LEIBNIZ INSTITUTE
SAFE
Sustainable Architecture for Finance in Europe

Massimiliano Caporina | Michele Costola

Time-Varying Granger Causality Tests for Applications in Global Crude Oil Markets: A Study on the DCC-MGARCH Hong Test

SAFE Working Paper No. 324

Leibniz Institute for Financial Research SAFE
Sustainable Architecture for Finance in Europe
info@safe-frankfurt.de | www.safe-frankfurt.de

Electronic copy available at: https://ssrn.com/abstract=3941778

Time-varying Granger causality tests for applications in global crude oil markets: A study on the DCC-MGARCH Hong test

Massimiliano Caporin[a], Michele Costola[b,c]

[a]*Department of Statistical Sciences, University of Padova, Padova, Italy*
[b]*Department of Economics, University Ca' Foscari Venezia, Venezia, Italy*
[c]*Leibniz Institute for Financial Research SAFE, Frankfurt am Main, Germany*

Abstract

Analysing causality among oil prices and, in general, among financial and economic variables is of central relevance in applied economics studies. The recent contribution of Lu et al. (2014) proposes a novel test for causality— the DCC-MGARCH Hong test. We show that the critical values of the test statistic must be evaluated through simulations, thereby challenging the evidence in papers adopting the DCC-MGARCH Hong test. We also note that rolling Hong tests represent a more viable solution in the presence of short-lived causality periods.

Keywords: Granger Causality, Hong test, DCC-GARCH, Oil market, COVID-19

JEL: C10, C13, C32, C58, Q43, Q47

1. Introduction

In a globalized economy, the study of spillovers among the prices of reference goods— drivers of possible shocks to both the real economy and the financial markets— is of central relevance. Oil prices represent one of these fundamental drivers, given their relevance in both the real and the financial cycles. However, oil is traded in several mercantile exchanges and with prices that reflect the different types of oil that can be extracted; classic examples are Brent and the West Texas Intermediate (WTI). In this setting, despite the fact that Brent and WTI being perceived as reference prices for the market, the study of information transmission among oil prices is relevant from an economic perspective in order to determine which price is mostly impacted by shocks— for example, associated with the disruption or reduction of the oil production or oil delivery— and how shocks are transmitted to other oil prices and subsequently to oil-derived productions; examples, in this regard, examples are given by Lu et al. (2014) and Caporin et al. (2019). Several studies have already addressed the issue of spillover or causality among oil prices. Among many others, we cite Lin and Tamvakis (2001, 2004), Hammoudeh and Li (2004), Bekiros and Diks (2008), and Geng et al. (2017). Different approaches have been considered for the analysis of the causality or spillovers, ranging from the standard causality testing put forward by Granger (1969) to a variety of generalizations including non-linear methods, quantile regression-based approaches, and wavelet transforms.

Email addresses: massimiliano.caporin@unipd.it (Massimiliano Caporin), michele.costola@unive.it (Michele Costola)

Electronic copy available at: https://ssrn.com/abstract=3941778

One the approaches that has recently received attention is included in Lu et al. (2014), in which a novel testing procedure has been put forward in order to test for dynamic causality among variables. The authors introduce a test combining the approach given by Hong (2001) and Hong et al. (2009) for spillover testing employing the dynamic conditional correlation (DCC) modeling strategy given by Engle and Sheppard (2001) and Engle (2002). In Lu et al. (2014) the proposed test, called DCC-MGARCH Hong (henceforth, the DCC-Hong), is used to assess the dynamic and contemporaneous spillover among different oil prices, the futures prices of Brent and WTI, and the Dubai and Tapis spot prices, thereby revealing the occurrence of relevant spillovers, both unidirectional and biderectional. The same testing procedure has also been used in different settings. Jammazi et al. (2017b) adopt the DCC-Hong test to study the causality between oil prices and stock markets at different time scales, while Jammazi et al. (2017a) focus on the relationship between the stock markets and interest rates. Kanda et al. (2018) analyze the causality between equity returns and currency returns, and Sibande et al. (2019) consider the relationship between stock market and unemployment; both studies focus on the UK and make use of a long time span involving two centuries of data. Gupta et al. (2019) study causality between oil prices and the US financial stress, and Coronado et al. (2020) correlate the US stock market and currency. Bathia et al. (2021) focus on unemployment and currency returns in the UK, while Gupta et al. (2021a) analyze the relationship between the US stock market movements and the presidential approval ratings. Further, Gupta et al. (2021b) monitor the impact of a news-based indicator of infectious diseases on the US treasury securities, while Zhang et al. (2021) evaluate the spillover between Bitcoin prices and internet attention.

While the DCC-Hong test is appealing from the empirical viewpoint, the statistical properties of the test are not known. In fact, even the authors acknowledge that in order to evaluate the null of absence of causality by using the DCC-Hong test statistic, the use of critical values based on the Normal distribution represents a rough approximation to the reality. By beginning from this observation, we provide a first contribution and obtain by simulation the critical values of the DCC-Hong test statistic. Our analyses indicate that the distribution of the test statistic (under the null hypothesis) is far from being Normal, is characterised by a large right tail, and depends on both the unconditional correlation between the analysed series and the sample size. Overall, we note that the use of Gaussian critical values lead to an over-rejection of the null hypothesis. When moving to the power of the test, the simulations we run compare the DCC-Hong to the rolling Hong test Hong (2001), thereby highlighting that when the sample size is relatively small, the DCC-Hong might be used, but when the sample larger than 200 observations, the rolling Hong test has better power in identifying the presence of causality. Further, while both tests do have good power in the presence of a strong causality between variables, the DCC-Hong has superior performances in cases where the causality link has a reduced intensity.

Thus, the simulation evidence thus challenges the validity of the empirical analyses based on the evaluation of the DCC-Hong test statistic and its comparison with Normal critical values, beginning from those in the paper by Lu et al. (2014). Therefore, we proceed to a replication of the evidence in Lu et al. (2014) and limiting our analyses to the causality between the Brent and WTI futures prices. We show that by using the simulated critical values, the evidence of causality dramatically reduces, and is focused on specific periods: the Iraq invasion in 2003, at the peak of oil prices in 2006, following the European sovereign crisis

Electronic copy available at: https://ssrn.com/abstract=3941778

in 2010, and after the production cuts during the Arab spring in 2011. On the contrary, the use of the DCC-Hong with normal critical values identified a much striking presence of causality which, unfortunately, is only apparent, due to the inappropriateness of the approximation of the test statistic with the Normal distribution. We also complement the replication with an analysis focused on a longer time sample, which also covers the COVID-19 pandemic. Both the rolling Hong and the DCC-Hong test identify occurrences of causality in the first part of 2020. In addition, the evaluation of the DCC-Hong test statistic over the entire sample leads us to identify a further weakness of this approach: as the DCC-Hong test is built on a set of estimated models, if the causality exists in limited periods of time, the estimated models would not identify its presence because they will be driven by the more relevant periods of non-causality. This calls for a rolling evaluation of the DCC-Hong test which, however, would limit its relevance, as the construction of the DCC-Hong test would become more computationally intensive than the simplest rolling Hong test. The only case in which the DCC-Hong test would provide a valuable information is for time series of reduced length.

The remainder of the paper is organized in the following manner: Section 2 reviews the Granger causality test given by Hong (2001) and Lu et al. (2014) and presents simulation evidences. Section 3 replicates some of the evidence in Lu et al. (2014) and includes insights on the causality on the period characterized by the diffusion of the COVID-19 pandemic. Section 4 concludes the paper.

2. Time-varying causality testing approaches

Similarly to Lu et al. (2014), we first introduce the Hong (2001) causality test. Let us denote by $x_{1,t}$ and $x_{2,t}$ the two series of interest and assume we are willing to evaluate if $x_{1,t}$ causes $x_{2,t}$ using a sample of T observations. Hong (2001) proposes a test for causality that generalizes the contribution of Cheung and Ng (1996). Hong (2001) indicates the introduction of a test for variance causality, building on the cross-correlation between two series of centered squared standardized innovations obtained by fitting ARMA-GARCH models on two time series. In our case, the focus is on mean causalityand, thus, we assume that the $x_{1,t}$ and $x_{2,t}$ series have been pre-filtered by appropriate ARMA-GARCH models. Therefore, we focus on their mean cross-correlation— that is, we do not square them.[1] The Hong (2001) test (hereafter, the Hong test) builds on the following statistic:

$$Q_H = \frac{T \sum_{j=1}^{T-1} k^2\left(\frac{j}{M}\right) \hat{\rho}_{2,1}^2\left(j\right) - C_{1T}\left(k\right)}{\sqrt{2D_{1T}\left(k\right)}}, \tag{1}$$

where

[1] We filter out the conditional variance dynamic in order to be coherent with the approach put forward by Lu et al. (2014) that eliminates the conditional variance dynamic before testing the null of zero cross-correlation.

Electronic copy available at: https://ssrn.com/abstract=3941778

$$C_{1T}\left(k\right)=\sum_{j=1}^{T-1}\left(1-\frac{j}{T}\right)k^{2}\left(\frac{1}{M}\right),\quad D_{1T}\left(k\right)=\sum_{j=1}^{T-1}\left(1-\frac{j}{T}\right)\left(1-\frac{j+1}{T}\right)k^{4}\left(\frac{1}{M}\right),$$

$$\hat{\rho}_{2,1}^{2}\left(j\right)=\frac{\sum_{t=j+1}^{T}x_{2,t}x_{1,t-j}}{\sqrt{\sum_{t=j+1}^{T}x_{1,t}^{2}}\sqrt{\sum_{t=j+1}^{T}x_{2,t}^{2}}},\quad k\left(z\right)=\left(1-\left|z\right|\right)\mathcal{I}\left(\left|z\right|<1\right),$$

and $j, M > 0$. Note that $\hat{\rho}_{2,1}^{2}\left(j\right)$ is the cross-correlation between $x_{1,t-j}$ and $x_{2,t}$, where the first term is lagged, and we adopt the Bartlett Kernel function $k\left(z\right)$ (it must be noted that other kernel functions can be used; see Hong, 2001). Finally, M is a lag truncation value inducing zero contribution for cross-correlations having a lag $j > M$— that is only the cross-correlations up to lag M contribute to the causality evaluation.

The Hong test statistic in 1 can be used to detect causality from $x_{2,t}$ to $x_{1,t}$. A statistic for detecting bidirectional causality is also available.

$$Q_{BH}=\frac{T\sum_{|j|=1}^{T-1}k^{2}\left(\frac{j}{M}\right)\hat{\rho}_{2,1}^{2}\left(j\right)-C_{2T}\left(k\right)}{\sqrt{2D_{2T}\left(k\right)}},\tag{2}$$

where

$$C_{2T}\left(k\right)=\sum_{|j|=1}^{T-1}\left(1-\frac{|j|}{T}\right)k^{2}\left(\frac{1}{M}\right),\quad D_{1T}\left(k\right)=\sum_{|j|=1}^{T-1}\left(1-\frac{|j|}{T}\right)\left(1-\frac{|j|+1}{T}\right)k^{4}\left(\frac{1}{M}\right).$$

The two test statistics have known asymptotic distributions, as derived by Hong (2001):

$$Q_{H}\longrightarrow\mathcal{N}\left(0,1\right),\quad Q_{BH}\longrightarrow\mathcal{N}\left(0,1\right).\tag{3}$$

For both test statistics, upper tail critical values must be used, as the rejection of the null is associated with large positive values of the test statistic: when we observe non-null and relevant cross-correlations, they are included, squared, in the test statistic.

In contrast to Lu et al. (2014), but coherently with Hong et al. (2009), we exclude the contemporaneous cross-correlation from the evaluation of causality. This is a clear difference between our approach and that of Lu et al. (2014), which might lead to a few differences in the actual data analyses; further details on this aspect are provided in the following section. The choice of excluding the contemporaneous correlation enables the detection of dynamic causality links among variables which are simultaneously traded. In Lu et al. (2014), the reader can find a discussion supporting the introduction of a variation of the Hong test to account for instantaneous correlation when markets are characterized by asynchronous trading; here, we do not consider such a possibility.[2]

Furthermore, similar to Lu et al. (2014), we apply the Hong test both at the entire sample

[2] In the following section, we also motivate our choice from an empirical perspective, when replicating Lu et al. (2014) study.

Electronic copy available at: https://ssrn.com/abstract=3941778

level as well as by resorting to a rolling procedure. In this case, the Hong test statistic is evaluated on samples of size $S < T$, rolled forward by one observation at a time, thereby leading to a total number of $S-T$ test statistics (from sample $[1, S]$ to sample $[T-S+1, T]$).

Lu et al. (2014) propose a novel approach to causality detection, the dynamic conditional correlation Hong test (or DCC-Hong). They motivate this addition to the literature due to the possible shortcoming of using cross-correlations evaluated on a rolling window: the lack of appropriate dynamic modeling of cross-correlations which makes them less reactive to the most recent information from the market. Inspired by the work of Engle (2002), where the author introduces the DCC as a more appropriate tool for correlation modeling, Lu et al. (2014) propose propose the evaluation of the evolution of cross-correlations in a dynamic manner. Mimicking the DCC model of Engle (2002), they suggest to first model the covariance matrix of $y_{t,j} = [x_{2,t} \quad x_{1,t-j}]'$ and call it $Q_t\left(j\right)$, as given below:

$$Q_t = \left(1-\alpha_j-\beta_j\right)\overline{R} + \alpha_j y_{t,j} y'_{t-1,j} + \beta_j Q_{t-1}, \tag{4}$$

$$r_{i,l,t}\left(j\right) = \left(1-\alpha_j-\beta_j\right)\overline{r}_{2,1} + \alpha_j x_{i,t-1} x_{l,t-j-1} + \beta r_{i,l,t-1}\left(j\right), \quad i,l = 1,2 \tag{5}$$

where $\overline{R}$ is the full-sample unconditional cross-correlation matrix (i.e. ones over the main diagonal and the cross-correlations off-diagonal). Building on the model estimates, they recover the cross-correlation by the following standardization

$$\rho_{2,1,t}\left(j\right) = \frac{r_{2,1,t}\left(j\right)}{\sqrt{r_{1,1,t}\left(j\right)}\sqrt{r_{2,2,t}\left(j\right)}}. \tag{6}$$

The test statistics suggested by Lu et al. (2014) are then set equivalent to Q_H and Q_{BH}:

$$Q_{DH,t} = \frac{T\sum_{j=1}^{T-1} k^2\left(\frac{j}{M}\right)\hat{\rho}^2_{2,1,t}\left(j\right) - C_{1T}\left(k\right)}{\sqrt{2D_{1T}\left(k\right)}}, \tag{7}$$

and

$$Q_{BDH,t} = \frac{T\sum_{|j|=1}^{T-1} k^2\left(\frac{j}{M}\right)\hat{\rho}^2_{2,1,t}\left(j\right) - C_{2T}\left(k\right)}{\sqrt{2D_{2T}\left(k\right)}}. \tag{8}$$

The use of the Bartlett Kernel in (7) and (8) allows the evaluation of only a small number of DCC-like models (i.e. only M), thereby making the entire procedure computationally feasible, even if more demanding than the rolling Hong ones. Given a sample of size T, the DCC filter is applied $2M$ times, with $2M$ different lead/lag values for j (M varies from $-M$ to M, excluding the zero value). Then, the DCC filter yields $2M$ estimated paths for the cross-correlations, thereby enabling the recovery of the test statistic for each point in time from $t = M+1$ to $t = T-M$. Such a feature of the DCC-Hong test makes it a proper alternative to the use of the rolling Hong procedure.

The introduction of the DCC-like dynamic in the cross-correlations appropriately enables the capturing of the dynamic evolution in the interdependence between two series. In this setting, the DCC model might be seen as a filter, thereby enabling the detection of whether a quantity of interest has a dynamic behaviour, even without adequately specifying a com-

Electronic copy available at: https://ssrn.com/abstract=3941778

plete model. We note that this is in line with the arguments of Caporin and McAleer (2012) for interpreting the DCC of Engle (2002) as a filter. Nevertheless, we emphasize that an appropriate model yielding dynamic cross-correlations is a VAR model with time-varying parameters (TVP-VAR). In fact, as the cross-correlation is a function of the model parameters, if the latter are dynamic then the former is also dynamic. We will use this intuition in the following section when designing the data-generating processes for our simulation study.

Building on heuristic arguments, Lu et al. (2014) suggest that the DCC-Hong tests (unidirectional and bidirectional) might be roughly approximated by a standardized Normal distribution under the null of the absence of causality. Lu et al. (2014) acknowledge that deriving the asymptotic distribution of the DCC-Hong test is particularly complex. In fact, under the null hypothesis, the β_j are nuisance parameters. Within the DCC model, such a situation requires specific procedures to test the null hypothesis of constant correlations, as indicated out by Engle and Sheppard (2001). The impact of the nuisance parameters under the null of constant correlation is also relevant in our case, as it corresponds to the presence of nuisance parameters in the evaluation of dynamic cross-correlations. Moreover, the influence of those nuisance parameters transmits to the dynamic cross-correlations and, subsequently, to the DCC-Hong test statistic. Therefore, the use of a Normal distribution, quoting Lu et al. (2014), enables only a *rough judgment*. We are aware that the derivation of the asymptotic distribution of the test statistic is complex, but adequate knowledge of its behavior is crucial for the appropriate use of the test. Therefore, in the following sub-section, we shed light on the distribution of the test statistic proposed by Lu et al. (2014) by resorting to a simulation study. This will allow us to recover simulated critical values under the null hypothesis, and to contrast the critical values under the assumption of Normality.

2.1. A simulation study

The first objective of our Monte Carlo study, is the evaluation of the size and power of the DCC-Hong test and to compare them with those of the (rolling) Hong test; for the latter, both a full-sample estimation of the test statistic and a rolling evaluation scheme must be taken into account. We consider different cases associated with alternative designs of the data-generating process. In all cases, we do not include a conditional variance dynamic, as our purpose is to test for causality in the mean of variables that have been pre-filtered with an ARMA-GARCH process.[3]

Case 1. The first simulation set focuses on the size and power of the Hong and DCC-Hong tests when the data generating process (GDP) is a VAR(1) model, namely

$$\begin{bmatrix} x_{1,t} \\ x_{2,t} \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} \phi_{1,1} & \phi_{1,2} \\ \phi_{2,1} & \phi_{2,2} \end{bmatrix} \begin{bmatrix} x_{1,t-1} \\ x_{2,t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix} \tag{9}$$

We simulate the innovation term ε_t from a Gaussian density with unit variances and correlation set to 0.5. In addition, for simplicity, we set the mean to be zero. Furthermore, for the parameters driving the dynamic, we set the diagonal coefficients $\phi_{1,1}$ and $\phi_{2,2}$ to be both equal to 0.5, while we always set $\phi_{2,1} = 0$. The coefficient $\phi_{1,2}$ is used to introduce

[3] Unreported results for the data-generating processes with innovations that include a GARCH term in the conditional variances provide results that are in keeping with those included in the present section.

Electronic copy available at: https://ssrn.com/abstract=3941778

Granger-type causality in the relationship between the two variables of the model: a non-null value implies causality from variable 2 to variable 1. The sample size of the simulated series takes a value of 500, 1000, or 2000, while we run 1000 replications in all experiments. A pre-sample of 1000 observations is introduced to avoid any dependence on starting values. We test for causality using the Hong test with the Bartlett kernel and the value of M equal to either 10 or 20. We report both the unidirectional tests as well as the bidirectional test. In order to filter out the serial dependence from each simulated series, we fit a simple AR(1) process and apply the Hong test on the innovations. For the Hong test, we evaluate the cross-correlations using the entire time series of the innovations. For the DCC-Hong test, we use the same M as for the Hong test and evaluate the test at the end of the sample. Moreover, for both tests, we use critical values associated with a standardized normal.

			Hong			Norm. DCC-Hong			Sim. DCC-Hong		
T	M	$\phi_{1,2}$	$2 \to 1$	$1 \to 2$	$1 \leftrightarrow 2$	$2 \to 1$	$1 \to 2$	$1 \leftrightarrow 2$	$2 \to 1$	$1 \to 2$	$1 \leftrightarrow 2$
500	10	0	0.065	0.053	0.065	0.287	0.276	0.377	0.050	0.056	0.057
1000	10	0	0.053	0.055	0.060	0.281	0.293	0.383	0.055	0.050	0.052
2000	10	0	0.046	0.058	0.036	0.230	0.276	0.342	0.046	0.044	0.050
500	20	0	0.068	0.059	0.061	0.404	0.401	0.543	0.051	0.061	0.051
1000	20	0	0.044	0.052	0.040	0.399	0.399	0.548	0.051	0.045	0.052
2000	20	0	0.048	0.056	0.055	0.347	0.390	0.502	0.047	0.050	0.054
500	10	0.2	0.935	0.047	0.852	0.918	0.270	0.903	0.438	0.051	0.238
1000	10	0.2	1.000	0.046	0.997	0.992	0.264	0.987	0.832	0.042	0.615
2000	10	0.2	1.000	0.055	1.000	0.999	0.264	1.000	0.987	0.042	0.964
500	20	0.2	0.874	0.053	0.739	0.928	0.396	0.913	0.250	0.049	0.149
1000	20	0.2	0.997	0.052	0.985	0.992	0.383	0.987	0.611	0.037	0.341
2000	20	0.2	1.000	0.056	1.000	0.998	0.372	1.000	0.960	0.052	0.809

Table 1: Size and power of the Hong and DCC-Hong tests under the DGP of Case 1 at the 5% confidence level. For the DCC-Hong test, we report the size and power obtained from normal critical values as well as for critical values recovered by simulations (with innovations' correlation set to 0.5).

Table 1 confirms the good size and power of the Hong test, which is in keeping with the evidences in Hong (2001). The size and power are not affected by the sample size and are only slightly impacted by the choice of M. On the contrary, when using the standardized normal quantiles, as in Lu et al. (2014), the DCC-Hong test is characterized by a clear oversize, which only marginally decreases with an increase in the sample size. Moreover, the size evidently reduces when M increases.

However, while for the Hong test, the asymptotic distribution has been derived in Hong (2001), for the DCC-Hong only heuristic arguments have been used to support the use of a standardized normal as an approximation. In Figures (1) and (2), we report the Kernel density estimates of the DCC-Hong test statistics (for two unidirectional tests and for the bidirectional test) corresponding to the simulations under the null hypotheses used in Table 1. The plot clearly reveals that the test statistic does not possess a standard Normal density.

Electronic copy available at: https://ssrn.com/abstract=3941778

We note a clear asymmetry and a very thick right tail. Moreover, by increasing the value of M, the deviation from the Gaussian increases. Such evidence gives rise to the oversize noted in Table 1. Therefore, our simulations suggest that the derivation of the proper asymptotic distribution of the test statistic might represent a challenging topic for research. In fact, asymptotic results are not yet available for the DCC model of Engle (2002), an aspect that makes the asymptotic analysis on the DCC-Hong test rather complex. Furthermore, following our discussion in the previous section, the non-standard form of the test statistic distribution could be a by-product of the presence of nuisance parameters in the DCC filter.

In keeping with the literature on testing in the presence of nuisance parameters, we recover critical values for the test by running simulations under the DGP of this first case. Given the presence of a large right tails in the kernel densities of Figures (1) and (2), we run 10,000 replications in order to recover critical values. Table 2 reports the simulated critical values for selected values of M, from the sample size T, of the correlation between the innovations, ρ, and for various confidence levels.

The differences between the critical values reported in table and the quantiles of the normal is striking. In addition, the distribution of the test statistics depends on the level of unconditional correlation between the series. Moreover, it appears that the distribution of the test statistic is impacted by the sample size. Our hypothesis is that this depends on the relevance of the upper tail of the test statistic, thereby requiring longer samples and a large number of simulations (larger than the one we adopt) to adequately measure the critical values. A further possibility, is that by adopting a longer time series, the estimation of the filter (i.e. of its parameter) underlying the DCC-Hong test is more precise with a longer sample.

The difference in the test statistic distribution from the normal hypothesized by Lu et al. (2014) challenges their findings and sheds additional light on the performances of the DCC-Hong test. In fact, by re-evaluating the DCC-Hong test size and power under Case 1 and using the simulated critical values, we do note that the test has an appropriate size (obviously) and its power properties are reasonable, even though the power of the Hong test is higher. The over-rejection that characterized the DCC-Hong test when using Gaussian quantiles has an impact not only on the evidence in Lu et al. (2014), but an all the studies, cited in the introduction, that adopt it.

Apart from the need to use simulated critical values, the simulation above already shows evidence that caution must be exercised in the use of the DCC-Hong testing approach, as the critical values depend on a number of elements. In the settings considered above, the power of the DCC-Hong test is in line with that of the Hong test. Moreover, the performances of the Hong test when the underlying cross-correlation functions are evaluated on a shorter window are discussed under the following case.

Case 2. In this second simulation set, we conduct a simple assessment of the appropriateness of the Hong and DCC-Hong tests when the causality changes over time in a simple way— that is with a structural break in the parameters. In detail, the DGP is equivalent to that of Case 1, but for parameter $\phi_{1,2}$ which takes a value of 0 up to the middle of the sample and a value of 0.2 or 0.7 afterwards. We maintain all the other settings as in Case 1 with the addition of a second implementation for the Hong test. In fact, apart from evaluating the test on the full residuals sample, we also evaluate the test by focusing on the last 100 observations. This is in line with the specification adopted by Lu et al. (2014) and will allow

Electronic copy available at: https://ssrn.com/abstract=3941778

			$T = 500$			$T = 1000$			$T = 2000$		
M	ρ	α	$2 \to 1$	$1 \to 2$	$1 \leftrightarrow 2$	$2 \to 1$	$1 \to 2$	$1 \leftrightarrow 2$	$2 \to 1$	$1 \to 2$	$1 \leftrightarrow 2$
10	0	0.1	4.42	4.19	5.41	4.35	4.72	5.59	4.29	4.44	5.48
10	0	0.05	6.67	6.49	7.55	6.84	6.98	7.96	6.58	6.91	8.02
10	0	0.01	13.42	13.55	13.65	14.76	15.65	15.55	15.82	16.22	15.90
10	0.25	0.1	4.29	4.16	5.30	4.43	4.57	5.62	4.22	4.32	5.28
10	0.25	0.05	6.62	6.40	7.51	6.67	6.94	8.03	6.53	6.58	7.97
10	0.25	0.01	13.71	13.68	13.41	14.27	15.46	15.58	15.54	14.37	15.47
10	0.5	0.1	3.97	3.90	4.94	4.12	3.94	5.03	4.06	3.99	5.04
10	0.5	0.05	6.14	6.02	7.15	6.24	6.30	7.38	6.19	6.43	7.43
10	0.5	0.01	12.88	12.47	13.39	14.13	13.28	14.88	14.30	14.37	15.06
10	0.75	0.1	3.56	3.53	4.42	3.56	3.46	4.59	3.47	3.49	4.48
10	0.75	0.05	5.54	5.53	6.72	5.77	5.51	6.94	5.64	5.57	6.89
10	0.75	0.01	12.27	12.05	13.55	12.89	12.03	14.16	13.62	12.30	15.06
20	0	0.1	5.62	5.53	6.89	5.86	5.83	7.33	5.71	5.56	7.10
20	0	0.05	8.03	7.97	9.16	8.55	8.57	10.12	8.53	8.41	9.88
20	0	0.01	14.61	14.71	15.16	16.78	16.58	17.52	16.08	17.65	17.53
20	0.25	0.1	5.58	5.45	6.79	5.77	5.57	6.99	5.83	5.69	7.14
20	0.25	0.05	8.04	7.99	9.31	8.31	8.15	9.72	8.42	8.33	9.71
20	0.25	0.01	15.03	14.92	15.27	15.60	16.66	17.05	16.34	15.77	16.60
20	0.5	0.1	5.38	5.32	6.60	5.40	5.38	6.79	5.36	5.32	6.74
20	0.5	0.05	7.68	7.67	9.00	7.85	7.68	9.10	8.03	7.89	9.51
20	0.5	0.01	14.58	14.69	15.68	14.76	15.25	16.17	15.45	15.84	17.02
20	0.75	0.1	5.09	5.03	6.44	5.07	5.07	6.55	5.07	4.95	6.51
20	0.75	0.05	7.37	7.32	9.08	7.49	7.26	8.96	7.82	7.50	9.58
20	0.75	0.01	14.71	14.36	16.66	15.23	14.59	16.72	17.15	16.13	19.60

Table 2: Simulated critical values for the DCC-Hong test under the null hypothesis and the data-generating process of Case 1. The values are based on 10.000 simulations for different levels of the unconditional correlation, M, and confidence levels.

Electronic copy available at: https://ssrn.com/abstract=3941778

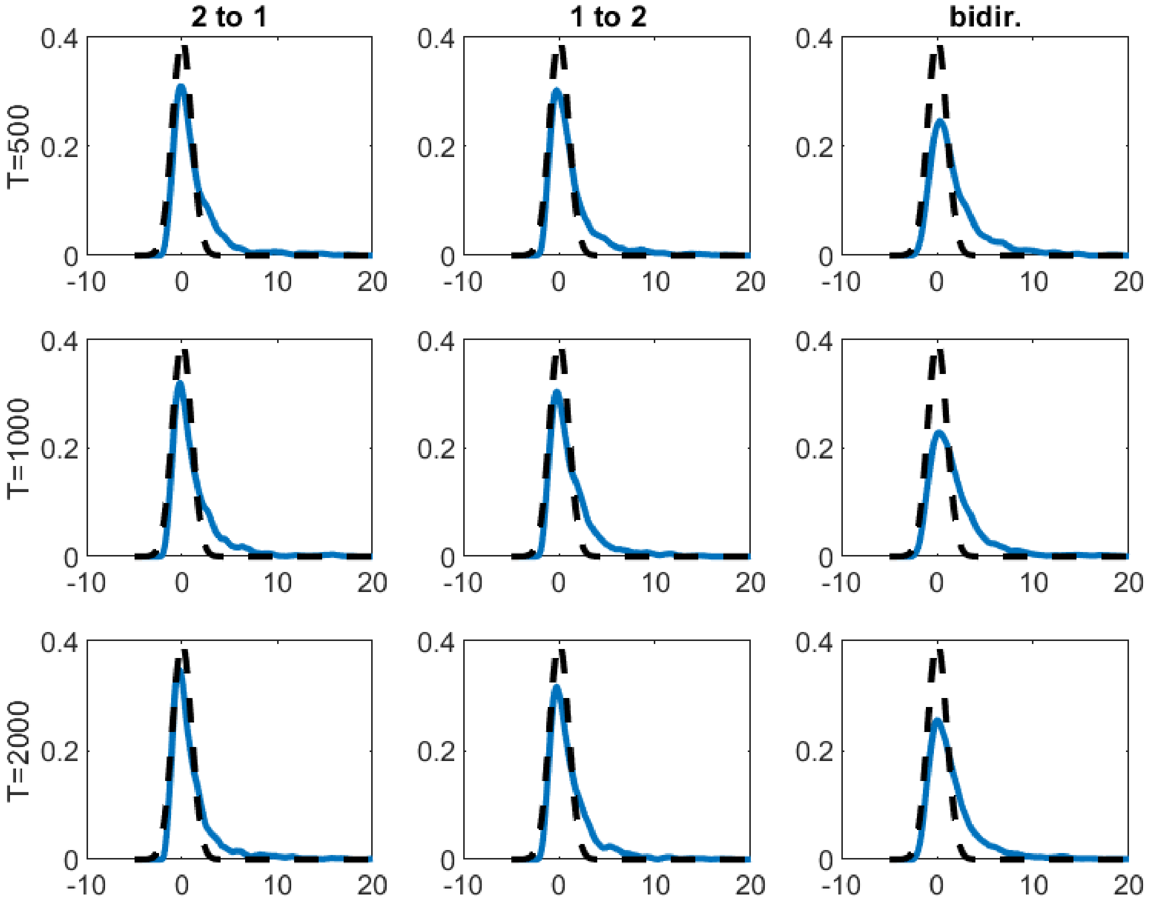

Figure 1: Kernel density estimate of the DCC-Hong test statistics (in blue) for different sample sizes under the data-generating process of Case 1 with $M = 10$; results based on 1000 experiments. Dashed black is used for the standardized Normal density.

Electronic copy available at: https://ssrn.com/abstract=3941778

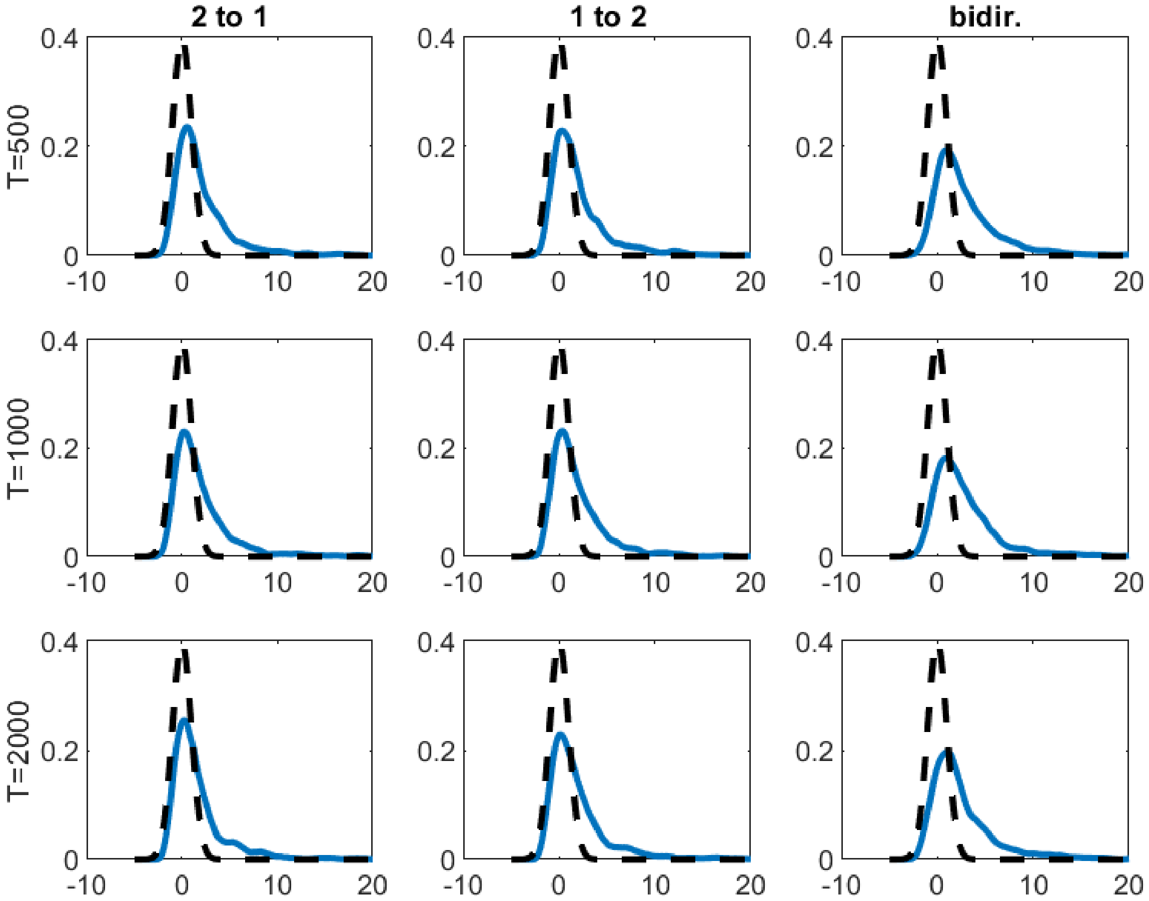

Figure 2: Kernel density estimate of the DCC-Hong test statistics (in blue) for different sample sizes under the data-generating process of Case 1 with $M = 20$; results based on 1000 experiments. Dashed black is used for the standardized Normal density.

Electronic copy available at: https://ssrn.com/abstract=3941778

us to validate the use of a rolling approach for the Hong test when the causality relationship changes after an external event (which causes a break in the relation between the two series). Moreover, this second specification for the Hong test will allow us to compare its performances to the DCC-Hong which, by construction, dynamically adapts to the evolution of the series: by using a shorter window for the Hong test evaluation, we induce the test statistic to be more reactive to possible changes in the time series.

			Hong : Full			Hong : 100			DCC-Hong		
T	M	$\phi_{1,2}$	$2 \to 1$	$1 \to 2$	$1 \leftrightarrow 2$	$2 \to 1$	$1 \to 2$	$1 \leftrightarrow 2$	$2 \to 1$	$1 \to 2$	$1 \leftrightarrow 2$
500	10	0.2	0.350	0.052	0.238	0.375	0.061	0.287	0.278	0.056	0.185
1000	10	0.2	0.618	0.050	0.456	0.363	0.061	0.277	0.515	0.048	0.403
2000	10	0.2	0.926	0.052	0.853	0.360	0.082	0.277	0.765	0.046	0.665
500	20	0.2	0.240	0.055	0.168	0.287	0.065	0.224	0.176	0.039	0.115
1000	20	0.2	0.489	0.057	0.361	0.293	0.074	0.204	0.366	0.055	0.242
2000	20	0.2	0.851	0.061	0.705	0.312	0.060	0.218	0.654	0.053	0.505
500	10	0.7	1.000	0.048	0.997	0.999	0.069	0.998	0.995	0.063	0.991
1000	10	0.7	1.000	0.049	1.000	0.999	0.066	0.996	1.000	0.042	1.000
2000	10	0.7	1.000	0.056	1.000	0.999	0.079	0.998	1.000	0.045	1.000
500	20	0.7	0.996	0.053	0.990	0.999	0.075	0.988	0.988	0.055	0.957
1000	20	0.7	1.000	0.060	1.000	0.996	0.072	0.987	0.998	0.055	0.997
2000	20	0.7	1.000	0.062	1.000	0.998	0.073	0.994	1.000	0.055	1.000

Table 3: Size and power of the Hong and DCC-Hong tests under the DGP of Case 2. For DCC-Hong, we adopt simulated critical values with the innovation correlation set to 0.5.

We begin by focusing on the size of the tests— that is, we consider the causality from variables 1 to 2, which is absent in all simulations, irrespective of the break. These results are in line with those of Case 1, both for the Hong and DCC-Hong tests. The results related to the power are interesting. In fact, when the causality is stronger ($\phi_{1,2} = 0.7$) both specifications of the Hong test as well as the DCC-Hong test have very good power. In contrast, for mild causality ($\phi_{1,2} = 0.2$), both tests suffer due to a decrease in power. Moreover, the power varies both with the sample length and the value of M. For the Hog test, when the sample size is relatively small— that is 100 observations— the power is limited, being approximately 30% for the case when $M = 20$ and somewhat larger for $M = 10$; this evidence challenges the use of the Hong test within a rolling scheme. In contrast, when focusing on the entire sample results for the Hong test and the DCC-Hong, we observe a larger power for the former, for all sample sizes and values of M. This evidence suggests that the Hong test performances are better when longer time series are available, while the DCC-Hong could be adopted with shorter time series to confirm the evidence of the Hong test.

Case 3. This last DGP is coherent with a smoother but continuous variation in the parameters driving the causality. We simulate time series from a time-varying parameters VAR (TVP-VAR) model defined in the following manner:

Electronic copy available at: https://ssrn.com/abstract=3941778

$$\begin{bmatrix} x_{1,t} \\ x_{2,t} \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} \phi_{1,1,t} & \phi_{1,2,t} \\ \phi_{2,1,t} & \phi_{2,2,t} \end{bmatrix} \begin{bmatrix} x_{1,t-1} \\ x_{2,t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix} \tag{10}$$

$$\begin{bmatrix} \phi_{1,1,t} \\ \phi_{1,2,t} \\ \phi_{2,1,t} \\ \phi_{2,2,t} \end{bmatrix} = \begin{bmatrix} \bar{\phi}_{1,1} \\ \bar{\phi}_{1,2} \\ \bar{\phi}_{2,1} \\ \bar{\phi}_{2,2} \end{bmatrix} + \rho \left(\begin{bmatrix} \phi_{1,1,t-1} \\ \phi_{1,2,t-1} \\ \phi_{2,1,t-1} \\ \phi_{2,2,t-1} \end{bmatrix} - \begin{bmatrix} \bar{\phi}_{1,1} \\ \bar{\phi}_{1,2} \\ \bar{\phi}_{2,1} \\ \bar{\phi}_{2,2} \end{bmatrix} \right) + \sigma \begin{bmatrix} \eta_{1,1,t} \\ \eta_{1,2,t} \\ \eta_{2,1,t} \\ \eta_{2,2,t} \end{bmatrix},$$

where the time-varying autoregressive parameters follow independent AR(1) processes with the same variance level, σ, for the evolution of AR parameters ($\eta_{i,j,t}$ is zero mean and unit variance i.i.d. noise). The parameter vector $\bar{\Phi}$ contains the the unconditional level of the time-varying parameters. We consider several cases for the model parameters by combining different levels of persistence of time-varying AR parameters (the value of ρ), thereby nearing a random walk dynamic which is often used in empirical applications, different levels of the parameters' innovation variance, and different unconditional levels of the time-varying parameter $\bar{\phi}_{1,2}$ to induce causality; the value of $\bar{\phi}_{2,1}$ is always set to 0, while $\bar{\phi}_{1,1} = \bar{\phi}_{2,2} = 0.5$. We note that these settings induce causality at the unconditional level (on the parameters) but the existence of causality might become less evident (or stronger) due to the time-varying evolution of the model parameters. The time-varying nature of the parameters also make the GDP coherent with the intuition that leads to the DCC-Hong test: it induces time-varying cross-correlation functions. Similar to the previous cases, we consider three different sample sizes, with $T = \{500, 1000, 2000\}$; two different levels for M, either 10 or 20; and the three tests for causality (two unidirectional and one bidirectional). Finally, for the Hong test, we run it by considering the entire sample size as well as last 100 or 200 observations only.

When the data do not unconditionally reveal the the existence of causality, both the Hong and DCC-Hong test exhibit good size properties which are not affected by different parameter settings nor by varying the size of the sample adopted for the evaluation of the test statistics.

Moving to the simulations assessing the power, we note again that both tests have appropriate power when the causality is stronger ($\phi_{1,2} = 0.7$); this result is confirmed for different sample sizes, different values of M, and different parameters settings. On the contrary, when the causality is mild ($\phi_{1,2} = 0.2$), we obtain results similar to those of Case 2. In fact, when focusing on the entire sample, the Hong test has better power than the DCC-Hong test, improving with the sample size and not much affected by the parameters settings. On the contrary, for small sample sizes, the power is relatively low, and worse than that of the DCC-Hong test. For the latter, the power increases with the sample size and is only marginally affected by the simulation settings, apart from a decrease with increasing M. We note that the simulations we provide also provide certain guidelines on the sample size at which the Hong test begins to have better power than the DCC-Hong test. In fact, when focusing on the results based on 200 observations and comparing them to the power of the DCC-Hong based on 500 observations, we note that the former has better power. Thus, we believe that with 200 observations and more, the power of the Hong test is higher than the power of the DCC-Hong test.

Electronic copy available at: https://ssrn.com/abstract=3941778

We link this evidence both to the test performances and to the DGP we consider, where causality is present at an unconditional level in the parameters: this might be better captured with longer samples. On the contrary, when samples are shorter, the parameters' dynamic impacts the test performances. This is coherent with the fact the power reduces when the causality is not strong ($\bar{\phi_{1,2}} = 0.2$), while for $\bar{\phi_{1,2}} = 0.7$ both tests have very good power levels.

The simulation result highlights the difficulties associated with the use of the Hong test within a rolling exercise, due to its low power when the causality is not strong and the rolling sample has a limited length. In this case, the DCC-Hong could be used as it has appropriate size and better power, when correct critical values are adopted. In contrast, when the sample size is larger than 200 observations, the Hong test has better power and is preferred.

3. Causality between oil prices: the case of Brent and WTI

In this section, we replicate some of the empirical evidence included in Lu et al. (2014). Specifically, we focus on the use of the rolling and DCC Hong tests and consider the two references indexes for the global crude oil markets: the West Texas Intermediate (WTI) crude oil and the Brent crude oil. In contrast to Lu et al. (2014), we do not jointly evaluate the causality between future prices, the WTI and Brent prices mentioned earlier, and spot prices, Dubai and Tapis spot prices, adopted in Lu et al. (2014). In fact, we prefer to avoid combining prices of different nature in the analyses, that is— on the one hand we do have a financial contract, while on the other hand we have a price derived from the exchange of large physical amounts of oil.

We downloaded the daily closing WTI and Brent futures prices from Bloomberg (CL1 Comdty and CO1 Comdty, respectively) from January 2002 to September 2021. Then, we computed the returns as $x_{i,t} = \log(P_{i,t}) - \log(P_{i,t-1})$, where $i = \{\text{WTI}, \text{BRENT}\}$. The analysis involves two periods: i) The period from 3 January 2002 to 19 March 2012, as considered in Lu et al. (2014); and ii) the entire period from 3 January 2002 to 2 September 2021, which also encompasses the outbreak of the COVID-19 pandemic. The former is included to highlight the differences in the test outcomes due to the introduction of contemporaneous term in the equations of the Rolling and DCC-Hong tests performed in Lu et al. (2014).[4] The descriptive statistics of returns of WTI and Brent are included in Table 6.

With regard to the first period, the returns show some discrepancies with respect to Lu et al. (2014). This could be attributed to the use of different data providers. In both samples, the mean of returns is small, the standard deviation is high, and the returns are not normally distributed, as shown by the Jarque-Bera test. This is also highlighted by the negative skewness and the large kurtosis. The augmented Dickey-Fuller (ADF) test reveals that the returns are stationary and the Box-Pierce statistics on the lag 5 and 10 confirms the presence of serial correlations.[5] Once again, data confirms that the ARMA-GARCH model represents an appropriate choice to account for the presence of heteroscedasticity and autocorrelations in the oil returns.

[4] In their work, the authors include the case $j = 0$, which corresponds to contemporaneous correlations.

[5] We have also performed the ADF test on the log prices. The test does not reject the null hypothesis of a unit root in both WTI and Brent for both periods.

Electronic copy available at: https://ssrn.com/abstract=3941778

					Hong : Full			Hong : 100			Hong : 200			DCC-Hong		
T	M	$\phi_{1,2}$	σ	ρ	$2 \to 1$	$1 \to 2$	$1 \leftrightarrow 2$	$2 \to 1$	$1 \to 2$	$1 \leftrightarrow 2$	$2 \to 1$	$1 \to 2$	$1 \leftrightarrow 2$	$2 \to 1$	$1 \to 2$	$1 \leftrightarrow 2$
500	10	0	0.001	0.9	0.065	0.065	0.058	0.069	0.071	0.079	0.080	0.070	0.082	0.063	0.049	0.054
1000	10	0	0.001	0.9	0.079	0.082	0.081	0.074	0.075	0.072	0.075	0.073	0.091	0.056	0.049	0.050
2000	10	0	0.001	0.9	0.101	0.107	0.118	0.070	0.078	0.085	0.087	0.075	0.089	0.055	0.057	0.064
500	20	0	0.001	0.9	0.069	0.073	0.062	0.080	0.074	0.078	0.083	0.070	0.086	0.061	0.043	0.059
1000	20	0	0.001	0.9	0.069	0.081	0.066	0.075	0.068	0.075	0.074	0.080	0.089	0.048	0.050	0.050
2000	20	0	0.001	0.9	0.094	0.090	0.101	0.066	0.063	0.068	0.079	0.074	0.088	0.051	0.061	0.051
500	10	0.2	0.001	0.9	0.747	0.055	0.612	0.212	0.068	0.187	0.385	0.072	0.288	0.270	0.049	0.171
1000	10	0.2	0.001	0.9	0.967	0.069	0.921	0.224	0.069	0.168	0.392	0.075	0.305	0.616	0.048	0.415
2000	10	0.2	0.001	0.9	1.000	0.082	1.000	0.195	0.080	0.170	0.382	0.074	0.311	0.911	0.062	0.788
500	20	0.2	0.001	0.9	0.659	0.060	0.498	0.172	0.072	0.145	0.320	0.069	0.259	0.161	0.050	0.104
1000	20	0.2	0.001	0.9	0.940	0.064	0.851	0.178	0.068	0.140	0.311	0.076	0.238	0.400	0.048	0.238
2000	20	0.2	0.001	0.9	1.000	0.076	0.993	0.159	0.072	0.125	0.317	0.070	0.236	0.764	0.059	0.515
500	10	0.7	0.001	0.9	1.000	0.039	1.000	0.997	0.078	0.988	1.000	0.064	1.000	0.999	0.044	0.998
1000	10	0.7	0.001	0.9	1.000	0.046	1.000	0.994	0.069	0.980	1.000	0.065	1.000	1.000	0.046	1.000
2000	10	0.7	0.001	0.9	1.000	0.055	1.000	0.992	0.066	0.987	1.000	0.065	1.000	1.000	0.067	1.000
500	20	0.7	0.001	0.9	1.000	0.050	1.000	0.994	0.080	0.967	1.000	0.066	1.000	0.999	0.043	0.996
1000	20	0.7	0.001	0.9	1.000	0.055	1.000	0.988	0.076	0.941	1.000	0.077	0.999	1.000	0.043	1.000
2000	20	0.7	0.001	0.9	1.000	0.066	1.000	0.986	0.069	0.960	1.000	0.057	1.000	1.000	0.056	1.000

Table 4: Size and power of the Hong and DCC-Hong tests under the DGP of Case 3 - first part. For DCC-Hong, we adopt simulated critical values with the innovation correlation set to 0.5.

Electronic copy available at: https://ssrn.com/abstract=3941778

					Hong : Full			Hong : 100			Hong : 200			DCC-Hong		
T	M	$\phi_{1,2}$	σ	ρ	$2 \rightarrow 1$	$1 \rightarrow 2$	$1 \leftrightarrow 2$	$2 \rightarrow 1$	$1 \rightarrow 2$	$1 \leftrightarrow 2$	$2 \rightarrow 1$	$1 \rightarrow 2$	$1 \leftrightarrow 2$	$2 \rightarrow 1$	$1 \rightarrow 2$	$1 \leftrightarrow 2$
500	10	0	0.0001	0.99	0.067	0.062	0.061	0.075	0.081	0.085	0.098	0.074	0.089	0.060	0.055	0.055
1000	10	0	0.0001	0.99	0.086	0.089	0.092	0.080	0.073	0.076	0.086	0.082	0.098	0.064	0.056	0.063
2000	10	0	0.0001	0.99	0.129	0.128	0.154	0.086	0.091	0.090	0.097	0.086	0.111	0.062	0.064	0.072
500	20	0	0.0001	0.99	0.071	0.065	0.068	0.083	0.080	0.077	0.084	0.085	0.099	0.063	0.047	0.060
1000	20	0	0.0001	0.99	0.076	0.094	0.082	0.075	0.072	0.076	0.088	0.082	0.098	0.050	0.062	0.061
2000	20	0	0.0001	0.99	0.117	0.127	0.138	0.082	0.077	0.085	0.090	0.077	0.108	0.058	0.062	0.053
500	10	0.2	0.0001	0.99	0.721	0.057	0.638	0.244	0.088	0.218	0.411	0.076	0.333	0.313	0.053	0.193
1000	10	0.2	0.0001	0.99	0.925	0.071	0.873	0.244	0.069	0.197	0.414	0.085	0.363	0.635	0.068	0.465
2000	10	0.2	0.0001	0.99	0.998	0.097	0.990	0.245	0.089	0.208	0.399	0.088	0.354	0.867	0.076	0.757
500	20	0.2	0.0001	0.99	0.672	0.057	0.523	0.205	0.080	0.181	0.352	0.081	0.279	0.186	0.045	0.116
1000	20	0.2	0.0001	0.99	0.888	0.082	0.820	0.217	0.070	0.168	0.353	0.079	0.290	0.448	0.064	0.287
2000	20	0.2	0.0001	0.99	0.993	0.111	0.978	0.197	0.082	0.174	0.353	0.086	0.300	0.780	0.066	0.505
500	10	0.7	0.0001	0.99	1.000	0.045	1.000	0.993	0.085	0.984	1.000	0.073	1.000	1.000	0.049	1.000
1000	10	0.7	0.0001	0.99	1.000	0.053	1.000	0.986	0.085	0.973	1.000	0.084	1.000	1.000	0.062	1.000
2000	10	0.7	0.0001	0.99	1.000	0.077	1.000	0.990	0.086	0.976	1.000	0.093	1.000	1.000	0.069	1.000
500	20	0.7	0.0001	0.99	1.000	0.047	1.000	0.989	0.091	0.956	1.000	0.072	1.000	1.000	0.041	0.996
1000	20	0.7	0.0001	0.99	1.000	0.074	1.000	0.974	0.079	0.936	1.000	0.090	0.998	1.000	0.058	1.000
2000	20	0.7	0.0001	0.99	1.000	0.096	1.000	0.982	0.079	0.950	1.000	0.092	1.000	1.000	0.062	1.000

Table 5: Size and power of Hong and DCC-Hong test under the DGP of Case 3 - second part. For DCC-Hong we adopt simulated critical values with innovation correlation set to 0.5.

Electronic copy available at: https://ssrn.com/abstract=3941778

Returns	WTI (I)	Brent (II)	WTI (III)	Brent (IV)
Mean	0.063	0.068	0.024	0.024
Maximum	16.410	12.707	31.963	19.077
Minimum	-13.065	-10.946	-34.542	-27.976
Standard deviation	2.427	2.196	2.637	2.262
Skewness	-0.022	-0.118	-0.478	-0.579
Kurtosis	7.171	5.754	28.054	16.026
JB	1929.676	847.656	134367.663	36554.987
	[0.0010]	[0.0010]	[0.0010]	[0.0010]
ADF	-53.254	-55.110	-72.636	-73.606
	[0.0010]	[0.0010]	[0.0010]	[0.0010]
Q(5)	21.571	24.511	45.805	12.122
	[0.0006]	[0.0002]	[0.0000]	[0.0332]
Q(10)	27.741	35.086	48.889	18.985
	[0.0020]	[0.0001]	[0.0000]	[0.0405]
Observations	2662	2662	5130	5130
Start	03-Jan-02	03-Jan-02	03-Jan-02	03-Jan-02
End	19-Mar-12	19-Mar-12	03-Sep-21	03-Sep-21

Table 6: Descriptive statistics for returns of WTI and Brent.

Notes: Columns I and II includes the descriptive statistics from from 3 January 2002 to 19 March 2012, the same period considered in Lu et al. (2014). Columns III and IV includes the full-sample from 3 January 2002 to 2 September 2021. JB refers to the Jarque-Bera normality test, ADF to the Dickey-Fuller unit root test, while Q(5) and Q(10) are the Box–Pierce statistics for 5th and 10th order serial correlations; values in [] are t-values. For the ADF test we use the standard specification without drift and trend components.

Electronic copy available at: https://ssrn.com/abstract=3941778

For the evaluation of causality, we follow Lu et al. (2014) in setting the lag truncation value, $M = 10$, in the use of the Bartlett kernel and in fixing the rolling subsample size for the rolling Hong tests, set to $S = 100$. In contrast with Lu et al. (2014), we do not include the contemporaneous causality. Lu et al. (2014) support such a choice as a consequence of the asynchronicity in the trading; we believe that such a choice is inappropriate given that the two future prices are recovered from exchanges based in the US and, therefore, can be safely treated as synchronous.[6]

With regard to the period considered in Lu et al. (2014), Figure 3 presents the unidirectional rolling Hong test from WTI to Brent (Brent ← WTI, first panel), the unidirectional rolling Hong test from Brent to WTI (Brent → WTI, second panel), and the bidirectional rolling Hong test (Brent ↔ WTI, third panel). The 1% normal quantile critical value for the rejection of the null hypothesis (no-tail causality) is indicated by the dashed red line. The Brent-WTI spread is reported at the bottom of the figure.

The unidirectional Brent ← WTI reveals that there are three statistically significant causality episodes from WTI to Brent. The first is a short-lived episode in mid-2006 when crude oil experienced an all-time record. The second episode occurred after May 2010 when crude oil prices dropped due to concerns about the economic growth in the European Union given the sovereign debt crisis in the peripheral countries. The last episode involves oil production cuts in February 2011 during the Arab Spring in Libya, Egypt, Yemen, Syria, and Bahrain.

The unidirectional Brent → WTI exhibits two statistically significant causality episodes from Brent to WTI. The first episode occurred during the invasion of Iraq in 2003. At that time, the country owned one of the largest oil reserves. The second episode refers to the cuts in oil production cuts during the Arab Spring as previously discussed.

The bidirectional Brent ↔ WTI indicates the three statistically significant causality episodes discussed for the unidirectional cases. The first is the invasion of Iraq in 2003. The second episode concerns the European sovereign debt crisis in May 2010. The third episode refers to the Arab spring in February 2011. As shown in Figure 1 in Lu et al. (2014), their values for the test statistic are considerably higher due to the contemporaneous correlation included in the rolling Hong tests. In this case, the role played by comovements is predominantly and is further clarified below for the case of the DCC-MGARCH.

Figure 4 presents the results for the DCC-MGARCH Hong tests. In addition, the dashed black line indicates the 1% simulated critical values according to the study performed in Section 2.1. It is worth noting that the simulated critical value is considerably higher than the one expressed by the normal quantile—that is, the black dashed line is well above the red dashed line.

Table 7 presents the rejection rate of the null hypothesis of no-tail causality for the DCC-MGARCH Hong test according to the Normal critical value (first column) and the simulated critical value (second column). As is evident, the rejection rate of the null hypothesis with the 1% normal quantile critical value is considerably higher than the one of the simulated critical values.

For the case of Brent ← WTI, the rejection rate is 100% for the normal quantile critical

[6] There is only a one-hour delay between the times in Chicago and New York.

Electronic copy available at: https://ssrn.com/abstract=3941778

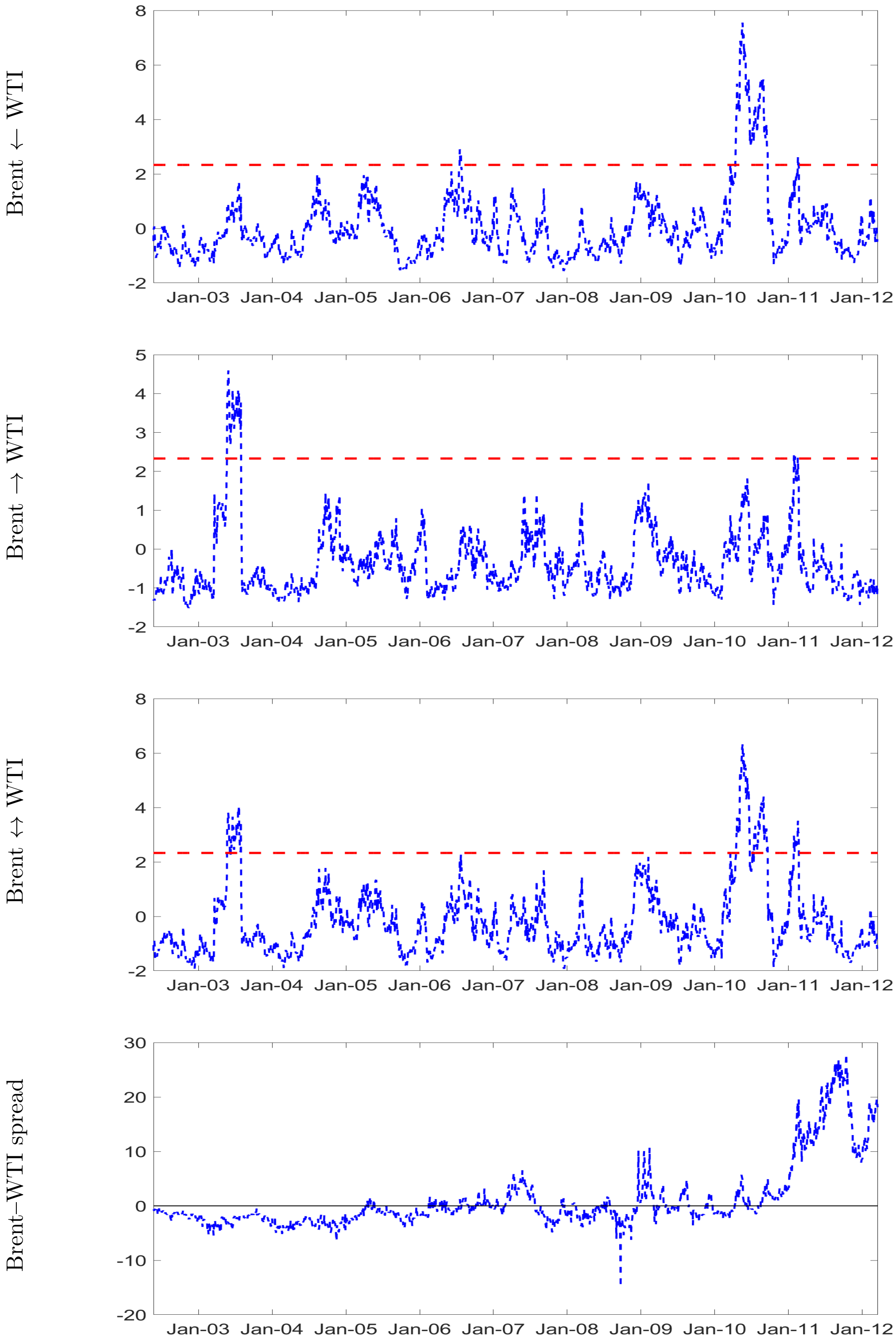

Figure 3: Rolling Hong tests with between Brent and WTI as in Lu et al. (2014).

Notes: The unidirectional rolling Hong test from WTI to Brent (first), the unidirectional rolling Hong test from Brent to WTI (second), and the bidirectional rolling Hong test. The dashed red line indicates the 1% normal quantile critical value. The forth panel includes the Brent-WTI spread. The period under consideration is from 3 January 2002 to 19 March 2012 as in Lu et al. (2014). The rolling sample size is equal to 100.

Electronic copy available at: https://ssrn.com/abstract=3941778

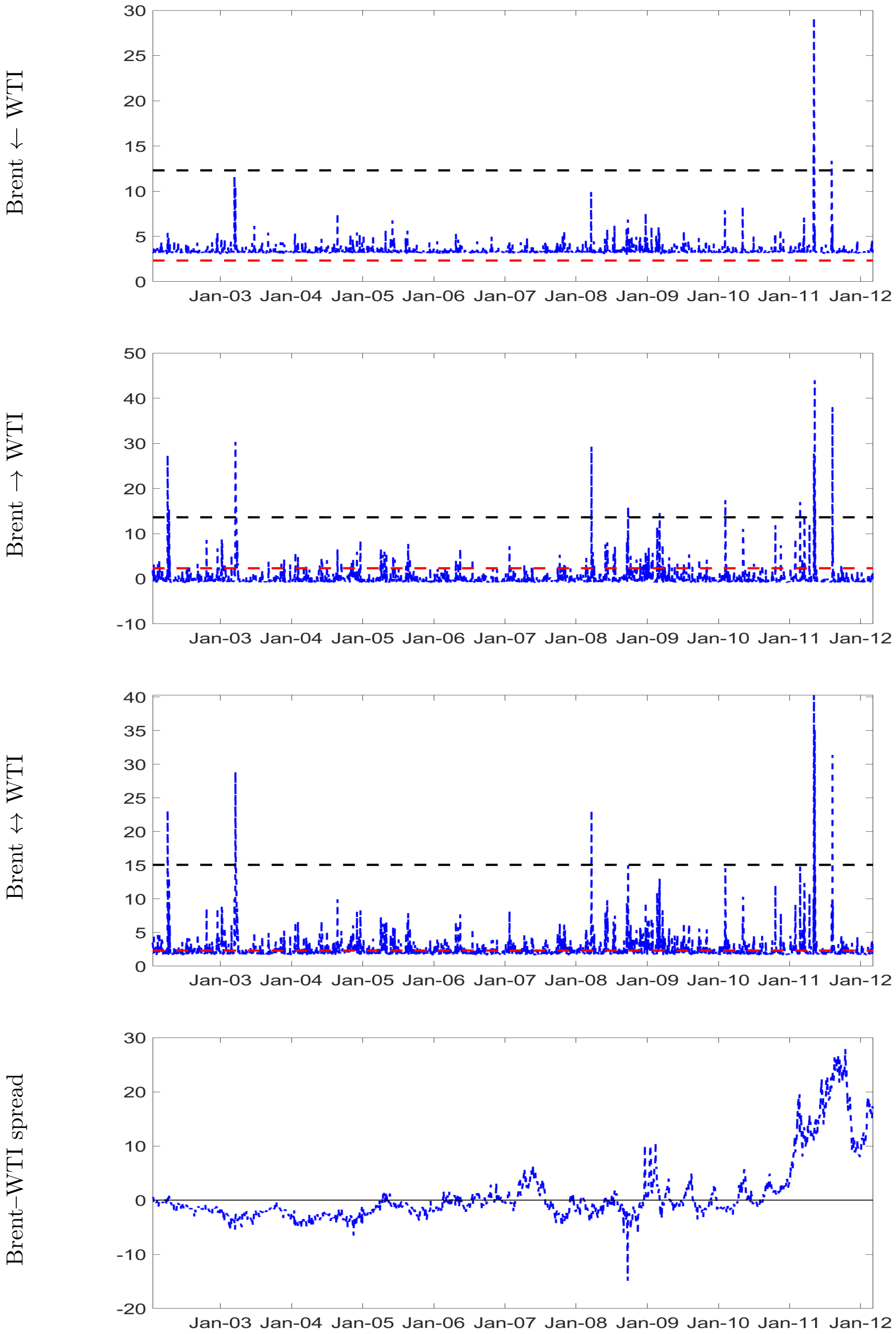

Figure 4: DCC-MGARCH Hong tests between Brent and WTI as in Lu et al. (2014).

Notes: The unidirectional DCC-MGARCH Hong test from WTI to Brent (first), the unidirectional DCC-MGARCH Hong test from Brent to WTI (second), and the bidirectional DCC-MGARCH Hong test. The dashed black (red) line indicates the 1% simulated (normal quantile) critical value. The forth panel includes the Brent-WTI spread. The period under consideration is from 3 January 2002 to 19 March 2012 as in Lu et al. (2014).

Electronic copy available at: https://ssrn.com/abstract=3941778

$\alpha = 1\%$	Normal	Simulated
Brent ← WTI	100.00%	0.11%
Brent → WTI	6.09%	0.57%
Brent ↔ WTI	33.72%	0.30%

Table 7: The rejection rate of the null hypothesis of no-tail causality for the DCC-MGARCH Hong tests according to the 1% normal critical value (first column) and the 1% simulated critical value (second column).

Notes: The unidirectional DCC-MGARCH Hong test from WTI to Brent (first row), the unidirectional DCC-MGARCH Hong test from Brent to WTI (second row), and the bidirectional DCC-MGARCH Hong test (third row). The period under consideration is from 3 January 2002 to 19 March 2012, as in Lu et al. (2014).

value, while it is 0.11% for the simulated quantile. This clearly leads to contrasting conclusions regarding the rejection of the null hypothesis and represents the key point of our exercise. The difference between the rejection rates of the two critical values is also visible for Brent → WTI and Brent ↔ WTI. The two cases show a rejection rate of 6.09% (0.57%) and 33.72% (0.30%) using the normal (simulated) quantile critical value, respectively. The simulated quantile critical value provides a number of statistically significant causality episodes and this is in line with the one for the rolling Hong test, even if the timing is not fully aligned. Clearly, the different timing also depends on the selected rolling subsample size.[7] For example, there are two further significant causality episodes for Brent → WTI and Brent ↔ WTI that are not detected in the rolling Hong test. The first concerns the oil embargo issued by Iraq in April 2002 and involved exports of approximately 1.5 million barrels of oil a day (1 million only to the US). The second involves a dispute between OPEC and the George W. Bush administration in March 2008 regarding the causes of the increase in oil prices.

Even in this case, the values for the test are considerably lower than the ones reported in Lu et al. (2014). Figure 5 presents the estimated dynamic conditional correlations, which range from -M to M with $M = 10$. The dashed red line indicates the contemporaneous correlation, the $j = 0$ case, included in the Hong test by Lu et al. (2014).

It is evident that the discrepancy in the results is due to the contemporaneous correlation included in the Hong test by Lu et al. (2014). The dynamic of the bidirectional test of Lu et al. (2014) (solid green line in Fig. 5 of the paper) is almost identical to the contemporaneous correlation depicted in our figure. As discussed above, we do believe that it does not represents an appropriate choice, since as contemporaneous correlations measure the comovements of two crude oil commodities that are synchronously traded in the US market. This is also confirmed by the difference in magnitude of the correlations. The dynamic contemporaneous correlation is, on average, 0.86 while the lead/lag correlations are, on average, close to zero.

Finally, Figures 6 and 7 include the results for the entire sample according to the rolling Hong and the DCC-MGARCH Hong tests, respectively. The dynamic of both Hong tests

[7] In Appendix Appendix A, we show that by selecting $S = 200$ the dynamic of the test changes.

Electronic copy available at: https://ssrn.com/abstract=3941778

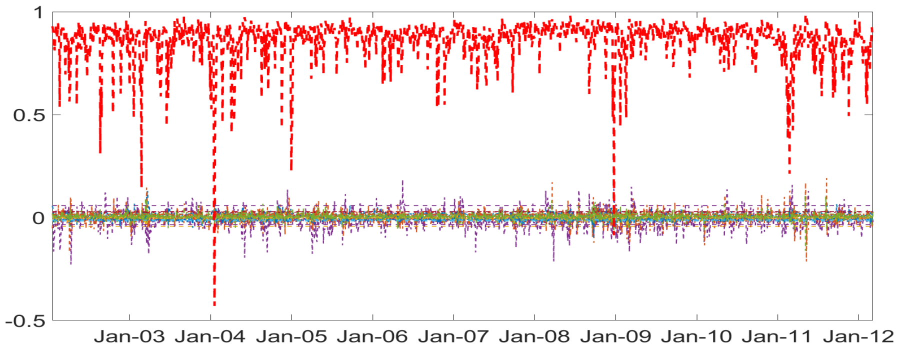

Figure 5: Estimated dynamic conditional correlations using the DCC model given by Engle (2002).

Notes: The DCC model provides $2M + 1$ estimated paths for the cross-correlation, which range from -M to M with $M = 10$. The dashed red line indicates the contemporaneous correlation, which represents the $j = 0$ case included in the Hong test by Lu et al. (2014). The period under consideration is from 3 January 2002 to 19 March 2012.

show a further statistically significant causal episode in June 2020, after the outbreak of COVID-19. Notably, the WTI experienced a negative price for the first time in history.

Table 8 presents the rejection rate of the null hypothesis of no tail causality for the DCC-MGARCH Hong test according to the Normal critical value (first column) and the simulated critical value (second column). Moreover for the entire sample, the rejection rate of the null hypothesis is considerably lower for the simulated critical values.

$\alpha = 1\%$	Normal	Simulated
Brent $\leftarrow$ WTI	4.32%	0.04%
Brent $\rightarrow$ WTI	1.53%	0.06%
Brent $\leftrightarrow$ WTI	2.99%	0.06%

Table 8: The rejection rate of the null hypothesis of no-tail causality for the DCC-MGARCH Hong tests according to the 1% Normal critical value (first column) and the 1% simulated critical value (second column).

Notes: The unidirectional DCC-MGARCH Hong test from WTI to Brent (first row), the unidirectional DCC-MGARCH Hong test from Brent to WTI (second row), and the bidirectional DCC-MGARCH Hong test (third row). The period under consideration is from 3 January 2002 to 2 September 2021.

It is worth noting that the DCC-MGARCH Hong tests provide different results for the subsample period from 3 January 2002 to 19 March 2012. As depicted in Figures 4 and 7, the value of the test is, on average, lower in the entire sample estimates with respect to the value obtained in the subsample estimates. In the former, the only significant causality episode is detected during the outbreak of COVID-19.

Smaller values of the statistic are yielded by lower dynamic conditional correlations obtained in the entire sample. If the causality represents short-lived periods, the identification of the significant episodes based on cross-correlations will fade away with the increase in T, since the estimated parameters of the DCC-GARCH model will be driven by periods of weak

Electronic copy available at: https://ssrn.com/abstract=3941778

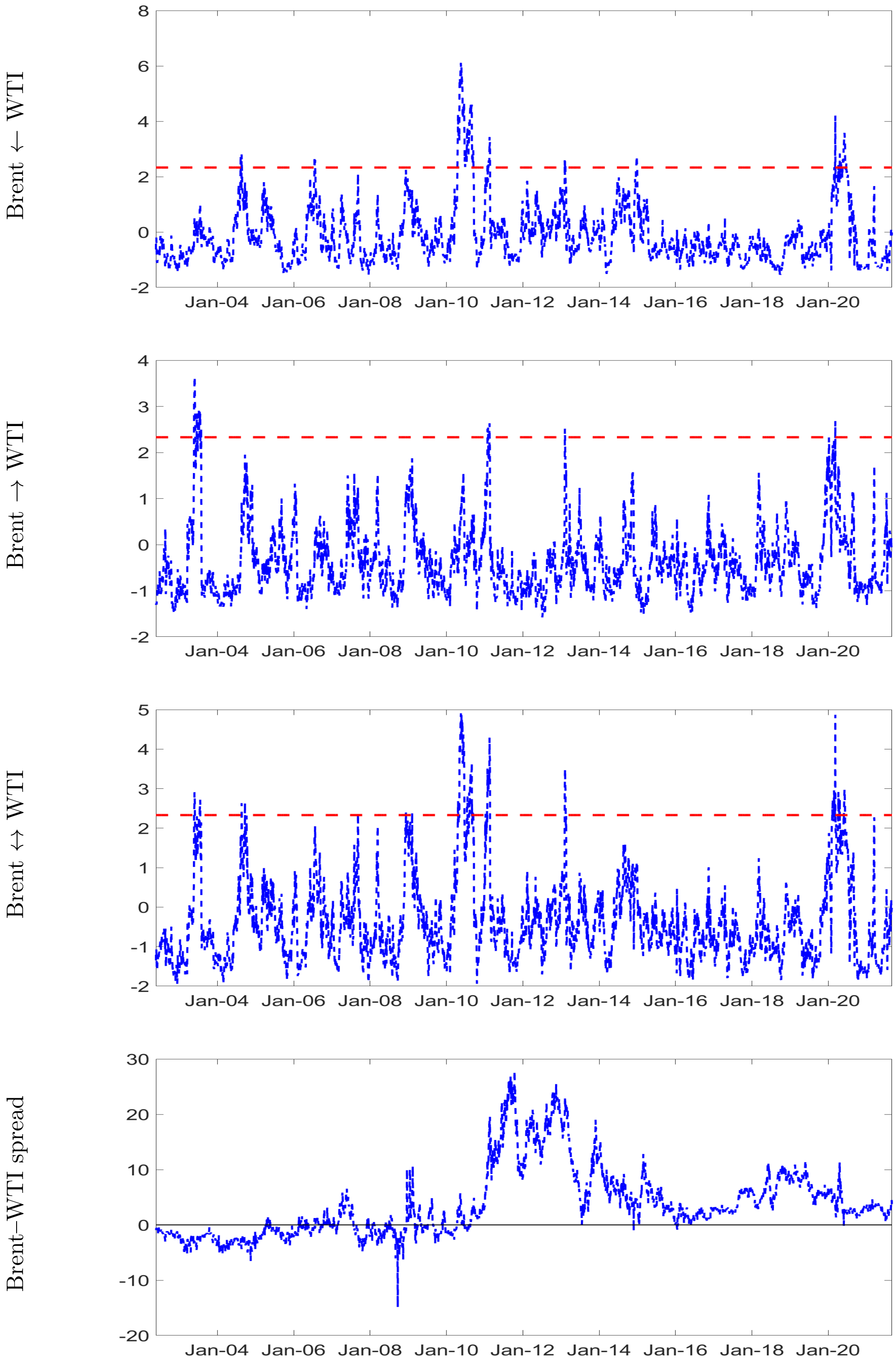

Figure 6: Rolling Hong tests between Brent and WTI for the entire sample.

Notes: The unidirectional rolling Hong test from WTI to Brent (first), the unidirectional rolling Hong test from Brent to WTI (second), and the bidirectional rolling Hong test. The dashed red line indicates the 1% normal quantile critical value. The forth panel includes the Brent-WTI spread. The period under consideration is from 3 January 2002 to 2 September 2021. The rolling sample size is equal to 100.

Electronic copy available at: https://ssrn.com/abstract=3941778

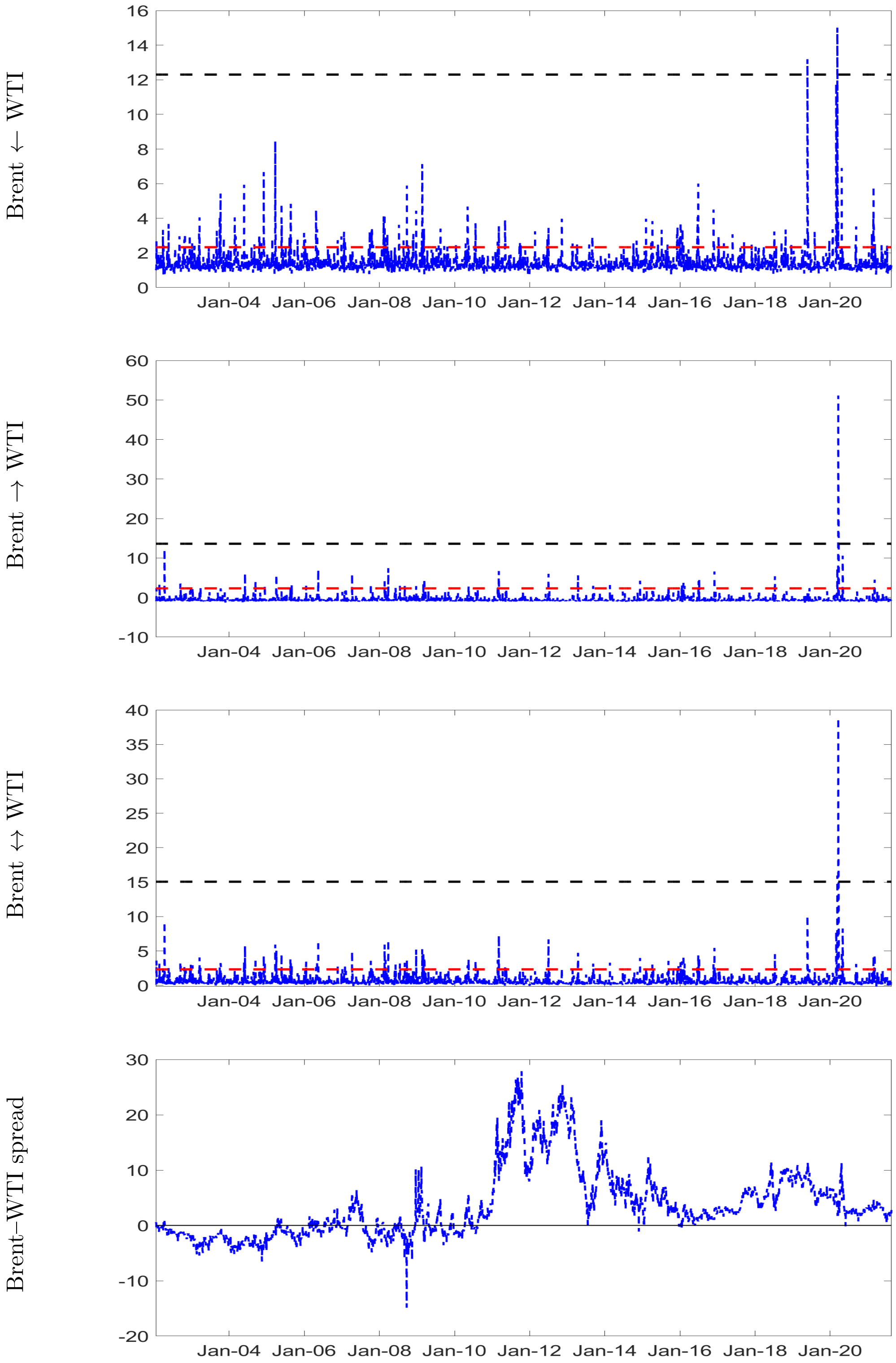

Figure 7: DCC-MGARCH Hong tests between Brent and WTI for the entire sample.

Notes: The unidirectional DCC-MGARCH Hong test from WTI to Brent (first), the unidirectional DCC-MGARCH Hong test from Brent to WTI (second), and the bidirectional DCC-MGARCH Hong test. The dashed black (red) line indicates the 1% simulated (normal quantile) critical value. The forth panel includes the Brent-WTI spread. The period under consideration is from 3 January 2002 to 2 September 2021.

Electronic copy available at: https://ssrn.com/abstract=3941778

dependence.

To preserve the identification of these transitory episodes, a rolling evaluation scheme of the DCC-Hong test should be adopted. Appendix B presents the rolling DCC-Hong tests for the entire sample and shows that short-lived episodes are also detected. However, a rolling DCC-Hong test would be more computationally intensive and a simple rolling Hong test would represent a more viable solution.

4. Conclusion

We show that the test statistic suggested by Lu et al. (2014) to detect causality has a non-standard distribution whose critical values must be recovered by simulations. Moreover, utilizing a Monte Carlo study, we show that a rolling application of the test proposed by Hong (2001) appears to be more appropriate with longer time series. Using simulated critical values, we replicate some of the evidence in Lu et al. (2014), thereby revealing striking differences in the detection of causality. Our results challenge the evidence reported in other studies that adopted the approach put forward by Lu et al. (2014).

Acknowledgments: the first author acknowledges financial support from Italian Ministry of University and Research project PRIN2017 HiDEA: Advanced Econometrics for High Frequency Data, 2017RSMPZZ.

References

Bathia, D., Demirer, R., Gupta, R., and Kotze, K. (2021). Unemployment fluctuations and currency returns in the United Kingdom: Evidence from over one and a half century of data. *Journal of Multinational Financial Management*, page 100679.

Bekiros, S. D. and Diks, C. G. (2008). The relationship between crude oil spot and futures prices: Cointegration, linear and nonlinear causality. *Energy Economics*, 30(5):2673–2685.

Caporin, M., Fontini, F., and Talebbeydokhti, E. (2019). Testing persistence of WTI and Brent long-run relationship after the shale oil supply shock. *Energy Economics*, 79:21–31.

Caporin, M. and McAleer, M. (2012). Do we really need both BEKK and DCC? A tale of two multivariate GARCH models. *Journal of Economic Surveys*, 26(4):736–751.

Cheung, Y.-W. and Ng, L. K. (1996). A causality-in-variance test and its application to financial market prices. *Journal of Econometrics*, 72(1-2):33–48.

Coronado, S., Gupta, R., Hkiri, B., and Rojas, O. (2020). Time-Varying Spillovers between Currency and Stock Markets in the USA: Historical Evidence From More than Two Centuries. *Advances in Decision Sciences*, 24(4):1–32.

Engle, R. (2002). Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *Journal of Business & Economic Statistics*, 20(3):339–350.

Electronic copy available at: https://ssrn.com/abstract=3941778

Engle, R. and Sheppard, K. (2001). Theoretical and empirical properites of dynamic conditional correlation multivariate GARCH. *NBER Working Paper Series*, 8554.

Geng, J.-B., Ji, Q., and Fan, Y. (2017). The relationship between regional natural gas markets and crude oil markets from a multi-scale nonlinear Granger causality perspective. *Energy Economics*, 67:98–110.

Granger, C. (1969). Investigating causal relationships by econometrics and cross-spectral models. *Econometrica*, 37:424–438.

Gupta, R., Kanda, P., Tiwari, A., and Wohar, M. (2019). Time-varying predictability of oil market movements over a century of data: The role of us financial stress. *North American Journal of Economics and Finance*, 50:100994.

Gupta, R., Kanda, P., and Wohar, M. (2021a). Predicting Stock Market Movements in the United States: The Role of Presidential Approval Ratings. *International Review of Finance*, 21-1:324–335.

Gupta, R., Subramanian, S., Bouri, E., and Ji, Q. (2021b). Infectious disease-related uncertainty and the safe-haven characteristic of US treasury securities. *International Review of Economics and Finance*, 71:289–298.

Hammoudeh, S. and Li, H. (2004). The impact of the Asian crisis on the behavior of US and international pretroleum prices. *Energy Economics*, 26:135–160.

Hong, Y. (2001). A test for volatility spillover with application to exchange rates. *Journal of Econometrics*, 103(1-2):183–224.

Hong, Y., Liu, Y., and Wang, S. (2009). Granger causality in risk and detection of extreme risk spillover between financial markets. *Journal of Econometrics*, 150(2):271–287.

Jammazi, R., Ferrer, R., Jareno, F., and Hammoudeh, S. (2017a). Main driving factors of the interest rate-stock market Granger causality. *International Review of Economics and Finance*, 52:260–280.

Jammazi, R., Ferrer, R., Jareno, F., and Shahzad, S. (2017b). Time-varying causality between crude oil and stock markets: What can we learn from a multiscale perspective? *International Review of Economics and Finance*, 49:453–483.

Kanda, P., Burke, M., and Gupta, R. (2018). Time-varying causality between equity and currency returns in the United Kingdom: Evidence from over two centuries of data. *Physica A*, 506:1060–1080.

Lin, S. and Tamvakis, M. (2001). Spillover effects in energy future markets. *Energy Economics*, 23:43–56.

Lin, S. and Tamvakis, M. (2004). Effects of NYMEX trading on ipe brent crude futures markets: a duration analysis. *Energy Policy*, 32:77–82.

Electronic copy available at: https://ssrn.com/abstract=3941778

Lu, F.-b., Hong, Y.-m., Wang, S.-y., Lai, K.-k., and Liu, J. (2014). Time-varying Granger causality tests for applications in global crude oil markets. *Energy Economics*, 42:289–298.

Sibande, X., Gupta, R., and Wohar, M. (2019). Time-varying causal relationship between stock market and unemployment in the United Kingdom: Historical evidence from 1855 to 2017. *Journal of Multinational Financial Management*, 49:81–88.

Zhang, X., Lu, F., Tao, R., and Wang, S. (2021). The time-varying causal relationship between the Bitcoin market and internet attention. *Financial Innovation*, 7:1–19.

Electronic copy available at: https://ssrn.com/abstract=3941778

Appendix A. Rolling Hong test with a subsample size equal to 200.

In this section, we perform the rolling Hong tests with the subsample size $S = 200$. Figures A.8 and A.9 show the unidirectional rolling Hong tests and the bidirectional rolling Hong test for the period considered in Lu et al. (2014) and the entire sample The period under consideration is from 3 January 2002 to 2 September 2021.

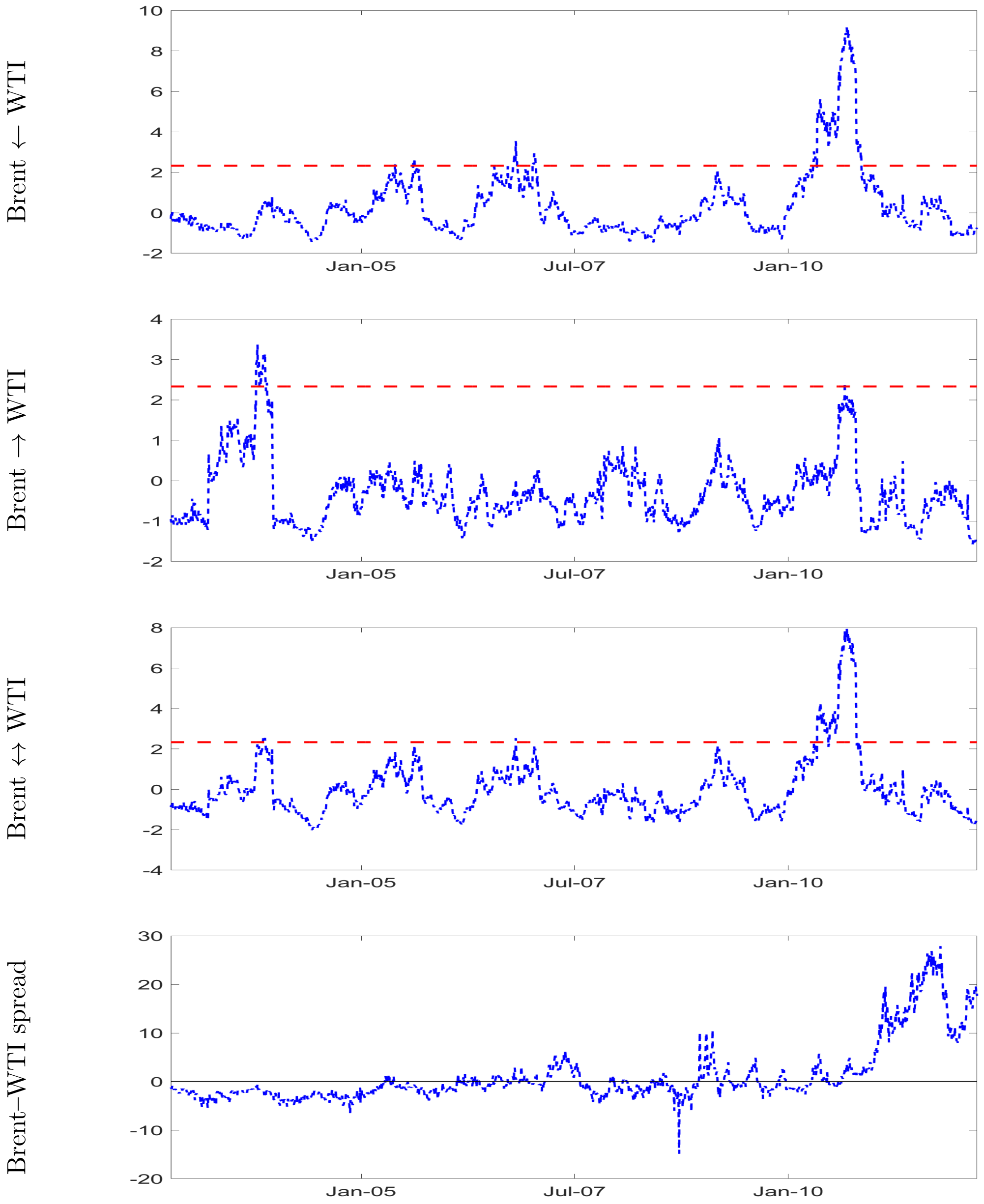

Figure A.8: Rolling Hong tests with between Brent and WTI as in Lu et al. (2014).

Notes: The unidirectional rolling Hong test from WTI to Brent (first), the unidirectional rolling Hong test from Brent to WTI (second), and the bidirectional rolling Hong test. The dashed red line indicates the 1% normal quantile critical value. The forth panel includes the Brent-WTI spread. The period under consideration is from 3 January 2002 to 19 March 2012 as in Lu et al. (2014). The rolling sample size is equal to 200.

Electronic copy available at: https://ssrn.com/abstract=3941778

Appendix B. Rolling DCC-MGARCH Hong test with a subsample size equal to 100.

In this section, we perform the rolling DCC-MGARCH Hong tests with the subsample size $S = 100$. Figures B.10 show the unidirectional rolling DCC-MGARCH Hong tests and the bidirectional rolling DCC-MGARCH Hong test from 3 January 3 2002 to 2 September 2021.

Electronic copy available at: https://ssrn.com/abstract=3941778

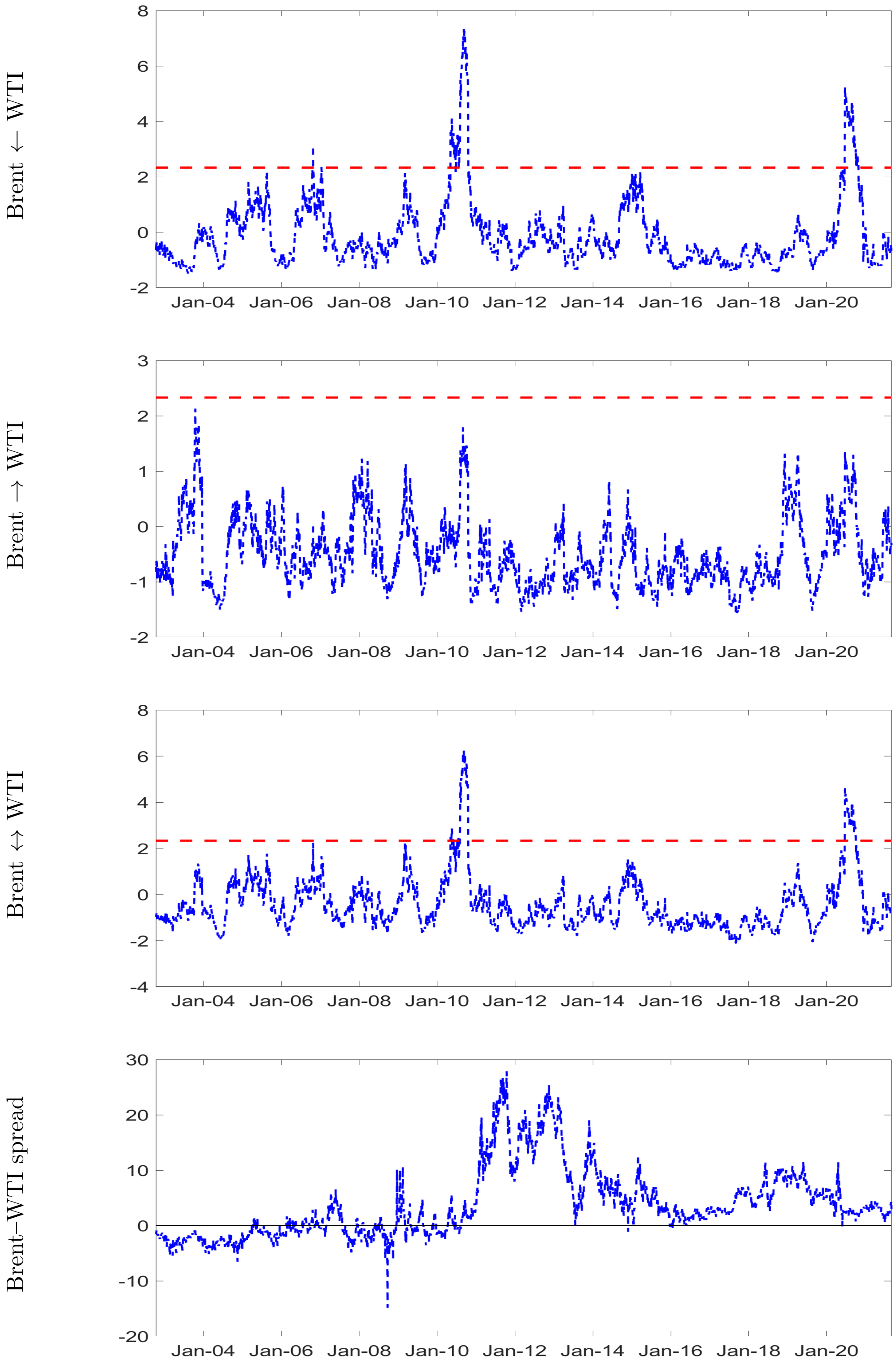

Figure A.9: Rolling Hong tests between Brent and WTI for the entire sample.

Notes: The unidirectional rolling Hong test from WTI to Brent (first), the unidirectional rolling Hong test from Brent to WTI (second), and the bidirectional rolling Hong test. The dashed red line indicates the 1% normal quantile critical value. The forth panel includes the Brent-WTI spread. The period under consideration is from 3 January 2002 to 2 September 2021. The rolling sample size is equal to 200.

Electronic copy available at: https://ssrn.com/abstract=3941778

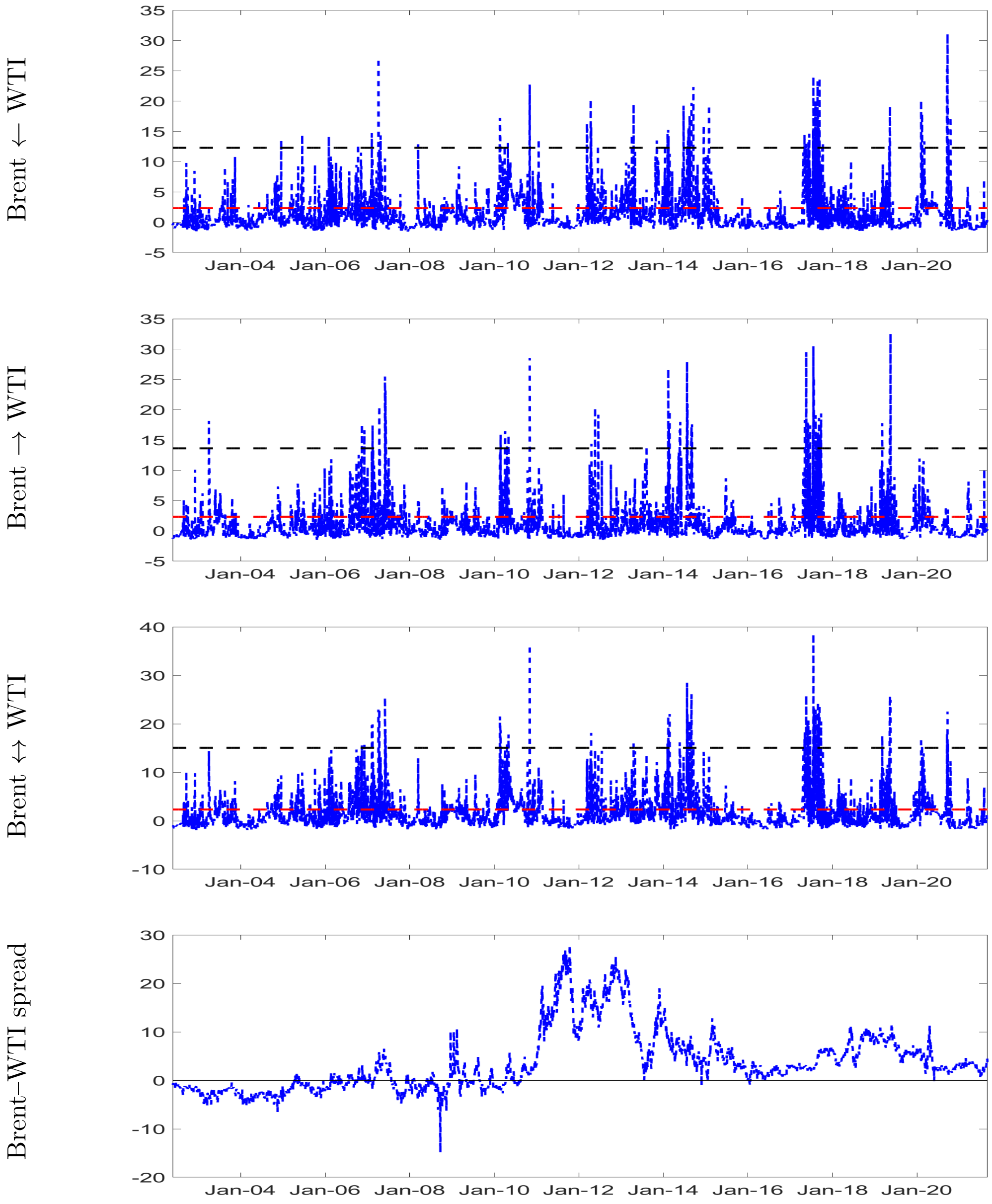

Figure B.10: Rolling DCC-Hong tests between Brent and WTI for the entire sample.

Notes: The unidirectional rolling DCC-MGARCH Hong test from WTI to Brent (first), the unidirectional rolling Hong test from Brent to WTI (second), and the bidirectional rolling Hong test. The dashed red line indicates the 1% normal quantile critical value. The forth panel includes the Brent-WTI spread. The period under consideration is from 3 January 2002 to 2 September 2021. The rolling sample size is equal to 100.

Electronic copy available at: https://ssrn.com/abstract=3941778

Recent Issues

No. 323	Alexandre Corhay, Thilo Kind, Howard Kung, Gonzalo Morales	Discount Rates, Debt Maturity, and the Fiscal Theory
No. 322	Gianluca Anese, Marco Corazza, Michele Costola, Loriana Pelizzon	Impact of Public News Sentiment on Stock Market Index Return and Volatility
No. 321	Ignazio Angeloni, Johannes Kasinger, Chantawit Tantasith	The Geography of Banks in the United States (1990-2020)
No. 320	Sebastian Steuer, Tobias H. Tröger	The Role of Disclosure in Green Finance
No. 319	Erik Theissen, Christian Westheide	Call of Duty: Designated Market Maker Participation in Call Auctions
No. 318	Kevin Bauer, Michael Kosfeld, Ferdinand von Siemens	Incentives, Self-Selection, and Coordination of Motivated Agents for the Production of Social Goods
No. 317	Volker Flögel, Christian Schlag, Claudia Zunft	Momentum-Managed Equity Factors
No. 316	Christian Mücke, Loriana Pelizzon, Vincenzo Pezone, Anjan Thako	The Carrot and the Stick: Bank Bailouts and the Disciplining Role of Board Appointments
No. 315	Kevin Bauer, Moritz von Zahn, Oliver Hinz	Expl(AI)ned: The Impact of Explainable Artificial Intelligence on Cognitive Processes
No. 314	Farshid Abdi, Mila Getmansky Sherman, Emily Kormanyos, Loriana Pelizzon, Zorka Simon	A Modern Take on Market Efficiency: The Impact of Trump's Tweets on Financial Markets
No. 313	Kevin Bauer, Andrej Gill	Mirror, Mirror on the Wall: Machine Predictions and Self-Fulfilling Prophecies
No. 312	Can Gao Ian Martin	Volatility, Valuation Ratios, and Bubbles: An Empirical Measure of Market Sentiment
No. 311	Wenhui Li, Christian Wilde	Separating the Effects of Beliefs and Attitudes on Pricing under Ambiguity
No. 310	Carmelo Latino, Loriana Pelizzon, Aleksandra Rzeźnik	The Power of ESG Ratings on Stock Markets
No. 309	Tabea Bucher-Koenen, Andreas Hackethal, Johannes Koenen, Christine Laudenbach	Gender Differences in Financial Advice

Electronic copy available at: https://ssrn.com/abstract=3941778