Received: 11 July 2026 | Revised: 14 September 2026 | Accepted: 23 September 2026 | Published online: 4 November 2026

RESEARCH ARTICLE

Journal of Data Science and Intelligent Systems
2026, Vol. 00(00) 1–13
DOI: 10.47852/bonviewJDSIS620211353

BON VIEW PUBLISHING

Fair Walk-Forward Forecasting of ICE Robusta Futures: Comparing Statistical, Machine-Learning, Deep-Learning, and Hybrid Models

Phung Duy Quang[1,*], Cao Ha Hai Dang[2], Nguyen Huu Thinh[1], Hoang Nam Quyen[3], Hoang Tuan Minh[3], Van Thi Thuan[4], Bui Thi Tam[5], Nguyen Viet Ha[3] and Tran Quang Thanh[6]

[1] *Faculty of Technology and Data Science, Foreign Trade University, Vietnam*

[2] *Faculty of Data Science and Artificial Intelligence, National Economics University, Vietnam*

[3] *School of Economics and International Business, Foreign Trade University, Vietnam*

[4] *Faculty of International Economics, Foreign Trade University, Vietnam*

[5] *Faculty of Business Administration, Foreign Trade University, Vietnam*

[6] *School of Advanced Education Programs, National Economics University, Vietnam*

Abstract: Distribution shifts can degrade nonlinear forecasting models trained on historical price levels, particularly relative to models that retain the most recently observed level. This study empirically compares 14 statistical, machine-learning, deep-learning, and Autoregressive Integrated Moving Average (ARIMA)–recurrent hybrid specifications for the daily ICE Futures Europe Robusta Coffee benchmark, rather than proposing a new forecasting algorithm. The synchronized dataset comprises 4,603 observations from 14 January 2008 to 28 January 2026. Specifications are selected using a chronological pure-training/validation split and evaluated once on a 300-observation blind test set under a common one-step-ahead walk-forward protocol. The blind-test price distribution shifts sharply from gross training (Kolmogorov–Smirnov statistic 0.955; standardized Wasserstein distance 4.274). Delta-Random Forest achieves the lowest mean absolute error (MAE) at 83.59 USD per metric ton, whereas Khashei–Bijari ARIMA–Long Short-Term Memory (LSTM) yields the lowest root mean squared error at 111.99. However, Delta-Random Forest improves MAE over Naive by only 0.27 USD per metric ton, and no selected pairwise comparison remains significant after Holm adjustment of the Diebold–Mariano tests. Grouped permutation importance shows that permuting lagged first-difference history increases Delta-Random-Forest MAE by only 0.76, while Volume and the three macro-financial regressors contribute negligibly. These findings support a cautious interpretation: differencing improves nonlinear learners' robustness to level-range shifts but does not establish economically or statistically meaningful predictive information beyond the most recently observed price.

Keywords: Robusta Coffee futures, walk-forward forecasting, hybrid models, distribution shift, target representation

1. Introduction

Forecasts of agricultural commodity prices can support procurement, inventory, hedging, and marketing decisions, but the usefulness of a forecast depends on both its accuracy and the decision context. Commodity series combine persistence, abrupt changes, time-varying volatility, and external financial influences. Contemporary reviews consequently recommend comparing statistical, machine-learning (ML), deep-learning (DL), and hybrid approaches under explicit and reproducible evaluation protocols [1, 2]. Recent large-scale agricultural applications also report substantial heterogeneity across commodities and model classes [3].

Coffee-specific evidence is narrower. Neural networks have produced competitive one-day forecasts for several exchange-traded commodities [4], and extreme learning machines have been applied to Arabica and Robusta prices [5]. Related studies of grain futures emphasize the roles of preprocessing and

***Corresponding author:** Phung Duy Quang, Faculty of Technology and Data Science, Foreign Trade University, Vietnam. Email: quangpd@ftu.edu.vn

© The Author(s) 2026. Published by BON VIEW PUBLISHING PTE. LTD. This is an open access article under the CC BY License (https://creativecommons.org/licenses/by/4.0/).

structural breaks [6], while interpretable multivariate methods have been proposed for corn futures [7]. Decomposition hybrids [8], metaheuristic transformers [9], and multivariate graph models [10] broaden the candidate set. These studies show that sophisticated methods can be useful, but they do not imply that complexity will beat a well-specified no-change benchmark for a particular daily series.

The dependent variable in this study is the ICE Futures Europe Robusta Coffee futures benchmark, quoted in US dollars per metric ton. The same instrument is commonly labeled "London Coffee" by retail data services. ICE describes the contract as a global benchmark for physical Robusta pricing [11]. The empirical evidence concerns this international futures benchmark; it does not contain Vietnamese export prices, domestic basis, VND/USD exchange rates, logistics costs, quality differentials, or export volumes. The revised title and claims therefore remove the earlier implication of a Vietnam-specific quantitative analysis.

Two design issues motivate the analysis. First, model classes must share an information-availability boundary at every forecast origin, although their predictor sets may differ. The comparison therefore concerns model specifications with different inputs under a common temporal cutoff, rather than algorithms given identical predictors. Second, nonlinear models fitted to a bounded transformation of a nonstationary price level may have to extrapolate when the evaluation period lies outside the training range. A first-difference target avoids direct level extrapolation because the reconstructed forecast retains the latest observed price. That reconstruction may protect accuracy under a level shift even when the change model contributes little information beyond a no-change forecast.

The study asks three research questions. RQ1: Under a common chronological walk-forward protocol, do recurrent hybrids or standalone ML/DL models significantly outperform Naive and Autoregressive Integrated Moving Average (ARIMA) benchmarks? RQ2: How does forecasting the first difference rather than the price level affect standalone ML/DL accuracy during distribution shift, and how much evidence remains beyond the latest observed level? RQ3: Are model rankings stable across 1, 5, 10, and 20-observation horizons?

The contribution is empirical and protocol-oriented rather than methodological. No new forecasting algorithm is proposed. The study (i) applies a shared forecast-origin information-availability boundary to 14 representative specifications with model-specific predictor sets, (ii) treats price level versus first difference as an explicit experimental factor, (iii) separates numerical ranking from multiplicity-controlled predictive-accuracy evidence, and (iv) adds quantified distribution-shift, computation-time, and model-interpretability analyses. The central claim is deliberately limited: first differencing greatly reduces the failure of the tested nonlinear level models under this sample's range shift, but the leading Delta and hybrid forecasts do not significantly or economically outperform Naive and ARIMA.

2. Literature Review and Research Gap

2.1. Benchmarks and fair forecast evaluation

ARIMA remains a useful benchmark because differencing and a parsimonious autoregressive–moving-average structure can represent short-memory dynamics transparently. Automated information-criterion searches reduce discretionary order choice [12]. For highly persistent prices, the Naive forecast, defined as the last observed value, is difficult to beat at short horizons. Forecasting competitions repeatedly show that simple methods and tree ensembles remain important comparators [13, 14].

Temporal evaluation must emulate actual deployment. Randomly mixing past and future observations can estimate a different prediction task from genuine forecasting. This study therefore separates pure training, validation, and a blind test chronologically, fixes all choices before the blind test, and updates only with information available at each origin. Statistical inference is also required because a numerical winner need not have a reproducible loss advantage.

2.2. Nonlinear models and target representation

Random Forest (RF) captures nonlinearities and interactions by averaging decorrelated trees; support vector regression (SVR) learns a nonlinear mapping through an epsilon-insensitive kernel objective. Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) recurrent networks learn sequential representations through gated state updates. Their flexibility can be valuable, but reviews caution that performance depends on scale, preprocessing, validation, and benchmark selection [2, 15, 16].

Recent agricultural evidence is mixed. Deep recurrent networks perform strongly for some daily spot-price panels [3, 17], while multivariate attention models exploit exogenous information [18]. Hybrid and optimized systems have reported gains for volatile series [19–21]. Other work finds ARIMA and recurrent methods close, with computational cost becoming relevant [22]. Coffee and grain-futures studies likewise emphasize that preprocessing and persistence influence the apparent advantage of nonlinear methods [4–6, 23].

Target scale is therefore a substantive modeling choice. A price-level learner trained after MinMax scaling is exposed to a bounded target range. A tree cannot extrapolate beyond its terminal-leaf responses, and an Radial Basis Function Support Vector Regression (RBF-SVR) generally reverts toward its learned support away from training data. By contrast, a change model forecasts $\Delta y_t = y_t - y_(t-1)$ and reconstructs $\hat{y}_t = y_(t-1) + \Delta\hat{y}_t$. The reconstruction imports the last observed level mechanically. A large level-versus-Delta improvement must consequently be interpreted as protection against level-range shift, not automatically as evidence that $\Delta\hat{y}_t$ adds large predictive content beyond $y_(t-1)$.

2.3. Hybrid models and the closest empirical studies

Zhang's sequential hybrid adds a neural forecast of ARIMA residuals to the ARIMA forecast [24]. Khashei and Bijari allow the nonlinear component to use both residual history and the original series [25]. Recent agricultural hybrids extend this idea with recurrent networks, lag selection, decomposition, and attention [8, 18, 19, 21]. These systems are plausible when a linear filter leaves exploitable nonlinear structure, but residual autocorrelation alone does not guarantee that a particular neural correction generalizes.

Table 1 locates the present study relative to the closest recent empirical designs. The research gap is not the absence of ARIMA–LSTM, RF, or first differences. It is the limited evidence on how these established components rank when (a) they share a forecast-origin information-availability boundary while using model-specific predictors, (b) target representation is varied explicitly, (c) the test distribution moves outside the training price range, (d) Naive and ARIMA remain in the comparison, and (e) numerical differences are subjected to multiplicity-controlled forecast tests.

Table 1
Recent empirical studies and the present framework

Study	Market/data	Design	Remaining gap addressed here
Xu and Zhang [4]	Coffee and seven commodities	Neural one-day forecasts	No common 14-model target ablation with multiplicity-controlled inference
Deina et al. [5]	Arabica and Robusta prices	Extreme learning-machine comparison	Different data frequency and no blind range-shift analysis
Brignoli et al. [6]	Grain futures	LSTM versus econometric models across horizons	No coffee application or explicit level-versus-Delta reconstruction test
Wu et al. [7]	Corn futures and external information	Interpretable temporal-fusion framework	No comparison with Naive, ARIMA and recurrent hybrids under a shared forecast-origin cutoff
Pandit et al. [8]	Monthly oilseed prices	CEEMDAN–TDNN and benchmarks	Different horizon/frequency and no out-of-range target ablation
Ray et al. [19]	Volatile agricultural price series	RF-assisted ARIMA–LSTM	No symmetric level/Delta tuning or Holm-controlled loss tests
Present study	Daily ICE Robusta benchmark	14 locked models; shared cutoff with model-specific inputs; blind walk-forward; target ablation	Adds shift effect sizes, DM/Holm tests, runtime, and Delta-RF interpretation

3. Data and Methods

3.1. Instrument, sources, and reproducibility limits

The modeling table follows the trading dates in an Investing.com historical export labeled "London Coffee Futures." This label refers to the ICE Futures Europe Robusta Coffee contract described by ICE [11]. Last Price is the reported daily close in USD per metric ton; Volume is the reported daily trading volume. The external regressors are the Nominal Broad US Dollar Index (DTWEXBGS), Effective Federal Funds Rate (EFFR), and West Texas Intermediate (WTI) spot price at Cushing (DCOILWTICO), retrieved through Federal Reserve Economic Data (FRED). Dataset links and access dates are provided in the Data Availability Statement.

Investing.com is a third-party retail aggregator, not the exchange's official settlement database. The downloaded snapshot may differ from later exports if the provider corrects history, and the authors cannot independently verify every observation against proprietary ICE records. To make this limitation auditable, the frozen processed dataset, preprocessing notebook, checksums, and prediction vectors are included in the public reproducibility repository. ICE documentation is used only to identify the contract, quotation unit, and benchmark role; no claim is made that the retail snapshot is an official ICE settlement record.

The authors used the provider-supplied historical series as downloaded; they did not select individual delivery contracts, splice contract-level files, or apply any rollover price adjustment. Parsing and date alignment preserved the exported daily Last Price values. The CSV contains no delivery-month identifiers, rollover flags, or adjustment factors, so the contract selected on each date cannot be reconstructed from this snapshot alone.

Investing.com's support documentation states that its website generally switches to the next futures contract during the weekend preceding expiry. A separate support page distinguishes real-time Contract for Difference (CFD) quotes from daily historical exchange data and notes that their contract transitions can differ. These general statements do not establish a date-by-date contract selection, and the documentation reviewed also does not establish whether this snapshot was back-adjusted at contract transitions. Accordingly, no volume-based selection rule, exact historical roll calendar, or adjusted/unadjusted classification is asserted. The source and documentation links are provided in the Data Availability Statement.

The forecasting target is therefore the provider-reported historical series, not an author-constructed continuous contract. Potential discontinuities at contract transitions cannot be separated from market price movements using this export alone. This limits contract-level and trading-return interpretations, while the reported model comparisons remain conditional on the same frozen target series.

The preprocessing aligns the three FRED series to coffee trading dates, forward-fills macro-financial observations unavailable on a trading date, and removes remaining initial incomplete rows. The resulting table contains 4,603 unique, strictly increasing daily observations from January 14, 2008, to January 28, 2026. Blank and explicit-zero Volume entries are both encoded as zero in the available processed dataset. This ambiguity is retained as a limitation and assessed through a without-Volume sensitivity analysis.

3.2. Chronological split and distribution shift

The last 300 observations form a blind test. The earlier 4,303 observations are divided chronologically into 3,657 pure training and 646 validation observations. The partitioning of the datasets is shown in Table 2. All hyperparameters, epoch counts, and ARIMA order choices use only pure training and validation. Once specifications are locked, scalers and models are refitted on gross training; the blind test is then evaluated once. Each ML/DL input uses a 60-observation lookback.

Table 2
Locked chronological data split

Split	Start	End	Observations
Pure-train	2008-01-14	2022-04-28	3,657
Validation	2022-04-29	2024-11-19	646
Blind test	2024-11-20	2026-01-28	300

Figure 1 shows the price path and locked split boundaries. The blind test contains a sustained higher price level than most of the gross training.

Distribution shift is quantified for every modeling variable using a two-sample Kolmogorov–Smirnov (KS) statistic, the first Wasserstein distance, a Wasserstein distance standardized by the gross-training standard deviation, and Jensen–Shannon (JS) divergence from common 30-bin empirical histograms. Holm adjustment is applied to the six KS p-values. These are descriptive two-sample diagnostics; serial dependence means their p-values should not be interpreted as independent-observation causal tests. Effect sizes carry the primary interpretation. The resulting distribution-shift diagnostics for all six modeling variables are summarized in Table 3.

Figure 2 visualizes the standardized shift. The price-level KS statistic is 0.955, the standardized Wasserstein distance is 4.274, and the JS divergence is 0.876. Fifty-six blind-test prices and 19 USD-Index values fall outside their gross-training ranges. Delta_Price also shifts, but much less in distributional shape than Last Price (JS 0.245 versus 0.876).

3.3. Forecasting models and residual construction

The Naive forecast is $\hat{y}_t = y_(t - 1)$. The ARIMA model is written as ψ(B)(1 − B)^d y_t = θ(B)ε_t where B is the backshift operator, $\varphi(B)$ is the autoregressive polynomial, $\theta(B)$ is the moving-average polynomial, d is the differencing order, and ε_t is the innovation. An Akaike information criterion (AIC) search on pure training over p, $q \leq 3$ and $d \leq 1$ selects ARIMA(0, 1, 1) [12].

Level RF and RBF-SVR use 60 lags of Last Price, Volume, USD_Index, EFFR, and Crude_Oil to forecast y_t. Two-layer LSTM and GRU networks use the same inputs. Zhang hybrids forecast the ARIMA residual from lagged residuals and the four non-price covariates; the predicted residual is added to the ARIMA forecast [24]. Khashei–Bijari (KB) hybrids additionally receive lagged Last Price [25]. Delta-RF, Delta-SVR, Delta-LSTM, and Delta-GRU use 60 lags of Delta_Price and the four covariates to forecast Δy_t and reconstruct the price from $y_(t - 1)$.

Residual construction is strictly chronological. During specification selection, ARIMA is fitted on pure training. The neural training targets are the in-sample fitted residuals within pure training; validation residual inputs are prior sequential one-step forecast errors, revealed only after each validation outcome. For final estimation, ARIMA is refitted on gross training and the hybrid networks are trained on its gross-training in-sample residuals. During blind-test walk-forward evaluation, the input at time t may contain forecast errors from earlier blind-test dates because those outcomes are already observed, but it never contains the current or a future error. Thus, the hybrid target and update rules match operational one-step forecasting without using a blind-test outcome before its forecast.

3.4. Symmetric hyperparameter search and neural training

RF and SVR tuning now use the same candidate budget for price-level and Delta targets. The RF grid contains 12

Figure 1
ICE Robusta price and locked chronological regions. The blind test begins after the dominant historical level range and contains 56 prices outside gross-training support

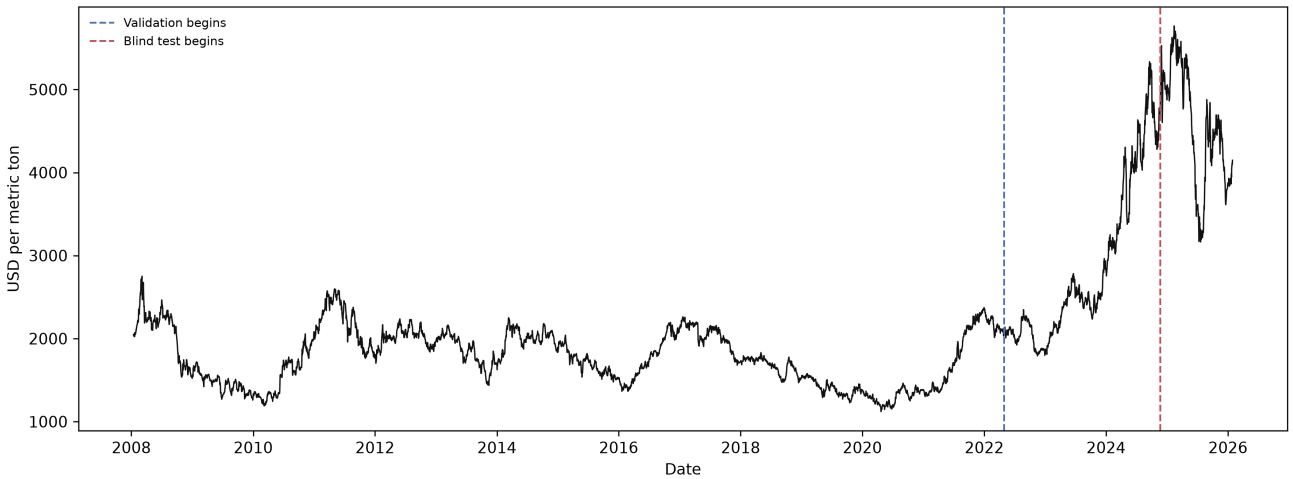
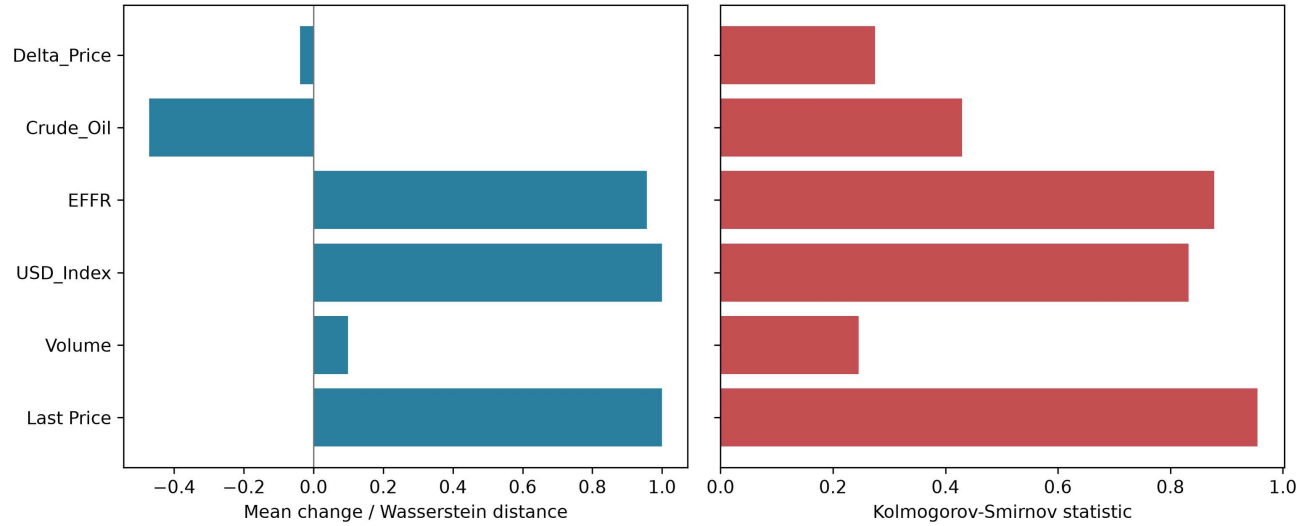


Table 3
Gross-training versus blind-test distribution shift

Feature	Train mean	Test mean	KS	Holm p	W/SD	JS	Outside
Last Price	1943.31	4574.63	0.955	<0.001	4.274	0.876	56
Volume	5694.32	5840.63	0.245	<0.001	0.317	0.143	0
USD_Index	106.06	122.96	0.832	<0.001	1.398	0.678	19
EFFR	1.17	4.21	0.878	<0.001	1.942	0.924	0
Crude_Oil	73.05	65.58	0.430	<0.001	0.697	0.478	0
Delta_Price	0.61	-1.69	0.275	<0.001	1.674	0.245	4

Note: W/SD is Wasserstein distance divided by the gross-training standard deviation; JS uses base-2 logarithms.

Figure 2
Quantified gross-training–blind-test distribution shift. The price level has the largest displacement; the effect sizes are descriptive and do not identify a causal event



configurations: 300 trees; maximum depth in {8, 12, unlimited}; minimum leaf size in {2, 3}; and maximum-feature rule in {square root, all features}. The SVR grid contains 18 configurations: C in {1, 10, 100}; ε in {0.01, 0.05}; and γ in {scale, 0.01, 0.1}. The sole selection criterion is validation mean absolute error (MAE) in reconstructed price units. The winning level RF uses depth 8, leaf 3, and all features; level SVR uses $C = 10$, $\varepsilon = 0.01$, and $\gamma = 0.01$. Delta-RF uses depth 8, leaf 3, and square-root feature subsampling; Delta-SVR uses $C = 1$, $\varepsilon = 0.01$, and $\gamma = 0.01$. The complete search tables are supplied with the code.

Each feature set has a separate MinMax scaler fitted only on the relevant training segment. All recurrent networks contain two LSTM or GRU layers of 50 units, each followed by dropout 0.20, and a one-unit dense output. Training uses Adam, Mean squared error (MSE) loss, batch size 32, no shuffling, a maximum of 100 epochs, and early stopping with patience 15 on validation loss. Locked epochs are 1 for level LSTM and GRU, 3 and 72 for Zhang LSTM and GRU, 12 and 11 for KB LSTM and GRU, and 36 and 2 for Delta-LSTM and Delta-GRU. Seed 42 defines the primary run.

3.5. Common one-step and multi-horizon protocols

Every primary blind-test forecast is one step ahead. At date t, all inputs end at $t - 1$. ARIMA produces the next forecast, after which the realized y_t updates the state-space filter without parameter re-estimation. ML/DL sequences use their respective predictor histories ending at that cutoff. For Delta models, $y_(t - 1)$ is added only after $\Delta\hat{y}_t$ is produced. This design holds the information-availability boundary constant across model classes, not the predictor sets. In particular, KB hybrids include lagged prices in the nonlinear component, whereas Zhang hybrids do not; Table 4 specifies the inputs for each model family.

Multi-horizon forecasts are generated recursively from 281 common rolling origins for h in {1, 5, 10, 20}. Within an origin, ARIMA produces a 20-step path. The KB hybrid feeds each predicted residual and predicted price back into its next input window; Delta-RF feeds each predicted difference back and accumulates it from the last known price. Future Volume and macro-financial variables are frozen at their origin values. The entire path is generated before any future actual is used. After

Table 4
Model inputs, symmetric search, and locked specifications

Family	Target/inputs	Locked specification
Naive	Price level	Last observed price
ARIMA	Price level	ARIMA(0, 1, 1), pure-training AIC
Level RF	60 lags of price + 4 covariates	300 trees; depth 8; leaf 3; all features
Level SVR	Same level inputs	RBF; $C = 10$; epsilon = 0.01; gamma = 0.01
Level LSTM/GRU	Same level inputs	2 × 50 units; dropout 0.20; epochs 1/1
Zhang hybrids	ARIMA residual + 4 covariates	LSTM/GRU epochs 3/72; residual added to ARIMA
KB hybrids	Zhang inputs + lagged price	LSTM/GRU epochs 12/11; residual added to ARIMA
Delta-RF	60 lags of Delta + 4 covariates	300 trees; depth 8; leaf 3; sqrt features
Delta-SVR	Same Delta inputs	RBF; $C = 1$; epsilon = 0.01; gamma = 0.01
Delta-LSTM/GRU	Same Delta inputs	2 × 50 units; dropout 0.20; epochs 36/2

Note: RF uses 12 candidates and SVR 18 candidates for each target representation; validation MAE in reconstructed price units is the selection criterion.

scoring that path, the rolling origin advances by one observation and only the newly available actual is assimilated. This is a controlled conditional scenario, not a claim that future covariates are known.

3.6. Evaluation, inference, structural diagnostics, and interpretability

Accuracy is summarized by MAE, root mean squared error (RMSE), mean absolute percentage error (MAPE), and R^2. MAE is the primary ranking measure because it is in USD per metric ton and is less dominated by extreme errors than RMSE. Diebold–Mariano (DM) tests compare loss differential $d_t = L(e_A, t) - L(e_B, t)$; negative means favor Model A [26]. A Newey–West-type long-run variance accounts for serial dependence [27], and Holm correction controls the family-wise error rate over 16 prespecified absolute- and squared-error comparisons [28]. Recent forecast-evaluation guidance likewise emphasizes suitable benchmarks, time-aware evaluation and multiplicity-aware statistical testing [26]. Failure to reject equal expected loss is not evidence of equivalence.

Pruned Exact Linear Time (PELT) change candidates are detected from the gross-training log-price series using an RBF cost, minimum segment length 90, and penalty 3 log(n). PELT is an optimization-based multiple-changepoint procedure whose selected candidates depend on the penalty specification [29]. Figure 3 displays their dates against the price level only for interpretability. The variable used by the algorithm is therefore log price throughout; the plot's vertical scale does not change the detection input.

Delta-RF interpretability is assessed in two complementary ways. First, impurity importances are summed across all 60 lags of each base feature. Second, grouped permutation importance permutes the full 60-lag block of one feature across blind-test sequences, repeats this 20 times, and reports the change in reconstructed-price MAE. Grouped permutation answers whether that feature block improves predictive loss conditional on the other blocks; it is not a causal effect and may be diluted by correlated predictors.

Wall-clock training and inference times are recorded for the locked specifications on the same Apple M4 MacBook Air with a 10-core CPU and 24 GB memory under macOS 26.5.2. The environment uses Python 3.13.7, NumPy 2.5.1, pandas 3.0.3, SciPy 1.18.0, scikit-learn 1.9.0, statsmodels 0.14.6, TensorFlow 2.21.0, ruptures 1.1.10, and Matplotlib 3.11.0. Timings are single-run descriptive measurements, exclude common data loading and sequence construction, and should not be treated as hardware-independent benchmarks. Hybrid timings include the shared ARIMA component.

4. Results

4.1. Residual diagnostics and motivation for hybrids

Figures 4 and 5 report first-difference dependence and ARIMA residual diagnostics before the forecast ranking. Most first-difference Autocorrelation Function (ACF)/Partial Autocorrelation Function (PACF) coefficients are small, whereas ARIMA residuals retain autocorrelation: Ljung–Box p-values are below 0.001 at lags 10 and 20, and rolling residual dispersion rises near the end of gross training. This motivates testing nonlinear corrections, but it does not show that a recurrent residual model will generalize.

4.2. Blind-test performance

Figure 6 overlays the actual blind-test path with Naive, ARIMA, KB ARIMA–LSTM, and Delta-RF. The first three forecast paths are visually close. Delta-RF likewise follows the prior observed level closely because its price reconstruction begins from that level.

Table 5 reports the locked 14-model comparison after the symmetric RF/SVR search. Delta-RF has the lowest MAE (83.59) and MAPE (1.86%), while KB ARIMA–LSTM has the lowest RMSE (111.99) and highest R^2 (0.974). Delta-LSTM, Naive, and ARIMA are within 0.28 USD per metric ton of the MAE leader. Delta-RF improves MAE over Naive by 0.27 (0.32%) and over ARIMA by 0.28 (0.34%). The improvement over Naive is only 0.006% of the blind-test mean price of 4,574.63 USD per metric ton, so the numerical ranking alone has negligible economic scale.

Figure 7 makes two points. First, the four-level neural/ML specifications remain much weaker during the higher-price blind test even after RF/SVR tuning. Second, the 10 leading Delta, hybrid, and benchmark specifications occupy a narrow loss band.

Figure 3
PELT candidates detected from gross-training log price and displayed on the price scale. The algorithm is fitted to log price; the vertical scale is used only to show candidate dates in familiar units

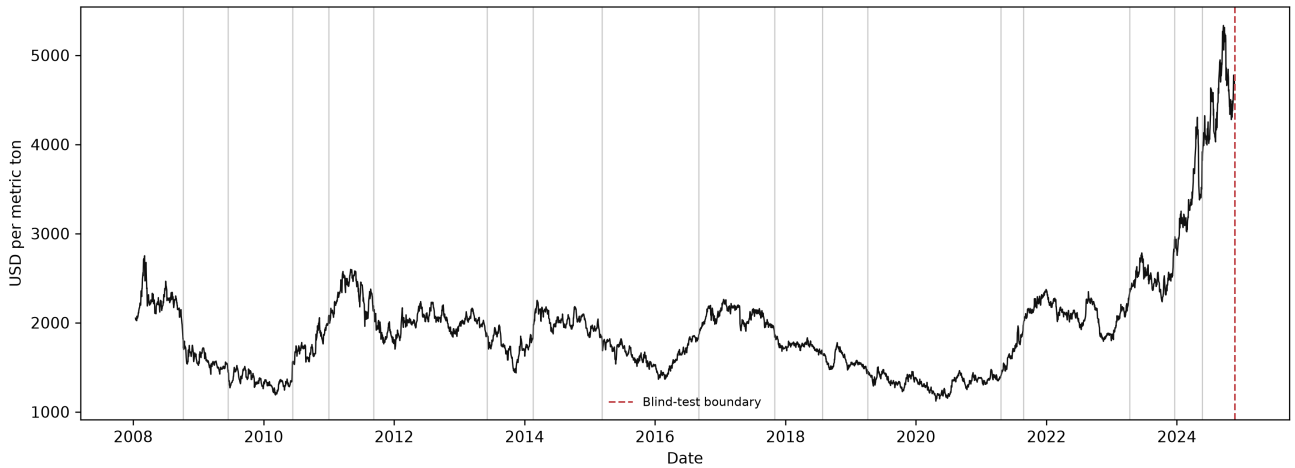


Figure 4

Price-level autocorrelation and first-difference partial autocorrelation. Persistence is strong in levels, whereas differenced dependence is comparatively compact

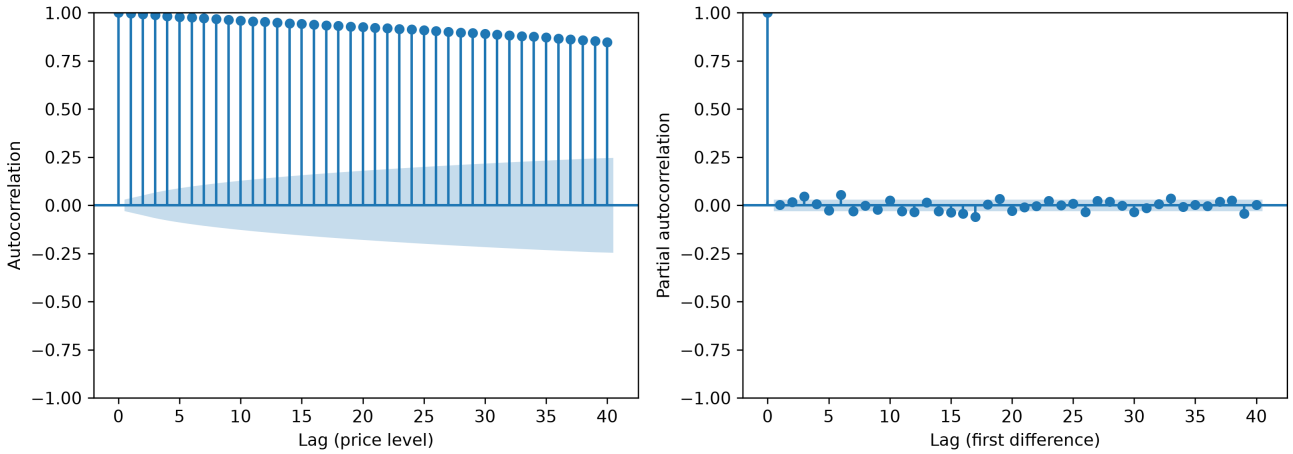


Figure 5

ARIMA residual autocorrelation and rolling dispersion. Remaining dependence motivates testing nonlinear corrections but does not guarantee out-of-sample hybrid gains

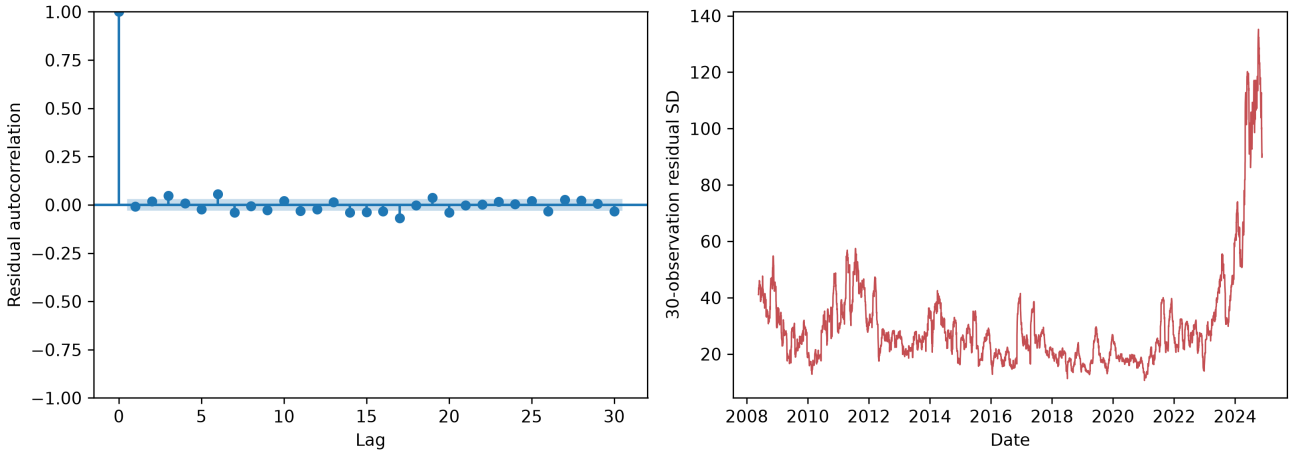


Figure 6

Observed prices and selected one-step walk-forward forecasts. The close overlap among benchmark and leading models is consistent with their small loss differences

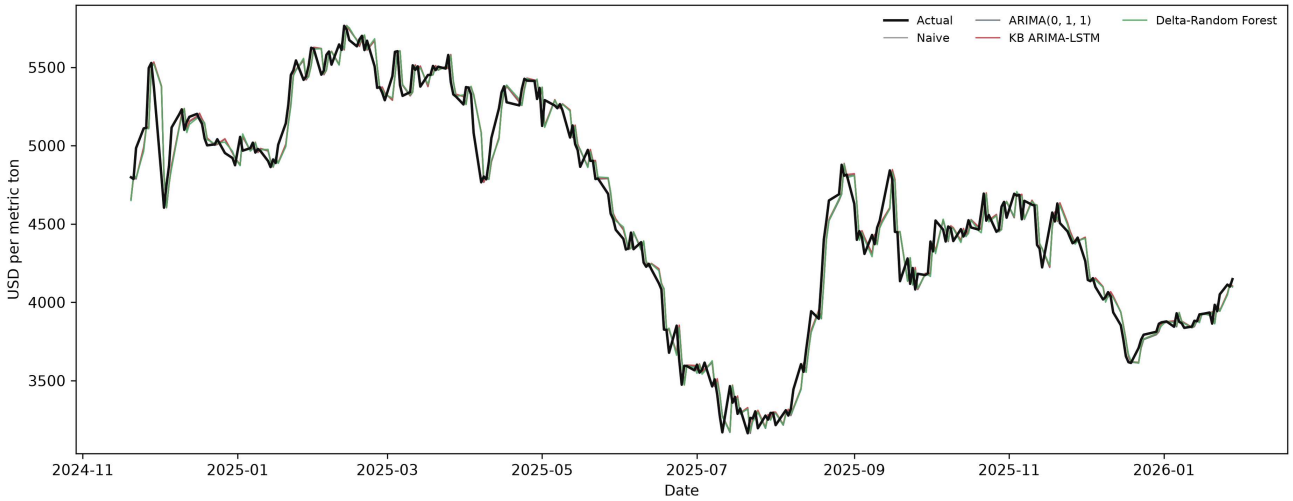
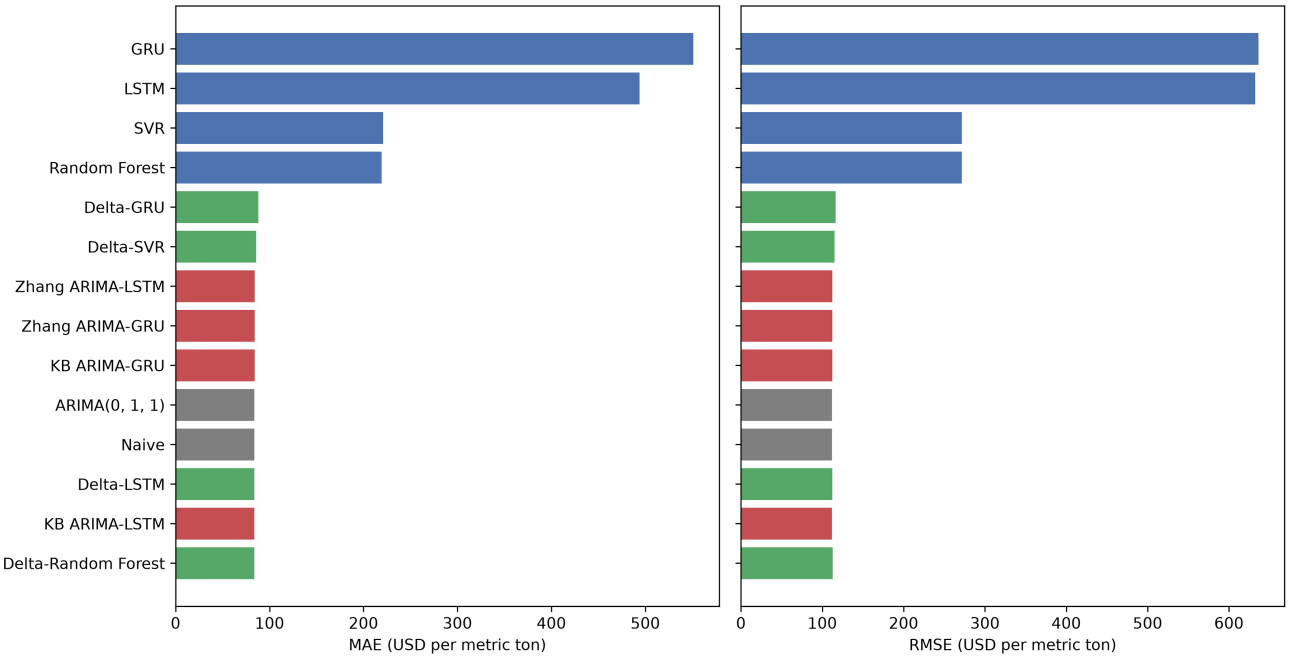


Table 5
Blind-test one-step performance (300 observations)

Rank	Model	MAE	RMSE	MAPE (%)	R^2
1	Delta-Random Forest	83.59	112.62	1.86	0.974
2	KB ARIMA–LSTM	83.76	111.99	1.87	0.974
3	Delta-LSTM	83.81	112.13	1.87	0.974
4	Naive	83.86	112.09	1.87	0.974
5	ARIMA(0, 1, 1)	83.87	112.08	1.87	0.974
6	KB ARIMA–GRU	84.06	112.29	1.87	0.974
7	Zhang ARIMA–GRU	84.23	112.45	1.88	0.974
8	Zhang ARIMA–LSTM	84.26	112.44	1.88	0.974
9	Delta-SVR	85.91	115.14	1.92	0.972
10	Delta-GRU	88.23	116.53	1.97	0.972
11	Random Forest	219.22	271.50	4.73	0.846
12	SVR	220.75	271.65	4.79	0.846
13	LSTM	493.67	632.03	12.40	0.168
14	GRU	551.02	636.43	13.44	0.156

Note: Lower MAE, RMSE, and MAPE are better. Numerical ranking does not imply a statistically significant loss difference.

Figure 7
Blind-test MAE and RMSE across all 14 specifications. Level ML/DL methods remain weak under range shift, while the leading Delta, hybrid, and benchmark models occupy a narrow loss band



4.3. Target representation and predictive-accuracy tests

Changing the target reduces MAE by 61.87% for RF, 61.08% for SVR, 83.02% for LSTM, and 83.99% for GRU relative to their tuned-level counterparts. These large within-algorithm reductions show that direct level prediction is fragile under the observed range shift. They do not show equivalent gains over Naive: the Delta reconstruction imports y_(t − 1), and Delta-RF beats Naive by only 0.27.

Table 6 confirms the distinction between a numerical winner and statistical evidence. Under absolute-error loss, all eight selected raw p-values exceed 0.59 and all Holm-adjusted p-values equal 1.000. Squared-error tests reach the same conclusion. Delta-RF, Delta-LSTM, and KB ARIMA–LSTM therefore show no statistically detectable advantage over Naive or ARIMA. The correct conclusion is "no detected difference," not equivalence and not proof that any benchmark is universally best.

4.4. Forecast horizon

Table 7 and Figure 8 extend evaluation to 281 rolling origins. Delta-RF has the lowest MAE at horizons 1, 10, and 20; Naive is lowest at horizon 5. KB ARIMA–LSTM has the lowest RMSE

Table 6
Selected Diebold–Mariano tests under absolute-error loss

Model A	Model B	Mean diff.	DM	Raw p	Holm p
Delta-Random Forest	Naive	–0.271	–0.516	0.606	1.000
Delta-Random Forest	ARIMA(0, 1, 1)	–0.283	–0.537	0.591	1.000
Delta-Random Forest	KB ARIMA–LSTM	–0.172	–0.275	0.783	1.000
Delta-LSTM	Naive	–0.054	–0.363	0.717	1.000
Delta-LSTM	ARIMA(0, 1, 1)	–0.066	–0.456	0.648	1.000
Delta-LSTM	KB ARIMA–LSTM	0.045	0.302	0.763	1.000
KB ARIMA–LSTM	Naive	–0.099	–0.389	0.698	1.000
KB ARIMA–LSTM	ARIMA(0, 1, 1)	–0.111	–0.446	0.656	1.000

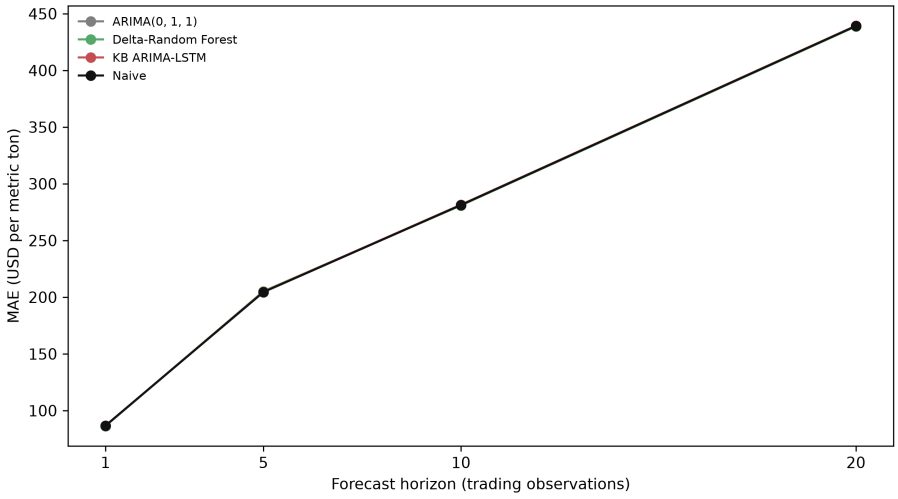
Note: Negative mean loss differences favor Model A. None is significant at 5% after Holm correction; squared-error tests give the same inferential conclusion.

Table 7
Recursive rolling-origin results by forecast horizon

h	Model	MAE	RMSE	MAPE (%)	R^2
1	Delta-Random Forest	86.42	115.51	1.92	0.972
1	KB ARIMA–LSTM	86.65	114.89	1.92	0.973
1	Naive	86.73	114.97	1.92	0.973
1	ARIMA(0, 1, 1)	86.74	114.96	1.93	0.973
5	Naive	204.54	271.05	4.53	0.850
5	ARIMA(0, 1, 1)	204.55	271.05	4.53	0.850
5	KB ARIMA–LSTM	204.79	270.92	4.54	0.850
5	Delta-Random Forest	205.24	272.34	4.54	0.848
10	Delta-Random Forest	280.92	364.66	6.33	0.728
10	Naive	281.26	365.53	6.35	0.727
10	ARIMA(0, 1, 1)	281.27	365.52	6.35	0.727
10	KB ARIMA–LSTM	281.57	365.52	6.36	0.727
20	Delta-Random Forest	438.78	551.29	10.18	0.380
20	Naive	439.15	554.78	10.21	0.372
20	ARIMA(0, 1, 1)	439.15	554.77	10.21	0.372
20	KB ARIMA–LSTM	439.40	554.76	10.22	0.372

Note: All models use 281 common origins. Future covariates are frozen at each origin; paths are generated recursively before the next actual is assimilated.

Figure 8
Recursive rolling-origin MAE by horizon. Ranking changes with horizon, and absolute error increases sharply; no method dominates across all horizons



at horizons 1 and 5. The Delta-RF MAE difference from Naive ranges from −0.37 to + 0.69 USD per metric ton, with the sign changing at $h = 5$. Forecast difficulty increases sharply: the best MAE rises from 86.42 at $h = 1$ to 438.78 at $h = 20$, while the best R^2 falls from approximately 0.973 to 0.380. There is no stable horizon-wide winner.

4.5. Interpretability and incremental information

Figure 9 and Table 8 show that Delta_Price accounts for 61.3% of grouped impurity importance. However, impurity importance can favor variables used frequently in tree splits. The blind-test grouped permutation result is more decision-relevant: permuting all 60 Delta_Price lags increases MAE by only 0.76 USD per metric ton on average (SD 0.36). EFFR and crude oil increase MAE by 0.04 and 0.03; USD_Index and Volume changes are centered near zero. Thus, the fitted Delta-RF relies mainly on lagged price changes, but their incremental predictive value beyond the last-level reconstruction is small.

4.6. Distribution shift, robustness, and the repeated Holm value

Figure 2 and Table 3 show that the largest standardized displacement occurs in Last Price; EFFR and USD_Index also change materially. These results explain why bounded level learners face an extrapolation problem, but they cannot identify which economic event caused the shift.

Five-seed checks leave the leading recurrent results stable: KB ARIMA–LSTM has mean MAE 83.82 (SD 0.10), and Delta-LSTM has mean 83.83 (SD 0.02). Removing Volume changes Delta-RF MAE from 83.59 to 83.99 and KB ARIMA–LSTM from 83.76 to 83.81. The two Holm-adjusted p-values are both 0.3181, but this is not a copying error. The raw p-values are 0.159052 and 0.288533. Holm first multiplies the smaller by two (0.318104) and the larger by one (0.288533) and then enforces nondecreasing step-down adjusted values; both therefore become 0.318104. The complete calculation is supplied in the reproducibility package.

4.7. Computational cost

Table 9 reports single-run wall-clock measurements and dominant complexity terms. Delta-RF trains in 0.84 s and generates all 300 blind-test predictions in 0.04 s. KB ARIMA–LSTM takes 28.14 s to train and 2.06 s for sequential test inference, including ARIMA. The 72-epoch Zhang ARIMA–GRU is the most expensive tested specification at 156.43 s. Because the leading loss differences are not significant, these costs favor Naive, ARIMA, or Delta-RF when transparency and maintenance matter more than a nominal metric rank.

Figure 9
Delta-Random-Forest grouped impurity and permutation importance. Lagged price changes dominate tree split use, but their mean blind-test permutation effect is only 0.76 USD per metric ton

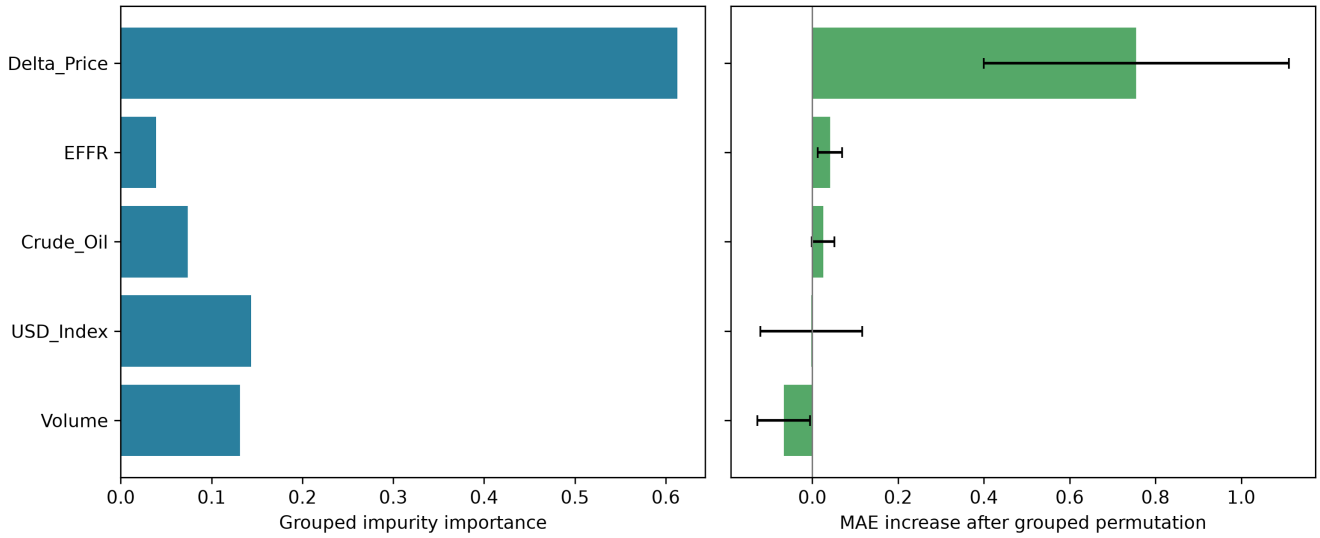


Table 8
Delta-Random-Forest grouped feature importance

Feature	Impurity (%)	Mean MAE +	SD	Min	Max
Delta_Price	61.3	0.755	0.356	–0.212	1.196
EFFR	3.9	0.041	0.029	–0.010	0.087
Crude_Oil	7.4	0.025	0.026	–0.020	0.078
USD_Index	14.4	–0.002	0.119	–0.240	0.225
Volume	13.1	–0.067	0.061	–0.194	0.051

Note: Grouped permutation uses 20 repeats and permutes all 60 lags of one base feature across blind-test sequences. Negative values indicate no incremental benefit in that permutation draw.

Table 9
Single-run computational cost of the locked specifications

Model	Training (*s*)	300-test inference (*s*)	Fitted size
Naive	0.00	0.00	0 parameters
ARIMA(0, 1, 1)	0.06	1.73	2 coefficients
Random Forest	13.45	0.03	300 trees; 90,804 nodes
SVR	0.49	0.03	876 support vectors
LSTM	2.77	0.27	31,451 neural parameters; 1 epochs
GRU	3.22	0.33	23,901 neural parameters; 1 epochs
Zhang ARIMA–LSTM	6.42	2.01	31,451 neural parameters; 3 epochs
Zhang ARIMA–GRU	156.43	2.20	23,901 neural parameters; 72 epochs
KB ARIMA–LSTM	28.14	2.06	31,651 neural parameters; 12 epochs
KB ARIMA–GRU	27.99	2.10	24,051 neural parameters; 11 epochs
Delta-Random Forest	0.84	0.04	300 trees; 16,560 nodes
Delta-SVR	1.13	0.13	3,335 support vectors
Delta-LSTM	79.46	0.34	31,451 neural parameters; 36 epochs
Delta-GRU	6.67	0.37	23,901 neural parameters; 2 epochs

Note: Dominant terms: RF $O(B\,N \log N$ m_try) training and $O(B \log N)$ inference; SVR between $O(N^2)$ and $O(N^3)$ training and O(Sd) inference; recurrent networks $O(E\,N\,W\,H(H + d))$ training; ARIMA uses iterative state-space likelihood and sequential state updates. Timings exclude common data loading and sequence construction.

5. Discussion

5.1. Complexity did not produce a robust advantage

The five closest methods—Delta-RF, KB ARIMA–LSTM, Delta-LSTM, Naive, and ARIMA—span only 0.28 USD per metric ton in MAE around a total error near 84. None differs significantly from the benchmarks. This is consistent with literature showing that model ranking depends on preprocessing, horizon, and series rather than architectural complexity alone [1, 6, 13, 14].

Residual autocorrelation and heteroscedasticity explain why a hybrid is worth testing, but not why it must win. A recurrent correction can fail when residual patterns are unstable, dominated by volatility rather than conditional mean, weak relative to noise, or different between training and deployment. Here, the ARIMA forecast already tracks the persistent level, and the neural correction is small. KB ARIMA–LSTM slightly lowers RMSE, suggesting fewer large misses, but its MAE improvement over ARIMA is only 0.11 and is not significant. The hybrid result is therefore marginal despite residual structure.

5.2. What the Delta result does and does not show

The level-versus-Delta contrast is large because the representations solve different extrapolation problems. Delta reconstruction carries forward the latest observed level, whereas the level RF, SVR, LSTM, and GRU must map scaled histories directly to a price outside much of their training support. Symmetric tuning improves level RF and especially level SVR but does not remove this structural disadvantage.

The revised interpretation is narrower than "target representation dominates architecture." The data show that differencing protects the tested nonlinear models from level-range shift. They do not show that Delta models contain substantial predictive information beyond the last observed price. Delta-RF's 0.27 MAE gain over Naive, nonsignificant DM tests, and 0.76 grouped permutation increment all support this caution. A fair ablation must therefore report both the level-versus-Delta improvement and the Delta-versus-Naive increment.

5.3. Practical and economic significance

For procurement, a benchmark-price forecast can inform short-horizon scenarios, timing discussions, and risk monitoring, but the relevant economic question is whether forecast improvements exceed basis risk, bid–ask spreads, transaction costs, and operational uncertainty. Here, a 0.27 USD-per-ton MAE reduction is approximately 0.006% of the blind-test mean price and is not statistically distinguishable from zero. It is too small to justify a more complex deployment on accuracy alone.

For hedging, Naive or ARIMA provides a transparent daily reference with minimal maintenance. Delta-RF is a reasonable low-cost nonlinear challenger when monitoring identifies level-range shift. KB ARIMA–LSTM may be considered when large errors, represented by RMSE, receive greater weight, but its added complexity is not supported by a significant loss improvement. For export planning, the ICE benchmark is only one component; local basis, VND/USD, quality, freight, contract maturity, inventory, and hedge ratios must be modeled separately. The present data do not estimate those channels.

5.4. Generalizability and limitations

The framework—chronological locking, shared forecast-origin information-availability boundaries, strong baselines, target ablation, shift diagnostics, and saved loss vectors—is transferable to other coffee contracts, agricultural commodities, and financial series. The fitted models and numerical rankings are not automatically transferable. Other markets differ in liquidity, units, seasonality, contract rolls, local basis, and covariate availability. External validity requires a fresh split, tuning, and inference analysis on each instrument.

The study has six main limitations. First, it models one benchmark series. Second, daily prices and Volume come from

a third-party snapshot rather than an independently verified exchange settlement dataset; the snapshot does not identify contract transitions or establish the provider's price-adjustment treatment. Third, blank and explicit-zero Volume observations share one code. Fourth, multi-horizon covariates are frozen rather than forecast jointly. Fifth, a 300-observation blind test has limited power for very small loss differences. Sixth, grouped feature importance is predictive and conditional, not causal. Probabilistic forecasts, transaction-cost-aware decisions, adaptive retraining, volatility models, and official multi-contract data remain for future work.

6. Conclusion

Under a common chronological evaluation protocol, Delta-RF achieves the lowest blind-test MAE, whereas KB ARIMA–LSTM yields the lowest RMSE. However, their numerical improvements over the Naive and ARIMA benchmarks are marginal, and none remains statistically significant after Holm adjustment. These results indicate that, over the blind-test period considered, increasing model complexity does not translate into a robust predictive advantage.

First-difference target formulations substantially reduce the forecast errors of the nonlinear level-based models under the pronounced level-range distribution shift observed in the test period. This finding, however, should be interpreted primarily as evidence of robustness rather than predictive superiority. Differencing preserves the most recently observed price level during reconstruction and thereby reduces the burden of extrapolating beyond the historical level range; nevertheless, the predicted price changes contain little measurable information beyond that already captured by the no-change benchmark. More broadly, the results emphasize that claims of architectural superiority in financial forecasting should be preceded by comparisons against strong baselines, explicit reconstruction rules, temporally valid evaluation protocols, effect-size-aware statistical inference, and reproducible prediction vectors.

The empirical evidence is specific to the international ICE Futures Europe Robusta Coffee benchmark and the evaluation period considered. Future research could extend the analysis by integrating official contract data with local basis and exchange-rate information, examining multiple coffee markets, retaining missingness indicators as potentially informative features, and evaluating probabilistic and adaptive forecasting approaches under changing market regimes.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The frozen processed analytical dataset, preprocessing and modeling code, environment specification, validation-search tables, saved blind-test prediction vectors, and revision-only analyses are available in the public reproducibility repository: https://github.com/hdanginneu/ice-robusta-walk-forward-forecasting.

Price and Volume were obtained from Investing.com, London Coffee Futures Historical Data (accessed January 28, 2026). The frozen analytical sample spans January 14, 2008, to January 28, 2026. Use of the provider's data remains subject to its terms.

The macro-financial series were retrieved through FRED, Federal Reserve Bank of St. Louis: Nominal Broad US Dollar Index (DTWEXBGS); EFFR; crude oil prices: WTI—Cushing, Oklahoma (DCOILWTICO) (all accessed March 1, 2026). These source records are publicly accessible.

Author Contribution Statement

Phung Duy Quang: Conceptualization, Writing – review & editing, Supervision, Project administration. **Cao Ha Hai Dang:** Methodology, Software, Formal analysis, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Nguyen Huu Thinh:** Conceptualization, Methodology, Validation, Writing – review & editing, Supervision. **Hoang Nam Quyen:** Investigation, Writing – original draft. **Hoang Tuan Minh:** Investigation, Data curation, Writing – review & editing. **Van Thi Thuan:** Investigation, Resources, Data curation, Writing – review & editing. **Bui Thi Tam:** Validation, Writing – review & editing, Visualization. **Nguyen Viet Ha:** Validation, Writing – review & editing. **Tran Quang Thanh:** Software, Validation, Writing – review & editing.

References

[1] Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Ben Taieb, S., . . . , & Ziel, F. (2022). Forecasting: Theory and practice. *International Journal of Forecasting*, *38*(3), 705–871. https://doi.org/10.1016/j.ijforecast.2021.11.001

[2] Zhang, C., Sjarif, N. N. A., & Ibrahim, R. (2024). Deep learning models for price forecasting of financial time series: A review of recent advancements: 2020–2022. *WIREs Data Mining and Knowledge Discovery*, *14*(1), e1519. https://doi.org/10.1002/widm.1519

[3] Manogna, R. L., Dharmaji, V., & Sarang, S. (2025). Enhancing agricultural commodity price forecasting with deep learning. *Scientific Reports*, *15*(1), 20903. https://doi.org/10.1038/s41598-025-05103-z

[4] Xu, X., & Zhang, Y. (2022). Commodity price forecasting via neural networks for coffee, corn, cotton, oats, soybeans, soybean oil, sugar, and wheat. *Intelligent Systems in Accounting, Finance and Management*, *29*(3), 169–181. https://doi.org/10.1002/isaf.1519

[5] Deina, C., Prates, do Amaral., H, M., Alves, C. H. R., Martins, M. S. R., Trojan, F., . . . , & Stevan, S. L., Jr. (2022). A methodology for coffee price forecasting based on extreme learning machines. *Information Processing in Agriculture*, *9*(4), 556–565. https://doi.org/10.1016/j.inpa.2021.07.003

[6] Brignoli, P. L., Varacca, A., Gardebroek, C., & Sckokai, P. (2024). Machine learning to predict grains futures prices. *Agricultural Economics*, *55*(3), 479–497. https://doi.org/10.1111/agec.12828

[7] Wu, B., Wang, Z., & Wang, L. (2024). Interpretable corn future price forecasting with multivariate time series. *Journal of Forecasting*, *43*(5), 1575–1594. https://doi.org/10.1002/for.3099

[8] Pandit, P., Sagar, A., Ghose, B., Paul, M., Kisi, O., Vishwakarma, D. K., . . . , & Yadav, K. K. (2024). Hybrid modeling approaches for agricultural commodity prices using CEEMDAN and time delay neural networks.

Scientific Reports, *14*(1), 26639. https://doi.org/10.1038/s41598-024-74503-4

[9] Nayak, G. H. H., Alam, M. W., Naik, B. S., Varshini, B. S., Avinash, G., Kumar, R. R., . . . , & Singh, K. N. (2025). Meta-transformer: Leveraging metaheuristic algorithms for agricultural commodity price forecasting. *Journal of Big Data*, *12*(1), 138. https://doi.org/10.1186/s40537-025-01196-5

[10] Min, Y., Kim, Y. R., Hyon, Y., Ha, T., Lee, S., Hyun, J., & Lee, M. R. (2025). RNN and GNN based prediction of agricultural prices with multivariate time series and its short-term fluctuations smoothing effect. *Scientific Reports*, *15*(1), 13681. https://doi.org/10.1038/s41598-025-97724-7

[11] Intercontinental Exchange. (n.d). *Robusta coffee futures: Contract specifications*. Retrieved from: https://www.ice.com/products/37089079/Robusta-Coffee-Futures/specs

[12] Sardar, I., Akbar, M. A., Leiva, V., Alsanad, A., & Mishra, P. (2023). Machine learning and automatic ARIMA/Prophet models-based forecasting of COVID-19: Methodology, evaluation, and case study in SAARC countries. *Stochastic Environmental Research and Risk Assessment*, *37*(1), 345–359. https://doi.org/10.1007/s00477-022-02307-x

[13] Januschowski, T., Wang, Y., Torkkola, K., Erkkilä, T., Hasson, H., & Gasthaus, J. (2022). Forecasting with trees. *International Journal of Forecasting*, *38*(4), 1473–1481. https://doi.org/10.1016/j.ijforecast.2021.10.004

[14] Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2022). M5 accuracy competition: Results, findings, and conclusions. *International Journal of Forecasting*, *38*(4), 1346–1364. https://doi.org/10.1016/j.ijforecast.2021.11.013

[15] Masini, R. P., Medeiros, M. C., & Mendes, E. F. (2023). Machine learning advances for time series forecasting. *Journal of Economic Surveys*, *37*(1), 76–111. https://doi.org/10.1111/joes.12429

[16] Benidis, K., Rangapuram, S. S., Flunkert, V., Wang, Y., Maddix, D., Turkmen, C., . . . , & Januschowski, T. (2022). Deep learning for time series forecasting: Tutorial and literature survey. *ACM Computing Surveys*, *55*(6), 1–36. https://doi.org/10.1145/3533382

[17] Manogna, R. L., & Mishra, A. K. (2021). Forecasting spot prices of agricultural commodities in India: Application of deep-learning models. *Intelligent Systems in Accounting, Finance and Management*, *28*(1), 72–83. https://doi.org/10.1002/isaf.1487

[18] Gu, Y. H., Jin, D., Yin, H., Zheng, R., Piao, X., & Yoo, S. J. (2022). Forecasting agricultural commodity prices using dual input attention LSTM. *Agriculture*, *12*(2), 256. https://doi.org/10.3390/agriculture12020256

[19] Ray, S., Lama, A., Mishra, P., Biswas, T., Sankar Das, S., & Gurung, B. (2023). An ARIMA–LSTM model for predicting volatile agricultural price series with random forest technique. *Applied Soft Computing, 149, Part A*, 110939. https://doi.org/10.1016/j.asoc.2023.110939

[20] Sari, M., Duran, S., Kutlu, H., Guloglu, B., & Atik, Z. (2024). Various optimized machine learning techniques to predict agricultural commodity prices. *Neural Computing and Applications*, *36*(19), 11439–11459. https://doi.org/10.1007/s00521-024-09679-x

[21] Avinash, G., Ramasubramanian, V., Ray, M., Paul, R. K., Godara, S., Nayak, G. H. H., . . . , & Iquebal, M. A. (2024). Hidden Markov guided deep learning models for forecasting highly volatile agricultural commodity prices. *Applied Soft Computing*, *158*, 111557. https://doi.org/10.1016/j.asoc.2024.111557

[22] Menculini, L., Marini, A., Proietti, M., Garinei, A., Bozza, A., Moretti, C., & Marconi, M. (2021). Comparing Prophet and deep learning to ARIMA in forecasting wholesale food prices. *Forecasting*, *3*(3), 644–662. https://doi.org/10.3390/forecast3030040

[23] Xu, X., & Zhang, Y. (2021). Corn cash price forecasting with neural networks. *Computers and Electronics in Agriculture*, *184*, 106120. https://doi.org/10.1016/j.compag.2021.106120

[24] Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, *50*, 159–175. https://doi.org/10.1016/S0925-2312(01)00702-0

[25] Khashei, M., & Bijari, M. (2011). A novel hybridization of artificial neural networks and ARIMA models for time series forecasting. *Applied Soft Computing*, *11*(2), 2664–2675. https://doi.org/10.1016/j.asoc.2010.10.015

[26] Hewamalage, H., Ackermann, K., & Bergmeir, C. (2023). Forecast evaluation for data scientists: Common pitfalls and best practices. *Data Mining and Knowledge Discovery*, *37*(2), 788–832. https://doi.org/10.1007/s10618-022-00894-5

[27] Newey, W. K., & West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, *55*(3), 703–708. https://doi.org/10.2307/1913610

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

[29] Quinn, M., Chung, A., & Glass, K. (2022). Automated selection of changepoints using empirical P-values and trimming. *JAMIA Open*, *5*(4), ooac090. https://doi.org/10.1093/jamiaopen/ooac090

How to Cite: Quang, P. D., Hai Dang, C. H., Thinh, N. H., Quyen, H. N., Minh, H. T., Thuan, V. T., …, & Thanh, T. Q. (2026). Fair Walk-Forward Forecasting of ICE Robusta Futures: Comparing Statistical, Machine-Learning, Deep-Learning, and Hybrid Models. *Journal of Data Science and Intelligent Systems*. https://doi.org/10.47852/bonviewJDSIS620211353