Université du Québec en Outaouais

Département des sciences administratives

Working Paper No. 41

Multi-Resolution Context in Intraday Probabilistic Return Forecasting: A Pre-Registered, Matched-Budget Measurement on 300 U.S. Instruments

CHRONOS-X: a Multi-Resolution Patch Transformer against matched baselines, 5 to 60 minutes ahead, 2018–2025

Simon-Pierre Boucher

Département des sciences administratives
Université du Québec en Outaouais

simon-pierre.boucher@uqo.ca

Gatineau – Pavillon Alexandre-Taché
283, boulevard Alexandre-Taché
Gatineau, Québec, Canada J9A 1L8

November 2026

Version 1.1

Abstract

Does multi-resolution context improve calibrated probabilistic forecasts of equity and ETF returns 5–60 minutes ahead? We test five pre-registered hypotheses with a Multi-Resolution Patch Transformer, its one-minute-only twin and ten baselines on 234 million one-minute bars of 300 instruments, with day-cluster inference. Over five seeds at a 12,000-step budget the contrast is -0.081 to -0.214 % of scaled CRPS, every seed pair negative: the pre-stated rule supports the hypothesis in 2024 at small magnitude; on 2025 these checkpoints keep the sign but fail it. A minute-scale Student-*t* HAR-RV matches the neural models at five minutes and leads by 0.21–0.49 % afterwards. Pooled training beats refitted baselines on unseen instruments only at five minutes, without penalty; no location forecast has economic content after costs; a rank-4 copula adds nothing. The full model gains 1.54 % and 0.30 % of scaled CRPS over a calibrated Student-*t* random walk at 5 and 60 minutes.

Keywords: intraday forecasting, probabilistic forecasting, transformers, multi-resolution representations, calibration, proper scoring rules, high-frequency data

JEL Classification: C45, C53, C58, G12, G17

1. Introduction

Every minute of a U.S. trading day leaves behind one bar—open, high, low, close, volume—for each of several thousand instruments. Foundation models for time series promise to learn representations from such streams and to forecast any series zero-shot (Ansari et al., 2024; Das et al., 2024), a growing literature applies deep sequence models to intraday prices (Sirignano and Cont, 2019; Zhang et al., 2019; Kolm et al., 2023), and multi-scale patch Transformers report gains over single-scale designs on the standard long-horizon benchmarks (Shabani et al., 2023; Zhang et al., 2024b; Chen et al., 2024; Du et al., 2023). Whether a compact model that reads several resolutions of minute-level history improves *calibrated* forecasts of the next hour, relative to its one-minute-only twin and to the volatility models an econometrician would write down, is an empirical question that is rarely posed with matched information sets, a matched training budget, proper scoring rules and inference that respects the dependence of overlapping intraday windows. This paper poses it in that form and reports what the data show.

We build a Multi-Resolution Patch Transformer (MRPT) inside the research package CHRONOS-X.[1] At a decision time inside the regular session the model reads the last 390 completed one-minute bars, the last five days of completed 5-minute aggregates and the last twenty days of 30-minute aggregates, each tokenised in patches with gated feature projections, resolution and temporal-gap embeddings; a shared pre-norm Transformer encoder and a direct decoder with sixty future-calendar queries emit a Student-t marginal for each of the next sixty minute returns, and a second-stage low-rank Gaussian copula turns the marginals into joint sample paths. All inputs are strictly causal, scale-free (divided by a trailing volatility computed at the origin) and masked where the source has no bar; the model carries no ticker identity. Five hypotheses were written before any model was trained: multi-resolution history helps at a matched architecture and training budget (H1), pooled training transfers to unseen instruments (H2, in two clauses), history-only regime conditioning improves calibration under stress (H3), coherent paths beat independent marginals on path scores (H4), and—exploratory—the forecasts carry economic content after costs (H5).

The data are 300 U.S. stocks and ETFs drawn at the end of 2017 from 2,147 candidates that met point-in-time liquidity rules, stratified by asset class, dollar volume, volatility and sector, with 60 instruments held out entirely: 234 million one-minute bars from a private FirstRate Data lakehouse (FinData OS) over 2018–2025, audited for timestamp conventions, session boundaries, early closes, daylight-saving time, missing minutes, splits and dividends. The draw is not point-in-time in one respect that matters: candidates were kept only if their realised regular-hours minute coverage over 2018–2025 reached 80 %, an availability filter computed ex post that removes 994 candidates, 66 of

[1] The name is unrelated to Chronos (Ansari et al., 2024) and to ChronosX (Pineda Arango et al., 2025), which adapts that pretrained model to exogenous variables; neither is evaluated here and the package shares no code with them. The package keeps its name; the paper title names the measurement.

which pass the same gate on the training years alone. The filter removes the names whose trading dried up, which are the episodes most likely to stress calibration and transfer; we therefore score the 66 excluded instruments as a second held-out set. Training uses 2018–2022, validation 2023, the development test 2024 and an untouched final period 2025; evaluation uses 986,468 origins in 2024 alone, every 15 minutes from 10:00 over 251 sessions. Baselines share the information set exactly: a random walk with a trailing scale (Gaussian, and Student-t with scale and tail calibrated on 2023), seasonal volatility, EWMA, HAR-RV at the minute scale, GARCH(1,1)-t, ridge, elastic net, LightGBM, an MLP and a TCN with the same Student-t head, and the one-minute-only MRPT as the matched-architecture control (same width, depth and optimiser steps; 7.4 % fewer parameters). Scores are the CRPS, the weighted interval score, quantile losses, coverage, PIT diagnostics, QLIKE, tail Brier scores and, for paths, energy and variogram scores. Inference treats the trading day as the cluster: the headline statistic is a day-cluster t on the loss differential, checked by a Diebold–Mariano test on daily mean differentials, with Holm adjustment inside named families and model confidence sets per horizon; H1 is additionally replicated over five training seeds, because one seed cannot separate an architectural effect from training noise.

The findings are negative for the architecture and informative about the data, and they are stated with their uncertainty. Every model except the corrected LightGBM at five minutes beats the Gaussian random walk at every horizon, by 2.50 to 4.14 % of scaled CRPS; against a Student-t random walk whose scale and tail are calibrated on 2023 the gain of the full 12,000-step model shrinks to 1.54 % of scaled CRPS at five minutes and 0.30 % at sixty (Table 8). The information in the minute stream is about dispersion: the four neural location forecasts have directional accuracy of 0.499–0.502, the linear and tree location forecasts reach 0.518 at five minutes, and with a one-minute execution delay none of the 7 location models has a net return after costs that survives in both 2024 and 2025 (5 of 224 cells positive in 2024 against about 11.2 expected false positives, 0 in 2025). With baselines built under one consistent retransformation, the neural models and HAR-RV differ by -0.07 % at five minutes (Holm $p = 0.148$) and HAR-RV with a kurtosis-matched Student-t is both sharper and better calibrated from fifteen minutes on, by 0.21–0.49 %. The contrast between MRPT and its one-minute-only twin—the multi-resolution context that defines the model—is -0.081 % at five minutes and -0.214 % at sixty minutes on average over five seeds, negative in every seed pair at every horizon (t across seeds -4.53 and -4.05, day-cluster t -13.6 and -11.2), so that the pre-stated seed-aware rule is met at 4 of the four horizons and Hypothesis 1 is supported in 2024. The effect is small: 1.5–3.4 times the seed half-band of 0.05–0.06 % that the paper fixed as its smallest effect size of interest, with per-seed contrasts spanning -0.15 % to -0.04 % at five minutes and -0.41 % to -0.13 % at sixty, and an order of magnitude below the gap to the random walk. In the untouched 2025 year the same eight seed checkpoints, scored once, keep the sign at every horizon (-0.116 % to -0.046 % on average, 4 of five seed pairs negative) but fail the pre-stated rule, because the t across

seeds stays below its 2.78 threshold; the 2024 verdict therefore does not replicate in 2025, and the main run alone—the one seed without a 2025 effect—is what the single-seed reading of version 1.0 showed. A 20.7-million-parameter model does no better than a 3.5-million one at the same 12,000-step budget. The remaining hypotheses are examined in Sections 7 and 8: zero-shot transfer to the held-out instruments and to the availability-excluded set, regime conditioning on stress days against a control that does not receive the regime statistics, the copula paths against independent sampling with diagnostics of the dependence the copula cannot see, and the untouched 2025 period.

The contribution is a pre-registered, matched-information-set, matched-budget measurement and its result—a multi-resolution effect that is statistically detectable in 2024, economically negligible beside what the heavy tail and the scale contribute, and absent from the untouched 2025 year—not a new state of the art. Three strands of published work are the reference points. Multi-scale patch Transformers (Shabani et al., 2023; Zhang et al., 2024b; Chen et al., 2024; Du et al., 2023; Xiao et al., 2025) document gains from multi-resolution context on long-horizon benchmarks; H1 is their standing claim, and the present measurement—a few hundredths to a few tenths of a percent of CRPS, above the seed band but an order of magnitude below the random-walk gap, on minute-level financial returns—is informative only because it is positioned against them. Pooled-model transfer across stocks is established for limit-order-book inputs (Sirignano and Cont, 2019) and for intraday realised variance (Zhang et al., 2024a); our second clause of H2 reproduces it for bars-only distributional forecasts and does not extend it. For daily realised variance, Christensen et al. (2023) and Zhang et al. (2024a) find that regularised regressions, trees and networks beat the HAR lineage, with larger gains at longer horizons; at 5–60-minute horizons, on full predictive distributions and with matched information sets, we find the opposite ordering from fifteen minutes on, and Section 3 sets out the design differences (horizon, target, loss, information set) that we believe reconcile the two. A fourth element is a set of measurements that we did not plan but that the protocol produced: the dependence of the model's PIT scores across future minutes, the share of the random-walk gain attributable to a calibrated heavy tail rather than to minute-level structure, and the behaviour of the nulls on the instruments the availability filter removed.

Section 2 states the hypotheses and their pre-specified tests, Section 3 places them in the literature, Sections 4–6 describe data, model and design, Sections 7–8 report the results, and Sections 9–11 discuss limitations, reproducibility and what the evidence does and does not say.

2. Research question and falsifiable hypotheses

At a decision time t inside the regular session, with every one-minute bar up to t complete, can a compact model that reads the recent minute-level history *and* coarser past-only aggregates produce better-calibrated and sharper predictive distributions for the next $H \in \{5, 15, 30, 60\}$ minutes than

the same model reading the minute stream only, or than well-specified non-neural volatility models? "Better" is defined by strictly proper scoring rules (Gneiting and Raftery, 2007; Matheson and Winkler, 1976), never by price-level errors: on a one-minute grid the random walk is a formidable point forecaster, and a model can look impressive on price-level mean squared error while learning nothing (Cont, 2001). We therefore train and score on return-based targets and evaluate full predictive distributions.

Let r_{t+h} denote the log return of the bar opening h minutes after t and $R_{t,H} = \sum_{h \le H} r_{t+h}$ the cumulative return. The model emits, for each future minute, a Student-t marginal with location, scale and degrees of freedom, and (in a second stage) a dependence structure across the H minutes. Five hypotheses were written before any model was trained; the project brief (`CLAUDE.md`) and the frozen v1.0 manuscript hold the original wording, which is the primary statement of each hypothesis below. Where the revision of this paper sharpened a statement to the test that was actually executed, the sharpened form follows as an *operationalisation adopted in revision*; no direction, no test statistic and no threshold was changed after the results were known, and Section 9 lists every deviation from the pre-registered protocol.

Hypothesis 1 (Multi-resolution history)**.** At equal parameter count and training budget, a model whose encoder reads the last 390 one-minute bars, the last five days of 5-minute aggregates and the last twenty days of 30-minute aggregates achieves a lower CRPS (and WIS) for $R_{t,H}$ than the same architecture reading the one-minute bars only.

Operationalisation adopted in revision. "Equal parameter count and training budget" is read as the same architecture (width, depth, decoder and head), the same optimiser steps and batch size, and a parameter count within 7.4 % of the full model's (the twin has fewer parameters because it has no 5- and 30-minute branches, an asymmetry that favours the null); the budget is the 12,000-step main-study budget.

Hypothesis 2 (Transfer to unseen instruments)**.** A model trained on a pool of instruments, without any ticker identity, scores better on instruments it has never seen than non-neural baselines fitted with the same information set, and its loss on unseen instruments is not materially worse than on pooled instruments.

Operationalisation adopted in revision. The statement has two clauses, (a) the pooled model beats the non-neural baselines refitted on the unseen instruments' own histories and (b) transfer costs nothing, which the v1.0 text tested and reported as one; they are tested and reported separately here.

Hypothesis 3 (History-only regime conditioning)**.** Conditioning the decoder on history-only regime statistics (trailing volatility and relative volume) improves calibration—measured by coverage gaps and the probability-integral-transform (PIT) dispersion—on days whose cross-sectional volatility is in the top decile, relative to the same model without regime conditioning.

Operationalisation adopted in revision. "Without regime conditioning" is read as a control that does not receive those statistics through any input; the v1.0 control received them through the global token, which confounded the test (Section 8).

Hypothesis 4 (Coherent joint paths)**.** A dependence-aware decoder (a Gaussian copula with a low-rank factor structure over the H future minutes, with the Stage-A marginals held fixed) achieves lower energy and variogram scores for the joint path, and lower CRPS for path functionals (maximum drawdown, realised variance), than sampling the same marginals independently.

Operationalisation adopted in revision. The factor structure has rank 4, as implemented; the verdict is specific to that copula (Section 9).

Hypothesis 5 (Economic content, exploratory)**.** Signals derived from the predictive median contain incremental information after an assumed round-trip cost; this hypothesis is exploratory because no quote data are available and costs are assumptions.

Operationalisation adopted in revision. The predictive median of every location model (the 7 models with a location forecast, not only MRPT-S) is tested, with a one-minute execution delay.

Each hypothesis has a pre-specified test. The primary statistic for a loss contrast is a day-cluster t (the v1.0 text named the origin-level Diebold–Mariano statistic; the change, made in revision before the seed study was run, is recorded in Section 9): the difference in mean loss between two models, with a standard error from the trading-day cluster bootstrap (all instruments and origins of one session form one cluster, because millions of overlapping windows are not independent observations). Two checks accompany it: a Diebold–Mariano test on the 251 daily mean loss differentials with a Newey–West variance (Diebold and Mariano, 1995; Harvey et al., 1997; Newey and West, 1987), and the original Diebold–Mariano statistic on the origin-level differential with an overlap-aware HAC variance, which is reported but not used for decisions because its Bartlett lag does not absorb day-level common shocks. Holm adjustments (Holm, 1979) are applied within named families (one family per evaluation split and loss), and a model confidence set at level 0.10 (Hansen et al., 2011) is computed per horizon from a moving-block bootstrap of daily mean losses. Models are never re-estimated during an evaluation year, so the comparisons are of the fixed-scheme kind (Giacomini and White, 2006); no Giacomini–White conditional test is run. For Hypothesis 1 the one-seed test is supplemented by a five-seed replication of both twins: the contrast is declared detectable at a horizon only if the mean of the per-seed differences is at least 2.78 standard errors from zero (a t with four degrees of freedom) *and* the day-cluster t on the seed-averaged differential exceeds 1.96 in absolute value. The smallest effect size of interest for H1 is the seed half-band of the full model (0.05–0.06 % of scaled CRPS); the minimum detectable effect at 80 % power from the day-cluster standard error is 0.02 % at five minutes and 0.07 % at sixty (Table B.7, Appendix). Null findings are reported as such, with these bounds.

3. Related work

Volatility models, machine learning and the HAR benchmark. The classical description of intraday returns combines a deterministic periodic component with persistent conditional heteroskedasticity (Andersen and Bollerslev, 1997; Bollerslev, 1986; Nelson, 1991; Engle, 2001). Realised measures built from intraday returns turned volatility into an (almost) observable quantity and showed that simple heterogeneous autoregressions forecast it well (Andersen et al., 2003; Barndorff-Nielsen and Shephard, 2002; Corsi, 2009; Bollerslev et al., 2016); Hansen and Lunde (2005) found that little beats a GARCH(1,1) once the comparison is fair, and Patton (2011) showed that only a few loss functions (QLIKE, MSE) rank variance forecasts consistently under noisy proxies. The HAR lineage is also the benchmark that the recent machine-learning literature sets out to beat, with mixed outcomes. Christensen et al. (2023) find that regularised regressions, trees and neural networks improve on HAR for the daily realised variance of DJIA constituents, with larger gains at longer horizons; Zhang et al. (2024a) report that neural networks trained on pooled intraday data dominate HAR-type models for intraday realised variance and transfer to stocks not in the training set; Reisenhofer et al. (2022) obtain modest gains with a convolutional network that nests HAR, which suggests that the structure the data reward is HAR-like; and Brini (2026) find no uniform gain of nine zero-shot foundation models over eight HAR-family specifications. Our result—HAR-RV with a Student-t body matching the neural models at five minutes and leading from fifteen minutes on—conflicts with the first two findings and is consistent with the last two. The designs differ in four ways that we believe explain the difference: the horizon (5–60 minutes against days), the target (the full distribution of the cumulative return, scored by CRPS and WIS, against a point forecast of realised variance scored by MSE or QLIKE), the information set (every model reads exactly the completed bars at the origin, and the neural models receive no feature that HAR-RV does not) and the inference (trading-day clusters on 300 instruments). We do not claim that one of these is decisive; Section 7 reports which published patterns reproduce and which do not.

Deep sequence models and multi-resolution Transformers. Sirignano and Cont (2019) documented that a single neural model trained on pooled limit-order-book data transfers across stocks, the closest antecedent of our second clause of Hypothesis 2, although with order-book inputs that we do not have; DeepLOB (Zhang et al., 2019) and deep order-flow imbalance models (Kolm et al., 2023) also use quote-level data, and Lucchese et al. (2024) measure short-horizon predictability from the order book with model confidence sets on a multi-horizon problem of the same kind as ours. With bars only, the relevant architectures are the generic sequence models: temporal convolutions (Bai et al., 2018), including the dilated causal design that Moreno-Pino and Zohren (2022) apply to intraday bars for day-ahead volatility and that our TCN baseline follows; DeepAR-style autoregressive

likelihood models (Salinas et al., 2020); N-BEATS (Oreshkin et al., 2020); and Transformer variants (Vaswani et al., 2017; Wen et al., 2023). Patch tokenisation (Nie et al., 2023) and variable-as-token inversion (Liu et al., 2024) are the designs our encoder adapts, and Zeng et al. (2023) is the reminder that linear baselines are competitive. MRPT belongs to the family of multi-scale patch Transformers—Scaleformer (Shabani et al., 2023), MTST (Zhang et al., 2024b), Pathformer (Chen et al., 2024), MultiResFormer (Du et al., 2023) and the multi-resolution patching of TimeFound (Xiao et al., 2025)—whose common claim is that reading several temporal resolutions improves long-horizon forecasts on the electricity, traffic, weather and ETT benchmarks. Hypothesis 1 transposes that claim to minute-level financial returns under a matched budget; the null we report is a counter-example on this data, not a verdict on those benchmarks. Pretrained foundation models for time series (Ansari et al., 2024; Das et al., 2024) and the finance-specific K-line foundation model Kronos (Shi et al., 2026) promise zero-shot forecasting; the evidence on daily financial data is mixed (Rahimikia et al., 2025; Brini, 2026) and direct application to prices appears to require fine-tuning (Fu et al., 2025). We discuss in Section 6 why they are not evaluated here, and we note that the matched-budget question of this paper differs from the question a zero-shot evaluation answers. Diffusion-based multivariate forecasters (Rasul et al., 2021) are a possible Stage-B alternative that we do not pursue given their compute cost. Surveys of financial machine learning (Gu et al., 2020; Kelly and Xiu, 2023; Lim and Zohren, 2021) stress the evaluation pitfalls we try to avoid: overlapping samples, survivorship, look-ahead and multiple testing (López de Prado, 2018; Bailey and López de Prado, 2014).

Scoring, inference and dependence. For distributional forecasts we rely on the theory of proper scoring rules (Matheson and Winkler, 1976; Gneiting and Raftery, 2007), the calibration–sharpness paradigm and PIT diagnostics (Gneiting et al., 2007), the weighted interval score (Bracher et al., 2021), closed-form CRPS for the Student-t family as implemented in `scoringRules` (Jordan et al., 2019; Taillardat et al., 2016), threshold-weighted variants (Gneiting and Ranjan, 2011), and multivariate energy and variogram scores (Scheuerer and Hamill, 2015). Inference on loss differentials uses the Diebold–Mariano statistic (Diebold and Mariano, 1995; Harvey et al., 1997) on daily mean differentials with a Newey–West variance (Newey and West, 1987), trading-day cluster bootstraps, Holm and Benjamini–Hochberg adjustments (Holm, 1979; Benjamini and Hochberg, 1995) and the model confidence set of Hansen et al. (2011); the fixed-scheme setting of Giacomini and White (2006) describes our comparisons but its conditional test is not run. Coherent joint paths require a dependence model. Copulas separate marginals from dependence (Sklar, 1959; Patton, 2006); our Stage B uses a Gaussian copula with a low-rank-plus-diagonal correlation over the sixty future minutes, trained on the probability-integral transforms of the observed targets with the marginals frozen, and Section 8 reports what such a copula can and cannot detect. Finally, the

high-frequency predictability literature finds large and pervasive short-horizon predictability from trades-and-quotes predictors (Ait-Sahalia et al., 2026; Kolm et al., 2023; Lucchese et al., 2024); our directional results, obtained from bars alone, are read against that finding in Section 9.

Table 1. Contribution matrix: the closest existing studies and this paper (horizon, target, loss, instruments, information set, finding)

Study	Horizon	Target	Loss / score	Instruments	Information set	Finding
Christensen et al. (2023)	1–22 days	daily realised variance (point)	MSE, QLIKE vs HAR lineage	DJIA constituents	daily RV built from intraday returns, minimal tuning	regularisation, trees and neural nets beat HAR; gains larger at longer horizons
Zhang et al. (2024a)	intraday (sub-daily RV)	intraday realised volatility (point)	MSE/QLIKE; pooled vs stock-specific	U.S. stocks	own history + market-volatility commonality proxy	pooled neural nets dominate; trained models transfer to stocks absent from training
Sirignano and Cont (2019)	next price move	direction of the next mid-price change	classification accuracy	≈ 1,000 U.S. stocks	limit-order-book states, pooled	a universal model trained across stocks beats stock-specific models
Lucchese et al. (2024)	multi-horizon (ticks to minutes)	mid-price return direction / magnitude	accuracy, Model Confidence Sets	many U.S. stocks	volume representation of the limit order book	short-horizon predictability from the book; universal trading patterns; multi-horizon modelling helps
Ait-Sahalia et al. (2026)	seconds to minutes	high-frequency returns and durations	out-of-sample R^2, classification	U.S. equities (ultra-high-frequency)	trades-and-quotes predictors	predictability is large, systematic and pervasive at short horizons
Kolm et al. (2023)	multiple short horizons	mid-price returns (alpha)	out-of-sample R^2	Nasdaq stocks	deep order-flow imbalance (LOB)	alpha concentrated at the shortest horizons, decaying with horizon
Moreno-Pino and Zohren (2022)	one day ahead	daily volatility (point)	point-forecast losses vs daily-data baselines	equities	intraday returns through dilated causal convolutions	dilated convolutions on high-frequency inputs beat daily-data baselines
Reisenhofer et al. (2022)	daily	daily realised volatility (point)	point-forecast losses vs HAR	U.S. equities and indices	HAR-structured convolutional network	modest gains over HAR with an interpretable HAR nesting
Zhang et al. (2024b); Chen et al. (2024); Shabani et al. (2023); Du et al. (2023)	96–720 steps	multivariate point forecasts	MSE/MAE	ETT, weather, traffic, electricity benchmarks	single series at several patch sizes / scales	multi-resolution or multi-scale patching improves on single-resolution PatchTST-type models
Xiao et al. (2025)	zero-shot, many horizons	univariate point/quantile forecasts	benchmark MSE/MAE, CRPS	public benchmarks	multi-resolution patching inside a pretrained foundation model	multi-resolution patching helps a time-series foundation model
Shi et al. (2026)	next bars, volatility	K-line (OHLCV) sequences	task-specific losses	multi-exchange candlesticks	pre-trained K-line tokeniser	a financial foundation model outperforms general time-series foundation models on K-line tasks
Rahimikia et al. (2025)	daily	excess returns	zero-shot, fine-tuned and from-scratch comparisons	global markets	time-series foundation models	mixed: foundation-model gains are not uniform across markets
Brini (2026)	daily (three horizons)	realised volatility (point)	pairwise and multi-model tests vs HAR family	50 assets (VOLARE)	nine zero-shot foundation models vs eight econometric specifications	no uniform gain over HAR; advantages concentrated in outlier assets
Fu et al. (2025)	daily prices	price series	forecast error after fine-tuning	financial series	TimesFM fine-tuned on prices	direct application unsatisfactory; fine-tuning needed
This paper	5, 15, 30, 60 minutes	full predictive distribution of the cumulative return (Student-t marginals per minute)	CRPS, WIS, coverage, PIT, QLIKE, path scores; day-cluster inference, Holm/BH, MCS	300 U.S. stocks and ETFs (60 held out; 2024 and untouched 2025)	completed one-minute bars only (no quotes), identical for every model; one-minute-only twin with the same architecture and optimiser steps	see Sections 7–8 and 11; every number is generated from the result files

Notes. Rows summarise what each study measured, as stated in the cited paper (verified records: `results/metrics/references_verification.json`); they are not re-estimations. The multi-scale patch-Transformer family (MTST, Pathformer, Scaleformer, MultiResFormer) is grouped in one row because the four papers share the design and the benchmarks. The matched-information-set, proper-scoring, day-cluster design of this paper differs from the daily-horizon point-volatility comparisons (Christensen et al.; Zhang et al.; HARNet; DeepVol; Brini) and from the order-book predictability studies (Sirignano and Cont; Lucchese et al.; Aït-Sahalia et al.; Kolm et al.) on horizon, target, loss and information set, which is where the conflicting "machine learning beats HAR" and "returns are predictable" findings are reconciled in Sections 7.3, 7.5 and 11.

4. Data

4.1. Source and connector

All bars come from FinData OS, a private lakehouse of FirstRate Data one-minute bars for U.S. stocks and ETFs (6.70 and 0.91 billion rows respectively at audit time, 2000–2026). The connector discovers the catalog, the published Arrow schema and the OpenAPI document, reads small verification pages through the paginated bars endpoint and performs bulk retrieval through asynchronous Parquet exports pinned to a dataset version. Every call keeps the request identifier, dataset identifier and version; every export file is schema-validated and hashed. The license is private use: no raw bar, grid or per-origin return leaves the compute node, and the release contains code, aggregate results, per-origin *scores* and synthetic fixtures only (Section 10).

4.2. Conventions verified on real rows

Timestamps are bar-*open* instants in UTC: the bar stamped 09:30 ET covers 09:30:00–09:30:59 and is complete at 09:31. A regular XNYS session is 09:30–16:00 ET, i.e. 391 bars on a normal day (slots $0, \ldots, 390$) and 211 on a 13:00 early close; the last slot is the closing print, whose volume is one to two orders of magnitude larger than a typical minute (6.7 million against a median of 0.12 million shares for AAPL on 2024-03-08). The closing print is excluded from context windows and targets. Extended hours (04:00–20:00 ET) are present at the source and excluded here. Zero-volume minutes are absent at the source and are never filled: missingness is a mask and a feature. Daylight-saving transitions are handled through the exchange calendar (first regular bar at 14:30Z on 2024-03-08 and 13:30Z on 2024-03-11). Stored bars are raw; splits are applied explicitly per session from the instrument master, which reproduces the server-side split-adjusted series exactly (AAPL 4:1 on 2020-08-31, ratio 4.000 on prices and volume). Cash dividends are not applied: intraday log returns within a session are invariant to the dividend factor and any dividend adjustment is window dependent; the overnight-gap feature across an ex-dividend date is left raw. Mergers and spin-offs are unavailable; a delisted series simply ends.

4.3. Universe: point-in-time selection rules and an ex-post availability filter

The selection date is 2017-12-29. Eligible instruments have a first bar at least six months before that date and a last bar at least six months after the sample start (a documented survival requirement), an average daily dollar volume of at least 5 million USD over December 2017 and at least fifteen daily bars in that window. These rules use only information available at the selection date and leave 2,147 candidates. From them we draw 300 instruments proportionally over asset class (75 % stocks, 25 % ETFs), dollar-volume tercile and daily-volatility tercile, with a round-robin over SIC divisions

(stocks, from SEC submissions) or ETF name categories inside each stratum. A first draw revealed regular-hours minute coverage between 6 % and 100 % over 2018–2025 (median 82 %): a quarter of the instruments traded in fewer than 60 % of regular minutes, which makes per-minute targets mostly masked. Before any model was trained we audited the coverage of all candidates and imposed an availability filter of at least 80 % regular-hours coverage over the full window 2018–2025, which leaves 1,152 candidates (Spearman correlation between 2017 dollar volume and coverage 0.76); the second draw keeps the same strata and seed. 60 instruments (20 %) are held out entirely for zero-shot evaluation.

The availability filter is not point-in-time: it reads the validation, test and final years. Recomputed on the training years alone (mean yearly coverage 2018–2022 at or above the gate), 1,178 candidates pass; 66 pass on the training years but are excluded by the full-window gate because their coverage fell below the gate in 2023, 2024 and 2025 (0.72, 0.63 and 0.64 on average); 40 are admitted only because of later years; and 12 of the 300 selected instruments fail the gate on the training years (27 have 2024 coverage below the gate, 20 have 2025 coverage below it). The filter removes 83 % of the candidates (Table C.1, Appendix) in the lowest 2017 dollar-volume tercile and 9 % in the highest. Its direction of bias is known: the names whose trading dried up are removed, and these are the episodes most likely to stress calibration and zero-shot transfer. We therefore export the 66 train-pass, full-fail instruments, refit the non-neural baselines on their own histories and score the frozen pooled checkpoints on them zero-shot as a second held-out set (Section 7.6). One candidate (GLOB) failed the coverage audit with an HTTP error and was excluded without a measured coverage; it is active at the source. Only 5 delisted candidates passed the filter, and 1 of the 300 selected instruments (CNO) carries a delisted flag in the vendor instrument master, with a master `last_date` of 2024-12-20 that the exported bars contradict: all 300 instruments have bars through 2025, and the five delisted candidates share that master date, which we read as an instrument-master artefact rather than five same-day delistings. FinData's delisted one-minute archives are only partly ingested, so the universe is close to a survivor sample, a limitation we carry through the paper.

4.4. Sample, grid and targets

The main sample spans 2018-01-02 to 2025-12-31 (2011 sessions): 233958776 exported rows, of which 220142835 fall in regular hours (93.7 % coverage on average, minimum 80.3 %; 13.8 million extended-hours rows excluded; zero integrity rejections and duplicates; 242 vendor-flagged rows in 105 instruments—60 price jumps and 182 outside-session stamps—kept with their flags; the counts are released as an aggregate). Each instrument is stored as a dense session $\times$ 391 grid of log open, high, low, close and log volume with a presence mask, plus past-only 5-minute (78 per day) and 30-minute (13 per day) aggregates with their own masks and a past-20-session seasonal profile of log volume and of squared returns. An origin (s, m) is session s at slot m, with bars $0, \ldots, m$

complete; it is eligible when $m \geq 30$, $m + 60$ does not cross the last tradable slot (389, or 209 on an early close), bar m is present and at least 80 % of the next 60 slots are present. The last rule conditions on future presence, identically for every model; it removes 7.2 % of the pooled training origins, 7.4 % of the validation origins, 10.9 % of the 2024 origins and 9.1 % of the 2025 origins (6.0–10.0 % on the held-out instruments); its effect on the scores was not measured (Table C.2, Appendix; Section 9). Targets are $y_h = \log C_{m+h} - \log C_{p(m+h-1)}$, where $p(\cdot)$ is the last present slot, with mask $\mathbb{1}[\text{bar } m + h \text{ present}]$, so that the masked sum over $h \leq H$ is exactly the H-minute log return of the last observed bar. Derived targets are the cumulative return $R_{t,H}$, the realised variance $\sum y_h^2$, high/low excursions, and the tail event $\left|R_{t,H}\right| > 2.5\,\sigma_t\sqrt{H}$, where σ_t is the trailing root-mean-square one-minute return in the 390-bar window (floored at 10^{-5}), computed from the past only.

4.5. Features

Every bar carries fifteen causal features: scaled return, range, upper and lower wicks, body, close location, log volume, relative volume against the seasonal profile, missing flag, session-boundary flag, log temporal gap (which encodes overnight and weekend gaps), minute fraction, time to close, weekday and early-close flag. Returns and ranges are divided by σ_t (times $\sqrt{5}$ and $\sqrt{30}$ for the aggregates), so inputs are scale free. Nine global history-only statistics—$\log \sigma_t$, the scaled overnight gap, the calendar position (minute fraction, time to close, weekday, early-close flag), the ratios of 30-minute and intraday realised variance to σ_t^2, and the mean relative volume over 30 minutes—feed a global token and the regime embedding; the four volatility-regime statistics among them (overnight gap, the two realised-variance ratios and relative volume) are the ones the H3 control removes (Section 8). A leakage gate in the test suite perturbs every bar after the origin and asserts that no input changes; a quarantined fixture that deliberately uses the next bar is rejected by the same test.

4.6. Splits

Sessions are split chronologically: training up to 2022-12-31, validation 2023, development test 2024 and an untouched final test 2025 evaluated once, with a one-session embargo at the start of each evaluation block. Held-out instruments are evaluated only with models trained on the pooled instruments. Training origins are 2 random eligible slots per instrument-session (586,202 training origins over 2018–2022; a run of 12,000 steps of 64 sees 768,000 examples, i.e. 1.31 passes; windows are indexed into memory-mapped grids). Every protocol constant below is regenerated from the run metadata (Table C.3, Appendix). Evaluation origins are every 15 minutes from slot 30 (10:00) to slot 315, i.e. at most 20 candidate slots on a normal session (8 on an early close), of which 16.4 per instrument-session are eligible on average in 2024 (986,468 origins over 251

sessions and 240 pooled instruments; 250,322 origins on the 60 held-out instruments) and 16.9 in 2025 (1,008,425 origins over 249 sessions; 254,334 held-out). The 2024 development test carries every headline table and every ablation; 2025 was scored once per final model and never used for any selection.

5. The Multi-Resolution Patch Transformer

5.1. Inputs and tokenisation

MRPT is a member of the multi-scale patch-Transformer family (Shabani et al., 2023; Zhang et al., 2024b; Chen et al., 2024; Du et al., 2023): each temporal resolution is tokenised in patches and the tokens share one encoder. Branch A reads the last 390 one-minute bars ending at the origin (crossing session boundaries, with boundary and gap flags), tokenised in patches of five bars (78 tokens). Branch B reads the last 390 completed 5-minute buckets (about five trading days; bucket j of the current session is complete only if slot $5j + 4 \leq m$), in patches of three (130 tokens). Branch C reads the last 260 completed 30-minute buckets (twenty days), in patches of two (130 tokens). No unfinished aggregate ever enters a window. Each patch is flattened, concatenated with its fraction of present bars and its maximal log gap, and projected through a gated linear unit (Dauphin et al., 2017) to d_{model}; a learned resolution embedding, a sinusoidal position measured as the distance to the origin in patch units, and a sinusoidal temporal-gap embedding are added. A global token projects the nine history-only statistics of Section 4. Patches without any present bar remain tokens but are masked in attention.

5.2. Backbone, fusion and regime conditioning

The default backbone is a shared pre-norm Transformer encoder (Vaswani et al., 2017; Ba et al., 2016) over the concatenated token sequence (339 tokens with the global token; 79 for the one-minute-only twin), with GELU activations (Hendrycks and Gimpel, 2016) and key-padding masks. The alternative fusion keeps a small encoder per resolution and lets the one-minute stream cross-attend to the two coarser streams before a final encoder. A regime embedding maps the global statistics through a two-layer MLP ("stats"), through a soft codebook of eight learned regimes ("learned"), or is disabled ("none"); it is added to the decoder queries. Because the global token carries the same statistics into the encoder, disabling the regime embedding alone does not remove them from the model; the control that does is described in Section 6. The core model carries no ticker identity; a ticker embedding exists only as an ablation.

5.3. Direct decoder and Student-t head (Stage A)

Sixty learned future-position queries, each augmented with the known future calendar position (minute fraction, time to close, weekday, early-close flag) and the regime embedding, pass through a two-layer Transformer decoder that cross-attends to the encoder memory. A linear head yields three numbers per future minute, parametrised in a scale-equivariant way:

$$\mu_h = \sigma_t \, a_h/10, \qquad s_h = \sigma_t \, \mathrm{softplus}(b_h + 0.5413) + 10^{-7}, \qquad \nu_h = 2.05 + \mathrm{softplus}(c_h + 1.5), \quad (1)$$

so that the scale starts at σ_t and the degrees of freedom start near 3.75 with finite variance; the location is damped by construction. The loss is the masked Student-t negative log-likelihood averaged over present target minutes, computed in single precision regardless of the device. Gaussian and quantile (pinball) heads are ablations. These are marginal predictions per minute; the cumulative-return distribution at H is obtained by Monte Carlo (200 draws) with independent minutes.

5.4. Dependence-aware paths (Stage B)

With the marginals frozen, a copula head produces, from the same decoder states, factor loadings $\Lambda \in \mathbb{R}^{H \times k}$ ($k = 4$) and a positive diagonal ψ, defining a correlation matrix $R = D^{-1/2}(\Lambda\Lambda^{\mathsf{T}} + \mathrm{diag}\,\psi)D^{-1/2}$ over the H future minutes. Training maximises the Gaussian-copula log-likelihood of $z_h = \Phi^{-1}(F_{t_{\nu_h}}((y_h - \mu_h)/s_h))$, with missing minutes marginalised exactly by replacing their rows and columns of R by the identity; the head is selected on the copula likelihood of a stratified validation sample of 50,000 2023 origins over all 240 pooled instruments, evaluated every 100 steps. Sampling draws $z = \Lambda\varepsilon + \psi^{1/2}\eta$, standardises, maps through Φ and the Student-t quantile function, and never needs a Cholesky factor. Joint samples give cumulative paths, maximum drawdown, realised volatility and interval coverage, scored against independent sampling of the same marginals. A Gaussian copula with a constant low-rank correlation measures linear dependence of the normal scores only; it is blind to dependence in z_h^2 or $|z_h|$ and to tail dependence, which Section 8 measures separately.

5.5. Sizes, training, selection and compute

Two presets are trained: Small ($d = 192$, 4 layers, 6 heads, FFN 768; 3,455,619 trainable parameters) and Base ($d = 384$, 8 layers, 8 heads, FFN 1536; 20,718,339), each with a two-layer decoder; the one-minute-only twin of Small has 3,200,259 parameters (no 5- and 30-minute patch projections or resolution embeddings) and the cross-attention variant 5,533,443. Parameter counts and measured throughput appear in Table 2. Optimisation uses AdamW (Loshchilov and Hutter, 2019) with $\beta = (0.9, 0.98)$, peak learning rate 3×10^{-4}, 400 warm-up steps and cosine decay to 10 % of the

peak rate, weight decay 0.01, gradient clipping at 1.0, batch 64, 12,000 optimiser steps for every main model (1.31 passes over the training origins) and 4,000 for the ablations, fixed seeds, and crash-safe checkpoints with resume. Mixed precision is not used. No hyper-parameter search was run: the learning rate, warm-up, weight decay, batch size and step budget were fixed before the main study and are the same for every neural model; the consequences of the fixed budget are measured in Section 8 with a five-seed replication and a learning curve to 34,000 steps.

Checkpoint selection deserves a precise statement. The main runs evaluate the validation likelihood every 1,000 steps with patience 4 on the first 60 batches of 64 rows of the instrument-ordered 2023 index, i.e. 3,840 origins that belong to a single instrument—an implementation artefact. Its consequence is bounded: re-scoring every saved checkpoint on a seeded stratified sample of 50,000 2023 origins across all 240 pooled instruments changes the selected checkpoint of none of the 12,000-step main models (the best checkpoint is the last one for MRPT-S, its twin, MRPT-B and the TCN, and an earlier one for the MLP), while 8 of the 15 4,000-step ablation models would have selected their last checkpoint instead of the saved one (Table 3); these ablations were not re-predicted and the single-instrument selection is a live source of their seed-level noise. Every run trained for this paper after the main study (seeds, learning curve, H3 controls, transfer run, copula head) selects on the stratified sample. Figure 4 shows the training and validation curves of the main models with the stratified re-score overlaid. The compute ledger (Table 4) lists every training run of the study with its node, chip, steps and wall-clock and the total in GPU-hours, of which 6.0 for the main study and ablations on one node; runs are compared at matched optimiser steps, never at matched wall-clock.

6. Experimental design

6.1. Baselines with a matched information set

Every baseline sees exactly the completed bars available at the origin and is fitted on the training period, with any free constant chosen on the 2023 validation period; nothing is re-estimated during 2024 or 2025. Non-neural: a random walk with zero drift and a Gaussian scale from the trailing root-mean-square one-minute return (the reference throughout); the same random walk with a Student-t body whose scale factor and degrees of freedom are chosen on 2023 by CRPS ("Student-t random walk"; its scale factor shrinks and its tail thickens with the horizon, 'results/metrics/recal/c_val.json'), which is the floor against which the contribution of minute-level structure is measured in Section 8; a seasonal-volatility model that sums the past-20-session profile of squared returns over the future slots, level-adjusted by the trailing 30-minute realised variance; an EWMA ($\lambda = 0.97$ per minute, Student-t with $\nu = 5$); HAR-RV at the minute scale (log realised variance over the next H minutes regressed by OLS on log trailing 30-minute, one-day and five-day realised variances and time of

Table 2. Measured training throughput, memory and cost of the model presets and of the five trained main-study models on the compute node (M3 Ultra)

Panel A. Presets benchmarked (synthetic batches; Apple M3 Ultra, torch 2.15.0, mps, FP32)

Preset	Batch	Parameters	Steps/s	Examples/s	Peak memory (MB)	Latency/origin (ms)	–
small	32	3,455,619	20.81	666	54	0.34	–
small	64	3,455,619	12.21	782	69	0.25	–
small	128	3,455,619	5.93	758	97	0.27	–
base	32	20,718,339	6.11	195	275	5.04	–
base	64	20,718,339	3.24	207	284	1.59	–
base	128	20,718,339	1.51	193	315	1.76	–

Panel B. Models trained in the main study (configs/main.yaml; training only, prediction excluded)

Model	Batch	Parameters	Steps	Wall-clock (s)	Examples/s	Peak memory (MB)	Best validation NLL
MRPT-S (1m+5m+30m)	64	3,455,619	12,000	1,759	437	63	-6.4535
MRPT-S (1m only)	64	3,200,259	12,000	1,088	706	87	-6.4520
MRPT-B (1m+5m+30m)	64	20,718,339	12,000	4,580	168	403	-6.4526
TCN (1m)	64	231,459	12,000	927	828	199	-6.4518
MLP (1m)	64	301,827	12,000	839	916	15	-6.4497

Notes. Panel A: measured on synthetic batches by `scripts/benchmark.py` (node m3u96b); data loader at 64 per batch: 0 workers: 3971 ex/s; 4 workers: 13016 ex/s; 8 workers: 38486 ex/s. The brief's 'spec' preset (d = 256, 6 layers; 7,677,955 parameters) was benchmarked but never trained and is therefore not listed. Panel B: `train_summary.json` of each run (wall-clock of the training loop including the periodic validation; prediction on the 2024 evaluation origins is extra; the best validation NLL is on the original 3,840-origin subset, see Table 3). The one-minute-only twin has 3,200,259 trainable parameters against 3,455,619 for the multi-resolution model (7.4 % fewer: no 5- and 30-minute patch projections or resolution embeddings) and the same width, depth, head and optimiser steps. Source: `results/metrics/benchmark_M3U96b.json`, `results/metrics/main/*/train_summary.json`.

Table 3. Checkpoint selection re-examined: validation NLL of every saved best and last checkpoint on the original 3,840-origin subset and on a seeded stratified sample of 50,000 2023 origins

Model	Saved best step	NLL, original subset (best)	NLL, stratified (best)	NLL, stratified (last)	Selected on the stratified subset
Panel A. Main study (12,000 steps)					
MLP (1m)	11,000	-6.44969	-6.11752	-6.11750	best (unchanged)
TCN (1m)	12,000	-6.45179	-6.11989	-6.11989	best (unchanged)
MRPT-S (1m+5m+30m)	12,000	-6.45346	-6.12091	-6.12091	best (unchanged)
MRPT-S (1m only)	12,000	-6.45202	-6.11949	-6.11949	best (unchanged)
MRPT-B (1m+5m+30m)	12,000	-6.45257	-6.11978	-6.11978	best (unchanged)
Panel B. Ablations (4,000 steps)					
MRPT-S (1m+5m)	3,500	-6.45123	-6.11855	-6.11825	best (unchanged)
MRPT-S cross-attention fusion	3,500	-6.45035	-6.11770	-6.11767	best (unchanged)
MRPT-S Gaussian head	3,500	-6.39999	-6.08111	-6.08161	last
MRPT-S no regime	3,500	-6.45086	-6.11827	-6.11834	last
MRPT-S learned regime	2,500	-6.45015	-6.11678	-6.11780	last
MRPT-S returns only	3,000	-6.44645	-6.11482	-6.11462	best (unchanged)
MRPT-S time only	3,000	-6.44640	-6.11476	-6.11463	best (unchanged)
MRPT-S OHLCV (no rel. volume)	3,500	-6.45000	-6.11793	-6.11751	best (unchanged)
MRPT-S shuffled minutes	3,500	-6.44936	-6.11716	-6.11720	last
MRPT-S + ticker id	4,000	-6.45159	-6.11891	-6.11891	best (unchanged)
MRPT-S reference (4k steps)	3,500	-6.45071	-6.11834	-6.11843	last
MRPT-S 1m only (4k)	3,500	-6.45048	-6.11769	-6.11783	last
MRPT-S reference, seed 2	3,500	-6.45127	-6.11818	-6.11828	last
MRPT-S reference, seed 3	3,000	-6.45062	-6.11826	-6.11826	last
MRPT-S + copula (joint sampling)	3,500	-6.45091	-6.11857	-6.11845	best (unchanged)

Notes. The original runs selected `best.pt` on the first 3,840 rows of the instrument-ordered 2023 validation index (one instrument; `val_max_batches` × batch 64). Every saved checkpoint was re-scored (m4mm, M4 Pro) on the same subset (reproducing the saved value exactly) and on a stratified sample of 50,000 origins across the 240 pooled instruments (seed 20261105). 0 of the five main-study models and 8 of the fifteen 4k ablation models would select the other checkpoint on the stratified subset; the stratified NLL differences are of order 10^{-4} and no ablation was re-predicted with the other checkpoint (not in the plan). The two subsets differ by ≈ 0.33 in level because the first is a single instrument. Source: `results/metrics/validation_rescore.csv`.

Table 4. Compute ledger by experiment: training wall-clock of every trained model of the study and of the review (prediction excluded)

Experiment	Runs	Nodes	Chips	GPU-hours
Main study (five models, 12k steps)	5	m3u96b	Apple M3 Ultra	2.6
Ablations (15 Small runs incl. copula head, 4k)	15	m3u96b	Apple M3 Ultra	3.4
Seed study of the H1 twins (seeds 2–5, 12k)	8	M4M64c, MacStudnpierre4, MacStudnpierre6	Apple M4 Max	3.3
Learning curve (continuations to 24k/36k)	4	M5M36, MacStudnpierre2	Apple M3 Ultra, Apple M5 Max	1.4
H2 transfer run (300 instruments, 12k)	1	m3u96b	Apple M3 Ultra	0.5
H3 control ablations (4 runs, 4k)	4	M5M36, MacStudnpierre2	Apple M3 Ultra, Apple M5 Max	0.9
Total	37			12.1

Notes. Sum of the training-loop wall-clock (`train_summary.json`, periodic validation included, prediction passes excluded) over the runs of each experiment; the baseline study (main + ablations, all on M3U96b) accounts for 6.0 of the 12.1 GPU-hours. Node and chip of every run, steps, parameters and throughput are in Table B.42; prediction and CPU re-scoring passes are in `results/metrics/review_runs*.jsonl`. Throughput is not comparable across chips; every comparison between models is at matched optimiser steps.

day); GARCH(1,1)-t on 5-minute training returns per instrument with the analytic variance term structure (Bollerslev, 1986; Hansen and Lunde, 2005); ridge, elastic net (Zou and Hastie, 2005) and LightGBM (Ke et al., 2017) on nineteen origin features, predicting the scaled cumulative return and the log scaled realised variance. The four log-variance regressions (HAR-RV, ridge, elastic net, LightGBM) are retransformed with one consistent, estimated correction: the predicted variance is $\exp(\hat{f})$ times the Duan smearing factor $\bar{c} = \text{mean}(\exp(\hat{\varepsilon}))$ of the training residuals (1.10–1.86 across models and horizons, largest at five minutes), and the Student-t degrees of freedom are fitted by the method of moments on returns standardised by the corrected variance (4.00–7.14; Tables A.1 and A.2). The correction matters (Table 5; fitted factors in Table A.2, Appendix): under a lognormal mean correction applied to the prediction but not to the residuals used for ν, HAR-RV's 50 % intervals covered 60 % of five-minute outcomes and the tree and linear regressions' 50 % intervals 37 %; with the consistent correction they cover 51 % and 48 %, HAR-RV's five-minute scaled CRPS changes by -1.76 % (0.4615 to 0.4534) and ridge's by -1.57 %, which is the whole of the five-minute lead that the sequence models hold over an inconsistently constructed HAR-RV; the corrected contrast is reported in Section 7.2. LightGBM's five-minute variance model remains mis-specified after the correction (smearing factor 1.863, 50 % coverage 74 %): a share 0.053 % of its five-minute variance forecasts lies below one hundredth of $\sigma_t^2 H$, which dominates its QLIKE (Table B.20); its five-minute row is a property of that model, not of the comparison. Neural, on the one-minute branch only: an MLP on the last sixty bars and a dilated temporal convolutional network (Bai et al., 2018; Moreno-Pino and Zohren, 2022), both with the same Student-t head and loss as the main model, and the one-minute-only MRPT itself, the matched-architecture patch-Transformer control (Nie et al., 2023): same width, depth, head and optimiser steps, with 7.4 % fewer parameters (3,200,259 against 3,455,619) and 79 rather than 339 tokens per step (1,088 against 1,759 s of training wall-clock). Equal steps rather than equal parameters or equal compute define the matched budget; the asymmetry favours the null, since the richer model has more parameters and more tokens per step.

A scalar recalibration accompanies every model as a companion, not a replacement: a scale factor c (and, for the parametric models, ν) is chosen on 100,000 stratified 2023 validation origins by CRPS and applied unchanged in 2024 and 2025. The neural models are the least calibrated of the set once the baselines are corrected (50 % coverage 55–59 % in 2024, scale factors below one) and gain a few tenths of a percent of CRPS on validation from the rescaling; the corrected HAR-RV is already calibrated ($c = 0.94$ to 0.92, with gains an order of magnitude smaller; Table 11). The recalibrated contrasts are reported next to the raw ones wherever the five-minute ranking is at stake (Table 11).

Pretrained foundation models (Ansari et al., 2024; Das et al., 2024; Shi et al., 2026) are not evaluated: their weights were not available offline on the compute node, their tokenisers assume a single real-valued series without missingness masks or multi-resolution context, their terms were not reviewed within this project, and the published evidence on financial data is mixed and points to fine-tuning rather than zero-shot use (Rahimikia et al., 2025; Brini, 2026; Fu et al., 2025). The matched-budget question asked here is a different question from the one a zero-shot evaluation answers; their absence is a stated limitation, not evidence about them. iTransformer (Liu et al., 2024) is not implemented. EGARCH forecasts beyond one step require simulation and were too slow for the number of evaluation origins. Table A.1 lists the specifications.

Table 5. Reconciliation of the HAR/ridge/elastic-net/LightGBM baselines: v1.0 construction versus the consistent retransformation (Duan smearing, ν on corrected residuals, no log H column)

Model	$H = 5$		$H = 15$		$H = 30$		$H = 60$	
	scaled CRPS	cov. 50	scaled CRPS	cov. 50	scaled CRPS	cov. 50	scaled CRPS	cov. 50
Panel A. 2024 development test, pooled instruments: scaled CRPS (before → after) and 50 % coverage (before → after)								
HAR-RV	0.4615 → 0.4534	0.60 → 0.51	0.4403 → 0.4404	0.50 → 0.51	0.4295 → 0.4295	0.48 → 0.50	0.4166 → 0.4167	0.49 → 0.50
Ridge	0.4603 → 0.4530	0.37 → 0.48	0.4405 → 0.4405	0.47 → 0.51	0.4297 → 0.4297	0.48 → 0.51	0.4167 → 0.4170	0.49 → 0.51
Elastic net	0.4603 → 0.4531	0.37 → 0.48	0.4406 → 0.4405	0.47 → 0.51	0.4296 → 0.4297	0.48 → 0.51	0.4167 → 0.4170	0.49 → 0.52
LightGBM	0.4591 → 0.4974	0.38 → 0.74	0.4415 → 0.4400	0.44 → 0.48	0.4293 → 0.4295	0.49 → 0.51	0.4166 → 0.4169	0.50 → 0.52
Panel B. 2025 final period, pooled instruments: scaled CRPS (before → after) and 50 % coverage (before → after)								
HAR-RV	0.4674 → 0.4597	0.60 → 0.50	0.4450 → 0.4451	0.50 → 0.51	0.4352 → 0.4352	0.48 → 0.49	0.4232 → 0.4233	0.49 → 0.50
Ridge	0.4663 → 0.4591	0.37 → 0.48	0.4449 → 0.4449	0.47 → 0.52	0.4350 → 0.4351	0.48 → 0.51	0.4232 → 0.4235	0.49 → 0.51
Elastic net	0.4663 → 0.4592	0.37 → 0.48	0.4449 → 0.4449	0.47 → 0.52	0.4351 → 0.4351	0.48 → 0.51	0.4232 → 0.4235	0.49 → 0.51
LightGBM	0.4655 → 0.5036	0.38 → 0.74	0.4457 → 0.4443	0.44 → 0.48	0.4347 → 0.4348	0.48 → 0.51	0.4232 → 0.4235	0.50 → 0.52

Notes. 'Before' = the v1.0 baselines (tag `review-baseline-20261103`: HAR variance $\exp(A\beta + \frac{1}{2}s^2)$ with ν from residuals standardised by $\exp(A\beta)$; trees $\exp(\hat{f})\sigma^2$ with no correction); 'after' = the same regressions with one estimated retransformation for every model, $\hat{c}_{\text{sm}} = \overline{\exp(\text{resid})}$ applied to $\exp(A\beta)$ (Duan 1983), ν by the method of moments on the returns standardised by the corrected variance, and the redundant log H column removed (Table A.2 lists the fitted factors). Everything else — training period, regressors, evaluation origins — is identical; the corrected rows are the ones used in every table of the paper. Scaled CRPS = CRPS$/(\sigma_t\sqrt{H})$; cov. 50 = empirical coverage of the central 50 % interval (nominal 0.50). The LightGBM row at $H = 5$ remains mis-specified after the correction (over-dispersed; see the variance-floor diagnostics in the QLIKE table). Held-out rows are in the source file. Source: `results/metrics/recal/baseline_fix_delta.csv`.

6.2. Scores and conventions

The primary scores are the sample-based CRPS of the cumulative return at each horizon (200 draws, energy-form estimator), the weighted interval score (Bracher et al., 2021) over eleven central intervals, the mean pinball loss at seven quantile levels, empirical coverage of the 50/80/95 %

central intervals, PIT dispersion and Kolmogorov–Smirnov statistics (Gneiting et al., 2007), and sharpness (90 % interval width). For parametric forecasts the closed-form Student-t CRPS (Jordan et al., 2019) and the log score are also computed; the log score is reported only where comparable. Secondary: absolute and squared errors of the cumulative return, QLIKE on the realised variance with positive forecasts (Patton, 2011), the Brier score of the tail event, directional accuracy against the zero-drift reference, and compute cost. Path-level scores (energy and variogram, $p = 0.5$; Scheuerer and Hamill, 2015) and CRPS of maximum drawdown and realised volatility are computed for the Stage-B comparison. Two CRPS conventions exist and the paper uses one of them for every number it states: *scaled* CRPS, $\mathrm{CRPS}/(\sigma_t\sqrt{H})$, which makes instruments and horizons comparable and is the convention of every level table, every relative contrast, every cluster statistic and every figure in the main text, unless a table note says otherwise. The unscaled CRPS, which weights high-volatility origins and days more heavily, is reported in companion tables in the appendix with its own contrasts; no ranking in the main text changes between the two conventions (the largest rank change among the main 2024 models is 3 positions and no contrast against the random walk changes sign), but individual relative differences do, so a number from one convention is never quoted next to a level from the other. Relative changes are $100 \times (\bar{L}_{\mathrm{model}} - \bar{L}_{\mathrm{ref}})/\bar{L}_{\mathrm{ref}}$, negative when the model is better; QLIKE, whose mean is negative, is compared in levels only.

6.3. Inference

All intervals are trading-day cluster bootstraps (1,000 resamples; all instruments and origins of one session form one cluster), and every pairwise contrast reports the day-cluster t as its headline statistic. A Diebold–Mariano test on the 251 daily mean loss differentials with a Newey–West variance (five lags) accompanies it in the result files and in the table notes. The origin-level Diebold–Mariano statistic with a Bartlett-kernel HAC variance whose lag equals the overlap of consecutive evaluation windows times the number of instruments is also recorded; its lag spans fifteen to sixty calendar minutes of a day ordered by instrument, so it does not absorb day-level common volatility shocks and overstates the evidence—its absolute value exceeds the day-cluster t by a median factor of 1.16 over the 720 CRPS contrasts (both conventions) of the pooled, held-out and recalibrated pairwise files of the main and final studies, in 90 % of them (Table B.8, Appendix, whose note names the files)—and it is never the basis of a verdict. Holm adjustments are applied within one family per evaluation split and loss, whose size is stated in each table note (60 tests for the 2024 pooled CRPS family, 60 for 2025, 68 for the ablations, 28 for each second-held-out family); Benjamini–Hochberg values are in the result files. A model confidence set at level 0.10 (Hansen et al., 2011) over the main models is computed per horizon from a moving-block bootstrap of the daily mean losses (block 5 days, with block 20 as a check). Models are frozen through the evaluation years, so the comparisons are of the fixed-scheme kind and the nested twins need no nested-model correction. Subgroup results (asset

class, dollar-volume and volatility tercile, sector, time of day, year, stress days defined by the top decile of cross-sectional σ_t) use the same machinery, and the stress-day calibration contrasts of Hypothesis 3 are differences of absolute coverage gaps between two models with paired day-cluster intervals. Economic evaluation (Hypothesis 5) trades the sign of the predictive median of each of the 7 location models (MRPT-S, MRPT-B, MLP, TCN, ridge, elastic net, LightGBM) only when its magnitude exceeds k times an assumed round-trip cost ($k \in \{1, 2\}$; costs of 0/1/2.5/5/10 basis points), enters one minute after the origin so that the first future minute is excluded from the realised return, and reports gross and net returns per trade with day-clustered intervals; with 280 cells per evaluation set the expected number of intervals excluding zero by chance at the 5 % level is about 11.2, and a cell counts as evidence only if it is positive in both 2024 and 2025. It is a sensitivity analysis, not a strategy.

6.4. Ablations, controls and falsification

At the Small scale on the development test, each trained for 4,000 steps with the same seed and compared with a reference trained under the same budget: resolutions (1m only, 1m+5m, 1m+5m+30m), fusion (shared vs cross-attention), head (Gaussian vs Student-t), regime embedding (none, statistics, learned), feature sets (returns only, OHLCV without relative volume, time only), ticker embedding, two further seeds of the reference, shuffled minute order inside the one-minute window at training and evaluation (whole rows, and price/volume columns only with the calendar columns in place: if the model only used marginal statistics, shuffling would not hurt), and the copula head against independent sampling. Three controls were added for the hypotheses that the ablations above cannot identify: for Hypothesis 3, the global token is reduced to $\log \sigma_t$ and the calendar and the model is trained with and without the regime statistics in the regime embedding, so that the two runs differ only in whether the statistics enter at all; a time-only variant with the same reduced global token measures what the calendar and the trailing scale alone achieve; and the price/volume-only shuffle isolates the order of the bars from their provenance. For Hypothesis 1, both twins are retrained with four further seeds at the 12,000-step budget and continued from their checkpoints to 24,000 and 36,000 steps under the same early-stopping rule; for Hypothesis 2, MRPT-S is trained once on all 300 instruments and compared with the zero-shot checkpoint on the identical held-out origins. Zero-shot evaluation on the 60 held-out instruments, on the 66 availability-excluded instruments and on the untouched 2025 period completes the protocol.

7. Results

7.1. Pilot

The pilot (20 instruments, September–December 2022, test on the last ten sessions, 4659 origins) validated the pipeline from export to LaTeX table. Its scores are in Table B.40 (Appendix) and are not a preview of the main study: on that small sample the neural models were within 0.32 % of one another in scaled CRPS, LightGBM was worse than the random walk at every horizon and HAR-RV at five minutes, and the pilot used a different evaluation stride (10 minutes) and number of draws (256). Nothing was tuned on it.

7.2. Development test (2024)

Table 6 reports the scaled CRPS of the cumulative return on the 986,468 evaluation origins of 2024 (240 pooled instruments, 16.4 origins per instrument-session on average, at most 20 per session, 251 sessions), with trading-day cluster-bootstrap intervals and the five-seed band of the two H1 twins; Table 7 the relative change against the Gaussian random walk with its day-cluster t; Table 9 the model confidence sets. Figure 1 summarises the first two.

The Gaussian random walk with a trailing one-minute scale is beaten at every horizon by every model except LightGBM at five minutes, by 2.50 % (seasonal volatility, five minutes) to 4.14 % (HAR-RV, sixty minutes) of scaled CRPS, with day-cluster t statistics of at least 25.5 in absolute value in every such cell; the corrected LightGBM is +5.86 % worse than the random walk at five minutes (Section 6.1) and in line with the other regressions from fifteen minutes on. Most of this gain is the heavy tail and the scale, not minute-level structure: the Student-t random walk, whose scale and degrees of freedom were chosen on 2023, is itself -2.08 % to -3.38 % of scaled CRPS from the Gaussian one (relative change, negative = better), and the MRPT-S contrast with it is -1.54 % at five minutes, -1.10 % at fifteen, -0.72 % at thirty and -0.30 % at sixty (day-cluster t -25.6, -15.8, -8.9 and -3.2). The information in the minute stream concerns the dispersion of the next hour; its directional content is examined in Section 7.5.

At five minutes the neural family (MRPT-S, MRPT-B, TCN, MLP) and the consistently constructed regressions are within a few hundredths of a percent of one another (Table 8): MRPT-S against HAR-RV is -0.07 % [-0.13, -0.01] (day-cluster t -2.2, Holm $p = 0.148$ in the family of 60), against ridge +0.01 % and against elastic net +0.00 %. A lead of the sequence models at five minutes appears only after every model is rescaled on 2023: MRPT-S_{recal} against HAR-RV_{recal} is -0.40 % (day-cluster t -14.6), against $ridge_{recal}$ -0.13 % (t -3.6, Holm $p = 0.00481$), with the same signs in 2025 (Tables 11 and B.1); it is a lead of a few tenths of a percent that the raw comparison does not show, and we report it with that qualification. From fifteen minutes on the feature-based volatility

regressions—HAR-RV, ridge, elastic net and LightGBM—are the best models: HAR-RV against MRPT-S is 0.4167 against 0.4187 at sixty minutes, a difference of +0.49 % [+0.34, +0.62] (t = 6.6), and 0.21–0.49 % over the three longer horizons; the recalibrated contrast at sixty minutes is -0.02 % (t -0.5), so the rescaling closes the gap at the longest horizon without reversing it. The model confidence sets (Table 9; level 0.10, block 5 days, daily mean unscaled CRPS) say the same with one exhibit per horizon: the Gaussian random walk is outside the set at every horizon, MRPT-S is a member at five minutes only, and from fifteen minutes on the sets contain only variance regressions. The 90 % model confidence set on daily mean CRPS (block length 5) contains at H = 5 Ridge, Elastic net, MRPT-S (1m+5m+30m); at H = 15 LightGBM alone; at H = 30 HAR-RV, Ridge, Elastic net, LightGBM; at H = 60 HAR-RV alone. MRPT-S is better than GARCH(1,1)-t, EWMA, the TCN and the MLP at every horizon (contrasts from -0.16 % to -1.06 %, all with Holm $p < 0.01$).

Hypothesis 1 is the central test. On the single main run the contrast between the multi-resolution MRPT-S and its one-minute-only twin is -0.04 % at five minutes and -0.13 % at sixty (day-cluster t -5.1 and -5.1). One run cannot separate an architectural effect from training noise, so both twins were retrained with four further seeds (Table 10; levels per seed in Table B.9, Appendix). Over the five seed pairs the contrast is negative in every pair at every horizon, with mean -0.081 % (range -0.15 % to -0.04 %) at five minutes, -0.117 % at fifteen, -0.152 % at thirty and -0.214 % (range -0.41 % to -0.13 %) at sixty; the t statistic across seeds is -4.53, -3.18, -3.20 and -4.05 (four degrees of freedom) and the day-cluster t on the seed-averaged differential -13.6, -14.0, -13.0 and -11.2. By the pre-stated seed-aware rule of Section 2, Hypothesis 1 is supported in 2024 on scaled CRPS: the multi-resolution context lowers scaled CRPS at 4 of the four horizons (2 of four on unscaled CRPS, where the seed-to-seed dispersion is larger), by -0.081 % to -0.214 % on average over 5 seeds, with every seed pair negative at every horizon. In the untouched 2025 period the same eight seed checkpoints, scored once, give a mean contrast of -0.116 % to -0.046 % (at least 4 of the 5 seed pairs negative at every horizon, up to 1 positive at one horizon): the mean keeps the 2024 sign at every horizon but the seed-to-seed dispersion is too large for the rule ($|t_{\text{seed}}|$ of 1.9 to 2.3 against 2.78, day-cluster t of 6.5 to 7.9), so the 2024 verdict is not replicated by the pre-stated rule in 2025. The effect is small on every yardstick fixed in advance: it is 1.5–3.4 times the seed half-band of the full model (0.05–0.06 %, the smallest effect size of interest; 0.07–0.14 % for the twin) and 3.0–4.5 times the minimum detectable effect of Table B.7, while the per-seed contrasts span -0.15 % to -0.04 % at five minutes and -0.41 % to -0.13 % at sixty, so that a single seed can misstate it by a factor of two or more. It is also an order of magnitude smaller than the gap between any neural model and the random walk and smaller than the gap between the neural models and HAR-RV at sixty minutes. We therefore read Hypothesis 1 as supported by its pre-registered test in 2024 at a magnitude that is statistically detectable but economically negligible: at a matched 12,000-step budget (1.31 passes, no hyper-parameter search), five days of 5-minute aggregates and twenty days

of 30-minute aggregates add a few hundredths to a few tenths of a percent of CRPS to what the last 390 one-minute bars already contain. Whether that remains true at larger budgets is examined in Section 8, and the untouched 2025 year, where the effect does not recur (Section 7.6), is the caveat that qualifies the verdict. Size does not matter at this budget either: the 20.7-million-parameter Base model and the 3.5-million-parameter Small model differ by +0.05 % at sixty minutes (Holm $p = 0.148$), with Base the worse of the two, and Base's validation likelihood is also the worse at 12,000 steps; the Base model was not trained beyond that budget, so the statement is conditional on it.

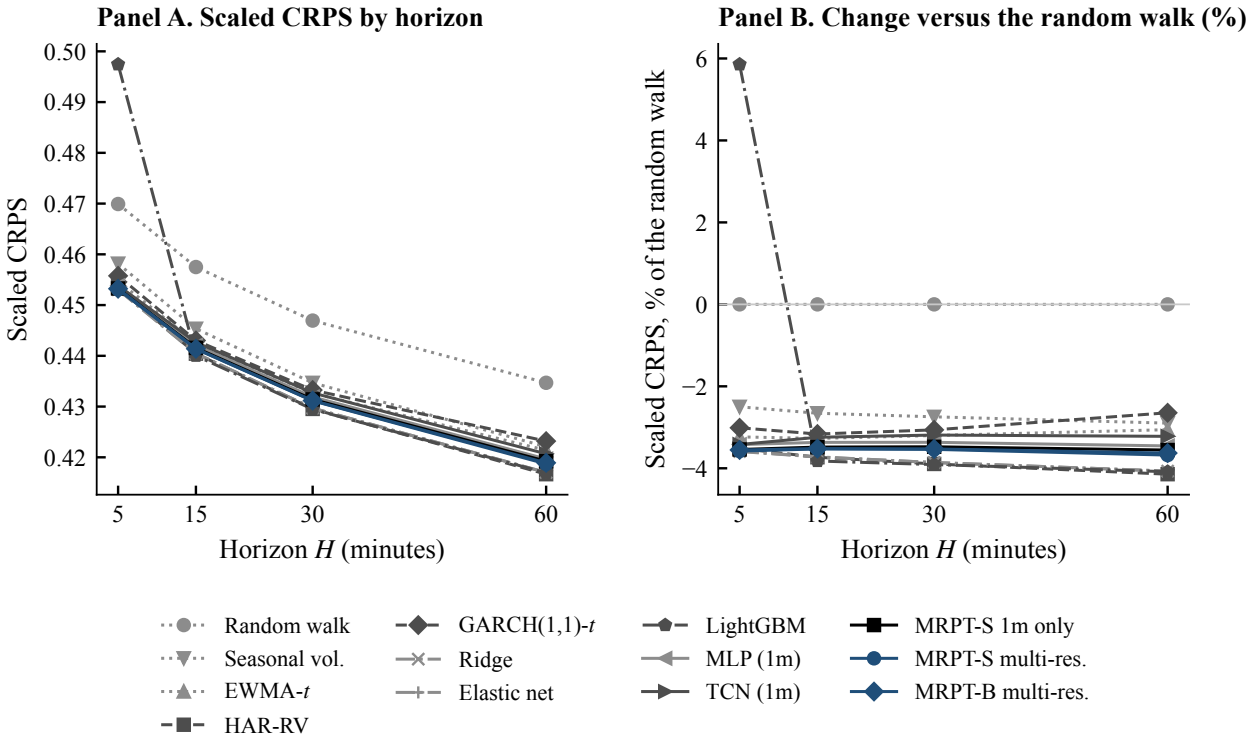


Figure 1. Development test 2024 (240 instruments, 986,468 origins, 251 sessions). Panel A: scaled CRPS of the cumulative-return forecast by horizon with trading-day cluster-bootstrap intervals. Panel B: relative change in scaled CRPS versus the Gaussian random walk with a trailing scale; negative is better. HAR-RV, ridge, elastic net and LightGBM are the consistently corrected baselines of Section 6.1. Source: `results/metrics/main/leaderboard_test.csv`.

7.3. Calibration

Table 12 and Figure 2 show empirical coverage of the 50/80/95 % central intervals and the PIT Kolmogorov–Smirnov statistic. The Gaussian random walk is over-dispersed everywhere (50 % intervals cover 68 % of outcomes at sixty minutes); the Student-t random walk calibrated on 2023 is nominal in the centre (51 %). The Student-t neural models and GARCH/EWMA are nearly nominal

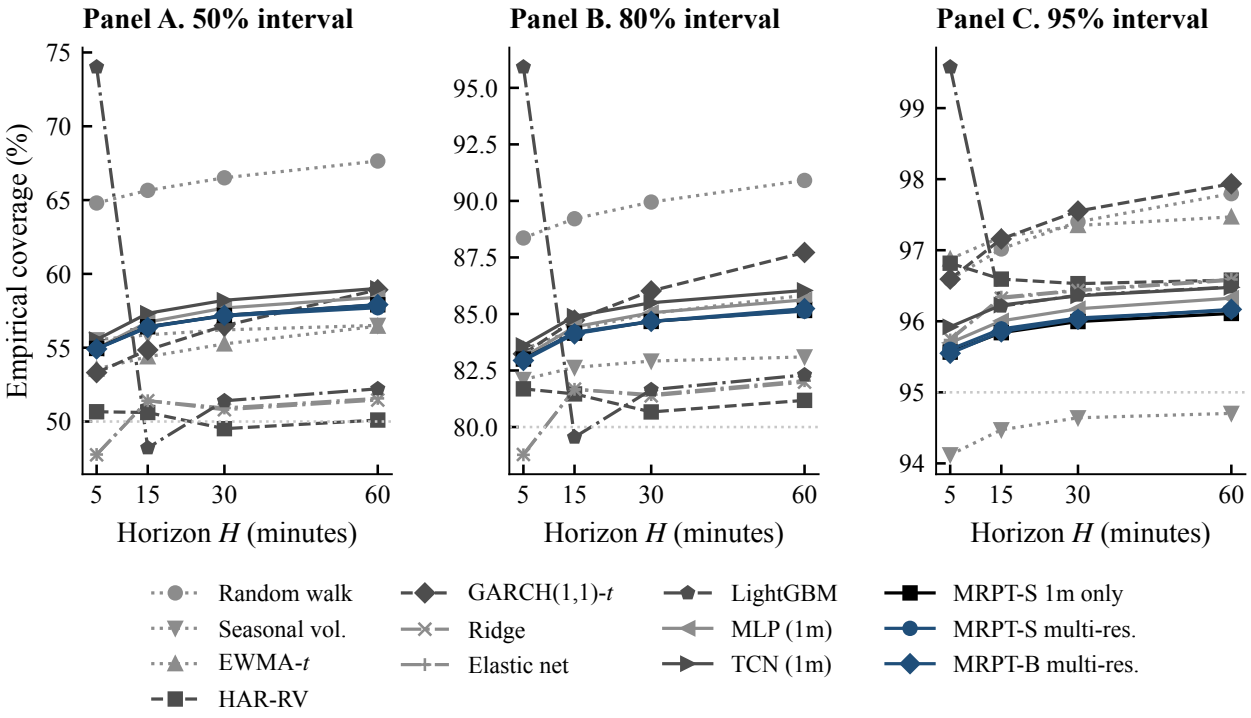


Figure 2. Empirical coverage of the central 50 %, 80 % and 95 % prediction intervals by horizon (dotted line = nominal), development test 2024, 986,468 origins; the random walk is the Gaussian reference.

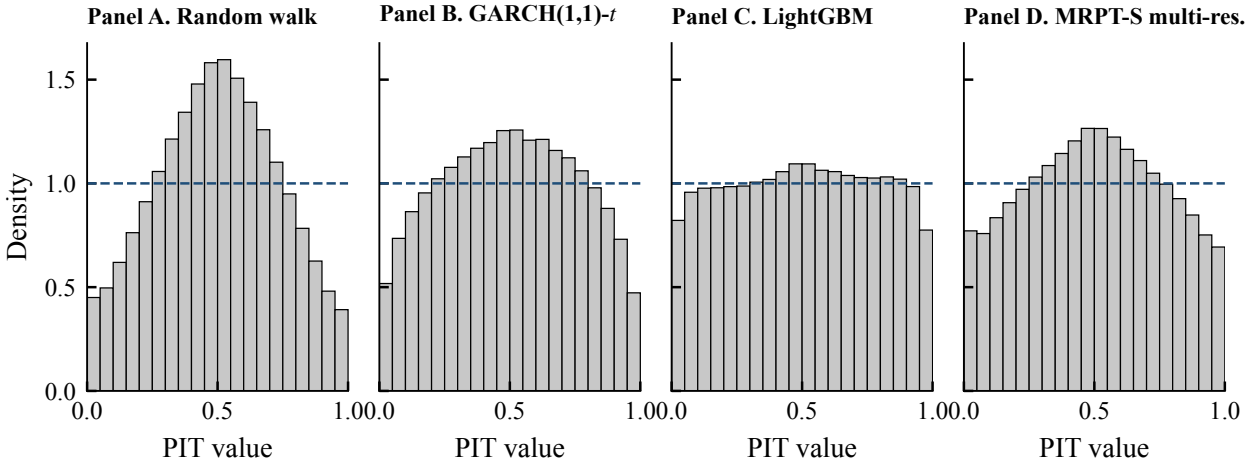


Figure 3. PIT histograms at $H = 60$ for the random walk, GARCH(1,1)-t, LightGBM and MRPT-S (dashed line = uniform density), development test 2024.

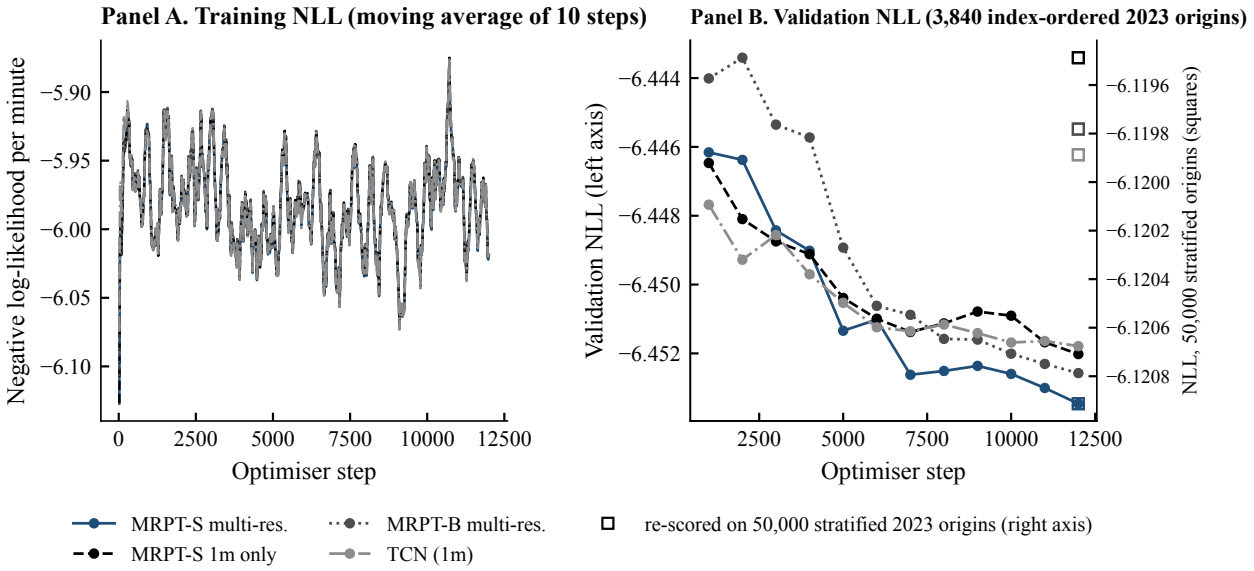


Figure 4. Training (moving average) and validation negative log-likelihood per optimiser step, main study. Panel B shows the likelihood on the 3,840-origin single-instrument subset used for selection and, as hollow markers on the right axis, the re-score of the saved best and last checkpoints on the stratified 50,000-origin 2023 sample.

Table 6. Main study: mean scaled CRPS of the cumulative-return forecast by horizon (minutes)

Model	Parameters	$H = 5$	$H = 15$	$H = 30$	$H = 60$
Panel A. Models as specified (baselines) or trained (neural models)					
Random walk (trailing RMS)	–	0.4699 [0.4648, 0.4754]	0.4575 [0.4515, 0.4639]	0.4469 [0.4412, 0.4538]	0.4347 [0.4271, 0.4441]
Seasonal volatility	–	0.4582 [0.4527, 0.4641]	0.4453 [0.4390, 0.4519]	0.4347 [0.4285, 0.4418]	0.4221 [0.4139, 0.4320]
EWMA-t	–	0.4547 [0.4492, 0.4606]	0.4425 [0.4363, 0.4491]	0.4326 [0.4264, 0.4396]	0.4214 [0.4133, 0.4314]
HAR-RV	–	0.4534 [0.4478, 0.4594]	0.4404 [0.4339, 0.4472]	0.4295 [0.4231, 0.4368]	0.4167 [0.4081, 0.4271]
GARCH(1,1)-t	–	0.4558 [0.4502, 0.4616]	0.4430 [0.4367, 0.4497]	0.4333 [0.4271, 0.4403]	0.4232 [0.4153, 0.4328]
Ridge	–	0.4530 [0.4474, 0.4591]	0.4405 [0.4340, 0.4473]	0.4297 [0.4233, 0.4370]	0.4170 [0.4086, 0.4273]
Elastic net	–	0.4531 [0.4474, 0.4592]	0.4405 [0.4340, 0.4473]	0.4297 [0.4233, 0.4369]	0.4170 [0.4086, 0.4273]
LightGBM	–	0.4974 [0.4925, 0.5026]	0.4400 [0.4335, 0.4468]	0.4295 [0.4230, 0.4366]	0.4169 [0.4086, 0.4271]
MLP (1m)	301,827	0.4538 [0.4484, 0.4598]	0.4421 [0.4357, 0.4486]	0.4319 [0.4258, 0.4388]	0.4197 [0.4116, 0.4291]
TCN (1m)	231,459	0.4538 [0.4484, 0.4597]	0.4427 [0.4364, 0.4491]	0.4327 [0.4265, 0.4396]	0.4207 [0.4126, 0.4303]
MRPT-S (1m+5m+30m)	3,455,619	0.4531 [0.4476, 0.4590]	0.4414 [0.4351, 0.4479]	0.4311 [0.4249, 0.4381]	0.4187 [0.4105, 0.4284]
MRPT-S (1m only)	3,200,259	0.4533 [0.4478, 0.4592]	0.4416 [0.4352, 0.4481]	0.4314 [0.4251, 0.4382]	0.4192 [0.4110, 0.4290]
MRPT-B (1m+5m+30m)	20,718,339	0.4532 [0.4478, 0.4591]	0.4414 [0.4351, 0.4479]	0.4312 [0.4249, 0.4381]	0.4189 [0.4107, 0.4287]
Panel B. Recalibrated on the 2023 validation origins (scalar scale c and ν per model and horizon; Section 6)					
Random walk, Student-t (2023-calibrated)	–	0.4602 [0.4545, 0.4663]	0.4463 [0.4396, 0.4533]	0.4343 [0.4277, 0.4419]	0.4200 [0.4113, 0.4305]

Notes. development test 2024, 986468 evaluation origins over 251 sessions. Brackets: 95 % trading-day cluster-bootstrap intervals (all instruments and origins of one session form one cluster). Convention: scaled CRPS, i.e. CRPS divided by $\sigma_t\sqrt{H}$. Parameters: trainable parameters of the trained models (`train_summary.json`); baselines have none in this sense.

Table 7. Main study: scaled CRPS relative to Random walk (trailing RMS)

Model	$H = 5$	$H = 15$	$H = 30$	$H = 60$
Panel A. Models as specified (baselines) or trained (neural models)				
Seasonal volatility	-2.50 [-30.9]	-2.66 [-32.8]	-2.74 [-30.3]	-2.90 [-25.5]
EWMA-t	-3.23 [-42.2]	-3.27 [-40.4]	-3.20 [-34.0]	-3.06 [-25.9]
HAR-RV	-3.51 [-43.7]	-3.73 [-42.7]	-3.89 [-37.0]	-4.14 [-30.0]
GARCH(1,1)-t	-3.01 [-40.4]	-3.16 [-40.3]	-3.06 [-35.4]	-2.65 [-26.0]
Ridge	-3.59 [-39.3]	-3.72 [-43.9]	-3.86 [-39.0]	-4.07 [-32.1]
Elastic net	-3.58 [-39.3]	-3.72 [-43.9]	-3.86 [-39.2]	-4.07 [-32.1]
LightGBM	+5.86 [63.1]	-3.82 [-41.6]	-3.91 [-40.8]	-4.08 [-34.4]
MLP (1m)	-3.42 [-49.5]	-3.37 [-45.9]	-3.37 [-41.3]	-3.45 [-34.6]
TCN (1m)	-3.42 [-49.7]	-3.24 [-44.7]	-3.19 [-39.0]	-3.22 [-32.6]
MRPT-S (1m+5m+30m)	-3.58 [-49.2]	-3.53 [-46.0]	-3.54 [-40.5]	-3.68 [-34.3]
MRPT-S (1m only)	-3.54 [-49.5]	-3.48 [-45.9]	-3.48 [-40.1]	-3.55 [-34.1]
MRPT-B (1m+5m+30m)	-3.55 [-49.7]	-3.52 [-46.5]	-3.53 [-41.1]	-3.63 [-34.2]
Panel B. Recalibrated on the 2023 validation origins (scalar scale c and ν per model and horizon; Section 6)				
Random walk, Student-t (2023-calibrated)	-2.08 [-31.3]	-2.45 [-28.6]	-2.84 [-28.0]	-3.38 [-25.0]
Elastic net (recalibrated)	-3.61 [-39.5]	-3.77 [-42.0]	-3.89 [-38.3]	-4.11 [-31.0]
HAR-RV (recalibrated)	-3.36 [-43.9]	-3.69 [-44.3]	-3.86 [-39.3]	-4.13 [-30.7]
LightGBM (recalibrated)	+5.02 [54.9]	-3.83 [-45.2]	-3.94 [-39.8]	-4.14 [-33.1]
MRPT-B (1m+5m+30m) (recalibrated)	-3.71 [-49.4]	-3.81 [-44.6]	-3.92 [-38.4]	-4.13 [-31.5]
MRPT-S (1m only) (recalibrated)	-3.70 [-48.4]	-3.77 [-44.4]	-3.87 [-39.0]	-4.05 [-32.1]
MRPT-S (1m+5m+30m) (recalibrated)	-3.74 [-49.2]	-3.82 [-45.7]	-3.93 [-39.2]	-4.16 [-32.3]
Ridge (recalibrated)	-3.62 [-39.7]	-3.77 [-42.1]	-3.89 [-38.3]	-4.11 [-31.0]

Notes. development test 2024, 986468 evaluation origins over 251 sessions. Cells: percentage change in mean scaled CRPS, i.e. CRPS divided by $\sigma_t\sqrt{H}$ versus Random walk (trailing RMS), i.e. $100 \times (\bar{L}_{\text{model}} - \bar{L}_{\text{ref}})/\bar{L}_{\text{ref}}$ (negative = better); in brackets the day-cluster t-statistic of the mean loss differential (its trading-day bootstrap standard error, 1,000 resamples; every origin of one session forms one cluster). Convention: changes and tests on scaled CRPS; these contrasts against the random walk are not part of a Holm family (the pre-registered pairwise contrasts with Holm adjustment are in the pairwise tables). Panel B lists the same forecasts after the scalar recalibration of Section 6 and the Student-t random walk whose scale and degrees of freedom were fitted on 2023; it is reported for completeness and is not used for any ranking in the text. Intervals are in the result file.

at 95 % but over-dispersed in the centre (50 % coverage 55–59 % for the neural models): their degrees of freedom and scale produce tails that are right and a body that is too wide, which the scalar recalibration of Section 6.1 corrects toward nominal with the CRPS gains reported in Table 11. With the consistent retransformation HAR-RV, ridge and elastic net are close to nominal at every horizon (HAR-RV 50 % coverage 51 % at five minutes and 50 % at sixty), and LightGBM is over-dispersed at five minutes (74 %) and nominal from fifteen minutes on. The scores and the calibration therefore agree: at every horizon a well-specified variance regression with a kurtosis-matched Student-t is at least as well calibrated as the sequence models, and from fifteen minutes on it is also sharper. The PIT histograms of Figure 3 show the same shapes at sixty minutes.

Table 8. Main study: pre-registered pairwise contrasts — percentage change in mean scaled CRPS, day-cluster t and Holm-adjusted significance

Contrast (first minus second, % of the second)	$H = 5$	$H = 15$	$H = 30$	$H = 60$
H1: MRPT-S (1m+5m+30m) vs MRPT-S (1m only)	-0.04 [-5.1]***	-0.05 [-5.1]***	-0.07 [-4.8]***	-0.13 [-5.1]***
Size: MRPT-B vs MRPT-S	+0.03 [2.7]**	+0.01 [0.9]	+0.01 [0.9]	+0.05 [2.2]
MRPT-S vs HAR-RV	-0.07 [-2.2]	+0.21 [6.2]***	+0.37 [6.9]***	+0.49 [6.6]***
MRPT-S vs ridge	+0.01 [0.3]	+0.20 [6.2]***	+0.33 [7.2]***	+0.41 [6.1]***
MRPT-S vs elastic net	+0.00 [0.0]	+0.20 [6.1]***	+0.33 [7.3]***	+0.41 [6.1]***
MRPT-S vs LightGBM	-8.91 [-117.0]***	+0.31 [7.4]***	+0.39 [9.6]***	+0.42 [8.1]***
MRPT-S vs GARCH(1,1)-t	-0.59 [-19.4]***	-0.38 [-11.7]***	-0.49 [-14.4]***	-1.06 [-20.5]***
MRPT-S vs EWMA-t	-0.36 [-16.2]***	-0.26 [-11.5]***	-0.35 [-11.2]***	-0.64 [-12.4]***
MRPT-S vs seasonal volatility	-1.11 [-39.3]***	-0.90 [-33.9]***	-0.82 [-26.0]***	-0.80 [-18.6]***
MRPT-S vs TCN	-0.17 [-13.0]***	-0.29 [-21.1]***	-0.36 [-19.2]***	-0.47 [-15.8]***
MRPT-S vs MLP	-0.16 [-6.0]***	-0.17 [-5.7]***	-0.18 [-5.3]***	-0.23 [-5.0]***
MRPT-S vs random walk	-3.58 [-49.2]***	-3.53 [-46.0]***	-3.54 [-40.5]***	-3.68 [-34.3]***
MRPT-S vs Student-t random walk (2023-calibrated)	-1.54 [-25.6]***	-1.10 [-15.8]***	-0.72 [-8.9]***	-0.30 [-3.2]**
HAR-RV vs GARCH(1,1)-t	-0.52 [-19.7]***	-0.59 [-17.5]***	-0.86 [-17.0]***	-1.54 [-19.1]***
HAR-RV vs random walk	-3.51 [-43.7]***	-3.73 [-42.7]***	-3.89 [-37.0]***	-4.14 [-30.0]***

Notes. development test 2024, 986468 evaluation origins over 251 sessions. Cells: $100 \times (\bar{L}_{\text{first}} - \bar{L}_{\text{second}})/\bar{L}_{\text{second}}$ on scaled CRPS, i.e. CRPS divided by $\sigma_t\sqrt{H}$ (negative = the first model is better); in brackets the day-cluster t-statistic (mean loss differential over its trading-day bootstrap standard error, 1,000 resamples; every origin of one session is one cluster). Stars: Holm-adjusted p-value over the family of 60 tests (15 pairs × 4 horizons, adjusted on the day-cluster p): *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$. The HAC Diebold–Mariano t on the pooled series and the daily-mean DM t are in the result file (`results/metrics/main/pairwise_test.csv`); a contrast whose magnitude lies inside the five-seed band of Table 10 is not distinguishable from training randomness whatever its t.

Table 9. Main study: model confidence sets (MCS p-values; † = member of the 90 % set)

Model	CRPS				WIS			
	$H = 5$	$H = 15$	$H = 30$	$H = 60$	$H = 5$	$H = 15$	$H = 30$	$H = 60$
Random walk (trailing RMS)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Seasonal volatility	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
EWMA-t	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
HAR-RV	0.02	0.00	1.00†	1.00†	0.00	0.00	0.85†	1.00†
GARCH(1,1)-t	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Ridge	1.00†	0.00	0.42†	0.01	0.32†	0.00	0.17†	0.32†
Elastic net	0.20†	0.00	0.42†	0.01	0.16†	0.00	0.39†	0.32†
LightGBM	0.00	1.00†	0.42†	0.01	0.00	1.00†	1.00†	0.32†
MLP (1m)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TCN (1m)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MRPT-S (1m+5m+30m)	0.20†	0.00	0.00	0.00	1.00†	0.00	0.00	0.00
MRPT-S (1m only)	0.02	0.00	0.00	0.00	0.32†	0.00	0.00	0.00
MRPT-B (1m+5m+30m)	0.05	0.00	0.00	0.00	0.32†	0.00	0.00	0.00
Random walk, Student-t (2023-calibrated)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Notes. development test 2024, 986468 evaluation origins over 251 sessions. Model confidence set of Hansen, Lunde and Nason (2011) over the 14 models with a per-origin file, computed on the 251 daily mean losses (unscaled CRPS and WIS of the cumulative return) with the $T_{\max}$ statistic and a moving-block bootstrap of the days (block length 5, 1,000 resamples; block 20 in the result file); a model is in the 90 % set (†) when its MCS p-value is at least 0.10; the Student-t random walk calibrated on 2023 (last row) is a reference, not a competitor. Source: `results/metrics/main/mcs_test.csv`.

Table 10. Seed study of the H1 twins (5 seeds each at 12,000 steps, 2024 development test; Panel D: the same checkpoints on the untouched 2025 period): seed envelope of the levels and seed-aware test of the multi-resolution contrast

	$H = 5$	$H = 15$	$H = 30$	$H = 60$
Panel A. Scaled CRPS per twin: five-seed mean [min, max] (half-band, % of the mean)				
MRPT-S (1m+5m+30m)	0.4529 [0.4526, 0.4531] (0.055 %)	0.4412 [0.4409, 0.4414] (0.052 %)	0.4311 [0.4309, 0.4314] (0.059 %)	0.4188 [0.4186, 0.4191] (0.062 %)
MRPT-S (1m only)	0.4532 [0.4529, 0.4535] (0.066 %)	0.4417 [0.4414, 0.4422] (0.093 %)	0.4317 [0.4314, 0.4324] (0.115 %)	0.4197 [0.4192, 0.4204] (0.137 %)
Panel B. H1 contrast per seed pair, % of the 1m-only scaled CRPS [day-cluster t]				
Seed 1 (main study)	-0.039 [-5.1]	-0.047 [-5.1]	-0.065 [-4.8]	-0.127 [-5.1]
Seed 2	-0.062 [-8.0]	-0.089 [-9.2]	-0.138 [-10.5]	-0.226 [-10.8]
Seed 3	-0.146 [-15.7]	-0.260 [-20.6]	-0.336 [-19.9]	-0.413 [-17.0]
Seed 4	-0.071 [-10.1]	-0.106 [-12.0]	-0.123 [-11.0]	-0.166 [-9.5]
Seed 5	-0.087 [-9.4]	-0.084 [-6.7]	-0.099 [-5.6]	-0.136 [-4.6]
Panel C. Seed-aware test of H1 (5 seed pairs, 4 df) and day-cluster test of the seed-averaged differential				
Mean contrast over seeds (%, scaled CRPS)	-0.081	-0.117	-0.152	-0.214
Seed range (min, max)	[-0.146, -0.039]	[-0.260, -0.047]	[-0.337, -0.065]	[-0.413, -0.127]
t_{seed} (4 df) [p]	-4.53 [0.011]	-3.17 [0.034]	-3.19 [0.033]	-4.05 [0.016]
t_{cl} on the seed-averaged differential	-13.6	-14.0	-13.0	-11.2
Detectable by the pre-stated rule	yes	yes	yes	yes
Same contrast on unscaled CRPS (%) [t_{seed}]	-0.088 [-3.54]	-0.132 [-2.47]	-0.168 [-2.33]	-0.231 [-2.83]
Panel D. The same seed checkpoints scored once on the untouched 2025 period (scaled CRPS): contrast per seed pair [day-cluster t] and seed-aware test				
Seed 1 (main study), 2025	+0.004 [0.4]	+0.022 [2.0]	+0.017 [1.2]	-0.005 [-0.2]
Seed 2, 2025	-0.021 [-2.3]	-0.045 [-3.8]	-0.073 [-4.9]	-0.142 [-6.3]
Seed 3, 2025	-0.115 [-12.4]	-0.200 [-14.2]	-0.242 [-11.8]	-0.293 [-11.9]
Seed 4, 2025	-0.048 [-6.3]	-0.093 [-9.5]	-0.104 [-7.9]	-0.121 [-6.6]
Seed 5, 2025	-0.049 [-5.0]	-0.034 [-2.8]	-0.021 [-1.3]	-0.015 [-0.6]
Mean contrast over seeds, 2025 (%)	-0.046	-0.070	-0.085	-0.116
Seed range (min, max), 2025	[-0.115, +0.004]	[-0.201, +0.022]	[-0.242, +0.017]	[-0.294, -0.005]
t_{seed} (4 df) [p], 2025	-2.29 [0.084]	-1.88 [0.134]	-1.90 [0.130]	-2.21 [0.092]
t_{cl} on the seed-averaged differential, 2025	-6.9	-7.9	-7.0	-6.5
Detectable by the pre-stated rule, 2025	no	no	no	no
Same contrast on unscaled CRPS (%) [t_{seed}], 2025	-0.055 [-2.03]	-0.081 [-1.83]	-0.096 [-1.67]	-0.121 [-1.79]

Notes. Seeds 2–5 of MRPT-S (1m+5m+30m) and of its one-minute-only twin were trained with the main-study configuration (12,000 steps, batch 64, early stopping on a seeded stratified subset of 50,000 2023 validation origins) on M4 Max nodes; seed 1 is the main-study run. Panel A: mean, minimum and maximum of the scaled CRPS over the 5 seeds and the half-band $(\max - \min)/(2\,\text{mean})$ in percent, the smallest effect size of interest used in the text. Panel B: the H1 contrast of each seed pair with its day-cluster t (1,000 trading-day bootstrap resamples). Panel C: t_{seed} = mean of the 5 per-seed differentials over their standard error across seeds (t with 4 degrees of freedom); t_{cl} = day-cluster t of the seed-averaged loss differential; the pre-stated rule for 'detectable' is |t_seed| > 2.78 (t, 4 df) and |t_cl| > 1.96. Convention: scaled CRPS (unscaled where stated). Panel D (revision package P3b): the eight seed checkpoints of seeds 2–5 were scored exactly once on the 2025 final origins with the main study's prediction path (no selection; seed 1 is the main study's 2025 run); the per-seed pairs come from `seeds/pairwise_final.csv` and the seed-aware statistics from `seed_contrast_final.csv`, the 2025 levels per seed are in Table B.10 (Appendix). Source: `results/metrics/seeds/seed_bands.csv`, `seed_contrast.csv`, `seeds/pairwise_test.csv` (per-seed pairs).

7.4. Subgroups and stress days

Splitting the MRPT-S versus HAR-RV contrast at sixty minutes by subgroup (relative change in unscaled CRPS, positive when HAR-RV is better),[2] HAR-RV's advantage is concentrated in low-dollar-volume stocks (+1.3 %), low-volatility instruments (+1.0 %), fixed-income ETFs (+1.7 %) and the stocks without a SIC code (+2.7 %); MRPT-S is better on broad-equity, sector and volatility-linked ETFs (-0.4 %, -0.3 % and -1.0 %) and the two are tied on high-dollar-volume instruments (day-cluster t -0.3). Afternoon origins favour HAR-RV (+1.0 %), morning origins less (+0.2 %). On stress days (top decile of cross-sectional σ_t, 24 sessions and 100,846 origins) the ordering is unchanged and every model's coverage moves by at most a few percentage points: no model breaks under stress, and none gains a special advantage there.

[2] `results/metrics/main/subgroups_mrpt_small_vs_har_rv_test.csv`, which carries both CRPS conventions in a `metric` column; the v1.1 draft quoted the unscaled values under a scaled label.

Table 11. Main study: scalar recalibration fitted on 2023 — scaled CRPS and 50 % coverage before and after

Model	$H = 5$		$H = 15$		$H = 30$		$H = 60$	
	orig.	recal.	orig.	recal.	orig.	recal.	orig.	recal.
Scaled CRPS								
HAR-RV	0.4534	0.4541	0.4404	0.4406	0.4295	0.4297	0.4167	0.4167
Ridge	0.4530	0.4529	0.4405	0.4402	0.4297	0.4295	0.4170	0.4168
Elastic net	0.4531	0.4530	0.4405	0.4402	0.4297	0.4295	0.4170	0.4168
LightGBM	0.4974	0.4935	0.4400	0.4400	0.4295	0.4294	0.4169	0.4167
MRPT-S (1m+5m+30m)	0.4531	0.4523	0.4414	0.4400	0.4311	0.4294	0.4187	0.4166
MRPT-S (1m only)	0.4533	0.4525	0.4416	0.4403	0.4314	0.4297	0.4192	0.4171
MRPT-B (1m+5m+30m)	0.4532	0.4525	0.4414	0.4401	0.4312	0.4294	0.4189	0.4168
Coverage of the central 50 % interval (nominal 0.50)								
HAR-RV	0.507	0.545	0.506	0.530	0.495	0.521	0.501	0.523
Ridge	0.477	0.496	0.514	0.509	0.508	0.513	0.515	0.518
Elastic net	0.478	0.496	0.514	0.509	0.509	0.513	0.516	0.518
LightGBM	0.740	0.742	0.482	0.519	0.514	0.516	0.522	0.519
MRPT-S (1m+5m+30m)	0.549	0.523	0.564	0.529	0.571	0.527	0.577	0.533
MRPT-S (1m only)	0.549	0.524	0.564	0.529	0.572	0.528	0.579	0.535
MRPT-B (1m+5m+30m)	0.549	0.524	0.564	0.529	0.572	0.528	0.579	0.535

Notes. development test 2024, 986468 evaluation origins over 251 sessions. 'orig.' = the model as trained/fitted; 'recal.' = the same forecasts with a scalar scale c (and, for the parametric baselines, degrees of freedom ν) chosen per model and horizon on a seeded stratified subset of 100,000 validation origins of 2023 to minimise mean CRPS (`results/metrics/recal/c_val.json`); the 2024 and 2025 scores are then computed once with the fitted constants (`scripts/recalibrate.py`, `scripts/review/rescore_scaled.py`). Convention: scaled CRPS. Source: the `_recal` rows of the leaderboard file.

Table 12. Main study: empirical 50/80/95% coverage and PIT KS statistic

Model	$H = 5$				$H = 15$				$H = 30$				$H = 60$			
	50	80	95	KS	50	80	95	KS	50	80	95	KS	50	80	95	KS
Panel A. Models as specified (baselines) or trained (neural models)																
Random walk (trailing RMS)	0.65	0.88	0.97	0.078	0.66	0.89	0.97	0.079	0.67	0.90	0.97	0.083	0.68	0.91	0.98	0.089
Seasonal volatility	0.56	0.82	0.94	0.031	0.56	0.83	0.94	0.031	0.56	0.83	0.95	0.033	0.57	0.83	0.95	0.034
EWMA-t	0.54	0.83	0.97	0.023	0.54	0.84	0.97	0.024	0.55	0.85	0.97	0.028	0.56	0.86	0.97	0.034
HAR-RV	0.51	0.82	0.97	0.014	0.51	0.81	0.97	0.011	0.50	0.81	0.97	0.009	0.50	0.81	0.97	0.010
GARCH(1,1)-t	0.53	0.83	0.97	0.021	0.55	0.85	0.97	0.027	0.56	0.86	0.98	0.035	0.59	0.88	0.98	0.047
Ridge	0.48	0.79	0.96	0.015	0.51	0.82	0.96	0.010	0.51	0.81	0.96	0.010	0.51	0.82	0.97	0.012
Elastic net	0.48	0.79	0.96	0.015	0.51	0.82	0.96	0.010	0.51	0.81	0.96	0.010	0.52	0.82	0.97	0.013
LightGBM	0.74	0.96	1.00	0.124	0.48	0.80	0.96	0.011	0.51	0.82	0.96	0.010	0.52	0.82	0.96	0.015
MLP (1m)	0.55	0.83	0.96	0.029	0.57	0.84	0.96	0.034	0.58	0.85	0.96	0.039	0.58	0.86	0.96	0.043
TCN (1m)	0.56	0.84	0.96	0.032	0.57	0.85	0.96	0.037	0.58	0.85	0.96	0.042	0.59	0.86	0.96	0.046
MRPT-S (1m+5m+30m)	0.55	0.83	0.96	0.029	0.56	0.84	0.96	0.033	0.57	0.85	0.96	0.037	0.58	0.85	0.96	0.040
MRPT-S (1m only)	0.55	0.83	0.96	0.030	0.56	0.84	0.96	0.033	0.57	0.85	0.96	0.037	0.58	0.85	0.96	0.041
MRPT-B (1m+5m+30m)	0.55	0.83	0.96	0.029	0.56	0.84	0.96	0.033	0.57	0.85	0.96	0.037	0.58	0.85	0.96	0.041
Panel B. Recalibrated on the 2023 validation origins (scalar scale c and ν per model and horizon; Section 6)																
Random walk, Student-t (2023-calibrated)	0.51	0.80	0.95	0.015	0.51	0.80	0.95	0.008	0.51	0.80	0.95	0.008	0.51	0.80	0.95	0.009

Notes. development test 2024, 986468 evaluation origins over 251 sessions. Empirical coverage of the central 50/80/95 % prediction intervals of the cumulative return (nominal 0.50/0.80/0.95) and the Kolmogorov–Smirnov distance of the PIT values from uniformity; no scaling is involved.

7.5. Directional and economic content (Hypothesis 5)

The directional accuracy of the predictive medians, on origins with a non-zero realised return and with day-cluster intervals (Table 13), separates the models into two groups. The four neural location forecasts, whose location is damped by construction, sit at 0.499–0.502 at every horizon (MRPT-S 0.5007 [0.4996, 0.5018] at five minutes; the largest t against one half among the neural models and horizons is 2.1, for MRPT-B at sixty minutes). The linear and tree location forecasts reach 0.5181 [0.5152, 0.5209] for ridge at five minutes (t 12.5), with elastic net and LightGBM within a tenth of a point, 0.5101 at fifteen minutes, 0.5046 at thirty and 0.4985 at sixty; the pattern repeats at 0.515 on the held-out instruments and 0.519 in 2025. A hit rate of 0.5181 at five minutes that decays to one half by sixty minutes is the signature of short-horizon reversal in bar closes—bid–ask bounce—and the negative first-order dependence of the model's own PIT scores reported in Section 8 is the same phenomenon seen from the other side.

Whether that signal has economic content is the question of Hypothesis 5, which we test on all 7 location models with a one-minute execution delay, costs of 0/1/2.5/5/10 basis points and two thresholds, of which 224 cells per evaluation set carry a positive cost (Table 14). Of the 224 cells with a cost of at least 1 bp, 5 have a net interval above zero in 2024 (against about 11 expected false positives at the 5 % level) and none in 2025; no cell is positive in both years. The 2024 positive cells are sparse cells of a few hundred trades (ridge and elastic net at fifteen minutes at 2.5–5 basis points, MRPT-B at sixty minutes at 10 basis points; 'results/metrics/main/economic_test.csv'); on the held-out instruments 0 cells are positive in 2024 and 1 in 2025. In the dense cells (one basis point, $k = 1$, 1,012 to 298,881 trades) no net mean is positive with an interval excluding zero (MRPT-S at sixty minutes +0.1 bp [-1.7, +1.9]), and the ridge signal at five minutes is negative net of one basis point (-0.8 bp [-1.4, -0.2]). We read this as a null: the linear and tree location forecasts carry a small directional signal in bar closes that does not survive a one-basis-point cost with a one-minute delay, and the neural location forecasts carry none; whether quote-based execution would change the first statement is a question for data this paper does not have (Section 9).

7.6. Zero-shot transfer, the availability-excluded instruments and the untouched 2025 period

Hypothesis 2 has two clauses and they receive different verdicts (Table 15). Clause (a) asks whether the pooled model beats, on instruments it has never seen, the non-neural baselines refitted on those instruments' own histories. On the 60 held-out instruments (250,322 origins in 2024, Table 16) the MRPT-S contrast with HAR-RV is -0.20 % at five minutes (day-cluster t -4.7, Holm $p = 7.66e-05$) and is not distinguishable from it afterwards: -0.04 % at fifteen minutes, +0.06 % at thirty and +0.02 % at sixty (Holm $p = 1$, 1 and 1); against ridge the contrast is -0.15 % at five minutes and +0.12 % at thirty (Holm $p = 0.697$), against LightGBM -0.21 % and -0.02 %. In 2025 the held-out

Table 13. Directional accuracy of every model with a location forecast: share of origins whose predictive median has the sign of the realised cumulative return, with day-cluster intervals and a test against one half

Model	$H = 5$	$H = 15$	$H = 30$	$H = 60$
Panel A. 2024 development test, pooled instruments (up to 961153 *origins with a non-zero realised return, 251 sessions)*				
Ridge	0.5181 [0.5152, 0.5209]***	0.5101 [0.5073, 0.5128]***	0.5046 [0.5015, 0.5078]***	0.4985 [0.4930, 0.5047]
Elastic net	0.5184 [0.5156, 0.5213]***	0.5103 [0.5076, 0.5129]***	0.5053 [0.5021, 0.5084]***	0.4985 [0.4927, 0.5049]
LightGBM	0.5180 [0.5153, 0.5206]***	0.5085 [0.5062, 0.5108]***	0.5025 [0.4991, 0.5057]	0.4988 [0.4935, 0.5037]
MLP (1m)	0.5003 [0.4993, 0.5014]	0.5000 [0.4990, 0.5010]	0.4998 [0.4985, 0.5011]	0.4992 [0.4974, 0.5011]
TCN (1m)	0.5007 [0.4995, 0.5018]	0.5008 [0.4998, 0.5019]	0.5010 [0.4999, 0.5022]	0.5016 [0.5001, 0.5032]**
MRPT-S (1m+5m+30m)	0.5007 [0.4996, 0.5018]	0.5006 [0.4994, 0.5017]	0.5006 [0.4993, 0.5020]	0.5016 [0.4998, 0.5035]
MRPT-S (1m only)	0.4998 [0.4987, 0.5008]	0.5010 [0.4998, 0.5022]*	0.5009 [0.4996, 0.5022]	0.5011 [0.4994, 0.5028]
MRPT-B (1m+5m+30m)	0.5005 [0.4994, 0.5015]	0.5008 [0.4997, 0.5018]	0.5008 [0.4995, 0.5020]	0.5020 [0.5001, 0.5038]**
Panel B. 2024, held-out instruments (up to 246163 *origins with a non-zero realised return, 251 sessions)*				
Ridge	0.5153 [0.5117, 0.5191]***	0.5062 [0.5027, 0.5096]***	0.5047 [0.5009, 0.5089]**	0.5044 [0.4976, 0.5119]
Elastic net	0.5155 [0.5120, 0.5195]***	0.5060 [0.5025, 0.5094]***	0.5053 [0.5014, 0.5094]***	0.5044 [0.4975, 0.5121]
LightGBM	0.5121 [0.5091, 0.5155]***	0.5029 [0.5000, 0.5057]**	0.5003 [0.4965, 0.5043]	0.4990 [0.4946, 0.5037]
MLP (1m)	0.4993 [0.4971, 0.5014]	0.5007 [0.4987, 0.5027]	0.4989 [0.4968, 0.5010]	0.4999 [0.4974, 0.5028]
TCN (1m)	0.5016 [0.4996, 0.5038]	0.5014 [0.4994, 0.5035]	0.4997 [0.4977, 0.5018]	0.5012 [0.4989, 0.5037]
MRPT-S (1m+5m+30m)	0.5009 [0.4987, 0.5030]	0.5020 [0.5000, 0.5041]**	0.5002 [0.4980, 0.5023]	0.5022 [0.4995, 0.5049]
MRPT-S (1m only)	0.5002 [0.4983, 0.5022]	0.5022 [0.5003, 0.5041]**	0.4999 [0.4978, 0.5018]	0.5006 [0.4982, 0.5031]
MRPT-B (1m+5m+30m)	0.5004 [0.4982, 0.5025]	0.5011 [0.4992, 0.5031]	0.4999 [0.4978, 0.5019]	0.5020 [0.4995, 0.5047]
MRPT-S trained on all 300 instruments	0.5000 [0.4979, 0.5022]	0.5013 [0.4994, 0.5032]	0.5008 [0.4989, 0.5028]	0.5021 [0.4997, 0.5047]
Panel C. 2025 final period, pooled instruments (up to 985166 *origins with a non-zero realised return, 249 sessions)*				
Ridge	0.5190 [0.5156, 0.5222]***	0.5111 [0.5077, 0.5143]***	0.5031 [0.4997, 0.5065]*	0.4979 [0.4909, 0.5044]
Elastic net	0.5192 [0.5158, 0.5224]***	0.5111 [0.5076, 0.5144]***	0.5024 [0.4989, 0.5058]	0.4972 [0.4899, 0.5039]
LightGBM	0.5184 [0.5154, 0.5211]***	0.5114 [0.5084, 0.5140]***	0.5052 [0.5015, 0.5085]***	0.5011 [0.4963, 0.5062]
MLP (1m)	0.5004 [0.4995, 0.5014]	0.5003 [0.4992, 0.5015]	0.4999 [0.4985, 0.5012]	0.5008 [0.4989, 0.5027]
TCN (1m)	0.5007 [0.4997, 0.5017]	0.5003 [0.4991, 0.5013]	0.5006 [0.4994, 0.5018]	0.5010 [0.4993, 0.5027]
MRPT-S (1m+5m+30m)	0.5006 [0.4996, 0.5016]	0.5005 [0.4995, 0.5016]	0.5009 [0.4997, 0.5021]	0.5014 [0.4997, 0.5032]
MRPT-S (1m only)	0.5009 [0.5000, 0.5019]*	0.5005 [0.4994, 0.5015]	0.5008 [0.4995, 0.5019]	0.5022 [0.5005, 0.5039]**
MRPT-B (1m+5m+30m)	0.4995 [0.4986, 0.5005]	0.5008 [0.4998, 0.5019]	0.5008 [0.4996, 0.5020]	0.5021 [0.5006, 0.5037]***
Panel D. 2025, held-out instruments (up to 250672 *origins with a non-zero realised return, 249 sessions)*				
Ridge	0.5141 [0.5101, 0.5178]***	0.5069 [0.5031, 0.5103]***	0.5061 [0.5020, 0.5102]***	0.5021 [0.4955, 0.5092]
Elastic net	0.5142 [0.5101, 0.5179]***	0.5070 [0.5031, 0.5104]***	0.5061 [0.5020, 0.5102]***	0.5020 [0.4951, 0.5091]
LightGBM	0.5131 [0.5100, 0.5161]***	0.5052 [0.5020, 0.5083]***	0.5034 [0.4990, 0.5077]	0.4995 [0.4948, 0.5045]
MLP (1m)	0.5000 [0.4977, 0.5022]	0.5000 [0.4980, 0.5021]	0.5017 [0.4996, 0.5038]	0.4995 [0.4968, 0.5024]
TCN (1m)	0.5006 [0.4985, 0.5028]	0.4995 [0.4976, 0.5015]	0.5009 [0.4990, 0.5030]	0.5000 [0.4978, 0.5023]
MRPT-S (1m+5m+30m)	0.5002 [0.4981, 0.5024]	0.5002 [0.4981, 0.5025]	0.5021 [0.5002, 0.5043]**	0.5012 [0.4984, 0.5036]
MRPT-S (1m only)	0.5016 [0.4993, 0.5039]	0.4996 [0.4976, 0.5018]	0.5027 [0.5007, 0.5050]**	0.5013 [0.4988, 0.5037]
MRPT-B (1m+5m+30m)	0.4999 [0.4978, 0.5020]	0.4995 [0.4976, 0.5015]	0.5012 [0.4989, 0.5036]	0.5016 [0.4992, 0.5038]
MRPT-S trained on all 300 instruments	0.5002 [0.4981, 0.5023]	0.4992 [0.4971, 0.5013]	0.5026 [0.5006, 0.5049]**	0.5001 [0.4976, 0.5026]

Notes. Origins with a zero realised return are excluded; the random walk, seasonal, EWMA, HAR-RV and GARCH baselines have no location forecast ($\mu = 0$) and are not listed. Brackets: 95 % trading-day cluster-bootstrap interval (1,000 resamples); stars from the day-cluster t of (accuracy − 0.5): *** $|t| > 2.58$, ** $|t| > 1.96$, * $|t| > 1.64$ (no multiplicity adjustment; 28 cells per panel). The neural location is damped by construction ($\mu = \sigma_t\, a/10$). Whether the 51–52 % hit rate of the linear and tree models at five minutes survives transaction costs is the object of Table 14. Source: `results/metrics/main/dir_acc_test*.csv`, `final/dir_acc_final*.csv`.

contrast with HAR-RV is -0.22 % at five minutes and +0.17 % at sixty (t 2.8, Holm $p = 0.108$), HAR-RV ahead. Clause (a) is therefore supported at five minutes and rejected from fifteen minutes on. MRPT-S is better than GARCH(1,1)-t (contrast -0.80 % to -1.30 %), EWMA (-0.49 % to -0.87 %), the TCN and the MLP at every horizon, and than its one-minute-only twin (-0.04 % to -0.12 %).

Clause (b) asks whether transfer costs anything. Two tests answer it. First, the difference between the relative gain over the random walk on the held-out instruments and on the pooled instruments, with a joint day-cluster interval, is +0.07 percentage points [-0.07, +0.20] at five minutes and +0.07 [-0.08, +0.25] at sixty for MRPT-S (day-cluster t 1.0 and 0.9; the intervals contain zero at every horizon for

Table 14. Hypothesis 5 with the one-minute execution delay on every model with a location forecast: net return of a sign rule after round-trip costs (2024 development test, pooled instruments; counts for the four evaluation sets)

Model	$H = 5$	$H = 15$	$H = 30$	$H = 60$
Panel A. Net return per trade at a 1 bp round-trip cost, $k = 1$ (bp) [95 % CI] (trades)				
Ridge	-0.78 [-1.36, -0.20] (41110)	-0.11 [-1.14, +1.00] (54397)	-0.35 [-1.45, +0.88] (71468)	-1.17 [-2.62, +0.35] (150009)
Elastic net	-0.81 [-1.37, -0.20] (39026)	+0.11 [-0.96, +1.31] (49129)	-0.06 [-1.37, +1.29] (56750)	-1.12 [-2.78, +0.57] (126936)
LightGBM	-0.92 [-1.27, -0.54] (76960)	-0.96 [-1.59, -0.35] (116394)	-0.80 [-1.52, -0.14] (171899)	-0.92 [-1.82, +0.09] (298881)
MLP (1m)	-2.74 [-9.75, +5.08] (1163)	-0.46 [-2.95, +2.03] (20973)	-0.09 [-1.70, +1.72] (89217)	-0.57 [-1.81, +0.97] (263211)
TCN (1m)	-5.53 [-13.87, +2.70] (1202)	-0.03 [-2.30, +2.33] (17050)	-0.56 [-2.19, +1.08] (72183)	-0.77 [-2.20, +0.63] (208852)
MRPT-S (1m+5m+30m)	-9.58 [-16.16, -2.63] (1012)	-0.91 [-3.32, +1.44] (14656)	-0.70 [-2.77, +1.41] (55223)	+0.07 [-1.74, +1.87] (166278)
MRPT-B (1m+5m+30m)	-3.25 [-11.97, +5.30] (1061)	-0.71 [-3.33, +1.77] (19383)	-0.18 [-2.72, +2.23] (52948)	+0.70 [-1.17, +2.58] (141285)

Panel B. Cells of the full grid (7 models × 4 H × 4 costs × 2 thresholds = 224 per split)			
Split	Net interval excludes zero	of which net-positive	Expected false positives (5 %)
2024 development test, pooled instruments	33 of 224	5	11.2
2024, held-out instruments	33 of 224	0	11.2
2025 final period, pooled instruments	16 of 224	0	11.2
2025, held-out instruments	29 of 224	1	11.2

Notes. Sign rule: at every origin whose predictive median exceeds k times the round-trip cost in absolute value, buy (sell) one unit and unwind after H minutes; the realised return is measured from the close of the bar after the origin (one-minute execution delay), so the first future minute never enters the profit. Net return = gross return minus the round-trip cost, averaged over the traded origins; brackets: 95 % trading-day cluster-bootstrap interval (1,000 resamples). Panel A shows the densest cells (1 bp, $k = 1$); Panel B counts, over the 224 cells with a cost of at least 1 bp per split, how many have an interval excluding zero, how many of those are positive, and the number of false positives expected at the 5 % level if every cell were a true null (0.05×224). Positive cells — 2024 development test, pooled instruments: MRPT-B (1m+5m+30m), $H = 60$, 10 bp, $k = 2$: +121.7 bp [+27.8, +203.8], 201 trades; Ridge, $H = 15$, 2.5 bp, $k = 2$: +29.4 bp [+7.9, +51.2], 521 trades; Ridge, $H = 15$, 5 bp, $k = 1$: +26.9 bp [+5.4, +48.7], 521 trades; Elastic net, $H = 15$, 2.5 bp, $k = 2$: +32.4 bp [+4.1, +56.2], 455 trades; Elastic net, $H = 15$, 5 bp, $k = 1$: +29.9 bp [+1.6, +53.7], 455 trades; 2024, held-out instruments: none; 2025 final period, pooled instruments: none; 2025, held-out instruments: MRPT-B (1m+5m+30m), $H = 5$, 2.5 bp, $k = 1$: +76.6 bp [+34.5, +164.9], 31 trades. The conditional authorization trigger of the review (a linear or tree cell positive in both 2024 and 2025) was not met. The result files hold 308 rows per split: the 224 cells, the 56 cost-free rows (0 bp, $k = 1, 2$) and 28 diagnostic rows written with `cost_bps = -1`, $k = 0$ (median–realised correlation and slope on every origin, no trade); the last two groups are not cells of the grid. Source: `results/metrics/main/economic_test*.csv`, `final/economic_final*.csv` and their `_summary.json`.

MRPT-S, MRPT-B, the twin and the MLP), whereas the baselines refitted on the held-out instruments' own histories show a negative gap ('results/metrics/main/transfer_gap_test_heldout.csv'). Second, MRPT-S was trained once more on all 300 instruments, so that the 60 are in-sample for it, and scored on the identical 250,322 held-out origins: the in-sample model is not better than the zero-shot checkpoint—+0.01 % at five minutes, +0.04 % at thirty and +0.09 % at sixty (t 5.3; the same step budget spread over more instruments)—and shows the same pattern against HAR-RV (-0.19 % at five minutes, +0.11 % at sixty). Clause (b) is supported: a pooled model without ticker identity forecasts unseen liquid instruments as well as it forecasts the instruments it was trained on. Two caveats apply. The held-out ETFs are not all unrelated to the pooled ones: 12 of the 18 held-out ETFs share an index family with at least one pooled ETF (Table B.13, Appendix; for example a leveraged S&P 500 fund held out while SPY and other S&P 500 funds are pooled), so "unseen" means an unseen ticker, not an unseen index. And the held-out instruments passed the same availability filter as the pooled ones.

The 66 instruments that the availability filter removed although they pass it on the training years are the harder test (Table 17; 63,182 origins in 2024, baselines refitted on their own histories with the corrected construction). The neural checkpoints applied zero-shot improve on the Gaussian random walk by 4.5–5.2 % of scaled CRPS (gain, positive = better), more than on the pooled instruments,

and the Student-t random walk and the regressions improve by similar amounts; the contrast with HAR-RV is -0.82 % at five minutes (t -10.8) and +0.08 %, +0.33 % and +0.36 % at fifteen, thirty and sixty minutes (Holm p = 0.795, 0.0806 and 0.107), with the same ordering in 2025 (-0.85 % at five minutes, +0.17 % at sixty). The multi-resolution twin contrast on these instruments is -0.07 % to -0.15 % (Holm p = 0.107 to 0.0342). On the instruments whose liquidity fell, then, the ordering of the main study survives and the gaps are somewhat wider; the filter did not hide a reversal.

The untouched 2025 period (Table 18; 1,008,425 pooled origins over 249 sessions and 254,334 held-out origins; models and baselines frozen from the main study) repeats the 2024 ordering. Every model except LightGBM at five minutes beats the Gaussian random walk, HAR-RV and MRPT-S are within -0.07 % of each other at five minutes (-0.36 % after recalibration), and HAR-RV, ridge and elastic net lead from fifteen minutes on (0.4233 against 0.4256 at sixty minutes, +0.56 %, t = 8.6). On the main run the multi-resolution contrast loses its sign and its significance: MRPT-S against its twin is -0.01 % to +0.02 % across horizons (-0.01 % at sixty minutes, Holm p = 1), and on the held-out instruments -0.02 % to -0.00 %, single-seed numbers below the minimum detectable effect of 0.03–0.06 % (Table B.7, Appendix). The eight seed checkpoints of Section 7.2, scored once on the same 2025 origins with no selection (Table 10, Panel D; levels per seed in Table B.10, Appendix), show that the main run is the one seed without a 2025 effect: the mean contrast over the five seed pairs is -0.046 % at five minutes and -0.116 % at sixty, with 4 of the five pairs negative at every horizon and a day-cluster t on the seed-averaged differential of -6.9 and -6.5—the 2024 sign at about half the 2024 magnitude—but the t across seeds (-2.29, -1.88, -1.90 and -2.21) stays below the 2.78 threshold at every horizon, so the pre-stated rule is met at 0 of the four horizons and the 2024 verdict on Hypothesis 1 is not replicated in 2025. The seed half-band of the full model in 2025 is 0.06–0.09 %, wider than in 2024. Base and Small remain indistinguishable. One nuance appears on 2025 stress days (top decile of cross-sectional volatility, 101,561 origins): at sixty minutes HAR-RV's 50/80/95 % intervals cover 46/78/95 % of outcomes while MRPT-S covers 55/84/95 %; the regression's variance forecasts lag on the days when volatility jumps and the sequence model's wider body is, on those days, closer to nominal at 80 %, with the same scaled CRPS (0.4554 against 0.4559).

8. Ablations, controls and falsification

Every variant in Table 20 is a Small model trained for 4,000 steps with the same seed and compared, on the 2024 development test, with a reference trained under exactly the same budget (MRPT-S, 4k); two further seeds of the reference give a noise floor at that budget, and the five-seed band of the 12,000-step twins (Section 7.2) gives the floor at the full budget. Figure 5 shows the relative scaled-CRPS changes at sixty minutes against the reference with their day-cluster intervals and the

Table 15. Hypothesis 2 in its two pre-registered clauses: (a) superiority over matched non-neural baselines on instruments never seen in training, (b) absence of a transfer penalty (60 held-out instruments, 2024 and 2025)

	$H = 5$	$H = 15$	$H = 30$	$H = 60$
Panel A. H2a: zero-shot MRPT-S vs baselines refitted on the held-out histories, % of the baseline's scaled CRPS [t_{cl}]				
2024: vs HAR-RV	-0.20 [-4.7]***	-0.04 [-1.0]	+0.06 [1.0]	+0.02 [0.3]
2024: vs Ridge	-0.15 [-3.2]**	+0.02 [0.5]	+0.12 [2.0]	+0.07 [1.0]
2024: vs Elastic net	-0.15 [-3.2]**	+0.02 [0.4]	+0.11 [1.9]	+0.08 [1.2]
2024: vs LightGBM	-0.21 [-4.6]***	-0.02 [-0.4]	+0.00 [0.1]	-0.02 [-0.3]
2025: vs HAR-RV	-0.22 [-4.8]***	+0.00 [0.1]	+0.07 [1.2]	+0.17 [2.8]
2025: vs Ridge	-0.14 [-3.0]*	+0.13 [3.1]**	+0.17 [3.3]**	+0.21 [3.6]**
2025: vs Elastic net	-0.15 [-3.2]**	+0.13 [3.1]**	+0.15 [3.0]*	+0.22 [3.7]***
2025: vs LightGBM	-0.21 [-4.2]***	+0.05 [0.8]	+0.02 [0.3]	+0.04 [0.5]
Panel B. H2b: transfer gap = held-out gain vs RW minus pooled gain vs RW, percentage points [95 % CI]				
2024: MRPT-S (1m+5m+30m)	+0.07 [-0.07, +0.20]	+0.03 [-0.08, +0.16]	+0.02 [-0.11, +0.16]	+0.07 [-0.08, +0.25]
2024: MRPT-S (1m only)	+0.06 [-0.07, +0.19]	+0.02 [-0.09, +0.14]	+0.01 [-0.12, +0.15]	+0.07 [-0.09, +0.24]
2024: MRPT-B (1m+5m+30m)	+0.06 [-0.07, +0.20]	-0.00 [-0.11, +0.12]	-0.01 [-0.14, +0.13]	+0.07 [-0.10, +0.24]
2024: TCN (1m)	+0.08 [-0.04, +0.22]	+0.15 [+0.04, +0.27]	+0.17 [+0.04, +0.31]	+0.24 [+0.09, +0.40]
2024: MLP (1m)	-0.02 [-0.16, +0.11]	-0.08 [-0.22, +0.06]	-0.12 [-0.27, +0.03]	-0.09 [-0.26, +0.09]
2024: HAR-RV	-0.15 [-0.28, -0.03]	-0.39 [-0.51, -0.27]	-0.51 [-0.65, -0.36]	-0.60 [-0.75, -0.44]
2024: Ridge	-0.23 [-0.35, -0.08]	-0.31 [-0.43, -0.19]	-0.41 [-0.56, -0.27]	-0.46 [-0.63, -0.29]
2025: MRPT-S (1m+5m+30m)	-0.01 [-0.11, +0.09]	-0.01 [-0.12, +0.09]	-0.04 [-0.17, +0.09]	-0.10 [-0.23, +0.04]
2025: MRPT-S (1m only)	-0.03 [-0.13, +0.07]	-0.03 [-0.13, +0.07]	-0.06 [-0.17, +0.07]	-0.11 [-0.24, +0.04]
2025: MRPT-B (1m+5m+30m)	-0.02 [-0.12, +0.08]	-0.03 [-0.14, +0.08]	-0.03 [-0.15, +0.08]	-0.09 [-0.22, +0.04]
2025: TCN (1m)	-0.03 [-0.13, +0.08]	+0.02 [-0.08, +0.12]	+0.05 [-0.06, +0.18]	+0.03 [-0.10, +0.15]
2025: MLP (1m)	-0.08 [-0.18, +0.03]	-0.11 [-0.21, -0.00]	-0.12 [-0.24, +0.01]	-0.18 [-0.31, -0.04]
2025: HAR-RV	-0.22 [-0.31, -0.13]	-0.30 [-0.40, -0.20]	-0.37 [-0.49, -0.23]	-0.50 [-0.65, -0.35]
2025: Ridge	-0.34 [-0.44, -0.24]	-0.23 [-0.33, -0.12]	-0.32 [-0.44, -0.20]	-0.46 [-0.60, -0.31]
Panel C. H2b, paired: MRPT-S trained on all 300 vs zero-shot MRPT-S on the same held-out origins, % [t_{cl}]				
2024: all-300 model vs zero-shot MRPT-S	+0.01 [1.3]	+0.02 [1.6]	+0.04 [3.1]**	+0.09 [5.3]***
2024: all-300 model vs HAR-RV	-0.19 [-4.8]***	-0.03 [-0.6]	+0.10 [1.8]	+0.11 [1.7]
2025: all-300 model vs zero-shot MRPT-S	-0.01 [-0.3]	+0.02 [1.9]	+0.03 [1.9]	+0.06 [2.9]*
2025: all-300 model vs HAR-RV	-0.23 [-5.4]***	+0.03 [0.6]	+0.09 [1.7]	+0.22 [3.8]***

Notes. Sixty held-out instruments (250,322 origins in 2024, 254,334 in 2025). Panel A: percentage change in mean scaled CRPS of the zero-shot MRPT-S relative to each baseline refitted on the held-out instruments' own 2018–2022 histories with the corrected construction (negative = MRPT-S better); day-cluster t in brackets (1,000 trading-day resamples); stars = Holm-adjusted p over the held-out family (72 tests in 2024, 72 in 2025): *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$. Clause (a) is supported only where the entry is negative and starred. Panel B: transfer gap = (gain vs the random walk on the held-out instruments) − (gain on the pooled instruments), in percentage points of unscaled CRPS, with a 95 % interval from a joint trading-day bootstrap of both samples; an interval containing zero means no measurable transfer penalty. Panel C: MRPT-S trained once on all 300 instruments (seed 1, 12,000 steps) scored on the identical held-out origins and compared with the zero-shot checkpoint by the paired day-cluster test. Source: `results/metrics/main/pairwise_test_heldout.csv`, `final/pairwise_final_heldout.csv`, `*/transfer_gap_*_heldout.csv`.

seed band; all contrasts below are relative changes in scaled CRPS against the 4k reference, positive when the variant is worse, with Holm adjustment in one family of 68 tests.

Resolutions at the short budget (H1 again). At 4,000 steps removing the 5- and 30-minute branches costs +0.12 % at five minutes and +0.15 % at sixty, removing only the 30-minute branch +0.09 % and +0.07 %; the two extra seeds of the reference lie at -0.03 % and +0.14 % at sixty minutes, so the one-seed resolution effects at this budget sit at the edge of the seed spread. At the full budget the five-seed contrast of Section 7.2 is smaller and of the same sign.

Table 16. Main study (held-out instruments): mean scaled CRPS of the cumulative-return forecast by horizon (minutes)

Model	Parameters	$H = 5$	$H = 15$	$H = 30$	$H = 60$
Panel A. Models as specified (baselines) or trained (neural models)					
Random walk (trailing RMS)	–	0.4744 [0.4686, 0.4803]	0.4653 [0.4584, 0.4716]	0.4541 [0.4468, 0.4617]	0.4415 [0.4316, 0.4521]
Seasonal volatility	–	0.4623 [0.4562, 0.4686]	0.4531 [0.4458, 0.4602]	0.4419 [0.4342, 0.4500]	0.4289 [0.4184, 0.4399]
EWMA-*t*	–	0.4593 [0.4533, 0.4655]	0.4508 [0.4435, 0.4579]	0.4404 [0.4327, 0.4485]	0.4287 [0.4184, 0.4398]
HAR-RV	–	0.4580 [0.4517, 0.4643]	0.4490 [0.4414, 0.4563]	0.4378 [0.4299, 0.4461]	0.4249 [0.4144, 0.4361]
GARCH(1,1)-*t*	–	0.4607 [0.4545, 0.4670]	0.4517 [0.4443, 0.4591]	0.4415 [0.4339, 0.4496]	0.4306 [0.4203, 0.4414]
Ridge	–	0.4578 [0.4514, 0.4641]	0.4487 [0.4412, 0.4560]	0.4376 [0.4297, 0.4459]	0.4247 [0.4142, 0.4359]
Elastic net	–	0.4578 [0.4514, 0.4641]	0.4487 [0.4412, 0.4560]	0.4376 [0.4297, 0.4459]	0.4247 [0.4142, 0.4358]
LightGBM	–	0.4580 [0.4519, 0.4643]	0.4489 [0.4413, 0.4562]	0.4381 [0.4302, 0.4464]	0.4251 [0.4147, 0.4361]
MLP (1m)	301,827	0.4582 [0.4520, 0.4646]	0.4499 [0.4428, 0.4569]	0.4393 [0.4317, 0.4470]	0.4264 [0.4162, 0.4371]
TCN (1m)	231,459	0.4578 [0.4516, 0.4641]	0.4497 [0.4425, 0.4567]	0.4392 [0.4316, 0.4471]	0.4265 [0.4164, 0.4373]
MRPT-S (1m+5m+30m)	3,455,619	0.4571 [0.4510, 0.4634]	0.4488 [0.4415, 0.4559]	0.4381 [0.4304, 0.4460]	0.4250 [0.4150, 0.4357]
MRPT-S (1m only)	3,200,259	0.4573 [0.4511, 0.4636]	0.4490 [0.4417, 0.4561]	0.4384 [0.4307, 0.4463]	0.4255 [0.4154, 0.4364]
MRPT-B (1m+5m+30m)	20,718,339	0.4573 [0.4512, 0.4635]	0.4490 [0.4417, 0.4561]	0.4383 [0.4306, 0.4462]	0.4253 [0.4151, 0.4362]
MRPT-S trained on all 300 instruments	–	0.4571 [0.4510, 0.4635]	0.4489 [0.4416, 0.4560]	0.4383 [0.4306, 0.4462]	0.4254 [0.4154, 0.4362]
Panel B. Recalibrated on the 2023 validation origins (scalar scale c and ν per model and horizon; Section 6)					
Random walk, Student-*t* (2023-calibrated)	–	0.4655 [0.4590, 0.4719]	0.4556 [0.4478, 0.4629]	0.4429 [0.4348, 0.4515]	0.4282 [0.4171, 0.4394]

Notes. held-out instruments, 2024, 250322 evaluation origins over 251 sessions. Brackets: 95 % trading-day cluster-bootstrap intervals (all instruments and origins of one session form one cluster). Convention: scaled CRPS, i.e. CRPS divided by $\sigma_t\sqrt{H}$. Parameters: trainable parameters of the trained models (`train_summary.json`); baselines have none in this sense.

Table 17. Second held-out set (66 availability-excluded instruments) (held-out instruments): mean scaled CRPS of the cumulative-return forecast by horizon (minutes)

Model	$H = 5$	$H = 15$	$H = 30$	$H = 60$
Random walk (trailing RMS)	0.4523 [0.4447, 0.4601]	0.4398 [0.4324, 0.4483]	0.4322 [0.4240, 0.4428]	0.4194 [0.4082, 0.4342]
Seasonal volatility	0.4370 [0.4290, 0.4455]	0.4243 [0.4164, 0.4335]	0.4169 [0.4084, 0.4280]	0.4032 [0.3914, 0.4190]
EWMA-*t*	0.4329 [0.4249, 0.4413]	0.4218 [0.4138, 0.4310]	0.4158 [0.4069, 0.4268]	0.4040 [0.3924, 0.4199]
HAR-RV	0.4321 [0.4241, 0.4404]	0.4176 [0.4093, 0.4272]	0.4105 [0.4012, 0.4223]	0.3968 [0.3842, 0.4136]
GARCH(1,1)-*t*	0.4320 [0.4236, 0.4406]	0.4206 [0.4124, 0.4301]	0.4144 [0.4055, 0.4257]	0.4026 [0.3905, 0.4185]
Ridge	0.4296 [0.4214, 0.4382]	0.4172 [0.4088, 0.4267]	0.4103 [0.4010, 0.4219]	0.3967 [0.3842, 0.4136]
Elastic net	0.4297 [0.4214, 0.4382]	0.4171 [0.4088, 0.4266]	0.4102 [0.4010, 0.4219]	0.3967 [0.3841, 0.4135]
LightGBM	0.4669 [0.4595, 0.4745]	0.4180 [0.4093, 0.4281]	0.4111 [0.4016, 0.4228]	0.3970 [0.3845, 0.4135]
MLP (1m)	0.4292 [0.4211, 0.4377]	0.4180 [0.4099, 0.4272]	0.4116 [0.4027, 0.4227]	0.3982 [0.3862, 0.4142]
TCN (1m)	0.4286 [0.4207, 0.4370]	0.4187 [0.4107, 0.4280]	0.4127 [0.4037, 0.4238]	0.3999 [0.3881, 0.4158]
MRPT-S (1m+5m+30m)	0.4286 [0.4207, 0.4368]	0.4180 [0.4099, 0.4273]	0.4118 [0.4029, 0.4230]	0.3983 [0.3864, 0.4142]
MRPT-S (1m only)	0.4289 [0.4209, 0.4372]	0.4184 [0.4103, 0.4276]	0.4121 [0.4031, 0.4233]	0.3989 [0.3868, 0.4150]
MRPT-B (1m+5m+30m)	0.4288 [0.4209, 0.4372]	0.4183 [0.4102, 0.4276]	0.4120 [0.4030, 0.4232]	0.3988 [0.3868, 0.4149]

Notes. held-out instruments, 2024, 63182 evaluation origins over 251 sessions. Brackets: 95 % trading-day cluster-bootstrap intervals (all instruments and origins of one session form one cluster). Convention: scaled CRPS, i.e. CRPS divided by $\sigma_t\sqrt{H}$.

Budget: a learning curve of the twins. Whether the null of Hypothesis 1 is an artefact of the 12,000-step budget was tested by continuing both twins from their checkpoints under the same early-stopping rule, now on the stratified validation sample (Table 21). The continuations are warm restarts of the cosine schedule (learning rate $9.89e-05$ at the restart against a peak of 3×10^{-4}), not fresh runs, and early stopping ended them before the nominal budget: MRPT-S stopped at 34,000 steps of the 36,000-step budget and the one-minute twin at 30,000. Scaled CRPS moves by less than a tenth of a percent along the curve (MRPT-S at five minutes 0.4531, 0.4529 and 0.4528; at sixty

Table 18. Final period: mean scaled CRPS of the cumulative-return forecast by horizon (minutes)

Model	$H = 5$	$H = 15$	$H = 30$	$H = 60$
Panel A. Models as specified (baselines) or trained (neural models)				
Random walk (trailing RMS)	0.4759 [0.4683, 0.4828]	0.4621 [0.4543, 0.4697]	0.4521 [0.4439, 0.4602]	0.4409 [0.4316, 0.4512]
Seasonal volatility	0.4655 [0.4575, 0.4732]	0.4505 [0.4427, 0.4585]	0.4406 [0.4319, 0.4497]	0.4291 [0.4188, 0.4401]
EWMA-t	0.4609 [0.4529, 0.4682]	0.4470 [0.4391, 0.4548]	0.4380 [0.4293, 0.4469]	0.4278 [0.4175, 0.4387]
HAR-RV	0.4597 [0.4515, 0.4672]	0.4451 [0.4369, 0.4532]	0.4352 [0.4259, 0.4443]	0.4233 [0.4126, 0.4347]
GARCH(1,1)-t	0.4616 [0.4537, 0.4690]	0.4468 [0.4388, 0.4547]	0.4380 [0.4292, 0.4468]	0.4286 [0.4183, 0.4397]
Ridge	0.4591 [0.4509, 0.4667]	0.4449 [0.4368, 0.4529]	0.4351 [0.4259, 0.4442]	0.4235 [0.4130, 0.4347]
Elastic net	0.4592 [0.4509, 0.4667]	0.4449 [0.4368, 0.4529]	0.4351 [0.4260, 0.4442]	0.4235 [0.4130, 0.4347]
LightGBM	0.5036 [0.4964, 0.5103]	0.4443 [0.4362, 0.4524]	0.4348 [0.4257, 0.4438]	0.4235 [0.4130, 0.4347]
MLP (1m)	0.4600 [0.4519, 0.4672]	0.4465 [0.4385, 0.4545]	0.4371 [0.4282, 0.4458]	0.4260 [0.4158, 0.4369]
TCN (1m)	0.4600 [0.4520, 0.4671]	0.4471 [0.4391, 0.4550]	0.4379 [0.4291, 0.4467]	0.4271 [0.4169, 0.4380]
MRPT-S (1m+5m+30m)	0.4594 [0.4514, 0.4666]	0.4462 [0.4382, 0.4540]	0.4368 [0.4278, 0.4456]	0.4256 [0.4154, 0.4366]
MRPT-S (1m only)	0.4594 [0.4514, 0.4666]	0.4461 [0.4381, 0.4540]	0.4367 [0.4278, 0.4456]	0.4257 [0.4154, 0.4367]
MRPT-B (1m+5m+30m)	0.4595 [0.4516, 0.4667]	0.4462 [0.4383, 0.4541]	0.4368 [0.4279, 0.4456]	0.4257 [0.4155, 0.4366]
Panel B. Recalibrated on the 2023 validation origins (scalar scale c and ν per model and horizon; Section 6)				
Random walk, Student-t (2023-calibrated)	0.4666 [0.4583, 0.4741]	0.4510 [0.4424, 0.4593]	0.4398 [0.4305, 0.4489]	0.4266 [0.4159, 0.4382]

Notes. untouched final period 2025, 1008425 evaluation origins over 249 sessions. Brackets: 95 % trading-day cluster-bootstrap intervals (all instruments and origins of one session form one cluster). Convention: scaled CRPS, i.e. CRPS divided by $\sigma_t \sqrt{H}$.

Table 19. Final period: empirical 50/80/95% coverage and PIT KS statistic

Model	$H = 5$				$H = 15$				$H = 30$				$H = 60$			
	50	80	95	KS	50	80	95	KS	50	80	95	KS	50	80	95	KS
Panel A. Models as specified (baselines) or trained (neural models)																
Random walk (trailing RMS)	0.64	0.88	0.96	0.074	0.66	0.89	0.97	0.080	0.66	0.90	0.97	0.083	0.68	0.91	0.98	0.089
Seasonal volatility	0.56	0.82	0.94	0.032	0.56	0.83	0.95	0.034	0.56	0.83	0.95	0.035	0.57	0.83	0.95	0.037
EWMA-t	0.53	0.83	0.97	0.021	0.54	0.84	0.97	0.025	0.55	0.85	0.97	0.029	0.56	0.86	0.97	0.035
HAR-RV	0.50	0.81	0.97	0.017	0.51	0.81	0.97	0.010	0.49	0.81	0.96	0.009	0.50	0.81	0.96	0.010
GARCH(1,1)-t	0.52	0.82	0.96	0.020	0.54	0.84	0.97	0.021	0.55	0.85	0.97	0.028	0.57	0.86	0.97	0.039
Ridge	0.48	0.79	0.96	0.016	0.52	0.82	0.96	0.012	0.51	0.81	0.96	0.013	0.51	0.82	0.96	0.016
Elastic net	0.48	0.79	0.96	0.016	0.52	0.82	0.96	0.012	0.51	0.81	0.96	0.013	0.51	0.82	0.96	0.016
LightGBM	0.74	0.96	1.00	0.122	0.48	0.80	0.96	0.012	0.51	0.82	0.96	0.013	0.52	0.82	0.96	0.017
MLP (1m)	0.55	0.83	0.96	0.027	0.57	0.85	0.96	0.036	0.58	0.85	0.96	0.041	0.58	0.86	0.96	0.044
TCN (1m)	0.55	0.84	0.96	0.030	0.58	0.85	0.96	0.039	0.58	0.86	0.96	0.042	0.59	0.86	0.96	0.047
MRPT-S (1m+5m+30m)	0.55	0.83	0.96	0.028	0.57	0.85	0.96	0.035	0.57	0.85	0.96	0.038	0.58	0.85	0.96	0.041
MRPT-S (1m only)	0.55	0.83	0.96	0.027	0.57	0.84	0.96	0.034	0.57	0.85	0.96	0.037	0.58	0.85	0.96	0.040
MRPT-B (1m+5m+30m)	0.55	0.83	0.96	0.028	0.57	0.84	0.96	0.035	0.57	0.85	0.96	0.039	0.58	0.85	0.96	0.041
Panel B. Recalibrated on the 2023 validation origins (scalar scale c and ν per model and horizon; Section 6)																
Random walk, Student-t (2023-calibrated)	0.51	0.80	0.95	0.018	0.51	0.80	0.95	0.010	0.51	0.80	0.95	0.011	0.51	0.80	0.95	0.011

Notes. untouched final period 2025, 1008425 evaluation origins over 249 sessions. Empirical coverage of the central 50/80/95 % prediction intervals of the cumulative return (nominal 0.50/0.80/0.95) and the Kolmogorov–Smirnov distance of the PIT values from uniformity; no scaling is involved.

minutes 0.4187, 0.4188 and 0.4185), the slope of the stratified validation likelihood per thousand steps at the last budget is -0.00001 for MRPT-S and +0.00006 for the twin, and the sign of the contrast with HAR-RV at thirty and sixty minutes is unchanged along the curve (Table 21). The H1 gap at the last budget is +0.05 % at five minutes, -0.01 % at fifteen, -0.09 % at thirty and -0.25 % at sixty; at five and fifteen minutes it lies outside the five-seed band of the 12,000-step contrast (with a sign change at five minutes), inside it at thirty and sixty. By the rule fixed before the runs (gap inside the band at every horizon) the budget condition therefore stays in the statement of the result (inside at 2 of the four horizons; outside at $H = 5$, $H = 15$): the null is established at a 12,000-step budget

Table 20. Ablations (Small, 4,000 steps): mean scaled CRPS of the cumulative-return forecast by horizon (minutes)

Model	Parameters	$H = 5$	$H = 15$	$H = 30$	$H = 60$
Random walk (trailing RMS)	–	0.4699 [0.4648, 0.4754]	0.4575 [0.4515, 0.4639]	0.4469 [0.4412, 0.4538]	0.4347 [0.4271, 0.4441]
Seasonal volatility	–	0.4582 [0.4527, 0.4641]	0.4453 [0.4390, 0.4519]	0.4347 [0.4285, 0.4418]	0.4221 [0.4139, 0.4320]
EWMA-t	–	0.4547 [0.4492, 0.4606]	0.4425 [0.4363, 0.4491]	0.4326 [0.4264, 0.4396]	0.4214 [0.4133, 0.4314]
MRPT-S (1m+5m)	3,330,819	0.4537 [0.4482, 0.4596]	0.4418 [0.4355, 0.4483]	0.4316 [0.4253, 0.4384]	0.4191 [0.4110, 0.4289]
MRPT-S cross-attention fusion	5,533,443	0.4535 [0.4480, 0.4594]	0.4416 [0.4353, 0.4481]	0.4313 [0.4251, 0.4382]	0.4190 [0.4109, 0.4288]
MRPT-S Gaussian head	3,455,619	0.4546 [0.4492, 0.4606]	0.4424 [0.4361, 0.4489]	0.4320 [0.4258, 0.4389]	0.4195 [0.4115, 0.4292]
MRPT-S no regime	3,416,643	0.4534 [0.4479, 0.4593]	0.4414 [0.4352, 0.4480]	0.4312 [0.4249, 0.4380]	0.4188 [0.4107, 0.4286]
MRPT-S learned regime	3,421,643	0.4531 [0.4476, 0.4591]	0.4410 [0.4347, 0.4476]	0.4306 [0.4243, 0.4376]	0.4182 [0.4100, 0.4281]
MRPT-S returns only	3,455,619	0.4542 [0.4487, 0.4601]	0.4424 [0.4362, 0.4489]	0.4322 [0.4260, 0.4392]	0.4201 [0.4119, 0.4299]
MRPT-S time only	3,455,619	0.4542 [0.4488, 0.4601]	0.4424 [0.4362, 0.4489]	0.4323 [0.4260, 0.4392]	0.4200 [0.4119, 0.4299]
MRPT-S OHLCV (no rel. volume)	3,455,619	0.4536 [0.4481, 0.4595]	0.4418 [0.4355, 0.4482]	0.4316 [0.4253, 0.4385]	0.4193 [0.4113, 0.4291]
MRPT-S shuffled minutes	3,455,619	0.4538 [0.4483, 0.4598]	0.4420 [0.4357, 0.4484]	0.4318 [0.4255, 0.4388]	0.4196 [0.4115, 0.4294]
MRPT-S + ticker id	3,501,699	0.4534 [0.4479, 0.4593]	0.4418 [0.4356, 0.4483]	0.4318 [0.4256, 0.4387]	0.4199 [0.4118, 0.4297]
MRPT-S reference (4k steps)	3,455,619	0.4533 [0.4478, 0.4592]	0.4415 [0.4352, 0.4480]	0.4312 [0.4249, 0.4381]	0.4188 [0.4107, 0.4286]
MRPT-S 1m only (4k)	3,200,259	0.4538 [0.4483, 0.4597]	0.4420 [0.4358, 0.4485]	0.4318 [0.4256, 0.4386]	0.4195 [0.4113, 0.4293]
MRPT-S reference, seed 2	3,455,619	0.4531 [0.4476, 0.4591]	0.4411 [0.4347, 0.4476]	0.4308 [0.4245, 0.4379]	0.4187 [0.4103, 0.4287]
MRPT-S reference, seed 3	3,455,619	0.4531 [0.4476, 0.4590]	0.4413 [0.4350, 0.4479]	0.4312 [0.4250, 0.4380]	0.4194 [0.4115, 0.4290]
MRPT-S + copula (joint sampling)	3,456,584	0.4534 [0.4479, 0.4593]	0.4417 [0.4355, 0.4481]	0.4316 [0.4254, 0.4385]	0.4194 [0.4113, 0.4291]
MRPT-S + copula, independent sampling	3,456,584	0.4535 [0.4480, 0.4593]	0.4416 [0.4354, 0.4481]	0.4314 [0.4252, 0.4383]	0.4191 [0.4110, 0.4288]

Notes. development test 2024, 986468 evaluation origins over 251 sessions. Brackets: 95 % trading-day cluster-bootstrap intervals (all instruments and origins of one session form one cluster). Convention: scaled CRPS, i.e. CRPS divided by $\sigma_t\sqrt{H}$. Parameters: trainable parameters of the trained models (`train_summary.json`); baselines have none in this sense.

with no hyper-parameter search, the continuations do not overturn it, and they do not establish it at larger budgets either. Base was not continued beyond 12,000 steps.

What the model uses. The destructive ablations locate the information. A model that sees only the calendar and missingness flags—no returns, ranges or volumes—with the *full* global token (which carries the trailing scale, the realised-variance ratios and relative volume, i.e. the HAR regressors) is worse than the reference by +0.21 % to +0.29 %; with the global token reduced to $\log \sigma_t$ and the calendar, so that the model has nothing but a scale and a clock, the cost is +0.56 % to +0.50 % (day-cluster t 12.6 and 7.1). Returns-only inputs cost +0.20 % to +0.29 %; dropping relative volume +0.07 % to +0.12 %; shuffling the order of the 390 one-minute bars at training and evaluation time +0.11 % to +0.18 % when whole rows are permuted (the calendar and gap columns travel with the bars, so provenance stays readable) and +0.09 % to +0.20 % when only the price and volume columns are permuted and the calendar columns stay in place (t 9.0 and 10.0). The falsification test passes in the sense intended: the order of the price and volume columns matters, by one to two tenths of a percent.

How much of the gain over the random walk is minute-level structure depends on which random walk (Table 22). Against the Gaussian random walk the scale-and-clock model achieves +85 % to +89 % of the full model's gain. Against the Student-t random walk calibrated on 2023, which is the fairer floor because the Student-t body is something every baseline can have, the full model's gain is 1.67 % at five minutes, 1.23 % at fifteen, 0.82 % at thirty and 0.38 % at sixty, and the scale-and-clock

model's share of it is +61 %, +48 % and +31 % at five, fifteen and thirty minutes and negative at sixty (it is +0.23 % worse than the Student-t random walk there, t 3.0). In other words, most of the headline gain over the Gaussian random walk is the heavy tail and the scale of a well-calibrated random walk; what the price-and-volume history adds on top of a scale and a clock is +0.50 % to +0.56 % of CRPS across horizons, and that is the quantity a sequence model, or a variance regression, can extract from bars.

Fusion, head, regime (H3), identity. Cross-attention fusion of separate per-scale encoders is slightly worse than the shared backbone (+0.05 % to +0.04 %, with 60 % more parameters). The Gaussian head is worse than the Student-t head by +0.32 % of unscaled CRPS at five minutes and +0.20 % at sixty minutes (day-cluster $t = 20.2$ and 6.5, Holm $p < 10^{-9}$ in the single ablation family of 68 tests). Disabling the regime embedding while the global token still carries the statistics changes CRPS by +0.02 % to -0.00 %, which says only that the same statistics entering twice are not better than once. Hypothesis 3 is tested by the control of Section 6: two models with the reduced global token, one receiving the volatility-regime statistics through the regime embedding and one receiving them nowhere, trained on the same node with the same seed, compared on the 24 stress days by the paired day-cluster difference of their absolute coverage gaps (Table 23; levels in Table B.12, Appendix). The model with the statistics has the *larger* gap at every horizon: for the 50 % interval the difference is +0.49 percentage points [+0.34, +0.65] at five minutes, +0.36 at fifteen, +0.31 at thirty and +0.16 [-0.05, +0.39] at sixty, with the interval excluding zero at every horizon but the longest; the 80 % gap moves the same way (+0.40 points at five minutes, the interval again excluding zero at every horizon but the longest) and the 95 % gap (+0.19 points at five minutes) and the PIT-variance gap exclude zero at every horizon. On the stress days the model that receives the regime statistics is farther from nominal coverage than the control at 4 of 4 horizons for both the 50 and 80 % intervals (the interval of the difference excludes zero at 3 of 4 horizons for the 50 % interval, at 3 of 4 for at least one of the two) and closer at 0 of 4. By the pre-stated rule (closer to nominal at two or more horizons) the statistics make stress-day calibration closer to nominal at 0 horizons: Hypothesis 3 is rejected, and the sign is the opposite of the hypothesis. The CRPS contrast between the two controls is +0.01 % at five minutes and -0.07 % at sixty, i.e. no consistent benefit either way, and both reduced-token models are within +0.13 % of the full reference. A learned eight-code regime codebook improves CRPS by -0.03 % to -0.16 % and lowers central coverage toward nominal; the effect is a single-seed number at the 4,000-step budget, inside the band spanned by the two extra seeds, and we do not read it as an architectural finding. A ticker-identity embedding hurts (+0.03 % to +0.25 %): instrument identity over-fits at this scale, which is also why the core model has none.

Joint paths (H4). The copula head trained on the probability-integral transforms of the Stage-A marginals and selected on the stratified validation copula likelihood is numerically the identity: on a stratified sample of validation origins the mean norm of its factor loadings is 0.082 against a mean diagonal of 2.48, an implied mean off-diagonal correlation of 0.00025. Correlated sampling accordingly changes the energy score, the variogram score and the CRPS of maximum drawdown and realised variance by amounts whose intervals contain zero (Table 24), and the CRPS of the cumulative return is +0.07 % worse at sixty minutes [+0.04, +0.10]. Hypothesis 4 is rejected for this specification. It would be wrong to conclude that the next hour's minute returns are conditionally independent given the history: the diagnostics on the model's own normal scores z_h along the sixty future minutes, on 986,468 origins with day-cluster intervals, show a first-order autocorrelation of -0.069 [-0.072, -0.066] (negative, the bid–ask bounce of bar closes; -0.005 at lag two), an autocorrelation of z_h^2 of 0.095 [0.087, 0.104] and of $|z_h|$ of 0.085 at lag one (Table 25) (within-hour volatility clustering that the marginals do not remove), a share 1.00 of days rejecting a Ljung–Box test on z_h^2 at ten lags, and a ratio of the realised variance of $R_{t,H}$ to the variance implied by independent minutes of 0.887 [0.863, 0.913] at five minutes and 0.819 [0.780, 0.863] at sixty, so the independent-minute aggregate overstates the dispersion of the cumulative return, consistent with the negative first-order dependence (the main model gives the same picture). A Gaussian copula with a constant rank-4 correlation captures none of this: it is blind to second-moment dependence and a rank-4 factor cannot represent banded short-lag structure. The correct statement is that a low-rank Gaussian copula captures no dependence that the path scores reward, while first-order and second-moment dependence of the scores exist and were not modelled; a t-copula or an ARCH-type dependence head was not tried.

9. Limitations

Survivorship and the availability filter. The universe requires six months of presence after the sample start and, more importantly, 80 % regular-hours minute coverage over 2018–2025, a filter that reads the evaluation years and removes 66 candidates that pass it on the training years (Section 4.3); only 5 delisted candidates passed it, because the vendor's delisted one-minute archives are only partly ingested. The direction of the bias is that the instruments whose trading dried up—the episodes most likely to stress calibration and transfer—are under-represented. The zero-shot scores on the 66 excluded instruments (Section 7.6) bound the effect for the names the filter removed for post-2022 reasons, but not for instruments that left the sample altogether. The results therefore describe liquid instruments that survived the window, and the transfer claim (Hypothesis 2) is a claim about unseen *liquid survivors*. The eligibility rule that requires 80 % of the next 60 slots to be present also conditions on the future, identically for every model; it removes 10.9 % of the 2024

Table 21. Learning curve of the H1 twins (P4): scaled CRPS, HAR-RV contrast and H1 gap on the 2024 development test at the 12k, 24k and 36k-step budgets

Run	Best validation NLL	$H = 5$	$H = 15$	$H = 30$	$H = 60$
Panel A. Scaled CRPS and validation NLL per twin and budget (steps reached in parentheses)					
MRPT-S (1m+5m+30m), 12k budget (12,000)	-6.45346	0.4531	0.4414	0.4311	0.4187
MRPT-S (1m+5m+30m), 24k budget (24,000)	-6.12207	0.4529	0.4413	0.4312	0.4188
MRPT-S (1m+5m+30m), 36k budget (34,000)	-6.12231	0.4528	0.4411	0.4310	0.4185
MRPT-S (1m only), 12k budget (12,000)	-6.45202	0.4533	0.4416	0.4314	0.4192
MRPT-S (1m only), 24k budget (20,000)	-6.11987	0.4532	0.4418	0.4319	0.4199
MRPT-S (1m only), 36k budget (30,000)	-6.12069	0.4526	0.4411	0.4313	0.4195
Panel B. MRPT-S (1m+5m+30m) vs HAR-RV per budget, % of HAR-RV's unscaled CRPS [day-cluster t]					
12k budget		-1.91 [-51.7]	+0.34 [8.8]	+0.53 [9.3]	+0.65 [8.9]
24k budget		-1.95 [-52.4]	+0.34 [8.3]	+0.56 [9.6]	+0.72 [9.3]
36k budget		-1.99 [-49.5]	+0.28 [6.7]	+0.49 [8.0]	+0.62 [7.4]
Panel C. H1 gap per budget: multi-resolution minus 1m-only, % of the 1m-only unscaled CRPS [day-cluster t]					
12k budget		-0.039 [-4.4]	-0.037 [-3.7]	-0.050 [-3.4]	-0.102 [-4.1]
24k budget		-0.083 [-5.8]	-0.156 [-8.5]	-0.219 [-7.0]	-0.329 [-6.1]
36k budget		+0.047 [3.8]	-0.011 [-0.9]	-0.086 [-4.5]	-0.254 [-7.5]
Five-seed range of the 12k contrast (unscaled)		[-0.183, -0.039]	[-0.342, -0.037]	[-0.450, -0.050]	[-0.549, -0.102]
36k gap inside the seed range		no	no	yes	yes

Notes. The 12k runs are the main-study models; the 24k and 36k rows are warm-restart continuations from the 12k and 24k checkpoints (optimiser and scheduler state restored, cosine schedule rebuilt to the new budget, learning rate at the restart $\approx 0.56\times$ the peak), early-stopped with patience 4 (1,000-step evaluations) on the stratified 50,000-origin 2023 validation subset; the steps actually reached are in parentheses (the 36k budget stopped at 34k and 30k, the 1m-only 24k budget at 20k). Validation NLL of the 12k rows is on the original 3,840-origin subset and is not comparable with the stratified values of the continuations. Panels B and C are on unscaled CRPS with the day-cluster t in brackets (1,000 trading-day resamples); the seed range is the minimum and maximum of the five-seed 12k contrast of Table 10 on the same convention. The pre-stated budget verdict (both criteria: 36k gap inside the seed range at every horizon, and the sign of the HAR-RV contrast at 30 and 60 minutes unchanged) decides whether the 12,000-step budget condition leaves the abstract. Source: `results/metrics/budget/learning_curve.csv`, `budget_verdict.json`.

Table 22. Where the gain over the random walk comes from: the time-only model (calendar, missingness and $\log \sigma_t$ only, no price or volume history) against the full 4k reference, measured from two reference points (2024, unscaled CRPS)

	$H = 5$	$H = 15$	$H = 30$	$H = 60$
Panel A. Reference: Student-t random walk with scale and ν fitted on 2023 (the floor every model should clear)				
Gain of the full model (4k reference), % of the reference CRPS	+1.67	+1.23	+0.82	+0.38
Gain of the time-only model (reduced global token), %	+1.03	+0.60	+0.25	-0.14
Share of the full gain achieved by time only	+61 %	+48 %	+31 %	-36 %
Remainder attributable to the price/volume history, % of the reference CRPS	+0.65	+0.64	+0.56	+0.52
t_{cl}: time only vs full; time only vs reference	13.1; -19.8	11.5; -10.0	8.8; -3.9	6.8; 1.9
Panel B. Reference: Gaussian random walk with the trailing one-minute scale (the v1.0 reference)				
Gain of the full model (4k reference), % of the reference CRPS	+4.10	+4.12	+4.15	+4.34
Gain of the time-only model (reduced global token), %	+3.47	+3.50	+3.61	+3.85
Share of the full gain achieved by time only	+85 %	+85 %	+87 %	+89 %
Remainder attributable to the price/volume history, % of the reference CRPS	+0.63	+0.62	+0.54	+0.50
t_{cl}: time only vs full; time only vs reference	13.1; -41.3	11.5; -41.6	8.8; -40.0	6.8; -33.7

Notes. Gain = $100 \times (\bar{L}_{\mathrm{ref}} - \bar{L}_{\mathrm{model}})/\bar{L}_{\mathrm{ref}}$ on unscaled CRPS over the 986,468 origins of 2024; share = gain of the time-only model divided by the gain of the full model (a negative share means the time-only model is worse than the reference). The time-only model here receives the reduced global token of Table 23 (not the HAR regressors of the v1.0 variant); the Student-t random walk is the Gaussian random walk with a scalar scale and degrees of freedom chosen on the 2023 validation origins (`scripts/recalibrate.py`). Day-cluster t-statistics from the pairwise file (1,000 trading-day resamples). Source: `results/metrics/ablations_h3/decomposition.csv`.

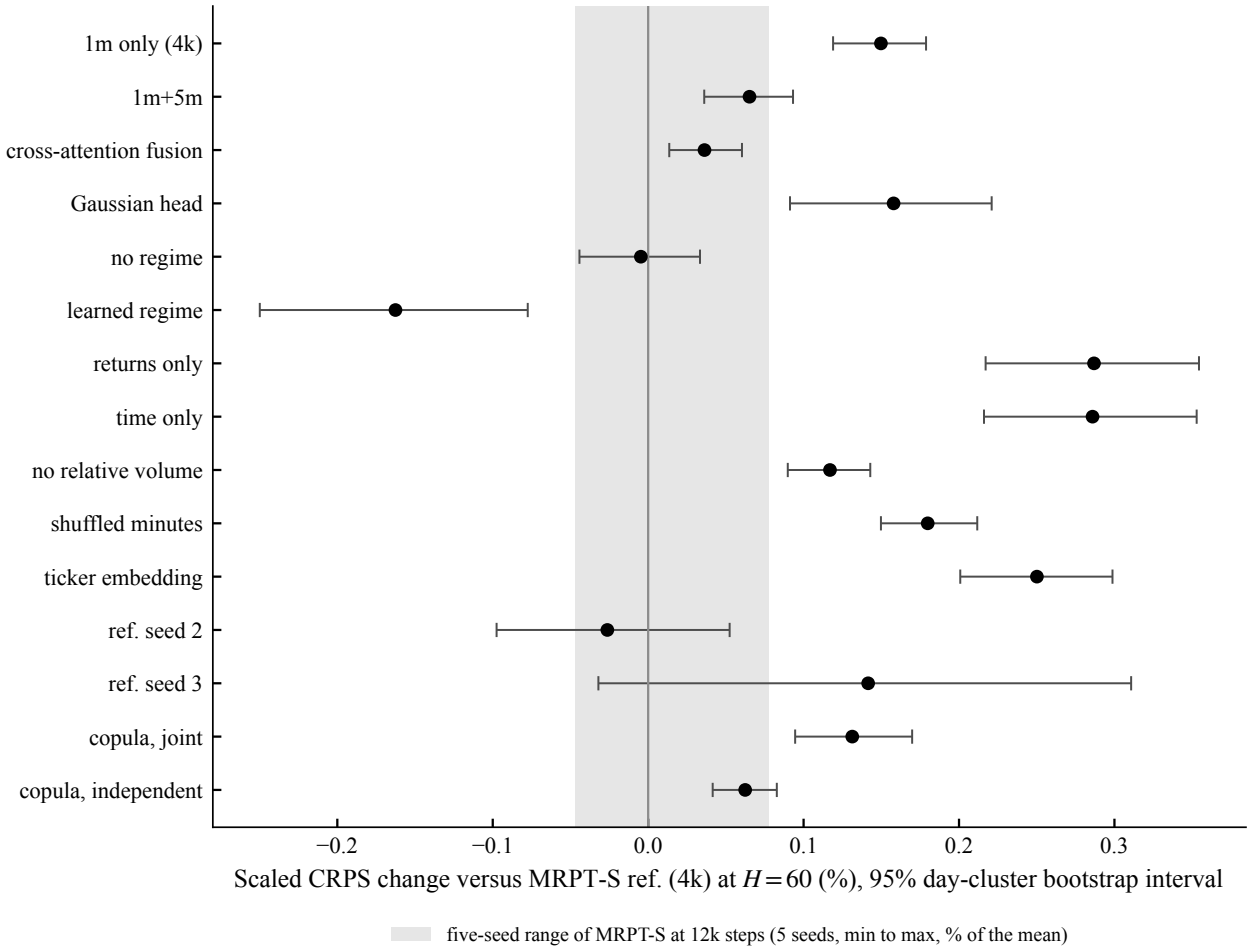


Figure 5. Ablations at the Small scale (4,000 steps each): relative change in scaled CRPS at $H = 60$ versus the equal-budget reference, with 95 % day-cluster bootstrap intervals; the two extra seeds of the reference and the five-seed band of the 12,000-step MRPT-S are shown as the noise floors.

origins, and its effect on the scores was not measured.

Bars are not quotes. FirstRate bars carry no bid–ask spread, trades or order flow. The high-frequency predictability literature finds large and pervasive short-horizon predictability from trades-and-quotes predictors (Ait-Sahalia et al., 2026; Lucchese et al., 2024; Kolm et al., 2023); the directional accuracy of 0.518 for the linear and tree location forecasts at five minutes, decaying to one half by sixty minutes and not surviving a one-basis-point cost with a one-minute delay, is consistent with the reading that the exploitable signal lives in quotes and order flow and that bar closes show its bounce. That reading is a hypothesis for data this paper does not have, not a finding; the economic evaluation of Hypothesis 5 rests on assumed costs and bar-close execution.

Table 23. Hypothesis 3 on an unconfounded control: regime statistics entering the decoder versus the same network without them, both with the global token reduced to $\log \sigma_t$ and the calendar (Small, 4,000 steps, seed 1, 2024)

	$H = 5$	$H = 15$	$H = 30$	$H = 60$
Panel A. Stress days (24 sessions, 100846 *origins): |gap| regime-statistics model − |gap| no-regime model, pp [95 % CI]*				
Central 50 % interval	+0.49 [+0.34, +0.65]	+0.36 [+0.20, +0.54]	+0.31 [+0.15, +0.50]	+0.16 [-0.05, +0.39]
Central 80 % interval	+0.40 [+0.26, +0.54]	+0.38 [+0.23, +0.53]	+0.33 [+0.16, +0.52]	+0.14 [-0.03, +0.33]
Central 95 % interval	+0.19 [+0.12, +0.25]	+0.13 [+0.07, +0.19]	+0.11 [+0.05, +0.17]	+0.08 [+0.01, +0.16]
PIT variance (nominal 1/12)	+0.0009 [+0.0006, +0.0012]	+0.0007 [+0.0004, +0.0010]	+0.0007 [+0.0003, +0.0010]	+0.0004 [+0.0001, +0.0008]
Gaps on stress days, regime-statistics model (50/80/95, pp)	+4.5/+3.1/+0.9	+6.2/+4.6/+1.3	+6.8/+5.1/+1.5	+7.8/+6.0/+2.0
Gaps on stress days, no-regime model (50/80/95, pp)	+4.1/+2.7/+0.7	+5.9/+4.2/+1.1	+6.5/+4.8/+1.3	+7.6/+5.9/+1.9
Panel B. All 2024 days: the same |gap| differences, pp [95 % CI]				
Central 50 % interval	+0.39 [+0.35, +0.44]	+0.17 [+0.12, +0.22]	+0.10 [+0.06, +0.15]	-0.05 [-0.11, +0.01]
Central 80 % interval	+0.32 [+0.28, +0.37]	+0.16 [+0.12, +0.20]	+0.08 [+0.04, +0.13]	-0.04 [-0.08, +0.01]
Central 95 % interval	+0.16 [+0.14, +0.18]	+0.06 [+0.04, +0.09]	+0.03 [+0.01, +0.05]	-0.03 [-0.05, -0.00]
Panel C. Scaled CRPS contrasts of the same runs, all 2024 origins, % [t_{cl}], Holm stars				
Regime statistics vs no regime (both with the reduced global token)	+0.015 [2.1]	-0.004 [-0.5]	-0.023 [-2.3]	-0.067 [-6.1]***
Regime statistics, reduced global token vs the 4k reference	-0.045 [-5.6]***	-0.031 [-3.0]**	-0.000 [-0.0]	+0.059 [2.7]**
No regime, reduced global token vs the 4k reference	-0.060 [-7.4]***	-0.027 [-2.6]*	+0.022 [1.5]	+0.126 [5.2]***
Shuffled price/volume columns (calendar in place) vs the 4k reference	+0.088 [9.0]***	+0.108 [7.0]***	+0.128 [6.9]***	+0.197 [10.0]***
Time only, reduced global token vs the 4k reference	+0.561 [12.6]***	+0.560 [11.7]***	+0.517 [8.9]***	+0.502 [7.1]***

Notes. The pre-registered test of H3 compares, on stress days, the absolute deviation of the empirical coverage from its nominal level (the 'coverage gap') of the regime-conditioned model with that of the control; a negative difference means the regime statistics bring coverage closer to nominal. In the v1.0 ablation the control still received the nine regime statistics through the global token; here both runs use a global token reduced to [$\log \sigma_t$, minute fraction, time to close, weekday, early-close flag] and differ only in the regime conditioning. Intervals: paired trading-day cluster bootstrap of the |gap| difference (1,000 resamples); PIT variance gap = $|\mathrm{Var}(\mathrm{PIT}) - 1/12|$. Panel C: scaled CRPS contrasts with the day-cluster t and Holm-adjusted stars over the H3 family (32 tests): $^{***}\ p < 0.01$, $^{**}\ p < 0.05$, $^{*}\ p < 0.10$. The rule of the improvement plan (supported when the regime-conditioned model is closer to nominal at ≥ 2 horizons for the 50 and 80 % intervals) is evaluated in the text. Source: `results/metrics/ablations_h3/stress_contrast_test.csv`, `ablations_h3/pairwise_test.csv`.

Table 24. Joint paths versus independent sampling of the same Student-t marginals (MRPT-S + copula, 2024 development test, 986468 origins, 60-minute paths, 200 draws)

Score	Joint (copula)	Independent	Joint − independent
Energy score	3.4903e-03	3.4904e-03	-9.2×10^{-8} [-2.3×10^{-7}, $+5.0 \times 10^{-8}$]
Variogram score ($p = 0.5$)	1.5113e-04	1.5114e-04	-3.8×10^{-9} [-1.0×10^{-8}, $+2.2 \times 10^{-9}$]
CRPS, maximum drawdown	1.4632e-03	1.4633e-03	-1.8×10^{-7} [-6.2×10^{-7}, $+2.8 \times 10^{-7}$]
CRPS, realised variance	1.1881e-05	1.1869e-05	$+1.2 \times 10^{-8}$ [-2.8×10^{-9}, $+3.0 \times 10^{-8}$]

Notes. Means over the evaluation origins; brackets are 95 % trading-day cluster-bootstrap intervals of the paired difference (joint minus independent sampling of the same marginals). Energy and variogram scores are multivariate proper scores of the 60-minute path; the CRPS of maximum drawdown and of realised variance are path functionals. Source: `results/metrics/ablations/path_scores_mrpt_small_copula_test.json` (generated).

Deviations from the pre-registered protocol. Section 2 states each hypothesis in its pre-registered wording and, where the revision sharpened it to the executed test, the operationalisation adopted in revision: the H1 parameter clause ("equal" read as within 7.4 %), the two clauses of H2, the H3 control that receives the regime statistics through no input, the rank-4 copula of H4, and the seven location models with a one-minute delay of H5. The primary test statistic was changed in revision from the origin-level HAC Diebold–Mariano statistic of v1.0 to the day-cluster t with Holm families, before the seed study was run and because the HAC statistic overstates the evidence (Table B.8); the v1.0 statistic is kept in every result file. The five-seed replication of H1, its scoring on 2025, the learning curve, the second held-out set, the recalibration companion, the unconfounded H3 control and the consistent baseline construction were added in revision, after the v1.0 results were known,

Table 25. Dependence diagnostics of the probability-integral transforms (H4): what the rank-4 Gaussian copula could and could not model (2024 development test)

	lag 1	lag 2	lag 3	lag 10
Panel A. Autocorrelation of the PIT normal scores along the path [95 % CI]				
Copula model, z_h	-0.069 [-0.072, -0.066]	-0.005 [-0.007, -0.003]	-0.003 [-0.005, -0.001]	+0.001 [-0.000, +0.003]
Copula model, z_h^2	+0.095 [+0.087, +0.104]	+0.067 [+0.059, +0.076]	+0.062 [+0.053, +0.071]	+0.038 [+0.032, +0.044]
Copula model, $\lvert z_h \rvert$	+0.085 [+0.080, +0.089]	+0.052 [+0.048, +0.057]	+0.047 [+0.043, +0.052]	+0.030 [+0.027, +0.034]
Main MRPT-S, z_h	-0.069 [-0.072, -0.065]	-0.005 [-0.007, -0.003]	-0.003 [-0.005, -0.001]	+0.001 [-0.001, +0.003]
Main MRPT-S, z_h^2	+0.089 [+0.084, +0.096]	+0.062 [+0.056, +0.069]	+0.057 [+0.051, +0.063]	+0.034 [+0.030, +0.038]
Main MRPT-S, $\lvert z_h \rvert$	+0.082 [+0.078, +0.086]	+0.050 [+0.046, +0.054]	+0.045 [+0.041, +0.049]	+0.028 [+0.026, +0.031]
Panel B. Variance ratio, realised over implied by independent minutes [95 % CI]				
	$H = 5$	$H = 15$	$H = 30$	$H = 60$
Copula model	0.887 [0.863, 0.913]	0.847 [0.823, 0.875]	0.835 [0.801, 0.874]	0.819 [0.780, 0.863]
Main MRPT-S	0.877 [0.854, 0.901]	0.835 [0.809, 0.863]	0.820 [0.787, 0.860]	0.805 [0.768, 0.850]
Panel C. Other diagnostics				
Share of days rejecting Ljung–Box $Q(10)$ on z_h^2 at 5 %	1.00			
Variance of the per-minute PIT (1/12 = 0.0833 if calibrated)	0.0809			
Copula head: mean $\lVert \lambda_h \rVert$, mean ψ, implied off-diagonal correlation	0.082, 2.48, 0.00025 (20000 stratified validation origins)			

Notes. For every origin of the 2024 development test (986468 origins over 251 sessions) and every future minute $h = 1, \ldots, 60$ the normal score $z_h = \Phi^{-1}(F_h(y_h))$ is computed from the model's Student-t marginal. Panel A: autocorrelation of z_h, z_h^2 and $\lvert z_h \rvert$ at lag k along the path, pooled over origins; the first row is the copula model (MRPT-S + copula, the marginals the copula head was trained on), the second the main MRPT-S model. A Gaussian copula with a constant correlation matrix can represent linear dependence of z_h only; dependence of z_h^2 and $\lvert z_h \rvert$ (second-moment, within-hour volatility clustering) and tail dependence are outside its reach. Panel B: realised variance of the H-minute cumulative return over the variance implied by independent minutes; a ratio below one is what negative first-order dependence produces. Intervals: trading-day cluster bootstrap (1,000 resamples). Panel C: the retrained copula head (validation = copula NLL on 50,000 stratified 2023 origins) selects the same checkpoint as the original; its loadings imply a correlation matrix numerically equal to the identity. Source: `results/metrics/ablations/copula_diagnostics.csv`, `.json`.

and are labelled as such wherever they appear; none of them changed a direction or a threshold. The 2025 period was scored once per model at every stage.

One budget, few seeds. The verdict on Hypothesis 1—supported in 2024 at a small magnitude—is established at a 12,000-step budget (1.31 passes) with no hyper-parameter search, replicated over five seeds, and followed along a warm-restart learning curve that early stopping ended at 34,000 steps; fresh runs with a cosine schedule matched to the larger budgets, a learning-rate sweep, a continuation of the Base model beyond 12,000 steps and seeds for the regime, fusion and feature ablations were not run. The size statement is therefore "at a matched 12,000-step budget", the ablation effects at 4,000 steps are single-seed numbers read against a two-seed floor, and although the eight seed checkpoints were scored once on 2025 (Table 10, Panel D), the 2025 evidence is one year of five seed pairs whose dispersion is too large for the pre-stated rule: the non-replication is a failure to detect, not a detected absence. The 2024 development test carries every headline table and every ablation; the hypotheses, tests and budgets were fixed before it was used, but no result in Sections 7–8 other than the 2025 columns is out of sample with respect to the choice of what to report.

Validation subsample. The main runs selected checkpoints on 3,840 index-ordered 2023 origins of a single instrument (Section 5.5). The re-score on a stratified sample shows that no 12,000-step model would have selected a different checkpoint, but 8 of the 15 short-budget ablations would have,

and those ablations were not re-predicted.

Baselines. HAR-RV, ridge, elastic net and LightGBM are fitted once on 2018–2022 and frozen through 2025; an expanding-window yearly refit, which could only help them, was not run, nor was a maximum-likelihood estimate of their degrees of freedom (the method-of-moments values are in Table A.1). The corrected LightGBM remains mis-specified at five minutes (Section 6.1); its five-minute row should be read as a property of the tree variance model under this retransformation. Pretrained time-series foundation models, Kronos, iTransformer and EGARCH were not evaluated (Section 6); their absence is a limitation of this version, not evidence about them.

Dependence across minutes. Hypothesis 4 was tested with a rank-4 Gaussian copula with a constant correlation, which cannot represent the second-moment and banded short-lag dependence that the PIT diagnostics of Section 8 show to exist; a t-copula, an ARCH-type dependence head or a diffusion decoder were not tried. The statement that coherent paths add nothing is specific to that copula.

Regular hours only. Origins whose horizon crosses the close are excluded, extended hours are not studied, and the closing print is excluded from context and targets. The model has never seen the first thirty minutes as an origin (the first origin is at 10:00), although those minutes are in its context.

Missingness and targets. Missing minutes are masked rather than filled; the target of a minute that follows a gap is the return accrued since the last observed bar, so its scale is slightly larger than a one-minute return. With coverage above 80 % and the future-presence rule this affects a small fraction of targets, and it affects all models identically, but it is a modelling convention, not a property of the market.

Corporate actions. Splits are applied explicitly; dividends are not, and the overnight gap across an ex-dividend date is left raw. Mergers and spin-offs end a series without notice.

Compute and reproduction tolerance. All runs used Apple Silicon (MPS) without mixed precision; the review runs were spread over several nodes (Table 4) and are compared at matched optimiser steps. The deterministic baselines reproduce across chips exactly; a retrained neural model reproduces to within the seed band of Section 7.2, which is the tolerance a replicator should expect, and MPS against CPU or CUDA kernels was not measured. The brief's upper range (50–100 million parameters) was not explored.

Multiplicity and overlap. Evaluation origins overlap within a session and are cross-sectionally dependent; all decisions rest on day-cluster statistics with Holm adjustment inside named families, and the origin-level HAC Diebold–Mariano statistic is reported only as the anticonservative check it is. The effective sample size is far smaller than the number of origins suggests, and the minimum detectable effect of Section 2 is the right scale against which to read every null in this paper.

10. Reproducibility and release

The repository `wp41_uqo` contains the package `chronos_x` (data connector, calendar, quality checks, corporate actions, grids and windows; models; training; scoring and inference; baselines), the numbered scripts that produced every artifact cited here, the configurations that freeze the protocol (`configs/pilot`, `main`, `ablations`, `baselines`, `final`, and the review configurations for the seeds, learning curve, H3 controls, transfer run and second held-out set), the test suite (calendar and timestamp conventions, grid construction on synthetic bars, leakage gate with a quarantined future-feature fixture, model shapes and parameter counts, numerical checks of the Student-t density and CRPS against Monte Carlo, copula sampling and likelihood, the cluster-t, daily Diebold–Mariano and model-confidence-set routines, live API contract tests), the universe manifests with their hashes and the instrument-master version, the provenance records of every API call (endpoint, parameters, dataset version and request identifier; no bars), the phase reports, and the LaTeX source whose tables, figures and macros—every number in the text included—are generated from the saved results by the scripts named in `docs/REPRODUCIBILITY.md`. A synthetic cache in the exact FinData schema lets anyone run the entire pipeline without a license (`configs/synthetic_smoke.yaml`). Raw bars, grids and per-origin returns never leave the compute node; the released artifacts are code, aggregate scores and figures of scores. The FinData dataset versions used are 17375 (stocks) and 5162 (ETFs); the API audit, the coverage audit and the vendor-flag counts are saved with the results.

Every inferential exhibit of this paper depends on per-origin score arrays that are too large for the repository. They are packed in a restricted archive (171 files, 27.6 GB: per-origin CRPS, WIS, PIT, coverage indicators, QLIKE and tail Brier scores per horizon, with an origin index reduced to the symbol, session, slot and trailing scale; no returns, no bars, no parameters), from which the pairwise tests of the main study regenerate with a maximum difference of 0.0 over 48 rows. The archive is available from the author under the FirstRate licence terms. The environment is pinned (`uv.lock` and a frozen requirement list; Python 3.12.15, PyTorch 2.15.0 with MPS, macOS 26.4) and recorded per node. Reproduction tolerances are measured, not assumed: tables and macros regenerated from the committed aggregates are byte-identical; the deterministic baselines re-run on a different chip reproduce the committed scores exactly; a retrained neural model reproduces to within the five-seed band of Table B.9. Seeds, checkpoints, the per-node ledger of every review run

and the exact commands are listed in `docs/REPRODUCIBILITY.md`.

11. Conclusion

We asked whether a compact, transferable, multi-resolution Transformer can learn minute-level structure that improves calibrated forecasts of the next 5 to 60 minutes, and we tested the question against matched baselines, with proper scoring rules and day-cluster inference, on 234 million one-minute bars of 300 U.S. stocks and ETFs, with 60 unseen instruments, 66 instruments the availability filter had removed, and an untouched year. The answers come with their uncertainty. Minute-level history improves distributional forecasts over a Gaussian trailing-scale random walk by 2.50 to 4.14 % of scaled CRPS for every model except the corrected LightGBM at five minutes, but most of that is the heavy tail and the scale: over a Student-t random walk calibrated on the validation year the full 12,000-step model's gain is 1.54 % of scaled CRPS at five minutes and 0.30 % at sixty (Table 8). The improvement is a dispersion improvement: the neural location forecasts have no directional content, the linear and tree location forecasts have a small one (0.518 at five minutes) that does not survive a one-basis-point cost with a one-minute delay, and no cell of the economic test is positive in both evaluation years. With baselines built under one consistent retransformation, the sequence models and a minute-scale HAR-RV with a kurtosis-matched Student-t differ by -0.07 % at five minutes, a lead of a few tenths of a percent for the sequence models appears only after every model is rescaled on the validation year, and HAR-RV is sharper and better calibrated from fifteen minutes on, by 0.21–0.49 %. The multi-resolution context that defines the architecture lowers scaled CRPS against the one-minute-only twin by -0.081 % to -0.214 % on average over five seeds, negative in every seed pair and detectable by the pre-registered rule at every horizon in 2024, so that Hypothesis 1 is supported at that magnitude—1.5–3.4 times the seed half-band—at a 12,000-step budget that a warm-restart learning curve did not overturn and did not extend; in the untouched year the same checkpoints keep the sign (-0.116 % to -0.046 %) but fail the rule, so the verdict is a 2024 result that 2025 does not confirm. Model size does not matter between 3.5 and 20.7 million parameters at that budget. Regime statistics worsen rather than improve stress-day calibration once the control does not receive them elsewhere. A rank-4 Gaussian copula adds nothing to independent sampling, while the model's own PIT scores show first-order dependence of -0.069 and second-moment dependence of 0.095 at one minute that it cannot represent. Pooled training transfers to unseen liquid instruments without penalty, and on the instruments whose liquidity fell the ordering of the main study survives with wider gaps.

The central empirical statement of this paper is therefore not that a Transformer predicts financial markets, nor that it cannot. It is that, for liquid U.S. instruments at the one-minute scale and with bars as the only input, what minute-level history adds to a calibrated heavy-tailed random walk is

1.54 % of scaled CRPS at five minutes and 0.30 % at sixty for the 12,000-step model, that a variance regression extracts it at least as well as a sequence model beyond five minutes, and that reading coarser past-only aggregates on top of the last 390 bars adds a further -0.081 % to -0.214 % in 2024—detectable, a few multiples of the seed half-band, and absent in 2025. Three directions follow from the limitations rather than from the results: order-book or quote data, which the literature suggests carry the microstructure signal that bars only show the bounce of; the delisted instruments that this universe could not include; and a dependence model that can see the second-moment structure the Gaussian copula cannot. The package, the protocol, the per-origin score archive and these results are released so that all three can be tested the same way.

References

Ait-Sahalia, Yacine, Jianqing Fan, Lirong Xue, and Xiaonan Zhu, "How and When Are High-Frequency Stock Returns Predictable?," *Management Science*, 2026, *72* (9), 7792–7815.

Andersen, Torben G. and Tim Bollerslev, "Intraday periodicity and volatility persistence in financial markets," *Journal of Empirical Finance*, 1997, *4* (2-3), 115–158.

_ , _ , Francis X. Diebold, and Paul Labys, "Modeling and Forecasting Realized Volatility," *Econometrica*, 2003, *71* (2), 579–625.

Ansari, Abdul Fatir, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang, "Chronos: Learning the Language of Time Series," *Transactions on Machine Learning Research*, 2024. arXiv:2403.07815.

Ba, Jimmy Lei, Jamie Ryan Kiros, and Geoffrey E. Hinton, "Layer Normalization," arXiv:1607.06450 2016.

Bai, Shaojie, J. Zico Kolter, and Vladlen Koltun, "An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling," arXiv:1803.01271 2018.

Bailey, David H. and Marcos López de Prado, "The Deflated Sharpe Ratio: Correcting for Selection Bias, Backtest Overfitting, and Non-Normality," *The Journal of Portfolio Management*, 2014, *40* (5), 94–107.

Barndorff-Nielsen, Ole E. and Neil Shephard, "Econometric Analysis of Realized Volatility and its Use in Estimating Stochastic Volatility Models," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2002, *64* (2), 253–280.

Benjamini, Yoav and Yosef Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 1995, *57* (1), 289–300.

Bollerslev, Tim, "Generalized autoregressive conditional heteroskedasticity," *Journal of Econometrics*, 1986, *31* (3), 307–327.

_ , Andrew J. Patton, and Rogier Quaedvlieg, "Exploiting the errors: A simple approach for improved volatility forecasting," *Journal of Econometrics*, 2016, *192* (1), 1–18.

Bracher, Johannes, Evan L. Ray, Tilmann Gneiting, and Nicholas G. Reich, "Evaluating epidemic forecasts in an interval format," *PLOS Computational Biology*, 2021, *17* (2), e1008618.

Brini, Alessio, "Forecasting Realized Volatility with Time Series Foundation Models: A Comparison with Econometric Benchmarks," arXiv:2607.05291 2026.

Chen, Peng, Yingying Zhang, Yunyao Cheng, Yang Shu, Yihang Wang, Qingsong Wen, Bin Yang, and Chenjuan Guo, "Pathformer: Multi-scale Transformers with Adaptive Pathways for Time Series Forecasting," in "International Conference on Learning Representations (ICLR 2024)" 2024. arXiv:2402.05956.

Christensen, Kim, Mathias Siggaard, and Bezirgen Veliyev, "A Machine Learning Approach to Volatility Forecasting," *Journal of Financial Econometrics*, 2023, *21* (5), 1680–1727.

Cont, R., "Empirical properties of asset returns: stylized facts and statistical issues," *Quantitative Finance*, 2001, *1* (2), 223–236.

Corsi, F., "A Simple Approximate Long-Memory Model of Realized Volatility," *Journal of Financial Econometrics*, 2009, *7* (2), 174–196.

Das, Abhimanyu, Weihao Kong, Rajat Sen, and Yichen Zhou, "A decoder-only foundation model for time-series forecasting," in "International Conference on Machine Learning (ICML 2024)" 2024, pp. 10148–10167. arXiv:2310.10688.

Dauphin, Yann N., Angela Fan, Michael Auli, and David Grangier, "Language Modeling with Gated Convolutional Networks," in "International Conference on Machine Learning (ICML 2017)" 2017, pp. 933–941. arXiv:1612.08083.

Diebold, Francis X. and Roberto S. Mariano, "Comparing Predictive Accuracy," *Journal of Business & Economic Statistics*, 1995, *13* (3), 253–263.

Du, Linfeng, Ji Xin, Alex Labach, Saba Zuberi, Maksims Volkovs, and Rahul G. Krishnan, "MultiResFormer: Transformer with Adaptive Multi-Resolution Modeling for General Time Series Forecasting," arXiv:2311.18780 2023.

Engle, Robert, "GARCH 101: The Use of ARCH/GARCH Models in Applied Econometrics," *Journal of Economic Perspectives*, 2001, *15* (4), 157–168.

Fu, Xinghong, Masanori Hirano, and Kentaro Imajo, "Financial Fine-tuning a Large Time Series Model," in "2025 IEEE Symposium on Computational Intelligence for Financial Engineering and Economics (CIFEr)" 2025, pp. 1–9. arXiv:2412.09880.

Giacomini, Raffaella and Halbert White, "Tests of Conditional Predictive Ability," *Econometrica*, 2006, *74* (6), 1545–1578.

Gneiting, Tilmann and Adrian E Raftery, "Strictly Proper Scoring Rules, Prediction, and Estimation," *Journal of the American Statistical Association*, 2007, *102* (477), 359–378.

_ **and Roopesh Ranjan**, "Comparing Density Forecasts Using Threshold- and Quantile-Weighted Scoring Rules," *Journal of Business & Economic Statistics*, 2011, *29* (3), 411–422.

_ **, Fadoua Balabdaoui, and Adrian E. Raftery**, "Probabilistic Forecasts, Calibration and Sharpness," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2007, *69* (2), 243–268.

Gu, Shihao, Bryan Kelly, and Dacheng Xiu, "Empirical Asset Pricing via Machine Learning," *The Review of Financial Studies*, 2020, *33* (5), 2223–2273.

Hansen, Peter R. and Asger Lunde, "A forecast comparison of volatility models: does anything beat a GARCH(1,1)?," *Journal of Applied Econometrics*, 2005, *20* (7), 873–889.

_ **,** _ **, and James M. Nason**, "The Model Confidence Set," *Econometrica*, 2011, *79* (2), 453–497.

Harvey, David, Stephen Leybourne, and Paul Newbold, "Testing the equality of prediction mean squared errors," *International Journal of Forecasting*, 1997, *13* (2), 281–291.

Hendrycks, Dan and Kevin Gimpel, "Gaussian Error Linear Units (GELUs)," arXiv:1606.08415 2016.

Holm, Sture, "A Simple Sequentially Rejective Multiple Test Procedure," *Scandinavian Journal of Statistics*, 1979, *6* (2), 65–70. JSTOR stable 4615733.

Jordan, Alexander, Fabian Krüger, and Sebastian Lerch, "Evaluating Probabilistic Forecasts with scoringRules," *Journal of Statistical Software*, 2019, *90* (12).

Ke, Guolin, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in "Advances in Neural Information Processing Systems 30" 2017, pp. 3146–3154.

Kelly, Bryan and Dacheng Xiu, "Financial Machine Learning," *Foundations and Trends® in Finance*, 2023, *13* (3-4), 205–363.

Kolm, Petter N., Jeremy Turiel, and Nicholas Westray, "Deep order flow imbalance: Extracting alpha at multiple horizons from the limit order book," *Mathematical Finance*, 2023, *33* (4), 1044–1081.

Lim, Bryan and Stefan Zohren, "Time-series forecasting with deep learning: a survey," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 2021, *379* (2194), 20200209.

Liu, Yong, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long, "iTransformer: Inverted Transformers Are Effective for Time Series Forecasting," in "International Conference on Learning Representations (ICLR 2024)" 2024. arXiv:2310.06625.

López de Prado, Marcos, *Advances in Financial Machine Learning*, Hoboken, NJ: Wiley, 2018.

Loshchilov, Ilya and Frank Hutter, "Decoupled Weight Decay Regularization," in "International Conference on Learning Representations (ICLR 2019)" 2019. arXiv:1711.05101.

Lucchese, Lorenzo, Mikko S. Pakkanen, and Almut E.D. Veraart, "The short-term predictability of returns in order book markets: A deep learning perspective," *International Journal of Forecasting*, 2024, *40* (4), 1587–1621.

Matheson, James E. and Robert L. Winkler, "Scoring Rules for Continuous Probability Distributions," *Management Science*, 1976, *22* (10), 1087–1096.

Moreno-Pino, Fernando and Stefan Zohren, "DeepVol: Volatility Forecasting from High-Frequency Data with Dilated Causal Convolutions," arXiv:2210.04797 2022.

Nelson, Daniel B., "Conditional Heteroskedasticity in Asset Returns: A New Approach," *Econometrica*, 1991, *59* (2), 347–370.

Newey, Whitney K. and Kenneth D. West, "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 1987, *55* (3), 703.

Nie, Yuqi, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam, "A Time Series is Worth 64 Words: Long-term Forecasting with Transformers," in "International Conference on Learning Representations (ICLR 2023)" 2023. arXiv:2211.14730.

Oreshkin, Boris N., Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio, "N-BEATS: Neural basis expansion analysis for interpretable time series forecasting," in "International Conference on Learning Representations (ICLR 2020)" 2020. arXiv:1905.10437.

Patton, Andrew J., "Modelling Asymmetric Exchange Rate Dependence," *International Economic Review*, 2006, *47* (2), 527–556.

_ , "Volatility forecast comparison using imperfect volatility proxies," *Journal of Econometrics*, 2011, *160* (1), 246–256.

Pineda Arango, Sebastian, Pedro Mercado, Shubham Kapoor, Abdul Fatir Ansari, Lorenzo Stella, Huibin Shen, Hugo Senetaire, Caner Turkmen, Oleksandr Shchur, Danielle C. Maddix, Michael Bohlke-Schneider, Yuyang Wang, and Syama Sundar Rangapuram, "ChronosX: Adapting Pretrained Time Series Models with Exogenous Variables," in "International Conference on Artificial Intelligence and Statistics (AISTATS 2025)" 2025, pp. 2242–2250. arXiv:2503.12107.

Rahimikia, Eghbal, Hao Ni, and Weiguan Wang, "Re(Visiting) Time Series Foundation Models in Finance," arXiv:2511.18578 2025.

Rasul, Kashif, Calvin Seward, Ingmar Schuster, and Roland Vollgraf, "Autoregressive Denoising Diffusion Models for Multivariate Probabilistic Time Series Forecasting," in "International Conference on Machine Learning (ICML 2021)" 2021, pp. 8857–8868. arXiv:2101.12072.

Reisenhofer, Rafael, Xandro Bayer, and Nikolaus Hautsch, "HARNet: A Convolutional Neural Network for Realized Volatility Forecasting," SSRN working paper 4116642 2022.

Salinas, David, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski, "DeepAR: Probabilistic forecasting with autoregressive recurrent networks," *International Journal of Forecasting*, 2020, *36* (3), 1181–1191.

Scheuerer, Michael and Thomas M. Hamill, "Variogram-Based Proper Scoring Rules for Probabilistic Forecasts of Multivariate Quantities," *Monthly Weather Review*, 2015, *143* (4), 1321–1334.

Shabani, Amin, Amir Abdi, Lili Meng, and Tristan Sylvain, "Scaleformer: Iterative Multi-scale Refining Transformers for Time Series Forecasting," in "International Conference on Learning Representations (ICLR 2023)" 2023. arXiv:2206.04038.

Shi, Yu, Zongliang Fu, Shuo Chen, Bohan Zhao, Wei Xu, Changshui Zhang, and Jian Li, "Kronos: A Foundation Model for the Language of Financial Markets," *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026, *40* (30), 25366–25373.

Sirignano, Justin and Rama Cont, "Universal features of price formation in financial markets: perspectives from deep learning," *Quantitative Finance*, 2019, *19* (9), 1449–1459.

Sklar, Abe, "Fonctions de répartition à *n* dimensions et leurs marges," *Publications de l'Institut de Statistique de l'Université de Paris*, 1959, *8*, 229–231.

Taillardat, Maxime, Olivier Mestre, Michaël Zamo, and Philippe Naveau, "Calibrated Ensemble Forecasts Using Quantile Regression Forests and Ensemble Model Output Statistics," *Monthly Weather Review*, 2016, *144* (6), 2375–2393.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention Is All You Need," in "Advances in Neural Information Processing Systems 30 (NeurIPS 2017)" 2017, pp. 5998–6008. arXiv:1706.03762.

Wen, Qingsong, Tian Zhou, Chaoli Zhang, Weiqi Chen, Ziqing Ma, Junchi Yan, and Liang Sun, "Transformers in Time Series: A Survey," in "Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence" 2023, pp. 6778–6786.

Xiao, Congxi, Jingbo Zhou, Yixiong Xiao, Xinjiang Lu, Le Zhang, and Hui Xiong, "TimeFound: A Foundation Model for Time Series Forecasting," arXiv:2503.04118 2025.

Zeng, Ailing, Muxi Chen, Lei Zhang, and Qiang Xu, "Are Transformers Effective for Time Series Forecasting?," *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023, *37* (9), 11121–11128.

Zhang, Chao, Yihuang Zhang, Mihai Cucuringu, and Zhongmin Qian, "Volatility Forecasting with Machine Learning and Intraday Commonality," *Journal of Financial Econometrics*, 2024, *22* (2), 492–530.

Zhang, Yitian, Liheng Ma, Soumyasundar Pal, Yingxue Zhang, and Mark Coates, "Multi-resolution Time-Series Transformer for Long-term Forecasting," in "International Conference on Artificial Intelligence and Statistics (AISTATS 2024)" 2024, pp. 4222–4230. arXiv:2311.04147.

Zhang, Zihao, Stefan Zohren, and Stephen Roberts, "DeepLOB: Deep Convolutional Neural Networks for Limit Order Books," *IEEE Transactions on Signal Processing*, 2019, *67* (11), 3001–3012.

Zou, Hui and Trevor Hastie, "Regularization and Variable Selection Via the Elastic Net," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2005, *67* (2), 301–320.

Appendices

A. Baseline specifications

Table A.1. Baseline specifications (information set = completed bars at the origin; tuned on training/validation periods only)

Model	Specification
Random walk	window: 390 one-minute bars; distribution: Gaussian, zero mean, scale = trailing RMS one-minute return x sqrt(H)
Seasonal volatility	profile: past-20-session mean squared return per slot (past only); level_adjustment: trailing 30-minute RV over the seasonal counterpart, clipped to [0.25, 4]
EWMA-t	lambda: 0.97 per minute; nu: 5; history: last 3 sessions of one-minute returns
HAR-RV	regressors: log RV 30 min, log RV 390 min, log RV 5 days, time of day, time of day squared, log H; fit: OLS on training origins, one regression per horizon; nu: method of moments on standardised training residuals, clipped to [2.5, 50]
GARCH(1,1)-t	model: GARCH(1,1) with Student-t innovations on 5-minute training returns, one fit per instrument (arch 8); forecast: analytic variance term structure over ceil(H/5) steps; own filter over the last 400 five-minute returns
Ridge	alpha: 1.0; features: 19 origin features (src/chronos_x/baselines/features.py); targets: scaled cumulative return (location) and log scaled realised variance (scale); nu from residual kurtosis
Elastic net	alpha: 0.001; l1_ratio: 0.5
LightGBM	n_estimators: 400; learning_rate: 0.03; num_leaves: 31; subsample: 0.8; colsample_bytree: 0.8
MLP	inputs: last 60 one-minute bars (15 features each) + 9 global features + future calendar; hidden: 256 x 2 (GELU, dropout 0.1); head: Student-t per future minute, masked NLL
TCN	channels: 96; blocks: 6 dilated causal convolutions (kernel 3, dilation 1..32) over 390 one-minute bars; head: Student-t per future minute, masked NLL
Not run	*chronos_timesfm*: pretrained foundation models not evaluated: weights not available offline on the compute node; tokenisers assume a single real-valued series without missingness masks; licenses/terms not reviewed within this project; *egarch*: multi-step forecasts require simulation; too slow for the number of evaluation origins; GARCH(1,1)-t reported; *patchtst_itransformer*: the one-minute-only MRPT variant is the equal-parameter patch-transformer control; iTransformer not implemented

Notes. Specifications as configured in `configs/baselines.yaml`; every baseline is fitted once on the pooled 2018–2022 training origins (held-out and look-ahead instruments: refitted on their own histories) and frozen through 2025. The HAR-RV, ridge, elastic-net and LightGBM variance forecasts use the consistent retransformation of Section 6 (Duan smearing, ν by the method of moments on the corrected residuals, no log H column); the fitted factors are in Table A.2 and the before/after reconciliation in Table 5.

Table A.2. Fitted retransformation and tail parameters of the regression baselines: Duan smearing factor, method-of-moments degrees of freedom, and the scale and ν chosen by the 2023 recalibration

Model	$H = 5$				$H = 15$				$H = 30$				$H = 60$			
	$\hat{c}_{\text{sm}}$	$\hat{\nu}_{\text{mom}}$	ν_{recal}	c_{recal}	$\hat{c}_{\text{sm}}$	$\hat{\nu}_{\text{mom}}$	ν_{recal}	c_{recal}	$\hat{c}_{\text{sm}}$	$\hat{\nu}_{\text{mom}}$	ν_{recal}	c_{recal}	$\hat{c}_{\text{sm}}$	$\hat{\nu}_{\text{mom}}$	ν_{recal}	c_{recal}
HAR-RV	1.568	6.06	8	0.94	1.226	5.15	8	0.94	1.161	5.50	6	0.96	1.135	6.45	8	0.92
Ridge	1.635	5.73	8	0.90	1.219	5.29	8	0.92	1.152	5.58	8	0.92	1.127	6.51	8	0.92
Elastic net	1.636	5.73	8	0.90	1.219	5.28	8	0.92	1.153	5.58	8	0.92	1.127	6.49	8	0.92
LightGBM	1.863	4.00	8	0.86	1.198	5.63	8	0.94	1.130	5.89	8	0.94	1.105	7.14	8	0.94

Notes. $\hat{c}_{\text{sm}}$ = mean of exp(residual) of the log-variance regression on the training origins (2018–2022), multiplying $\exp(A\beta)$ (HAR, ridge, elastic net) or $\exp(\hat{f})\,\sigma^2$ (LightGBM); $\hat{\nu}_{\text{mom}} = 4 + 6/(\kappa - 3)$ from the kurtosis κ of the returns standardised by the corrected variance, clipped to $[2.5, 50]$; ν_{recal} and c_{recal} = the degrees of freedom (grid $\{3, 4, 5, 6, 8, 12, \infty\}$) and scalar scale (grid $[0.6, 1.6]$, step 0.02) that minimise the mean CRPS on a seeded stratified subset of 100,000 2023 validation origins, used only for the recalibrated rows. The v1.0 construction under-stated the variance most at five minutes (factor 1.6–1.9). Source: `results/metrics/recal/nu_hat.csv`, `c_val.json`.

B. Additional tables

The tables of this appendix complete the main text under the conventions of Section 6.2: the unscaled-CRPS companions of the main contrasts (relative changes and day-cluster statistics on unscaled CRPS, which weights high-volatility origins and days more heavily), the weighted interval score and QLIKE in levels, the held-out and 2025 calibration and recalibration tables, the second held-out set in 2025, the H3-control, seed and learning-curve tables, the pairwise contrast tables of every split, the minimum detectable effects, the Diebold–Mariano inflation audit, the ETF family

overlap and the pilot. Each note states the sample, the loss convention, the inference procedure and the Holm family.

Table B.1. Main study: pre-registered pairwise contrasts after recalibrating every model on 2023 — percentage change in mean scaled CRPS, day-cluster t and Holm-adjusted significance

Contrast (first minus second, % of the second)	$H = 5$	$H = 15$	$H = 30$	$H = 60$
H1: MRPT-S vs 1m-only twin (both recalibrated)	-0.04 [-3.1]**	-0.06 [-4.3]***	-0.07 [-4.5]***	-0.11 [-4.4]***
Size: MRPT-B vs MRPT-S	+0.03 [2.1]	+0.01 [0.7]	+0.02 [1.1]	+0.03 [1.4]
MRPT-S vs HAR-RV	-0.40 [-14.6]***	-0.14 [-5.8]***	-0.07 [-2.4]	-0.02 [-0.5]
MRPT-S vs HAR-RV	-0.23 [-8.3]***	+0.17 [5.5]***	+0.34 [7.4]***	+0.48 [6.7]***
MRPT-S vs ridge	-0.13 [-3.6]***	-0.05 [-1.7]	-0.04 [-1.3]	-0.05 [-1.1]
MRPT-S vs elastic net	-0.14 [-3.9]***	-0.05 [-1.8]	-0.04 [-1.4]	-0.05 [-1.0]
MRPT-S vs LightGBM	-8.34 [-98.1]***	+0.01 [0.3]	+0.00 [0.1]	-0.02 [-0.5]
MRPT-S vs LightGBM	-8.19 [-108.8]***	+0.31 [9.3]***	+0.41 [9.5]***	+0.48 [8.2]***
MRPT-S vs Student-t random walk (2023-calibrated)	-1.70 [-31.1]***	-1.40 [-23.9]***	-1.13 [-17.6]***	-0.80 [-11.0]***
MRPT-S vs Student-t random walk (2023-calibrated)	-1.54 [-25.6]***	-1.10 [-15.8]***	-0.72 [-8.9]***	-0.30 [-3.2]**
MRPT-S (1m+5m+30m) (recalibrated) vs MRPT-S (1m+5m+30m)	-0.17 [-10.5]***	-0.30 [-14.4]***	-0.41 [-13.8]***	-0.50 [-13.2]***
HAR-RV (recalibrated) vs Random walk, Student-t (2023-calibrated)	-1.31 [-22.4]***	-1.26 [-22.6]***	-1.06 [-20.0]***	-0.78 [-16.2]***

Notes. development test 2024, 986468 evaluation origins over 251 sessions. Cells: $100 \times (\bar{L}_{\text{first}} - \bar{L}_{\text{second}})/\bar{L}_{\text{second}}$ on scaled CRPS, i.e. CRPS divided by $\sigma_t\sqrt{H}$ (negative = the first model is better); in brackets the day-cluster t-statistic (mean loss differential over its trading-day bootstrap standard error, 1,000 resamples; every origin of one session is one cluster). Stars: Holm-adjusted p-value over the family of 48 tests (12 pairs × 4 horizons, adjusted on the day-cluster p): *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$. The HAC Diebold–Mariano t on the pooled series and the daily-mean DM t are in the result file (`results/metrics/main/pairwise_test_recal.csv`); a contrast whose magnitude lies inside the five-seed band of Table 10 is not distinguishable from training randomness whatever its t.

Table B.2. Main study (held-out instruments): pre-registered pairwise contrasts — percentage change in mean scaled CRPS, day-cluster t and Holm-adjusted significance

Contrast (first minus second, % of the second)	$H = 5$	$H = 15$	$H = 30$	$H = 60$
H1: MRPT-S (1m+5m+30m) vs MRPT-S (1m only)	-0.04 [-3.8]***	-0.05 [-4.0]***	-0.07 [-4.1]***	-0.12 [-4.1]***
Size: MRPT-B vs MRPT-S	+0.04 [2.8]*	+0.05 [3.2]**	+0.04 [2.6]	+0.07 [2.5]
MRPT-S vs HAR-RV	-0.20 [-4.7]***	-0.04 [-1.0]	+0.06 [1.0]	+0.02 [0.3]
MRPT-S vs ridge	-0.15 [-3.2]**	+0.02 [0.5]	+0.12 [2.0]	+0.07 [1.0]
MRPT-S vs elastic net	-0.15 [-3.2]**	+0.02 [0.4]	+0.11 [1.9]	+0.08 [1.2]
MRPT-S vs LightGBM	-0.21 [-4.6]***	-0.02 [-0.4]	+0.00 [0.1]	-0.02 [-0.3]
MRPT-S vs GARCH(1,1)-t	-0.80 [-17.7]***	-0.65 [-14.9]***	-0.78 [-16.7]***	-1.30 [-19.9]***
MRPT-S vs EWMA-t	-0.49 [-13.5]***	-0.44 [-12.5]***	-0.53 [-11.8]***	-0.87 [-14.4]***
MRPT-S vs seasonal volatility	-1.14 [-26.1]***	-0.96 [-24.2]***	-0.87 [-17.3]***	-0.91 [-15.5]***
MRPT-S vs TCN	-0.15 [-8.1]***	-0.21 [-10.8]***	-0.27 [-10.4]***	-0.35 [-8.9]***
MRPT-S vs MLP	-0.25 [-3.6]**	-0.24 [-3.4]**	-0.28 [-3.5]**	-0.33 [-3.7]***
MRPT-S vs random walk	-3.66 [-39.3]***	-3.55 [-38.4]***	-3.53 [-33.3]***	-3.74 [-28.0]***
MRPT-S vs Student-t random walk (2023-calibrated)	-1.80 [-24.0]***	-1.48 [-16.8]***	-1.09 [-11.1]***	-0.74 [-6.4]***
HAR-RV vs GARCH(1,1)-t	-0.60 [-17.0]***	-0.61 [-15.3]***	-0.83 [-16.1]***	-1.31 [-17.0]***
HAR-RV vs random walk	-3.46 [-35.0]***	-3.50 [-34.5]***	-3.58 [-30.9]***	-3.76 [-25.8]***
H2 paired: trained on all 300 vs zero-shot MRPT-S	+0.01 [1.3]	+0.02 [1.6]	+0.04 [3.1]**	+0.09 [5.3]***
Trained on all 300 vs HAR-RV	-0.19 [-4.8]***	-0.03 [-0.6]	+0.10 [1.8]	+0.11 [1.7]
Trained on all 300 vs random walk	-3.64 [-39.8]***	-3.53 [-38.5]***	-3.49 [-33.2]***	-3.65 [-27.9]***

Notes. held-out instruments, 2024, 250322 evaluation origins over 251 sessions. Cells: $100 \times (\bar{L}_{\text{first}} - \bar{L}_{\text{second}})/\bar{L}_{\text{second}}$ on scaled CRPS, i.e. CRPS divided by $\sigma_t\sqrt{H}$ (negative = the first model is better); in brackets the day-cluster t-statistic (mean loss differential over its trading-day bootstrap standard error, 1,000 resamples; every origin of one session is one cluster). Stars: Holm-adjusted p-value over the family of 72 tests (18 pairs × 4 horizons, adjusted on the day-cluster p): *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$. The HAC Diebold–Mariano t on the pooled series and the daily-mean DM t are in the result file (`results/metrics/main/pairwise_test_heldout.csv`); a contrast whose magnitude lies inside the five-seed band of Table 10 is not distinguishable from training randomness whatever its t.

Table B.3. Final period: pre-registered pairwise contrasts — percentage change in mean scaled CRPS, day-cluster t and Holm-adjusted significance

Contrast (first minus second, % of the second)	$H = 5$	$H = 15$	$H = 30$	$H = 60$
H1: MRPT-S (1m+5m+30m) vs MRPT-S (1m only)	+0.00 [0.4]	+0.02 [2.0]	+0.02 [1.2]	-0.01 [-0.2]
Size: MRPT-B vs MRPT-S	+0.03 [2.6]	+0.02 [1.8]	+0.01 [0.6]	+0.01 [0.4]
MRPT-S vs HAR-RV	-0.07 [-2.1]	+0.23 [5.6]***	+0.37 [6.5]***	+0.56 [8.6]***
MRPT-S vs ridge	+0.05 [1.3]	+0.28 [8.7]***	+0.39 [8.5]***	+0.51 [9.0]***
MRPT-S vs elastic net	+0.05 [1.2]	+0.28 [8.7]***	+0.38 [8.4]***	+0.51 [8.9]***
MRPT-S vs LightGBM	-8.79 [-99.8]***	+0.41 [9.9]***	+0.45 [11.3]***	+0.50 [9.0]***
MRPT-S vs GARCH(1,1)-t	-0.49 [-10.7]***	-0.16 [-3.4]**	-0.29 [-6.4]***	-0.69 [-10.1]***
MRPT-S vs EWMA-t	-0.34 [-13.5]***	-0.18 [-6.6]***	-0.28 [-8.7]***	-0.50 [-9.8]***
MRPT-S vs seasonal volatility	-1.31 [-20.7]***	-0.97 [-20.2]***	-0.88 [-18.6]***	-0.80 [-13.6]***
MRPT-S vs TCN	-0.13 [-9.3]***	-0.20 [-13.1]***	-0.26 [-13.1]***	-0.34 [-11.7]***
MRPT-S vs MLP	-0.13 [-6.3]***	-0.09 [-4.3]***	-0.08 [-3.3]**	-0.07 [-2.4]
MRPT-S vs random walk	-3.47 [-46.1]***	-3.46 [-39.6]***	-3.40 [-34.2]***	-3.46 [-27.5]***
MRPT-S vs Student-t random walk (2023-calibrated)	-1.55 [-24.3]***	-1.07 [-12.2]***	-0.69 [-8.0]***	-0.23 [-2.5]
HAR-RV vs GARCH(1,1)-t	-0.42 [-9.6]***	-0.39 [-7.3]***	-0.65 [-10.6]***	-1.24 [-14.9]***
HAR-RV vs random walk	-3.41 [-39.4]***	-3.68 [-36.1]***	-3.76 [-29.0]***	-3.99 [-25.0]***

Notes. untouched final period 2025, 1008425 evaluation origins over 249 sessions. Cells: $100 \times (\bar{L}_{\text{first}} - \bar{L}_{\text{second}})/\bar{L}_{\text{second}}$ on scaled CRPS, i.e. CRPS divided by $\sigma_t\sqrt{H}$ (negative = the first model is better); in brackets the day-cluster t-statistic (mean loss differential over its trading-day bootstrap standard error, 1,000 resamples; every origin of one session is one cluster). Stars: Holm-adjusted p-value over the family of 60 tests (15 pairs × 4 horizons, adjusted on the day-cluster p): *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$. The HAC Diebold–Mariano t on the pooled series and the daily-mean DM t are in the result file (`results/metrics/final/pairwise_final.csv`); a contrast whose magnitude lies inside the five-seed band of Table 10 is not distinguishable from training randomness whatever its t.

Table B.4. Final period: pre-registered pairwise contrasts after recalibrating every model on 2023 — percentage change in mean scaled CRPS, day-cluster t and Holm-adjusted significance

Contrast (first minus second, % of the second)	$H = 5$	$H = 15$	$H = 30$	$H = 60$
H1: MRPT-S vs 1m-only twin (both recalibrated)	-0.00 [-0.1]	-0.01 [-0.9]	+0.01 [0.6]	-0.03 [-1.3]
Size: MRPT-B vs MRPT-S	+0.04 [2.1]	+0.04 [2.9]*	+0.00 [0.1]	+0.00 [0.2]
MRPT-S vs HAR-RV	-0.36 [-13.8]***	-0.15 [-6.1]***	-0.08 [-3.0]*	+0.03 [0.9]
MRPT-S vs HAR-RV	-0.21 [-7.3]***	+0.18 [5.2]***	+0.35 [7.4]***	+0.55 [9.1]***
MRPT-S vs ridge	-0.08 [-2.2]	-0.00 [-0.0]	-0.01 [-0.4]	+0.04 [0.9]
MRPT-S vs elastic net	-0.08 [-2.4]	-0.00 [-0.1]	-0.02 [-0.5]	+0.03 [0.8]
MRPT-S vs LightGBM	-8.21 [-86.2]***	+0.07 [2.7]	+0.05 [2.0]	+0.03 [0.9]
MRPT-S vs LightGBM	-8.07 [-94.0]***	+0.41 [12.8]***	+0.48 [11.0]***	+0.55 [8.9]***
MRPT-S vs Student-t random walk (2023-calibrated)	-1.70 [-29.4]***	-1.40 [-19.2]***	-1.12 [-16.8]***	-0.74 [-10.5]***
MRPT-S vs Student-t random walk (2023-calibrated)	-1.55 [-24.3]***	-1.07 [-12.2]***	-0.69 [-8.0]***	-0.23 [-2.5]
MRPT-S (1m+5m+30m) (recalibrated) vs MRPT-S (1m+5m+30m)	-0.16 [-10.0]***	-0.34 [-13.4]***	-0.43 [-12.3]***	-0.51 [-12.6]***
HAR-RV (recalibrated) vs Random walk, Student-t (2023-calibrated)	-1.34 [-22.3]***	-1.25 [-18.0]***	-1.04 [-17.4]***	-0.77 [-14.1]***

Notes. untouched final period 2025, 1008425 evaluation origins over 249 sessions. Cells: $100 \times (\bar{L}_{\text{first}} - \bar{L}_{\text{second}})/\bar{L}_{\text{second}}$ on scaled CRPS, i.e. CRPS divided by $\sigma_t\sqrt{H}$ (negative = the first model is better); in brackets the day-cluster t-statistic (mean loss differential over its trading-day bootstrap standard error, 1,000 resamples; every origin of one session is one cluster). Stars: Holm-adjusted p-value over the family of 48 tests (12 pairs × 4 horizons, adjusted on the day-cluster p): *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$. The HAC Diebold–Mariano t on the pooled series and the daily-mean DM t are in the result file (`results/metrics/final/pairwise_final_recal.csv`); a contrast whose magnitude lies inside the five-seed band of Table 10 is not distinguishable from training randomness whatever its t.

Table B.5. Final period (held-out instruments): pre-registered pairwise contrasts — percentage change in mean scaled CRPS, day-cluster t and Holm-adjusted significance

Contrast (first minus second, % of the second)	$H = 5$	$H = 15$	$H = 30$	$H = 60$
H1: MRPT-S (1m+5m+30m) vs MRPT-S (1m only)	-0.02 [-1.3]	-0.00 [-0.3]	-0.01 [-0.7]	-0.02 [-1.0]
Size: MRPT-B vs MRPT-S	+0.05 [2.8]	+0.04 [2.6]	+0.02 [1.2]	+0.02 [0.9]
MRPT-S vs HAR-RV	-0.22 [-4.8]***	+0.00 [0.1]	+0.07 [1.2]	+0.17 [2.8]
MRPT-S vs ridge	-0.14 [-3.0]*	+0.13 [3.1]**	+0.17 [3.3]**	+0.21 [3.6]**
MRPT-S vs elastic net	-0.15 [-3.2]**	+0.13 [3.1]**	+0.15 [3.0]*	+0.22 [3.7]***
MRPT-S vs LightGBM	-0.21 [-4.2]***	+0.05 [0.8]	+0.02 [0.3]	+0.04 [0.5]
MRPT-S vs GARCH(1,1)-t	-0.72 [-13.1]***	-0.45 [-8.1]***	-0.61 [-11.4]***	-1.00 [-12.9]***
MRPT-S vs EWMA-t	-0.48 [-13.4]***	-0.34 [-9.7]***	-0.48 [-11.7]***	-0.69 [-11.2]***
MRPT-S vs seasonal volatility	-1.30 [-17.3]***	-1.00 [-16.7]***	-0.90 [-15.6]***	-0.80 [-11.0]***
MRPT-S vs TCN	-0.13 [-7.6]***	-0.17 [-8.5]***	-0.19 [-8.0]***	-0.23 [-7.0]***
MRPT-S vs MLP	-0.17 [-6.6]***	-0.13 [-5.0]***	-0.12 [-4.3]***	-0.11 [-3.2]**
MRPT-S vs random walk	-3.52 [-39.9]***	-3.51 [-34.5]***	-3.44 [-30.4]***	-3.47 [-24.3]***
MRPT-S vs Student-t random walk (2023-calibrated)	-1.83 [-22.7]***	-1.41 [-14.7]***	-1.08 [-11.1]***	-0.62 [-5.7]***
HAR-RV vs GARCH(1,1)-t	-0.50 [-10.0]***	-0.45 [-7.4]***	-0.68 [-10.5]***	-1.16 [-14.5]***
HAR-RV vs random walk	-3.30 [-33.4]***	-3.51 [-29.7]***	-3.51 [-25.8]***	-3.63 [-22.8]***
H2 paired: trained on all 300 vs zero-shot MRPT-S	-0.01 [-0.3]	+0.02 [1.9]	+0.03 [1.9]	+0.06 [2.9]*
Trained on all 300 vs HAR-RV	-0.23 [-5.4]***	+0.03 [0.6]	+0.09 [1.7]	+0.22 [3.8]***
Trained on all 300 vs random walk	-3.53 [-39.9]***	-3.49 [-34.5]***	-3.42 [-30.4]***	-3.41 [-25.1]***

Notes. held-out instruments, 2025, 254334 evaluation origins over 249 sessions. Cells: $100 \times (\bar{L}_{\text{first}} - \bar{L}_{\text{second}})/\bar{L}_{\text{second}}$ on scaled CRPS, i.e. CRPS divided by $\sigma_t\sqrt{H}$ (negative = the first model is better); in brackets the day-cluster t-statistic (mean loss differential over its trading-day bootstrap standard error, 1,000 resamples; every origin of one session is one cluster). Stars: Holm-adjusted p-value over the family of 72 tests (18 pairs × 4 horizons, adjusted on the day-cluster p): *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$. The HAC Diebold–Mariano t on the pooled series and the daily-mean DM t are in the result file (`results/metrics/final/pairwise_final_heldout.csv`); a contrast whose magnitude lies inside the five-seed band of Table 10 is not distinguishable from training randomness whatever its t.

Table B.6. Second held-out set (66 availability-excluded instruments) (held-out instruments): pre-registered pairwise contrasts — percentage change in mean scaled CRPS, day-cluster t and Holm-adjusted significance

Contrast (first minus second, % of the second)	$H = 5$	$H = 15$	$H = 30$	$H = 60$
H1: MRPT-S (1m+5m+30m) vs MRPT-S (1m only)	-0.07 [-2.5]	-0.09 [-3.6]***	-0.08 [-2.4]	-0.15 [-3.1]**
Size: MRPT-B vs MRPT-S	+0.05 [1.7]	+0.07 [2.9]*	+0.03 [1.1]	+0.13 [3.1]**
MRPT-S vs HAR-RV	-0.82 [-10.8]***	+0.08 [0.9]	+0.33 [2.7]*	+0.36 [2.6]
MRPT-S vs ridge	-0.25 [-2.8]*	+0.19 [2.2]	+0.38 [3.4]**	+0.38 [2.9]*
MRPT-S vs elastic net	-0.25 [-2.9]*	+0.20 [2.4]	+0.38 [3.4]**	+0.40 [3.0]**
MRPT-S vs LightGBM	-8.20 [-56.1]***	-0.00 [-0.0]	+0.19 [1.3]	+0.33 [2.6]
MRPT-S vs random walk	-5.24 [-43.0]***	-4.96 [-44.5]***	-4.71 [-34.6]***	-5.04 [-35.2]***

Notes. held-out instruments, 2024, 63182 evaluation origins over 251 sessions. Cells: $100 \times (\bar{L}_{\text{first}} - \bar{L}_{\text{second}})/\bar{L}_{\text{second}}$ on scaled CRPS, i.e. CRPS divided by $\sigma_t\sqrt{H}$ (negative = the first model is better); in brackets the day-cluster t-statistic (mean loss differential over its trading-day bootstrap standard error, 1,000 resamples; every origin of one session is one cluster). Stars: Holm-adjusted p-value over the family of 28 tests (7 pairs × 4 horizons, adjusted on the day-cluster p): *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$. The HAC Diebold–Mariano t on the pooled series and the daily-mean DM t are in the result file (`results/metrics/lookahead66/pairwise_test_heldout.csv`); a contrast whose magnitude lies inside the five-seed band of Table 10 is not distinguishable from training randomness whatever its t.

Table B.7. Minimum detectable effect of the pre-registered nulls: observed contrast / smallest contrast detectable at 80 % power (percent of the reference's scaled CRPS)

Null (first model minus second, % of the second)	$H = 5$	$H = 15$	$H = 30$	$H = 60$
H1 multi-resolution vs 1m-only (2024)	-0.039 / 0.022	-0.047 / 0.026	-0.065 / 0.038	-0.127 / 0.070
H1 multi-resolution vs 1m-only (2025)	+0.004 / 0.028	+0.022 / 0.030	+0.017 / 0.039	-0.005 / 0.061
Base vs Small (2024)	+0.028 / 0.029	+0.008 / 0.027	+0.011 / 0.035	+0.051 / 0.064
no-regime vs full (ablations 4k)	+0.016 / 0.018	-0.007 / 0.020	-0.009 / 0.029	-0.005 / 0.057
H2a MRPT-S vs HAR-RV on held-out instruments (2024)	-0.201 / 0.119	-0.045 / 0.128	+0.058 / 0.166	+0.018 / 0.191
MRPT-S vs HAR-RV (2024)	-0.068 / 0.085	+0.211 / 0.095	+0.370 / 0.149	+0.487 / 0.205

Notes. MDE = 2.8 x se_cl (alpha 5 %, power 80 %), where the standard error is the trading-day cluster-bootstrap standard error of the mean loss differential on the stated split; it quantifies the evaluation-origin uncertainty only (one trained network per model). The training-seed uncertainty, which is of the same order, is in Table 10: a contrast is reported as 'not detectable' when it fails either the MDE here or the seed-aware rule there. Convention: scaled CRPS. Source: `results/metrics/mde.csv` (`scripts/review/mde.py`).

Table B.8. Inflation of the v1.0 Diebold–Mariano statistic relative to the day-cluster statistic over every pairwise contrast file (CRPS, both conventions)

File	Contrasts	Median $\lvert t_{\mathrm{DM}}\rvert/\lvert t_{\mathrm{cl}}\rvert$	Share > 1 (%)	Maximum	$\lvert t_{\mathrm{DM}}\rvert > 1.96$	$\lvert t_{\mathrm{cl}}\rvert > 1.96$
`ablations/pairwise_test.csv`	136	1.19	96	1.63	119	117
`ablations_h3/pairwise_test.csv`	64	1.24	97	1.77	59	57
`final/pairwise_final.csv`	120	1.19	85	1.73	103	101
`final/pairwise_final_heldout.csv`	144	1.13	87	1.76	106	105
`final/pairwise_final_recal.csv`	96	1.10	78	1.64	65	64
`lookahead66/pairwise_final_heldout.csv`	56	1.03	70	1.38	35	34
`lookahead66/pairwise_test_heldout.csv`	56	1.13	86	1.26	46	45
`main/pairwise_test.csv`	120	1.21	96	1.68	110	110
`main/pairwise_test_heldout.csv`	144	1.16	97	1.72	116	111
`main/pairwise_test_recal.csv`	96	1.18	96	1.38	69	67
`seeds/pairwise_final.csv`	96	1.17	85	1.43	89	88
`seeds/pairwise_test.csv`	96	1.26	100	1.45	92	91
All files	1224	1.17	90	1.77		

Notes. t_{DM} = the v1.0 Diebold–Mariano statistic on the pooled per-origin loss differential with a Bartlett HAC variance whose lag is $\lceil H/\mathrm{stride}\rceil\times$ (number of instruments), i.e. 15–60 calendar minutes of a (date, slot, instrument)-ordered panel, which does not absorb day-level common shocks; t_{cl} = the mean differential over its trading-day cluster-bootstrap standard error (1,000 resamples), the statistic every test in this version uses. Both are kept in every result file. The median and share quoted in Section 6.3 are computed over the main and final files only (pooled, held-out and recalibrated families, both CRPS conventions): `main/pairwise_test.csv`, `main/pairwise_test_heldout.csv`, `main/pairwise_test_recal.csv`, `final/pairwise_final.csv`, `final/pairwise_final_heldout.csv`, `final/pairwise_final_recal.csv` (720 contrasts). Source: `results/metrics/*/pairwise_*.csv`.

Table B.9. Seed study (12,000 steps): mean scaled CRPS of the cumulative-return forecast by horizon (minutes)

Model	Parameters	$H = 5$	$H = 15$	$H = 30$	$H = 60$
MRPT-S seed 2	3,455,619	0.4529 [0.4474, 0.4588]	0.4413 [0.4349, 0.4478]	0.4312 [0.4249, 0.4381]	0.4189 [0.4108, 0.4286]
MRPT-S seed 3	3,455,619	0.4528 [0.4473, 0.4588]	0.4411 [0.4347, 0.4476]	0.4309 [0.4246, 0.4378]	0.4187 [0.4105, 0.4284]
MRPT-S seed 4	3,455,619	0.4526 [0.4471, 0.4585]	0.4409 [0.4346, 0.4475]	0.4309 [0.4245, 0.4378]	0.4186 [0.4105, 0.4283]
MRPT-S seed 5	3,455,619	0.4529 [0.4474, 0.4588]	0.4414 [0.4351, 0.4478]	0.4314 [0.4251, 0.4383]	0.4191 [0.4111, 0.4286]
MRPT-S 1m only, seed 2	3,200,259	0.4532 [0.4477, 0.4590]	0.4416 [0.4353, 0.4482]	0.4318 [0.4255, 0.4387]	0.4199 [0.4117, 0.4296]
MRPT-S 1m only, seed 3	3,200,259	0.4535 [0.4481, 0.4594]	0.4422 [0.4359, 0.4487]	0.4324 [0.4261, 0.4392]	0.4204 [0.4123, 0.4300]
MRPT-S 1m only, seed 4	3,200,259	0.4529 [0.4474, 0.4588]	0.4414 [0.4351, 0.4479]	0.4314 [0.4251, 0.4382]	0.4193 [0.4112, 0.4290]
MRPT-S 1m only, seed 5	3,200,259	0.4533 [0.4479, 0.4592]	0.4418 [0.4355, 0.4482]	0.4318 [0.4255, 0.4386]	0.4197 [0.4116, 0.4293]

Notes. development test 2024, 986468 evaluation origins over 251 sessions. Brackets: 95 % trading-day cluster-bootstrap intervals (all instruments and origins of one session form one cluster). Seeds 2–5 of the two H1 twins (12,000 steps, stratified validation); seed 1 is the main-study run of Table 6. The `seeds/pairwise_test.csv` contrasts of each seed against HAR-RV were recomputed against the corrected HAR-RV arrays (Stage E, M3U96b; the pass-1 file had used the pre-correction arrays); the seed-aware H1 contrast is in Table 10. Convention: scaled CRPS, i.e. CRPS divided by $\sigma_t\sqrt{H}$. Parameters: trainable parameters of the trained models (`train_summary.json`); baselines have none in this sense.

Table B.10. Seed study (12,000 steps): mean scaled CRPS of the cumulative-return forecast by horizon (minutes)

Model	Parameters	$H = 5$	$H = 15$	$H = 30$	$H = 60$
MRPT-S seed 2	3,455,619	0.4591 [0.4511, 0.4663]	0.4460 [0.4381, 0.4538]	0.4368 [0.4278, 0.4456]	0.4258 [0.4155, 0.4368]
MRPT-S seed 3	3,455,619	0.4591 [0.4512, 0.4663]	0.4458 [0.4379, 0.4537]	0.4366 [0.4277, 0.4455]	0.4256 [0.4153, 0.4366]
MRPT-S seed 4	3,455,619	0.4589 [0.4508, 0.4661]	0.4455 [0.4376, 0.4535]	0.4363 [0.4274, 0.4453]	0.4253 [0.4151, 0.4364]
MRPT-S seed 5	3,455,619	0.4592 [0.4512, 0.4664]	0.4461 [0.4382, 0.4540]	0.4370 [0.4281, 0.4458]	0.4261 [0.4159, 0.4371]
MRPT-S 1m only, seed 2	3,200,259	0.4592 [0.4512, 0.4664]	0.4462 [0.4382, 0.4540]	0.4371 [0.4282, 0.4460]	0.4264 [0.4161, 0.4375]
MRPT-S 1m only, seed 3	3,200,259	0.4597 [0.4517, 0.4668]	0.4467 [0.4388, 0.4546]	0.4377 [0.4288, 0.4465]	0.4268 [0.4167, 0.4378]
MRPT-S 1m only, seed 4	3,200,259	0.4591 [0.4511, 0.4663]	0.4460 [0.4380, 0.4539]	0.4368 [0.4279, 0.4457]	0.4259 [0.4156, 0.4369]
MRPT-S 1m only, seed 5	3,200,259	0.4594 [0.4514, 0.4666]	0.4463 [0.4383, 0.4541]	0.4371 [0.4282, 0.4459]	0.4262 [0.4160, 0.4372]

Notes. untouched final period 2025, 1008425 evaluation origins over 249 sessions. Brackets: 95 % trading-day cluster-bootstrap intervals (all instruments and origins of one session form one cluster). Seeds 2–5 of the two H1 twins (the 12,000-step checkpoints of the seed study) scored exactly once on the untouched 2025 origins (revision package P3b, no selection); seed 1 is the main-study run of Table 18. The seed-aware H1 contrast on 2025 is Panel D of Table 10. Convention: scaled CRPS, i.e. CRPS divided by $\sigma_t\sqrt{H}$. Parameters: trainable parameters of the trained models (`train_summary.json`); baselines have none in this sense.

Table B.11. Learning curve (continued training): mean scaled CRPS of the cumulative-return forecast by horizon (minutes)

Model	Parameters	$H = 5$	$H = 15$	$H = 30$	$H = 60$
MRPT-S continued to 24k steps	3,455,619	0.4529 [0.4474, 0.4588]	0.4413 [0.4350, 0.4478]	0.4312 [0.4249, 0.4381]	0.4188 [0.4106, 0.4284]
MRPT-S continued to 36k steps	3,455,619	0.4528 [0.4473, 0.4587]	0.4411 [0.4348, 0.4476]	0.4310 [0.4247, 0.4380]	0.4185 [0.4104, 0.4281]
MRPT-S 1m only, 24k steps	3,200,259	0.4532 [0.4478, 0.4591]	0.4418 [0.4355, 0.4483]	0.4319 [0.4257, 0.4387]	0.4199 [0.4119, 0.4297]
MRPT-S 1m only, 36k steps	3,200,259	0.4526 [0.4471, 0.4585]	0.4411 [0.4348, 0.4477]	0.4313 [0.4250, 0.4381]	0.4195 [0.4113, 0.4291]

Notes. development test 2024, 986468 evaluation origins over 251 sessions. Brackets: 95 % trading-day cluster-bootstrap intervals (all instruments and origins of one session form one cluster). Convention: scaled CRPS, i.e. CRPS divided by $\sigma_t\sqrt{H}$. Parameters: trainable parameters of the trained models (`train_summary.json`); baselines have none in this sense.

Table B.12. H3 control ablations (Small, 4,000 steps): empirical 50/80/95% coverage and PIT KS statistic

Model	$H = 5$				$H = 15$				$H = 30$				$H = 60$			
	50	80	95	KS	50	80	95	KS	50	80	95	KS	50	80	95	KS
MRPT-S regime stats, reduced global token	0.55	0.83	0.96	0.027	0.56	0.84	0.96	0.033	0.57	0.85	0.96	0.037	0.58	0.86	0.96	0.044
MRPT-S no regime, reduced global token	0.54	0.83	0.95	0.026	0.56	0.84	0.96	0.032	0.57	0.85	0.96	0.037	0.58	0.86	0.96	0.045
MRPT-S time only, reduced global token	0.56	0.84	0.96	0.036	0.58	0.85	0.96	0.042	0.59	0.86	0.96	0.047	0.60	0.87	0.97	0.050
MRPT-S shuffled price/volume columns	0.55	0.84	0.96	0.032	0.57	0.85	0.96	0.037	0.58	0.85	0.96	0.042	0.59	0.86	0.96	0.047

Notes. development test 2024, 986468 evaluation origins over 251 sessions. Empirical coverage of the central 50/80/95 % prediction intervals of the cumulative return (nominal 0.50/0.80/0.95) and the Kolmogorov–Smirnov distance of the PIT values from uniformity; no scaling is involved.

Table B.13. Held-out ETFs that share an index family with a pooled ETF: what 'never seen in training' means for the transfer test

Held-out ETF	Category	Index family	Pooled ETFs in the same family	Pooled tickers
SPHD	volatility_alt	sp500;dividend;low_volatility	8	IVE;SDS;SH;SPY;SPYD;UPRO;USMV;VIG
SSO	leveraged_inverse	sp500	6	IVE;SDS;SH;SPY;SPYD;UPRO
SCHF	international	developed_ex_us	3	IXUS;SPDW;VEU
VEA	international	developed_ex_us	3	IXUS;SPDW;VEU
SPSB	leveraged_inverse	us_corporate_bond	2	IGSB;LQD
TLT	fixed_income	us_treasury	2	IEF;SHV
IAU	commodity	gold	1	GLD
SCHA	broad_equity	us_smallcap	1	TNA
SCHH	other	us_real_estate	1	VNQ
TZA	leveraged_inverse	us_smallcap	1	TNA
UDOW	leveraged_inverse	dow30	1	SDOW
XLRE	sector	us_real_estate	1	VNQ
BOTZ	international	–	0	–
ITB	broad_equity	homebuilders	0	–
PSQ	leveraged_inverse	nasdaq100	0	–
SOXX	sector	semiconductor	0	–
VMBS	other	mbs	0	–
XRT	sector	retail	0	–

Notes. 12 of the 18 held-out ETFs track an index (or a leveraged/inverse version of it) that a pooled ETF also tracks; their minute returns are therefore strongly related to series the model was trained on, and the held-out contrasts of Table 15 are, for these names, a test of transfer across related rather than independent instruments. Families were assigned from the vendor names (`scripts/review/lookahead_universe.py`). Source: `results/metrics/etf_family_overlap.csv`.

Table B.14. Main study: CRPS relative to Random walk (trailing RMS)

Model	$H = 5$	$H = 15$	$H = 30$	$H = 60$
Panel A. Models as specified (baselines) or trained (neural models)				
Seasonal volatility	-2.89 [-28.7]	-3.09 [-30.4]	-3.22 [-28.7]	-3.44 [-25.0]
EWMA-t	-3.65 [-37.2]	-3.69 [-37.1]	-3.60 [-32.3]	-3.47 [-25.6]
HAR-RV	-4.03 [-39.1]	-4.33 [-39.8]	-4.54 [-36.2]	-4.87 [-30.9]
GARCH(1,1)-t	-3.60 [-35.7]	-3.77 [-36.5]	-3.70 [-33.1]	-3.40 [-26.0]
Ridge	-4.10 [-36.9]	-4.31 [-39.8]	-4.51 [-37.1]	-4.80 [-32.1]
Elastic net	-4.09 [-36.9]	-4.31 [-39.9]	-4.51 [-37.2]	-4.80 [-32.0]
LightGBM	+5.44 [55.7]	-4.40 [-39.0]	-4.50 [-38.0]	-4.74 [-33.1]
MLP (1m)	-3.94 [-42.7]	-3.93 [-41.0]	-3.94 [-38.2]	-4.07 [-33.1]
TCN (1m)	-3.93 [-42.1]	-3.75 [-38.8]	-3.68 [-34.9]	-3.73 [-30.4]
MRPT-S (1m+5m+30m)	-4.05 [-41.5]	-4.03 [-39.8]	-4.05 [-35.6]	-4.27 [-32.2]
MRPT-S (1m only)	-4.02 [-41.8]	-4.00 [-40.6]	-4.00 [-36.4]	-4.17 [-32.5]
MRPT-B (1m+5m+30m)	-4.04 [-41.5]	-4.04 [-40.3]	-4.06 [-36.6]	-4.23 [-32.3]
Panel B. Recalibrated on the 2023 validation origins (scalar scale c and ν per model and horizon; Section 6)				
Random walk, Student-t (2023-calibrated)	-2.46 [-33.6]	-2.92 [-31.8]	-3.36 [-31.5]	-3.98 [-28.6]
Elastic net (recalibrated)	-4.10 [-36.8]	-4.37 [-38.9]	-4.55 [-36.6]	-4.84 [-31.3]
HAR-RV (recalibrated)	-3.84 [-38.7]	-4.27 [-40.3]	-4.50 [-37.5]	-4.85 [-31.3]
LightGBM (recalibrated)	+4.61 [48.9]	-4.38 [-40.8]	-4.53 [-37.5]	-4.80 [-32.2]
MRPT-B (1m+5m+30m) (recalibrated)	-4.22 [-41.3]	-4.37 [-39.9]	-4.49 [-35.9]	-4.77 [-31.5]
MRPT-S (1m only) (recalibrated)	-4.21 [-41.5]	-4.32 [-40.2]	-4.44 [-36.4]	-4.71 [-31.9]
MRPT-S (1m+5m+30m) (recalibrated)	-4.23 [-41.6]	-4.36 [-40.4]	-4.50 [-36.3]	-4.80 [-31.9]
Ridge (recalibrated)	-4.11 [-36.8]	-4.37 [-38.9]	-4.55 [-36.6]	-4.84 [-31.3]

Notes. development test 2024, 986468 evaluation origins over 251 sessions. Cells: percentage change in mean unscaled CRPS of the cumulative return (the per-origin loss on which every test is computed) versus Random walk (trailing RMS), i.e. $100 \times (\bar{L}_{\text{model}} - \bar{L}_{\text{ref}})/\bar{L}_{\text{ref}}$ (negative = better); in brackets the day-cluster t-statistic of the mean loss differential (its trading-day bootstrap standard error, 1,000 resamples; every origin of one session forms one cluster). Convention: changes and tests on unscaled CRPS; these contrasts against the random walk are not part of a Holm family (the pre-registered pairwise contrasts with Holm adjustment are in the pairwise tables). Panel B lists the same forecasts after the scalar recalibration of Section 6 and the Student-t random walk whose scale and degrees of freedom were fitted on 2023; it is reported for completeness and is not used for any ranking in the text. Intervals are in the result file.

Table B.15. Main study (held-out): CRPS relative to Random walk (trailing RMS)

Model	$H = 5$	$H = 15$	$H = 30$	$H = 60$
Panel A. Models as specified (baselines) or trained (neural models)				
Seasonal volatility	-2.97 [-21.7]	-3.03 [-23.8]	-3.14 [-21.7]	-3.38 [-18.3]
EWMA-t	-3.54 [-27.0]	-3.45 [-27.0]	-3.33 [-23.5]	-3.23 [-18.5]
HAR-RV	-3.88 [-28.8]	-3.93 [-28.9]	-4.03 [-26.5]	-4.27 [-23.3]
GARCH(1,1)-t	-3.28 [-24.4]	-3.20 [-24.8]	-3.06 [-21.8]	-2.83 [-16.8]
Ridge	-3.87 [-27.6]	-4.01 [-29.2]	-4.09 [-26.7]	-4.34 [-23.4]
Elastic net	-3.87 [-27.8]	-4.00 [-29.2]	-4.09 [-26.6]	-4.35 [-23.5]
LightGBM	-3.83 [-27.5]	-3.90 [-29.4]	-3.96 [-26.0]	-4.24 [-23.0]
MLP (1m)	-3.92 [-29.9]	-3.85 [-29.8]	-3.82 [-27.0]	-3.98 [-23.6]
TCN (1m)	-4.02 [-31.2]	-3.90 [-30.7]	-3.85 [-27.3]	-3.98 [-23.9]
MRPT-S (1m+5m+30m)	-4.12 [-30.7]	-4.07 [-31.0]	-4.08 [-27.6]	-4.34 [-24.4]
MRPT-S (1m only)	-4.08 [-31.2]	-4.02 [-30.6]	-4.02 [-27.5]	-4.25 [-24.5]
MRPT-B (1m+5m+30m)	-4.10 [-30.6]	-4.04 [-30.6]	-4.05 [-27.6]	-4.29 [-24.2]
MRPT-S trained on all 300 instruments	-4.11 [-30.9]	-4.04 [-30.6]	-4.03 [-27.4]	-4.25 [-24.1]
Panel B. Recalibrated on the 2023 validation origins (scalar scale c and ν per model and horizon; Section 6)				
Random walk, Student-t (2023-calibrated)	-2.17 [-23.5]	-2.37 [-22.1]	-2.78 [-21.2]	-3.40 [-19.7]
Elastic net (recalibrated)	-3.93 [-26.6]	-4.01 [-28.5]	-4.12 [-25.7]	-4.44 [-22.0]
HAR-RV (recalibrated)	-3.91 [-28.1]	-3.95 [-28.8]	-4.05 [-25.8]	-4.36 [-21.7]
LightGBM (recalibrated)	-3.90 [-27.5]	-3.91 [-28.9]	-3.97 [-25.3]	-4.30 [-21.9]
Ridge (recalibrated)	-3.93 [-26.6]	-4.02 [-28.4]	-4.12 [-25.6]	-4.43 [-22.0]

Notes. held-out instruments, 2024, 250322 evaluation origins over 251 sessions. Cells: percentage change in mean unscaled CRPS of the cumulative return (the per-origin loss on which every test is computed) versus Random walk (trailing RMS), i.e. $100 \times (\bar{L}_{\text{model}} - \bar{L}_{\text{ref}})/\bar{L}_{\text{ref}}$ (negative = better); in brackets the day-cluster t-statistic of the mean loss differential (its trading-day bootstrap standard error, 1,000 resamples; every origin of one session forms one cluster). Convention: changes and tests on unscaled CRPS; these contrasts against the random walk are not part of a Holm family (the pre-registered pairwise contrasts with Holm adjustment are in the pairwise tables). Panel B lists the same forecasts after the scalar recalibration of Section 6 and the Student-t random walk whose scale and degrees of freedom were fitted on 2023; it is reported for completeness and is not used for any ranking in the text. Intervals are in the result file.

Table B.16. Main study (held-out): scaled CRPS relative to Random walk (trailing RMS)

Model	$H = 5$	$H = 15$	$H = 30$	$H = 60$
Panel A. Models as specified (baselines) or trained (neural models)				
Seasonal volatility	-2.55 [-25.8]	-2.62 [-27.4]	-2.68 [-24.5]	-2.86 [-20.4]
EWMA-t	-3.18 [-34.1]	-3.12 [-32.4]	-3.01 [-27.3]	-2.90 [-20.5]
HAR-RV	-3.46 [-35.0]	-3.50 [-34.5]	-3.58 [-30.9]	-3.76 [-25.8]
GARCH(1,1)-t	-2.89 [-31.0]	-2.91 [-31.4]	-2.77 [-27.0]	-2.48 [-20.1]
Ridge	-3.52 [-33.5]	-3.57 [-35.2]	-3.64 [-31.5]	-3.80 [-26.6]
Elastic net	-3.52 [-33.8]	-3.56 [-35.2]	-3.64 [-31.4]	-3.82 [-26.8]
LightGBM	-3.46 [-34.5]	-3.53 [-34.9]	-3.53 [-29.9]	-3.72 [-26.1]
MLP (1m)	-3.41 [-32.2]	-3.31 [-30.6]	-3.26 [-27.2]	-3.42 [-24.0]
TCN (1m)	-3.51 [-39.4]	-3.34 [-37.3]	-3.27 [-32.1]	-3.40 [-26.8]
MRPT-S (1m+5m+30m)	-3.66 [-39.3]	-3.55 [-38.4]	-3.53 [-33.3]	-3.74 [-28.0]
MRPT-S (1m only)	-3.62 [-39.3]	-3.50 [-37.2]	-3.46 [-32.4]	-3.63 [-27.4]
MRPT-B (1m+5m+30m)	-3.62 [-39.1]	-3.50 [-37.7]	-3.49 [-33.2]	-3.67 [-27.5]
MRPT-S trained on all 300 instruments	-3.64 [-39.8]	-3.53 [-38.5]	-3.49 [-33.2]	-3.65 [-27.9]
Panel B. Recalibrated on the 2023 validation origins (scalar scale c and ν per model and horizon; Section 6)				
Random walk, Student-t (2023-calibrated)	-1.89 [-24.5]	-2.09 [-21.8]	-2.46 [-21.1]	-3.03 [-18.8]
Elastic net (recalibrated)	-3.48 [-30.9]	-3.57 [-33.7]	-3.67 [-29.4]	-3.92 [-24.0]
HAR-RV (recalibrated)	-3.47 [-33.6]	-3.52 [-34.1]	-3.60 [-29.5]	-3.86 [-23.2]
LightGBM (recalibrated)	-3.52 [-33.9]	-3.53 [-33.8]	-3.54 [-28.5]	-3.80 [-24.0]
Ridge (recalibrated)	-3.49 [-30.9]	-3.58 [-33.7]	-3.67 [-29.4]	-3.92 [-24.0]

Notes. held-out instruments, 2024, 250322 evaluation origins over 251 sessions. Cells: percentage change in mean scaled CRPS, i.e. CRPS divided by $\sigma_t\sqrt{H}$ versus Random walk (trailing RMS), i.e. $100 \times (\bar{L}_{\text{model}} - \bar{L}_{\text{ref}})/\bar{L}_{\text{ref}}$ (negative = better); in brackets the day-cluster t-statistic of the mean loss differential (its trading-day bootstrap standard error, 1,000 resamples; every origin of one session forms one cluster). Convention: changes and tests on scaled CRPS; these contrasts against the random walk are not part of a Holm family (the pre-registered pairwise contrasts with Holm adjustment are in the pairwise tables). Panel B lists the same forecasts after the scalar recalibration of Section 6 and the Student-t random walk whose scale and degrees of freedom were fitted on 2023; it is reported for completeness and is not used for any ranking in the text. Intervals are in the result file.

Table B.17. Main study (held-out): model confidence sets (MCS p-values; † = member of the 90 % set)

Model	CRPS				WIS			
	$H = 5$	$H = 15$	$H = 30$	$H = 60$	$H = 5$	$H = 15$	$H = 30$	$H = 60$
Random walk (trailing RMS)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Seasonal volatility	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
EWMA-t	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
HAR-RV	0.00	0.01	0.39†	0.33†	0.00	0.01	0.06	0.10
GARCH(1,1)-t	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Ridge	0.00	0.68†	1.00†	0.84†	0.00	0.04	0.63†	0.45†
Elastic net	0.00	0.62†	0.75†	1.00†	0.00	0.03	0.55†	0.50†
LightGBM	0.00	0.01	0.17†	0.33†	0.00	0.01	0.06	0.37†
MLP (1m)	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00
TCN (1m)	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00
MRPT-S (1m+5m+30m)	1.00†	1.00†	0.75†	0.93†	1.00†	1.00†	1.00†	1.00†
MRPT-S (1m only)	0.00	0.62†	0.20†	0.24†	0.00	0.04	0.55†	0.37†
MRPT-B (1m+5m+30m)	0.44†	0.68†	0.39†	0.33†	0.41†	0.17†	0.63†	0.50†
MRPT-S trained on all 300 instruments	0.44†	0.68†	0.39†	0.33†	0.41†	0.17†	0.63†	0.45†
Random walk, Student-t (2023-calibrated)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Notes. held-out instruments, 2024, 250322 evaluation origins over 251 sessions. Model confidence set of Hansen, Lunde and Nason (2011) over the 15 models with a per-origin file, computed on the 251 daily mean losses (unscaled CRPS and WIS of the cumulative return) with the $T_{\max}$ statistic and a moving-block bootstrap of the days (block length 5, 1,000 resamples; block 20 in the result file); a model is in the 90 % set (†) when its MCS p-value is at least 0.10; the Student-t random walk calibrated on 2023 (last row) is a reference, not a competitor. Source: `results/metrics/main/mcs_test_heldout.csv`.

Table B.18. Main study (held-out): scalar recalibration fitted on 2023 — scaled CRPS and 50 % coverage before and after

Model	$H = 5$		$H = 15$		$H = 30$		$H = 60$	
	orig.	recal.	orig.	recal.	orig.	recal.	orig.	recal.
Scaled CRPS								
HAR-RV	0.4580	0.4580	0.4490	0.4489	0.4378	0.4377	0.4249	0.4245
Ridge	0.4578	0.4579	0.4487	0.4487	0.4376	0.4374	0.4247	0.4242
Elastic net	0.4578	0.4579	0.4487	0.4487	0.4376	0.4374	0.4247	0.4242
LightGBM	0.4580	0.4577	0.4489	0.4489	0.4381	0.4380	0.4251	0.4248
Coverage of the central 50 % interval (nominal 0.50)								
HAR-RV	0.506	0.497	0.495	0.500	0.506	0.497	0.525	0.504
Ridge	0.500	0.479	0.497	0.491	0.507	0.496	0.525	0.503
Elastic net	0.500	0.479	0.497	0.491	0.507	0.497	0.525	0.503
LightGBM	0.493	0.495	0.495	0.492	0.503	0.497	0.521	0.502

Notes. held-out instruments, 2024, 250322 evaluation origins over 251 sessions. 'orig.' = the model as trained/fitted; 'recal.' = the same forecasts with a scalar scale c (and, for the parametric baselines, degrees of freedom ν) chosen per model and horizon on a seeded stratified subset of 100,000 validation origins of 2023 to minimise mean CRPS (`results/metrics/recal/c_val.json`); the 2024 and 2025 scores are then computed once with the fitted constants (`scripts/recalibrate.py`, `scripts/review/rescore_scaled.py`). Convention: scaled CRPS. Source: the `_recal` rows of the leaderboard file.

Table B.19. Main study: mean WIS ($\times 10^4$) of the cumulative-return forecast by horizon (minutes)

Model	$H = 5$	$H = 15$	$H = 30$	$H = 60$
Panel A. Models as specified (baselines) or trained (neural models)				
Random walk (trailing RMS)	7.462 [7.322, 7.608]	12.559 [12.304, 12.834]	17.325 [16.944, 17.700]	23.787 [23.225, 24.432]
Seasonal volatility	7.238 [7.101, 7.381]	12.157 [11.911, 12.421]	16.748 [16.377, 17.129]	22.946 [22.374, 23.564]
EWMA-*t*	7.182 [7.047, 7.324]	12.089 [11.844, 12.350]	16.701 [16.326, 17.089]	22.978 [22.410, 23.612]
HAR-RV	7.153 [7.017, 7.295]	12.000 [11.753, 12.262]	16.520 [16.135, 16.910]	22.607 [22.031, 23.247]
GARCH(1,1)-*t*	7.189 [7.053, 7.332]	12.083 [11.837, 12.346]	16.689 [16.316, 17.072]	23.003 [22.445, 23.623]
Ridge	7.146 [7.009, 7.290]	11.999 [11.752, 12.260]	16.524 [16.142, 16.909]	22.619 [22.048, 23.257]
Elastic net	7.147 [7.010, 7.291]	12.000 [11.753, 12.261]	16.522 [16.140, 16.909]	22.620 [22.050, 23.259]
LightGBM	7.940 [7.797, 8.090]	11.989 [11.742, 12.253]	16.519 [16.141, 16.910]	22.624 [22.055, 23.267]
MLP (1m)	7.151 [7.016, 7.293]	12.038 [11.794, 12.299]	16.605 [16.235, 16.981]	22.768 [22.209, 23.380]
TCN (1m)	7.151 [7.017, 7.293]	12.062 [11.819, 12.323]	16.653 [16.286, 17.033]	22.855 [22.295, 23.468]
MRPT-S (1m+5m+30m)	7.141 [7.006, 7.282]	12.022 [11.779, 12.281]	16.583 [16.213, 16.963]	22.718 [22.149, 23.339]
MRPT-S (1m only)	7.144 [7.010, 7.286]	12.028 [11.786, 12.288]	16.593 [16.221, 16.974]	22.745 [22.177, 23.365]
MRPT-B (1m+5m+30m)	7.142 [7.007, 7.284]	12.022 [11.780, 12.281]	16.582 [16.212, 16.965]	22.730 [22.160, 23.348]
Panel B. Recalibrated on the 2023 validation origins (scalar scale c and ν per model and horizon; Section 6)				
Random walk, Student-*t* (2023-calibrated)	7.286 [7.147, 7.432]	12.203 [11.945, 12.471]	16.751 [16.356, 17.151]	22.845 [22.264, 23.486]

Notes. development test 2024, 986468 evaluation origins over 251 sessions. Brackets: 95 % trading-day cluster-bootstrap intervals (all instruments and origins of one session form one cluster). Convention: weighted interval score of the cumulative return (unscaled). Values are multiplied by 10^4 (raw WIS of order 10^{-3}).

Table B.20. Main study: mean QLIKE of the realised-variance forecast by horizon (minutes)

Model	$H = 5$	$H = 15$	$H = 30$	$H = 60$
Panel A. Models as specified (baselines) or trained (neural models)				
Random walk (trailing RMS)	-12.301 [-12.336, -12.265]	-11.244 [-11.277, -11.205]	-10.582 [-10.616, -10.543]	-9.931 [-9.967, -9.890]
Seasonal volatility	-12.386 [-12.419, -12.349]	-11.327 [-11.364, -11.283]	-10.655 [-10.695, -10.606]	-9.992 [-10.036, -9.932]
EWMA-*t*	-12.429 [-12.460, -12.393]	-11.354 [-11.391, -11.311]	-10.671 [-10.709, -10.626]	-9.993 [-10.034, -9.941]
HAR-RV	-12.424 [-12.458, -12.387]	-11.359 [-11.403, -11.306]	-10.694 [-10.736, -10.645]	-10.045 [-10.086, -9.995]
GARCH(1,1)-*t*	-12.339 [-12.377, -12.297]	-11.273 [-11.312, -11.229]	-10.592 [-10.631, -10.547]	-9.906 [-9.946, -9.858]
Ridge	-12.447 [-12.479, -12.409]	-11.385 [-11.422, -11.342]	-10.715 [-10.753, -10.669]	-10.057 [-10.098, -10.005]
Elastic net	-12.446 [-12.479, -12.408]	-11.385 [-11.422, -11.341]	-10.714 [-10.753, -10.668]	-10.057 [-10.097, -10.004]
LightGBM	-4.472 [-7.937, -0.651]	-11.200 [-11.379, -10.941]	-10.724 [-10.763, -10.681]	-10.070 [-10.109, -10.025]
MLP (1m)	-12.459 [-12.494, -12.421]	-11.389 [-11.425, -11.343]	-10.717 [-10.754, -10.670]	-10.056 [-10.097, -10.007]
TCN (1m)	-12.470 [-12.502, -12.433]	-11.394 [-11.431, -11.351]	-10.720 [-10.760, -10.674]	-10.062 [-10.102, -10.015]
MRPT-S (1m+5m+30m)	-12.476 [-12.508, -12.440]	-11.408 [-11.444, -11.367]	-10.737 [-10.774, -10.694]	-10.080 [-10.118, -10.035]
MRPT-S (1m only)	-12.473 [-12.505, -12.436]	-11.403 [-11.440, -11.359]	-10.730 [-10.769, -10.685]	-10.072 [-10.111, -10.021]
MRPT-B (1m+5m+30m)	-12.473 [-12.504, -12.435]	-11.404 [-11.441, -11.361]	-10.733 [-10.772, -10.688]	-10.075 [-10.115, -10.026]
Panel B. Recalibrated on the 2023 validation origins (scalar scale c and ν per model and horizon; Section 6)				
Random walk, Student-*t* (2023-calibrated)	-12.318 [-12.354, -12.281]	-11.265 [-11.304, -11.220]	-10.612 [-10.652, -10.564]	-9.978 [-10.021, -9.927]

Notes. development test 2024, 986468 evaluation origins over 251 sessions. Brackets: 95 % trading-day cluster-bootstrap intervals (all instruments and origins of one session form one cluster). Convention: QLIKE of the realised variance, $\log v + \mathrm{RV}/v$ without the data-only constant (levels can be negative). LightGBM variance-floor diagnostics (`results/metrics/recal/lightgbm_floor.csv`): $H = 5$: 0.053 % of origins with a variance forecast below $0.01\ \sigma_t^2 H$ (0.26 % below $0.1\ \sigma_t^2 H$), mean QLIKE -4.47 raw, -11.83 with the variance floored at $0.01\ \sigma_t^2 H$, -11.88 winsorised at the 1st/99th percentiles (maximum per-origin QLIKE 1025991); $H = 15$: 0.005 % of origins with a variance forecast below $0.01\ \sigma_t^2 H$ (0.05 % below $0.1\ \sigma_t^2 H$), mean QLIKE -11.20 raw, -11.39 with the variance floored at $0.01\ \sigma_t^2 H$, -11.41 winsorised at the 1st/99th percentiles (maximum per-origin QLIKE 110429); $H = 30$: 0.001 % of origins with a variance forecast below $0.01\ \sigma_t^2 H$ (0.03 % below $0.1\ \sigma_t^2 H$), mean QLIKE -10.72 raw, -10.73 with the variance floored at $0.01\ \sigma_t^2 H$, -10.74 winsorised at the 1st/99th percentiles (maximum per-origin QLIKE 1801); $H = 60$: 0.000 % of origins with a variance forecast below $0.01\ \sigma_t^2 H$ (0.01 % below $0.1\ \sigma_t^2 H$), mean QLIKE -10.07 raw, -10.07 with the variance floored at $0.01\ \sigma_t^2 H$, -10.09 winsorised at the 1st/99th percentiles (maximum per-origin QLIKE 393). The $H = 5$ mean is dominated by a few near-zero variance forecasts (no forecast is below $10^{-10}\ \sigma_t^2 H$).

Table B.21. Main study: WIS relative to Random walk (trailing RMS)

Model	$H = 5$	$H = 15$	$H = 30$	$H = 60$
Panel A. Models as specified (baselines) or trained (neural models)				
Seasonal volatility	-3.00 [-27.8]	-3.20 [-29.6]	-3.33 [-27.8]	-3.54 [-23.8]
EWMA-t	-3.75 [-36.0]	-3.75 [-35.8]	-3.60 [-30.8]	-3.40 [-24.0]
HAR-RV	-4.14 [-38.1]	-4.45 [-39.4]	-4.64 [-35.7]	-4.96 [-30.1]
GARCH(1,1)-t	-3.66 [-34.3]	-3.79 [-34.9]	-3.67 [-31.4]	-3.29 [-24.1]
Ridge	-4.22 [-35.9]	-4.46 [-39.3]	-4.63 [-36.5]	-4.91 [-31.3]
Elastic net	-4.22 [-35.9]	-4.45 [-39.4]	-4.63 [-36.6]	-4.91 [-31.2]
LightGBM	+6.41 [57.8]	-4.54 [-39.0]	-4.65 [-37.8]	-4.89 [-32.8]
MLP (1m)	-4.17 [-42.7]	-4.15 [-41.1]	-4.15 [-38.1]	-4.28 [-32.8]
TCN (1m)	-4.16 [-41.8]	-3.96 [-38.8]	-3.88 [-34.7]	-3.92 [-30.1]
MRPT-S (1m+5m+30m)	-4.30 [-41.2]	-4.28 [-39.8]	-4.28 [-35.5]	-4.50 [-31.9]
MRPT-S (1m only)	-4.26 [-41.5]	-4.23 [-40.6]	-4.22 [-36.1]	-4.38 [-32.1]
MRPT-B (1m+5m+30m)	-4.28 [-41.4]	-4.28 [-40.3]	-4.29 [-36.4]	-4.44 [-32.0]
Panel B. Recalibrated on the 2023 validation origins (scalar scale c and ν per model and horizon; Section 6)				
Random walk, Student-t (2023-calibrated)	-2.35 [-31.7]	-2.84 [-29.5]	-3.31 [-29.3]	-3.96 [-26.7]
Elastic net (recalibrated)	-4.27 [-35.6]	-4.55 [-38.1]	-4.71 [-35.6]	-5.00 [-30.0]
HAR-RV (recalibrated)	-4.02 [-37.9]	-4.45 [-39.9]	-4.64 [-36.8]	-5.00 [-30.0]
LightGBM (recalibrated)	+5.07 [48.4]	-4.58 [-40.6]	-4.71 [-37.1]	-4.98 [-31.6]
MRPT-B (1m+5m+30m) (recalibrated)	-4.44 [-40.7]	-4.57 [-39.3]	-4.67 [-34.9]	-4.96 [-30.2]
MRPT-S (1m only) (recalibrated)	-4.42 [-40.8]	-4.52 [-39.6]	-4.62 [-35.3]	-4.88 [-30.5]
MRPT-S (1m+5m+30m) (recalibrated)	-4.46 [-40.9]	-4.57 [-40.0]	-4.69 [-35.5]	-4.99 [-30.8]
Ridge (recalibrated)	-4.28 [-35.7]	-4.55 [-38.2]	-4.71 [-35.6]	-4.99 [-30.1]

Notes. development test 2024, 986468 evaluation origins over 251 sessions. Cells: percentage change in mean weighted interval score of the cumulative return (unscaled) versus Random walk (trailing RMS), i.e. $100 \times (\bar{L}_{\text{model}} - \bar{L}_{\text{ref}})/\bar{L}_{\text{ref}}$ (negative = better); in brackets the day-cluster t-statistic of the mean loss differential (its trading-day bootstrap standard error, 1,000 resamples; every origin of one session forms one cluster). Convention: changes and tests on unscaled WIS; these contrasts against the random walk are not part of a Holm family (the pre-registered pairwise contrasts with Holm adjustment are in the pairwise tables). Panel B lists the same forecasts after the scalar recalibration of Section 6 and the Student-t random walk whose scale and degrees of freedom were fitted on 2023; it is reported for completeness and is not used for any ranking in the text. Intervals are in the result file.

Table B.22. Main study: QLIKE relative to Random walk (trailing RMS)

Model	$H = 5$	$H = 15$	$H = 30$	$H = 60$
Panel A. Models as specified (baselines) or trained (neural models)				
Seasonal volatility	-0.0847 [-19.5]	-0.0831 [-16.8]	-0.0733 [-11.2]	-0.0602 [-6.3]
EWMA-t	-0.1280 [-33.4]	-0.1100 [-26.0]	-0.0896 [-17.7]	-0.0616 [-10.0]
HAR-RV	-0.1233 [-17.3]	-0.1151 [-8.6]	-0.1116 [-10.3]	-0.1139 [-16.4]
GARCH(1,1)-t	-0.0379 [-4.0]	-0.0294 [-4.1]	-0.0101 [-1.4]	+0.0255 [3.7]
Ridge	-0.1454 [-33.2]	-0.1414 [-26.6]	-0.1328 [-22.8]	-0.1258 [-19.1]
Elastic net	-0.1450 [-33.2]	-0.1410 [-26.4]	-0.1323 [-22.5]	-0.1254 [-18.8]
LightGBM	+7.8295 [4.0]	+0.0436 [0.3]	-0.1417 [-29.2]	-0.1390 [-28.9]
MLP (1m)	-0.1582 [-32.3]	-0.1451 [-29.2]	-0.1349 [-24.6]	-0.1245 [-18.5]
TCN (1m)	-0.1692 [-45.6]	-0.1502 [-30.6]	-0.1379 [-22.6]	-0.1306 [-24.1]
MRPT-S (1m+5m+30m)	-0.1752 [-42.8]	-0.1640 [-41.4]	-0.1547 [-37.4]	-0.1481 [-32.9]
MRPT-S (1m only)	-0.1718 [-42.9]	-0.1590 [-34.1]	-0.1486 [-28.6]	-0.1401 [-24.4]
MRPT-B (1m+5m+30m)	-0.1715 [-42.8]	-0.1604 [-35.3]	-0.1508 [-30.6]	-0.1434 [-25.9]
Panel B. Recalibrated on the 2023 validation origins (scalar scale c and ν per model and horizon; Section 6)				
Random walk, Student-t (2023-calibrated)	-0.0169 [-10.1]	-0.0213 [-5.0]	-0.0299 [-5.7]	-0.0465 [-7.1]
Elastic net (recalibrated)	-0.1221 [-20.8]	-0.1313 [-18.3]	-0.1234 [-15.2]	-0.1180 [-12.6]
HAR-RV (recalibrated)	-0.1337 [-16.5]	-0.1153 [-7.5]	-0.1123 [-9.4]	-0.1081 [-11.4]
LightGBM (recalibrated)	+8.5796 [4.1]	+0.0713 [0.5]	-0.1351 [-22.5]	-0.1332 [-21.4]
MRPT-B (1m+5m+30m) (recalibrated)	-0.1591 [-36.1]	-0.1435 [-24.3]	-0.1255 [-17.5]	-0.1200 [-14.4]
MRPT-S (1m only) (recalibrated)	-0.1590 [-34.3]	-0.1420 [-22.6]	-0.1234 [-16.3]	-0.1164 [-13.4]
MRPT-S (1m+5m+30m) (recalibrated)	-0.1626 [-35.2]	-0.1482 [-29.5]	-0.1303 [-22.6]	-0.1246 [-19.0]
Ridge (recalibrated)	-0.1228 [-21.0]	-0.1318 [-18.5]	-0.1239 [-15.4]	-0.1184 [-12.8]

Notes. development test 2024, 986468 evaluation origins over 251 sessions. Cells: level difference in mean QLIKE versus Random walk (trailing RMS) (negative = better); relative changes are not reported because the reference mean is not positive; in brackets the day-cluster t-statistic of the mean loss differential (its trading-day bootstrap standard error, 1,000 resamples; every origin of one session forms one cluster). Convention: changes and tests on unscaled QLIKE; these contrasts against the random walk are not part of a Holm family (the pre-registered pairwise contrasts with Holm adjustment are in the pairwise tables). Panel B lists the same forecasts after the scalar recalibration of Section 6 and the Student-t random walk whose scale and degrees of freedom were fitted on 2023; it is reported for completeness and is not used for any ranking in the text. Intervals are in the result file.

Table B.23. Main study (held-out): empirical 50/80/95% coverage and PIT KS statistic

Model	$H = 5$				$H = 15$				$H = 30$				$H = 60$			
	50	80	95	KS	50	80	95	KS	50	80	95	KS	50	80	95	KS
Panel A. Models as specified (baselines) or trained (neural models)																
Random walk (trailing RMS)	0.64	0.88	0.96	0.075	0.65	0.89	0.97	0.073	0.65	0.90	0.97	0.078	0.67	0.91	0.98	0.084
Seasonal volatility	0.55	0.82	0.94	0.026	0.54	0.82	0.94	0.026	0.55	0.82	0.94	0.028	0.55	0.82	0.95	0.029
EWMA-t	0.53	0.83	0.97	0.018	0.53	0.84	0.97	0.019	0.54	0.84	0.97	0.024	0.55	0.85	0.97	0.030
HAR-RV	0.51	0.81	0.96	0.007	0.50	0.80	0.96	0.007	0.51	0.81	0.96	0.009	0.52	0.83	0.96	0.016
GARCH(1,1)-t	0.53	0.83	0.96	0.018	0.54	0.84	0.97	0.022	0.55	0.85	0.97	0.029	0.58	0.87	0.98	0.042
Ridge	0.50	0.80	0.96	0.006	0.50	0.80	0.96	0.007	0.51	0.81	0.96	0.010	0.53	0.83	0.96	0.017
Elastic net	0.50	0.80	0.96	0.007	0.50	0.80	0.96	0.007	0.51	0.81	0.96	0.010	0.52	0.83	0.96	0.017
LightGBM	0.49	0.81	0.97	0.011	0.49	0.80	0.96	0.007	0.50	0.81	0.96	0.008	0.52	0.82	0.96	0.015
MLP (1m)	0.54	0.82	0.96	0.023	0.55	0.84	0.96	0.029	0.56	0.84	0.96	0.033	0.57	0.85	0.96	0.037
TCN (1m)	0.55	0.83	0.96	0.027	0.56	0.84	0.96	0.031	0.57	0.85	0.96	0.035	0.58	0.85	0.96	0.039
MRPT-S (1m+5m+30m)	0.54	0.82	0.95	0.024	0.55	0.83	0.96	0.026	0.56	0.84	0.96	0.030	0.56	0.84	0.96	0.033
MRPT-S (1m only)	0.54	0.82	0.95	0.023	0.55	0.83	0.96	0.026	0.56	0.84	0.96	0.029	0.56	0.85	0.96	0.034
MRPT-B (1m+5m+30m)	0.54	0.82	0.95	0.023	0.55	0.83	0.96	0.026	0.56	0.84	0.96	0.030	0.57	0.85	0.96	0.035
MRPT-S trained on all 300 instruments	0.54	0.82	0.95	0.022	0.55	0.83	0.96	0.026	0.56	0.84	0.96	0.030	0.57	0.85	0.96	0.035
Panel B. Recalibrated on the 2023 validation origins (scalar scale c and ν per model and horizon; Section 6)																
Random walk, Student-t (2023-calibrated)	0.50	0.80	0.95	0.007	0.50	0.79	0.95	0.006	0.50	0.79	0.95	0.008	0.50	0.80	0.95	0.008

Notes. held-out instruments, 2024, 250322 evaluation origins over 251 sessions. Empirical coverage of the central 50/80/95 % prediction intervals of the cumulative return (nominal 0.50/0.80/0.95) and the Kolmogorov–Smirnov distance of the PIT values from uniformity; no scaling is involved.

Table B.24. Final period: CRPS relative to Random walk (trailing RMS)

Model	$H = 5$	$H = 15$	$H = 30$	$H = 60$
Panel A. Models as specified (baselines) or trained (neural models)				
Seasonal volatility	-2.52 [-17.7]	-2.85 [-21.6]	-2.89 [-19.4]	-3.07 [-18.0]
EWMA-t	-3.51 [-29.5]	-3.59 [-22.6]	-3.39 [-23.8]	-3.14 [-18.5]
HAR-RV	-3.84 [-29.1]	-4.05 [-28.1]	-4.12 [-24.7]	-4.37 [-21.3]
GARCH(1,1)-t	-3.38 [-28.9]	-3.68 [-19.7]	-3.44 [-24.0]	-3.09 [-17.3]
Ridge	-3.97 [-28.3]	-4.14 [-28.1]	-4.19 [-26.4]	-4.39 [-23.7]
Elastic net	-3.96 [-28.3]	-4.14 [-28.1]	-4.19 [-26.5]	-4.39 [-23.7]
LightGBM	+5.32 [41.2]	-4.22 [-25.6]	-4.22 [-25.8]	-4.35 [-20.7]
MLP (1m)	-3.84 [-31.2]	-3.84 [-28.0]	-3.81 [-27.1]	-3.85 [-23.6]
TCN (1m)	-3.89 [-31.4]	-3.74 [-25.8]	-3.60 [-26.9]	-3.55 [-23.4]
MRPT-S (1m+5m+30m)	-3.96 [-30.2]	-3.89 [-27.9]	-3.85 [-26.7]	-3.91 [-22.0]
MRPT-S (1m only)	-3.97 [-30.7]	-3.91 [-28.3]	-3.88 [-26.8]	-3.92 [-22.3]
MRPT-B (1m+5m+30m)	-3.94 [-31.0]	-3.87 [-28.7]	-3.84 [-27.2]	-3.89 [-23.5]
Panel B. Recalibrated on the 2023 validation origins (scalar scale c and ν per model and horizon; Section 6)				
Random walk, Student-t (2023-calibrated)	-2.26 [-20.4]	-2.65 [-21.5]	-3.00 [-20.4]	-3.56 [-19.4]
Elastic net (recalibrated)	-3.97 [-28.2]	-4.18 [-28.5]	-4.22 [-25.9]	-4.42 [-23.0]
HAR-RV (recalibrated)	-3.69 [-29.3]	-4.02 [-28.3]	-4.12 [-25.6]	-4.37 [-21.6]
LightGBM (recalibrated)	+4.50 [38.2]	-4.25 [-25.7]	-4.25 [-25.4]	-4.39 [-20.1]
MRPT-S (1m+5m+30m) (recalibrated)	-4.10 [-30.4]	-4.18 [-30.2]	-4.23 [-26.6]	-4.36 [-21.6]
Ridge (recalibrated)	-3.98 [-28.2]	-4.18 [-28.6]	-4.22 [-25.9]	-4.42 [-23.0]

Notes. untouched final period 2025, 1008425 evaluation origins over 249 sessions. Cells: percentage change in mean unscaled CRPS of the cumulative return (the per-origin loss on which every test is computed) versus Random walk (trailing RMS), i.e. $100 \times (\bar{L}_{\text{model}} - \bar{L}_{\text{ref}})/\bar{L}_{\text{ref}}$ (negative = better); in brackets the day-cluster t-statistic of the mean loss differential (its trading-day bootstrap standard error, 1,000 resamples; every origin of one session forms one cluster). Convention: changes and tests on unscaled CRPS; these contrasts against the random walk are not part of a Holm family (the pre-registered pairwise contrasts with Holm adjustment are in the pairwise tables). Panel B lists the same forecasts after the scalar recalibration of Section 6 and the Student-t random walk whose scale and degrees of freedom were fitted on 2023; it is reported for completeness and is not used for any ranking in the text. Intervals are in the result file.

Table B.25. Final period: scaled CRPS relative to Random walk (trailing RMS)

Model	$H = 5$	$H = 15$	$H = 30$	$H = 60$
Panel A. Models as specified (baselines) or trained (neural models)				
Seasonal volatility	-2.20 [-21.3]	-2.52 [-25.7]	-2.54 [-23.3]	-2.68 [-20.3]
EWMA-t	-3.15 [-40.3]	-3.28 [-34.4]	-3.13 [-29.5]	-2.97 [-22.2]
HAR-RV	-3.41 [-39.4]	-3.68 [-36.1]	-3.76 [-29.0]	-3.99 [-25.0]
GARCH(1,1)-t	-3.00 [-36.6]	-3.31 [-32.9]	-3.13 [-29.6]	-2.79 [-21.4]
Ridge	-3.53 [-36.8]	-3.73 [-37.7]	-3.78 [-31.5]	-3.95 [-27.3]
Elastic net	-3.52 [-36.8]	-3.73 [-37.7]	-3.77 [-31.6]	-3.95 [-27.4]
LightGBM	+5.82 [61.6]	-3.86 [-35.9]	-3.83 [-32.7]	-3.93 [-26.7]
MLP (1m)	-3.35 [-47.6]	-3.38 [-41.2]	-3.33 [-35.8]	-3.39 [-29.0]
TCN (1m)	-3.35 [-47.4]	-3.26 [-38.5]	-3.15 [-33.8]	-3.13 [-27.3]
MRPT-S (1m+5m+30m)	-3.47 [-46.1]	-3.46 [-39.6]	-3.40 [-34.2]	-3.46 [-27.5]
MRPT-S (1m only)	-3.48 [-47.1]	-3.48 [-40.7]	-3.42 [-35.1]	-3.45 [-27.9]
MRPT-B (1m+5m+30m)	-3.44 [-46.7]	-3.44 [-40.1]	-3.40 [-34.7]	-3.45 [-28.1]
Panel B. Recalibrated on the 2023 validation origins (scalar scale c and ν per model and horizon; Section 6)				
Random walk, Student-t (2023-calibrated)	-1.95 [-25.7]	-2.42 [-23.8]	-2.73 [-22.0]	-3.24 [-21.1]
Elastic net (recalibrated)	-3.54 [-36.9]	-3.78 [-36.5]	-3.80 [-30.7]	-3.98 [-26.5]
HAR-RV (recalibrated)	-3.27 [-40.7]	-3.64 [-37.3]	-3.74 [-31.0]	-3.98 [-25.6]
LightGBM (recalibrated)	+5.00 [54.2]	-3.86 [-38.5]	-3.86 [-32.0]	-3.98 [-25.7]
MRPT-S (1m+5m+30m) (recalibrated)	-3.62 [-45.6]	-3.78 [-39.5]	-3.82 [-32.3]	-3.95 [-25.9]
Ridge (recalibrated)	-3.55 [-37.0]	-3.78 [-36.5]	-3.81 [-30.8]	-3.98 [-26.5]

Notes. untouched final period 2025, 1008425 evaluation origins over 249 sessions. Cells: percentage change in mean scaled CRPS, i.e. CRPS divided by $\sigma_t\sqrt{H}$ versus Random walk (trailing RMS), i.e. $100 \times (\bar{L}_{\text{model}} - \bar{L}_{\text{ref}})/\bar{L}_{\text{ref}}$ (negative = better); in brackets the day-cluster t-statistic of the mean loss differential (its trading-day bootstrap standard error, 1,000 resamples; every origin of one session forms one cluster). Convention: changes and tests on scaled CRPS; these contrasts against the random walk are not part of a Holm family (the pre-registered pairwise contrasts with Holm adjustment are in the pairwise tables). Panel B lists the same forecasts after the scalar recalibration of Section 6 and the Student-t random walk whose scale and degrees of freedom were fitted on 2023; it is reported for completeness and is not used for any ranking in the text. Intervals are in the result file.

Table B.26. Final period: model confidence sets (MCS p-values; † = member of the 90 % set)

Model	CRPS				WIS			
	$H = 5$	$H = 15$	$H = 30$	$H = 60$	$H = 5$	$H = 15$	$H = 30$	$H = 60$
Random walk (trailing RMS)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Seasonal volatility	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
EWMA-t	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
HAR-RV	0.06	0.02	0.07	0.79†	0.00	0.01	0.01	0.56†
GARCH(1,1)-t	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Ridge	1.00†	0.02	0.13†	0.94†	0.23†	0.02	0.03	1.00†
Elastic net	0.95†	0.02	0.09	1.00†	0.12†	0.01	0.03	0.93†
LightGBM	0.00	1.00†	1.00†	0.47†	0.00	1.00†	1.00†	0.93†
MLP (1m)	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
TCN (1m)	0.06	0.00	0.00	0.00	0.12†	0.00	0.00	0.00
MRPT-S (1m+5m+30m)	0.76†	0.00	0.00	0.00	0.77†	0.00	0.00	0.00
MRPT-S (1m only)	0.95†	0.00	0.00	0.00	1.00†	0.00	0.00	0.00
MRPT-B (1m+5m+30m)	0.36†	0.00	0.00	0.00	0.23†	0.00	0.00	0.00
Random walk, Student-t (2023-calibrated)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Notes. untouched final period 2025, 1008425 evaluation origins over 249 sessions. Model confidence set of Hansen, Lunde and Nason (2011) over the 14 models with a per-origin file, computed on the 249 daily mean losses (unscaled CRPS and WIS of the cumulative return) with the $T_{\max}$ statistic and a moving-block bootstrap of the days (block length 5, 1,000 resamples; block 20 in the result file); a model is in the 90 % set (†) when its MCS p-value is at least 0.10; the Student-t random walk calibrated on 2023 (last row) is a reference, not a competitor. Source: `results/metrics/final/mcs_final.csv`.

Table B.27. Final period: scalar recalibration fitted on 2023 — scaled CRPS and 50 % coverage before and after

Model	$H = 5$		$H = 15$		$H = 30$		$H = 60$	
	orig.	recal.	orig.	recal.	orig.	recal.	orig.	recal.
Scaled CRPS								
HAR-RV	0.4597	0.4603	0.4451	0.4453	0.4352	0.4352	0.4233	0.4233
Ridge	0.4591	0.4590	0.4449	0.4447	0.4351	0.4349	0.4235	0.4233
Elastic net	0.4592	0.4591	0.4449	0.4447	0.4351	0.4350	0.4235	0.4233
LightGBM	0.5036	0.4997	0.4443	0.4443	0.4348	0.4347	0.4235	0.4233
MRPT-S (1m+5m+30m)	0.4594	0.4587	0.4462	0.4446	0.4368	0.4349	0.4256	0.4235
Coverage of the central 50 % interval (nominal 0.50)								
HAR-RV	0.503	0.542	0.506	0.530	0.493	0.518	0.499	0.521
Ridge	0.478	0.496	0.516	0.511	0.507	0.512	0.514	0.517
Elastic net	0.478	0.496	0.516	0.511	0.508	0.512	0.515	0.518
LightGBM	0.740	0.742	0.483	0.520	0.512	0.514	0.521	0.518
MRPT-S (1m+5m+30m)	0.550	0.524	0.569	0.533	0.574	0.529	0.580	0.535

Notes. untouched final period 2025, 1008425 evaluation origins over 249 sessions. 'orig.' = the model as trained/fitted; 'recal.' = the same forecasts with a scalar scale c (and, for the parametric baselines, degrees of freedom ν) chosen per model and horizon on a seeded stratified subset of 100,000 validation origins of 2023 to minimise mean CRPS (`results/metrics/recal/c_val.json`); the 2024 and 2025 scores are then computed once with the fitted constants (`scripts/recalibrate.py`, `scripts/review/rescore_scaled.py`). Convention: scaled CRPS. Source: the `_recal` rows of the leaderboard file.

Table B.28. Final period (held-out instruments): mean scaled CRPS of the cumulative-return forecast by horizon (minutes)

Model	$H = 5$	$H = 15$	$H = 30$	$H = 60$
Panel A. Models as specified (baselines) or trained (neural models)				
Random walk (trailing RMS)	0.4813 [0.4728, 0.4897]	0.4682 [0.4598, 0.4771]	0.4586 [0.4495, 0.4679]	0.4471 [0.4369, 0.4580]
Seasonal volatility	0.4705 [0.4614, 0.4794]	0.4563 [0.4478, 0.4658]	0.4468 [0.4367, 0.4569]	0.4351 [0.4236, 0.4474]
EWMA-t	0.4666 [0.4576, 0.4754]	0.4533 [0.4447, 0.4626]	0.4449 [0.4351, 0.4551]	0.4346 [0.4235, 0.4465]
HAR-RV	0.4654 [0.4562, 0.4743]	0.4518 [0.4428, 0.4614]	0.4425 [0.4323, 0.4529]	0.4309 [0.4193, 0.4432]
GARCH(1,1)-t	0.4677 [0.4588, 0.4766]	0.4538 [0.4451, 0.4632]	0.4455 [0.4358, 0.4555]	0.4360 [0.4250, 0.4481]
Ridge	0.4650 [0.4559, 0.4740]	0.4512 [0.4424, 0.4607]	0.4420 [0.4320, 0.4524]	0.4307 [0.4192, 0.4430]
Elastic net	0.4650 [0.4559, 0.4740]	0.4512 [0.4424, 0.4607]	0.4421 [0.4320, 0.4525]	0.4307 [0.4192, 0.4430]
LightGBM	0.4653 [0.4561, 0.4743]	0.4516 [0.4427, 0.4612]	0.4427 [0.4324, 0.4531]	0.4315 [0.4200, 0.4439]
MLP (1m)	0.4651 [0.4562, 0.4740]	0.4524 [0.4438, 0.4618]	0.4433 [0.4335, 0.4533]	0.4321 [0.4207, 0.4443]
TCN (1m)	0.4650 [0.4562, 0.4738]	0.4526 [0.4441, 0.4619]	0.4436 [0.4340, 0.4536]	0.4326 [0.4213, 0.4447]
MRPT-S (1m+5m+30m)	0.4644 [0.4556, 0.4732]	0.4518 [0.4432, 0.4611]	0.4428 [0.4330, 0.4528]	0.4316 [0.4202, 0.4437]
MRPT-S (1m only)	0.4644 [0.4556, 0.4733]	0.4518 [0.4433, 0.4611]	0.4428 [0.4330, 0.4529]	0.4317 [0.4203, 0.4438]
MRPT-B (1m+5m+30m)	0.4646 [0.4558, 0.4735]	0.4519 [0.4434, 0.4613]	0.4429 [0.4331, 0.4529]	0.4317 [0.4203, 0.4438]
MRPT-S trained on all 300 instruments	0.4643 [0.4555, 0.4732]	0.4519 [0.4433, 0.4613]	0.4429 [0.4331, 0.4529]	0.4319 [0.4204, 0.4440]
Panel B. Recalibrated on the 2023 validation origins (scalar scale c and ν per model and horizon; Section 6)				
Random walk, Student-t (2023-calibrated)	0.4730 [0.4637, 0.4823]	0.4582 [0.4489, 0.4682]	0.4476 [0.4374, 0.4579]	0.4343 [0.4229, 0.4470]

Notes. held-out instruments, 2025, 254334 evaluation origins over 249 sessions. Brackets: 95 % trading-day cluster-bootstrap intervals (all instruments and origins of one session form one cluster). Convention: scaled CRPS, i.e. CRPS divided by $\sigma_t \sqrt{H}$.

Table B.29. Final period (held-out): scaled CRPS relative to Random walk (trailing RMS)

Model	$H = 5$	$H = 15$	$H = 30$	$H = 60$
Panel A. Models as specified (baselines) or trained (neural models)				
Seasonal volatility	-2.25 [-19.3]	-2.54 [-22.5]	-2.56 [-20.3]	-2.69 [-17.8]
EWMA-t	-3.06 [-33.5]	-3.18 [-29.3]	-2.98 [-24.9]	-2.80 [-18.5]
HAR-RV	-3.30 [-33.4]	-3.51 [-29.7]	-3.51 [-25.8]	-3.63 [-22.8]
GARCH(1,1)-t	-2.82 [-30.9]	-3.07 [-27.4]	-2.85 [-24.2]	-2.49 [-17.2]
Ridge	-3.39 [-33.6]	-3.64 [-31.9]	-3.60 [-27.4]	-3.67 [-23.8]
Elastic net	-3.38 [-33.6]	-3.63 [-31.8]	-3.59 [-27.2]	-3.68 [-23.8]
LightGBM	-3.32 [-32.4]	-3.56 [-28.2]	-3.46 [-23.9]	-3.50 [-20.6]
MLP (1m)	-3.36 [-38.9]	-3.39 [-34.0]	-3.32 [-30.0]	-3.36 [-25.0]
TCN (1m)	-3.40 [-40.1]	-3.34 [-34.0]	-3.25 [-30.2]	-3.25 [-24.6]
MRPT-S (1m+5m+30m)	-3.52 [-39.9]	-3.51 [-34.5]	-3.44 [-30.4]	-3.47 [-24.3]
MRPT-S (1m only)	-3.50 [-39.9]	-3.50 [-34.6]	-3.43 [-30.2]	-3.45 [-24.3]
MRPT-B (1m+5m+30m)	-3.47 [-39.6]	-3.47 [-34.7]	-3.42 [-30.5]	-3.45 [-24.9]
MRPT-S trained on all 300 instruments	-3.53 [-39.9]	-3.49 [-34.5]	-3.42 [-30.4]	-3.41 [-25.1]
Panel B. Recalibrated on the 2023 validation origins (scalar scale c and ν per model and horizon; Section 6)				
Random walk, Student-t (2023-calibrated)	-1.72 [-19.9]	-2.13 [-18.8]	-2.39 [-17.5]	-2.87 [-16.7]
Elastic net (recalibrated)	-3.37 [-30.8]	-3.65 [-30.3]	-3.63 [-25.3]	-3.79 [-21.4]
HAR-RV (recalibrated)	-3.31 [-31.9]	-3.53 [-29.4]	-3.52 [-24.6]	-3.72 [-20.4]
LightGBM (recalibrated)	-3.37 [-31.5]	-3.56 [-27.6]	-3.49 [-22.9]	-3.58 [-19.1]
Ridge (recalibrated)	-3.37 [-30.7]	-3.65 [-30.4]	-3.63 [-25.4]	-3.78 [-21.4]

Notes. held-out instruments, 2025, 254334 evaluation origins over 249 sessions. Cells: percentage change in mean scaled CRPS, i.e. CRPS divided by $\sigma_t\sqrt{H}$ versus Random walk (trailing RMS), i.e. $100 \times (\bar{L}_{\text{model}} - \bar{L}_{\text{ref}})/\bar{L}_{\text{ref}}$ (negative = better); in brackets the day-cluster t-statistic of the mean loss differential (its trading-day bootstrap standard error, 1,000 resamples; every origin of one session forms one cluster). Convention: changes and tests on scaled CRPS; these contrasts against the random walk are not part of a Holm family (the pre-registered pairwise contrasts with Holm adjustment are in the pairwise tables). Panel B lists the same forecasts after the scalar recalibration of Section 6 and the Student-t random walk whose scale and degrees of freedom were fitted on 2023; it is reported for completeness and is not used for any ranking in the text. Intervals are in the result file.

Table B.30. Final period (held-out): empirical 50/80/95% coverage and PIT KS statistic

Model	$H = 5$				$H = 15$				$H = 30$				$H = 60$			
	50	80	95	KS	50	80	95	KS	50	80	95	KS	50	80	95	KS
Panel A. Models as specified (baselines) or trained (neural models)																
Random walk (trailing RMS)	0.64	0.88	0.96	0.072	0.65	0.89	0.97	0.077	0.65	0.89	0.97	0.080	0.67	0.90	0.97	0.085
Seasonal volatility	0.55	0.82	0.94	0.029	0.55	0.82	0.94	0.030	0.55	0.82	0.95	0.031	0.55	0.83	0.94	0.033
EWMA-t	0.52	0.83	0.97	0.017	0.53	0.84	0.97	0.022	0.54	0.84	0.97	0.024	0.55	0.85	0.97	0.032
HAR-RV	0.50	0.80	0.96	0.015	0.50	0.80	0.96	0.008	0.50	0.81	0.96	0.012	0.52	0.83	0.96	0.019
GARCH(1,1)-t	0.52	0.82	0.96	0.016	0.53	0.83	0.97	0.020	0.54	0.84	0.97	0.026	0.57	0.86	0.97	0.038
Ridge	0.50	0.81	0.96	0.015	0.50	0.81	0.96	0.009	0.51	0.81	0.96	0.012	0.53	0.83	0.96	0.019
Elastic net	0.50	0.81	0.96	0.015	0.50	0.81	0.96	0.009	0.51	0.81	0.96	0.012	0.53	0.83	0.96	0.019
LightGBM	0.49	0.81	0.97	0.012	0.50	0.81	0.96	0.009	0.50	0.81	0.96	0.010	0.52	0.82	0.96	0.016
MLP (1m)	0.54	0.83	0.96	0.024	0.56	0.84	0.96	0.033	0.56	0.85	0.96	0.036	0.57	0.85	0.96	0.042
TCN (1m)	0.54	0.83	0.96	0.027	0.57	0.85	0.96	0.036	0.57	0.85	0.96	0.038	0.58	0.86	0.96	0.043
MRPT-S (1m+5m+30m)	0.54	0.83	0.96	0.024	0.56	0.84	0.96	0.031	0.56	0.84	0.96	0.033	0.57	0.85	0.96	0.037
MRPT-S (1m only)	0.54	0.82	0.96	0.023	0.56	0.84	0.96	0.031	0.56	0.84	0.96	0.033	0.57	0.85	0.96	0.037
MRPT-B (1m+5m+30m)	0.54	0.83	0.96	0.025	0.56	0.84	0.96	0.032	0.56	0.84	0.96	0.034	0.57	0.85	0.96	0.038
MRPT-S trained on all 300 instruments	0.54	0.82	0.96	0.022	0.56	0.84	0.96	0.031	0.56	0.85	0.96	0.034	0.57	0.85	0.96	0.039
Panel B. Recalibrated on the 2023 validation origins (scalar scale c and ν per model and horizon; Section 6)																
Random walk, Student-t (2023-calibrated)	0.50	0.80	0.95	0.015	0.50	0.80	0.95	0.009	0.50	0.79	0.95	0.012	0.50	0.80	0.95	0.013

Notes. held-out instruments, 2025, 254334 evaluation origins over 249 sessions. Empirical coverage of the central 50/80/95 % prediction intervals of the cumulative return (nominal 0.50/0.80/0.95) and the Kolmogorov–Smirnov distance of the PIT values from uniformity; no scaling is involved.

Table B.31. Second held-out set (66 availability-excluded instruments) (held-out): scaled CRPS relative to Random walk (trailing RMS)

Model	$H = 5$	$H = 15$	$H = 30$	$H = 60$
Seasonal volatility	-3.39 [-29.9]	-3.52 [-31.6]	-3.52 [-27.2]	-3.86 [-24.7]
EWMA-t	-4.28 [-39.9]	-4.10 [-38.1]	-3.79 [-28.2]	-3.66 [-23.7]
HAR-RV	-4.46 [-42.0]	-5.04 [-36.7]	-5.03 [-29.4]	-5.37 [-26.1]
GARCH(1,1)-t	-4.48 [-33.2]	-4.37 [-31.2]	-4.11 [-25.2]	-4.01 [-22.2]
Ridge	-5.01 [-39.4]	-5.14 [-37.6]	-5.07 [-30.9]	-5.40 [-27.6]
Elastic net	-5.00 [-39.5]	-5.16 [-37.9]	-5.08 [-30.7]	-5.41 [-27.6]
LightGBM	+3.22 [20.9]	-4.96 [-30.3]	-4.89 [-25.5]	-5.34 [-29.0]
MLP (1m)	-5.10 [-44.6]	-4.96 [-45.1]	-4.77 [-37.1]	-5.04 [-35.8]
TCN (1m)	-5.25 [-44.8]	-4.81 [-42.3]	-4.52 [-34.2]	-4.63 [-33.4]
MRPT-S (1m+5m+30m)	-5.24 [-43.0]	-4.96 [-44.5]	-4.71 [-34.6]	-5.04 [-35.2]
MRPT-S (1m only)	-5.18 [-42.9]	-4.88 [-43.4]	-4.64 [-34.3]	-4.89 [-33.8]
MRPT-B (1m+5m+30m)	-5.20 [-44.8]	-4.89 [-43.9]	-4.68 [-34.0]	-4.91 [-32.5]

Notes. held-out instruments, 2024, 63182 evaluation origins over 251 sessions. Cells: percentage change in mean scaled CRPS, i.e. CRPS divided by $\sigma_t\sqrt{H}$ versus Random walk (trailing RMS), i.e. $100 \times (\bar{L}_{\text{model}} - \bar{L}_{\text{ref}})/\bar{L}_{\text{ref}}$ (negative = better); in brackets the day-cluster t-statistic of the mean loss differential (its trading-day bootstrap standard error, 1,000 resamples; every origin of one session forms one cluster). Convention: changes and tests on scaled CRPS; these contrasts against the random walk are not part of a Holm family (the pre-registered pairwise contrasts with Holm adjustment are in the pairwise tables). Intervals are in the result file.

Table B.32. Second held-out set (66 availability-excluded instruments) (held-out): empirical 50/80/95% coverage and PIT KS statistic

Model	$H = 5$				$H = 15$				$H = 30$				$H = 60$			
	50	80	95	KS	50	80	95	KS	50	80	95	KS	50	80	95	KS
Random walk (trailing RMS)	0.68	0.89	0.97	0.095	0.69	0.90	0.97	0.094	0.69	0.91	0.98	0.096	0.70	0.92	0.98	0.102
Seasonal volatility	0.58	0.83	0.94	0.047	0.58	0.84	0.95	0.040	0.58	0.84	0.95	0.040	0.58	0.84	0.95	0.041
EWMA-t	0.57	0.85	0.97	0.042	0.57	0.86	0.97	0.038	0.58	0.86	0.97	0.040	0.59	0.87	0.98	0.048
HAR-RV	0.57	0.85	0.97	0.040	0.51	0.81	0.96	0.009	0.51	0.81	0.96	0.009	0.52	0.82	0.96	0.015
GARCH(1,1)-t	0.52	0.81	0.95	0.020	0.52	0.82	0.96	0.011	0.53	0.83	0.96	0.017	0.55	0.85	0.97	0.030
Ridge	0.53	0.82	0.96	0.018	0.50	0.80	0.96	0.008	0.50	0.80	0.96	0.011	0.52	0.82	0.96	0.013
Elastic net	0.53	0.82	0.96	0.019	0.50	0.81	0.96	0.008	0.50	0.80	0.96	0.011	0.52	0.82	0.96	0.013
LightGBM	0.73	0.94	0.99	0.114	0.47	0.78	0.95	0.019	0.48	0.78	0.95	0.015	0.51	0.81	0.96	0.012
MLP (1m)	0.55	0.82	0.95	0.030	0.56	0.84	0.96	0.033	0.57	0.84	0.96	0.037	0.58	0.85	0.96	0.044
TCN (1m)	0.55	0.83	0.96	0.033	0.57	0.85	0.96	0.037	0.58	0.85	0.96	0.042	0.59	0.86	0.97	0.049
MRPT-S (1m+5m+30m)	0.55	0.82	0.95	0.031	0.57	0.84	0.96	0.034	0.57	0.84	0.96	0.037	0.58	0.85	0.96	0.041
MRPT-S (1m only)	0.55	0.82	0.95	0.031	0.57	0.84	0.96	0.034	0.57	0.84	0.96	0.038	0.58	0.85	0.96	0.044
MRPT-B (1m+5m+30m)	0.55	0.82	0.95	0.032	0.57	0.84	0.96	0.035	0.57	0.84	0.96	0.039	0.58	0.85	0.96	0.044

Notes. held-out instruments, 2024, 63182 evaluation origins over 251 sessions. Empirical coverage of the central 50/80/95 % prediction intervals of the cumulative return (nominal 0.50/0.80/0.95) and the Kolmogorov–Smirnov distance of the PIT values from uniformity; no scaling is involved.

Table B.33. Second held-out set (66 availability-excluded instruments) (held-out instruments): mean scaled CRPS of the cumulative-return forecast by horizon (minutes)

Model	$H = 5$	$H = 15$	$H = 30$	$H = 60$
Random walk (trailing RMS)	0.4562 [0.4456, 0.4690]	0.4451 [0.4332, 0.4578]	0.4361 [0.4252, 0.4479]	0.4257 [0.4131, 0.4393]
Seasonal volatility	0.4414 [0.4300, 0.4549]	0.4293 [0.4165, 0.4424]	0.4204 [0.4087, 0.4331]	0.4095 [0.3957, 0.4236]
EWMA-t	0.4377 [0.4264, 0.4509]	0.4271 [0.4147, 0.4395]	0.4196 [0.4079, 0.4319]	0.4108 [0.3972, 0.4250]
HAR-RV	0.4369 [0.4256, 0.4500]	0.4234 [0.4104, 0.4365]	0.4148 [0.4026, 0.4279]	0.4041 [0.3899, 0.4186]
GARCH(1,1)-t	0.4365 [0.4248, 0.4501]	0.4249 [0.4124, 0.4373]	0.4173 [0.4052, 0.4300]	0.4082 [0.3943, 0.4227]
Ridge	0.4344 [0.4228, 0.4477]	0.4229 [0.4100, 0.4360]	0.4146 [0.4024, 0.4277]	0.4041 [0.3900, 0.4184]
Elastic net	0.4344 [0.4228, 0.4478]	0.4229 [0.4100, 0.4359]	0.4146 [0.4025, 0.4277]	0.4041 [0.3900, 0.4184]
LightGBM	0.4752 [0.4649, 0.4871]	0.4238 [0.4108, 0.4370]	0.4154 [0.4032, 0.4288]	0.4042 [0.3900, 0.4186]
MLP (1m)	0.4340 [0.4225, 0.4476]	0.4237 [0.4109, 0.4366]	0.4154 [0.4037, 0.4281]	0.4052 [0.3919, 0.4194]
TCN (1m)	0.4334 [0.4222, 0.4468]	0.4238 [0.4113, 0.4366]	0.4160 [0.4043, 0.4287]	0.4063 [0.3929, 0.4207]
MRPT-S (1m+5m+30m)	0.4332 [0.4220, 0.4465]	0.4232 [0.4106, 0.4362]	0.4152 [0.4033, 0.4280]	0.4048 [0.3912, 0.4193]
MRPT-S (1m only)	0.4333 [0.4221, 0.4466]	0.4234 [0.4107, 0.4363]	0.4152 [0.4034, 0.4280]	0.4049 [0.3912, 0.4194]
MRPT-B (1m+5m+30m)	0.4335 [0.4223, 0.4468]	0.4236 [0.4109, 0.4367]	0.4155 [0.4036, 0.4282]	0.4052 [0.3917, 0.4196]

Notes. held-out instruments, 2025, 83322 evaluation origins over 249 sessions. Brackets: 95 % trading-day cluster-bootstrap intervals (all instruments and origins of one session form one cluster). Convention: scaled CRPS, i.e. CRPS divided by $\sigma_t\sqrt{H}$.

Table B.34. H3 control ablations (Small, 4,000 steps): mean scaled CRPS of the cumulative-return forecast by horizon (minutes)

Model	Parameters	$H = 5$	$H = 15$	$H = 30$	$H = 60$
MRPT-S regime stats, reduced global token	3,455,619	0.4531 [0.4476, 0.4590]	0.4413 [0.4350, 0.4479]	0.4312 [0.4249, 0.4381]	0.4191 [0.4110, 0.4288]
MRPT-S no regime, reduced global token	3,416,643	0.4530 [0.4475, 0.4589]	0.4414 [0.4351, 0.4479]	0.4313 [0.4250, 0.4381]	0.4194 [0.4113, 0.4290]
MRPT-S time only, reduced global token	3,455,619	0.4558 [0.4504, 0.4617]	0.4439 [0.4376, 0.4505]	0.4334 [0.4273, 0.4406]	0.4210 [0.4128, 0.4309]
MRPT-S shuffled price/volume columns	3,455,619	0.4537 [0.4482, 0.4596]	0.4420 [0.4357, 0.4484]	0.4318 [0.4255, 0.4386]	0.4197 [0.4115, 0.4294]

Notes. development test 2024, 986468 evaluation origins over 251 sessions. Brackets: 95 % trading-day cluster-bootstrap intervals (all instruments and origins of one session form one cluster). Convention: scaled CRPS, i.e. CRPS divided by $\sigma_t\sqrt{H}$. Parameters: trainable parameters of the trained models (`train_summary.json`); baselines have none in this sense.

Table B.35. Ablations (Small, 4,000 steps): empirical 50/80/95% coverage and PIT KS statistic

Model	$H = 5$				$H = 15$				$H = 30$				$H = 60$			
	50	80	95	KS	50	80	95	KS	50	80	95	KS	50	80	95	KS
Random walk (trailing RMS)	0.65	0.88	0.97	0.078	0.66	0.89	0.97	0.079	0.67	0.90	0.97	0.083	0.68	0.91	0.98	0.089
Seasonal volatility	0.56	0.82	0.94	0.031	0.56	0.83	0.94	0.031	0.56	0.83	0.95	0.033	0.57	0.83	0.95	0.034
EWMA-t	0.54	0.83	0.97	0.023	0.54	0.84	0.97	0.024	0.55	0.85	0.97	0.028	0.56	0.86	0.97	0.034
MRPT-S (1m+5m)	0.55	0.83	0.96	0.030	0.57	0.84	0.96	0.034	0.57	0.85	0.96	0.039	0.58	0.85	0.96	0.041
MRPT-S cross-attention fusion	0.55	0.83	0.96	0.029	0.56	0.84	0.96	0.033	0.57	0.85	0.96	0.038	0.58	0.85	0.96	0.042
MRPT-S Gaussian head	0.57	0.84	0.95	0.037	0.57	0.85	0.96	0.038	0.58	0.85	0.96	0.040	0.58	0.85	0.96	0.044
MRPT-S no regime	0.54	0.83	0.95	0.027	0.56	0.84	0.96	0.031	0.57	0.84	0.96	0.035	0.58	0.85	0.96	0.039
MRPT-S learned regime	0.54	0.82	0.95	0.021	0.55	0.83	0.95	0.029	0.56	0.84	0.96	0.035	0.57	0.84	0.96	0.040
MRPT-S returns only	0.56	0.84	0.96	0.034	0.57	0.85	0.96	0.041	0.58	0.86	0.96	0.045	0.59	0.86	0.96	0.049
MRPT-S time only	0.56	0.84	0.96	0.034	0.57	0.85	0.96	0.040	0.58	0.86	0.96	0.045	0.59	0.86	0.96	0.049
MRPT-S OHLCV (no rel. volume)	0.55	0.83	0.96	0.031	0.57	0.85	0.96	0.036	0.58	0.85	0.96	0.041	0.58	0.86	0.96	0.044
MRPT-S shuffled minutes	0.55	0.83	0.96	0.031	0.57	0.85	0.96	0.037	0.58	0.85	0.96	0.042	0.58	0.86	0.96	0.046
MRPT-S + ticker id	0.55	0.83	0.96	0.029	0.57	0.84	0.96	0.034	0.58	0.85	0.96	0.039	0.59	0.86	0.96	0.044
MRPT-S reference (4k steps)	0.54	0.83	0.95	0.027	0.56	0.84	0.96	0.032	0.57	0.84	0.96	0.037	0.58	0.85	0.96	0.042
MRPT-S 1m only (4k)	0.55	0.83	0.96	0.030	0.57	0.84	0.96	0.035	0.57	0.85	0.96	0.039	0.58	0.85	0.96	0.043
MRPT-S reference, seed 2	0.54	0.82	0.95	0.029	0.56	0.84	0.96	0.034	0.57	0.84	0.96	0.041	0.58	0.85	0.96	0.048
MRPT-S reference, seed 3	0.54	0.82	0.95	0.020	0.56	0.84	0.96	0.037	0.56	0.84	0.96	0.045	0.57	0.85	0.96	0.056
MRPT-S + copula (joint sampling)	0.55	0.83	0.96	0.030	0.57	0.84	0.96	0.036	0.58	0.85	0.96	0.041	0.58	0.86	0.96	0.045
MRPT-S + copula, independent sampling	0.55	0.83	0.96	0.030	0.57	0.84	0.96	0.036	0.57	0.85	0.96	0.040	0.58	0.85	0.96	0.044

Notes. development test 2024, 986468 evaluation origins over 251 sessions. Empirical coverage of the central 50/80/95 % prediction intervals of the cumulative return (nominal 0.50/0.80/0.95) and the Kolmogorov–Smirnov distance of the PIT values from uniformity; no scaling is involved.

Table B.36. Ablations (Small, 4,000 steps): mean WIS ($\times 10^4$) of the cumulative-return forecast by horizon (minutes)

Model	$H = 5$	$H = 15$	$H = 30$	$H = 60$
Random walk (trailing RMS)	7.462 [7.322, 7.608]	12.559 [12.304, 12.834]	17.325 [16.944, 17.700]	23.787 [23.225, 24.432]
Seasonal volatility	7.238 [7.101, 7.381]	12.157 [11.911, 12.421]	16.748 [16.377, 17.129]	22.946 [22.374, 23.564]
EWMA-t	7.182 [7.047, 7.324]	12.089 [11.844, 12.350]	16.701 [16.326, 17.089]	22.978 [22.410, 23.612]
MRPT-S (1m+5m)	7.146 [7.011, 7.288]	12.026 [11.785, 12.286]	16.586 [16.218, 16.966]	22.727 [22.161, 23.342]
MRPT-S cross-attention fusion	7.142 [7.008, 7.284]	12.018 [11.777, 12.279]	16.576 [16.207, 16.956]	22.717 [22.150, 23.333]
MRPT-S Gaussian head	7.161 [7.026, 7.304]	12.041 [11.801, 12.301]	16.604 [16.237, 16.984]	22.751 [22.184, 23.365]
MRPT-S no regime	7.140 [7.006, 7.283]	12.016 [11.771, 12.276]	16.571 [16.201, 16.954]	22.711 [22.142, 23.331]
MRPT-S learned regime	7.138 [7.004, 7.280]	12.004 [11.761, 12.265]	16.550 [16.179, 16.933]	22.671 [22.097, 23.289]
MRPT-S returns only	7.153 [7.018, 7.295]	12.040 [11.795, 12.303]	16.608 [16.237, 16.991]	22.769 [22.210, 23.390]
MRPT-S time only	7.154 [7.019, 7.296]	12.042 [11.797, 12.305]	16.611 [16.239, 16.994]	22.773 [22.214, 23.395]
MRPT-S OHLCV (no rel. volume)	7.144 [7.010, 7.286]	12.024 [11.782, 12.284]	16.587 [16.221, 16.968]	22.736 [22.170, 23.352]
MRPT-S shuffled minutes	7.148 [7.014, 7.290]	12.030 [11.787, 12.292]	16.596 [16.228, 16.979]	22.754 [22.185, 23.378]
MRPT-S + ticker id	7.142 [7.008, 7.284]	12.029 [11.785, 12.288]	16.601 [16.233, 16.982]	22.777 [22.210, 23.392]
MRPT-S reference (4k steps)	7.138 [7.004, 7.281]	12.013 [11.770, 12.273]	16.568 [16.198, 16.949]	22.703 [22.134, 23.324]
MRPT-S 1m only (4k)	7.147 [7.013, 7.289]	12.031 [11.788, 12.291]	16.592 [16.223, 16.972]	22.742 [22.178, 23.359]
MRPT-S reference, seed 2	7.139 [7.004, 7.281]	12.009 [11.766, 12.273]	16.565 [16.190, 16.951]	22.715 [22.141, 23.353]
MRPT-S reference, seed 3	7.137 [7.004, 7.279]	12.013 [11.771, 12.271]	16.573 [16.207, 16.950]	22.740 [22.167, 23.347]
MRPT-S + copula (joint sampling)	7.143 [7.009, 7.284]	12.023 [11.780, 12.283]	16.590 [16.218, 16.968]	22.748 [22.181, 23.366]
MRPT-S + copula, independent sampling	7.143 [7.009, 7.285]	12.022 [11.780, 12.283]	16.584 [16.216, 16.963]	22.729 [22.161, 23.344]

Notes. development test 2024, 986468 evaluation origins over 251 sessions. Brackets: 95 % trading-day cluster-bootstrap intervals (all instruments and origins of one session form one cluster). Convention: weighted interval score of the cumulative return (unscaled). Values are multiplied by 10^4 (raw WIS of order 10^{-3}).

Table B.37. Seed study (12,000 steps): empirical 50/80/95% coverage and PIT KS statistic

Model	$H = 5$				$H = 15$				$H = 30$				$H = 60$			
	50	80	95	KS	50	80	95	KS	50	80	95	KS	50	80	95	KS
MRPT-S seed 2	0.54	0.83	0.95	0.027	0.56	0.84	0.96	0.033	0.57	0.84	0.96	0.039	0.58	0.85	0.96	0.045
MRPT-S seed 3	0.54	0.82	0.95	0.025	0.56	0.84	0.96	0.030	0.56	0.84	0.96	0.033	0.57	0.85	0.96	0.036
MRPT-S seed 4	0.54	0.82	0.95	0.022	0.55	0.83	0.96	0.027	0.56	0.84	0.96	0.032	0.57	0.85	0.96	0.036
MRPT-S seed 5	0.54	0.83	0.96	0.026	0.56	0.84	0.96	0.032	0.57	0.85	0.96	0.038	0.58	0.85	0.96	0.042
MRPT-S 1m only, seed 2	0.54	0.83	0.96	0.028	0.56	0.84	0.96	0.035	0.57	0.85	0.96	0.042	0.58	0.86	0.96	0.049
MRPT-S 1m only, seed 3	0.55	0.83	0.96	0.030	0.57	0.85	0.96	0.036	0.58	0.85	0.96	0.040	0.59	0.86	0.96	0.044
MRPT-S 1m only, seed 4	0.54	0.82	0.95	0.025	0.56	0.84	0.96	0.031	0.57	0.85	0.96	0.035	0.58	0.85	0.96	0.040
MRPT-S 1m only, seed 5	0.55	0.83	0.96	0.029	0.57	0.84	0.96	0.034	0.58	0.85	0.96	0.039	0.58	0.86	0.96	0.044

Notes. development test 2024, 986468 evaluation origins over 251 sessions. Empirical coverage of the central 50/80/95 % prediction intervals of the cumulative return (nominal 0.50/0.80/0.95) and the Kolmogorov–Smirnov distance of the PIT values from uniformity; no scaling is involved.

Table B.38. Seed study (12,000 steps), untouched 2025 period: empirical 50/80/95% coverage and PIT KS statistic

Model	$H = 5$				$H = 15$				$H = 30$				$H = 60$			
	50	80	95	KS	50	80	95	KS	50	80	95	KS	50	80	95	KS
MRPT-S seed 2	0.54	0.83	0.96	0.024	0.57	0.84	0.96	0.034	0.57	0.85	0.96	0.038	0.58	0.85	0.96	0.046
MRPT-S seed 3	0.54	0.82	0.95	0.024	0.56	0.84	0.96	0.031	0.57	0.84	0.96	0.035	0.57	0.85	0.96	0.039
MRPT-S seed 4	0.54	0.82	0.95	0.021	0.56	0.84	0.96	0.030	0.56	0.84	0.96	0.034	0.57	0.85	0.96	0.038
MRPT-S seed 5	0.54	0.83	0.96	0.025	0.57	0.84	0.96	0.034	0.57	0.85	0.96	0.037	0.58	0.85	0.96	0.043
MRPT-S 1m only, seed 2	0.54	0.83	0.96	0.023	0.56	0.84	0.96	0.034	0.57	0.85	0.96	0.039	0.58	0.85	0.96	0.047
MRPT-S 1m only, seed 3	0.55	0.83	0.96	0.029	0.57	0.85	0.96	0.038	0.58	0.85	0.96	0.042	0.59	0.86	0.96	0.046
MRPT-S 1m only, seed 4	0.54	0.82	0.95	0.023	0.56	0.84	0.96	0.032	0.57	0.85	0.96	0.037	0.58	0.85	0.96	0.040
MRPT-S 1m only, seed 5	0.55	0.83	0.96	0.026	0.57	0.85	0.96	0.035	0.57	0.85	0.96	0.038	0.58	0.86	0.96	0.042

Notes. untouched final period 2025, 1008425 evaluation origins over 249 sessions. Empirical coverage of the central 50/80/95 % prediction intervals of the cumulative return (nominal 0.50/0.80/0.95) and the Kolmogorov–Smirnov distance of the PIT values from uniformity; no scaling is involved.

Table B.39. Learning curve (continued training): empirical 50/80/95% coverage and PIT KS statistic

Model	$H = 5$				$H = 15$				$H = 30$				$H = 60$			
	50	80	95	KS	50	80	95	KS	50	80	95	KS	50	80	95	KS
MRPT-S continued to 24k steps	0.55	0.83	0.96	0.029	0.56	0.84	0.96	0.033	0.57	0.85	0.96	0.037	0.58	0.85	0.96	0.040
MRPT-S continued to 36k steps	0.54	0.83	0.96	0.029	0.56	0.84	0.96	0.034	0.57	0.84	0.96	0.039	0.57	0.85	0.96	0.043
MRPT-S 1m only, 24k steps	0.55	0.83	0.96	0.026	0.57	0.84	0.96	0.037	0.57	0.85	0.96	0.043	0.58	0.85	0.96	0.050
MRPT-S 1m only, 36k steps	0.54	0.82	0.95	0.024	0.56	0.83	0.96	0.030	0.57	0.84	0.96	0.036	0.58	0.85	0.96	0.042

Notes. development test 2024, 986468 evaluation origins over 251 sessions. Empirical coverage of the central 50/80/95 % prediction intervals of the cumulative return (nominal 0.50/0.80/0.95) and the Kolmogorov–Smirnov distance of the PIT values from uniformity; no scaling is involved.

Table B.40. Pilot: mean scaled CRPS of the cumulative-return forecast by horizon (minutes)

Model	Parameters	$H = 5$	$H = 15$	$H = 30$	$H = 60$
Random walk (trailing RMS)	–	0.4803 [0.4618, 0.5012]	0.4874 [0.4641, 0.5093]	0.4952 [0.4688, 0.5225]	0.4819 [0.4381, 0.5395]
Seasonal volatility	–	0.4746 [0.4545, 0.4957]	0.4793 [0.4526, 0.5047]	0.4870 [0.4569, 0.5190]	0.4767 [0.4268, 0.5419]
EWMA-t	–	0.4699 [0.4502, 0.4912]	0.4775 [0.4517, 0.5023]	0.4868 [0.4561, 0.5196]	0.4773 [0.4276, 0.5428]
HAR-RV	–	0.4833 [0.4648, 0.5032]	0.4761 [0.4492, 0.5019]	0.4845 [0.4540, 0.5168]	0.4728 [0.4223, 0.5391]
GARCH(1,1)-t	–	0.4714 [0.4522, 0.4932]	0.4783 [0.4545, 0.5018]	0.4876 [0.4602, 0.5167]	0.4789 [0.4331, 0.5388]
Ridge	–	0.4734 [0.4518, 0.4968]	0.4765 [0.4494, 0.5028]	0.4842 [0.4533, 0.5170]	0.4723 [0.4216, 0.5368]
Elastic net	–	0.4732 [0.4516, 0.4967]	0.4764 [0.4494, 0.5026]	0.4841 [0.4532, 0.5169]	0.4725 [0.4217, 0.5371]
LightGBM	–	0.5090 [0.4871, 0.5325]	0.5034 [0.4756, 0.5308]	0.5116 [0.4812, 0.5471]	0.5019 [0.4484, 0.5765]
MLP (1m)	301,827	0.4689 [0.4497, 0.4908]	0.4769 [0.4522, 0.5015]	0.4847 [0.4559, 0.5157]	0.4743 [0.4250, 0.5390]
TCN (1m)	231,459	0.4687 [0.4497, 0.4896]	0.4767 [0.4526, 0.5008]	0.4838 [0.4552, 0.5144]	0.4743 [0.4252, 0.5388]
MRPT-S (1m+5m+30m)	3,455,619	0.4697 [0.4509, 0.4903]	0.4763 [0.4526, 0.5001]	0.4831 [0.4553, 0.5135]	0.4736 [0.4238, 0.5381]
MRPT-S (1m only)	3,200,259	0.4693 [0.4506, 0.4899]	0.4765 [0.4527, 0.5006]	0.4842 [0.4558, 0.5148]	0.4745 [0.4245, 0.5395]

Notes. development test 2024, 4659 evaluation origins over 9 sessions. Brackets: 95 % trading-day cluster-bootstrap intervals (all instruments and origins of one session form one cluster). Convention: scaled CRPS, i.e. CRPS divided by $\sigma_t\sqrt{H}$. Parameters: trainable parameters of the trained models (`train_summary.json`); baselines have none in this sense.

Table B.41. Pilot: empirical 50/80/95% coverage and PIT KS statistic

Model	$H = 5$				$H = 15$				$H = 30$				$H = 60$			
	50	80	95	KS	50	80	95	KS	50	80	95	KS	50	80	95	KS
Random walk (trailing RMS)	0.62	0.87	0.96	0.070	0.61	0.87	0.96	0.063	0.60	0.86	0.96	0.068	0.61	0.87	0.97	0.075
Seasonal volatility	0.53	0.80	0.94	0.029	0.53	0.80	0.94	0.025	0.52	0.79	0.93	0.026	0.52	0.79	0.93	0.039
EWMA-t	0.51	0.81	0.97	0.017	0.50	0.80	0.96	0.012	0.49	0.80	0.96	0.027	0.51	0.80	0.96	0.034
HAR-RV	0.63	0.92	0.99	0.082	0.48	0.78	0.96	0.021	0.46	0.77	0.95	0.039	0.48	0.77	0.95	0.033
GARCH(1,1)-t	0.54	0.85	0.97	0.035	0.56	0.85	0.98	0.037	0.56	0.85	0.98	0.043	0.59	0.87	0.98	0.065
Ridge	0.41	0.70	0.91	0.074	0.44	0.74	0.94	0.042	0.44	0.75	0.94	0.052	0.46	0.75	0.94	0.040
Elastic net	0.41	0.71	0.91	0.075	0.44	0.74	0.94	0.041	0.44	0.74	0.94	0.053	0.46	0.75	0.94	0.038
LightGBM	0.35	0.64	0.89	0.110	0.41	0.71	0.92	0.063	0.46	0.74	0.92	0.044	0.45	0.75	0.92	0.042
MLP (1m)	0.51	0.80	0.95	0.021	0.53	0.81	0.95	0.024	0.51	0.80	0.94	0.030	0.53	0.80	0.94	0.043
TCN (1m)	0.54	0.82	0.96	0.027	0.55	0.82	0.96	0.034	0.52	0.81	0.95	0.032	0.54	0.81	0.95	0.050
MRPT-S (1m+5m+30m)	0.54	0.83	0.96	0.035	0.55	0.83	0.96	0.034	0.53	0.82	0.95	0.037	0.54	0.81	0.95	0.053
MRPT-S (1m only)	0.55	0.83	0.96	0.039	0.56	0.83	0.96	0.038	0.53	0.82	0.95	0.040	0.54	0.81	0.95	0.053

Notes. development test 2024, 4659 evaluation origins over 9 sessions. Empirical coverage of the central 50/80/95 % prediction intervals of the cumulative return (nominal 0.50/0.80/0.95) and the Kolmogorov–Smirnov distance of the PIT values from uniformity; no scaling is involved.

Table B.42. Compute ledger by run: every trained model of the study and of the review runs (training wall-clock; prediction excluded)

Experiment	Model	Node (hostname)	Chip	Steps	Parameters	Wall-clock (s)	Examples/s
main	MLP (1m)	m3u96b	Apple M3 Ultra	12,000	301,827	839	916
main	MRPT-B (1m+5m+30m)	m3u96b	Apple M3 Ultra	12,000	20,718,339	4,580	168
main	MRPT-S (1m+5m+30m)	m3u96b	Apple M3 Ultra	12,000	3,455,619	1,759	437
main	MRPT-S (1m only)	m3u96b	Apple M3 Ultra	12,000	3,200,259	1,088	706
main	TCN (1m)	m3u96b	Apple M3 Ultra	12,000	231,459	927	828
ablations	MRPT-S (1m+5m)	m3u96b	Apple M3 Ultra	4,000	3,330,819	720	356
ablations	MRPT-S 1m only (4k)	m3u96b	Apple M3 Ultra	4,000	3,200,259	618	414
ablations	MRPT-S reference (4k steps)	m3u96b	Apple M3 Ultra	4,000	3,455,619	842	304
ablations	MRPT-S + copula (joint sampling)	m3u96b	Apple M3 Ultra	4,000	3,456,584	855	299
ablations	MRPT-S cross-attention fusion	m3u96b	Apple M3 Ultra	4,000	5,533,443	794	322
ablations	MRPT-S Gaussian head	m3u96b	Apple M3 Ultra	4,000	3,455,619	853	300
ablations	MRPT-S learned regime	m3u96b	Apple M3 Ultra	4,000	3,421,643	853	300
ablations	MRPT-S no regime	m3u96b	Apple M3 Ultra	4,000	3,416,643	855	299
ablations	MRPT-S OHLCV (no rel. volume)	m3u96b	Apple M3 Ultra	4,000	3,455,619	855	299
ablations	MRPT-S returns only	m3u96b	Apple M3 Ultra	4,000	3,455,619	854	300
ablations	MRPT-S reference, seed 2	m3u96b	Apple M3 Ultra	4,000	3,455,619	858	299
ablations	MRPT-S reference, seed 3	m3u96b	Apple M3 Ultra	4,000	3,455,619	855	300
ablations	MRPT-S shuffled minutes	m3u96b	Apple M3 Ultra	4,000	3,455,619	853	300
ablations	MRPT-S + ticker id	m3u96b	Apple M3 Ultra	4,000	3,501,699	854	300
ablations	MRPT-S time only	m3u96b	Apple M3 Ultra	4,000	3,455,619	853	300
seeds	MRPT-S 1m only, seed 2	M4M64c	Apple M4 Max	12,000	3,200,259	932	824
seeds	MRPT-S 1m only, seed 3	M4M64c	Apple M4 Max	12,000	3,200,259	931	825
seeds	MRPT-S 1m only, seed 4	M4M64c	Apple M4 Max	12,000	3,200,259	938	819
seeds	MRPT-S 1m only, seed 5	M4M64c	Apple M4 Max	12,000	3,200,259	932	824
seeds	MRPT-S seed 2	MacStudnpierre6	Apple M4 Max	12,000	3,455,619	2,062	372
seeds	MRPT-S seed 3	MacStudnpierre6	Apple M4 Max	12,000	3,455,619	2,023	380
seeds	MRPT-S seed 4	MacStudnpierre4	Apple M4 Max	12,000	3,455,619	2,034	378
seeds	MRPT-S seed 5	MacStudnpierre4	Apple M4 Max	12,000	3,455,619	2,024	379
budget	MRPT-S 1m only, 24k steps	M5M36	Apple M5 Max	20,000	3,200,259	743	689
budget	MRPT-S 1m only, 36k steps	M5M36	Apple M5 Max	30,000	3,200,259	912	702
budget	MRPT-S continued to 24k steps	MacStudnpierre2	Apple M3 Ultra	24,000	3,455,619	1,928	398
budget	MRPT-S continued to 36k steps	MacStudnpierre2	Apple M3 Ultra	34,000	3,455,619	1,564	409
transfer300	MRPT-S trained on all 300 instruments	m3u96b	Apple M3 Ultra	12,000	3,455,619	1,902	404
ablations_h3	MRPT-S no regime, reduced global token	MacStudnpierre2	Apple M3 Ultra	4,000	3,416,643	640	400
ablations_h3	MRPT-S regime stats, reduced global token	MacStudnpierre2	Apple M3 Ultra	4,000	3,455,619	640	400
ablations_h3	MRPT-S shuffled price/volume columns	M5M36	Apple M5 Max	4,000	3,455,619	933	274
ablations_h3	MRPT-S time only, reduced global token	M5M36	Apple M5 Max	4,000	3,455,619	936	273
Total	37 runs					43,639 (12.1 GPU-h; baseline study 6.0 GPU-h)	

Notes. Wall-clock of the training loop from `train_summary.json` of every run (including the periodic validation passes; the prediction passes over the evaluation origins are not included). Node = hostname recorded by the run; the chip comes from `results/metrics/review_runs*.jsonl` (ledger of the review runs) and from `benchmark_M3U96b.json` for the baseline study. Continued runs (24k, 36k) report the steps of the continuation segment. Throughput is not comparable across chips; comparisons between models are at matched optimiser steps, never at matched wall-clock.

C. Universe and protocol

Table C.1. The coverage gate of the universe is an ex-post availability filter: membership counts when the regular-hours one-minute coverage threshold (80 %) is computed over 2018–2025 versus over the training years only

Statistic	Count
Candidates eligible on the 2017-12-29 rules (liquidity, price, history)	2,147
Pass the coverage gate computed over 2018–2025 (the production gate)	1,152
Would pass the same gate computed over the training years 2018–2022 only	1,178
Pass on the training years but excluded by the full-window gate (look-ahead exclusions; the second held-out set)	66
Fail on the training years but admitted by the full-window gate	40
Candidates removed by measured coverage below the gate	994
of which with coverage between 50 % and the gate	666
of which with coverage below 50 %	328
Coverage-audit failures assigned coverage 0 (GLOB; active instrument, query error)	1
Instruments selected	300
selected although below the gate on the training years	12
selected with 2024 coverage below the gate	27
selected with 2025 coverage below the gate	20
selected with at least one evaluation year (2023–2025) below the gate	35
Delisted candidates passing the gate (instrument-master date 2024-12-20; bars continue through 2025)	5
Delisted instruments among the selected (CNO; 2025 bars present)	1
Spearman correlation of 2017 dollar volume with measured coverage over the candidates	0.755

Notes. The universe was drawn on 2017-12-29 information for liquidity, price and history, then filtered on the realised share of regular-hours minutes with a bar over 2018–2025 (`scripts/audit_coverage.py`); the gate therefore reads the evaluation years, including the untouched 2025, and removes names whose trading dried up afterwards. The counts recompute the gate from the audited coverage file on the training years alone; the 66 instruments that pass on the training years but fail on the full window are scored zero-shot as a second held-out set (Table 17). Source: `results/metrics/universe_gate_sensitivity.csv` (`scripts/review/lookahead_universe.py`) from `coverage_audit_candidates.csv` and the manifest.

Table C.2. Share of candidate origins removed by the future-presence rule (at least 80 % of the next 60 one-minute slots must carry a bar), by instrument role and split

Role	Split	Instruments	Candidate origins	Eligible	Removed	Removed (%)
pooled	train	240	82,907,077	76,967,041	5,940,036	7.2
pooled	val	240	1,123,311	1,040,345	82,966	7.4
pooled	test	240	1,106,611	986,468	120,143	10.9
pooled	final	240	1,109,784	1,008,425	101,359	9.1
held-out	train	60	20,531,715	18,654,708	1,877,007	9.1
held-out	val	60	283,116	266,237	16,879	6.0
held-out	test	60	278,001	250,322	27,679	10.0
held-out	final	60	278,653	254,334	24,319	8.7

Notes. Candidate origins = every slot from 30 to the last slot whose 60-minute horizon stays inside the session (stride 1 for training, counted before the per-session cap; the evaluation stride 15 for validation, test and final); the rule is common to all models and is applied at the origin, not at the forecast. The sensitivity with the rule removed (`min_future_present = 0`) was not run (optional package P8b). Source: `results/metrics/origin_eligibility.csv` (`scripts/review/future_presence_share.py`).

Table C.3. Protocol constants as run, regenerated from the run metadata and configurations (the values every sentence of the paper uses)

Constant	Value	Source record
Evaluation stride between origins (minutes)	15	main/run_meta_test.json config.eval.stride
Candidate origin slots per normal session	20	arange(m_min, 389 - H + 1, stride) from main/run_meta_test.json config.windows + eval.stride
Candidate origin slots per early-close session	8	arange(m_min, 209 - H + 1, stride)
First origin slot (minutes after the open)	30	config.windows.m_min
Last origin slot on a normal session	315	last eligible slot on a normal session
Future-presence rule (% of the next 60 slots)	80	config.windows.min_future_present
Realised origins per instrument-session, 2024	16.4	origins.test / (n_pooled x n_days)
Predictive draws per origin	200	main/run_meta_test.json config.eval.n_samples
Horizons (minutes)	5, 15, 30, 60	main/run_meta_test.json config.eval.horizons
One-minute context (bars)	390	config.windows.L_A
Five-minute context (bars)	390	config.windows.L_B
Thirty-minute context (bars)	260	config.windows.L_C
Tokens per step, multi-resolution	339	L_A/5 + L_B/3 + L_C/2 + 1 (patch sizes of model.tex)
Tokens per step, 1m only	79	L_A/5 + 1
Training origins (pooled)	586,202	main/run_meta_test.json origins.train
Validation origins (pooled, 2023)	1,040,345	main/run_meta_test.json origins.val
Test origins (pooled, 2024)	986,468	main/run_meta_test.json origins.test
Final origins (pooled, 2025)	1,008,425	main/run_meta_test.json origins.final
Test origins (held-out, 2024)	250,322	main/run_meta_test_heldout.json origins.test
Final origins (held-out, 2025)	254,334	main/run_meta_test_heldout.json origins.final
Sessions, 2024	251	main/leaderboard_test.csv n_days
Sessions, 2025	249	final/leaderboard_final.csv n_days
Optimiser steps, main study	12,000	config.train.max_steps
Batch size	64	config.train.batch_size
Passes over the training origins	1.31	max_steps x batch_size / origins.train
Peak learning rate	3×10^{-4}	config.train.lr
Warm-up steps	400	config.train.warmup_steps
Weight decay	0.01	config.train.weight_decay
Gradient clipping	1.0	config.train.grad_clip
Validation every (steps)	1,000	config.train.eval_every
Early-stopping patience (evaluations)	4	config.train.patience
Validation subset of the original runs (origins)	3,840	val_max_batches x batch_size (trainer.py:95 uses batch_size)
Optimiser steps, ablations	4,000	ablations/run_meta_test.json config.train.max_steps
Parameters, MRPT-S (1m+5m+30m)	3,455,619	main/mrpt_small/train_summary.json n_params
Parameters, MRPT-S (1m only)	3,200,259	main/mrpt_small_1m_only/train_summary.json n_params
Parameters, MRPT-B	20,718,339	main/mrpt_base/train_summary.json n_params
Parameters, TCN	231,459	main/tcn/train_summary.json n_params
Parameters, MLP	301,827	main/mlp/train_summary.json n_params
Parameters, cross-attention variant	5,533,443	ablations/mrpt_small_cross/train_summary.json n_params
Training seed, main study	1	config.train.seed

Notes. Generated by scripts/review/protocol_constants.py into results/protocol_constants.json and paper/protocol_macros.tex; a replicator following these values runs the experiment of the result files (the v1.0 text stated a 10-minute stride, 36 and 24 origins per session, 252 days and 256 draws, none of which matched the configuration).

Table C.4. Universe allocation: instruments per asset class × dollar-volume tercile × daily-volatility tercile (selection date 2017-12-29; 60 of the 300 instruments are held out)

Asset class	ADV tercile	Vol. low	Vol. mid	Vol. high
etf	high	8	8	8
etf	low	9	8	9
etf	mid	9	8	8
stock	high	25	25	25
stock	low	25	25	25
stock	mid	25	25	25
held out		stocks 42, ETFs 18		

Notes. Instruments drawn on 2017-12-29 information (dollar volume, price, history) in a stratified design (asset class × 2017 dollar-volume tercile × 2017 daily-volatility tercile), then filtered on the realised 2018–2025 regular-hours one-minute coverage (an ex-post availability filter, Table C.1); 300 instruments, of which 60 held out for the transfer test and 1 delisted in the vendor instrument master (bars through 2025 nonetheless). Source: data/manifests/universe_main.json.

Table C.5. Sector (SIC division) and ETF category counts in the universe

Asset class	Group	*n*
etf	broad_equity	12
etf	commodity	5
etf	fixed_income	8
etf	international	13
etf	leveraged_inverse	14
etf	other	10
etf	sector	10
etf	volatility_alt	3
stock	construction	13
stock	finance_insurance_realestate	30
stock	manufacturing	34
stock	mining	19
stock	retail	26
stock	services	32
stock	transport_utilities	32
stock	unknown	25
stock	wholesale	14

Notes. Stocks are grouped by SIC division from the vendor reference ('unknown' = no SIC code in the instrument master); ETFs by the category assigned from the vendor name. Counts over the 300 selected instruments (pooled and held out). Source: `data/manifests/universe_main.json`.

Pooled instruments (240). ABT, ACH, ACM, ADP, AG, AGG, ALB, ALK, AM, AMLP, APA, ARCC, ARKK, ARMK, AU, AUPH, AXTA, BB, BCE, BDN, BF.B, BHP, BKD, BLDR, BLMN, BND, BOX, BTU, CAH, CB, CG, CHD, CHTR, CI, CMCSA, CME, CNC, CNDT, CNO, CNX, COR, CPRT, CVBF, CVE, CX, CZR, D, DB, DBC, DDD, DGX, DHI, DIS, DKS, DLR, DXC, EBS, EFX, EHC, EMLC, EMN, EOG, EQNR, EQT, ERIC, EWA, EWT, EWY, FAST, FAZ, FEZ, FHN, FLR, FPE, FSLR, FXI, GDDY, GEO, GIL, GLD, GLNG, GNTX, GOOGL, GOOS, GPC, GSAT, GTN, HAL, HMC, HOMB, HOPE, HSBC, HSIC, HXL, IEF, IGSB, ILF, INDA, IRBT, IRDM, ITUB, IVE, IVZ, IWF, IXUS, IYE, J, JPM, KBE, KBR, KBWB, KIM, KLAC, KMI, KSS, LBTYA, LDOS, LEN, LNT, LOW, LQD, LRCX, LYG, M, MAA, MAC, MINT, MOMO, MPC, MRSH, MUR, MUX, NAVI, NEM, NI, NOW, NRG, NTNX, ODFL, OIH, OKE, OPLN, OPTU, OXY, PAGP, PANW, PBA, PBR, PBR.A, PDBC, PEP, PFF, PFG, PFGC, PHM, PK, POR, PWR, RCI, RIO, RIOT, ROL, ROST, RPM, RUN, SABR, SAN, SBH, SBSW, SBUX, SCHG, SCO, SDOW, SDS, SFM, SH, SHEL, SHV, SHW, SJNK, SLF, SMFG, SPDW, SPEM, SPGI, SPWR, SPY, SPYD, SRPT, SSNC, STAG, TAP, TDC, TEL, TIP, TJX, TMO, TNA, TPR, TRI, TTE, TU, TXRH, UNFI, UPRO, URBN, USFD, USMV, VALE, VDE, VEU, VFH, VGK, VIAV, VIG, VIV, VNQ, VTEB, VTR, VWO, VXX, VZ, WAB, WDAY, WEN, WERN, WES, WIT, WTRG, WU, WWW, WY, XBI, XLP, XLU, XME, XPO, XYL, ZION, ZTO

Held-out instruments (60). ACAD, AEE, AEP, AVT, BDX, BHC, BOTZ, BRK.B, CAKE, CAT, CNI, CVNA, ENB, FNV, FTI, GPK, H, HD, HDB, HLF, HTGC, IAU, ING, ITB, KBH, KRG, LKQ, MCK, MSFT,

MTZ, ORCL, PDM, PSQ, PTEN, RBA, SCHA, SCHF, SCHH, SOXX, SPHD, SPSB, SSO, SYY, TLT, TOL, TROW, TSCO, TZA, UAL, UDOW, UE, VEA, VMBS, VMC, W, WMT, WRB, XLRE, XRT, YUM