimation of a risk-neutral density, are required. Third, the rationale of using a fixed effective range target is clear. In times of high liquidity, more options are included in the original OVX calculation, which increases, *ceteris paribus*, the magnitude of the extracted measure. This effect is largely alleviated by keeping the strike range, normalized by the level of risk-neutral volatility (σ_{BS}), relatively stable across time and largely independent from market liquidity.²³

Targeting fixed values of the effective range statistic determines the minimum and maximum strikes that can be considered in the OVX calculation. Of course, since observed strike prices discrete, the actual range of (normalized) strikes will not be constant. Figure 2 depicts the actual effective range for the OVX , OVX_2 , and the OVX_1 measures computed using the observed strike prices included in each VIX formula. It is clear that the effective range for the OVX measure exhibits substantial time-variation throughout our sample. Notably, in some cases the range of (normalized) strikes changes markedly from 1 month to the next. Unsurprisingly, the effective ranges for OVX_2 and the OVX_1 measures are very stable across time.

We now examine the information content of the OVX indices by looking at the adjusted R^2 of forecast regressions on realized variance. We consider both univariate and bivariate models, that is, regression models that only use OVX indices and those that include the HAR forecast as an additional regressor. As shown in Table 10, with the exception of the OVX_1 , every univariate model has a higher adjusted R^2 compared to its bivariate counterpart. The ranking within each of these groups is identical for both types of models, namely OVX_1 provides the best results followed by $OVX_2, OVX_{1.5}, OVX_{2.5}, OVX_3$, and, lastly, the OVX . So, with the exception of OVX_2 , the wider the cut-off range the more noisy the option-implied information about future realized variance. The most important empirical finding, however,

²³It should be noted that the simplicity provided by the ER approach is not without cost. For example, unlike our GB2-RND approach, it does not take into account the effect of risk-neutral skewness in determining the strike/corridor range.

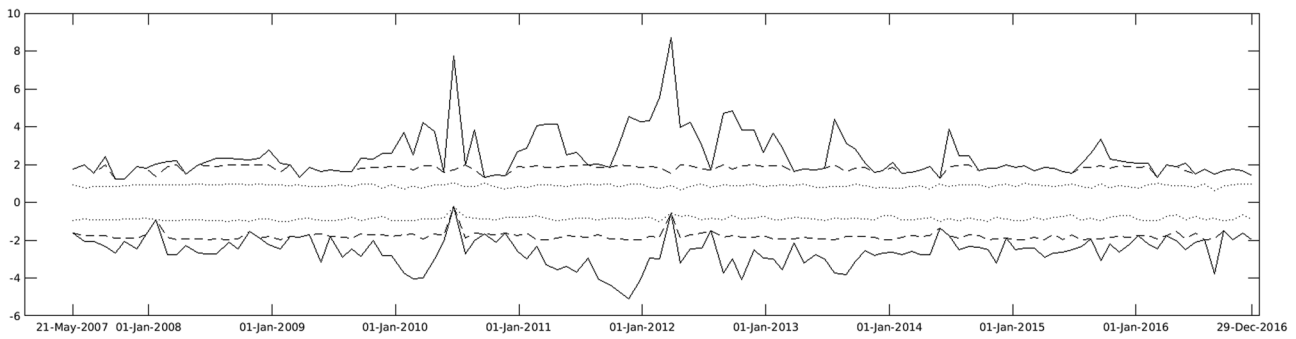


FIGURE 2 Effective (normalized) strike range corresponding to alternative crude-oil volatility indices, namely the OVX (solid line), OVX1 (dashed line) and the OVX2 (dotted line)

TABLE 10 Information content tests for alternative OVX indices

	OVX	OVX ₁	OVX ₁₅	OVX ₂	OVX ₂₅	OVX ₃
Panel A: Adjusted R^2 (%)						
Mincer–Zarnowitz	51.790	53.320	52.530	52.844	52.070	51.806
Augmented	51.540	53.656	52.425	52.720	51.883	51.552
Panel B: Encompassing Regressions						
P value (OVX)		.199	.334	.080	.269	.806
P value (column measure)		.072	.253	.064	.223	.742

Note: Panel A displays the adjusted R^2 values from Mincer–Zarnowitz regressions involving the alternative OVX measures discussed in Section 5.2. It also reports the adjusted R^2 s from augmented regressions that combine each OVX measure with the HAR forecast. Panel B presents the p values from the bivariate encompassing regressions described in Equation (19). The top row corresponds to the p values of the OVX parameter while the bottom row corresponds to the p value of the competing OVX measure displayed at the top of the column.

is that the cut-off algorithm employed by the CBOE appears to result in a measure that provides the least informative forecasts compared to any of the ER-based alternatives. It is also worth noting that while models that include the OVX_1 generate the highest adjusted R^2 in these comparisons, they still rank below our CIV1, CIV2, and CIV3 measures.

We conclude our analysis by conducting a series of encompassing regression tests where we combine the OVX measure with each of the competing indices. Specifically, we compare the information content of OVX forecasts, denoted as f^{OVX} , against those of a reference ER-based measure, denoted as f^{er} , by estimating

$$RV_{t+1,t+22} = \beta_0 + \beta_1 f_{t,t+22}^{OVX} + \beta_2 f_{t,t+22}^{er} + \epsilon_{t+22}. \quad (19)$$

The estimation results are presented in Panel B of Table 10. When the OVX is combined with either $OVX_{1.5}$, $OVX_{2.5}$ or OVX_3 neither β_1 or β_2 are significant, although the latter has slightly lower p values than the former for all three cases. On the contrary, when the OVX is combined with the $OVX_{2.5}$ measure both β_1 and β_2 are significant at the 10% level, with p values of 8% and 6.5% respectively. Finally, and most notably, our empirical results show that the OVX_1 measure encompasses the OVX at the 10% level since β_2 is significant (p value of 7%) while β_1 is not (p value of 20%).

6 | ROBUSTNESS CHECKS

In this section, we analyze the robustness of our results to: (a) the construction of alternative option-implied measures; (b) the size of the rolling-window used to estimate the models; (c) the choice of out-of-sample period; (d) the method used to bias correct the OIVs; and (e) the inclusion of overnight returns. In the interests of brevity, we only discuss the

results for our main data set, that is the full crude-oil sample spanning the 1996–2016 period. Detailed tables, displaying the estimation results of Sections 6.1–6.5, are provided in the supplementary online appendix.²⁴

6.1 | Alternative option-implied measures

When constructing our option-implied measures, several methodological choices have to be made. The two most notable ones, were presented in Section 3.3. First, a model that can generate option prices for arbitrary strikes is needed. Second, one needs to select the barrier levels that define each corridor measure. Below, we explain the rationale of our decisions and empirically investigate the robustness of our results by contrasting them with two reasonable alternatives.

6.1.1 | Cubic spline implied volatility function

In theory, a continuum of observed option prices is required in order evaluate Equations (4) and (5) and calculate model-free and corridor implied volatilities. Since this is not feasible in practice, some assumptions are needed to generate option prices at arbitrary strikes. A wide collection of alternative methods can be employed for this task, with the cubic spline implied volatility function approach being quite popular in the literature, especially for the case of S&P 500 options.²⁵ Instead of fitting a polynomial function to our implied volatility data, we have opted for a parametric GB2 risk-neutral density (RND) estimation instead.²⁶

The key finding of our paper is that CIV measures contain more information about future crude-oil volatility compared to other natural benchmarks, that is, forecasts obtained through the HAR model or the ATMIV, OVX, and MFIV measures. Among these benchmarks, MFIV is based on the GB2 RND approach we have adopted. It is, therefore, possible that our CIV measures have outperformed MFIV simply because of our poor modeling assumption.

To empirically investigate whether our RND choice had an impact on our findings, we have constructed a MFIV measure ($MFIV_c$) using a cubic spline approach as in Jiang and Tian (2005) and compared its performance in predicting future crude-oil variance using regression analysis. What we found was that, for both univariate and bivariate regressions (which include the lagged RV), the $MFIV_c$ had a lower R^2 compared to any other CIV measure. Encompassing regressions that combine $MFIV_c$ with each of the alternative OIV measures further underscored these results. In particular, the parameter corresponding to the $MFIV_c$ measure was always insignificant, while the narrow corridor measures CIV1 and CIV2 encompassed the $MFIV_c$ measure at the 10% level.

6.1.2 | Corridor selection vis-à-vis Andersen and Bondarenko (2007)

As noted in Section 3.3, the corridor width of the CIVs we examine are determined by the quantiles of the risk-neutral distribution F_Q , that is for a given quartile level p , we discard options that have strike prices either lower than $F_Q^{-1}(p)$ or higher than $F_Q^{-1}(1 - p)$. Our resulting CIV measures are obtained by setting $p = .45, .35, .25$ or $.1$, so our selection slightly differs from that of Andersen and Bondarenko (2007). Specifically, in their study Andersen and Bondarenko (2007) set $p = .25, .10, .05$ or $.025$, so our empirical results do not include their two widest corridor measures. Our choice was guided by the fact that crude-oil options are far less liquid compared to those written on the S&P 500, so option prices at extreme strikes would generally be either highly unreliable or not traded at all. With that in mind, we have opted to examine two narrow corridors measures instead, that is, CIV1 and CIV2.

To verify that the exclusion of extreme corridor measures is innocuous, we construct the two “extreme” CIV measures lacking from our CIV collection, that is $CIV_{.05}$ and $CIV_{.025}$ (obtained by setting $p = .05$ and $.025$, respectively), and examine their performance using regression analysis. In particular, we make comparisons in terms the adjusted R^2 of forecast regressions as well as conduct inference via encompassing specifications.

²⁴The online appendix can be found at bit.ly/AppendixJFM.

²⁵See, for example, Jiang and Tian (2005), Bliss and Panigirtzoglou (2004), and Dotsis and Vlastakis (2016), among others.

²⁶The cubic spline method is also not the preferred approach of two studies closely related to ours, namely Andersen and Bondarenko (2007) and Andersen et al. (2015).

Regressing the realized monthly variance series on either $CIV_{.05}$ or $CIV_{.025}$ and comparing the resulting adjusted R^2 with those of the other OIV alternatives revealed that, for both cases, their performance is inferior to all other CIV measures. The same is true when lagged RV is included in the regression models. Moreover, encompassing regressions demonstrate that every CIV measure encompasses $CIV_{.05}$ and $CIV_{.025}$ at the 10% level, while for the case of the CIV1 this is true at the 5% level. All in all, and in line with the main findings of our paper, very wide corridors tent to be associated with inferior forecast performance and that is demonstrably true for extreme measures such as $CIV_{.05}$ and $CIV_{.025}$.

6.2 | Estimation window

Thus far, the parameters for our bias-correction and HAR models have been estimated using rolling windows of 60 observations, or approximately 5 years' worth of data. To examine the robustness of our results to this modeling choice, we vary the estimation window between 66 and 96 observations, or between approximately 5 and 8 years, and then re-evaluate our forecasts using the MSE, QLIKE, and realized utility loss functions.²⁷

Reassuringly, our main conclusions continue to hold. In particular, for all estimation windows, the BC-CIV1 model has the lowest MSE, while the HAR-CIV1 has the lowest QLIKE and highest realized utility. Therefore, our results are robust to the choice of estimation window.

6.3 | Subsample analysis

In our analysis in Section 4, our out-of-sample period started in January 2001 and ended in April 2016. To check the robustness of our results to variations in the out-of-sample period, we vary the out-of-sample start date to be either January 2002, January 2003, ..., or January 2012. This ensures we have subsamples that both include and exclude the 2007–2008 financial crisis. We do not consider a start date beyond 2012 because this would result in an insufficient number of out-of-sample observations.

Using the MSE, QLIKE, and realized utilities as measures of forecast accuracy, we find that the pattern observed over the full out-of-sample period is retained in each of the subsamples; the BC-CIV1 provides the most accurate forecasts according to the MSE while the HAR-CIV1 forecasts are the most accurate according to the QLIKE and have the highest realized utility. Therefore, our results appear to be robust to the choice of different out-of-sample periods.

6.4 | Alternative bias correction procedure

We have also conducted robustness checks with respect to our bias correction technique. Thus far, the bias in OIVs has been corrected for by generating out-of-sample forecasts through either univariate (Equation (11)) or augmented (Equation (12)) regression models. We now examine an alternative technique inspired by Prokopczuk and Simen (2014), who show that the forecasting performance of the MFIV can be improved significantly by making a non-parametric adjustment for the variance risk premium. We apply their technique to all our OIVs. Since this method relies on averages of ratios, we will refer to it as a relative bias-correction.

More precisely, to implement the relative bias-correction, the average ratio of OIV to RV must be computed to get an estimate of the relative bias, which, in our case, is estimated using a window of τ monthly nonoverlapping observations,

$$RB_t^{IV} = \frac{1}{\tau - 1} \sum_{s=1}^{\tau-1} \frac{OIV_{t-(22s), t-(22s)+22}}{RV_{t-(22s)+1, t-(22s)+22}}. \quad (20)$$

The relative bias-corrected OIV can then be estimated as follows:

$$RBCIV_t = \frac{OIV_{t, t+22}}{RB_t^{IV}}. \quad (21)$$

²⁷We do not use an estimation window below 60 observations since this results in poorly estimated parameters and recalcitrant forecasts.

We evaluate the performance of the relative bias-correction forecasts using the MSE, QLIKE and realized utility loss functions. As benchmarks, we include our best-performing (nonrelative) bias-corrected and HAR-based forecasts, that is BC-CIV1 and HAR-CIV1. To investigate the effect of the window size used in the relative bias-correction, we make comparisons for $\tau = \{6, 12, \dots, 60\}$.²⁸

When the MSE is used a ranking criterion we find that the relative bias-corrected forecasts are more accurate than the BC-CIV1 and HAR-CIV1 forecasts, for all values of τ . Except for when $\tau = 24$, the lowest MSE is generated by a narrow corridor measure (RBC-CIV2) which is consistent with our findings that narrow CIV measures are more informative about future volatility compared to ATMIV or MFIV. Notably, under either the QLIKE or realized utility loss functions, the HAR-CIV1 forecasts provide the most accurate forecasts for all values of τ . So, overall, our main conclusions concerning the information content of narrow CIV measures and particularly CIV1 remain largely unaltered.

6.5 | Including overnight returns

Finally, we examine whether our results are robust to the inclusion of overnight returns, that is the return between the close of the market on day $t - 1$ and the open of the market on day t , in the computation of RV. Thus far, RV has been computed using intraday returns, meaning that our RV estimates exclude volatility accumulated overnight, which we denote overnight IVAR. This is important for two reasons. First, our sample covers the period January 1996 to April 2016, and therefore, includes a time period before June 2006, when our crude-oil futures only traded 09:00–14:30 ET. Thus, our forecasting analysis spans a time period when the exclusion of overnight IVAR may have a material impact on our results. Second, our OIVs provide risk-neutral expectations about the total IVAR over 22-day forecast horizons. Consequently, our forecasting models that employ OIVs have not been fully utilizing the information available and our comparisons of forecast accuracy have not been for time periods that match the forecast horizon of the OIVs.

To compute RVs that estimate IVAR inclusive of the overnight period, we follow Hansen and Lunde (2005) and use three alternative RV estimators. The first is the *irregular* RV where the squared overnight return is simply added to the intraday RV measure. The second is the *scaled* RV, where an estimate of overnight IVAR is incorporated into RV by scaling-up the intraday RV. Specifically, the intraday RV is multiplied by a constant that ensures that the resulting estimator has the correct expected value (i.e., IVAR). The third estimator is the *combination* RV which is similar to the irregular RV, except that the weights applied to the squared overnight returns and the intraday RV are allowed to vary from unity. In particular, Hansen and Lunde (2005) demonstrate how to select these weights in an optimal way, that is compute the weights that minimize the variance of the estimated RV measure while keeping its expected value equal to the total IVAR.

To evaluate the robustness of our results to the inclusion of overnight returns, we repeat our forecasting exercises using the three alternative IVAR estimators described above. We find that forecasts that employ CIV1 in their construction deliver the most accurate forecasts irrespective of whether IVAR is estimated using the scaled RV or combination RV estimator. This is because these CIV1-based models generate the lowest MSE and QLIKE losses, as well as the highest realized utility. Again, these results support our conclusion that using CIV1 leads to superior volatility forecasts.

When the irregular RV is used, the HAR-ATMIV and HAR-MFIV appear to deliver the most accurate forecasts under the QLIKE and realized utility loss functions, respectively. However, Ahoniemi and Lanne (2013) provide empirical evidence that both the scaled RV and combination RV provide more reliable results when selecting among competing volatility forecasts due to the higher level of noise in the irregular RV. Therefore, we do not believe the results using the irregular RV provide sufficiently strong evidence to contradict our conclusion that using CIV1 leads to the most accurate volatility forecasts.

7 | CONCLUSION

In this paper, we evaluated, using both economic and statistical criteria, the information content of monthly crude-oil volatility forecasts extracted from the prices of traded options. We examined a variety of alternative option-implied

²⁸These effectively correspond to window sizes of 0.5, 1, 1.5, ..., 5 years.

measures including Black–Scholes ATMIV, model-free volatility expectations (MFIV), CBOE's "oil-VIX" (OVX) and, notably, CIV. Besides stand-alone comparisons, option-implied forecasts were also contrasted, and combined, with those obtained by a realized volatility model (HAR) that utilizes high-frequency return information.

Our key finding is that a particular CIV measure (CIV1), that utilizes a narrow range of option contracts, consistently generates the most accurate forecasts compared to all other alternatives. In Mincer–Zarnowitz regressions, CIV1 achieves the highest R^2 , while encompassing regression tests show that CIV1 subsumes the information contained in ATMIV, MFIV, OVX, and HAR forecasts. Furthermore, under either a symmetric (MSE) or an asymmetric (QLIKE) loss function, CIV1-based forecasts deliver the lowest forecast errors. In terms of economic significance, incorporating the CIV1 into the HAR model leads to forecasts that generate a significantly higher realized utility, even when transaction costs are taken into account. All these findings remain intact for both our full sample (1996–2016) and subsample (2007–2016) that only contains data after the commodities' "financialization" period.

Our results also provide valuable insights regarding the information content of the OVX index. In terms of predicting crude-oil volatility, we find that the OVX-based forecasts perform rather poorly. Regression-based tests show that the OVX is encompassed by CIV1, while OVX-based forecasts are typically the least accurate according to either statistical or economic loss functions. To the best of our knowledge, this is the first time the reliability of the OVX measure has been scrutinized, so the concerns we raise have direct implications for practitioners who often rely on the CBOE's volatility indices.

Finally, we repeat the same empirical exercise for the case of the USO ETF volatility (2007–2016). Since the USO ETF attempts to track the price of crude-oil and its options underlie the construction of the OVX index, it becomes possible to further scrutinize our results. Notably, find that the same CIV1 measure continues to generate the most accurate volatility forecasts compared to any other alternative. This underscores the robustness of our findings. The performance of the OVX, while somewhat improved in this data set, continues to be unimpressive. Our complementary analysis on the cut-off point algorithm applied by the CBOE suggests that it introduces noise that impedes its forecast accuracy.

Overall, this paper contributes to the academic literature that assesses the forward-looking information embedded in the prices of crude-oil options. Given that measuring crude-oil risk is of paramount importance for a variety of economic agents, our empirical study is of value for policy-makers and investors alike.

ACKNOWLEDGMENTS

The authors would like to thank Robert Webb (editor) and an anonymous referee for thoughtful recommendations on an earlier version of this paper. We would also like to thank participants at the IFABS 2019 Conference in Angers for useful comments and suggestions.

DATA AVAILABILITY STATEMENT

The crude-oil and USO ETF option data that support the findings of this study are available from Bloomberg. Intraday crude-oil futures and USO ETF price data were obtained from TickDataTM. Restrictions apply to the availability of the aforementioned data sets, which were used under license for this study.

ORCID

Leonidas Tsiaras  <http://orcid.org/0000-0002-4154-4210>

Dudley Gilder  <http://orcid.org/0000-0001-7039-762X>

REFERENCES

- Agnolucci, P. (2009). Volatility in crude-oil futures: A comparison of the predictive ability of GARCH and implied volatility models. *Energy Economics*, 31(2), 316–321. <https://doi.org/10.1016/j.eneco.2008.11.001>. <http://www.sciencedirect.com/science/article/pii/S0140988308001655>
- Ahoniemi, K., & Lanne, M. (2013). Overnight stock returns and realized volatility. *International Journal of Forecasting*, 29(4), 592–604. <https://doi.org/10.1016/j.ijforecast.2013.03.006>. <http://www.sciencedirect.com/science/article/pii/S0169207013000472>
- Andersen, T. G., Bollerslev, T., Diebold, F. X., & Labys, P. (2000). Great realizations. *RISK*, 13, 105–108.
- Andersen, T. G., Bollerslev, T., Diebold, F. X., & Labys, P. (2001). The distribution of realized exchange rate volatility. *Journal of the American Statistical Association*, 96(453), 42–55. <https://doi.org/10.1198/016214501750332965>
- Andersen, T. G., Bollerslev, T., Diebold, F. X., & Labys, P. (2003). Modeling and forecasting realized volatility. *Econometrica*, 71(2), 579–625. <https://doi.org/10.1111/1468-0262.00418>
- Andersen, T. G., & Bondarenko, O. (2007). Construction and interpretation of model-free implied volatility. In I. Nelken (Ed.), *Volatility as an asset class* (pp. 141–181). London: Risk Publications.

- Andersen, T. G., Bondarenko, O., & Gonzalez-Perez, M. T. (2015). Exploring return dynamics via corridor implied volatility. *The Review of Financial Studies*, 28(10), 2902–2945. <https://doi.org/10.1093/rfs/hhv033>
- Aruoba, S. B., Diebold, F. X., & Scotti, C. (2009). Real-time measurement of business conditions. *Journal of Business & Economic Statistics*, 27(4), 417–427. <https://doi.org/10.1198/jbes.2009.07205>
- Barndorff-Nielsen, O. E., & Shephard, N. (2002). Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(2), 253–280. <https://doi.org/10.1111/1467-9868.00336>
- Barone-Adesi, G., & Whaley, R. E. (1987). Efficient analytic approximation of American option values. *The Journal of Finance*, 42(2), 301–320. <https://doi.org/10.1111/j.1540-6261.1987.tb02569.x>
- Beckers, S. (1981). Standard deviations implied in option prices as predictors of future stock price variability. *Journal of Banking and Finance*, 5(3), 363–381. [https://doi.org/10.1016/0378-4266\(81\)90032-7](https://doi.org/10.1016/0378-4266(81)90032-7)
- Bliss, R. R., & Panigirtzoglou, N. (2004). Option-implied risk aversion estimates. *The Journal of Finance*, 59(1), 407–446. <https://doi.org/10.1111/j.1540-6261.2004.00637.x>
- Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3), 307–327. [https://doi.org/10.1016/0304-4076\(86\)90063-1](https://doi.org/10.1016/0304-4076(86)90063-1)
- Bollerslev, T., Hood, B., Huss, J., & Pedersen, L. H. (2018). Risk everywhere: Modeling and managing volatility. *Review of Financial Studies*, 31, 2729–2773.
- Bollerslev, T., Patton, A. J., & Quaedvlieg, R. (2016). Exploiting the errors: A simple approach for improved volatility forecasting. *Journal of Econometrics*, 192(1), 1–18. <https://doi.org/10.1016/j.jeconom.2015.10.007>. <http://www.sciencedirect.com/science/article/pii/S0304407615002584>
- Bollerslev, T., Tauchen, G., & Zhou, H. (2009). Expected stock returns and variance risk premia. *The Review of Financial Studies*, 22(11), 4463–4492. <https://doi.org/10.1093/rfs/hhp008>
- Bookstaber, R. M., & McDonald, J. B. (1987). A general distribution for describing security price returns. *Journal of Business*, 60(3), 401–424.
- Britten-Jones, M., & Neuberger, A. (2000). Option prices, implied price processes, and stochastic volatility. *The Journal of Finance*, 55(2), 839–866. <https://doi.org/10.1111/0022-1082.00228>
- Busch, T., Christensen, B. J., & Nielsen, M. Ø. (2011). The role of implied volatility in forecasting future realized volatility and jumps in foreign exchange, stock, and bond markets. *Journal of Econometrics*, 160(1), 48–57. <https://doi.org/10.1016/j.jeconom.2010.03.014>. <http://www.sciencedirect.com/science/article/pii/S0304407610000564>
- Caldeira, J. F., Moura, G. V., Nogales, F. J., & Santos, A. A. P. (2017). Combining multivariate volatility forecasts: An economic-based approach. *Journal of Financial Econometrics*, 15(2), 247–285. <https://doi.org/10.1093/jjfinec/nbw010>
- Carr, P., Lee, R., & Wu, L. (2012). Variance swaps on time-changed Lévy processes. *Finance and Stochastics*, 16(2), 335–355. <https://doi.org/10.1007/s00780-011-0157-9>
- Carr, P., & Madan, D. (1998). Towards a theory of volatility trading. *Volatility: New Estimation Techniques for Pricing Derivatives*, 29, 417–427.
- Carr, P., & Wu, L. (2009). Variance risk premiums. *The Review of Financial Studies*, 22(3), 1311–1341. <https://doi.org/10.1093/rfs/hhn038>
- Chiras, D. P., & Manaster, S. (1978). The information content of option prices and a test of market efficiency. *Journal of Financial Economics*, 6(2), 213–234. [https://doi.org/10.1016/0304-405X\(78\)90030-2](https://doi.org/10.1016/0304-405X(78)90030-2). <http://www.sciencedirect.com/science/article/pii/0304405X78900302>
- Christoffersen, P., Jacobs, K., & Chang, B. Y. (2013). Chapter 10—Forecasting with option-implied information. In G. Elliott & A. Timmermann (Eds.), *Handbook of economic forecasting* (pp. 581–656). Amsterdam: Elsevier.
- Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2), 174–196. <https://doi.org/10.1093/jjfinec/nbp001>
- Crude-oil Futures Options. Bloomberg; 1996–2016. Retrieved from www.Bloomberg.com
- Day, T. E., & Lewis, C. M. (1993). Forecasting futures market volatility. *Journal of Derivatives*, 1(2), 33–50. <https://doi.org/10.3905/jod.1993.407876>. <http://jod.ijournals.com/content/1/2/33>
- Demeterfi, K., Derman, E., Kamal, M., & Zou, J. (1999). A guide to volatility and variance swaps. *The Journal of Derivatives*, 6(4), 9–32. <https://doi.org/10.3905/jod.1999.319129>. <http://jod.ijournals.com/content/6/4/9>
- Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, 13, 253–263. <https://doi.org/10.1198/073500102753410444>. <https://amstat.tandfonline.com/doi/abs/10.1198/073500102753410444#.WzUZO4onZtQ>
- Dotsis, G., & Vlastakis, N. (2016). Corridor volatility risk and expected returns. *Journal of Futures Markets*, 36(5), 488–505. <https://doi.org/10.1002/fut.21738>
- Feunou, B., Jahan-Parvar, M. R., & Okou, C. (2017). Downside variance risk premium. *Journal of Financial Econometrics*, 16(3), 341–383. <https://doi.org/10.1093/jjfinec/nbx020>
- Fung, H., Lie, C., & Moreno, A. (1990). The forecasting performance of the implied standard deviation in currency options. *Managerial Finance*, 16(3), 24–29. <https://doi.org/10.1108/eb013646>
- Gemmell, G. (1986). The forecasting performance of stock options on the London traded options market. *Journal of Business Finance and Accounting*, 13(4), 535–546. <https://doi.org/10.1111/j.1468-5957.1986.tb00516.x>
- Griffin, J. M., & Shams, A. (2018). Manipulation in the VIX? *The Review of Financial Studies*, 31(4), 1377–1417. <https://doi.org/10.1093/rfs/hhx085>
- Hansen, P. R., & Lunde, A. (2005). A realized variance for the whole day based on intermittent high-frequency data. *Journal of Financial Econometrics*, 3(4), 525–554. <https://doi.org/10.1093/jjfinec/nbi028>

- Haugom, E., Langeland, H., Molnár, P., & Westgaard, S. (2014). Forecasting volatility of the U.S. oil market. *Journal of Banking and Finance*, 47, 1–14. <https://doi.org/10.1016/j.jbankfin.2014.05.026>. <http://www.sciencedirect.com/science/article/pii/S0378426614001939>
- Henderson, B. J., Pearson, N. D., & Wang, L. (2014). New evidence on the financialization of commodity markets. *The Review of Financial Studies*, 28(5), 1285–1311. <https://doi.org/10.1093/rfs/hhu091>
- Intraday Crude-oil Futures Prices. TickDataTM; 1996–2016. Retrieved from www.tickdata.com
- Intraday USO ETF Prices. TickDataTM; 2007–2016. Retrieved from www.tickdata.com
- Jiang, G. J., & Tian, Y. S. (2005). The model-free implied volatility and its information content. *Review of Financial Studies*, 18(4), 1305–1342. Retrieved from <https://academic.oup.com/rfs/article/18/4/1305/1595745>. <https://doi.org/10.1093/rfs/hhi027>
- Kang, S. H., Kang, S.-M., & Yoon, S.-M. (2009). Forecasting volatility of crude oil markets. *Energy Economics*, 31(1), 119–125. <https://doi.org/10.1016/j.eneco.2008.09.006>. Retrieved from <http://www.sciencedirect.com/science/article/pii/S0140988308001539>
- Latané, H. A., & Rendleman, R. J. (1976). Standard deviations of stock price ratios implied in option prices. *The Journal of Finance*, 31(2), 369–381. <https://doi.org/10.1111/j.1540-6261.1976.tb01892.x>
- Liu, L. Y., Patton, A. J., & Sheppard, K. (2015). Does anything beat 5-minute RV? A comparison of realized measures across multiple asset classes. *Journal of Econometrics*, 187(1), 293–311. <https://doi.org/10.1016/j.jeconom.2015.02.008>. <http://www.sciencedirect.com/science/article/pii/S0304407615000329>
- Locke, P. R., & Venkatesh, P. C. (1997). Futures market transaction costs. *Journal of Futures Markets*, 17(2), 229–245.
- Martens, M., & Zein, J. (2004). Predicting financial volatility: High-frequency time-series forecasts vis-à-vis implied volatility. *Journal of Futures Markets*, 24(11), 1005–1028. <https://doi.org/10.1002/fut.20126>
- Meddahi, N. (2002). A theoretical comparison between integrated and realized volatility. *Journal of Applied Econometrics*, 17(5), 479–508. <https://doi.org/10.1002/jae.689>
- Melick, W. R., & Thomas, C. P. (1997). Recovering an asset's implied pdf from option prices: An application to crude oil during the gulf crisis. *Journal of Financial and Quantitative Analysis*, 32(1), 91–115. <https://doi.org/10.2307/2331318>. Retrieved from <https://www.cambridge.org/core/journals/journal-of-financial-and-quantitative-analysis/article/recovering-an-assets-implied-pdf-from-option-prices-an-application-to-crude-oil-during-the-gulf-crisis/8D2B398E4952712ABC9B9D6028FA6270>
- Mincer, J., & Zarnowitz, V. (1969). The evaluation of economic forecasts. In J. Mincer (Ed.), *Economic forecasts and expectations Economic forecasts and expectations* (pp. 3–46). Cambridge: National Bureau of Economic Research.
- Newey, W. K., & West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3), 703–708. Retrieved from <https://search.proquest.com/docview/214867593?accountid=15390>
- Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, 160(1), 246–256. <https://doi.org/10.1016/j.jeconom.2010.03.034>. Retrieved from <http://www.sciencedirect.com/science/article/pii/S030440761000076X>
- Patton, A. J., & Sheppard, K. (2009). Evaluating volatility and correlation forecasts. In T. Mikosch, J.-P. Kreiß, R. A. Davis & T. G. Andersen (Eds.), *Handbook of financial time series* (pp. 801–838). Berlin, Heidelberg: Springer.
- Patton, A. J., & Sheppard, K. (2015). Good volatility, bad volatility: Signed jumps and the persistence of volatility. *Review of Economics and Statistics*, 97(3), 683–697. https://doi.org/10.1162/REST_a_00503
- Prokopczuk, M., & Simen, C. W. (2014). The importance of the volatility risk premium for volatility forecasting. *Journal of Banking and Finance*, 40, 303–320. <https://doi.org/10.1016/j.jbankfin.2013.12.002>. <https://www.sciencedirect.com/science/article/pii/S0378426613004640>
- Prokopczuk, M., Symeonidis, L., & Wese, S. C. (2016). Do jumps matter for volatility forecasting? Evidence from energy markets. *Journal of Futures Markets*, 36(8), 758–792. <https://doi.org/10.1002/fut.21759>
- Sadorsky, P. (2006). Modeling and forecasting petroleum futures volatility. *Energy Economics*, 28(4), 467–488. <https://doi.org/10.1016/j.eneco.2006.04.005>. <http://www.sciencedirect.com/science/article/pii/S014098830600048X>
- Sévi, B. (2014). Forecasting the volatility of crude oil futures using intraday data. *European Journal of Operational Research*, 235(3), 643–659. <https://doi.org/10.1016/j.ejor.2014.01.019>. <http://www.sciencedirect.com/science/article/pii/S037722171400040X>
- Singleton, K. J. (2014). Investor flows and the 2008 boom/bust in oil prices. *Management Science*, 60(2), 300–318. <https://doi.org/10.1287/mnsc.2013.1756>
- Szakmary, A., Ors, E., Kim, J. K., & Davidson, W. N. (2003). The predictive power of implied volatility: Evidence from 35 futures markets. *Journal of Banking and Finance*, 27(11), 2151–2175. [https://doi.org/10.1016/S0378-4266\(02\)00323-0](https://doi.org/10.1016/S0378-4266(02)00323-0). <http://www.sciencedirect.com/science/article/pii/S0378426602003230>
- Taylor, S. J. (2005). *Asset price dynamics, volatility, and prediction*. Princeton, NJ: Princeton University Press.
- Taylor, S. J., Yadav, P. K., & Zhang, Y. (2010). The information content of implied volatilities and model-free volatility expectations: Evidence from options written on individual stocks. *Journal of Banking and Finance*, 34(4), 871–881. <https://doi.org/10.1016/j.jbankfin.2009.09.015>. <https://www.sciencedirect.com/science/article/pii/S0378426609002489>
- Todorov, V. (2009). Variance risk-premium dynamics: The role of jumps. *The Review of Financial Studies*, 23(1), 345–383. <https://doi.org/10.1093/rfs/hhp035>
- Trippi, R. (1977). A test of option market efficiency using a random-walk valuation model. *Journal of Economics and Business*, 29, 93–98.
- USO ETF Options. Bloomberg; 2007–2016. Retrieved from www.Bloomberg.com
- Wang, Y., Liu, L., Ma, F., & Wu, C. (2016). What the investors need to know about forecasting oil futures return volatility. *Energy Economics*, 57, 128–139. <https://doi.org/10.1016/j.eneco.2016.05.004>. <http://www.sciencedirect.com/science/article/pii/S0140988316301049>

Zhang, L., Mykland, P. A., & Aït-Sahalia, Y. (2005). A tale of two time scales. *Journal of the American Statistical Association*, 100(472), 1394–1411. <https://doi.org/10.1198/016214505000000169>

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Gilder D, Tsiaras L. Volatility forecasts embedded in the prices of crude-oil options. *J Futures Markets*. 2020;1–33. <https://doi.org/10.1002/fut.22114>