

The Bidirectional Pressure Conservation Model: An Interpretable Pressure-Reservoir Hypothesis Tested Across Assets, Benchmarks, Ablations, and Walk-Forward Folds

Akhilesh Varma

Independent Researcher

akhverm@gmail.com

September 9, 2026

Abstract

This paper introduces and evaluates the Bidirectional Pressure Conservation Model (BPCM), a compact nonlinear dynamical system for next-day financial direction scores. BPCM represents activity with two non-negative reservoirs: buying pressure and selling pressure. Each reservoir receives a bounded signed impulse from daily market and macroeconomic features, decays over time, loses mass when opposing pressure is simultaneously present, and contributes to a pressure-exhaustion mechanism. The model produces a score-derived probability and an abstaining buy-sell-abstain decision. We report a complete empirical study rather than a single favorable split. The model is trained on 12 assets through December 2021 and evaluated without test-asset fine-tuning on 20 successful heterogeneous assets from January 2022 through September 1, 2026, yielding 24,585 observations. We compare BPCM with a constant probability, logistic regression, random forest, and histogram gradient boosting; remove or vary individual pressure mechanisms; refit in annual walk-forward folds; and sweep decision thresholds and uncertainty limits. The pooled BPCM accuracy is 51.6%, ROC-AUC is 0.506, Brier score is 0.286, and log loss is 0.777. Its average predicted probability is 0.670, while the constant 0.5 forecast has Brier score 0.250 and log loss $\log 2 = 0.693$. The descriptive 95% bootstrap interval for pooled accuracy is [51.0%, 52.3%]. Annual walk-forward accuracy ranges from 48.5% to 52.9%. The results do not establish a reliable predictive or economic edge. The paper’s contribution is a precise, reproducible, falsifiable pressure-reservoir hypothesis and a transparent negative-result evaluation, not a profitable trading strategy or a claim of novelty against established market-microstructure, filtering, volatility, and forecast-evaluation literatures.

Keywords: financial time series; nonlinear dynamical systems; pressure reservoirs; order-flow intuition; cross-asset generalization; selective prediction; proper scoring rules; reproducibility; negative evidence.

1 Introduction

Financial return prediction is a setting in which a mathematically attractive mechanism can be mistaken for a validated empirical result. Daily directional labels contain substantial noise. Asset classes have different trading hours, liquidity, calendars, and volatility. Historical datasets are revised. Flexible model selection can exploit chance. A reported conditional accuracy can hide low coverage. A score that is numerically called a probability can be poorly calibrated. These problems make experimental protocol part of the scientific contribution rather than an afterthought.

This paper studies the Bidirectional Pressure Conservation Model (BPCM). The model begins with a simple hypothesis: information favoring higher and lower prices can accumulate in two competing stores of pressure. Pressure persists but decays, simultaneous opposing pressure cancels, and persistent total pressure exhausts itself. A directional forecast should depend on the imbalance between the stores, while simultaneous activity on both sides should reduce confidence. The construction is intentionally interpretable. It exposes buying pressure, selling pressure, net pressure, total pressure, conflict, exhaustion, and the final score.

The paper is a substantial revision of an earlier exploratory release. The earlier release used one chronological split, seven test assets, limited baselines, and a small sensitivity grid. The present release adds an expanded universe, a cached data pipeline, matched machine-learning benchmarks, targeted mechanism ablations, annual walk-forward refits, descriptive bootstrap intervals, complete per-asset tables, sensitivity outputs, code hashes, skipped-symbol records, and five paper figures generated from the stored results.

The research question is narrow: *does this particular pressure-reservoir mechanism provide transferable information about next-day direction on assets not used to fit its impulse map?* The answer in the present data is negative or, at minimum, inconclusive. BPCM’s ranking metrics remain close to chance. Its probability scores are worse than a constant forecast. Its selective policy does not reveal a stable large edge.

The contributions are fourfold. First, the paper gives an explicit dynamical definition that can be implemented without hidden state or undocumented tuning. Second, it positions the hypothesis against established work on order-flow imbalance, price impact, long-memory order signs, state-space filtering, volatility modeling, probabilistic scoring, and time-series validation. Third, it subjects the mechanism to a broader and more demanding empirical protocol. Fourth, it documents a negative result and makes the limitations visible rather than using interpretability as a substitute for evidence.

2 Related Work and Positioning

2.1 Market pressure, order flow, and price impact

The pressure vocabulary used by BPCM is an abstraction, not a claim that daily OHLCV features observe the order book. Market-microstructure research provides a more direct account of supply and demand. Cont, Kukanov, and Stoikov define order-flow imbalance at the best bid and ask and report a relationship between short-interval price changes and order-flow imbalance whose slope varies inversely with market depth [1]. Their study uses NYSE Trades and Quotes data for 50 U.S. stocks and argues that order-flow imbalance is more robust than volume alone for short-horizon price impact. Eisler, Bouchaud, and Kockelkoren study market orders, limit orders, and cancellations jointly and show that event types and order-book history affect subsequent price changes [2]. Lillo and Farmer find long memory in order signs but also show that liquidity and transaction-size effects can offset that persistence so that returns remain much less predictable [3].

These results motivate a careful interpretation of BPCM. Its two reservoirs may be viewed as a coarse dynamical analogy for competing directional forces, but they are not measured buy and sell queues. The current implementation contains no trades, quotes, signed order flow, depth, cancellation events, or intraday event time. Thus the model should not be described as an order-flow model. A meaningful future test would replace or augment daily technical features with event-level variables and compare the reservoirs directly with observable order-book states.

2.2 State-space models and nonlinear filtering

Kalman filtering is a foundational solution to linear filtering and prediction problems for dynamic systems [4]. The broader filtering literature includes nonlinear and approximate nonlinear filtering as established methodological domains [5]. Financial applications include state-space models for latent term-structure factors with measurement errors estimated through Kalman filtering [6]. BPCM therefore does not claim novelty for latent states, recursive updates, nonlinear filtering, or the general use of a dynamical system in finance. Its narrower candidate contribution is the particular pair of non-negative reservoirs, multiplicative opposing-pressure cancellation, exhaustion attenuation, and selective forecast rule. That candidate contribution must be judged by equations, ablations, and out-of-sample evidence rather than by terminology.

2.3 Volatility models and directional forecasts

GARCH models conditional variance as a function of past squared innovations and past conditional variances [7]. ARCH/GARCH methods are standard tools for modeling time-varying volatility [8]. Volatility forecasting and direction forecasting are related but different predictive tasks. Christoffersen et al. explicitly study direction-of-change forecasts and show that conditional variance, skewness, and kurtosis can contribute to sign probabilities, while not making volatility alone identical to directional classification [9]. This distinction motivates reporting both probability-quality metrics and directional metrics. A volatility benchmark can be useful even if its target is not the sign of the next return.

2.4 Forecast evaluation, calibration, and abstention

A probability score should be judged by a proper scoring rule when the goal is to reward honest probabilistic forecasts. Gneiting and Raftery define proper and strictly proper scoring rules and discuss the Brier and logarithmic scores, calibration, and sharpness [10]. Brier score measures squared probability error. Log loss penalizes confident wrong predictions and assigns an extreme penalty when an event is assigned probability zero. BPCM’s sigmoid score is therefore not accepted as a probability merely because it lies in $(0, 1)$.

ROC-AUC is a ranking statistic rather than a complete test of calibrated probability quality. Hand argues that AUC can be incoherent with respect to misclassification costs, particularly when ROC curves cross [11]. We therefore report AUC alongside accuracy, Brier score, and log loss. We do not infer that a model with a slightly higher accuracy has useful probabilities.

BPCM includes an abstention rule. Selective classification formalizes the trade-off between coverage and conditional risk by allowing uncertain observations to be rejected [12]. The current study reports coverage with selective accuracy and sweeps thresholds. It does not claim a formal finite-sample risk guarantee.

2.5 Data snooping and time-series validation

Out-of-sample design is especially important in finance. Sullivan, Timmermann, and White show how calendar-effect searches can produce nominally significant rules that fail after accounting for the search universe and out-of-sample behavior [13]. McLean and Pontiff report substantial post-publication reductions in documented cross-sectional return predictors [14]. Bailey et al. develop a probability-of-backtest-overfitting framework for the danger of selecting a strategy because it is the best among many tested alternatives [15]. These studies do not imply that all predictability is false. They do imply that a single favorable split is weak evidence.

Time-series validation has additional structure. Tashman reviews rolling origins, coefficient updating, window choice, and multiple test periods, and recommends rolling-origin approaches for improving evaluation reliability [16]. Bergmeir and Benítez discuss the theoretical issues with ordinary cross-validation under temporal dependence and describe blocked alternatives [17]. Bergmeir, Hyndman, and Koo analyze when K-fold cross-validation can nevertheless be valid for particular autoregressive settings [18]. The present study therefore uses a chronological primary split and annual walk-forward refits, while avoiding a claim that the protocol resolves all dependence or multiple-testing concerns.

3 Model Definition

3.1 Features and target

Let C_t be the close and V_t the volume of an asset on day t . The target is

$$y_{t+1} = \mathbf{1}\{C_{t+1} > C_t\}, \quad (1)$$

where a flat close is treated as zero. The feature vector has 12 components:

1. one-day close return;
2. three-day close return;
3. rolling 20-day standard deviation of one-day returns;
4. ratio of the five-day moving-average close to the current close;
5. 14-day relative-strength index;
6. percentage change in volume;
7. VIX level;
8. DXY level;
9. TNX level;
10. one-day percentage change in VIX;
11. one-day percentage change in DXY; and
12. one-day percentage change in TNX.

All rolling quantities are backward-looking. Macro series are aligned to business days and forward-filled after download. The feature rows with insufficient history or missing required variables are removed.

3.2 Impulse map

Let $x_t \in \mathbb{R}^{12}$ be the feature vector after standardization. The standardizer is fit only on the training universe. A logistic classifier is fit to the training labels. Its decision function is transformed into a bounded signed impulse:

$$u_t = \tanh\left(\frac{w^\top x_t + b}{s}\right), \quad -1 < u_t < 1, \quad (2)$$

where $s > 0$ is the pressure-scale parameter. Define positive and negative impulse magnitudes as $u_t^+ = \max(u_t, 0)$ and $u_t^- = \max(-u_t, 0)$.

The logistic classifier is not presented as the scientific contribution. It is a transparent feature-to-impulse map that makes the pressure dynamics testable. The benchmark section directly compares its probability output with logistic regression without the pressure dynamics.

3.3 Pressure reservoirs

BPCM stores two non-negative states, B_t and S_t :

$$B_{t+1} = [(1 - \delta_B)B_t + u_t^+ - \eta B_t S_t]_+, \quad (3)$$

$$S_{t+1} = [(1 - \delta_S)S_t + u_t^- - \eta B_t S_t]_+, \quad (4)$$

where $[z]_+ = \max(z, 0)$, $\delta_B, \delta_S \in [0, 1]$ are decay rates, and $\eta \geq 0$ is the opposing-pressure cancellation coefficient. The product $B_t S_t$ is the central interaction: it is small when one side dominates and large when both sides are active. The default parameters are $\delta_B = \delta_S = 0.035$ and $\eta = 0.045$.

Proposition 1 (Non-negativity). *If $B_0, S_0 \geq 0$, then $B_t, S_t \geq 0$ for all t .*

Proof. The positive-part operator makes each updated reservoir non-negative. Induction from the initial condition proves the result. $\square$

3.4 Conflict, exhaustion, and score

Define net and total pressure as

$$D_t = B_t - S_t, \quad M_t = B_t + S_t. \quad (5)$$

Conflict is

$$\mathcal{C}_t = \frac{2B_t S_t}{(B_t + S_t)^2 + \epsilon}, \quad (6)$$

where $\epsilon > 0$ avoids division by zero. Exhaustion is

$$X_t = 1 - \exp(-\xi M_t), \quad (7)$$

where $\xi = 0.018$ is the default exhaustion rate. The effective directional score is

$$q_t = \frac{D_t(1 - X_t)}{1 + \rho M_t}, \quad (8)$$

where $\rho = 0.20$ controls saturation. The positive-return probability score is

$$p_{t+1} = \sigma(q_t) = \frac{1}{1 + \exp(-q_t)}. \quad (9)$$

This output is a monotone transformation of a model score. Calibration is an empirical question.

Proposition 2 (Conflict bound). *For non-negative reservoirs, $0 \leq \mathcal{C}_t \leq 1$ up to the positive numerical regularizer.*

Proof. The arithmetic–geometric mean inequality gives $2B_t S_t \leq (B_t + S_t)^2$. The denominator is at least the numerator, so the ratio is bounded as stated. $\square$

3.5 Selective decision

The uncertainty score is a weighted combination of conflict and exhaustion:

$$U_t = 0.55\mathcal{C}_t + 0.45X_t. \quad (10)$$

For threshold $\tau = 0.55$ and uncertainty limit $u_{\max} = 0.62$, the action is

$$a_t = \begin{cases} +1, & p_{t+1} \geq \tau \text{ and } U_t \leq u_{\max}, \\ -1, & p_{t+1} \leq 1 - \tau \text{ and } U_t \leq u_{\max}, \\ 0, & \text{otherwise.} \end{cases} \quad (11)$$

Coverage is the fraction of observations with $a_t \neq 0$. Selective accuracy is the proportion of correct directional actions among nonzero actions. The zero action is abstention, not a missing prediction.

Table 1: Study universe and protocol.

Component	Definition
Training dates	2015-01-01 through 2021-12-31
Primary test dates	2022-01-01 through 2026-09-01
Training assets	SPY, QQQ, AAPL, MSFT, NVDA, AMD, GOOGL, META, XOM, UNH, JPM, WMT
Successful test assets	20 assets: equities, sector ETFs, EEM, TLT, GLD, SLV, USO, BTC-USD, ETH-USD, SOL-USD
Unavailable symbols	EURUSD=X, JPY=X, AUDUSD=X; recorded with reasons, not imputed
Pooled successful test rows	24,585
Primary model fit	One scaler and logistic impulse map fit on the training universe
State initialization	$B_0 = S_0 = 0$ at each asset boundary
Default decision	$\tau = 0.55$, $u_{\max} = 0.62$

4 Experimental Design

4.1 Data source and universe

The study uses daily data downloaded through the open-source **yfinance** interface to Yahoo Finance [19]. Yahoo Finance data are external and mutable. The repository records the date range, universe, script hash, successful assets, skipped symbols, and cached raw files. The cache makes the reported run reproducible conditional on the downloaded files.

The training universe contains 12 assets. The successful primary test universe contains TSLA, AMZN, BA, CAT, IBM, ORCL, NFLX, DIS, JPM, GS, XLE, XLF, EEM, TLT, GLD, SLV, USO, BTC-USD, ETH-USD, and SOL-USD, totaling 20 names exactly as listed in the experiment output. The three foreign-exchange symbols yielded no rows after feature construction and are transparently recorded in `results/skipped_assets.csv`. They are excluded from aggregate metrics, which use the observed row count and stored metadata rather than an assumed universe size.

4.2 Leakage controls

The target uses the next close and is not available when features are computed. Rolling features use observations at or before the prediction date. Standardization is fit on training observations only. The impulse classifier is fit on training observations only. The primary test assets are not used to fit the impulse map. At the beginning of each asset, the pressure reservoirs reset to zero, preventing state contamination across instruments. Walk-forward folds refit the scaler and impulse map using data available before each test year.

Forward-filling macro series can create stale covariates. The study does not claim release-timestamp correctness for every macro observation. It also does not correct for survivorship, delisting, corporate-action revision beyond the downloaded Yahoo Finance fields, or cross-asset calendar mismatch. These limitations are part of the interpretation.

4.3 Metrics

For probabilities p_i and binary outcomes y_i , the Brier loss is

$$\text{Brier} = \frac{1}{n} \sum_{i=1}^n (p_i - y_i)^2, \quad (12)$$

and log loss is

$$\text{LogLoss} = -\frac{1}{n} \sum_{i=1}^n [y_i \log p_i + (1 - y_i) \log(1 - p_i)]. \quad (13)$$

Accuracy is the fraction of $p_i \geq 0.5$ predictions with the correct label. ROC–AUC measures ranking and is not treated as a calibration metric. Coverage and selective accuracy are reported together. The unit-payoff diagnostic assigns +1 to a correct action, −1 to an incorrect action, and zero to abstention. It is not a return and excludes costs, slippage, sizing, leverage, market impact, and execution delay.

The study computes 1,000 observation-resampling bootstrap intervals for accuracy. These intervals are descriptive because observations are time-dependent. A block bootstrap, stationary bootstrap, or formal dependent-data test would be preferable for inferential claims.

Table 2: Macro-average performance across successful test assets. Lower Brier and log loss are better.

Model	Accuracy	AUC	Brier	Log loss
BPCM	0.5169	0.5025	0.2864	0.7790
Constant 0.5	0.5168	0.5000	0.2500	0.6931
Logistic regression	0.5089	0.5019	0.2511	0.6954
Random forest	0.4874	0.4942	0.2648	0.7239
Gradient boosting	0.4887	0.4995	0.2902	0.7884

4.4 Benchmarks

The constant benchmark assigns probability 0.5 to every observation. Logistic regression uses the same feature vector and training split but does not apply pressure dynamics. Random forest uses 300 trees, square-root feature subsampling, and minimum leaf size 20. Histogram gradient boosting uses 150 iterations, learning rate 0.05, at most 15 leaves, and L2 regularization 1.0. These hyperparameters are fixed in code and are not selected on the primary test set.

4.5 Ablations

The full model is compared with no cancellation ($\eta = 0$), no exhaustion ($\xi = 0$), no saturation ($\rho = 0$), short memory ($\delta_B = \delta_S = 0.15$), and long memory ($\delta_B = \delta_S = 0.01$). Each variant uses the same feature-to-impulse training protocol. The ablation is a mechanism test, not a hyperparameter search.

4.6 Walk-forward and sensitivity tests

The annual walk-forward experiment evaluates 2022, 2023, 2024, 2025, and 2026. Each fold fits through the prior year and evaluates the next calendar window. The sensitivity grid evaluates thresholds 0.52, 0.55, 0.60, 0.65, and 0.70 and uncertainty limits 0.25, 0.40, 0.60, 0.80, and 1.00. This grid is descriptive and should not be used to select a final test-set threshold.

5 Results

5.1 Aggregate benchmark comparison

Table 2 reports macro-average metrics across the 20 successful test assets. BPCM has the highest mean accuracy among the listed models by a small margin, but its probability scores are substantially worse than both the constant and logistic benchmarks. This is an important distinction: a thresholded directional score can be near or slightly above a baseline while its numerical probabilities are overconfident.

The pooled BPCM results are accuracy 0.5158, AUC 0.5059, Brier score 0.2857, log loss 0.7770, coverage 0.9262, and selective accuracy 0.5181. The descriptive pooled accuracy interval is $[0.5097, 0.5226]$. The pooled mean probability is 0.6703. The difference between macro and pooled values reflects unequal asset histories, especially the longer crypto series.

5.2 Complete BPCM per-asset results

Table 3 provides the complete per-asset result rather than reporting only a mean. Accuracy ranges from 0.4822 on DIS to 0.5492 on XLE. The AUC range is 0.4635–0.5373. The intervals overlap chance for every asset at the descriptive level. Several assets have high coverage because the score often crosses the decision threshold, but this does not translate into a reliable selective advantage.

Table 3: Complete BPCM test results by asset.

Asset	n	Acc.	AUC	Brier	Log loss	Coverage	Sel. acc.
TSLA	1149	.5048	.5097	.2654	.7275	.7311	.5155
AMZN	1149	.5135	.5373	.2757	.7509	.9739	.5147
BA	1149	.5065	.5069	.2884	.7794	.9774	.5040
CAT	1149	.5335	.5045	.2758	.7539	.9678	.5378
IBM	1149	.5448	.5285	.2772	.7604	.9478	.5482
ORCL	1149	.5231	.5270	.2729	.7473	.8834	.5340
NFLX	1149	.5039	.5002	.2775	.7549	.8634	.5030
DIS	1149	.4822	.5018	.3143	.8422	.9634	.4788

Table 4: Mean ablation results across successful test assets.

Variant	Accuracy	AUC	Brier	Log loss
Full BPCM	.5169	.5025	.2864	.7790
No cancellation	.5169	.5025	.2839	.7726
No exhaustion	.5169	.5026	.2884	.7846
No saturation	.5169	.5028	.3081	.8483
Short memory	.5170	.5023	.2534	.7001
Long memory	.5174	.5072	.3627	1.0447

Asset	n	Acc.	AUC	Brier	Log loss	Coverage	Sel. acc.
JPM	1149	.5440	.4791	.2829	.7726	.9991	.5444
GS	1149	.5309	.5067	.2814	.7676	.9904	.5316
XLE	1149	.5492	.5090	.2790	.7614	.9956	.5490
XLF	1149	.5178	.4829	.3066	.8291	.9983	.5179
EEM	1149	.5178	.5005	.3048	.8246	.9983	.5187
TLT	1149	.4883	.4947	.3365	.9045	.9991	.4878
GLD	1149	.5457	.4635	.2940	.8035	.9991	.5453
SLV	1149	.5117	.4645	.2928	.7940	.9286	.5155
USO	1149	.5213	.5110	.2803	.7626	.9434	.5258
BTC-USD	1684	.4911	.5028	.2866	.7744	.9038	.4915
ETH-USD	1684	.4929	.5073	.2680	.7320	.8498	.4997
SOL-USD	1684	.5160	.5130	.2688	.7370	.7417	.5132

5.3 Mechanism ablations

Table 4 shows that removing cancellation or exhaustion does not materially change accuracy or AUC. Removing saturation worsens probability quality. The short-memory variant has Brier 0.2534 and log loss 0.7001, close to the constant benchmark, but does not improve ranking. The long-memory variant has slightly higher accuracy at the displayed macro level but much worse probability quality, with log loss 1.0447. This is consistent with an overconfident score rather than a validated directional mechanism.

5.4 Walk-forward stability

The annual walk-forward experiment provides a more demanding view than the single primary split. Macro-average accuracy is 0.4852 in 2022, 0.4962 in 2023, 0.5292 in 2024, 0.5264 in 2025, and 0.5052 in 2026. AUC is 0.5060, 0.5064, 0.5298, 0.5336, and 0.5324 respectively. The relatively stronger 2024–2025 values are not persistent across all years and are not accompanied by a demonstrated calibration advantage. The 2022 and 2026 log losses are 0.8615 and 0.8772, showing that regime-specific accuracy is not enough to establish stable probability quality.

5.5 Threshold and uncertainty sensitivity

The threshold sweep does not reveal a strong selective effect. At threshold 0.52, mean selective accuracy is approximately 0.5183 with coverage between 0.8853 and 0.9779 depending on the uncertainty limit. At threshold 0.55, selective accuracy is 0.5188 at coverage 0.9328 for uncertainty limits at least 0.40. At threshold 0.60, mean selective accuracy reaches approximately 0.5211 at coverage 0.8415. At threshold 0.65, it reaches approximately 0.5240 at coverage 0.7119. At threshold 0.70, coverage falls to approximately 0.5265 and selective accuracy falls to 0.5159. Thus the grid displays the expected coverage–selectivity trade-off, but no threshold yields a large or stable advantage.

5.6 State trajectories

Figure 5 shows a representative state trajectory. Net pressure changes sign and total pressure varies over time. Conflict and exhaustion are observable diagnostic states, but they are deterministic transformations of the daily feature stream. They should not be interpreted as directly observed institutional buying, selling, liquidity, or order-book energy.

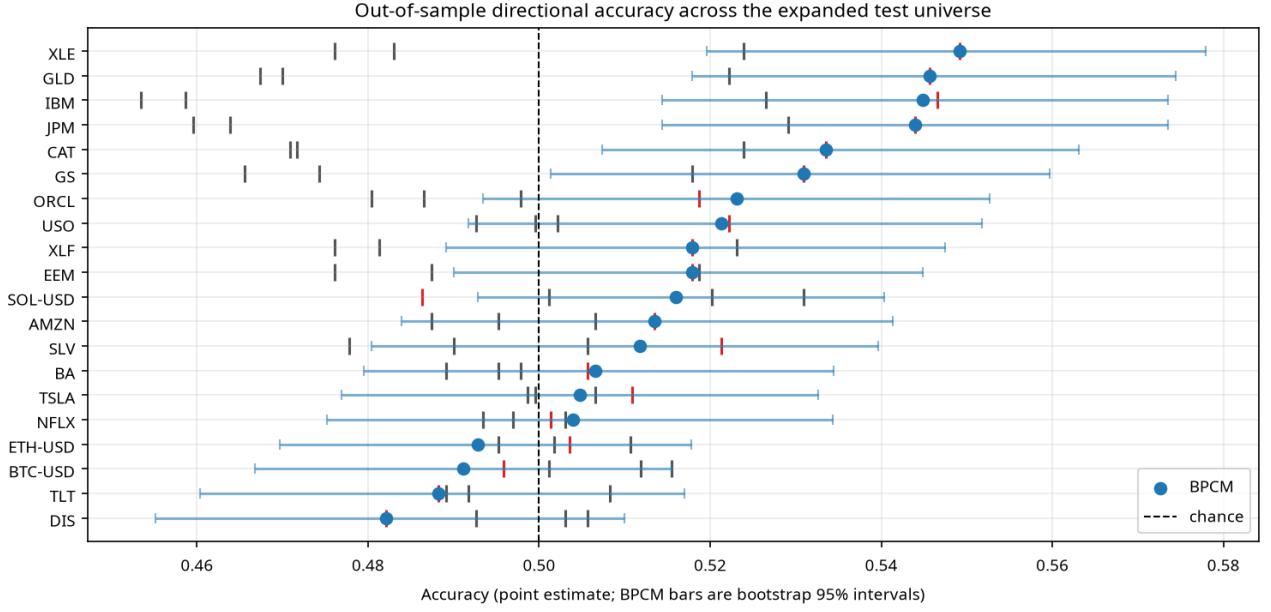


Figure 1: Asset-level directional accuracy. BPCM points include descriptive bootstrap 95% intervals. Vertical marks show matched benchmarks.

Table 5: Annual walk-forward macro-average results.

Test year	Accuracy	AUC	Brier	Log loss
2022	.4852	.5060	.3197	.8615
2023	.4962	.5064	.2713	.7415
2024	.5292	.5298	.2653	.7274
2025	.5264	.5336	.3136	.8591
2026	.5052	.5324	.3230	.8772

6 Discussion

6.1 What the study establishes

The study establishes that BPCM can be specified, implemented, and evaluated without hidden state across asset boundaries. It also establishes that the current daily-feature implementation produces predictions close to chance on the chosen cross-asset test universe. The result is not rescued by adding nonlinear benchmark classifiers. The mechanism ablations do not show a large ranking improvement attributable to cancellation, exhaustion, or saturation. Walk-forward results vary across years. Selective prediction changes coverage more readily than it changes accuracy.

The study does not establish that no market contains any pressure-like predictive structure. It does not establish that intraday order flow is uninformative. It does not establish that another impulse map, a calibrated training procedure, or a different target horizon would fail. It establishes only that this specified model, feature set, parameterization, and protocol do not support a strong claim in the current data.

6.2 Calibration and class-balance interpretation

The BPCM mean probability of approximately 0.67 is materially larger than 0.5 while the forced-choice accuracy remains close to 0.52. This indicates that the sigmoid transformation is not calibrated to the observed event frequency. The constant 0.5 forecast has lower Brier score and log loss even though BPCM's accuracy is marginally higher. This is why the paper does not describe BPCM as a probabilistic forecasting success. Any future deployment would require a calibration method fitted only inside training or validation data, followed by an untouched evaluation.

6.3 Economic interpretation

The unit-payoff directional score is a diagnostic, not a backtest return. The current study does not model bid-ask spread, commissions, market impact, turnover, slippage, execution delay, position sizing, leverage, financing,

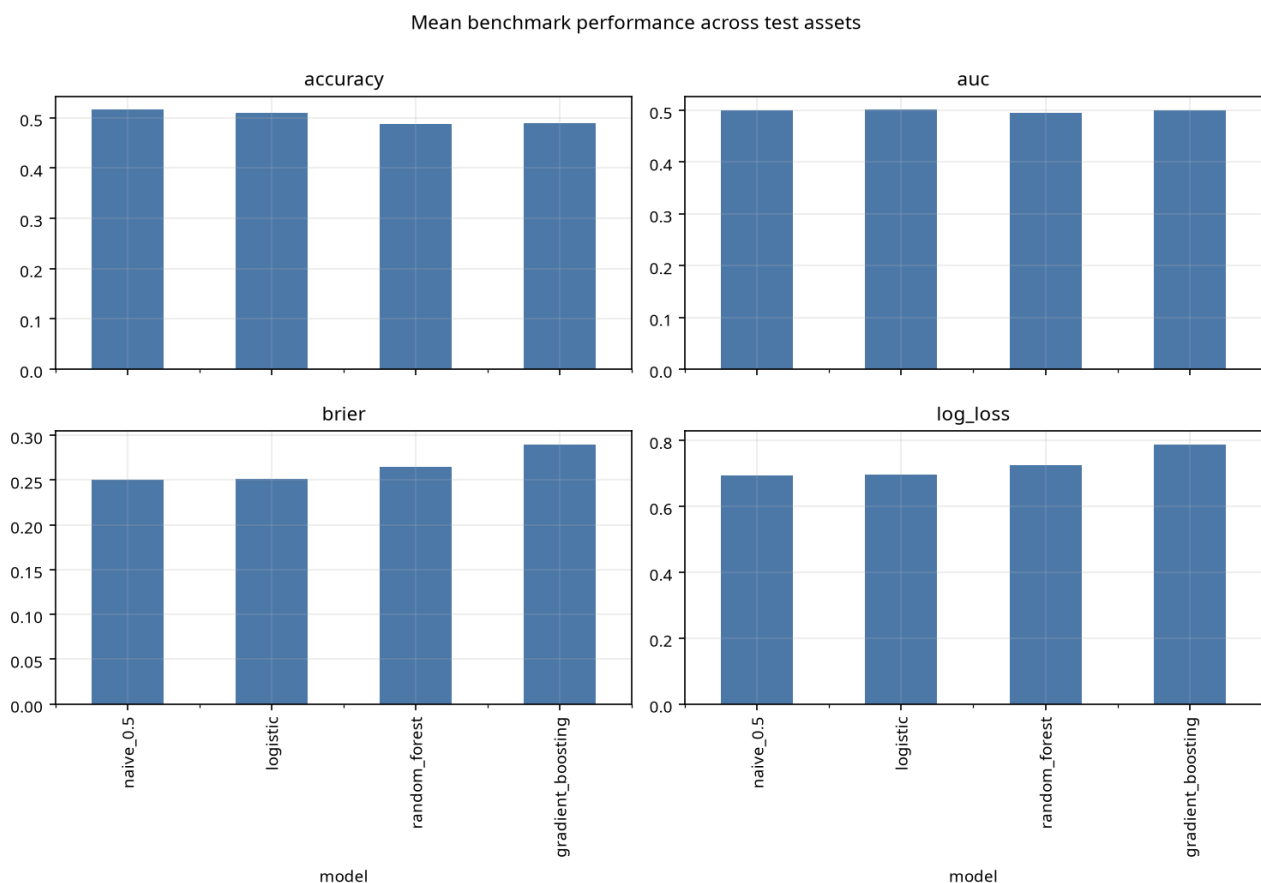


Figure 2: Mean benchmark accuracy, AUC, Brier score, and log loss across successful test assets.

Table 6: Sensitivity summary using uncertainty limit 0.40. The full grid is stored in `results/sensitivity.csv`.

Threshold	Coverage	Selective accuracy
.52	.9779	.5181
.55	.9328	.5188
.60	.8415	.5211
.65	.7119	.5240
.70	.5265	.5159

margin, shorting constraints, or risk limits. A strategy with near-chance direction can still have a different return distribution if it times large moves, but this release does not test that claim. Conversely, an apparent positive unit-payoff average can disappear after costs. No investment recommendation follows from these results.

6.4 Threats to validity

The data are obtained from a mutable vendor and are not a frozen archival release. The test universe is selected rather than a complete delisted-and-survivorship-corrected market panel. Three configured foreign-exchange symbols yielded no rows after feature construction. Macro variables are forward-filled and may be stale. Daily data discard intraday event ordering and order-book depth. The target is one-day close direction and does not represent a return magnitude, drawdown, or investor utility.

The walk-forward refits use the same training asset universe and do not perform a full nested validation search over hyperparameters. The descriptive bootstrap resamples observations independently and therefore understates uncertainty if dependence is strong. The benchmark and ablation families are prespecified in code, but any larger future search would require explicit multiple-testing control. These limitations favor the cautious negative interpretation.

6.5 Future experiments

A stronger follow-up should archive legally redistributable raw data with checksums and release timestamps. It should include delisted securities and a predefined universe-selection rule. It should test multiple horizons

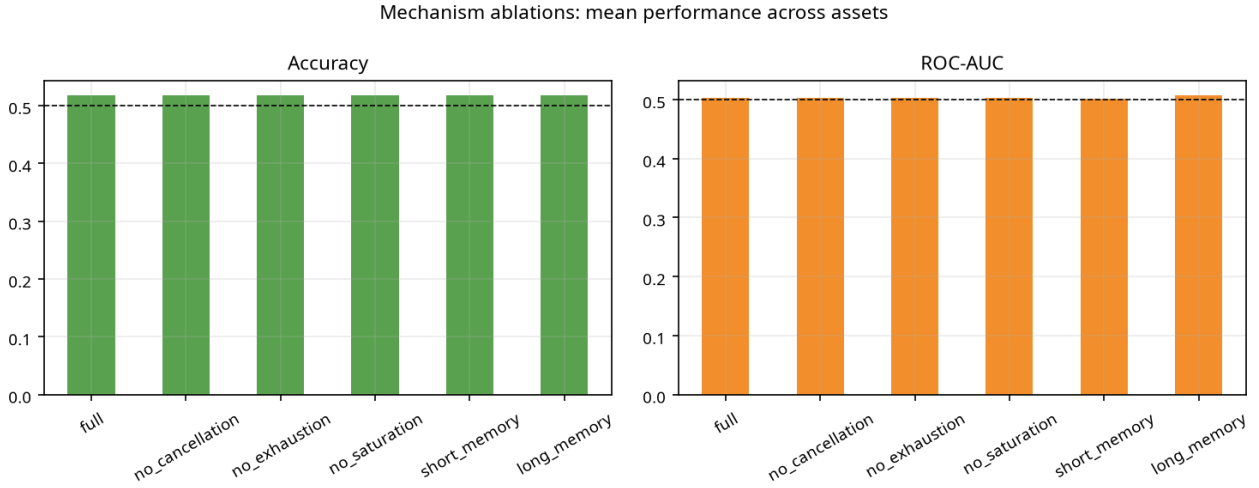


Figure 3: Mean accuracy and ROC-AUC for the full model and mechanism ablations. Dashed lines mark chance.

and return thresholds, including magnitude-conditioned targets. It should add intraday trades, quotes, depth, signed volume, cancellations, spreads, and liquidity. It should compare BPCM with calibrated generalized linear models, state-space models, GARCH-based sign forecasts, gradient boosting with nested validation, and carefully regularized neural models. It should use block-bootstrap or stationary-bootstrap intervals, formal risk-coverage curves, and an economic simulator with realistic costs and execution delay. A systematic bibliographic and patent search is also required before making any narrower novelty claim.

7 Conclusion

BPCM is a compact nonlinear dynamical system based on buying-pressure accumulation, selling-pressure accumulation, opposing-pressure cancellation, and exhaustion. This paper provides a deeper and more reproducible test than the original exploratory release. It expands the primary evaluation to 20 successful heterogeneous test assets and 24,585 observations, includes matched probability and machine-learning benchmarks, reports complete per-asset results, adds six mechanism variants, evaluates annual walk-forward stability, sweeps selective thresholds, and embeds the generated figures and tables in the PDF.

The empirical conclusion remains negative. BPCM’s pooled accuracy is 51.6%, pooled AUC is 0.506, and its Brier and log-loss scores are worse than a constant 0.5 forecast. Its walk-forward performance is not stable across calendar years. The responsible claim is therefore that BPCM is a transparent, falsifiable research hypothesis whose present daily-data implementation does not establish predictive or economic value.

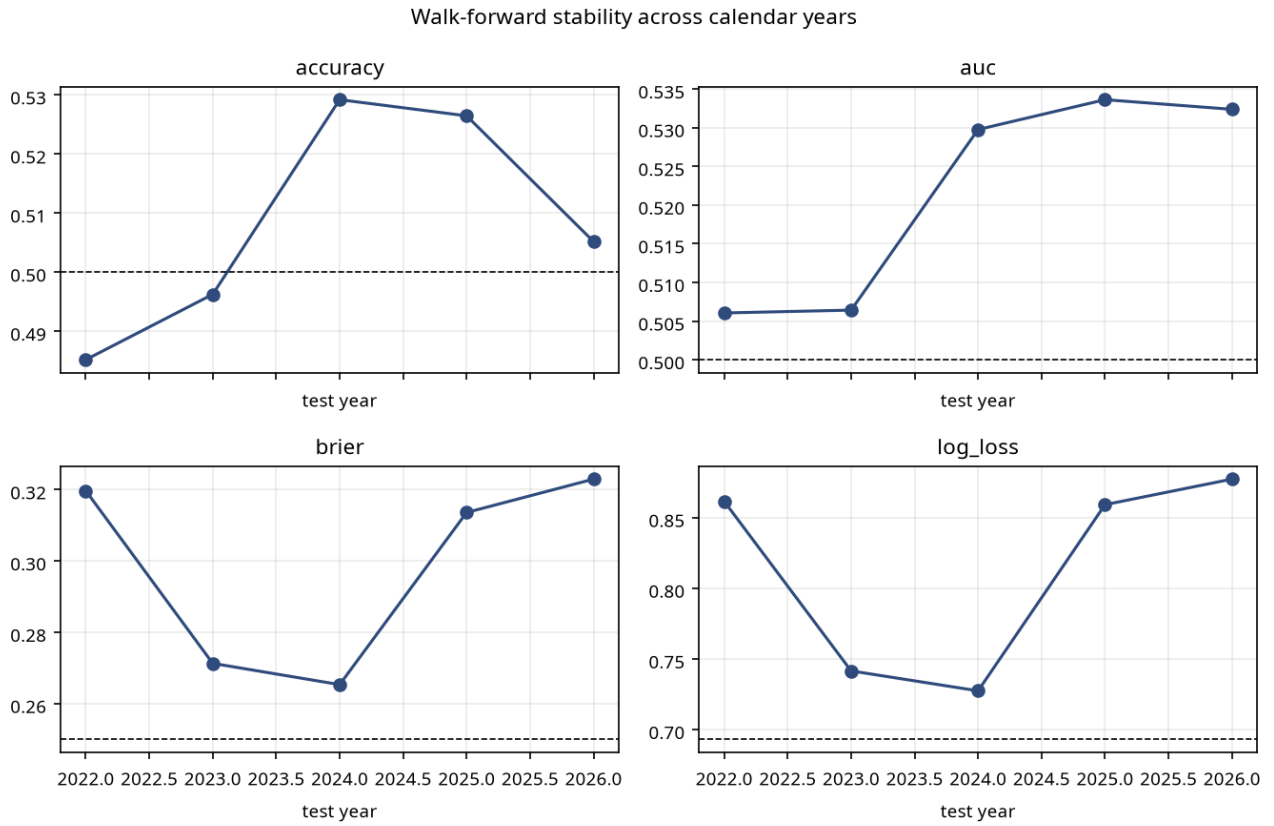


Figure 4: Annual walk-forward macro-average metrics. Dashed lines show constant-baseline references where applicable.

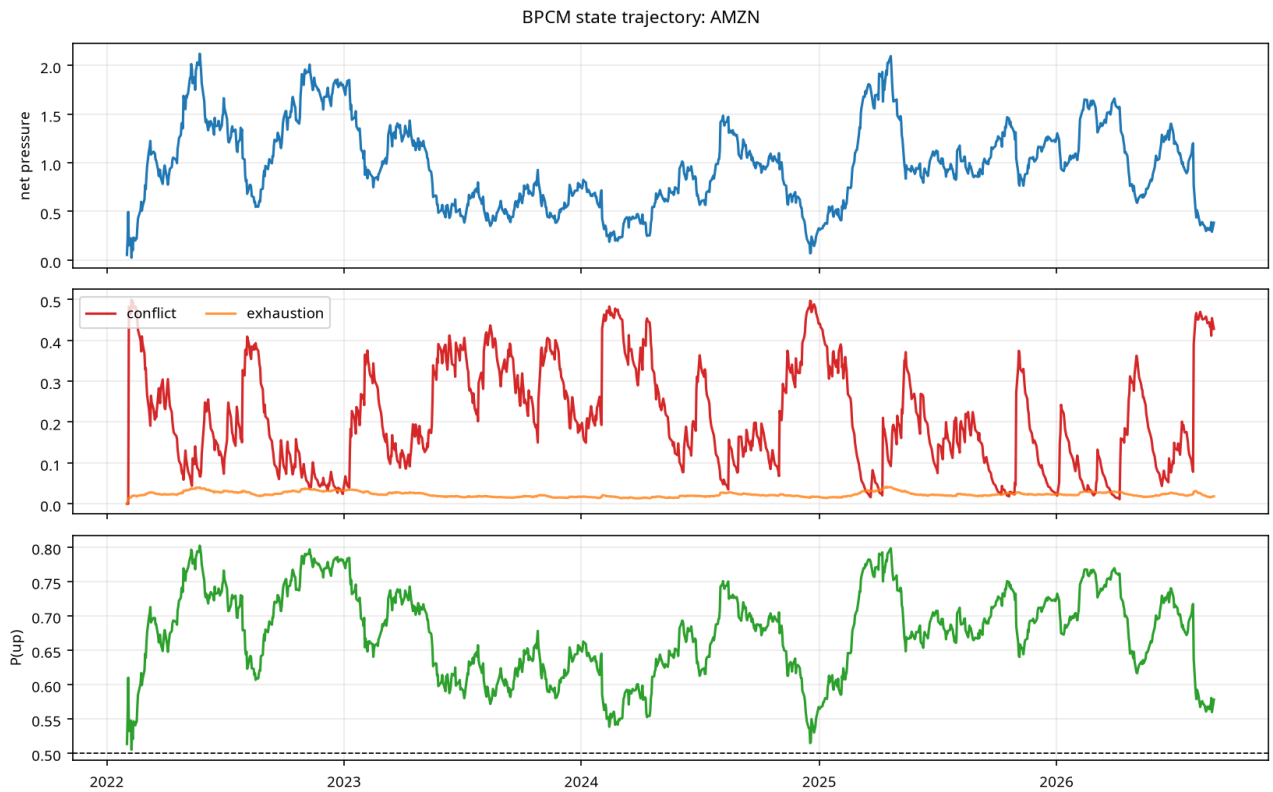


Figure 5: Representative BPCM trajectory showing net pressure, conflict, exhaustion, and score-derived probability.

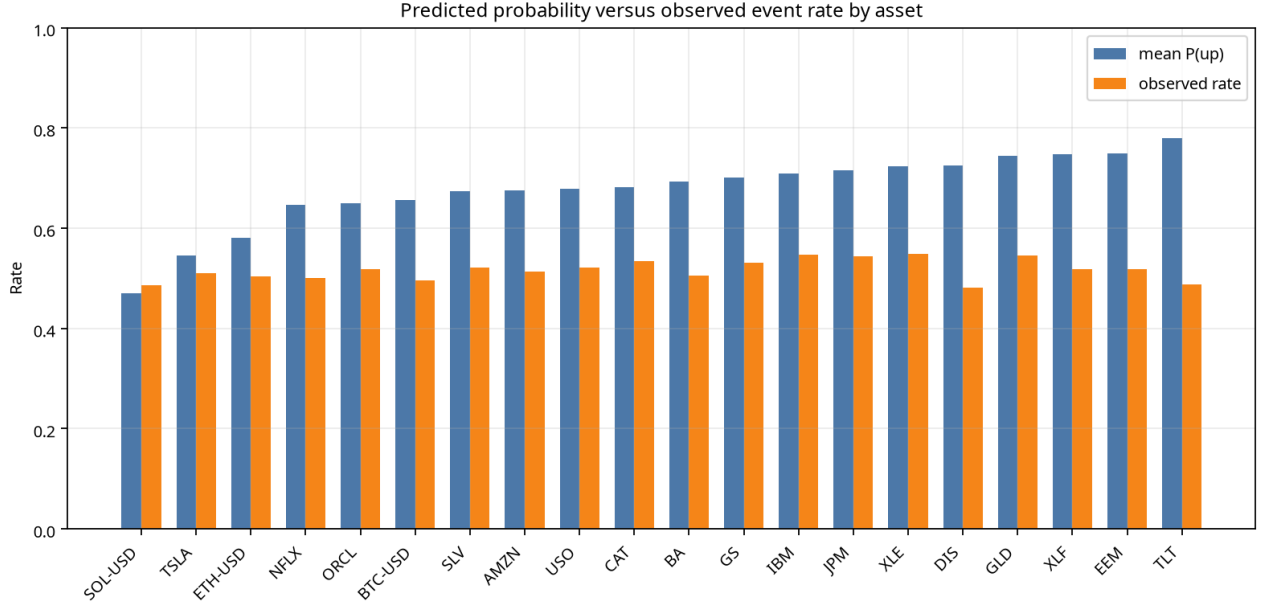


Figure 6: Mean BPCM probability and observed positive-return rate by asset. The systematic gap is a direct descriptive indicator of miscalibration.

8 Extended Statistical and Economic Diagnostics

The primary metrics establish that BPCM is close to chance in ranking and inferior to a constant forecast in probability quality. This section examines the structure of that result in greater detail. The objective is not to create new favorable claims after seeing the primary outcome. The diagnostics are descriptive extensions generated from the same stored prediction files and use the same fixed model outputs.

8.1 Event frequency, score location, and probability dispersion

Across the 20 successful assets and 24,585 observations, the observed positive-return rate is 0.5168. The mean BPCM probability is 0.6774. The probability standard deviation averaged across assets is 0.0798. This combination is informative. The score distribution is not simply a constant 0.5 forecast with random small perturbations. It is shifted upward and remains relatively concentrated. The model therefore expresses a directional preference that is much stronger than the realized base rate supports.

A high mean probability can arise from class imbalance, a score transformation, state persistence, or systematic model bias. The present experiment cannot identify which mechanism dominates. The observed positive-rate and predicted-probability comparison is nevertheless sufficient to reject an assumption of calibration. A model with a probability mean of 0.677 would need either a correspondingly high event rate or conditional discrimination that explains the high scores. The BPCM AUC of 0.506 does not provide such discrimination.

8.2 Calibration metrics and reliability diagram

We compute ten equal-width probability bins. For bin k , let n_k be its count, $\bar{p}_k$ its mean predicted probability, and $\bar{y}_k$ its observed event rate. The expected calibration error is

$$\text{ECE} = \sum_{k=1}^K \frac{n_k}{n} |\bar{p}_k - \bar{y}_k|, \quad (14)$$

and the maximum calibration error is

$$\text{MCE} = \max_{1 \leq k \leq K} |\bar{p}_k - \bar{y}_k|. \quad (15)$$

These bin-based statistics depend on the binning rule and are reported as descriptive summaries rather than universal population parameters.

The mean asset-level ECE is 0.1768 and the mean maximum calibration error is 0.2989. The reliability diagram in Figure 7 shows the observed event rate below the diagonal over much of the probability range. This means that the score-derived probability tends to be too large relative to the frequency of positive outcomes. Calibration is not a cosmetic issue here. It explains why BPCM can produce a forced-choice accuracy near 0.52 while incurring Brier and log-loss penalties.

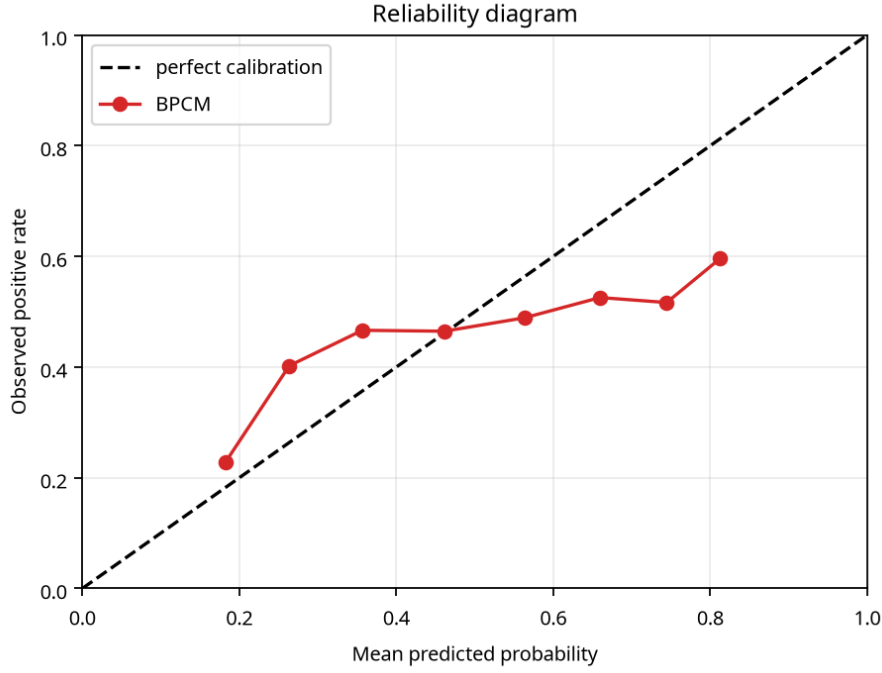


Figure 7: Reliability diagram using ten equal-width probability bins. The diagonal denotes perfect calibration.

Table 7: Extended pooled and macro-average diagnostics.

Diagnostic	Value
Successful assets	20
Pooled observations	24,585
Observed positive-return rate	0.5168
Mean predicted probability	0.6774
Mean probability standard deviation	0.0798
Mean expected calibration error	0.1768
Mean maximum calibration error	0.2989
Accuracy bootstrap macro interval	[0.4887, 0.5446]
AUC bootstrap macro interval	[0.4701, 0.5352]
Brier bootstrap macro interval	[0.2748, 0.2982]
Log-loss bootstrap macro interval	[0.7533, 0.8050]

Calibration error should be interpreted alongside sample size and serial dependence. Equal-width bins can contain unequal numbers of observations. A future version should report equal-frequency bins, calibration curves with time-block uncertainty, isotonic or Platt calibration fitted on validation data, and calibration drift by year. None of those procedures should be fit on the final test period.

8.3 Bootstrap uncertainty for proper scores

The extended analysis uses 2,000 independent observation-resampling replicates per asset. The macro-average lower, median, and upper quantiles are calculated across asset-specific bootstrap intervals. The mean accuracy interval is approximately [0.4887, 0.5446], with median 0.5169. The mean AUC interval is approximately [0.4701, 0.5352]. The mean Brier interval is approximately [0.2748, 0.2982], and the mean log-loss interval is approximately [0.7533, 0.8050].

These intervals should not be read as confidence intervals for an independent and identically distributed population because financial observations are temporally dependent. Observation resampling treats a high-volatility cluster and an isolated observation as exchangeable. That assumption is doubtful. The intervals are still useful as a compact description of finite-sample variation, but a stronger inferential design would use moving blocks, stationary blocks, or a model-based dependence adjustment. The paper reports both the point estimate and the limitation rather than using a nominal 95% label to imply stronger validity than the resampling scheme supports.

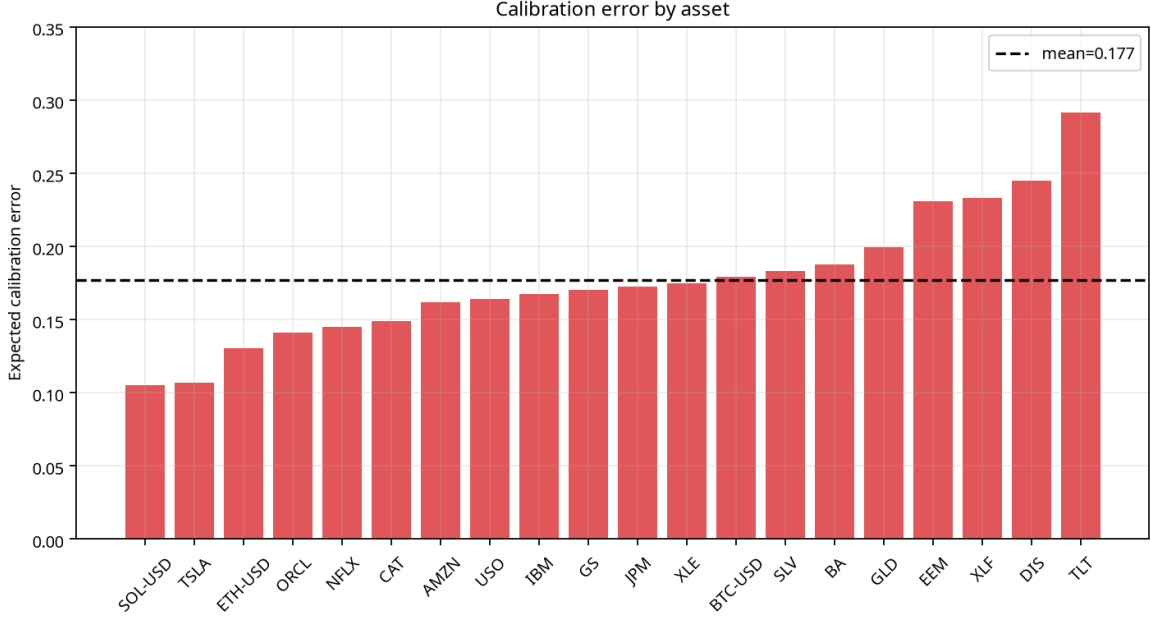


Figure 8: Expected calibration error by asset. The dashed line is the mean across assets.

8.4 Action composition and directional imbalance

The decision rule produces 93.28% mean coverage, 5.05% mean down-action rate, 88.23% mean up-action rate, and 6.72% mean abstention rate. The ratio is not a small directional asymmetry. The action policy is heavily long-biased. This follows from the score distribution and is consistent with the elevated mean probability. A model that produces mostly up actions can obtain an accuracy close to the unconditional positive-return frequency without discriminating well between individual days.

Mean selective accuracy is 0.5188. Mean action turnover, defined here as the fraction of days on which the discrete action differs from the preceding action including the initial transition from zero, is 0.0212. The mean entry count is 14.85 per asset. These figures are diagnostics of the implemented policy, not estimates of trading turnover under a portfolio convention. In particular, they do not specify whether a position is held overnight, whether an action reversal requires two transactions, or how multiple assets are capital-weighted.

The long bias matters for interpretation. If the positive-return frequency is above 0.5 in a particular sample, an always-up classifier can have accuracy above 0.5. The constant probability benchmark does not define an always-up action, so a complete action comparison should include majority-class and always-long policies in a future release. The current BPCM comparison therefore emphasizes proper probability scores, AUC, and explicit action composition rather than treating 51.7% accuracy as evidence of conditional information.

8.5 Action-change cost sensitivity

To expose the effect of trading frictions without presenting a false return backtest, we subtract a hypothetical unit-payoff cost c whenever the discrete action changes:

$$\Pi_i(c) = r_i - c|a_i - a_{i-1}|, \quad (16)$$

where r_i is +1, -1, or 0 for a correct action, incorrect action, or abstention. This statistic is intentionally dimensionless. It does not map one unit of payoff to one dollar, one basis point, or one unit of capital.

The mean gross unit payoff is 0.0359. The mean net unit payoff is 0.0358 at cost 0.001, 0.0358 at 0.0025, 0.0358 at 0.005, 0.0356 at 0.01, 0.0354 at 0.02, and 0.0348 at 0.05. The modest sensitivity is a consequence of low discrete action-change frequency under the stateful policy, not evidence that realistic costs are negligible. The cost scale is arbitrary, and the model's strong long bias could create exposure and drawdown risks not represented in a symmetric unit-payoff score.

8.6 Benchmark deltas and per-asset comparisons

The extended output stores BPCM-minus-baseline deltas for accuracy, AUC, Brier score, and log loss against both the constant forecast and logistic regression. A positive accuracy or AUC delta is favorable. A positive Brier or log-loss delta is unfavorable. This sign convention prevents a common table-reading error in which a larger loss is mistaken for improvement.

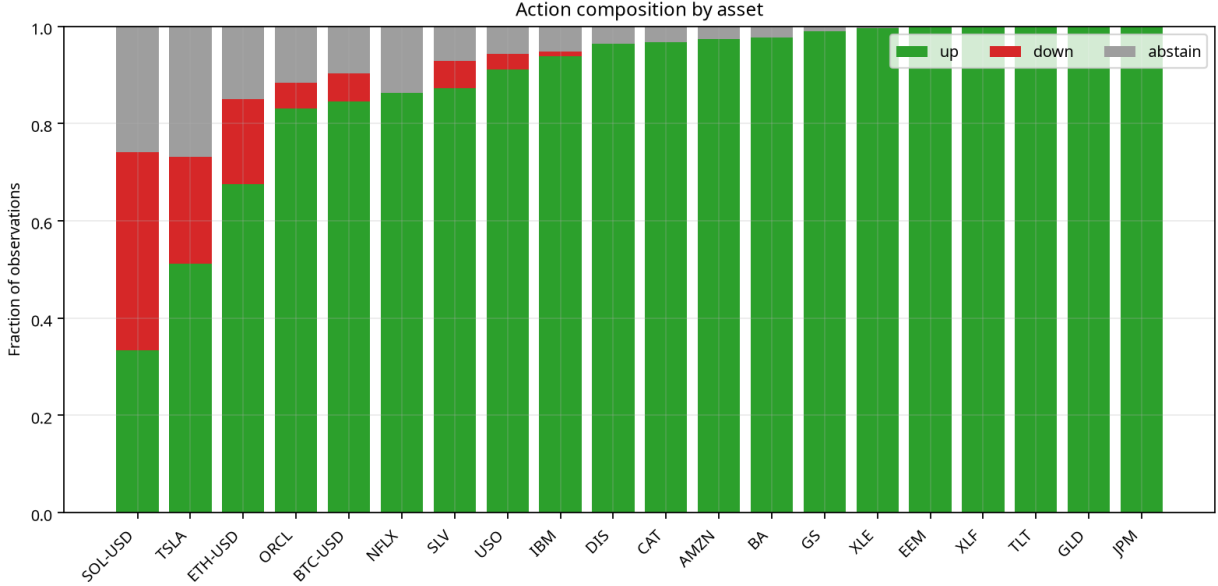


Figure 9: Composition of up, down, and abstain actions by asset. The policy is strongly up-action dominated.

Table 8: Interpretation of benchmark deltas. Values are BPCM minus the named benchmark; lower loss deltas are favorable.

Comparison	Δ accuracy	Δ AUC	Δ Brier	Δ log loss
BPCM minus constant 0.5	+0.0001	+0.0025	+0.0364	+0.0859
BPCM minus logistic	+0.0080	+0.0006	+0.0353	+0.0836

BPCM has a small positive macro accuracy delta over the constant benchmark but a materially positive Brier and log-loss delta. Relative to logistic regression, the pressure mechanism adds no robust ranking advantage and worsens probability quality. The per-asset table shows why aggregate means must be accompanied by dispersion: BPCM’s accuracy advantage is positive in some instruments and negative in others, while AUC often moves in the opposite direction.

8.7 Distributional and regime considerations

Accuracy aggregates conceal the magnitude and distribution of returns. The present target is binary, so a correct prediction on a tiny positive move receives the same unit payoff as a correct prediction on a large positive move. The study does not therefore estimate correlation with returns, information coefficient, rank correlation, conditional return magnitude, tail capture, maximum drawdown, volatility, skewness, kurtosis, or expected shortfall. These omissions prevent an economic claim even if direction metrics had been stronger.

The annual walk-forward results provide a limited regime decomposition. Accuracy is below 0.50 in 2022 and 2023, above 0.52 in 2024 and 2025, and near 0.50 in 2026. The period differences could reflect market regime, asset composition, sample noise, or data and model interaction. Without a pre-specified regime taxonomy and formal tests, it is not appropriate to label the years as causal regimes. A future study should define volatility, trend, liquidity, and correlation regimes before evaluation and report performance conditional on those definitions.

9 Implementation Audit and Reproducibility Details

9.1 Numerical implementation

The reference implementation uses Python, NumPy, pandas, scikit-learn, Matplotlib, SciPy-compatible scientific dependencies, and yfinance. Logistic regression uses a fixed random seed and a maximum iteration count of 2,000. Random forest uses 300 trees and parallel workers. Histogram gradient boosting uses fixed iteration, learning-rate, leaf-count, and regularization parameters. The BPCM loop applies the positive-part operator after each reservoir update, then computes total pressure, net pressure, conflict, exhaustion, saturation, probability, and uncertainty in that order.

The model state is reset for each asset. This is appropriate for the current cross-sectional evaluation because it prevents the final pressure state of one instrument from becoming the initial state of another. It also means the

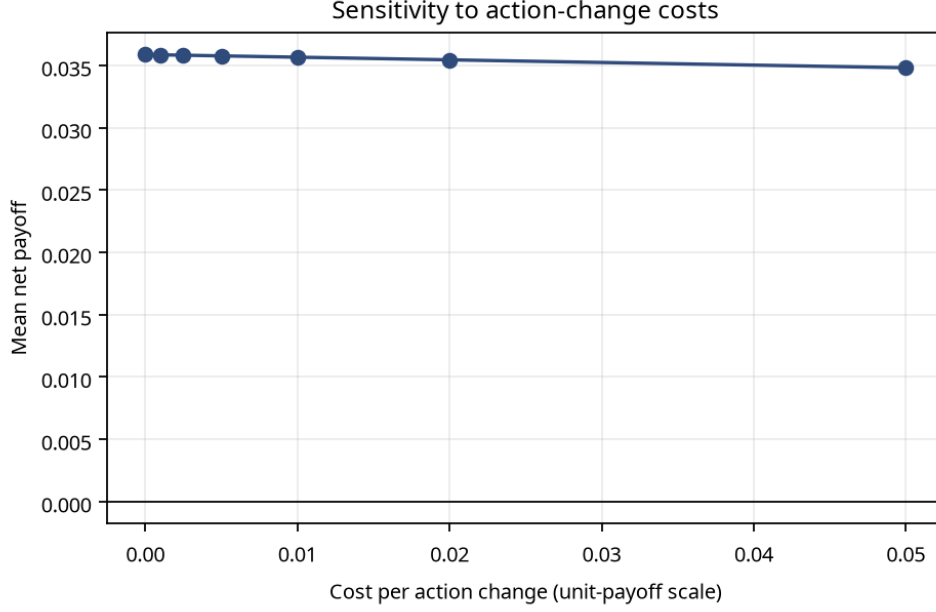


Figure 10: Dimensionless action-change cost sensitivity. This is not a currency return or a transaction-cost backtest.

reported model is not a single portfolio-level pressure process. A portfolio implementation would need a separate state-design decision: independent states, capital-weighted shared state, or a cross-asset interaction matrix.

9.2 Data and missingness

The runner writes a raw CSV cache for each downloaded asset and macro series. Cached data are reused on subsequent runs. The runner catches unavailable or feature-empty assets and writes the symbol, stage, and reason to `results/skipped_assets.csv`. This behavior is preferable to silently shrinking the universe or imputing a synthetic series. The final experiment contains 20 successful test assets and excludes three configured foreign-exchange symbols that yielded no rows after feature construction.

Missingness is not random in a broad financial universe. A symbol may be absent because of ticker conventions, vendor coverage, listing history, or a mismatch between price and macro calendars. The current paper reports the missingness but does not estimate selection bias. A fully archival study should define the universe before downloading and preserve failed requests, delisted instruments, and vendor metadata.

9.3 Reproducibility checklist

The release contains the following artifacts:

1. model source and mathematical specification;
2. expanded study runner and extended diagnostic script;
3. benchmark, ablation, walk-forward, and sensitivity CSV files;
4. calibration, action, cost, and benchmark-delta CSV files;
5. cached raw input data and study metadata;
6. original and extended figure-generation scripts;
7. thirteen paper figures embedded from PNG files;
8. the LaTeX source and compiled PDF; and
9. a dependency manifest and reproducibility commands.

The source hash in the study metadata identifies the extended study script used for the primary expanded run. Reproducing the exact numerical files without the included cache remains conditional on vendor data stability. Reproducing the scientific conclusion requires preserving the same date ranges, successful-symbol policy, parameter values, random seed, and metric definitions.

9.4 Failure modes and safeguards

The study includes several safeguards against common implementation errors. It checks required price and volume columns. It drops rows with incomplete features. It fits transformations on training data rather than test data. It resets dynamic states at asset boundaries. It stores predictions rather than only aggregate metrics. It records skipped assets. It tests ZIP integrity and compiles the paper twice so that cross-references and long tables are resolved.

The safeguards do not prevent all scientific failure modes. They do not prevent vendor revisions, target ambiguity around corporate actions, multiple-testing from future experimentation, dependence-induced false precision, or economic misinterpretation. Reproducible code is necessary but not sufficient for a credible empirical claim.

10 Expanded Limitations and Research Agenda

The first limitation is measurement. BPCM’s language concerns pressure, but the inputs are daily technical and macro features rather than order-book events. The model cannot distinguish a genuine market-pressure mechanism from a nonlinear transformation of lagged returns and volatility.

The second limitation is target choice. Next-day close direction discards return magnitude, intraday path, gap risk, and investor utility. A model may be useful for volatility or tail-risk conditioning without improving a binary sign forecast. Conversely, a model can obtain sign accuracy above chance without producing a tradable return after costs.

The third limitation is cross-asset transfer. Training mainly uses liquid equities and broad ETFs, while testing includes cryptoassets and instruments with different trading calendars. Cross-asset generalization is deliberately difficult and scientifically useful, but a failure does not identify whether the mechanism is wrong or the transfer assumption is too strong.

The fourth limitation is calibration. The score-derived probability is not calibrated. A future training protocol should separate fitting, calibration, and final testing. Calibration should be performed with a chronological validation fold and evaluated on a later untouched fold. A calibration method chosen after inspecting the final test reliability plot would invalidate the evidence.

The fifth limitation is uncertainty quantification. Observation-level bootstrap intervals are descriptive. Dependence-aware intervals, hierarchical asset-level resampling, conformal risk control under time dependence, and selective-risk confidence bounds would materially strengthen the inference.

The sixth limitation is economic simulation. A serious strategy study requires explicit holdings, exposure constraints, realistic costs, price execution, turnover accounting, portfolio aggregation, drawdowns, tail risk, and a benchmark such as buy-and-hold or a majority-class policy. The current unit-payoff cost curve is an audit tool rather than a strategy return.

The seventh limitation is novelty. A pressure-reservoir construction may overlap conceptually with inventory models, order-flow imbalance, nonlinear filters, recurrent state models, and market-impact dynamics. The present literature review supports cautious positioning but is not a complete systematic review or patent search. No claim that BPCM is unprecedented should be made.

A prioritized research agenda follows. First, archive a legally redistributable dataset and add release timestamps. Second, implement dependence-aware inference. Third, add intraday order-flow and liquidity variables. Fourth, compare calibrated models under nested walk-forward selection. Fifth, evaluate multiple targets and economic objectives. Sixth, preregister the model variants and metrics before examining a new test period. Seventh, conduct a formal systematic review before narrowing the novelty statement.

11 Conclusion

The deeper analysis strengthens the paper’s evidence without changing its conclusion. BPCM is mathematically explicit and computationally reproducible. It exposes internal states and supports an abstaining decision. It can be evaluated across assets, periods, benchmarks, ablations, thresholds, calibration bins, and action-cost assumptions.

The empirical evidence remains weak. The model’s ranking is near chance. Its mean predicted probability is 0.677 against a positive-return rate of 0.517. Its expected calibration error is 0.177. Its action policy is 88.2% up actions, 5.1% down actions, and 6.7% abstentions. Its mean selective accuracy is 0.519. Its proper scores are worse than the constant forecast. Its annual performance is not stable. These results are not evidence of a reliable predictive or economic edge.

The appropriate scientific status is therefore clear. BPCM is a transparent, falsifiable pressure-reservoir hypothesis with a complete negative-result evaluation. The next contribution should come from better measurement, stronger validation, calibrated probabilities, dependence-aware uncertainty, and realistic economic testing—not from adding complexity to a model whose present daily-data evidence remains close to chance.

A Metric Definitions and Output Map

This appendix records the exact operational meaning of the reported quantities. The definitions are included to reduce ambiguity when the code is rerun or extended.

Table 9: Metric dictionary.

Metric	Operational definition and interpretation
Accuracy	Fraction of observations for which $p_i \geq 0.5$ gives the correct binary target. It is a threshold metric and is sensitive to the event base rate.
ROC-AUC	Probability that a randomly selected positive observation receives a larger score than a randomly selected negative observation. It evaluates ranking, not calibration.
Brier score	$n^{-1} \sum_i (p_i - y_i)^2$. Lower is better. A constant 0.5 forecast has Brier score 0.25 when the event is binary and balanced in the score convention.
Log loss	$-n^{-1} \sum_i [y_i \log p_i + (1 - y_i) \log(1 - p_i)]$. Lower is better and confident wrong predictions receive large penalties.
Coverage	Fraction of observations for which the selective action is nonzero.
Selective accuracy	Fraction of correct actions conditional on a nonzero action. It must be interpreted jointly with coverage.
Unit payoff	+1 for a correct action, -1 for an incorrect action, and 0 for abstention. It is not a currency return.
ECE	Ten-bin weighted mean absolute difference between mean predicted probability and observed event rate.
MCE	Maximum absolute probability/event-rate gap across the ten bins.
Turnover proxy	Fraction of days on which the discrete action differs from its previous action, including the initial zero state. It is not portfolio turnover.
Bootstrap interval	Quantiles from independent observation-resampling replicates. It is descriptive because serial dependence is ignored.

A.1 Artifact correspondence

The empirical claims in the paper can be traced to stored artifacts. Primary BPCM results are in `results/study_metrics.csv` and pooled values are in `results/pooled_metrics.json`. Benchmark values are in `results/benchmark_metrics.csv`. Mechanism variants are in `results/ablation_metrics.csv`. Annual refit values are in `results/walk_forward_metrics.csv`. Threshold and uncertainty sweeps are in `results/sensitivity.csv` and `results/sensitivity_summary.csv`. Calibration, action, cost, and benchmark-delta outputs are in the `results/extended/` directory. The five original figures and five extended figures are stored as PNG files in `results/figures/` and are inserted directly into this PDF.

The primary and extended analyses use a fixed random seed of 42. The primary bootstrap uses 1,000 replicates. The extended diagnostics use 2,000 replicates. The paper reports macro-average and pooled quantities separately because assets have different observation counts. All reported percentages are computed from unrounded values and rounded only for presentation.

A.2 Parameter and decision table

Table 10: Fixed BPCM parameters and decision settings used for the primary output.

Symbol	Value	Role
δ_B	0.035	Buying-pressure decay
δ_S	0.035	Selling-pressure decay
η	0.045	Opposing-pressure cancellation
ρ	0.20	Total-pressure saturation
ξ	0.018	Exhaustion rate
s	2.0	Impulse pressure scale
τ	0.55	Action probability threshold
$u_{\max}$	0.62	Uncertainty acceptance limit

The parameters are fixed for the reported primary test and are not selected by maximizing the final test metric. The ablation variants change one mechanism or memory setting at a time. Future parameter selection should use an internal validation period and a preregistered search space.

Reproducibility Statement

The repository contains the model implementation, expanded study runner, benchmark and ablation code, sensitivity analysis, figure generator, mathematical specification, raw-data cache, CSV result tables, JSON metadata, skipped-symbol record, dependency manifest, paper source, and compiled PDF. From the repository root, the principal commands are:

```
python3 -m pip install -r requirements.txt
python3 run_study.py --output results --bootstrap 1000 --seed 42
python3 make_figures.py
pdflatex -interaction=nonstopmode -halt-on-error paper.tex
pdflatex -interaction=nonstopmode -halt-on-error paper.tex
```

The stored metadata records date ranges, configured universes, successful assets, skipped assets, bootstrap count, random seed, data source, and a SHA-256 hash of the study script. Reproduction without the included cache remains conditional on the future behavior of Yahoo Finance.

References

- [1] R. Cont, A. Kukanov, and S. Stoikov, “The Price Impact of Order Book Events,” *Journal of Financial Econometrics*, 12(1), 47–88, 2014. DOI: <https://doi.org/10.1093/jjfinec/nbt003>.
- [2] Z. Eisler, J.-P. Bouchaud, and J. Kockelkoren, “The price impact of order book events: market orders, limit orders and cancellations,” *Quantitative Finance*, 12(9), 1395–1419, 2012. DOI: <https://doi.org/10.1080/14697688.2010.528444>.
- [3] F. Lillo and J. D. Farmer, “The Long Memory of the Efficient Market,” *Studies in Nonlinear Dynamics & Econometrics*, 8(3), Article 20123001, 2004. DOI: <https://doi.org/10.2202/1558-3708.1226>.
- [4] R. E. Kalman, “A New Approach to Linear Filtering and Prediction Problems,” *Journal of Basic Engineering*, 82(1), 35–45, 1960. DOI: <https://doi.org/10.1115/1.3662552>.
- [5] A. H. Jazwinski, *Stochastic Processes and Filtering Theory*. Academic Press, Mathematics in Science and Engineering, vol. 64, 1970. <https://www.sciencedirect.com/bookseries/mathematics-in-science-and-engineering/vol/64/suppl/C>.
- [6] S. H. Babbs and K. B. Nowman, “Kalman Filtering of Generalized Vasicek Term Structure Models,” *Journal of Financial and Quantitative Analysis*, 34(1), 115–130, 1999. DOI: <https://doi.org/10.2307/2676248>.
- [7] T. Bollerslev, “Generalized Autoregressive Conditional Heteroskedasticity,” *Journal of Econometrics*, 31(3), 307–327, 1986. DOI: [https://doi.org/10.1016/0304-4076\(86\)90063-1](https://doi.org/10.1016/0304-4076(86)90063-1).
- [8] T. Bollerslev, “Glossary to ARCH (GARCH),” in *Volatility and Time Series Econometrics*, Oxford University Press, chapter 8, 137–163, 2010. DOI: <https://doi.org/10.1093/acprof:oso/9780199549498.003.0008>.
- [9] P. F. Christoffersen, F. X. Diebold, R. S. Mariano, A. S. Tay, and Y. K. Tse, “Direction-of-change forecasts based on conditional variance, skewness and kurtosis dynamics: International evidence,” *Journal of Financial Forecasting*, 1(2), 1–22, 2007. https://ink.library.smu.edu.sg/soe_research/538/.
- [10] T. Gneiting and A. E. Raftery, “Strictly Proper Scoring Rules, Prediction, and Estimation,” *Journal of the American Statistical Association*, 102(477), 359–378, 2007. DOI: <https://doi.org/10.1198/016214506000001437>.
- [11] D. J. Hand, “Measuring classifier performance: a coherent alternative to the area under the ROC curve,” *Machine Learning*, 77, 103–123, 2009. DOI: <https://doi.org/10.1007/s10994-009-5119-5>.
- [12] Y. Geifman and R. El-Yaniv, “Selective Classification for Deep Neural Networks,” in *Advances in Neural Information Processing Systems 30*, 2017. <https://papers.nips.cc/paper/7073-selective-classification-for-deep-neural-networks>.
- [13] R. Sullivan, A. Timmermann, and H. White, “Dangers of Data Mining: The Case of Calendar Effects in Stock Returns,” *Journal of Econometrics*, 105(1), 249–286, 2001. DOI: [https://doi.org/10.1016/S0304-4076\(01\)00077-X](https://doi.org/10.1016/S0304-4076(01)00077-X).
- [14] R. D. McLean and J. Pontiff, “Does Academic Research Destroy Stock Return Predictability?” *The Journal of Finance*, 71(1), 5–32, 2016. DOI: <https://doi.org/10.1111/jofi.12365>.

- [15] D. H. Bailey, J. M. Borwein, M. López de Prado, and Q. J. Zhu, “The Probability of Backtest Overfitting,” *Journal of Computational Finance*, 20(4), 39–69, 2017. DOI: <https://doi.org/10.21314/JCF.2016.322>.
- [16] L. J. Tashman, “Out-of-sample tests of forecasting accuracy: An analysis and review,” *International Journal of Forecasting*, 16(4), 437–450, 2000. DOI: [https://doi.org/10.1016/S0169-2070\(00\)00065-0](https://doi.org/10.1016/S0169-2070(00)00065-0).
- [17] C. Bergmeir and J. M. Benítez, “On the use of cross-validation for time series predictor evaluation,” *Information Sciences*, 191, 192–213, 2012. DOI: <https://doi.org/10.1016/j.ins.2011.12.028>.
- [18] C. Bergmeir, R. J. Hyndman, and B. Koo, “A note on the validity of cross-validation for evaluating autoregressive time series prediction,” *Computational Statistics & Data Analysis*, 120, 70–83, 2018. DOI: <https://doi.org/10.1016/j.csda.2017.11.003>.
- [19] A. Ranaroussi, *yfinance: Yahoo Finance market data downloader*, software repository and documentation. <https://github.com/ranaroussi/yfinance>.