Conditional threshold effects of stock market volatility on crude oil market volatility

Kaiji Motegi* and Shigeyuki Hamori†
Kobe University and Kobe University
July 17, 2023

Abstract
This paper analyzes conditional threshold effects of stock market volatility on crude oil market volatility. We use the conditional threshold autoregressive (CoTAR) model, a novel extension of TAR from a constant to time-varying threshold. The conditional threshold is specified as an empirical quantile of recent realizations of a threshold variable. CoTAR should fit asset return volatilities well, as investors’ perception on high versus low volatilities likely depends on time. We show via Monte Carlo simulations that the out-of-sample forecast accuracies of TAR and CoTAR can be correctly compared by the Diebold-Mariano test. Our empirical study reveals that CoTAR is on par with TAR for most cases and significantly better for some cases in terms of predictive ability. CoTAR often outperforms TAR when predicting a downside volatility measure; it is a useful finding which helps market participants and policymakers better manage downward risks.

JEL codes: C22, C53, C58.

Keywords: Conditional Threshold Autoregression (CoTAR), Diebold-Mariano test, good and bad volatilities, out-of-sample forecast, realized volatility (RV).

---

*Corresponding author: Graduate School of Economics, Kobe University. 2-1 Rokkodai-cho, Nada, Kobe, Hyogo 657-8501 Japan. E-mail: motegi@econ.kobe-u.ac.jp
†Graduate School of Economics, Kobe University. 2-1 Rokkodai-cho, Nada, Kobe, Hyogo 657-8501 Japan. E-mail: hamori@econ.kobe-u.ac.jp
1 Introduction

Measuring, modelling, and predicting the volatility of financial asset returns have been a central goal for econometricians, market participants, and policymakers. Realized volatility (RV) measures, which were originally put forward by Andersen and Bollerslev (1998), are often chosen as an objective variable of time series models and forecasting methods. RV is defined as the sum of squared high frequency returns of a target asset. RV is easy to compute from data, as far as asset returns are observed at a higher frequency than the target frequency. Besides, RV is guaranteed to well approximate the true volatility of asset returns given certain regularity conditions. Numerous researchers have been analyzing and extending RV in various directions; for example, Barndorff-Nielsen, Kinnebrock, and Shephard (2010) proposed realized semi-variances to distinguish “good volatility” and “bad volatility”.\footnote{Andersen and Benzoni (2009) provide an excellent overview and survey of RV. More recent developments of RV are covered by Ghysels and Marcellino (2018, Ch.14) among others.}

Crude oil is one of the most actively traded commodities in the world and a major source of the global energy supply. Modelling and forecasting crude oil RVs were first pursued by several teams; for example, Haugom, Langeland, Molnár, and Westgaard (2014) and Wen, Gong, and Cai (2016) use Corsi’s (2009) heterogeneous autoregressive (HAR) model and its variants to predict RVs of the West Texas Intermediate (WTI) futures market. They found that HAR-type models fit crude oil RVs well, inspiring many follow-up studies. Ma, Wei, Liu, and Huang (2018) is one of the earliest work that distinguished the “good” volatility and “bad” volatility of crude oil markets, using realized semi-variances. Their empirical results suggest that downside volatility has a stronger impact on the oil futures market volatility than upside volatility does.\footnote{Degiannakis and Filis (2017), Ma, Zhang, Huang, and Lai (2018), Luo, Ji, Klein, Todorova, and Zhang (2020), Gong and Lin (2021), and Yang, Wei, Li, Liu, and Wang (2021) also performed HAR-based studies on crude oil RVs.}

There is an insufficient amount of research on the regime-dependent nature of crude oil RVs. In general, financial indicators often exhibit heterogeneous properties across “low” and “high” volatility regimes. A potential source for the heterogeneity is that market participants behave differently during tranquil and crisis periods. Such an asymmetric pattern may well exist in crude oil markets, and Ma, Wahab, Huang, and Xu (2017) explored this possibility. They introduced Markov regime switching to the HAR-type analysis of crude oil RVs, documenting that the inclusion of the regime switching mechanism significantly improves out-of-sample performance.

Chen, Qiao, and Zhang (2022, CQZ2022) also focused on the regime-dependent...
characteristics of crude oil RVs. They used the threshold autoregressive (TAR) model of Tong (1978) to examine if crude oil RVs have heterogeneous properties when stock markets are in a low versus high volatility regime. CQZ2022 detected significant threshold effects of stock market RVs on crude oil RVs, implying that the crude oil volatility evolves asymmetrically when the overall financial market risk (measured by the stock market RVs) is low versus high. In addition, CQZ2022 showed that TAR achieves significantly high out-of-sample performance than benchmark AR models.

In TAR models, a target variable $y$ has heterogeneous structures when a threshold variable $x$ is below versus above a constant threshold $\mu$. Recently, TAR and related models have been extended in various ways to allow for a time-varying threshold $\mu_t$. First, thresholds are specified as a linear combination of observed covariates in Seo and Linton (2007), Dueker, Psaradakis, Sola, and Spagnolo (2013), and Yu and Fan (2021), among others. Second, Zhu, Chen, and Lin (2019) took a state space approach such that $\mu_t$ follows a latent AR process. Third, Yang, Lee, and Chen (2021) applied a Fourier approximation to $\mu_t$. Fourth, Motegi, Cai, Hamori, and Xu (2020) proposed the moving average threshold heterogeneous autoregression (MAT-HAR) by adding time-varying thresholds to each term of HAR, where the thresholds are specified as sample moving averages of $y$ aggregated at each sampling frequency.

Time-varying threshold models are expected to fit RV measures well, as a cut-off level of “low” and “high” volatilities may well change over time. Consider an RV shock of a certain magnitude; it might surprise investors during a tranquil period but might be nothing surprising during a crisis period, judging from recent observations of the RV measure. Indeed, Salisu, Gupta, and Ogbonna (2022) found via the MAT-HAR models that monthly RVs of U.S. stock markets have significant time-varying threshold effects. Thus, it is likely that fitting time-varying threshold models to crude oil RVs leads to significant improvement in in-sample and out-of-sample performance. The present paper fills this gap for the first time in the literature.

We use Conditional Threshold Autoregressive (CoTAR) models proposed by Motegi, Dennis, and Hamori (2022, MDH2022). In CoTAR models, the threshold is specified as an empirical quantile of recent observations of a threshold variable $x$. The conditional threshold $\mu_t$ represents a “normal” level of recent $x$ if a middle quantile is picked, and an “abnormal” level if a lower or upper quantile is picked. MDH2022 has demonstrated that their proposed methodology attains desired properties in both large and finite samples. An empirical study of MDH2022 revealed that there are significant conditional threshold effects in daily new confirmed cases of COVID-19 in
Japan. A virtue of CoTAR compared with other time-varying threshold models lies in its computational simplicity, as the conditional threshold can be computed from data without relying on advanced methods such as state space representation.

In our empirical study, the target variable $y$ is monthly RV measures of WTI spot returns and the threshold variable $x$ is monthly RV measures of S&P 500 Index (SPX) returns. These choices are identical to CQZ2022, hence our work is a novel extension of CQZ2022 from a constant threshold to a conditional threshold. Another contribution of ours is that we inspect the out-of-sample forecast performance of CoTAR, which has not been done in the existing literature. We show via Monte Carlo simulations that the well-known tests of Diebold and Mariano (1995, DM1995) operate well when we compare the relative predictive accuracies of TAR and CoTAR.

Our main empirical results are as follows. First, the threshold models (i.e., TAR and CoTAR) attain higher predictive power for crude oil RVs than benchmark non-threshold models, which highlights the importance of capturing threshold effects. Second, the DM test indicates that CoTAR is comparable with TAR for most cases and better for some cases in terms of forecasting crude oil RVs. In particular, we observe fairly strong evidence that CoTAR outperforms TAR when forecasting bad volatility. It is a useful finding which helps market participants and policymakers better monitor and control downside risks.

The remainder of the paper is organized as follows. In Section 2, the TAR and CoTAR models are presented and compared. In Section 3, we conduct Monte Carlo simulations to investigate the relative out-of-sample performance of the two models. In Section 4, we analyze threshold effects of stock market RVs on crude oil market RVs. Some concluding remarks are provided in Section 5.

We use the following notation throughout the paper. $\mathbb{R}$ is the set of real numbers. $\mathbb{N}$ is the set of natural numbers. $\lfloor a \rfloor$ is the largest integer not larger than $a \in \mathbb{R}$. The Euclidean norm of any $k$-dimensional vector $c \in \mathbb{R}^k$ is denoted as $\|c\| = (c^\top c)^{1/2}$. $1(A)$ is the indicator function which equals 1 if event $A$ occurs and 0 otherwise. $\#A$ is the number of elements of set $A$. $A \times B$ is the Cartesian product of sets $A$ and $B$.

## 2 TAR and CoTAR models

In this section, we present and compare the TAR and CoTAR models. We describe model specifications in Section 2.1, parameter estimation in Section 2.2, testing the null hypothesis of no threshold effects in Section 2.3, and comparing out-of-sample
2.1 Model specifications

Let $y_t$ and $x_t$ be a target variable and a threshold variable at time $t \in \{1, \ldots, n\}$, respectively. Tong’s (1978) TAR model with two regimes is specified as follows.

$$y_t = \begin{cases} 
\alpha_1 + \sum_{k=1}^{p} \phi_{1k} y_{t-k} + u_t & \text{if } x_{t-d} < \mu, \\
\alpha_2 + \sum_{k=1}^{p} \phi_{2k} y_{t-k} + u_t & \text{if } x_{t-d} \geq \mu,
\end{cases}$$

(1)

where $(\alpha_r, \phi_{r1}, \ldots, \phi_{rp}) \in \mathbb{R}^{p+1}$ are regression parameters in regime $r \in \{1, 2\}$; $d \in \mathbb{N}$ is the delay parameter; $\mu \in \mathbb{R}$ is the threshold parameter; $u_t$ is the error term.

A key feature of (1) is that $y$ has different autocorrelation structures below versus above the unconditional threshold $\mu$. The term “unconditional” means that $\mu$ is time-independent and chosen from the entire memory $X^n_t = \{x_1, \ldots, x_n\}$. The constant threshold $\mu$ might not be suitable for RV analyses since investors’ perception on low versus high volatilities likely depends on time.

Based on (1), MDH2022 proposed the CoTAR model by replacing $\mu$ with a conditional threshold $\mu_t$ which is time-dependent and chosen from a local memory $X^{t}_{t-m+1} = \{x_{t-m+1}, \ldots, x_t\}$:

$$y_t = \begin{cases} 
\alpha_1 + \sum_{k=1}^{p} \phi_{1k} y_{t-k} + u_t & \text{if } x_{t-d} < \mu_{t-d-1}(c), \\
\alpha_2 + \sum_{k=1}^{p} \phi_{2k} y_{t-k} + u_t & \text{if } x_{t-d} \geq \mu_{t-d-1}(c).
\end{cases}$$

(2)

The conditional threshold $\mu_{t}(c)$ is the $mc$-th smallest value (i.e., the 100$c$-percentile) of $X^{t}_{t-m+1}$; $m = \#X^{t}_{t-m+1}$ is the size of the local memory; $c \in \{1/m, 2/m, \ldots, 1\}$ signifies the relevant percentile, where the possible values $c$ can take are restricted to be discrete so that $mc \in \{1, \ldots, m\}$.

A unique feature of CoTAR lies in the specification of the conditional threshold $\mu_{t-d-1}(c)$. If $c = (2m)^{-1}(m+1)$, then $\mu_{t-d-1}(c)$ almost coincides with the median of $X^{t}_{t-d-m}$. In this case, regime 1 arises when $x$ is below the “normal” level given the local memory, while regime 2 arises when $x$ is above it. If $c$ is close to the lower bound $1/m$ or the upper bound 1, then a regime switch is triggered by a rare event of $x$ crossing an “abnormal” level given the local memory. Consider, for example, a CoTAR model with $m = 12$ and $c = 9/12 = 0.75$, where the target variable $y$ is a monthly log-RV of WTI and the threshold variable $x$ is a monthly log-RV of SPX. Then, the conditional
threshold is the 75-percentile of the recent 12-month observations of the log-RV of SPX, hence regime 2 represents periods of excessive volatility in stock markets (in the relative sense). The conditional threshold approach is intuitively reasonable, as investors likely assess the current status of financial markets relative to the recent past, not to a constant cut-off value.\(^3\)

### 2.2 Profiling estimation of parameters

To review the estimation of the TAR model, stack the parameters of (1) as follows.

\[
\begin{align*}
\beta_1 &= \begin{bmatrix} \alpha_1 \\ \phi_{11} \\ \vdots \\ \phi_{1p} \end{bmatrix}, \\
\beta_2 &= \begin{bmatrix} \alpha_2 \\ \phi_{21} \\ \vdots \\ \phi_{2p} \end{bmatrix}, \\
\beta &= \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}, \\
\gamma &= \begin{bmatrix} d \\ \mu \end{bmatrix}, \\
\theta &= \begin{bmatrix} \beta \\ \gamma \end{bmatrix}.
\end{align*}
\]

The entire parameter vector \(\theta\) is partitioned into the regression parameters \(\beta\) and the nuisance parameters \(\gamma\). To focus on the estimation of \(\theta\), we sidestep lag selection issues by assuming that \(p\) is known.

We can estimate \(\beta\) and \(\gamma\) via profiling, a well-known two-step approach which proceeds as follows. The choice set of the delay parameter \(d\) is \(\mathcal{D} = \{1, 2, \ldots, \bar{d}\}\), where the upper bound \(\bar{d} \in \mathbb{N}\) is pre-specified by the researcher. The choice set of the threshold parameter \(\mu\) is \(\mathcal{X} = \{x_t\}_{t=1}^n\). Let \(\bar{\Gamma} = \mathcal{D} \times \mathcal{X}\) be the entire choice set of \(\gamma\), and let \(\delta_r(\gamma)\) be the share of regime \(r \in \{1, 2\}\) relative to the whole sample given \(\gamma\). For some \(\gamma \in \bar{\Gamma}\), \(\delta_r(\gamma)\) may be too small to identify both regimes in finite samples. A practical compromise often made in the TAR literature is to restrict \(\bar{\Gamma}\) so that both regimes account for at least 15% of the entire sample:

\[
\Gamma = \{\gamma \in \bar{\Gamma} | \min\{\delta_1(\gamma), \delta_2(\gamma)\} > 0.15\}.
\]

For each fixed \(\gamma \in \Gamma\), the conditional least squares estimator for \(\beta\), denoted as \(\hat{\beta}(\gamma)\), is available in closed form. An optimal \(\gamma \in \Gamma\) that minimizes the sum of squared

---

\(^3\)One might be tempted to specify the conditional threshold as a conditional average of \(x\) (i.e., \(\mu_t = m^{-1} \sum_{t=0}^{m-1} x_{t-t}\)); indeed, Motegi, Cai, Hamori, and Xu (2020) took this approach to construct the MAT-HAR models. This is an intuitively plausible generalization of HAR, and there is a computational advantage that the percentile parameter \(c\) disappears. A possible limitation of the conditional average specification, however, is that the threshold level cannot be an “abnormal” level unlike the conditional quantile specification with \(c\) being away from 0.5.
errors is chosen as the profiling estimator for $\gamma$, and it is denoted as $\hat{\gamma}$. The profiling estimator for $\beta$ is given by $\hat{\beta} = \hat{\beta}(\hat{\gamma})$. Chan (1993) derived the asymptotic properties of $\hat{\theta} = (\hat{\beta}^\top, \hat{\gamma}^\top)^\top$ under some regularity conditions.

The parameters of the CoTAR model (2) can be estimated analogously. Assume as in MDH2022 that the memory size $m$ is known, and replace $\mu$ with $c$. The choice set of the percentile parameter $c$ is $C = \{1/m, 2/m, \ldots, 1\}$. The parameter space $\Gamma = \mathcal{D} \times C$ is restricted by (4) so that both regimes occur sufficiently many times. The remaining procedure is analogous to TAR. MDH2022 derived the asymptotic properties of the profiling estimator for CoTAR, following Chan’s (1993) derivation.

2.3 Testing the no-threshold-effect hypothesis

Given either the TAR model (1) or the CoTAR model (2), there are no threshold effects if and only if $H_0^* : \beta_1 = \beta_2$, where $\beta_r$ is defined in (3). Under $H_0^*$, the persistence structure of $y$ is identical for the two regimes, and TAR and CoTAR reduce to a single regime AR($p$). Hence, it is crucial to test the no-threshold-effect hypothesis $H_0^*$.

This is a non-standard testing problem, since the nuisance parameters $\gamma$ are not identified under $H_0^*$. Hansen (1996) proposed a novel solution of the wild-bootstrap test for TAR, and MDH2022 took the same approach for CoTAR.

The wild-bootstrap test for the no-threshold-effect hypothesis $H_0^*$ proceeds as follows. For each fixed value of nuisance parameters $\gamma \in \Gamma$, compute the Wald test statistic or Lagrange Multiplier (LM) test statistic associated with $H_0^*$. Transform the conditional Wald or LM test statistics across all possible values of $\gamma \in \Gamma$ into a single test statistic, where common transformations include supremum, average, and exponential. This test statistic is compared with wild-bootstrap test statistics which are computed under $H_0^*$. If the resulting bootstrap p-value (i.e., the frequency of the bootstrap test statistics exceeding the actual test statistic) is smaller than a nominal size $a \in (0, 1)$, then reject $H_0^*$ at the 100$a\%$ level. We refrain from presenting detailed procedures of the wild-bootstrap test, as they are well established in the literature (e.g., Hansen, 1996, Motegi, Dennis, and Hamori, 2022, Motegi and Dennis, 2022).

In the TAR framework, Hansen (1996) demonstrated that wild-bootstrap test for $H_0^*$ attains asymptotic validity under some regularity conditions. Similarly, MDH2022 showed that the CoTAR-based wild-bootstrap test for $H_0^*$ is asymptotically valid. Further, MDH2022 showed via Monte Carlo simulations that their test has accurate empirical size under $H_0^*$ and high empirical power under $H_1^* : \beta_1 \neq \beta_2$.

We have several remarks on the wild-bootstrap test for the no-threshold-effect
hypothesis $H^*_0$. First, this test is an in-sample analysis, hence it is not necessarily synonymous with the out-of-sample performance of TAR or CoTAR. Second, this test does not directly compare TAR and CoTAR. Testing $H^*_0$ given (1) amounts to evaluating the relative validity of no threshold effects (AR) versus constant threshold effects (TAR). Similarly, testing $H^*_0$ given (2) amounts to evaluating the relative validity of no threshold effects (AR) versus conditional threshold effects (CoTAR). These facts motivate a direct comparison of the out-of-sample forecast accuracies of TAR and CoTAR. Such a study is missing in the literature as CoTAR was proposed only recently. We fill this gap by exploiting the Diebold-Mariano tests.

2.4 Comparing out-of-sample forecast accuracies

Consider either the TAR model (1) or the CoTAR model (2). We perform the rolling window out-of-sample prediction based on observed $\{y_t, x_t\}_{t=1}^N$, where $N \in \mathbb{N}$ is the entire sample size. First, fit the model to $\{y_t\}_{t=1}^n$ and compute the profiling estimator as described in Section 2.2. The window size is fixed at $n = \lceil \tau N \rceil$ with $\tau \in (0, 1)$. Second, compute the one-step-ahead forecast of $y_{n+1}$, denoted as $\hat{y}_{n+1}$. Analogously, fit the model to $\{y_t\}_{t=n+2}^{n+T}$ and compute the one-step-ahead forecast of $y_{n+2}$, denoted as $\hat{y}_{n+2}$. Stop once $\{\hat{y}_t\}_{t=n+1}^{n+T}$ is obtained, where $T = N - n$ is the number of windows used for forecast evaluation.\(^4\) The forecast error is given by $\hat{e}_t = y_t - \hat{y}_t$ for $t \in \{n+1, \ldots, n+T\}$. The mean squared forecast error (MSE) is defined as $MSE_T = T^{-1} \sum_{t=n+1}^{n+T} \hat{e}_t^2$, and the root mean squared forecast error (RMSE) is defined as $RMSE_T = MSE_T^{1/2}$.

A simple way to compare the predictive accuracies of TAR and CoTAR is to compare their (R)MSEs. Further, DM1995 established a novel approach to test if the spread between the two MSEs is statistically significant. Compute $\hat{d}_t = (\hat{e}_t^{ TAR})^2 - (\hat{e}_t^{ CoTAR})^2$ for $t \in \{n+1, \ldots, n+T\}$, where $\hat{e}_t^{ TAR}$ and $\hat{e}_t^{ CoTAR}$ are the forecast errors associated with TAR and CoTAR, respectively. DM1995 proposed multiple ways to transform $\{\hat{d}_t\}_{t=n+1}^{n+T}$ into a single test statistic, including mean difference ($S_1$), sign ($S_2a$), and Wilcoxon’s signed rank ($S_{3a}$). The test statistic most commonly used in the literature

\(^4\)In this approach, the window size is fixed at $n$ as both the first and last time points of the window shift one by one. Another common approach is to fix the first point at $t = 1$ and shift the last point from $n$ to $N-1$, in which case the window size increases from $n$ to $N-1$. In the numerical and empirical studies of this paper, we focus on the former approach to save space. In extra studies not reported here, we find that the two approaches lead to qualitatively similar conclusions; results of the latter approach is available upon request.
is $S_1$, whose core component is as follows:

$$
\bar{d}_T = \frac{1}{T} \sum_{t=n+1}^{n+T} \hat{d}_t = MSE_{TAR}^T - MSE_{CoTAR}^T,
$$

where $MSE_{TAR}^T$ and $MSE_{CoTAR}^T$ are the MSEs of TAR and CoTAR, respectively. The $S_1$ test judges the spread between the two MSEs is statistically significant.

The null hypothesis of the DM tests, denoted as $H_0$, is that the TAR-based and CoTAR-based forecasts are as accurate as each other. DM1995 showed that, under $H_0$ and some mild regularity conditions, each test statistic follows the standard normal distribution asymptotically. We consider three alternative hypotheses separately:

- $H_1$: The two forecasts have different accuracies (two-sided test). Reject $H_0$ at the 100a% level if $|S| > \Phi^{-1}(1 - a/2)$, where $S$ is the test statistic (i.e., either $S_1$, $S_{2a}$, or $S_{3a}$) and $\Phi^{-1}(\cdot)$ is the inverse distribution function of the standard normal distribution.

- $H_{1}^{TAR}$: The TAR-based forecast is more accurate than the CoTAR-based forecast (one-sided test). Reject $H_0$ at the 100a% level if $S < \Phi^{-1}(a)$.

- $H_{1}^{CoTAR}$: The CoTAR-based forecast is more accurate than the TAR-based forecast (one-sided test). Reject $H_0$ at the 100a% level if $S > \Phi^{-1}(1 - a)$.

It is well known that the asymptotic DM tests are valid under remarkably general contexts (e.g., Diebold, 2015). In the next section, we show via Monte Carlo simulations that the DM test is indeed a compelling approach to compare the relative out-of-sample performance of TAR and CoTAR.

### 3 Monte Carlo simulations

In this section, we perform Monte Carlo simulations to compare the out-of-sample forecast performance of TAR and CoTAR. The simulation design is described in Section 3.1, and results are reported in Section 3.2.

#### 3.1 Simulation design

Assume that a data generating process (DGP) for a threshold variable $x$ is AR(1). Specifically, $x_t = \psi_0 x_{t-1} + \xi_t$, where $\psi_0 \in \{0.2, 0.8\}$ and $\xi_t \sim \mathcal{N}(0, 1)$. A DGP for a
target variable \(y\) is either AR, TAR, or CoTAR written in a unified formula:

\[
y_t = \begin{cases} 
\alpha_{10} + \phi_{10} y_{t-1} + \epsilon_t, & \text{if } x_{t-d_0} < (1-\iota)\mu_0 + \iota \mu_{t-d_0-1}(c_0), \\
\alpha_{20} + \phi_{20} y_{t-1} + \epsilon_t, & \text{if } x_{t-d_0} \geq (1-\iota)\mu_0 + \iota \mu_{t-d_0-1}(c_0),
\end{cases}
\]

(5)

where \(d_0 = 1, \mu_0 = 0, c_0 = 0.5, \iota \in \{0,1\}\), and \(\epsilon_t \overset{i.i.d.}\sim \mathcal{N}(0,1)\).

Let \(\beta_{r0} = (\alpha_{r0}, \phi_{r0})^\top\) be the true regression coefficients in regime \(r \in \{1,2\}\). When \(\beta_{10} = \beta_{20} = \beta_0\), there are no threshold effects. In this case, the threshold variable \(x\) does not affect the target variable \(y\) and (5) reduces to single-regime AR:

\[y_t = z_{t-1}^\top \beta_0 + \epsilon_t\] with \(z_{t-1} = (1, y_{t-1})^\top\). We consider two parameterizations: \(\beta_0 = (0,0.2)^\top\) (i.e., low persistence in \(y\)) and \(\beta_0 = (0,0.8)^\top\) (i.e., high persistence in \(y\)).

When \(\beta_{10} \neq \beta_{20}\), there are threshold effects under the true DGP. We consider two parameterizations:

**Weak threshold effects** \(\beta_{10} = (0,0.2)^\top\) and \(\beta_{20} = (0.6,0.6)^\top\).

**Strong threshold effects** \(\beta_{10} = (0,0.2)^\top\) and \(\beta_{20} = (0.8,0.8)^\top\).

Our terminology of “weak” and “strong” threshold effects is based on the distance between \(\beta_{10}\) and \(\beta_{20}\): \(||\beta_{10} - \beta_{20}|| \in \{0.721, 1\}\) under weak and strong threshold effects, respectively.

When there are threshold effects, \(x\) affects \(y\) by switching regimes. The role of the binary indicator \(\iota \in \{0,1\}\) in (5) is expressing TAR and CoTAR coherently. When \(\iota = 0\), (5) reduces to TAR with constant threshold \(\mu_0 = 0\). When \(\iota = 1\), (5) reduces to CoTAR with conditional threshold \(\mu_t(c_0)\) being the \(mc_0\)-th smallest value (or the 100\(c_0\)-percentile) of \(\{x_t, \ldots, x_{t-m+1}\}\), where the memory size is set to be \(m = 12\).

For monthly data, having \((m,c) = (12,0.5)\) implies that the conditional threshold is almost the median of recent 12-month observations of \(x\).

Generate a sample of \(\{y_t, x_t\}_{t=1}^N\) from each DGP considered, where the sample size is \(N \in \{180,360,540\}\). For monthly data, these sample sizes correspond to 15, 30, and 45 years, all of which are realistic values. To remove potential impacts of initial values, we generate 2\(N\) realizations and use the second half for analysis.

For each Monte Carlo sample, we fit either an TAR or CoTAR model. The TAR model is specified as in (1) with \(p = 1\):

\[
y_t = \begin{cases} 
\alpha_1 + \phi_1 y_{t-1} + u_t, & \text{if } x_{t-d} < \mu, \\
\alpha_2 + \phi_2 y_{t-1} + u_t, & \text{if } x_{t-d} \geq \mu.
\end{cases}
\]
The regression parameters $\boldsymbol{\beta} = (\alpha_1, \phi_1, \alpha_2, \phi_2)^\top$ and the nuisance parameters $\boldsymbol{\gamma} = (d, \mu)^\top$ are estimated via profiling as described in Section 2.2. The choice set of the delay parameter $d$ is set to be $\mathcal{D} = \{1, 2, 3\}$, and the restriction (4) is imposed so that both regimes appear sufficiently many times. The CoTAR model is specified as in (2) with $p = 1$:

$$
y_t = \begin{cases} 
\alpha_1 + \phi_1 y_{t-1} + u_t, & \text{if } x_{t-d} < \mu_{t-d-1}(c), \\
\alpha_2 + \phi_2 y_{t-1} + u_t, & \text{if } x_{t-d} \geq \mu_{t-d-1}(c). 
\end{cases}
$$

As in the TAR model, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma} = (d, c)^\top$ are estimated via profiling. The choice set of the delay parameter $d$ is $\mathcal{D} = \{1, 2, 3\}$, and the choice set of the percentile parameter $c$ is $\mathcal{C} = \{1/m, 2/m, \ldots, 1\}$ with $m = 12$.

For each model and Monte Carlo sample, we perform the rolling window out-of-sample prediction as described in Section 2.4. The window size is $n = \lfloor \tau N \rfloor$ with $\tau \in \{0.4, 0.6, 0.8\}$. Recall that the null hypothesis of the DM test, $H_0$, is that the TAR-based and CoTAR-based forecasts are as accurate as each other. We consider the three alternative hypotheses: $H_1$, $H_{\text{tar}}^1$, and $H_{\text{cotar}}^1$. When the true DGP is AR, we expect that $H_0$ should be rejected with probability approaching (p.a.) nominal size $a = 0.05$ as $N$ grows, whichever the alternative hypothesis is. When the true DGP is TAR, $H_0$ should be rejected in favor of $H_1$ and $H_{\text{tar}}^1$ with p.a. 1, and in favor of $H_{\text{cotar}}^1$ with p.a. 0. When the true DGP is CoTAR, $H_0$ should be rejected in favor of $H_1$ and $H_{\text{cotar}}^1$ with p.a. 1, and in favor of $H_{\text{tar}}^1$ with p.a. 0. To verify these conjectures, we compute rejection frequencies of the asymptotic DM test with mean difference statistic ($S_1$) after $J = 1000$ Monte Carlo trials.$^5$

3.2 Simulation results

The rejection frequencies under the AR, TAR, and CoTAR processes are reported in Tables 1-3, respectively. We begin with the cases where the true DGP is AR (Table 1). This DGP is included as a special case of both TAR and CoTAR models, and $H_0$ is true. The rejection frequencies should therefore be interpreted as empirical size. As expected, the empirical size reported in Table 1 is close to the nominal size $a = 0.05$ for all cases considered. Suppose for example that $N = 180$, $\phi_0 = 0.8$, $\psi_0 = 0.8$, and $\tau = 0.6$, then the empirical size is $\{0.049, 0.048, 0.048\}$ when the alternative hypotheses are $\{H_1, H_{\text{tar}}^1, H_{\text{cotar}}^1\}$, respectively. These results indicate that

$^5$In extra simulations not reported here, we found that the sign statistic ($S_{2a}$) and Wilcoxon’s signed rank statistic ($S_{3a}$) lead to much lower empirical power than $S_1$. Simulation results on $S_{2a}$ and $S_{3a}$ are available upon request.
the asymptotic DM test operates well under \( H_0 \).

Table 1: Rejection frequencies of the Diebold-Mariano test (DGP: AR)

<table>
<thead>
<tr>
<th>( \phi_0 )</th>
<th>( \psi_0 )</th>
<th>( \tau )</th>
<th>( N = 180 )</th>
<th>( N = 360 )</th>
<th>( N = 540 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 0.2 )</td>
<td>( 0.2 )</td>
<td>( 0.4 )</td>
<td>( 0.034 )</td>
<td>( 0.050 )</td>
<td>( 0.029 )</td>
</tr>
<tr>
<td>( 0.2 )</td>
<td>( 0.2 )</td>
<td>( 0.6 )</td>
<td>( 0.047 )</td>
<td>( 0.043 )</td>
<td>( 0.052 )</td>
</tr>
<tr>
<td>( 0.2 )</td>
<td>( 0.2 )</td>
<td>( 0.8 )</td>
<td>( 0.046 )</td>
<td>( 0.043 )</td>
<td>( 0.054 )</td>
</tr>
<tr>
<td>( 0.2 )</td>
<td>( 0.8 )</td>
<td>( 0.4 )</td>
<td>( 0.033 )</td>
<td>( 0.054 )</td>
<td>( 0.034 )</td>
</tr>
<tr>
<td>( 0.2 )</td>
<td>( 0.8 )</td>
<td>( 0.6 )</td>
<td>( 0.056 )</td>
<td>( 0.054 )</td>
<td>( 0.060 )</td>
</tr>
<tr>
<td>( 0.2 )</td>
<td>( 0.8 )</td>
<td>( 0.8 )</td>
<td>( 0.049 )</td>
<td>( 0.046 )</td>
<td>( 0.052 )</td>
</tr>
<tr>
<td>( 0.8 )</td>
<td>( 0.2 )</td>
<td>( 0.4 )</td>
<td>( 0.045 )</td>
<td>( 0.061 )</td>
<td>( 0.035 )</td>
</tr>
<tr>
<td>( 0.8 )</td>
<td>( 0.2 )</td>
<td>( 0.6 )</td>
<td>( 0.033 )</td>
<td>( 0.039 )</td>
<td>( 0.035 )</td>
</tr>
<tr>
<td>( 0.8 )</td>
<td>( 0.2 )</td>
<td>( 0.8 )</td>
<td>( 0.047 )</td>
<td>( 0.042 )</td>
<td>( 0.057 )</td>
</tr>
<tr>
<td>( 0.8 )</td>
<td>( 0.8 )</td>
<td>( 0.4 )</td>
<td>( 0.040 )</td>
<td>( 0.059 )</td>
<td>( 0.037 )</td>
</tr>
<tr>
<td>( 0.8 )</td>
<td>( 0.8 )</td>
<td>( 0.6 )</td>
<td>( 0.049 )</td>
<td>( 0.048 )</td>
<td>( 0.048 )</td>
</tr>
<tr>
<td>( 0.8 )</td>
<td>( 0.8 )</td>
<td>( 0.8 )</td>
<td>( 0.066 )</td>
<td>( 0.055 )</td>
<td>( 0.057 )</td>
</tr>
</tbody>
</table>

DGP for the target variable is AR: \( y_t = \phi_0 y_{t-1} + \epsilon_t \) with \( \phi_0 \in \{0.2,0.8\} \). DGP for the threshold variable: \( x_t = \psi_0 x_{t-1} + \xi_t \) with \( \psi_0 \in \{0.2,0.8\} \). Sample size: \( N \in \{180,360,540\} \). The rolling window out-of-sample prediction is performed, where the window size is \( n = \lfloor \tau N \rfloor \) with \( \tau \in \{0.4,0.6,0.8\} \). The TAR-based and CoTAR-based forecasts are compared by the Diebold-Mariano test. \( H_0 \): The two forecasts are as accurate as each other. \( H_1 \): The two forecasts have different accuracies. \( H_{\text{tar}}^1 \): TAR is more accurate than CoTAR. \( H_{\text{cotar}}^1 \): CoTAR is more accurate than TAR. This table reports rejection frequencies associated with the nominal size \( a = 0.05 \).

We next focus on the cases where the true DGP is TAR and hence \( H_1 \) and \( H_{\text{tar}}^1 \) are true (Table 2). When the alternative hypothesis is chosen to be \( H_{\text{cotar}}^1 \), rejections of \( H_0 \) almost never occur as expected. In what follows, we restrict our attention to \( H_1 \) and \( H_{\text{tar}}^1 \) and discuss how the empirical power of the DM test varies across cases. First, the empirical power is always higher when \( H_{\text{tar}}^1 \) is chosen over \( H_1 \). This is a well-known power gain from implementing a one-sided test instead of a two-sided test. The spread in power is approximately 10-15% points; for example, the empirical power is \( \{0.739,0.849\} \) under \( \{H_1,H_{\text{tar}}^1\} \) with \( N = 180 \), strong threshold effects, \( \psi_0 = 0.8 \), and \( \tau = 0.4 \). In what follows, we focus on \( H_{\text{tar}}^1 \) to keep our discussions concise.

Second, stronger threshold effects result in higher empirical power (Table 2). Other things being equal, the empirical power under strong threshold effects (i.e., \( \beta_2 = (0.8,0.8)^T \)) always surpasses the empirical power under weak threshold effects (i.e., \( \beta_2 = (0.6,0.6)^T \)). Similarly, the empirical power increases when the threshold
Table 2: Rejection frequencies of the Diebold-Mariano test (DGP: TAR)

<table>
<thead>
<tr>
<th>TE</th>
<th>ψ0</th>
<th>τ</th>
<th>$N = 180$</th>
<th>$N = 360$</th>
<th>$N = 540$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>$H_1$</td>
<td>$H_{1}^{\text{tar}}$</td>
<td>$H_{1}^{\text{cotar}}$</td>
</tr>
<tr>
<td>weak</td>
<td>0.2</td>
<td>0.4</td>
<td>0.236</td>
<td>0.335</td>
<td>0.005</td>
</tr>
<tr>
<td>weak</td>
<td>0.2</td>
<td>0.6</td>
<td>0.166</td>
<td>0.251</td>
<td>0.011</td>
</tr>
<tr>
<td>weak</td>
<td>0.2</td>
<td>0.8</td>
<td>0.104</td>
<td>0.180</td>
<td>0.015</td>
</tr>
<tr>
<td>weak</td>
<td>0.8</td>
<td>0.4</td>
<td>0.277</td>
<td>0.398</td>
<td>0.002</td>
</tr>
<tr>
<td>weak</td>
<td>0.8</td>
<td>0.6</td>
<td>0.211</td>
<td>0.333</td>
<td>0.001</td>
</tr>
<tr>
<td>weak</td>
<td>0.8</td>
<td>0.8</td>
<td>0.171</td>
<td>0.276</td>
<td>0.007</td>
</tr>
<tr>
<td>strong</td>
<td>0.2</td>
<td>0.4</td>
<td>0.487</td>
<td>0.613</td>
<td>0.000</td>
</tr>
<tr>
<td>strong</td>
<td>0.2</td>
<td>0.6</td>
<td>0.292</td>
<td>0.439</td>
<td>0.001</td>
</tr>
<tr>
<td>strong</td>
<td>0.2</td>
<td>0.8</td>
<td>0.152</td>
<td>0.301</td>
<td>0.003</td>
</tr>
<tr>
<td>strong</td>
<td>0.8</td>
<td>0.4</td>
<td>0.739</td>
<td>0.849</td>
<td>0.000</td>
</tr>
<tr>
<td>strong</td>
<td>0.8</td>
<td>0.6</td>
<td>0.600</td>
<td>0.737</td>
<td>0.000</td>
</tr>
<tr>
<td>strong</td>
<td>0.8</td>
<td>0.8</td>
<td>0.353</td>
<td>0.516</td>
<td>0.000</td>
</tr>
</tbody>
</table>

DGP for the target variable is TAR: $y_t = 0.2y_{t-1} + \epsilon_t$ if $x_{t-1} < 0$, and $y_t = \alpha_0 + \phi_2 0_{x_{t-1}} + \epsilon_t$ if $x_{t-1} \geq 0$. $(\alpha_0, \phi_2) \in \{(0.6, 0.6), (0.8, 0.8)\}$ under weak and strong threshold effects (TEs), respectively. DGP for the threshold variable: $x_t = \psi_0 x_{t-1} + \xi_t$ with $\psi_0 \in \{0.2, 0.8\}$. Sample size: $N \in \{180, 360, 540\}$. The rolling window out-of-sample prediction is performed, where the window size is $n = \lceil \tau N \rceil$ with $\tau \in \{0.4, 0.6, 0.8\}$.

The TAR-based and CoTAR-based forecasts are compared by the Diebold-Mariano test. $H_0$: The two forecasts are as accurate as each other. $H_1$: The two forecasts have different accuracies. $H_{1}^{\text{tar}}$: TAR is more accurate than CoTAR. $H_{1}^{\text{cotar}}$: CoTAR is more accurate than TAR. This table reports rejection frequencies associated with the nominal size $\alpha = 0.05$.

variable $x$ becomes more persistent (i.e., when $\psi_0$ increases from 0.2 to 0.8). These results imply that the greater persistence in $y$ and $x$ helps us distinguish the relative predictive performance of the TAR and CoTAR models. When $N = 180$ and $\tau = 0.4$, the empirical power is 0.335 under weak threshold effects with $\psi_0 = 0.2$; 0.398 under weak threshold effects with $\psi_0 = 0.8$; 0.613 under strong threshold effects with $\psi_0 = 0.2$; 0.849 under strong threshold effects with $\psi_0 = 0.8$.

Third and curiously, the empirical power is decreasing in $\tau \in \{0.4, 0.6, 0.8\}$ for all cases considered (Table 2). Taking $N = 180$, strong threshold effects, and $\psi_0 = 0.8$ as an example, the empirical power is $\{0.849, 0.737, 0.516\}$ for $\tau \in \{0.4, 0.6, 0.8\}$, respectively. Recall that the window size is $n \in \{72, 108, 144\}$ and the number of windows is $T \in \{108, 72, 36\}$ for $\tau \in \{0.4, 0.6, 0.8\}$. These results suggest that researchers should secure a sufficient number of windows to keep the DM test powerful. Choosing a too small window size, however, would trigger computational failure due to too little
Table 3: Rejection frequencies of the Diebold-Mariano test (DGP: CoTAR)

<table>
<thead>
<tr>
<th>TE</th>
<th>(\psi_0)</th>
<th>(\tau)</th>
<th>(N = 180)</th>
<th>(N = 360)</th>
<th>(N = 540)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>(H_1)</td>
<td>(H_1^{\text{tar}})</td>
<td>(H_1^{\text{cotar}})</td>
</tr>
<tr>
<td>weak</td>
<td>0.8</td>
<td>0.4</td>
<td>0.191</td>
<td>0.005</td>
<td>0.292</td>
</tr>
<tr>
<td>weak</td>
<td>0.8</td>
<td>0.6</td>
<td>0.246</td>
<td>0.004</td>
<td>0.371</td>
</tr>
<tr>
<td>weak</td>
<td>0.8</td>
<td>0.8</td>
<td>0.177</td>
<td>0.011</td>
<td>0.296</td>
</tr>
<tr>
<td>strong</td>
<td>0.2</td>
<td>0.4</td>
<td>0.434</td>
<td>0.000</td>
<td>0.595</td>
</tr>
<tr>
<td>strong</td>
<td>0.2</td>
<td>0.6</td>
<td>0.324</td>
<td>0.002</td>
<td>0.533</td>
</tr>
<tr>
<td>strong</td>
<td>0.2</td>
<td>0.8</td>
<td>0.140</td>
<td>0.001</td>
<td>0.283</td>
</tr>
<tr>
<td>strong</td>
<td>0.8</td>
<td>0.4</td>
<td>0.769</td>
<td>0.001</td>
<td>0.865</td>
</tr>
<tr>
<td>strong</td>
<td>0.8</td>
<td>0.6</td>
<td>0.697</td>
<td>0.000</td>
<td>0.834</td>
</tr>
<tr>
<td>strong</td>
<td>0.8</td>
<td>0.8</td>
<td>0.346</td>
<td>0.001</td>
<td>0.521</td>
</tr>
</tbody>
</table>

DGP for the target variable is CoTAR: \(y_t = 0.2y_{t-1} + \epsilon_t\) if \(x_{t-1} < \mu_{t-1}(c_0)\), and \(y_t = \alpha_{20} + \phi_{20}y_{t-1} + \epsilon_t\) if \(x_{t-1} \geq \mu_{t-1}(c_0)\), where \(\mu_1(c_0)\) is the \(mc_0\)-th smallest value of \(\{x_t, x_{t-1}, \ldots, x_{t-m+1}\}\) with \((m, c_0) = (12, 0.5)\). \((\alpha_{20}, \phi_{20}) \in \{(0.6, 0.6), (0.8, 0.8)\}\) under weak and strong threshold effects (TEs), respectively. DGP for the threshold variable: \(x_t = \psi_0 x_{t-1} + \xi_t\) with \(\psi_0 \in \{0.2, 0.8\}\). Sample size: \(N \in \{180, 360, 540\}\). The rolling window out-of-sample prediction is performed, where the window size is \(n = \lceil \tau N \rceil\) with \(\tau \in \{0.4, 0.6, 0.8\}\).

The TAR-based and CoTAR-based forecasts are compared by the Diebold-Mariano test. \(H_0\): The two forecasts are as accurate as each other. \(H_1\): The two forecasts have different accuracies. \(H_1^{\text{tar}}\): TAR is more accurate than CoTAR. \(H_1^{\text{cotar}}\): CoTAR is more accurate than TAR. This table reports rejection frequencies associated with the nominal size \(\alpha = 0.05\).

Information available for estimation. In fact, estimation would become unstable if \(\tau = 0.2\), in which case the window size is only \(n = 36\).

Fourth, the larger sample size leads to higher power for all cases (Table 2). See for example the cases with strong threshold effects, \(\psi_0 = 0.2\), and \(\tau = 0.4\); the empirical power is \(\{0.613, 0.877, 0.973\}\) for \(N \in \{180, 360, 540\}\), respectively. In summary, all our simulation results in Table 2 are plausible, and we can conclude that the DM test achieves desired power properties when the true DGP is TAR.

We next focus on the cases where the true DGP is CoTAR and hence \(H_1\) and \(H_1^{\text{cotar}}\) are true (Table 3). Key implications from Table 3 are mostly analogous to those from Table 2. When the alternative hypothesis is chosen to be \(H_1^{\text{tar}}\), rejections of \(H_0\) almost never occur. The empirical power is approximately 10-20\% points higher when \(H_1^{\text{cotar}}\) is chosen over \(H_1\); for example, the empirical power is \(\{0.346, 0.521\}\).
under \( \{H_1, H_1^{cotar}\} \) with \( N = 180 \), strong threshold effects, \( \psi_0 = 0.8 \), and \( \tau = 0.8 \). We restrict our attention to \( H_1^{cotar} \) hereafter.

The empirical power under strong threshold effects is always higher than the empirical power under weak threshold effects (Table 3). Besides, the empirical power increases when the threshold variable \( x \) becomes more persistent. These results reinforce our previous finding from Table 2 that the greater persistence in \( y \) and \( x \) magnifies the difference between the predictive accuracies of TAR and CoTAR.

Interestingly, the empirical power is decreasing in \( \tau \in \{0.4, 0.6, 0.8\} \) for most but not all cases (Table 3). Taking \( N = 180 \), strong threshold effects, and \( \psi_0 = 0.8 \) as an example, the empirical power is \( \{0.865, 0.834, 0.521\} \) for \( \tau \in \{0.4, 0.6, 0.8\} \), respectively. Replacing strong threshold effects with weak threshold effects, however, results in non-monotonic power of \( \{0.292, 0.371, 0.296\} \). These results suggest that the window size matters more when the DGP is CoTAR than when the DGP is TAR. This contrast seems reasonable because, as shown in model (6), \( m = 12 \) initial values must be secured to model conditional threshold effects.

Summarizing Tables 1-3, our overall conclusion is that the rolling window DM test achieves desired size properties when the true DGP is AR, and desired power properties when the true DGP is TAR or CoTAR.

## 4 Empirical analysis

In this section, we analyze the threshold effects of stock market volatility on crude oil market volatility. Using TAR, CQZ2022 found significant threshold effects of the RV of SPX on the RV of WTI. They also found that TAR attains a higher out-of-sample predictive accuracy than AR. Our conjecture is that the CoTAR-based forecast should be as accurate as or even more accurate than the TAR-based forecast, as conditional thresholds should well match an underlying mechanism of crude oil RVs. We describe our data and perform some preliminary analysis in Section 4.1, and conduct our main analysis in Section 4.2.

### 4.1 Data and preliminary analysis

We mimic the data of CQZ2022 as closely as possible. Crude oil prices are proxied by daily observations of “Cushing, OK WTI Spot Price FOB (Dollars per Barrel)”, which are publicly available at the website of U.S. Energy Information Administration with the ID code being RWTC. Stock prices are proxied by daily closing values of SPX in
terms of U.S. dollars, which can be downloaded freely at Investing.com. To formulate RV measures, we rely on notations commonly used in the Mixed Data Sampling (MIDAS) literature including Motegi and Dennis (2022). Let $\mathbb{L} = \{1, 2, \ldots, N\}$ be the set of months, and let $m_t \in \mathbb{N}$ be the number of business days in month $t \in \mathbb{L}$. Let $\mathbb{H}_t = \{t - 1 + 1/m_t, t - 1 + 2/m_t, \ldots, t\}$ be the set of days in month $t \in \mathbb{L}$. Let $P_t$ be the price of either WTI or SPX on day $t \in \mathbb{H}$, where $\mathbb{H} = \cup_{t \in \mathbb{L}} \mathbb{H}_t$. A daily return on day $t \in \mathbb{H}$ is defined as $r_t = (P_t - P_{t-1})/P_{t-1}$.

There are at least three common ways to measure volatility at month $t \in \mathbb{L}$:

$$RV_t^\pm = \sum_{j \in \mathbb{H}_t} r_j^2, \quad RV_t^+ = \sum_{j \in \mathbb{H}_t} r_j^2 1(r_j > 0), \quad RV_t^- = \sum_{j \in \mathbb{H}_t} r_j^2 1(r_j < 0). \quad (7)$$

The (total) realized variance, $RV_t^\pm$, is the most classical and commonly used RV measure in the literature; it takes into account positive and negative returns equally. The realized semi-variances, $RV^+$ and $RV^-$, were originally put forward by Barndorff-Nielsen, Kinnebrock, and Shephard (2010) as proxies of “good volatility” and “bad volatility”. By distinguishing good and bad volatilities, portfolio managers can handle upside and downside risks separately. It is also well known that time series properties of $RV^+$ and $RV^-$ often differ from each other. For these reasons, the realized semi-variances have been studied extensively in the volatility prediction literature.

In Figure 1, we plot monthly log-RV data from January 1990 through December 2021 for each of WTI and SPX; this sample period matches CQZ2022, and the sample size is $N = 384$ months. Several implications can be drawn from Figure 1. First, the crude oil market is riskier than the stock market for any RV measures considered. The oil volatility exceeds the stock volatility for $\{376, 367, 360\}$ months out of $N = 384$ in terms of $\{RV^\pm, RV^+, RV^-\}$, respectively.

Second, for both the oil and stock volatilities, the COVID-19 crisis in early 2020 is the largest shock in the entire sample period (Figure 1). In particular, the COVID-19 shock is by far the largest shock for WTI; other shocks including the global financial crisis (GFC) around 2008 had much smaller impacts on the oil volatility. GFC and COVID-19 are more comparable in magnitude for SPX, although the latter had a

---

6Another common definition is $\tilde{r}_t = \ln(P_t/P_{t-1})$, but we use $r_t = (P_t - P_{t-1})/P_{t-1}$ to better mimic CQZ2022. The spread between $\tilde{r}_t$ and $r_t$ increases during financial turmoil.

7See Chen and Ghysels (2011) and Patton and Sheppard (2015) for early contributions. Realized semi-variances of crude oil markets are analyzed by Ma, Wei, Liu, and Huang (2018), Gong and Lin (2021), Lyu, Wei, Hu, and Yang (2021), and CQZ2022, among others. According to CQZ2022’s out-of-sample analysis, the superiority of TAR relative to AR is more salient when the bad volatility in crude oil markets is forecasted than when the good volatility is forecasted.
WTI: West Texas Intermediate crude oil spot price (USD per barrel). SPX: Closing value of S&P 500 Index (USD). For each series, we compute monthly total, good, and bad volatilities (i.e., $RV^\pm$, $RV^+$, $RV^-$) from daily squared returns. This figure plots the log-RVs from January 1990 through December 2021 ($N = 384$ months).

slightly larger impact on the stock market volatility. These observations are unchanged.
whichever volatility measure in (7) is used.

Third, the bad volatility appears to be more volatile than the total and good volatilities (Figure 1). Besides, negative spikes are more frequently observed in $RV^-$ than in $RV^\pm$ and $RV^+$. These features appear for both WTI and SPX, and they suggest that empirical results of $RV^-$ may differ from those of $RV^\pm$ and $RV^+$.

In Table 4, we report sample statistics of the monthly log-RVs of WTI and SPX. Implications from Table 4 are consistent with our previous discussions on Figure 1. First, the mean, median, minimum, and maximum of the log-RV of WTI are all larger than those of the log-RV of SPX for any RV measures considered. This fact confirms our previous observation that the crude oil market is riskier than the stock market. Second, the skewness and kurtosis of the log-RV of WTI are both larger than those of the log-RV of SPX for any measures. This evidence is in line with the fact that the COVID-19 pandemic is an exceptionally large shock for the crude oil market but comparable with GFC for the stock market. Third, for both WTI and SPX, $\ln RV^-$ has the larger standard deviation than $\ln RV^\pm$ and $\ln RV^+$, verifying that the bad volatility is most volatile among the three measures. Fourth, all six series are positively skewed, and $RV^-$ has the smallest skewness. These facts are in line with Figure 1 since positive spikes occur more frequently than negative spikes for $RV^\pm$ and $RV^+$, while the frequencies of positive and negative spikes are closer in $RV^-$.  

Table 4: Sample statistics of the log-RV measures of WTI and SPX returns

<table>
<thead>
<tr>
<th>series</th>
<th>RV</th>
<th>mean</th>
<th>median</th>
<th>min</th>
<th>max</th>
<th>stdev</th>
<th>skew</th>
<th>kurt</th>
</tr>
</thead>
<tbody>
<tr>
<td>WTI</td>
<td>$RV^\pm$</td>
<td>-4.767</td>
<td>-4.848</td>
<td>-6.879</td>
<td>-0.153</td>
<td>0.909</td>
<td>0.876</td>
<td>5.680</td>
</tr>
<tr>
<td>WTI</td>
<td>$RV^+$</td>
<td>-5.491</td>
<td>-5.514</td>
<td>-8.057</td>
<td>-0.698</td>
<td>0.986</td>
<td>0.669</td>
<td>5.099</td>
</tr>
<tr>
<td>WTI</td>
<td>$RV^-$</td>
<td>-5.592</td>
<td>-5.627</td>
<td>-8.521</td>
<td>-1.022</td>
<td>1.028</td>
<td>0.494</td>
<td>4.601</td>
</tr>
<tr>
<td>SPX</td>
<td>$RV^\pm$</td>
<td>-6.508</td>
<td>-6.587</td>
<td>-8.803</td>
<td>-2.615</td>
<td>0.976</td>
<td>0.607</td>
<td>3.781</td>
</tr>
<tr>
<td>SPX</td>
<td>$RV^+$</td>
<td>-7.165</td>
<td>-7.234</td>
<td>-9.493</td>
<td>-3.425</td>
<td>0.939</td>
<td>0.663</td>
<td>4.203</td>
</tr>
<tr>
<td>SPX</td>
<td>$RV^-$</td>
<td>-7.433</td>
<td>-7.357</td>
<td>-11.25</td>
<td>-3.204</td>
<td>1.264</td>
<td>0.015</td>
<td>3.225</td>
</tr>
</tbody>
</table>

WTI: West Texas Intermediate crude oil spot price (USD per barrel). SPX: Closing value of S&P 500 Index (USD). For each series, we compute monthly total, good, and bad realized volatilities (i.e., $RV^\pm$, $RV^+$, $RV^-$) from daily squared returns. This table reports summary statistics of the log-RVs from January 1990 through December 2021 ($N = 384$ months).
4.2 Main analysis and results

To investigate threshold effects of the stock market volatility on the crude oil market volatility, we use the log-RV of SPX as the threshold variable $x$ and the log-RV of WTI as the target variable $y$. We analyze the three volatility measures in (7) separately. For each measure, we fit the TAR model (1) and the CoTAR model (2) to evaluate the validity of constant versus conditional threshold effects. The lag length is set to be $p = 2$ months for both models. The nuisance parameters of TAR are $\gamma = (d, \mu)^\top$; the choice set of delay parameter $d$ is $D = \{1, 2, 3, 4\}$; the choice set of threshold parameter $\mu$ is the entire memory $X^n_1 = \{x_1, \ldots, x_n\}$. The nuisance parameters of CoTAR are $\gamma = (d, c)^\top$; the choice set of $d$ is $D = \{1, 2, 3, 4\}$; the choice set of percentile parameter $c$ is $\{1/m, 2/m, \ldots, 1\}$ with the memory size being fixed at $m = 12$ months. For both models, we restrict the space of $\gamma$ by (4) so that both regimes appear sufficiently many times. We perform full sample analysis in Section 4.2.1, and rolling window out-of-sample analysis in Section 4.2.2.

4.2.1 Full sample analysis

In this section, we perform full sample analysis with the sample period being January 1990 through December 2021 ($n = 384$). For each of TAR and CoTAR, we estimate the regression parameters $(\beta_1, \beta_2)$ and the nuisance parameters $\gamma$ via profiling as described in Section 2.2. Further, we implement the wild bootstrap test for the no-threshold-effect hypothesis, $H_0^* : \beta_1 = \beta_2$, as described in Section 2.3. The average LM test statistic is used, as it has the best finite sample performance according to the Monte Carlo simulations of MDH2022. When computing LM statistics conditional on $\gamma$, we use a heteroscedasticity-robust covariance matrix to address potential conditional heteroscedasticity in the target variable $y$; see MDH2022 for detailed procedures. The number of bootstrap samples is $B = 5000$.

In Figure 2, we add the estimated thresholds to the time series plot of $x$. The constant threshold $\hat{\mu}$ is drawn for the TAR case, while the conditional threshold $\mu_t(\hat{c})$ is drawn for the CoTAR case. Figure 2 highlights the difference between the two models. For TAR, regime 1 (resp. regime 2) represents periods of low (resp. high) volatility relative to the entire memory $X^n_1$. For CoTAR, the interpretation of low volatility versus high volatility is relative to the local memory $X^{t-m+1}_t$.

Focusing on $RV^\pm$, regime 2 seems to last longer than regime 1 for the TAR case (Figure 2). In particular, regime 2 occurs for 86 consecutive months from January 1997 through February 2004, a period of the dot-com bubble and its burst. Besides, regime
Figure 2: The log-RVs of SPX and estimated thresholds based on TAR and CoTAR

TAR: \( y_t = z_{t-1}^\top \beta_1 + u_t \) if \( x_{t-d} < \mu \) and \( y_t = z_{t-1}^\top \beta_2 + u_t \) if \( x_{t-d} \geq \mu \), where \( z_{t-1} = (1, y_{t-1}, y_{t-2})^\top \).

CoTAR: \( y_t = z_{t-1}^\top \beta_1 + u_t \) if \( x_{t-d} < \mu_{t-d-1}(c) \) and \( y_t = z_{t-1}^\top \beta_2 + u_t \) if \( x_{t-d} \geq \mu_{t-d-1}(c) \), where \( \mu_t(c) \) is the \( mc \)-th smallest value of \( \{x_{t-m+1}, \ldots, x_t\} \) with \( m = 12 \). \( y \) is the log-RV of WTI, and \( x \) is the log-RV of SPX. For each model and RV measure, we implement full-sample profiling estimation and plot the estimated thresholds. Sample period: January 1990 – December 2021 (\( n = 384 \) months).
2 occurs for 33 consecutive months from July 2007 through March 2010, reflecting GFC. For the CoTAR case, regime 1 seems to last longer than regime 2, which is a notable contrast to TAR. Regime 1 occurs for 16 consecutive months from April 2003 through July 2004, as $RV^\pm$ of SPX decreases rapidly after the dot-com bubble burst. Similarly, regime 1 occurs for 13 consecutive months from April 2009 through April 2010 in response to the convergence of GFC. The weaker persistence of regime 2 relative to regime 1 given CoTAR is consistent with Ma, Wahab, Huang, and Xu (2017), who fitted HAR models with Markov regime switching to crude oil RVs and found that a high volatility regime is short-lived. Finally, we observe similar results when $RV^\pm$ is replaced with $RV^\pm$.

Results of the bad volatility $RV^-$ are in stark contrast with those of $RV^\pm$ and $RV^+$ (Figure 2). Given TAR, the estimated constant threshold $\hat{\mu}$ for $RV^-$ is located at a larger percentile of $X^m$ than for the other RV measures. Given CoTAR, the estimated conditional threshold $\mu_t(\hat{c})$ for $RV^-$ passes through smaller percentiles of $X_{t-m+1}$ than for the others. The unique empirical results of $RV^-$ relative to $RV^\pm$ and $RV^+$ are in line with Figure 1 and Table 4, where $RV^-$ exhibits several special features. They are also in line with previous studies documenting the asymmetry between upside and downside volatilities of crude oil markets (e.g., Ma, Wei, Liu, and Huang, 2018, Lyu, Wei, Hu, and Yang, 2021, CQZ2022).

Results of the full sample analysis are summarized in Table 5. We begin with discussing the results associated with $RV^\pm$. First, the estimated threshold is $\hat{\mu} = -7.107$ for TAR, roughly a 28 percentile of $X^m$. Indeed, the share of regime 1 to the whole sample, denoted as $\hat{\delta}_1$, is 0.282 given TAR. Second, the estimated percentile parameter is $\hat{c} = 0.667$ for CoTAR. This suggests that regime 1 is the majority given CoTAR; indeed, $\hat{\delta}_1 = 0.618$. Third, $(\hat{D}_1, \hat{D}_2) = (2.250, 5.612)$ given TAR, where $\hat{D}_r$ is the average duration of regime $r \in \{1, 2\}$ in terms of months. Given CoTAR, $(\hat{D}_1, \hat{D}_2) = (3.966, 2.409)$. As expected from Figure 2, regime 2 lasts longer than regime 1 on average for TAR, whereas regime 1 lasts longer than regime 2 for CoTAR. Fourth and importantly, when $RV^\pm$ is analyzed, the wild-bootstrap p-value with respect to the no-threshold-effect hypothesis $H_0^*$ is 0.051 for TAR and 0.068 for CoTAR (Table 5). For both models, $H_0^*$ is rejected at the 10% level, suggesting the presence of threshold effects. The evidence of constant threshold effects is consistent with CQZ2022, while the evidence of conditional threshold effects is a new finding.

As expected from Figure 2, analyzing $RV^+$ instead of $RV^\pm$ leads to qualitatively similar results (Table 5). In particular, the bootstrap p-value with respect to $H_0^*$ is
Table 5: Summary results of the full sample analysis (January 1990–December 2021)

<table>
<thead>
<tr>
<th>model</th>
<th>RV</th>
<th>$\hat{d}$</th>
<th>$\hat{\mu}$</th>
<th>$\hat{c}$</th>
<th>$\hat{\delta}_1$</th>
<th>$\hat{\delta}_{12}$</th>
<th>$\hat{D}_1$</th>
<th>$\hat{D}_2$</th>
<th>$\hat{p}(H_0^*)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAR</td>
<td>$RV^\pm$</td>
<td>1</td>
<td>-7.107</td>
<td>-</td>
<td>0.282</td>
<td>0.556</td>
<td>0.175</td>
<td>2.250</td>
<td>5.612</td>
</tr>
<tr>
<td>TAR</td>
<td>$RV^+$</td>
<td>1</td>
<td>-7.761</td>
<td>-</td>
<td>0.285</td>
<td>0.505</td>
<td>0.198</td>
<td>2.019</td>
<td>4.982</td>
</tr>
<tr>
<td>TAR</td>
<td>$RV^-$</td>
<td>3</td>
<td>-6.735</td>
<td>-</td>
<td>0.703</td>
<td>0.843</td>
<td>0.372</td>
<td>6.233</td>
<td>2.691</td>
</tr>
<tr>
<td>CoTAR</td>
<td>$RV^\pm$</td>
<td>2</td>
<td>-</td>
<td>0.667</td>
<td>0.618</td>
<td>0.748</td>
<td>0.411</td>
<td>3.966</td>
<td>2.409</td>
</tr>
<tr>
<td>CoTAR</td>
<td>$RV^+$</td>
<td>1</td>
<td>-</td>
<td>0.667</td>
<td>0.637</td>
<td>0.717</td>
<td>0.500</td>
<td>3.537</td>
<td>1.985</td>
</tr>
<tr>
<td>CoTAR</td>
<td>$RV^-$</td>
<td>1</td>
<td>-</td>
<td>0.250</td>
<td>0.250</td>
<td>0.333</td>
<td>0.223</td>
<td>1.500</td>
<td>4.429</td>
</tr>
</tbody>
</table>

TAR: $y_t = z_{t-1}^\top \beta_1 + u_t$ if $x_{t-d} < \mu$ and $y_t = z_{t-1}^\top \beta_2 + u_t$ if $x_{t-d} \geq \mu$, where $z_{t-1} = (1, y_{t-1}, y_{t-2})^\top$. CoTAR: $y_t = z_{t-1}^\top \beta_1 + u_t$ if $x_{t-d} < \mu_{t-d}(c)$ and $y_t = z_{t-1}^\top \beta_2 + u_t$ if $x_{t-d} \geq \mu_{t-d}(c)$, where $\mu_{t}(c)$ is the $mc$-th smallest value of $\{x_{t-m+1}, \ldots, x_1\}$ with $m = 12$. $y$ is the log-RV of WTI, and $x$ is the log-RV of SPX. For each model and RV, we report summary results of the full sample analysis. $(\hat{d}, \hat{\mu}, \hat{c})$ are the profiling estimators for $(d, \mu, c)$. $\hat{\delta}_1$ is the share of regime 1 to the whole sample. $\hat{\delta}_{12}$ is the empirical transition probability from regime $r \in \{1, 2\}$ to regime 1. $\hat{D}_r$ is the average duration of regime $r$. $\hat{p}(H_0^*)$ is the bootstrap p-value for the no-threshold-effect hypothesis $H_0^* : \beta_1 = \beta_2$. Asterisks ***, **, and * indicate a rejection of $H_0^*$ at the 1%, 5%, and 10% levels, respectively.

0.083 for TAR and 0.091 for CoTAR, rejections at the 10% level.

Results on the bad volatility $RV^-$ are considerably different from those on $RV^\pm$ and $RV^+$ (Table 5). First, the estimated threshold is $\hat{\mu} = -6.735$ for TAR, roughly a 70 percentile of $X_1^n$. This means regime 1 is the majority given TAR. Second, the estimated percentile parameter is $\hat{c} = 0.250$ for CoTAR, indicating regime 1 is the minority given CoTAR. Third, $(\hat{D}_1, \hat{D}_2) = (6.233, 2.691)$ for TAR, while $(\hat{D}_1, \hat{D}_2) = (1.500, 4.429)$ for CoTAR. These durations imply regime 1 lasts longer than regime 2 on average for TAR, whereas regime 2 lasts longer than regime 1 for CoTAR. Fourth, the bootstrap p-value with respect to $H_0^*$ is 0.250 for TAR and 0.221 for CoTAR. Hence, as far as $RV^-$ is concerned, we do not detect statistical evidence of either constant threshold effects or conditional threshold effects at any conventional levels.

4.2.2 Rolling window out-of-sample analysis

In this section, we perform rolling window out-of-sample prediction based on TAR and CoTAR. The entire sample period is January 1990 – December 2021 as in the full sample analysis ($N = 384$). The window size is $n = \lceil \tau N \rceil$, and the number of windows is $T = N - n$. We choose $\tau \in \{0.4, 0.6, 0.8\}$ as in the Monte Carlo simulations (Section 3). The window size is $n \in \{153, 230, 307\}$ and the number of windows is
$T \in \{231, 154, 77\}$ for $\tau \in \{0.4, 0.6, 0.8\}$, respectively. Taking $\tau = 0.4$ as an example, the first window is January 1990 – September 2002; the second window is February 1990 – October 2002; the 231st and last window is March 2009 – November 2021.

For each rolling window, we fit TAR and CoTAR with $y = \ln RV^{WTI}$ and $x = \ln RV^{SPX}$, execute the profiling estimation, and compute a one-step-ahead out-of-sample forecast of $y$. The specification of the two models is the same as in the full sample analysis. We consider the three alternative RV measures appearing in (7). We perform the asymptotic DM test with the mean-difference statistic $S_1$ as described in Section 2.4. The null hypothesis $H_0$ is that the TAR-based and CoTAR-based forecasts are as accurate as each other. We consider three alternative hypotheses: $H_1$ states that the two forecasts have different accuracies; $H_{\text{tar}}^1$ states that TAR is more accurate than CoTAR; $H_{\text{cotar}}^1$ states that CoTAR is more accurate than TAR.

In Figure 3, we plot empirical results with $\tau = 0.6$. First, the TAR and CoTAR forecasts for $RV^\pm$ and $RV^+$ seem fairly accurate. The two models deliver reasonable one-month-ahead forecasts despite the fact that the prediction periods contain GFC and the COVID-19 pandemic. For $RV^-$, the TAR and CoTAR forecasts tend to be conservative in both upward and downward directions. Second, the two models seem to perform roughly as well as each other for all RV measures. It is hard to visually judge the relative predictive accuracies of the two models, motivating a formal evaluation via the DM test.

In Table 6, we report the RMSEs of TAR and CoTAR as well as asymptotic p-values of the DM test. As reference values, we also report the RMSEs of an intercept-only model and an AR(2) model. Several remarks on Table 6 are in order. First, the threshold models (i.e., TAR and CoTAR) attain smaller RMSEs than the non-threshold models (i.e., intercept and AR) for all cases considered, which implies that volatility prediction improves by incorporating threshold effects. Taking $RV^\pm$ and $\tau = 0.6$ as an example, the RMSEs of the intercept, AR, TAR, and CoTAR forecasts are $\{0.982, 0.929, 0.776, 0.733\}$, respectively. There is apparently a large gap in RMSEs between the non-threshold and threshold models, which is indicative of substantial threshold effects. These results are partly consistent with CQZ2022, who found substantial gains in predictive power from using TAR instead of AR.

Second, $RMSE_{\text{tar}} > RMSE_{\text{cotar}}$ for three out of the nine cases considered (Table 6). According to the DM test, the spread in the RMSEs is statistically significant for all of the three cases, pointing to the superiority of CoTAR in terms of one-

---

8Results with $\tau \in \{0.4, 0.8\}$ are qualitatively similar to Figure 3, hence omitted to save space. These extra results are available upon request.
Table 6: Summary results of the rolling window out-of-sample analysis

<table>
<thead>
<tr>
<th></th>
<th>$\tau = 0.4$</th>
<th>$\tau = 0.6$</th>
<th>$\tau = 0.8$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(n = 153, $T = 231$)</td>
<td>(n = 230, $T = 154$)</td>
<td>(n = 307, $T = 77$)</td>
</tr>
<tr>
<td></td>
<td>$RV^\pm$</td>
<td>$RV^+$</td>
<td>$RV^-$</td>
</tr>
<tr>
<td>RMSE$^{inter}$</td>
<td>0.923</td>
<td>0.997</td>
<td>1.055</td>
</tr>
<tr>
<td>RMSE$^{ar}$</td>
<td>0.842</td>
<td>0.931</td>
<td>1.013</td>
</tr>
<tr>
<td>RMSE$^{tar}$</td>
<td>0.709</td>
<td>0.833</td>
<td>0.958</td>
</tr>
<tr>
<td>RMSE$^{cotar}$</td>
<td>0.722</td>
<td>0.847</td>
<td>0.990</td>
</tr>
<tr>
<td>$\hat{p}(H_1)$</td>
<td>0.466</td>
<td>0.589</td>
<td>0.229</td>
</tr>
<tr>
<td>$\hat{p}(H_1^{tar})$</td>
<td>0.233</td>
<td>0.294</td>
<td>0.114</td>
</tr>
<tr>
<td>$\hat{p}(H_1^{cotar})$</td>
<td>0.767</td>
<td>0.706</td>
<td>0.886</td>
</tr>
</tbody>
</table>

We perform rolling window out-of-sample forecast of the log-RVs of WTI, where the window size is $n = \lfloor \tau N \rfloor$ with $\tau \in \{0.4, 0.6, 0.8\}$ and the number of windows is $T = N - n$. Entire sample period: January 1990 – December 2021 ($N = 384$). We report the RMSEs of the intercept-only, AR, TAR, and CoTAR models. We also report asymptotic p-values of the Diebold-Mariano test. $H_0$: The TAR and CoTAR forecasts are as accurate as each other. $H_1$: The two forecasts have different accuracies. $H_{1^{tar}}$: TAR is more accurate than CoTAR. $H_{1^{cotar}}$: CoTAR is more accurate than TAR. Asterisks $^{***}$, $^{**}$, and $^*$ indicate a rejection of $H_0$ at the 1%, 5%, and 10% levels, respectively.

Month-ahead prediction. Focus again on $RV^\pm$ with $\tau = 0.6$, then $\hat{p}(H_1) = 0.019$ and $\hat{p}(H_{1^{cotar}}) = 0.010$. In this case, the null hypothesis of equal predictive accuracies, $H_0$, is rejected at the 5% level in favor of CoTAR (i.e., $H_1$ and $H_{1^{cotar}}$).

Third and interestingly, the other two cases with $RMSE^{tar} > RMSE^{cotar}$ arise for bad volatility $RV^-$ (Table 6). For $\tau = 0.8$, we have that $RMSE^{tar} = 1.111$, $RMSE^{cotar} = 1.042$, $\hat{p}(H_1) = 0.093$, and $\hat{p}(H_{1^{cotar}}) = 0.046$. In this case, $H_0$ is rejected in favor of $H_1$ at the 10% level and in favor of $H_{1^{cotar}}$ at the 5% level, again suggesting the higher out-of-sample performance of CoTAR. For $\tau = 0.6$, we have that $RMSE^{tar} = 1.020$, $RMSE^{cotar} = 0.998$, $\hat{p}(H_1) = 0.187$, and $\hat{p}(H_{1^{cotar}}) = 0.094$. The statistical significance is weaker than for $\tau = 0.8$, but we reject $H_0$ in favor of $H_{1^{cotar}}$ at the 10% level. CQZ2022 found that TAR is significantly superior to AR in terms of predicting the bad crude oil volatility. We have further revealed that CoTAR may perform even better than TAR, which is a valuable finding for business practitioners since predicting bad volatility plays a crucial role in financial risk management. It matters for policymakers as well, since a sharp decline in crude oil prices tends to be followed by a recession and deflation. The CoTAR-based prediction should help policymakers better monitor and combat slowdown in crude oil markets and the goods...
market in general.

Fourth, $RMSE^{tar} < RMSE^{cotar}$ for six cases, in only one of which $H_0$ is rejected at a conventional level (Table 6). The relevant case is $RV^+$ with $\tau = 0.6$, where $RMSE^{tar} = 0.856$, $RMSE^{cotar} = 0.883$, $\hat{p}(H_1) = 0.139$, and $\hat{p}(H_1^{tar}) = 0.070$. $H_0$ cannot be rejected in favor of $H_1$, but is rejected in favor of $H_1^{tar}$ at the 10% level. Except for this marginal rejection, there are no cases where $H_0$ is rejected in favor of $H_1^{tar}$. Hence, our overall conclusion is that CoTAR is at least on par with and sometimes better than TAR in terms of forecasting crude oil market volatility.

## 5 Conclusion

Modelling and predicting crude oil RVs are a vital area of research in financial econometrics and energy economics. A variety of models including HAR and TAR are employed to improve the out-of-sample forecast of crude oil RVs. Using TAR, CQZ2022 found significant threshold effects of stock market volatility on crude oil market volatility, and achieved the higher predictive accuracy than benchmark AR models. In the recent literature of TAR-type models, there is an increasing attention on extending a constant threshold $\mu$ to a time-varying threshold $\mu_t$. The time-varying threshold is supposed to better mimic adaptive reactions of market participants than the constant threshold. MDH2022 proposed a novel time-varying threshold model called CoTAR by specifying $\mu_t$ as an empirical quantile of recent observations of $x$.

In this paper, we investigate conditional threshold effects of stock market RVs on crude oil RVs based on the CoTAR model. Our main interest lies in comparing the out-of-sample forecast accuracies of TAR and CoTAR. We have shown via the Monte Carlo simulations that the relative performance of the two models can be correctly evaluated via the DM test. The DM test exhibits reasonable size and power in finite samples. This is a useful finding since the out-of-sample performance of CoTAR has never been studied in the literature. In our empirical application, the target variable $y$ is crude oil RVs and the threshold variable $x$ is stock market RVs. These choices are the same as CQZ2022, hence our work is a notable expansion of CQZ2022 from constant threshold effects (TAR) to conditional threshold effects (CoTAR).

Our empirical results are summarized as follows. First, the threshold models (i.e., TAR and CoTAR) attain higher predictive power for crude oil RVs than the non-threshold models (i.e., intercept and AR), which highlights the importance of capturing threshold effects. Second, the DM test indicates that CoTAR is at least
on par with and sometimes better than TAR in terms of forecasting crude oil RVs. In particular, we observe fairly strong evidence that CoTAR outperforms TAR when forecasting the bad volatility $RV^-$. CQZ2022 found that TAR predicts $RV^-$ better than AR, and we have further found that CoTAR outperforms TAR. Our finding helps market participants and policymakers better control downside risks.

Acknowledgements

We thank John W. Dennis and Munenori Ueda for helpful comments and discussions. The first author is grateful for the financial support of JSPS KAKENHI Grant-in-Aid for Challenging Research (Exploratory) with Grant Number 23K17555, Ishii Memorial Securities Research Promotion Foundation, Japan Securities Scholarship Foundation, Murata Science Foundation, and Nomura Foundation.

References


Hansen, B. E. (1996): “Inference when a nuisance parameter is not identified under the null hypothesis,” *Econometrica*, 64(2), 413–430.


Figure 3: Rolling window out-of-sample forecast of log-RVs of WTI ($\tau = 0.6$)

TAR: $\ln RV^\pm$

CoTAR: $\ln RV^\pm$

TAR: $\ln RV^+$

CoTAR: $\ln RV^+$

TAR: $\ln RV^-$

CoTAR: $\ln RV^-$

TAR: $y_t = z_{t-1}^\top \beta_1 + u_t$ if $x_{t-d} < \mu$ and $y_t = z_{t-1}^\top \beta_2 + u_t$ if $x_{t-d} \geq \mu$, where $z_{t-1} = (1, y_{t-1}, y_{t-2})^\top$.

CoTAR: $y_t = z_{t-1}^\top \beta_1 + u_t$ if $x_{t-d} < \mu_{t-d-1}(c)$ and $y_t = z_{t-1}^\top \beta_2 + u_t$ if $x_{t-d} \geq \mu_{t-d-1}(c)$, where $z_{t-1} = (1, y_{t-1}, y_{t-2})^\top$ and $\mu_t(c)$ is the $mc$-th smallest value of $\{x_{t-m+1}, \ldots, x_t\}$ with $m = 12$. $y$ is the log-RV of WTI, and $x$ is the log-RV of SPX. Entire sample period: January 1990 – December 2021 ($N = 384$). In this figure, we perform rolling window one-step-ahead out-of-sample forecast of $y$ based on either the TAR or CoTAR model, where the window size is $n = \lfloor \tau N \rfloor$ with $\tau = 0.6$. We plot actual $\{y_t\}_{t=1}^N$ with the blue, solid line and predicted $\{\hat{y}_t\}_{t=n+1}^N$ with the red, dashed line.