Does Causality Pay? Time-Varying Granger Causality as a Feature-Selection Device for Machine-Learning Forecasts of the Oil Price and the Euro-Dollar Rate

Dr. Mokaddem Tebra

Abstract: *The relationship between oil prices and exchange rates is well documented, yet its usefulness for forecasting remains uncertain. This study asks whether time-varying Granger-causality tests can improve machine-learning forecasts by deciding when one series should enter forecasts of the other. Using 6,341 daily observations of Brent crude oil and EUR/USD (4 January 1999 - 27 February 2026), rolling Toda-Yamamoto tests guide predictor selection in Ridge, Elastic Net, random-forest and gradient-boosting models. Forecasts are evaluated out of sample (from January 2012; horizons of 1, 5 and 22 trading days) under three strategies: unguided (both series), own-only (the target's history) and guided (cross-series predictors admitted only when non-causality is rejected). No model significantly outperforms the random walk. For EUR/USD, guided models reduce mean squared error by 1.48% relative to unguided ones, and two of twelve guided-versus-unguided comparisons survive Holm adjustment. However, own-only models achieve a slightly larger reduction (1.58%), guided models do not improve on them (average gain -0.11%), and randomly timed screens perform comparably (27.6% of them do at least as well). The gains therefore come from excluding predictors, not from the information in the test. Cross-series predictability is intermittent, and differences across episodes (COVID-19, the 2014-2016 oil collapse, the Russia-Ukraine war) are not statistically established. Causality screening is not a reliable basis for feature selection here.*

Keywords: Granger causality; Toda-Yamamoto; feature selection; machine learning; exchange-rate forecasting; Brent crude; random walk

JEL classification: C22; C53; F31; Q43

Dr. Mokaddem Tebra (m.tebra@univ-mascara.dz) Department of Economic Sciences, Mustapha Stambouli University of Mascara, Algeria
Laboratory of Sustainable Development Engineering, University Responsibility and Social Integration Oran 2

Introduction

The relationship between oil prices and exchange rates has attracted sustained attention in international economics and finance. Oil is traded internationally in US dollars, and changes in its price affect the external balances, income flows and portfolio positions of oil-importing and oil-exporting economies. These channels give a theoretical reason to expect oil prices and exchange rates to be related (Golub, 1983; Krugman, 1983). They do not imply that the relationship is stable over time or that it can be turned into reliable forecasts. Early empirical work found a link between the real oil price and the real effective exchange rate of the US dollar, with evidence of predictive influence running from oil to the exchange rate (Amano and van Norden, 1998). Later studies give a less uniform picture. Beckmann and Czudaj (2013) document differences in the direction and strength of the relationship across currencies and periods, and Beckmann, Czudaj and Arora (2020) emphasise the distinction between long-run connections and short-run dynamics. The empirical question is therefore not simply whether oil prices and exchange rates are related, but whether their interaction contains information that remains useful when forecasts are made outside the estimation sample.

The distinction matters particularly in exchange-rate forecasting. Since Meese and Rogoff (1983), the random walk has remained a demanding benchmark. Later research has identified circumstances in which economic and financial variables contain predictive information, but the evidence has often proved sensitive to the sample, the forecasting horizon and the evaluation procedure (Rossi, 2013). Related findings in commodity markets suggest that exchange rates can help predict commodity prices under certain conditions (Chen, Rogoff and Rossi, 2010), and forecast combinations have shown potential for improving oil-price forecasts at monthly horizons (Baumeister and Kilian, 2015). These results do not establish that the same relationships hold for daily Brent and EUR/USD forecasts. At this frequency the signal in returns is weak, and a model with many predictors may capture sample-specific variation rather than persistent predictive information.

Machine-learning methods offer a useful setting in which to examine the problem. Ridge regression, Elastic Net, random forests and gradient boosting can accommodate different forms of dependence and interaction among predictors (Gu, Kelly and Xiu, 2020). Their forecasting performance nevertheless depends on the information supplied to them, and adding variables does not necessarily improve accuracy, particularly when the variables contribute little beyond the target series' own history. Feature selection is therefore an important part of the forecasting design (Guyon and Elisseeff, 2003). One possible approach is to use Granger-causality tests as a preliminary screen: if the past values of one series improve the prediction of another, the corresponding predictors may be considered for inclusion. The Toda–Yamamoto procedure tests this predictive relationship in a levels VAR without the usual pre-testing for cointegration (Toda and Yamamoto, 1995). Throughout the paper a rejection of non-causality is read as Granger-causal predictability, not as evidence of economic or structural causation.

Using a causality test for this purpose raises a more demanding question than whether the test finds predictive relationships within a sample. A screen is useful to a forecaster only if the information guiding its decisions improves forecasts beyond what a simpler selection rule

would achieve. A guided model may outperform a model that always includes both series simply because it uses fewer predictors whenever the cross-series relationship is not detected. Such an improvement does not necessarily mean that the test has identified particularly informative periods; it may merely reflect the removal of variables that add estimation noise. Without an appropriate benchmark the two explanations cannot be told apart.

This study addresses the distinction through a three-way comparison. The first specification is unguided and includes predictors derived from both Brent and EUR/USD. The second is own-only, restricted to the history of the variable being forecast. The third is guided by rolling Toda–Yamamoto tests and admits predictors from the other series only when the null hypothesis of non-causality is rejected. The comparison between guided and unguided models measures the overall effect of applying the screen. The comparison between guided and own-only models is more informative about the incremental contribution of the test itself: if the guided specification does not improve on the own-only alternative, its advantage over the unguided model cannot be attributed convincingly to the timing information provided by the test.

The empirical analysis uses daily Brent crude oil prices and the EUR/USD exchange rate from 4 January 1999 to 27 February 2026. The forecasting exercise begins in January 2012 and considers horizons of 1, 5 and 22 trading days. Four machine-learning learners are evaluated under the three predictor strategies, alongside conventional benchmarks and a zero-return random walk. Out-of-sample forecast errors are assessed with Diebold–Mariano and Clark–West tests, with Holm adjustment for multiple comparisons. Placebo screens, alternative thresholds, alternative covariance estimators, a returns-based test, different lag rules and different window lengths examine whether the findings depend on a particular implementation choice. The analysis also considers selected historical episodes, namely the COVID-19 period, the 2014–2016 oil-price collapse and the Russia–Ukraine war. These episode comparisons are descriptive. They are intended to show how the predictive relationship and the usefulness of cross-series predictors vary across periods, not to identify the causal effects of the events themselves.

The study makes three contributions. First, it brings rolling Toda–Yamamoto tests into an out-of-sample machine-learning forecasting design for Brent and EUR/USD, extending the Toda–Yamamoto analysis of annual Algerian macroeconomic series in Mokaddem Faradji (2020) to a high-frequency international setting. Second, it separates the forecasting consequences of excluding cross-series predictors from the incremental value of using a causality test to decide when those predictors are admitted. Third, it examines the stability of the relationship across periods and tests whether the actual screening decisions outperform randomly timed ones. The design therefore evaluates not only whether a causality test detects predictive relationships, but also whether those detections give a practical advantage in forecasting.

The results offer a cautious answer. Guided models improve on their unguided counterparts for EUR/USD, but own-only models achieve a comparable and slightly larger average reduction in forecast error. Guided models do not improve on own-only models, and placebo screens perform similarly. No machine-learning specification is significantly better than the random walk. Reducing the predictor set can therefore be useful in this application, but the rolling

causality test provides no measurable additional forecasting advantage. The remainder of the paper is organised as follows. Section 2 reviews the literature and states the research gap. Section 3 presents the data and methodology. Section 4 reports the empirical results, including robustness checks and the episode analysis. Section 5 discusses the implications, and Section 6 concludes with the limitations of the study.

Literature review and research gap

2.1 Oil prices and the dollar: a link that moves

Early theory tied oil prices to the dollar through the current accounts and portfolios of oil exporters and importers (Golub, 1983; Krugman, 1983). Empirical studies since the late 1990s document co-movement whose direction and strength vary by sample, currency pair and method (Amano and van Norden, 1998; Zhang et al., 2008; Lizardo and Mollick, 2010; Reboredo, 2012). Beckmann and Czudaj (2013) stress that the pattern of causality is far from homogeneous across currencies and regimes, and that nonlinearities matter. The review of Beckmann, Czudaj and Arora (2020) reaches a similar conclusion: strong long-run linkages, time-varying short-run effects, and a weak mapping from in-sample relationships to out-of-sample forecasts.

Two further considerations bear on how a causality test between the two series should be read. First, co-movement need not reflect a direct link. The monetary policy of the Federal Reserve and the European Central Bank, interest-rate differentials and global risk appetite move both the dollar and commodity prices, so a rejection of non-causality can reflect a common driver as well as a transmission channel. Second, the structure of the oil market, with its futures and storage markets and the supply decisions of producers, means that the information reaching the exchange rate through oil may already be reflected in prices when it becomes observable at the daily frequency. A Granger-type test cannot separate these channels, which is one more reason to read its result as predictive rather than structural.

2.2 Predictability and the random-walk benchmark

Meese and Rogoff (1983) is the founding negative result, and Rossi (2013) shows, in her survey, how persistently it has resisted challenge: models that fit in sample tend to fail out of sample, and what predictability exists is sporadic. Two positive results are relevant here. Chen, Rogoff and Rossi (2010) find that the exchange rates of commodity-exporting economies help predict world commodity prices, and Baumeister and Kilian (2015) show that combining forecasts can beat the no-change benchmark for the real price of oil. Alquist, Kilian and Vigfusson (2013) survey the oil-price forecasting toolbox.

2.3 Machine learning, feature selection and causality

Ridge regression (Hoerl and Kennard, 1970), the LASSO and Elastic Net (Tibshirani, 1996; Zou and Hastie, 2005), random forests (Breiman, 2001) and gradient boosting (Friedman, 2001) are standard learners in the return-prediction literature (Gu, Kelly and Xiu, 2020). That evidence concerns monthly US stock returns with hundreds of predictors; whether it carries over to daily exchange-rate and oil returns with eighteen predictors is an open question. Choosing which predictors to feed them is the subject of the feature-selection literature (Guyon

and Elisseeff, 2003), which distinguishes filters, wrappers and embedded methods; causal feature selection, which asks which variables are causes or direct neighbours of the target, is a recognised branch (Guyon, Aliferis and Elisseeff, 2007). A parallel strand uses neural networks to *discover* nonlinear Granger causality (Tank et al., 2022). A Granger test used as a cheap, classical *filter* on the inputs of a forecasting model sits between these strands and, to my knowledge, has not been evaluated against an own-only benchmark in the oil–exchange-rate setting. On evaluation, out-of-sample R^2 against a naïve benchmark (Campbell and Thompson, 2008) and the Diebold–Mariano test (Diebold and Mariano, 1995), with the small-sample correction of Harvey, Leybourne and Newbold (1997), are the accepted tools; Clark and West (2007) provide an adjusted test for nested models.

2.4 Research gap

The literature leaves three questions open: whether Granger-causal predictability between oil and the euro–dollar rate is present often enough, and stably enough, to be useful to a forecaster; whether feeding the other series to a machine-learning model only when a causality test says so improves accuracy; and, if it does, whether the improvement reflects the information in the test or merely the removal of predictors. This paper addresses all three.

Data and methodology

3.1 Data

The data are the daily Brent crude spot price (USD per barrel, series DCOILBRENTEU, from the U.S. Energy Information Administration) and the daily spot rate of the US dollar per euro (series DEXUSEU, from the Board of Governors of the Federal Reserve System), both obtained from FRED. The second series is denoted EUR/USD throughout, in market notation, and a rise means a stronger euro. The two calendars differ because of national holidays and missing values; I keep only dates on which both are observed (Appendix A gives the counts). The sample starts on 4 January 1999, the first day of the euro, and ends on 27 February 2026, the last trading day of that month; **no observation after 27 February 2026 is used anywhere in the paper.** The sample end is fixed by design: observations later than that date, which are present in the source files, were discarded by a date filter in the data-preparation code. This leaves 6,341 common observations. Logarithmic prices are denoted B_t for Brent and E_t for EUR/USD, and daily returns are $100 \times \Delta \ln P_t$.

3.2 Unit roots and Toda–Yamamoto tests

Orders of integration are examined with the augmented Dickey–Fuller test (Dickey and Fuller, 1979; lag chosen by AIC, critical values from MacKinnon, 1996) and the KPSS test (Kwiatkowski et al., 1992). The Toda–Yamamoto procedure avoids pre-testing for cointegration by estimating a levels VAR of order $p + d_{\max}$, with $d_{\max}$ the highest order of integration suspected, and testing only the first p lags:

$$y_t = c + \sum_{i=1}^{p+d_{\max}} A_i\, y_{t-i} + u_t, \qquad y_t = (B_t,\ E_t)' \tag{1}$$

$$H_0 : \; a_{12,1} = \cdots = a_{12,p} = 0, \qquad W = (Rb)' \, [R V R']^{-1} (Rb) \; \sim \; \chi^2(p)$$

(2)

I set $d_{\max} = 1$ and choose p by the Bayesian information criterion (maximum six lags, capped at five in the forecasting design). In equation (2), $a_{12,i}$ is the coefficient on the i-th lag of the second variable in the equation of the first (and symmetrically for the opposite direction), b is the vector of estimated coefficients of the VAR equation, V its estimated covariance matrix and R the matrix that selects the p lags of the other series. Because daily financial series are heteroskedastic, V is by default the HC1 heteroskedasticity-robust estimator (White, 1980). Heteroskedasticity-robust errors do not correct for residual autocorrelation, and for this reason a Newey–West version (Bartlett kernel, five lags) is also used as a robustness check. Tests are run on the full sample, on sub-periods, and on rolling windows of 1,000 observations (about four years of trading days) moved forward by 21 days (one trading month). For comparison with the lag-length choice, fixed lags $p = 5$ and $p = 10$ are reported. A rejection of H_0 is evidence of Granger-causal predictability in the VAR sense; it is not a statement about economic or structural causation, and throughout the paper "causality" is shorthand for the former.

3.3 Forecasting design

For each series s the target is the cumulative log return over $h \in \{1, 5, 22\}$ trading days:

$$y^s_{t,h} = 100 \left[\ln P^s_{t+h} - \ln P^s_t \right] \tag{3}$$

At date t the predictors are nine variables per series, all computed from information available at t: the returns of the last five days, the 5-day and 22-day mean return, the 22-day standard deviation, and the deviation of the log price from its 250-day mean. Three feature sets are compared:

$$X^U = X^{own} \cup X^{other}, \quad X^O = X^{own}, \quad X^G = X^U \text{ if } p_{TY} < \alpha, \text{ and } X^O \text{ otherwise}$$

(4)

The *unguided* set X^U has 18 predictors, the *own-only* set X^O has the nine predictors of the target series, and the *guided* set X^G equals X^O when the screen is closed and X^U when it is open. Here p_{TY} is the p-value of the Toda–Yamamoto test of non-causality from the other series to the target, computed on the 1,000 most recent log-level observations ending at the re-estimation date, and $\alpha = 0.05$. The screen is computed once at each re-estimation date and held fixed until the next one, so the predictor set changes at most once per 63 trading days. The own-only set is the benchmark that isolates what the screen contributes: if guided models beat unguided ones only because they use fewer predictors, they will not beat own-only models.

Four learners are used, each in the three versions. Ridge regression chooses its penalty by generalised (leave-one-out) cross-validation over six values from 0.1 to 10,000 on standardised inputs, using the training window only. Elastic Net chooses its penalty and mixing parameter (0.2, 0.5, 0.8) by three-fold time-ordered cross-validation inside the training window. The random forest has 60 trees, depth 4, at least 40 observations per leaf, half of the predictors per split and half-sample bootstrap. Histogram-based gradient boosting has depth 3, learning rate

0.05, 80 iterations, L2 penalty 5 and at least 60 observations per leaf. Hyper-parameters are fixed in advance at conventional, deliberately modest settings and are not tuned on the test period. With only 60 trees a single random seed gave noticeably different forests (the gain of the guided over the unguided forest at the 5-day horizon for EUR/USD ranged from 0.5 to 1.6 per cent across four seeds), so the random-forest forecasts reported below are the average of four forests grown with seeds 42, 1, 2 and 3. The other learners are deterministic. Three linear benchmarks complete the set. AR is a direct OLS regression of $y_{t,h}$ on up to five own return lags, the number chosen by BIC; VAR adds the same number of lags of the other series; VAR-guided adds them only when the screen is open. The random walk (RW) is a zero return forecast.

The evaluation uses an expanding window. The first forecast is made on 3 January 2012, models are re-estimated every 63 trading days (56 re-estimations, 55 at $h = 22$) using only targets already realised at the re-estimation date, and the out-of-sample period ends on 27 February 2026, with 3,486, 3,482 and 3,465 forecasts at the three horizons. The counts differ because the last $h - 1$ forecast origins have no realised target at the end of the sample; for the same reason the last re-estimation block is not evaluated at $h = 22$.

3.4 Accuracy measures and tests

Accuracy is measured by the root mean squared error (RMSE) and the out-of-sample R^2 against the random walk (Campbell and Thompson, 2008):

$$R^2_{OS} = 1 - \frac{\sum_t e^2_{m,t}}{\sum_t e^2_{RW,t}} \tag{5}$$

Equal predictive accuracy is tested with the Diebold–Mariano statistic on the squared-error differential, where e is the forecast error and the subscripts 0 and 1 index the two models compared:

$$d_t = e^2_{0,t} - e^2_{1,t}, \qquad DM = \frac{D}{\sqrt{V(D)}} \tag{6}$$

Here D is the sample mean of d_t and $V(D)$ its Bartlett HAC variance of lag $h - 1$, which handles the overlap of multi-day forecasts (Newey and West, 1987); the Harvey–Leybourne–Newbold correction and Student-t critical values are used. Because the own-only model is nested in the unguided and guided models, the Clark–West adjusted statistic is also reported for these comparisons:

$$\psi_t = e^2_{r,t} - \left[e^2_{u,t} - (F_{r,t} - F_{u,t})^2 \right] \tag{7}$$

where r denotes the restricted (own-only) model, u the unrestricted one, e the forecast error and F the point forecast; the test is one-sided on the mean of ψ_t with the same HAC variance. Because several comparisons are run for each series, p-values are also adjusted with the Holm (1979) step-down procedure within each family of comparisons. A family consists of the

learner–horizon pairs of one comparison for one series (four learners at three horizons, that is twelve pairs, fewer when the two forecasts coincide); the three comparisons are not pooled. The share of correctly signed forecasts is reported in the text, and the mean absolute percentage error of the implied price-level forecast is not used for inference because it is dominated by today's price (Hyndman and Koehler, 2006).

3.5 Placebo and robustness design

Three exercises probe the guided models. First, a *placebo screen* opens at random with the same frequency as the actual screen (2,000 draws; seed 2026): if the actual screen finds useful moments, it should beat random timing. Second, the screen is varied: thresholds of 1% and 10%, a HAC covariance matrix, a Granger test on stationary returns without extra lags in place of the levels test, and windows of 250, 500, 750 and 2,000 observations, and lag rules chosen by AIC or HQ or fixed at 5 or 10. Because the screen does not change the learners' fitted models, only which of two stored forecasts is used at each date, these variants require no refitting. Third, the evaluation is repeated excluding forecast origins in 2020–2021 and the Toda–Yamamoto tests are repeated on sub-periods that exclude that episode. Fourth, three episodes are analysed separately. They are fixed in advance by the dates of external events rather than chosen from the data: the COVID-19 period (1 January 2020 – 31 December 2021), the Russia–Ukraine war (from the invasion of 24 February 2022 to the end of the sample) and the 2014–2016 oil-price collapse (1 June 2014 – 29 February 2016, during which Brent fell from 115.19 dollars on 19 June 2014 to 26.01 on 20 January 2016). Within each episode I report volatility, Toda–Yamamoto tests on the episode's own observations, and forecast accuracy for the forecasts whose origin falls in the episode. The rest of the out-of-sample period is labelled calm. A Chow-type Wald test with heteroskedasticity-robust errors, using daily returns and five lags of each series, checks whether the predictive coefficients inside each episode differ from those of the rest of the sample. Fifth, the equivalence of guided and own-only forecasts is tested with two one-sided tests (TOST): the mean squared-error difference, expressed as a percentage of the own-only MSE, is tested against margins of ±1% and ±2% with the HAC variance described above, and 90% confidence intervals are reported. The margins are conventions, not estimates of economic significance. Sixth, the screen is refreshed every day instead of every 63 days, choosing each day between the two stored forecasts.

Results

4.1 Descriptive statistics and unit roots

Figure 1 plots the two series and marks the start of the out-of-sample period. Table 1 summarises the returns and the unit-root evidence. Brent returns are extremely fat-tailed (kurtosis 70.9), with a large contribution from the spring of 2020: on 21 April 2020 the price fell from 17.36 to 9.12 dollars, a log return of −64.4 per cent, and rebounded by 41.2 per cent the next day. EUR/USD is far calmer (daily standard deviation 0.60 per cent). Returns are weakly positively correlated with each other (0.110), and the squared returns of both series show strong serial dependence, as expected from volatility clustering.

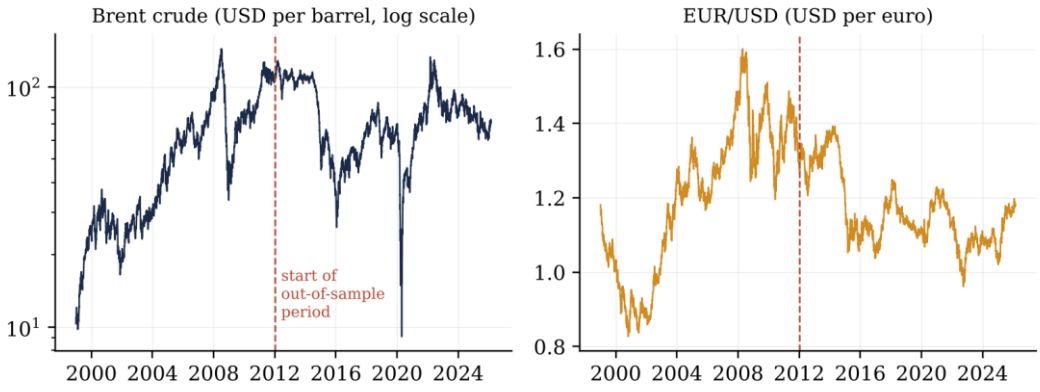

Figure 1. Brent crude and EUR/USD, daily, 4 January 1999 – 27 February 2026. The dashed line marks the start of the out-of-sample period (3 January 2012).

Table 1. Descriptive statistics and unit-root tests

Statistic	Brent	EUR/USD
Observations (log levels)	6,341	6,341
Mean daily return (%)	0.030	0.000
Standard deviation (%)	2.710	0.601
Skewness	−1.83	0.12
Kurtosis	70.92	5.91
Minimum (%)	−64.37 (21 Apr 2020)	−3.87 (16 Mar 2001)
Maximum (%)	41.20 (22 Apr 2020)	4.62 (19 Mar 2009)
Ljung–Box Q(10), returns	27.82 (p = 0.002)	10.87 (p = 0.368)
Ljung–Box Q(10), squared returns	1,328.2 (p < 0.001)	727.4 (p < 0.001)
ADF, log level, constant (lags)	−3.81 (14)	−1.88 (0)
ADF, log level, constant and trend	−3.68	−1.89
ADF, first difference, constant	−19.44	−78.92
KPSS, log level	5.514	2.787
KPSS, first difference	0.173	0.068

Notes: Sample 4 January 1999 – 27 February 2026. Returns are 100 × Δ ln P. ADF lags chosen by AIC (maximum 20); 5% critical values are −2.86 (constant) and −3.41 (constant and trend). The 5% critical value of the KPSS test (constant) is 0.463.

For EUR/USD the evidence is clear: the level is non-stationary and the first difference is stationary. For Brent the tests disagree. The ADF test, with 14 lags, rejects a unit root at the 5% level, while the KPSS test rejects stationarity decisively. The two readings may be reconciled

by long-memory or structural-break behaviour in the oil price, which I do not test, and they are the reason a procedure that does not require the integration order to be known is appropriate. The ADF rejection does not depend on the lag choice: with BIC selection (no lags) the statistic is −3.47, and with 5, 10 and 20 lags it is −3.45, −3.49 and −3.75, all below the −2.86 critical value; the KPSS test with a trend also rejects stationarity (2.287, against a 5% critical value of 0.146). I therefore proceed with the conservative choice $d_{\max} = 1$.

4.2 Granger-causal predictability between Brent and EUR/USD

Table 2 reports the Toda–Yamamoto tests. Over the full sample neither direction is significant, whichever lag rule is used: with the BIC lag ($p = 1$) the statistics are 0.03 (p = 0.855) for Brent → EUR/USD and 1.27 (p = 0.260) for EUR/USD → Brent. The sub-periods give a different picture. Evidence of causality from Brent to EUR/USD appears late in the sample: in 2020–2026 the BIC-lag test rejects at the 5% level ($W = 4.45$, p = 0.035), and in 2016–2026 at the 10% level (p = 0.098). The result is not robust to the exclusion of 2020–2021. In 2022–2026 the direction from Brent to EUR/USD is no longer significant (p = 0.132), while the reverse direction is ($W = 5.05$, p = 0.025); in 2016–2019 and in 2012–2019 neither direction is. With five or ten lags nothing is significant at 5% in any of the five main periods.

Table 2. Toda–Yamamoto causality tests by period and lag rule

Period	n	Lag rule	W: Brent → EUR/USD	p-value	W: EUR/USD → Brent	p-value
Full sample	6,341	BIC (p=1)	0.03	0.855	1.27	0.260
Full sample	6,341	Fixed p = 5	1.05	0.958	3.31	0.653
Full sample	6,341	Fixed p = 10	3.03	0.981	5.51	0.855
1999–2007	1,858	BIC (p=1)	0.25	0.616	0.23	0.634
1999–2007	1,858	Fixed p = 5	2.12	0.832	1.93	0.859
1999–2007	1,858	Fixed p = 10	3.27	0.974	4.72	0.909
2008–2015	1,990	BIC (p=1)	1.45	0.228	2.59	0.108
2008–2015	1,990	Fixed p = 5	3.26	0.661	7.32	0.198
2008–2015	1,990	Fixed p = 10	14.32	0.159	11.15	0.346
2016–Feb 2026	2,493	BIC (p=1)	2.74	0.098	2.41	0.121
2016–Feb 2026	2,493	Fixed p = 5	4.23	0.517	7.89	0.162
2016–Feb 2026	2,493	Fixed p = 10	14.93	0.135	9.40	0.495
2020–Feb 2026	1,505	BIC (p=1)	4.45	0.035	2.52	0.112
2020–Feb 2026	1,505	Fixed p = 5	6.76	0.239	8.30	0.141
2020–Feb 2026	1,505	Fixed p = 10	15.98	0.100	11.64	0.310
2012–2019	1,982	BIC (p=1)	0.26	0.610	0.34	0.561

Period	n	Lag rule	W: Brent → EUR/USD	p-value	W: EUR/USD → Brent	p-value
2016–2019	988	BIC (p=1)	0.84	0.359	0.03	0.868
2022–Feb 2026	1,016	BIC (p=1)	2.26	0.132	5.05	0.025

Notes: VAR in log levels with p + 1 lags, HC1 Wald tests on the first p lags, $\chi^2(p)$ p-values. BIC chooses p = 1 in every period. The last three rows add sub-periods that exclude 2020–2021.

Rolling windows give a sharper picture (Figure 2). Across 255 windows of 1,000 days, the BIC criterion selects one lag every time. Non-causality from Brent to EUR/USD is rejected at 5% in 11.8 per cent of the windows (21.6 per cent at 10%), and from EUR/USD to Brent in 4.7 per cent (8.2 per cent at 10%). With five lags the shares are 9.4 and 0.8 per cent. Overlapping windows are strongly dependent, so these proportions are descriptive. They indicate that Granger-causal predictability between the two series is episodic and that its direction is not stable across the sample. The 30 windows that reject non-causality from Brent to EUR/USD form seven episodes rather than thirty separate events: the longest runs from March 2013 to May 2014 (15 consecutive windows, about 315 trading days of window ends), two further episodes cover October 2023 – March 2024 and July – October 2024 (six and four windows), and the remaining rejections are isolated windows in 2012, July 2014 and June 2025. With six non-overlapping windows of 1,000 days, one rejects Brent → EUR/USD at 5% (the window ending in October 2024, p = 0.049) and one rejects EUR/USD → Brent (ending in July 2008, p = 0.048). BIC selects a single lag in every window and at every re-estimation date, whereas AIC varies between one and ten lags at the re-estimation dates (one lag 16 times, two lags 20, three lags 8, five lags once and ten lags 11 times); other lag rules are therefore examined in Section 4.5. Toda–Yamamoto inference does not require a stable VAR, but explosive roots would be a warning sign. The largest root modulus of the full-sample VAR is 0.999 and, across the rolling windows, it ranges from 0.986 to 1.012, exceeding 1.01 in one of the 255 windows.

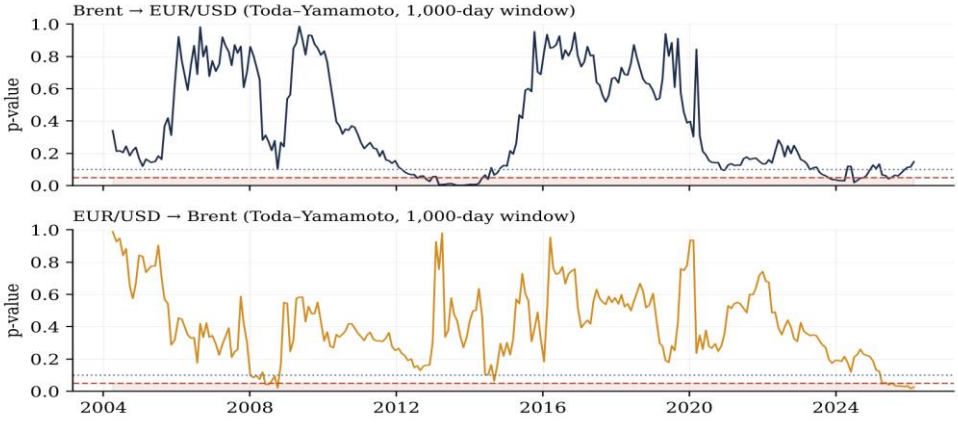

Figure 2. Rolling Toda–Yamamoto p-values (window 1,000 days, step 21 days, BIC lag). The dashed line is the 5% level and the dotted line the 10% level.

4.3 Out-of-sample accuracy

Tables 3 and 4 and Figure 3 report accuracy. The main result is the familiar one: no model beats the random walk in a statistically meaningful way. For EUR/USD every model has a negative out-of-sample R^2 at every horizon, and the random walk is significantly better at the 5% level in 35 of the 45 model–horizon cases (14 of 45 for Brent, Table 4). In no case is a model significantly better than the random walk. The deficit grows with model flexibility and with the horizon: from −0.1 per cent for the AR model at $h = 1$ to −24.2 per cent for unguided gradient boosting at $h = 22$. Elastic Net, which selects and shrinks predictors in one step, is the most accurate learner (R^2 between −0.06 and −1.66 per cent across its three versions), which is what one expects when almost nothing is predictable. Directional accuracy lies between 46 and 55 per cent in all cases.

Brent is more forgiving, and its long-horizon results deserve a closer look. At $h = 22$ the unguided Ridge model reaches an out-of-sample R^2 of 6.6 per cent and a hit rate of 55.1 per cent, the only sizeable positive R^2 in the study. The difference from the random walk is, however, far from significant (DM p = 0.436), and it is the best of 90 model–horizon–series combinations, so it is the kind of result that selection alone can produce. The own-only and guided Ridge versions reach only 1.3 per cent. I do not read any of this as evidence of predictability.

Table 3. Out-of-sample accuracy: EUR/USD (3 January 2012 – 27 February 2026)

Model	RMSE h=1	R^2oos h=1	RMSE h=5	R^2oos h=5	RMSE h=22	R^2oos h=22
Random walk	0.495	—	1.099	—	2.163	—
AR	0.495	-0.08*	1.100	-0.30*	2.181	-1.68**
VAR	0.496	-0.38**	1.101	-0.47**	2.184	-1.98**
VAR (guided)	0.495	-0.11**	1.100	-0.31*	2.181	-1.73**
Ridge – unguided	0.496	-0.36***	1.113	-2.70***	2.302	-13.30***
Ridge – own-only	0.496	-0.17**	1.108	-1.76***	2.241	-7.34***
Ridge – guided	0.496	-0.19**	1.109	-1.96***	2.249	-8.17***
Elastic Net – unguided	0.495	-0.06	1.101	-0.39**	2.180	-1.57*
Elastic Net – own-only	0.495	-0.08	1.101	-0.45*	2.181	-1.66*
Elastic Net – guided	0.495	-0.07	1.101	-0.43**	2.181	-1.66*
Random forest – unguided	0.497	-0.70***	1.114	-2.76***	2.299	-12.95**
Random forest – own-only	0.497	-0.55***	1.109	-1.95***	2.267	-9.86**
Random forest – guided	0.496	-0.49**	1.109	-1.90***	2.265	-9.64**

Model	RMSE h=1	R^2oos h=1	RMSE h=5	R^2oos h=5	RMSE h=22	R^2oos h=22
Gradient boosting – unguided	0.501	-2.19***	1.140	-7.58***	2.410	-24.22***
Gradient boosting – own-only	0.499	-1.69***	1.131	-5.98***	2.323	-15.35**
Gradient boosting – guided	0.499	-1.48***	1.131	-5.94***	2.334	-16.46**

*Notes: RMSE in percentage points of the h-day log return; R^2oos in per cent relative to the zero-return random walk. Stars give the two-sided Diebold–Mariano test against the random walk: * p < 0.10, ** p < 0.05, *** p < 0.01. A negative R^2oos with stars means that the random walk is significantly better. Unguided: all 18 predictors; own-only: the nine predictors of the target series; guided: the other series' predictors enter only when the Toda–Yamamoto screen rejects non-causality at 5%.*

Table 4. Out-of-sample accuracy: Brent (3 January 2012 – 27 February 2026)

Model	RMSE h=1	R^2oos h=1	RMSE h=5	R^2oos h=5	RMSE h=22	R^2oos h=22
Random walk	2.787	—	6.007	—	13.372	—
AR	2.788	-0.12	6.015	-0.28	13.430	-0.88*
VAR	2.788	-0.10	6.016	-0.30*	13.433	-0.91*
VAR (guided)	2.788	-0.13	6.015	-0.28	13.430	-0.88*
Ridge – unguided	2.788	-0.07	6.007	-0.01	12.920	6.64
Ridge – own-only	2.789	-0.17	6.021	-0.47	13.287	1.26
Ridge – guided	2.789	-0.17	6.022	-0.49	13.283	1.32
Elastic Net – unguided	2.787	-0.04*	6.014	-0.23**	13.424	-0.78*
Elastic Net – own-only	2.787	-0.04*	6.014	-0.23**	13.424	-0.78*
Elastic Net – guided	2.787	-0.04*	6.014	-0.23**	13.424	-0.78*
Random forest – unguided	2.790	-0.27	6.077	-2.35**	13.759	-5.88**
Random forest – own-only	2.790	-0.23*	6.067	-2.00***	13.731	-5.45***
Random forest – guided	2.790	-0.24*	6.065	-1.96***	13.720	-5.28***
Gradient boosting – unguided	2.801	-1.01	6.145	-4.66	13.930	-8.53

Model	RMSE h=1	R^2oos h=1	RMSE h=5	R^2oos h=5	RMSE h=22	R^2oos h=22
Gradient boosting – own-only	2.796	-0.67*	6.127	-4.02***	13.940	-8.68***
Gradient boosting – guided	2.796	-0.68**	6.127	-4.05***	13.933	-8.57***

Notes: As in Table 3.

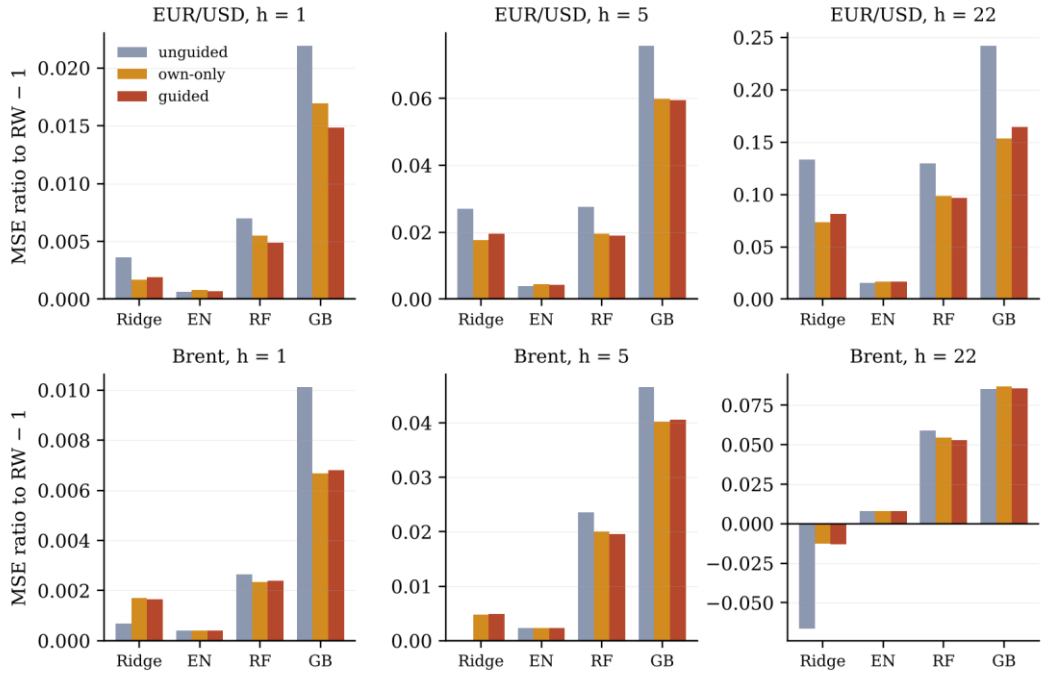

Figure 3. Mean squared error of each learner relative to the random walk, minus one (positive values mean the learner is worse). Grey: unguided; amber: own-only; red: guided. EN is Elastic Net and GB gradient boosting.

4.4 What does the screen add?

Table 5 summarises three comparisons for each series: own-only against unguided (the effect of dropping the other series), guided against unguided, and guided against own-only (the effect of the information in the screen). Table 6 gives the detail for EUR/USD, and Brent is in Appendix B.

Table 5. Summary of pairwise comparisons by series

Comparison	Series	n	Mean MSE gain (%)	Gains > 0	p < 0.05 (raw)	p < 0.05 (Holm)
Own-only vs unguided	EUR/USD	12	1.58	9 / 12	4	3
Own-only vs unguided	Brent	9	-0.39	5 / 12	0	0
Guided vs unguided	EUR/USD	12	1.48	9 / 12	3	2
Guided vs unguided	Brent	9	-0.37	5 / 12	0	0
Guided vs own-only	EUR/USD	11	-0.11	7 / 12	1	0
Guided vs own-only	Brent	9	0.03	5 / 12	0	0

Notes: A gain is 100 × (1 − MSE of the second model / MSE of the first), so positive values favour the second model named. n is the number of learner–horizon pairs for which the two forecasts differ (Elastic Net often selects no cross-series predictor, making guided and own-only forecasts identical); the gains count includes those pairs with zero gain. Raw and Holm-adjusted counts refer to two-sided Diebold–Mariano tests.

For EUR/USD the answer is clear. Dropping the Brent predictors altogether improves the MSE of the learners by 1.58 per cent on average, in nine of twelve cases, with four raw and three Holm-significant gains. The guided versions do almost the same (1.48 per cent, nine of twelve, three raw, two Holm-significant, namely Ridge at $h = 5$ and $h = 22$). But the guided models are no better than the own-only models: the average gain of guided over own-only is −0.11 per cent, seven of twelve are positive, and the single raw-significant difference (Ridge at $h = 5$, $p = 0.018$) favours the own-only model. The Clark–West test, which is better suited to the nested comparison, gives small p-values for gradient boosting at $h = 1$ (0.007) and the random forest at $h = 1$ (0.023), and weakly for gradient boosting at $h = 5$ (0.075); none survives a Holm adjustment across the eleven comparisons. For Brent, where the screen opens in only three of 56 re-estimations, all three comparisons are indistinguishable from zero (Table 5).

The −0.11 per cent difference between guided and own-only models is also tested for equivalence rather than only for a difference. With a margin of ±1 per cent of the own-only MSE, the TOST procedure establishes equivalence at the 5% level in nine of the eleven EUR/USD cases in which the two forecasts differ, and with a margin of ±2 per cent in ten of eleven; for Brent equivalence holds in all nine cases at both margins. The two exceptions are at the 22-day horizon (Ridge and gradient boosting), where the 90 per cent confidence intervals are wider: for gradient boosting the interval for the gain of guided over own-only models is −2.4 to +0.4 per cent. Equivalence is therefore well established at short horizons and not at the longest one, where the data cannot exclude a modest loss or gain from the screen. A difference of 1 per cent in MSE corresponds to a difference of about 0.5 per cent in RMSE; the full 90 per cent intervals are given in Appendix C so that readers can apply their own margin.

Table 6. EUR/USD: pairwise comparisons by learner and horizon

Learner	h	Own-only vs unguided: gain (%)	p	Guided vs unguided: gain (%)	p	Guided vs own-only: gain (%)	p	Clark–West p
Ridge	1	0.19	0.003	0.17	0.007	-0.02	0.146	0.907
Ridge	5	0.92	0.000	0.72	0.001	-0.20	0.018	0.980
Ridge	22	5.27	0.000	4.53	0.000	-0.77	0.126	0.823
Elastic Net	1	-0.02	0.462	-0.01	0.511	0.01	0.650	0.233
Elastic Net	5	-0.06	0.428	-0.05	0.200	0.02	0.808	0.312
Elastic Net	22	-0.09	0.128	-0.09	0.128	—	—	—
Random forest	1	0.15	0.297	0.21	0.121	0.06	0.096	0.023
Random forest	5	0.78	0.088	0.84	0.054	0.06	0.702	0.163
Random forest	22	2.73	0.133	2.93	0.099	0.20	0.644	0.172
Gradient boosting	1	0.49	0.219	0.69	0.067	0.21	0.069	0.007
Gradient boosting	5	1.48	0.161	1.52	0.134	0.04	0.901	0.075
Gradient boosting	22	7.14	0.039	6.24	0.064	-0.96	0.264	0.474

Notes: Gain in per cent of MSE; p is the two-sided Diebold–Mariano p-value with HAC variance (lag h − 1) and Harvey–Leybourne–Newbold correction; Clark–West p is one-sided for guided against own-only. A dash marks pairs in which the two forecasts are identical.

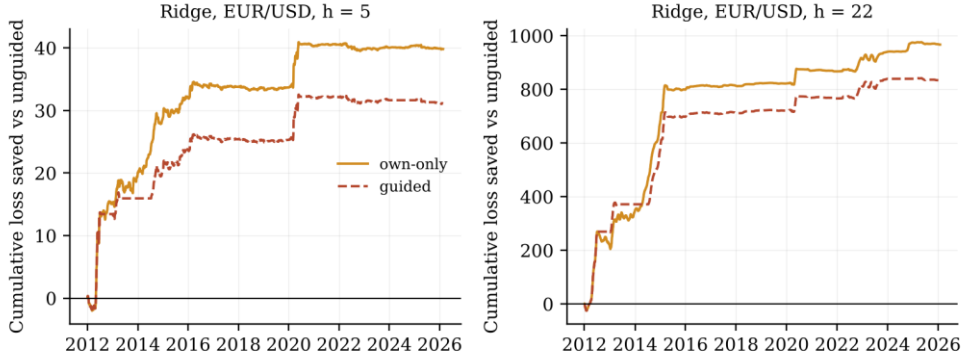

Figure 4. Cumulative squared-error loss saved relative to the unguided Ridge model for EUR/USD, by the own-only model (solid amber) and the guided model (dashed red), at horizons of 5 and 22 days. The two lines follow each other closely.

4.5 Robustness

Placebo screens. The actual screen opens in 21.4 per cent of the EUR/USD re-estimations. Drawing 2,000 random screens with the same frequency gives a mean gain over the unguided

models of 1.24 per cent, against 1.48 per cent for the actual screen, which lies at the 72.8th percentile of the placebo distribution. Against own-only models, the placebo mean is −0.36 per cent and the actual screen's is −0.11 per cent; 27.6 per cent of random screens do at least as well (Figure 5). The causality screen therefore cannot be distinguished from random timing.

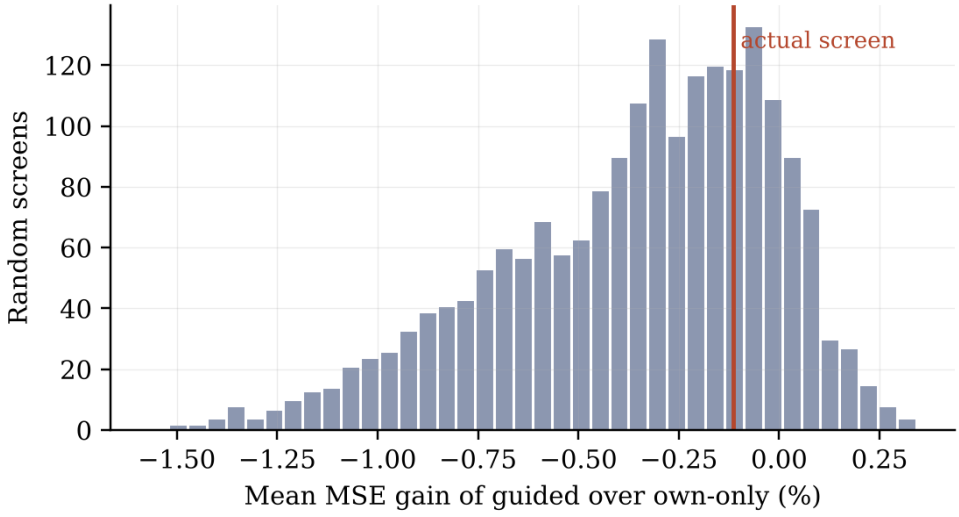

Figure 5. Distribution of the mean MSE gain of guided over own-only models (twelve EUR/USD learner–horizon cases) across 2,000 random screens with the same opening frequency as the actual screen. The vertical line is the actual screen.

Alternative screens. Table 7 varies the screen. Tightening the threshold to 1% opens it in 5.4 per cent of the EUR/USD re-estimations and makes guided models almost identical to own-only models (+0.08 per cent). Relaxing it to 10% opens it in 41.1 per cent of them and the guided models then lose to own-only models (−1.08 per cent). A HAC covariance matrix, a Granger test on returns instead of levels, and windows of 500 days give practically the same conclusions as the baseline; with a window of 2,000 days the screen opens in 7.1 per cent of the cases and the conclusion is unchanged. For Brent no variant produces a gain. Lag rules and shorter windows change the opening frequency but not the conclusion (Table 7). With the lag chosen by AIC the screen opens in 32.1 per cent of the EUR/USD forecasts and the gain over own-only models is +0.27 per cent in the pairs where the forecasts differ, with two raw-significant comparisons of eleven and none after the Holm adjustment; with fixed lags of 5 and 10 the gains are +0.05 and +0.13 per cent and none is significant, and with the HQ rule it is −0.04 per cent. Windows of 250 and 750 days give −0.13 and −0.34 per cent. Refreshing the screen every day rather than every 63 days gives the same picture: it opens in 19.4 per cent of the EUR/USD forecasts, the gain over own-only models is 0.03 per cent (positive in 75 per cent of the twelve cases) and the gain over unguided models 1.61 per cent; for Brent it opens in 5.0 per cent of the forecasts and the gains are −0.39 and 0.00 per cent.

Table 7. Robustness of the guided models to the specification of the screen

Screen	EUR/USD open (%)	Gain vs U (%)	Gain vs own (%)	Gain > 0 vs own (%)	Brent open (%)	Gain vs U (%)	Gain vs own (%)
Baseline: TY levels, HC, 1,000-day window, 5%	21.4	1.48	-0.11	58	5.4	-0.37	0.03
Threshold 1%	5.4	1.65	0.08	58	0.0	-0.39	0.00
Threshold 10%	41.1	0.55	-1.08	33	7.1	-0.41	-0.01
HAC covariance (Bartlett, 5 lags)	21.4	1.50	-0.08	58	3.6	-0.37	0.03
Granger test on returns (no d_max)	23.2	1.47	-0.12	58	3.6	-0.37	0.03
Window 500 days	16.1	1.49	-0.10	67	3.6	-0.41	-0.01
Window 2,000 days	7.1	1.50	-0.09	25	0.0	-0.39	0.00
Screen refreshed daily (instead of every 63 days)	19.4	1.61	0.03	75	5.0	-0.39	0.00
Window 250 days	16.1	1.46	-0.13	17	7.1	-0.27	0.13
Window 750 days	17.9	1.26	-0.34	42	5.4	-0.37	0.03
Lag chosen by AIC (up to 10)	32.1	1.82	0.25	67	0.0	-0.39	0.00
Lag chosen by HQ (up to 10)	21.4	1.55	-0.03	67	5.4	-0.37	0.03
Fixed p = 5	12.5	1.63	0.05	58	0.0	-0.39	0.00
Fixed p = 10	25.0	1.69	0.12	67	0.0	-0.39	0.00

Notes: Gains are averages over the twelve learner–horizon cases of each series (100 × (1 − MSE guided / MSE comparator)); U denotes unguided. Open (%) is the share of the 56 re-estimations (of the forecasts, in the last row) in which the screen opens.

Forest size. The random forests have 60 trees each and the reported forecasts average four of them (240 trees). To judge whether this is enough, the averages of seeds (42, 1) and (2, 3) were compared for EUR/USD: the gain of the guided over the unguided forest at the three horizons is 0.20, 1.01 and 3.10 per cent in the first half and 0.23, 0.70 and 2.66 per cent in the second, and the gain of guided over own-only forests stays below 0.3 percentage points in both. The halves agree to within about half a percentage point, which is small relative to the effects discussed, but a forest of several hundred trees was not grown. *Excluding 2020–2021.* Dropping forecast origins in 2020 and 2021, the guided models' average gain over unguided is 1.27 per cent for EUR/USD and over own-only −0.13 per cent; for Brent the figures are −0.02 and 0.05 per cent. The screen opens for EUR/USD in 2012–2014 (seven re-estimations) and in 2024–2025 (five), and the later windows contain 2020–2021 in their 1,000-day span, which is consistent with the sub-period results of Section 4.2. *Inference.* For the two comparisons that

survive the Holm adjustment in the guided-versus-unguided family (Ridge at $h = 5$ and $h = 22$), the p-values stay below 5 per cent when the HAC lag is raised from $h - 1$ to $2h$ and $5h$ (for example 0.0006, 0.0044 and 0.0147 for Ridge at $h = 5$). The conclusion that dropping the other series helps is therefore not an artefact of the overlap treatment.

4.6 COVID-19, the oil collapse and the war in Ukraine

Table 8 describes the three episodes and tests causality inside each. Volatility differs sharply across them and across the two series. Brent's daily standard deviation is 5.41 per cent during COVID-19, three times its calm-period level (1.78 per cent), and is 2.27 and 2.48 per cent in the war and the 2014–2016 collapse. EUR/USD is much steadier: 0.41 per cent during COVID-19, 0.50 in the war and 0.66 in 2014–2016, against 0.46 in calm periods. The euro–dollar rate was therefore most volatile in 2014–2016, not in the pandemic.

Table 8. Volatility and Toda–Yamamoto tests inside the episodes

Episode	n	Brent SD (%)	EUR/USD SD (%)	Brent → EUR/USD, BIC lag: W (p)	EUR/USD → Brent, BIC lag: W (p)	Brent → EUR/USD, 5 lags: W (p)	EUR/USD → Brent, 5 lags: W (p)
COVID-19	489	5.41	0.41	3.34 (0.068)	0.40 (0.528)	12.92 (0.024)	5.34 (0.376)
Russia–Ukraine war	980	2.27	0.50	2.03 (0.154)	5.95 (0.015)	5.85 (0.321)	7.49 (0.187)
2014–2016 oil collapse	433	2.48	0.66	0.01 (0.924)	0.02 (0.889)	3.21 (0.668)	4.43 (0.490)
Calm (test period outside the episodes)	1,585	1.78	0.46	—	—	—	—

Notes: n is the number of daily observations of the episode; SD is the standard deviation of daily returns. Tests use the log levels of the episode's own observations (VAR with p + 1 lags, HC1 Wald). With a Newey–West covariance matrix none of the rejections survives except EUR/USD → Brent in the war with the BIC lag (p = 0.017); in COVID-19 the five-lag rejection of Brent → EUR/USD has p = 0.119.

No episode shows robust Granger-causal predictability in either direction. In COVID-19 the Brent-to-EUR/USD test is marginal with the BIC lag (p = 0.068) and significant with five lags (p = 0.024), but neither survives a HAC covariance matrix (p = 0.108 and 0.119). In the war the only robust rejection runs the other way, from EUR/USD to Brent (p = 0.015, HAC p = 0.017), and it vanishes with five lags (p = 0.187). The 2014–2016 collapse shows nothing. This confirms the picture of Section 4.2: where predictability appears it is brief, fragile and not consistently directed.

Table 9. Forecast accuracy and the role of the other series, by episode

Episode	Series	n	Own-only vs unguided: gain (%)	Guided vs unguided: gain (%)	Guided vs own-only: gain (%)	Screen open (%)	Highest R^2oos (%)
COVID-19	EUR/USD	489	3.17	3.17	0.00	0	0.43
COVID-19	Brent	489	−0.70	−0.70	0.00	0	14.71
Russia–Ukraine war	EUR/USD	970	−0.63	−1.24	−0.60	32	2.70
Russia–Ukraine war	Brent	970	−1.50	−1.25	0.23	14	0.26
2014–2016 oil collapse	EUR/USD	433	3.47	3.47	0.00	7	−0.23
2014–2016 oil collapse	Brent	433	0.94	0.94	0.00	0	0.01
Calm periods	EUR/USD	1,585	1.14	1.27	0.13	26	0.05
Calm periods	Brent	1,585	−0.60	−0.60	0.00	0	0.18

Notes: Forecasts whose origin falls in the episode; n is the average number of forecasts per horizon. Gains are 100 × (1 − MSE of the second model / MSE of the first), averaged over the twelve learner–horizon cases, with identical-forecast pairs counted as zero. Screen open is the share of those forecasts made while the screen was open. Highest R^2oos is the largest out-of-sample R^2 against the random walk among the 45 model–horizon cases of the episode; no case beats the random walk significantly (smallest DM p-value among positive cases: 0.205).

Table 9 shows that no model beats the random walk in any episode. The largest positive R^2 is 14.7 per cent for Brent in COVID-19 (unguided Ridge at $h = 22$), but it is not significant (DM $p = 0.340$) and it is driven by March–June 2020: leaving those four months out, it falls to 0.9 per cent. The role of the other series changes with the episode. For EUR/USD, Brent's predictors hurt forecasts during COVID-19 and the 2014–2016 collapse: dropping them saves 3.2 and 3.5 per cent of the MSE on average, and in the pairs where the two models differ the unguided model is significantly worse in one of nine cases in COVID-19 and two of eleven in 2014–2016 after the Holm adjustment (mean loss from using Brent's predictors of 4.7 and 4.2 per cent). In the war the sign reverses. Using the other series' predictors improves forecasts in six of nine cases for Brent and seven of nine for EUR/USD (mean MSE gains of 1.8 and 0.8 per cent relative to own-only learners). Only one Diebold–Mariano test is significant before adjustment (EUR/USD, gradient boosting at $h = 5$, $p = 0.031$) and none after, but the Clark–West test is significant at 5% in three Brent cases and three EUR/USD cases. This is suggestive evidence, consistent with the war-period Toda–Yamamoto result, that the cross-series information was of some value in that episode.

The screen behaves sensibly but unhelpfully. It stays closed throughout COVID-19 and nearly so in the 2014–2016 collapse (7 per cent of EUR/USD forecasts), so guided models there equal own-only models and are 3.2 and 3.5 per cent better than unguided ones, a gain that is a form of protection against harmful predictors. During the war it opens for 32 per cent of the EUR/USD forecasts and 14 per cent of the Brent forecasts, which is too late and too rarely to capture the benefit: guided models are 1.2 per cent worse than unguided for EUR/USD, and 0.6 per cent worse than own-only. A screen with a 1,000-day look-back that refreshes quarterly reacts to a shift in the information content of the other series only after a long delay.

Table 10. Chow-type tests of equal predictive coefficients inside the episodes

Episode	Series	All coefficients: W (p)	Other series' lags only: W (p)
COVID-19	EUR/USD	17.4 (0.095)	11.0 (0.052)
Russia–Ukraine war	EUR/USD	8.6 (0.663)	7.2 (0.207)
2014–2016 oil collapse	EUR/USD	9.2 (0.600)	3.6 (0.606)
COVID-19	Brent	8.2 (0.695)	4.4 (0.491)
Russia–Ukraine war	Brent	9.1 (0.610)	2.8 (0.727)
2014–2016 oil collapse	Brent	13.6 (0.254)	3.8 (0.577)

Notes: Regression of the daily return of the series on its own five lags and five lags of the other series, with all coefficients interacted with an episode dummy; HC1 Wald tests of equality with the rest of the sample (12 and 5 restrictions in the first and second column respectively).

None of the six tests rejects at the 5% level (the smallest p-value is 0.052, for the other series' lags of EUR/USD in COVID-19), so the differences between episodes described above are not statistically established and should be read as descriptive.

4.7 Stability of the predictor set and cost

The screen opens for EUR/USD in 12 of the 56 re-estimations and switches state eight times, so that the average guided model has 10.9 predictors against nine for the own-only and 18 for the unguided model. For Brent it opens three times and switches three times, with 9.5 predictors on average. The Toda–Yamamoto test costs about 0.6 milliseconds per re-estimation, whereas fitting one learner takes roughly 0.08 seconds on average. The computational cost of the screen is therefore negligible; the question is whether it buys anything. Given that dropping the other series is cheaper still and captures the whole gain, it does not.

Discussion

Three findings stand out. The first is that the random walk remains unbeaten. This is the Meese–Rogoff result in a new guise, and it holds for flexible learners as well as linear ones. If anything, flexibility hurts: gradient boosting is the worst learner for EUR/USD at every horizon, and the most accurate one, Elastic Net, is the one that shrinks closest to a constant forecast. With daily returns whose predictable component is, at best, tiny, a flexible model has little to gain and a lot of noise to fit.

The second finding is that Granger-causal predictability is real but intermittent and unstable in direction. The full-sample test finds nothing, as the instability reported by Beckmann and Czudaj (2013) and Beckmann, Czudaj and Arora (2020) would lead one to expect, yet rolling windows show Brent helping to predict EUR/USD in about one window in nine. Sub-period results show that this is not stable. The Brent-to-EUR/USD evidence of 2020–2026 disappears when 2020–2021 is excluded, and in 2022–2026 the reverse direction is the significant one. A forecaster who relies on a single test will conclude there is nothing to exploit; one who tests repeatedly will find episodes whose direction changes. The dominant direction in this sample differs from that reported by Beckmann and Czudaj (2013), and I do not draw a conclusion from the discrepancy since the studies differ in currencies, frequency and sample.

The third finding answers the question in the title. Guidance does reduce the loss of machine-learning models for EUR/USD relative to feeding them everything, but the benefit is the benefit of not using the other series. Own-only models reach the same gain, guided models do not improve on them, and random timing of the screen performs about as well as the causality test. In feature-selection terms (Guyon and Elisseeff, 2003), the screen works as a filter, but it is a filter whose decisions carry no measurable information in this application. A possible reason is that Granger-causal predictability found in rolling windows is episodic and does not persist for the 63 days of the next forecasting period; I have not tested this explanation. Another is that the screen is tested on log levels while the learners predict returns, but the variant with a Granger test on returns gives the same results. What the data support is narrower than a claim of regularisation by causality: dropping weak predictors helps flexible learners when signals are scarce.

The episode analysis qualifies the picture without overturning it. COVID-19 and the war in Ukraine are the two shocks one would expect to link oil and the dollar, yet neither produces robust causality in both directions, and the models' errors are not smaller in them than in calm periods. What changes is the value of the cross-series predictors: harmful in COVID-19 and in the 2014–2016 collapse, mildly helpful in the war. If the information in the other series is regime-dependent in this way, a screen that looks back 1,000 days is badly suited to detect it, and a design with shorter windows, a regime indicator or an exogenous geopolitical-risk measure would be a better test of the idea. I do not test those here.

None of this should be oversold in either direction. The evidence does not show that the other series is useless: the Clark–West test is mildly supportive for gradient boosting at short horizons, and a ten-year test on one pair of series cannot rule out small effects. It shows that, in this setting, a causality screen does not improve on the simplest alternative, and that the random walk stays out of reach for every variant.

Conclusion and limitations

This study examined whether time-varying Granger-causality tests can improve machine-learning forecasts of Brent crude oil prices and the EUR/USD exchange rate by guiding the inclusion of cross-series predictors. The analysis combined daily data from 4 January 1999 to 27 February 2026 with an expanding-window forecasting exercise starting in January 2012.

Ridge regression, Elastic Net, random forests and gradient boosting were evaluated at horizons of 1, 5 and 22 trading days. The central issue was whether a causality-based screening rule can do more than reduce the number of predictors supplied to the forecasting models.

The findings lead to three main conclusions. First, the random walk remains a difficult benchmark. None of the machine-learning specifications significantly improves on it for either series at any horizon. The positive out-of-sample R^2 of the unguided Ridge model for Brent at the 22-day horizon (6.6 per cent) is noteworthy, but its difference from the random walk is not statistically significant (Diebold–Mariano $p = 0.436$), and it is the best of 90 model–horizon–series combinations. It cannot be treated as evidence of reliable forecasting superiority. More generally, the results underline how hard it is to extract stable information from daily financial returns, even with flexible learning algorithms.

Second, Granger-causal predictability between Brent and EUR/USD is intermittent rather than stable. The full-sample Toda–Yamamoto tests find no significant relationship in either direction under any of the lag specifications used. Rolling tests, by contrast, detect episodes in which predictability from one series to the other becomes significant, and the direction changes across sub-periods. In particular, the evidence of influence from Brent to EUR/USD in the later sample is sensitive to the inclusion of 2020–2021, while the 2022–2026 sub-period gives evidence in the reverse direction under the BIC-selected specification. These results caution against treating a full-sample relationship as representative of the whole period, and they show why the timing of predictive relationships deserves attention when forecasting procedures are designed.

Third, and most important for the question in the title, the causality screen does not show an incremental forecasting advantage. For EUR/USD, guided models reduce mean squared forecast error by 1.48 per cent on average relative to unguided models, and two of the twelve comparisons remain significant after Holm adjustment. Yet own-only models achieve an average reduction of 1.58 per cent relative to unguided models, and the guided specifications do not improve on them (average relative gain of −0.11 per cent). Randomly timed screens with the same opening frequency give broadly similar outcomes: 27.6 per cent of them do at least as well as the actual screen. The improvement associated with guidance is therefore better explained by the exclusion of cross-series predictors than by the information in the test. The Clark–West test, which suits these nested comparisons, is mildly supportive of some cross-series information for gradient boosting and the random forest at short horizons, but none of these results survives a Holm adjustment. For Brent the screen opens in only three of 56 re-estimations and produces no meaningful improvement. These conclusions are unchanged under other lag rules, window lengths and thresholds.

The episode analysis adds context. Cross-series predictors do not contribute uniformly across the periods examined. Brent's predictors tend to worsen EUR/USD forecasts during COVID-19 and the 2014–2016 oil-price collapse, whereas their contribution is modestly more favourable during the Russia–Ukraine war. These differences do not establish that the events changed the predictive relationship. The episode-specific tests give limited and sometimes specification-sensitive evidence, a Chow-type test does not reject stability of the predictive coefficients in any episode, and no model beats the random walk in any episode. The results

should be read as descriptive evidence of variation across periods, not as proof of event-specific causal effects.

For forecasting practice, the results suggest that adding information from a related financial market should be justified by its contribution to out-of-sample accuracy rather than by the existence of an in-sample or rolling-window predictive relationship. In the present application, restricting the predictor set to the target series' own history captures the observed improvement without a causality-based screen, and is cheaper still. This does not imply that cross-series information is never useful. It indicates that the particular rolling Toda–Yamamoto rule examined here does not identify the periods in which such information improves forecasts beyond the simpler own-only alternative. For policy-minded readers, the results offer no support for projections that rely on the oil price to predict the euro–dollar rate.

Several limitations should be borne in mind. The analysis covers one commodity and one nominal exchange-rate pair at daily frequency; results may differ for other currencies, commodities, real effective exchange rates or lower-frequency data. Only point forecasts of conditional means are evaluated, not volatility, density or tail-risk forecasts, where financial-market relationships may behave differently, and their economic value, for example in trading rules after transaction costs, is not assessed. The estimation window is expanding; rolling windows were not examined for the learners. Hyper-parameters were fixed in advance rather than tuned, which limits the risk of selecting specifications on test-period performance but may understate what the learners could achieve. A random forest of several hundred trees was not grown: four forests of 60 trees are averaged instead. Single-seed random-forest results proved sensitive to the seed, which is why four seeds are averaged; the averaged forest may still differ slightly from another draw.

The statistical evaluation has limitations too. Comparisons across horizons and learners are not independent, and the rolling causality windows overlap substantially. Holm's adjustment addresses multiple testing within the stated families of comparisons but does not remove all dependence among tests. Diebold–Mariano tests applied to nested comparisons can be undersized (Clark and West, 2007), which is why the Clark–West results are reported. The screening rule rests mainly on Toda–Yamamoto tests applied to log levels; the returns-based variant and alternative window lengths give similar broad findings, but other approaches to detecting nonlinear or time-varying predictive relationships may behave differently. Finally, the historical episodes are defined by calendar dates, no geopolitical-risk indicator is used, and some episode-specific tests rest on one to two years of data, so their results call for caution.

Future research could examine whether the distinction between information-based screening and simple dimensionality reduction persists across other currency pairs and commodity markets, and whether causality-guided selection is more useful for forecasting volatility or other features of the return distribution than for conditional mean returns. Screens applied to individual predictors, shorter or adaptive windows, regime-switching or time-varying-parameter VARs, and methods that let parameters evolve more quickly may show whether the limited performance observed here reflects the particular rule or a more general difficulty in turning time-varying predictive relationships into reliable forecasts. The principal conclusion

is specific but important. In daily Brent and EUR/USD forecasting, reducing the predictor set can improve some machine-learning forecasts, but the rolling Toda–Yamamoto screen does not show that its timing decisions add value beyond using the target series' own information. Detecting predictive relationships and exploiting them are separate empirical tasks, and success in the first does not guarantee success in the second.

Appendix A. Brent: pairwise comparisons by learner and horizon

Learner	h	Own-only vs unguided: gain (%)	p	Guided vs unguided: gain (%)	p	Guided vs own-only: gain (%)	p	Clark–West p
Ridge	1	-0.10	0.136	-0.10	0.152	0.01	0.589	0.245
Ridge	5	-0.46	0.729	-0.48	0.718	-0.02	0.816	0.401
Ridge	22	-5.76	0.277	-5.69	0.282	0.07	0.807	0.205
Elastic Net	1	—	—	—	—	—	—	—
Elastic Net	5	—	—	—	—	—	—	—
Elastic Net	22	—	—	—	—	—	—	—
Random forest	1	0.03	0.901	0.02	0.923	-0.01	0.505	0.713
Random forest	5	0.34	0.754	0.39	0.722	0.05	0.265	0.099
Random forest	22	0.41	0.872	0.58	0.822	0.16	0.378	0.115
Gradient boosting	1	0.34	0.692	0.33	0.705	-0.01	0.433	0.651
Gradient boosting	5	0.61	0.829	0.58	0.837	-0.03	0.685	0.530
Gradient boosting	22	-0.14	0.977	-0.04	0.993	0.10	0.584	0.157

Notes: As in Table 6. For Brent the screen opens in only three of 56 re-estimations.

Appendix B. Equivalence tests: guided versus own-only models

Series	h	Learner	Gain (%)	90% confidence interval	TOST p (±1%)	TOST p (±2%)
Brent	1	Gradient boosting	-0.01	[-0.05, 0.02]	0.000	0.000
Brent	1	Random forest	-0.01	[-0.02, 0.01]	0.000	0.000
Brent	1	Ridge	0.01	[-0.01, 0.02]	0.000	0.000
Brent	5	Gradient boosting	-0.03	[-0.15, 0.09]	0.000	0.000
Brent	5	Random forest	0.05	[-0.02, 0.11]	0.000	0.000
Brent	5	Ridge	-0.02	[-0.15, 0.11]	0.000	0.000
Brent	22	Gradient boosting	0.10	[-0.19, 0.39]	0.000	0.000

Series	h	Learner	Gain (%)	90% confidence interval	TOST p (±1%)	TOST p (±2%)
Brent	22	Random forest	0.16	[-0.14, 0.47]	0.000	0.000
Brent	22	Ridge	0.07	[-0.38, 0.51]	0.000	0.000
EUR/USD	1	Elastic Net	0.01	[-0.02, 0.04]	0.000	0.000
EUR/USD	1	Gradient boosting	0.21	[0.02, 0.40]	0.000	0.000
EUR/USD	1	Random forest	0.06	[0.00, 0.13]	0.000	0.000
EUR/USD	1	Ridge	-0.02	[-0.04, 0.00]	0.000	0.000
EUR/USD	5	Elastic Net	0.02	[-0.10, 0.13]	0.000	0.000
EUR/USD	5	Gradient boosting	0.04	[-0.45, 0.53]	0.001	0.000
EUR/USD	5	Random forest	0.06	[-0.18, 0.29]	0.000	0.000
EUR/USD	5	Ridge	-0.20	[-0.35, -0.06]	0.000	0.000
EUR/USD	22	Gradient boosting	-0.96	[-2.37, 0.45]	0.482	0.113
EUR/USD	22	Random forest	0.20	[-0.50, 0.90]	0.030	0.000
EUR/USD	22	Ridge	-0.77	[-1.60, 0.05]	0.326	0.007

Notes: Gain of the guided over the own-only model in per cent of the own-only MSE, with a 90 per cent confidence interval from the HAC variance; TOST p is the larger of the two one-sided p-values for the stated margin. Pairs in which the two forecasts coincide are omitted.

References

Alquist, R., Kilian, L., & Vigfusson, R. J. (2013). Forecasting the price of oil. In G. Elliott & A. Timmermann (Eds.), *Handbook of Economic Forecasting* (Vol. 2, pp. 427–507). Elsevier.

Amano, R. A., & van Norden, S. (1998). Oil prices and the rise and fall of the US real exchange rate. *Journal of International Money and Finance, 17*(2), 299–316.

Baumeister, C., & Kilian, L. (2015). Forecasting the real price of oil in a changing world: A forecast combination approach. *Journal of Business & Economic Statistics, 33*(3), 338–351.

Beckmann, J., & Czudaj, R. (2013). Is there a homogeneous causality pattern between oil prices and currencies of oil importers and exporters? *Energy Economics, 40*, 665–678.

Beckmann, J., Czudaj, R. L., & Arora, V. (2020). The relationship between oil prices and exchange rates: Revisiting theory and evidence. *Energy Economics, 88*, 104772. https://doi.org/10.1016/j.eneco.2020.104772

Board of Governors of the Federal Reserve System (US). U.S. dollars to euro spot exchange rate [DEXUSEU]. FRED, Federal Reserve Bank of St. Louis. https://fred.stlouisfed.org/series/DEXUSEU

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Campbell, J. Y., & Thompson, S. B. (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *Review of Financial Studies, 21*(4), 1509–1531.

Chen, Y.-C., Rogoff, K. S., & Rossi, B. (2010). Can exchange rates forecast commodity prices? *The Quarterly Journal of Economics, 125*(3), 1145–1194.

Clark, T. E., & West, K. D. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics, 138*(1), 291–311.

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association, 74*(366), 427–431.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics, 13*(3), 253–263.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29*(5), 1189–1232.

Golub, S. S. (1983). Oil prices and exchange rates. *The Economic Journal, 93*(371), 576–593.

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *Review of Financial Studies, 33*(5), 2223–2273.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*, 1157–1182.

Guyon, I., Aliferis, C., & Elisseeff, A. (2007). Causal feature selection. In H. Liu & H. Motoda (Eds.), *Computational Methods of Feature Selection*. Chapman & Hall/CRC.

Harvey, D., Leybourne, S., & Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting, 13*(2), 281–291.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55–67.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*(2), 65–70.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting, 22*(4), 679–688.

Krugman, P. (1983). Oil and the dollar. In J. S. Bhandari & B. H. Putnam (Eds.), *Economic Interdependence and Flexible Exchange Rates*. MIT Press.

Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics, 54*(1–3), 159–178.

Lizardo, R. A., & Mollick, A. V. (2010). Oil price fluctuations and U.S. dollar exchange rates. *Energy Economics, 32*(2), 399–408.

MacKinnon, J. G. (1996). Numerical distribution functions for unit root and cointegration tests. *Journal of Applied Econometrics, 11*(6), 601–618.

Meese, R. A., & Rogoff, K. (1983). Empirical exchange rate models of the seventies: Do they fit out of sample? *Journal of International Economics, 14*(1–2), 3–24.

Mokaddem Faradji, T. (2020). Etude économétrique de l'effet du PIB par habitant sur le chômage en Algérie par la modélisation ARDL, test de cointégration aux bornes, test de causalité au sens de Toda et Yamamoto. *Nour Journal of Economic Studies* [مجلة نور للدراسات الاقتصادية], *6*(11).

Newey, W. K., & West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica, 55*(3), 703–708.

Reboredo, J. C. (2012). Modelling oil price and exchange rate co-movements. *Journal of Policy Modeling, 34*(3), 419–440.

Rossi, B. (2013). Exchange rate predictability. *Journal of Economic Literature, 51*(4), 1063–1119.

Tank, A., Covert, I., Foti, N., Shojaie, A., & Fox, E. B. (2022). Neural Granger causality. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(8), 4267–4279.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B, 58*(1), 267–288.

Toda, H. Y., & Yamamoto, T. (1995). Statistical inference in vector autoregressions with possibly integrated processes. *Journal of Econometrics, 66*(1–2), 225–250.

U.S. Energy Information Administration. Crude oil prices: Brent – Europe [DCOILBRENTEU]. FRED, Federal Reserve Bank of St. Louis. https://fred.stlouisfed.org/series/DCOILBRENTEU

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica, 48*(4), 817–838.

Zhang, Y.-J., Fan, Y., Tsai, H.-T., & Wei, Y.-M. (2008). Spillover effect of US dollar exchange rate on oil prices. *Journal of Policy Modeling, 30*(6), 973–991.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B, 67*(2), 301–320.