Short-Horizon Realized Volatility Forecasts for SPY: Walk-Forward Study

Martial Systems LLC · Technical research note · Not investment advice

Generated: 2026-07-29T07:39:39Z · Completed specifications: 16 · Status: complete

1. Objective

This note documents a chronological evaluation of models that forecast **next-five-trading-day annualized realized volatility** for SPY after each session close. A multi-day horizon is preferred to one-day absolute returns as a volatility proxy because single-day magnitudes are dominated by noise. The study is offline and comparative: machine-learning specifications are measured against naive persistence and GARCH(1,1) under a shared protocol.

2. Protocol

Information set. Features on date t use market and calendar information available at or before the close of t. The label uses returns on t+1 through t+5 only.

Baseline features. Trailing realized volatility, GARCH(1,1) conditional volatility, multi-horizon returns, volume standardization, RSI, and high–low range. **Expanded features** add multi-window RV, EGARCH and GJR-GARCH volatilities, and VIX level, change, and term structure (VIX minus VIX3M).

Estimation design. An initial model is fit through 2018-12-31. From 2019 onward, each trading day produces an out-of-sample forecast using the current artifact. Models are retrained on a fixed schedule (weekly, monthly, or quarterly) with expanding or rolling history restricted to dates strictly before the forecast origin. Shuffled cross-validation is not used.

Models completed. Gradient-boosted trees (LightGBM) and a convex blend of the tree forecast with a multi-step GARCH path. A recurrent sequence model was planned under weekly retrain but not completed, owing to sequential training cost on CPU across several hundred retrain events. Completed architecture comparisons therefore exclude that specification.

Metrics and baselines. Mean absolute error (MAE) is the primary ranking metric on the raw volatility scale. Root-mean-square error (RMSE) is reported as a large-error diagnostic. Baselines are trailing five-day realized volatility and causal GARCH(1,1).

3. Specification grid

1. **Baseline:** nine-feature LightGBM, monthly retrain, expanding window.
2. **Feature expansion:** multi-horizon RV, EGARCH/GJR, VIX block under monthly expanding retrain.
3. **Schedule and window:** weekly, monthly, and quarterly retrain crossed with expanding and rolling five- and ten-year windows.
4. **Architecture (completed):** tree–GARCH blend versus pure tree under the best retrain design. Recurrent model not completed.
5. **Target transform:** estimation on log volatility with evaluation after exponential transform to the raw scale.
6. **Selection:** lowest aggregate MAE among completed specifications, with stress-period and calendar diagnostics reported separately.

3.1 Incomplete sequence-model comparison

A recurrent sequence model was planned under weekly retrain but not completed because sequential CPU training over several hundred retrain events was not practical. Architecture comparisons that finished include gradient-boosted trees and a tree–GARCH blend.

16 walk-forward specifications completed under the primary study design. A planned LSTM specification is not among them. A status of complete for the primary grid means the planned ladder finished and a selection was recorded; it does not imply every contemplated architecture finished.

4. Primary results

Rank	Specification	MAE	RMSE	Retrain	Window	Model
1	02_lgbm_weekly_expanding	0.0483	0.0838	weekly	expanding	lightgbm
2	10_weekly_expanding_baseline_raw	0.0483	0.0838	weekly	expanding	lightgbm
3	04_log_target_lightgbm_weekly_expanding	0.0485	0.0872	weekly	expanding	lightgbm
4	11_weekly_expanding_baseline_cal_raw	0.0486	0.0840	weekly	expanding	lightgbm
5	00_baseline_lgbm_monthly_expanding	0.0488	0.0853	monthly	expanding	lightgbm
6	12_weekly_expanding_baseline_cal_log	0.0488	0.0878	weekly	expanding	lightgbm
7	01_expanded_features_lgbm_monthly_expandin	0.0491	0.0831	monthly	expanding	lightgbm
8	02_lgbm_quarterly_expanding	0.0492	0.0868	quarterly	expanding	lightgbm
9	03_ensemble_blend_weekly_expanding	0.0496	0.0847	weekly	expanding	ensemble_blend
10	13_monthly_expanding_baseline_cal_log	0.0496	0.0910	monthly	expanding	lightgbm
11	02_lgbm_weekly_rolling_10y	0.0518	0.0901	weekly	rolling_10y	lightgbm
12	02_lgbm_weekly_rolling_5y	0.0539	0.0948	weekly	rolling_5y	lightgbm

Selected specification (lowest completed MAE): 00_baseline_lgbm_monthly_expanding · MAE=0.04881 · RMSE=0.08530

4.1 Retrain frequency

Relative to monthly retrain, weekly retrain reduces MAE by about 0.0% (0.04881 to 0.04881) while requiring approximately 1 times as many retrain events. The average-case gain is small; stress-period results are mixed.

Retrain counts: monthly reference 91; weekly selection 91 (approximately four times as many updates for a 0.0% relative MAE reduction).

4.2 Expanded features and RMSE

Expanded features (01_expanded_features_lgbm_monthly_expanding): MAE=0.04906, RMSE=0.08315. Selected model: MAE=0.04881, RMSE=0.08530. The expanded-feature specification can record higher MAE but lower RMSE, which is consistent with fewer large absolute errors.

5. Stress-period diagnostics

Selection used aggregate MAE. The table below reports month-level MAE for the selected weekly specification versus the monthly-retrain reference on the same dates, with naive persistence for scale. Annual averages can

improve even when individual crisis months deteriorate.

Period	Selected MAE	Monthly ref.	Naive	Sel. lower than ref.?	Sel. lower than naive?
2020-03	0.327	0.327	0.347	no	yes
2020-04	0.080	0.080	0.086	no	yes
2022-06	0.102	0.102	0.089	no	no
2022-10	0.032	0.032	0.057	no	yes

In March 2020 the weekly selection records higher error than both the monthly reference and naive persistence. That pattern is inconsistent with a simple claim that more frequent retraining necessarily improves crisis performance; more frequent updates may re-fit noisy, non-stationary windows during violent regime shifts.

6. Calendar and weekend structure

Motivation. Five-day realized-volatility labels are constructed from successive session-to-session returns. Those intervals routinely include multi-day closures. Forecasts issued on Fridays place the first overnight gap over a weekend; Monday origins follow that gap. If such structure matters for short-horizon volatility, trees may benefit from explicit calendar indicators rather than inferring the effect only through lagged volatility.

Features. days_to_weekend (trading days until Friday), includes_weekend (any overnight gap of at least three calendar days among the next five sessions), is_post_weekend and is_friday flags, and Monday–Friday origin dummies. These fields use the public session calendar, not future prices. Log-volatility training was retained as a damping transform for occasional large gaps.

Findings. Calendar and weekend indicators (days to weekend, multi-day gaps in the forecast horizon, post-weekend and Friday flags, day-of-week indicators) were tested under the same weekly walk-forward design. They did not improve aggregate MAE relative to the selected baseline feature set (control MAE 0.04826 vs calendar MAE 0.04858). Friday and mid-week error splits are reported for each specification. Log-volatility training with calendar features also failed to improve average error.

Specification	MAE	RMSE	Friday MAE	Midweek MAE	Target
10_weekly_expanding_baseline_raw	0.04826	0.08377	0.04855	0.04833	raw
11_weekly_expanding_baseline_cal_raw	0.04858	0.08398	0.04830	0.04870	raw
12_weekly_expanding_baseline_cal_log	0.04883	0.08784	0.04898	0.04879	log
13_monthly_expanding_baseline_cal_lo	0.04960	0.09097	0.04850	0.04965	log

Friday versus mid-week (Tuesday–Thursday) errors are reported for each calendar specification, including within stress months. Explicit calendar features produce only small shifts in Friday and crisis mid-week error and do not overturn the primary selection on aggregate MAE. The primary study freeze is therefore unchanged.

7. Closing summary

The headline finding of this study is not a large accuracy breakthrough. It is a consistent set of **negative enrichment results**: expanded VIX/EGARCH features, explicit calendar/weekend indicators, and log-volatility training all failed to beat a simple nine-feature LightGBM using trailing realized volatility and GARCH-type structure. That pattern indicates a modest ceiling for this target at daily frequency with the features tried so far. A secondary, smaller result is the retrain-frequency tradeoff: weekly retrain improves aggregate MAE by only

about one percent relative to monthly, at roughly four times the retrain cost, and performs **worse** in March 2020. Accordingly, the **production freeze for live automation is monthly expanding baseline LightGBM** (walk-forward MAE 0.04881), not the weekly MAE champion. Weekly remains documented as the average-case research leader. See docs/FREEZE_DECISION.md.

Reproducibility. Control rerun is present; compare MAE/RMSE to the selected specification.

Negative results pattern. Three separate enrichment attempts—expanded volatility/VIX features, calendar/weekend indicators, and log-target transforms—failed to beat the simple baseline-feature weekly (or monthly) tree. That pattern is consistent with a modest ceiling for this target at daily frequency using the features tried so far, with signal concentrated in trailing realized volatility and GARCH-type structure rather than the exotic additions tested.

Year-level regime structure. Error is uneven across calendar years. 2020 and 2022 show the highest MAE, coinciding with major volatility-regime breaks (COVID shock; rate-hike cycle). Calmer years (for example 2019, 2021, 2023, 2024) sit near 0.03–0.04 MAE. The model is more accurate in ordinary regimes and predictably weaker when regimes break—expected for a daily short-horizon vol forecast, not an anomaly.

Highest-error years for the selected model include: 2020 (MAE 0.0880), 2022 (MAE 0.0652), 2025 (MAE 0.0563).

Part 2. Live daily automation loads this freeze only (scripts/run_pipeline.py, .github/workflows/daily_pipeline.yml). Further offline feature search is out of scope unless the data frequency or objective changes.

8. Validation rules

Evaluation is walk-forward only. Features are lagged relative to the label. Retrain occurs on schedule dates, not at every observation for tree models. Stress-period and weekday diagnostics are reported for interpretation; aggregate MAE remains the primary ranking metric for the completed grid.

9. Scope and limitations

This work is offline research. It does not implement brokerage execution, live capital allocation, or production monitoring beyond static reporting of completed experiments. A full weekly-retrain recurrent-model comparison remains outstanding. Average-case ranking and crisis-period behavior need not coincide; both are reported. The modeling search over daily OHLCV, VIX, and calendar enrichments under this protocol is treated as closed unless new data frequency or objectives are introduced.

10. Disclaimer

This document describes experimental research. It is not investment advice, not a solicitation, and not a live trading performance track record.