Proceedings of the 5th International Conference on Financial Technology and Business Analysis
DOI: 10.54254/2754-1169/2026.38677

Asset-Specific, State-Dependent, and Target-Sensitive Volatility Modeling: Evidence from GARCH-Family Forecasts Across Asset Classes

Yilun Wu

Johns Hopkins University School of Advanced International Studies, Washington, USA
mmmelainemmm@gmail.com

Abstract. This paper studies whether the relative performance of volatility models differs across asset classes and whether the differences can help improve forecasting. The study uses daily data for U.S. equities (SPY), investment-grade bonds (AGG), broad commodities (DBC), gold (GLD), and listed real estate (IYR) and compares GARCH, EGARCH, FIGARCH, and component GARCH within a common recursive out-of-sample framework. Historical MS-GARCH models are used to identify high- and low-volatility periods, while VIX-based and asset-specific volatility states are used to examine changes in forecast performance across market conditions. Forecast routing, GAS Tree parameter adaptation, and machine-learning-assisted model selection are also examined. SPY shows recurring support for EGARCH, but the best-performing fixed specification differs across assets. Some model comparisons show state-dependent performance, but routing and machine-learning selectors do not consistently outperform strong fixed benchmarks. GAS Tree provides selective gains, especially for AGG, but its advantage over Standard GAS reverses under the Garman–Klass proxy. Model suitability therefore depends on the asset, market conditions, and evaluation target.

Keywords: volatility forecasting, GARCH-family models, cross-asset volatility, regime dependence, adaptive forecasting

1. Introduction

Portfolio construction involves combining assets with different sources of risk to achieve a desired balance between risk and return. A diversified portfolio may include equities, fixed-income securities, commodities, precious metals, and real estate, whose volatility patterns can differ in terms of asymmetry, persistence, and responses to changing market conditions. Therefore, for portfolio risk analysis, the question is not only how volatile each asset is, but whether the same type of model is able to capture these differences.

To examine this question, five widely traded instruments are considered to represent different sources of risk that may appear within a diversified portfolio: SPY for U.S. equities, AGG for U.S. investment-grade fixed income, DBC for broad commodities, GLD for gold, and IYR for listed real

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2754-1169/2026.38677

estate. VIX, the volatility index, is included in the exploratory analysis and later used as an observable indicator of overall market stress.

All five assets are compared using the same data treatment, forecasting design, and loss function. The Generalized Autoregressive Conditional Heteroskedasticity (GARCH) framework and its extensions provide a useful basis as different specifications are designed to capture persistence, asymmetry, long memory, long-run and short-run components, and regime changes.

Therefore, the first goal is to test whether GARCH-family model suitability differs across assets. Volatility characteristics are examined using 10-, 15-, and 20-year samples to see which patterns remain similar across historical windows and which change. A common recursive out-of-sample experiment then compares forecasting performance.

This then leads to the second question of whether model performance changes with market conditions. To examine this, the common-market and asset-specific volatility states are distinguished using historical VIX conditions and the recent volatility of each asset to identify normal and stressful periods.

The study then uses forecast routing and machine-learning selectors to test whether state information can improve selection among GARCH-family specifications while GAS Tree allows observable conditions to modify parameters within the variance model.

The results suggest that no single GARCH specification performs best across all five asset classes and that model performance can change across volatility states. However, hard routing and machine-learning-based model selection generally do not provide reliable improvements over strong fixed GARCH-family benchmarks. GAS Tree provides better results in some cases, especially for AGG, although these improvements are not consistent across all settings.

The study does not aim to identify the best volatility model in general. Terms such as "preferred" and "best-performing" refer only to the specifications and forecasting design examined here. The goal is to provide a systematic comparison of GARCH-family framework across several major asset classes, examine which volatility characteristics matter across representative portfolio asset classes, when the relative performance of alternative GARCH specifications changes, and whether observable information about those changes can be used productively.

2. Literature review

Bollerslev [1] extended the ARCH framework introduced by Engle [2] by allowing past conditional variances to affect current variance; Nelson [3] introduced EGARCH to capture asymmetric responses to positive and negative return shocks; Baillie et al. [4] introduced FIGARCH to incorporate long-memory behavior; and Lee and Engle [5] separated conditional variance into permanent and temporary components to describe long-run and short-run changes in volatility.

These models belong to the broader GARCH family but describe different volatility structures. Therefore, whether their forecasting performance can produce comparable across different financial assets is the question. Hansen and Lunde [6] compare 330 ARCH-type models and find that GARCH(1,1) is difficult to outperform for exchange-rate volatility, while models incorporating leverage effects perform better for IBM returns.

However, model suitability may also change over time within the same asset. Haas et al. [7] address regime variation with a Markov-switching GARCH model that allows regime-specific variance dynamics.

If relative model performance varies across states, then the next question is how state information should enter the forecasting model. Creal et al. [8] develop the Generalized Autoregressive Score (GAS) framework, which allows time-varying parameters to be updated using the score of the

DOI: 10.54254/2754-1169/2026.38677

predictive likelihood. More recently, Patton and Simsek [9] combine GAS models with decision trees and random forests, allowing parameters to change according to observable state variables while keeping the main time-series structure of the model.

These studies provide the basis for examining differences across assets, changes across volatility states, and methods that adapt to those changes. This paper brings these questions into a common forecasting framework. It compares the same GARCH-family specifications across assets and observable states, then tests whether state information improves forecasts through model selection or parameter adaptation.

3. Methodology

3.1. Preliminary diagnostics and garch-family specifications

The study examines the 10-, 15-, and 20-year return samples using augmented Dickey–Fuller tests, return autocorrelation and partial autocorrelation functions, squared-return autocorrelations, and ARCH-LM tests. Standardized and squared standardized residuals are checked using Ljung-Box tests for serial correlation. Appendix B.1 reports preliminary diagnostic statistics and Appendix B.2 summarizes selected results from model checks.

Four fixed GARCH-family specifications are estimated: GARCH(1,1), EGARCH(1,1), FIGARCH, and component GARCH (CGARCH). All four models use an estimated constant mean and variance-normalized Student-t innovations. Appendix A.1 provides detailed model specifications. The exploratory analysis considers likelihood, AIC, BIC, parameter estimates, and residual diagnostics. Information criteria are compared within compatible asset, sample, and likelihood settings. Appendix A.6 reports differences in historical sample endpoints and model coverage.

3.2. Historical regime analysis

Following Haas et al. [7] and Ardia et al. [10], a two-state Markov-switching GARCH (MS-GARCH) model is estimated using centered returns with regime-specific sGARCH(1,1) dynamics, Student-t innovations, and first-order Markov transitions. Supplementary three-state models are compared using AIC and BIC when both specifications yield valid estimates.

Labels are assigned after estimation based on the return dispersion of observations classified into each state. The analysis reports hard sample shares, stationary probabilities, and expected durations. Smoothed probabilities are used to describe historical regimes, while the observable forecasting states used in the out-of-sample analysis are constructed separately from VIX and asset-volatility signals. Appendix A.2 provides definitions of these measures, while Appendix B.4 provides a brief two- versus three-state model-count comparison.

3.3. Recursive forecasts, evaluation, and observable states

For each of the five investable assets, the study uses the first 1,000 returns as the initial estimation window. The window is then expanded recursively over time, with parameters refitted every 20 trading days and the conditional-variance filter updated daily. Therefore, a "fixed model" refers to a fixed specification rather than fixed coefficients. VIX acts as a market-conditional signal instead of serving as an additional forecast target.

DOI: 10.54254/2754-1169/2026.38677

The primary volatility proxy is the next-day squared percentage log return. Forecast accuracy using QLIKE is evaluated as:

$$L_{i,m,t} = \log(\widehat{h}_{i,m,t}) + \frac{v_{i,t}}{\widehat{h}_{i,m,t}}, \quad v_{i,t} = r_{i,t}^2 \tag{1}$$

$\widehat{h}_{i,m,t}$ is model m 's one-step-ahead variance forecast for asset i on date t, formed using information available through $t-1$, $v_{i,t}$ is the squared percentage log return on date t. QLIKE is used to evaluate forecasts with imperfect volatility proxies [11]. Lower values indicate better forecasting performance. Pairwise comparisons use five-lag Newey–West heteroskedasticity and autocorrelation consistent (HAC) standard errors [12]. Appendix A.3 provides proxy scaling and implementation details.

Two observable volatility states are constructed with information available before each forecast is evaluated: Common Stress is defined when the expanding historical percentile of VIX is over 0.80; Asset Stress is defined when the expanding percentile of an asset's 21-day rolling volatility is over 0.80; remaining valid observations are classified as Normal. The percentile calculations begin at the 252nd observation of the master sample and, at each date, are calculated using all available non-missing values of the relevant signal observed up to that date. Both state variables are then lagged by one trading day before being used in the forecast evaluation. Combining the Common and Asset states produces four Dual-State categories.

To test whether the relative model performance varies across states, daily loss differentials between two models are regressed on the lagged stress indicator:

$$d_{i,t}^{m,n} = L_{i,m,t} - L_{i,n,t} = \gamma_0 + \gamma_1 \mathrm{Stress}_{t-1} + u_t \tag{2}$$

The intercept refers to the Normal-state differential, the slope measures its change during Stress, and the sum is the stress-state differential. A statistically significant stress coefficient with a sign change indicates a ranking reversal; a significant change without reversal indicates a stronger or weaker existing advantage. This does not require each state-specific differential to be separately significant. The reported p-values are unadjusted for multiple testing.

Observable states and past forecast losses are then used to select among the four existing GARCH-family forecasts. The routing rules select among the four existing GARCH-family forecasts without estimating a new volatility process. Rolling-Best selects the model with the lowest mean QLIKE over the preceding 63 trading days, while Common-State, Asset-State, and Dual-State routers use historical performance within the corresponding state. Dual-EWMA gives greater weight to recent performance within the same Dual State. All decisions use only information available before the forecast outcome is observed.

Routing performance is compared with the fixed GARCH-family model having the lowest average QLIKE over the same evaluation period. Because this benchmark is selected retrospectively, it provides a relatively strong comparison for evaluating whether ex-ante routing adds forecasting value. Appendix A.3 provides volatility-proxy construction, HAC implementation, and routing details.

DOI: 10.54254/2754-1169/2026.38677

3.4. Adaptation extensions

3.4.1. GAS and GAS tree

Standard GAS uses a Student-t likelihood and a unit-scaled score recursion for log variance [8]. The shallow GAS Tree follows the parameter-localization framework of Patton and Simsek [9], keeping a single chronological variance filter while allowing the variance intercept, score-response coefficient, and autoregressive coefficient to differ between two leaves. The conditional mean and degrees of freedom remain the same across both leaves.

Possible splits variables use lagged expanding VIX or asset-volatility percentiles, with thresholds of 0.60, 0.70, 0.80, and 0.90. A Tree is activated only if it has lower AIC than the comparable Standard GAS model. Once selected, the split variable and threshold are fixed, while model coefficients refitted every 20 trading days on an expanding sample. Appendix A.4 documents estimation and numerical implementation.

Activated Trees are compared with Standard GAS and fixed-model benchmarks on matched dates. The same GAS and GAS Tree forecasts are then rescored using the Garman-Klass variance proxy [13] as an additional robustness check, without split reselection or model retuning. GLD does not activate a Tree. Therefore, Standard GAS and the four fixed specifications are compared separately. Appendix Table B7 reports the initial Tree-selection results, while Table 6 reports the volatility-proxy robustness results.

3.4.2. Machine-learning-assisted model selection

The machine-learning analysis uses XGBoost [14] and Random Forest [15] to predict the future relative losses of the four fixed GARCH-family specifications rather than volatility itself. Predictors include lagged return characteristics, rolling volatility, asset and market volatility conditions, and asset identity. All predictors use information available before the target forecast period.

For each candidate specification, the target is the average of its one-step-ahead QLIKE loss over the next five dates minus the average loss across the four candidate models over the same dates:

$$Y^{rel}_{i,m,t} = \frac{1}{5}\sum_{j=0}^{4} L_{i,m,t+j} - \frac{1}{|\mathcal{M}|}\sum_{k\in\mathcal{M}}\left(\frac{1}{5}\sum_{j=0}^{4} L_{i,k,t+j}\right) \quad (3)$$

where $L_{i,m,t+j}$ denotes the QLIKE loss for asset i under model m on date $t+j$, and $\mathcal{M} = \{\mathrm{GARCH}, \mathrm{EGARCH}, \mathrm{FIGARCH}, \mathrm{CGARCH}\}$ is the fixed-model candidate set. The target removes the average loss shared by the candidates to isolate relative model performance. At each date, the model with the lowest predicted relative loss is selected and evaluated using its realized five-date average QLIKE. This averages five one-step-ahead losses, not a single five-day volatility forecast.

The final models use the conditions-only information set; historical model-loss features examined during specification development are excluded from the final regressions. Training targets end in 2019, validation covers 2020–2022, and testing begins in 2023. Complete five-date target windows are restricted to their assigned chronological subsamples, with overlapping boundary observations removed. XGBoost uses the validation sample for early stopping, while Random Forest uses the training sample for estimation and is then applied directly to the test period. Appendix A.5 provides details on feature construction, encoding, hyperparameters, and sample-boundary treatment.

The study examined several machine-learning specifications while using results from the same test period as a reference. Therefore, the study treats the final machine-learning results as

DOI: 10.54254/2754-1169/2026.38677

exploratory rather than as results from a completely untouched holdout sample.

Table 1. Forecasting samples and design settings

Exercise	Sample per asset	Core setting
Fixed GARCH forecasts	4,152 dates	1,000 initial returns; 20-day refit; daily 1-step forecast
Routing comparison	4,030 dates	63-day Rolling-Best; 20 observations per state; EWMA 0.95
Selected-Tree comparison	3,901 dates	Depth 1; locked initial split; matched forecast dates
ML complete feature panel	4,089 rows	Lagged conditions merged with fixed-model losses
ML training/validation/test	2,432/752/893	Chronological split; eight boundary rows purged

Notes: Detailed coverage and exclusion accounting are reported in Appendix A.6.

4. Data summary

4.1. Asset selection and sample construction

Daily market data were obtained from Yahoo Finance for five exchange-traded funds: SPDR S&P 500 ETF Trust (SPY) for U.S. equities; iShares Core U.S. Aggregate Bond ETF (AGG) for U.S. investment-grade fixed income; Invesco DB Commodity Index Tracking Fund (DBC) for broad commodities; SPDR Gold Shares (GLD) for gold, and iShares U.S. Real Estate ETF (IYR) for listed real estate. The CBOE Volatility Index (VIX) is included as an observable indicator of overall market stress when constructing market-wide volatility states.

Adjusted closing prices are converted to continuously compounded percentage returns:

$$r_{i,t} = 100\log\left(\frac{P_{i,t}^{Adj}}{P_{i,t-1}^{Adj}}\right), \tag{4}$$

where $P_{i,t}^{Adj}$ denotes the adjusted closing price of asset i on trading day t. The series are aligned by trading date to ensure consistency across assets. The longest common sample begins in February 2006, when DBC becomes available.

The analysis considers 10-, 15-, and 20-year historical windows, containing 2,658, 3,916, and 5,152 common daily return observations, respectively. These windows are used to assess the stability of volatility characteristics across historical samples, while the longest common sample provides the basis for the main out-of-sample forecasting analysis.

4.2. Descriptive statistics

Table 2 summarizes the return distributions across the three sample windows. Unconditional volatility differs across assets, with AGG consistently having the lowest daily standard deviation and IYR showing the highest, especially in the 20-year sample. VIX log changes also show much higher volatility than the five investable asset returns.

The return distributions also differ in skewness and tail behavior. All five assets are negatively skewed and have high kurtosis across the three windows, whereas VIX remains positively skewed. AGG is notable for combining relatively low overall volatility with strong negative skewness and heavy tails, especially in the 20-year sample.

DOI: 10.54254/2754-1169/2026.38677

IYR shows greater sensitivity, with its daily standard deviation rising from approximately 1.28 in the 10-year sample to 1.78 in the 20-year sample. Yet, the sample windows do not have exactly the same endpoints, reflecting both sample length and historical coverage.

Table 2. Cross-window descriptive statistics

Asset	Window	Mean	Standard Deviation	Skewness	Kurtosis
VIX	10-Year	-0.0083	8.0064	1.2685	11.0370
	15-Year	-0.0015	7.8568	1.1217	9.7264
	20-Year	0.0047	7.6790	1.0297	9.3800
AGG	10-Year	0.0071	0.3350	-1.1440	22.8567
	15-Year	0.0088	0.3007	-1.1250	24.2909
	20-Year	0.0119	0.3344	-2.1167	57.9875
SPY	10-Year	0.0558	1.1240	-0.5944	18.0084
	15-Year	0.0522	1.0766	-0.5811	16.1459
	20-Year	0.0417	1.2156	-0.3033	17.3108
DBC	10-Year	0.0363	1.1356	-0.7155	7.1599
	15-Year	0.0058	1.0897	-0.6475	7.0893
	20-Year	0.0083	1.2196	-0.4813	6.1822
GLD	10-Year	0.0483	1.0269	-0.6705	10.9177
	15-Year	0.0253	1.0542	-0.7058	10.1694
	20-Year	0.0365	1.1633	-0.4651	9.6778
IYR	10-Year	0.0249	1.2803	-1.4414	26.6547
	15-Year	0.0286	1.2264	-1.1944	23.9854
	20-Year	0.0277	1.7780	-0.6921	22.8344

Notes: Mean and standard deviation are reported in percentage-return units. VIX statistics refer to percentage log changes in the VIX index rather than the VIX level itself. Kurtosis is reported as Pearson kurtosis, for which the Gaussian benchmark is 3.

5. Empirical results

5.1. Cross-asset structure and historical regimes

All ADF tests reject the unit-root null hypothesis, while ARCH-LM tests show ARCH effects across all asset-window combinations. Some serial correlation remains in raw returns, especially for SPY, AGG, IYR, and VIX, whereas the Ljung–Box tests for DBC and GLD do not reject the null of no serial correlation at the 5% level. Appendix Table B1 reports the detailed diagnostic statistics.

The estimated volatility patterns differ across assets. SPY shows negative asymmetry across all three historical windows, with EGARCH signed-shock coefficients of similar magnitude. DBC shows persistent fractional dynamics, while IYR combines negative asymmetry with greater sensitivity to the historical window. These estimates describe patterns in the observed data rather than causal economic mechanisms. Appendix Table B3 reports selected cross-window coefficient estimates.

The results for AGG and GLD are less stable across models and sample windows. FIGARCH fails to produce a valid estimate for AGG in the short window, while its longer-window estimates are sensitive to parameter boundaries. In the 20-year sample, none of the AGG specifications fully

DOI: 10.54254/2754-1169/2026.38677

removes the remaining ARCH effects. GLD's historical model assessment also varies across windows and is subject to residual-adequacy qualifications. These findings suggest that in-sample results alone may not provide enough evidence to assign a single historical volatility structure to each asset.

The MS-GARCH results provide additional evidence of regime heterogeneity. AGG experiences relatively infrequent but persistent high-volatility periods in several windows, while DBC and GLD show greater variation in regime occupancy across samples. IYR has a relatively small share of observations classified as high volatility in the longer sample. State-label sensitivity for SPY and VIX further cautions against interpreting estimated high-volatility shares mechanically as the proportion of time spent in economic crises. Appendix Tables B2-B3 report historical model assessments and selected cross-window coefficients. Appendix Table B4 reports historical regime occupancy and duration results, while Appendix B.4 provides additional comparison of the number of states.

5.2. Fixed-model performance and state dependence

Table 3 shows that overall QLIKE is lowest for EGARCH on SPY, GARCH on AGG, and CGARCH on DBC, GLD, and IYR. However, the differences between models are not uniformly large or statistically significant. Therefore, numerical leadership is distinguished from formal evidence of forecasting superiority.

Table 3. Numerical fixed-model winners by asset and evaluation setting

Asset	Overall	QLIKE	Common Normal	Common Stress	Asset Normal	Asset Stress
SPY	EGARCH	0.713	EGARCH	FIGARCH	EGARCH	FIGARCH
AGG	GARCH	-1.822	EGARCH	CGARCH	GARCH	CGARCH
DBC	CGARCH	1.050	EGARCH	CGARCH	CGARCH	CGARCH
GLD	CGARCH	0.945	CGARCH	EGARCH	GARCH	EGARCH
IYR	CGARCH	1.045	CGARCH	EGARCH	CGARCH	FIGARCH

Notes: Primary squared-return target; 4,152 overall forecast dates per asset, with subsets for state-specific rankings. QLIKE is the overall numerical winner's loss. A winner label alone does not establish statistical superiority.

Under Common Normal conditions, EGARCH significantly outperforms FIGARCH, CGARCH, and GARCH for SPY. DBC shows significant advantages for CGARCH over several other models, with some of these differences becoming much stronger during Asset Stress. AGG also shows evidence that the relative performance of GARCH and CGARCH changes with its asset-specific volatility state. In contrast, the overall QLIKE values of the competing models remain relatively close for GLD and IYR. Appendix Table B5 reports the overall loss levels and selected pairwise comparisons.

Common-market and asset-specific stress are not equivalent. Figure 1 shows the strongest overlap between Common and Asset Stress for SPY and the least for GLD.

DOI: 10.54254/2754-1169/2026.38677

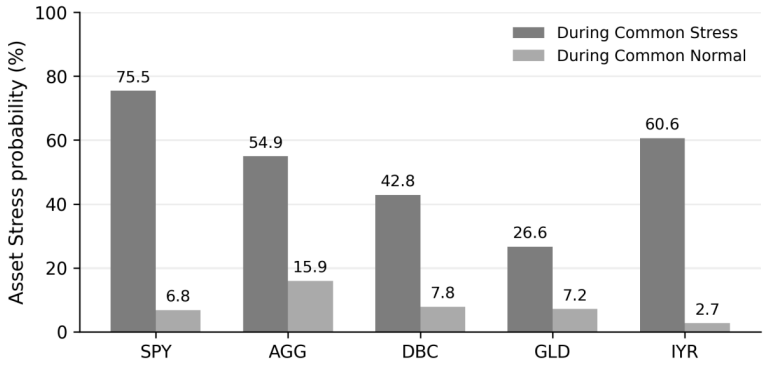


Figure 1. Asset-specific stress conditional on the common-market volatility state

Notes: Bars show Asset Stress percentages conditional on Common Stress and Common Normal. Each asset's sample requires both states to be observed. Classifications are contemporaneous here but lagged one day in forecasting. These descriptive samples differ from the out-of-sample forecasting panels.

Table 4 reports selected model pairs with significant state-dependent shifts in their QLIKE differentials. AGG exhibits a GARCH–CGARCH ranking reversal across its Asset state, while GLD shows a similar reversal between GARCH and FIGARCH across the Common state. DBC primarily exhibits amplification rather than reversal: the relative advantages among several specifications become significantly larger under stress. SPY and IYR show no significant direct changes in the reported full-sample regressions despite changes in some numerical state-specific winners.

Table 4. Reported state-dependent shifts in pairwise relative QLIKE

Asset	State	Model pair	Reported pattern	p-value
AGG	Asset	GARCH/CGARCH	Winner reversal	0.0306
DBC	Asset	FIGARCH/CGARCH	CGARCH amplification	3.06e-06
DBC	Asset	GARCH/CGARCH	CGARCH amplification	0.000292
DBC	Asset	GARCH/FIGARCH	GARCH amplification	0.0181
DBC	Common	GARCH/CGARCH	CGARCH amplification	0.0114
DBC	Common	FIGARCH/CGARCH	CGARCH amplification	0.0341
GLD	Common	GARCH/FIGARCH	Winner reversal	0.0291

Notes: p-values test the Stress coefficient in Equation (2) using five-lag Newey–West HAC inference and are unadjusted for multiple testing. A reversal combines a significant state shift with an estimated sign change; individual state differentials need not be significant. Amplification is a significant change in advantage magnitude without sign reversal.

5.3. Adaptive forecasting

Relative model performance can change across volatility states. The next question is whether this information helps a forecaster choose a model before its forecast outcome is known. Three approaches are considered: forecast routing, parameter adaptation through the GAS Tree, and machine-learning-assisted model selection.

DOI: 10.54254/2754-1169/2026.38677

5.3.1. Forecast routing

The routing methods produce some small numerical improvements over the fixed-model benchmarks, but most of these improvements are not statistically significant. Different routing rules achieve the lowest numerical loss across assets, with Dual-EWMA performing particularly well for AGG and DBC. For each asset, at least one router has lower mean QLIKE than the four fixed GARCH-family specifications. However, the statistical tests do not show that these improvements are reliable across the assets.

Table 5. Routing performance relative to fixed benchmarks

Asset	Fixed benchmark	Hard Dual p-value	Dual-EWMA − Fixed Mean Loss Differential	EWMA p-value
SPY	EGARCH	0.716	-0.00769	0.664
AGG	GARCH	0.574	-0.00492	0.2446
DBC	CGARCH	0.811	-0.00266	0.365
GLD	CGARCH	0.187	-0.00027	0.963
IYR	CGARCH	0.791	+0.00142	0.687

Notes: Negative Dual-EWMA-minus-fixed QLIKE differentials favor Dual-EWMA. The p-values compare Hard Dual and Dual-EWMA separately with the fixed benchmark, not the numerically best router. Benchmarks are selected retrospectively on the matched panel. Appendix Table B6 reports all routing mean losses.

The Dual-EWMA for AGG has lower mean QLIKE than Hard Dual routing but does not significantly outperform the fixed GARCH benchmark. DBC shows a similar pattern: Dual-EWMA has the lowest numerical mean QLIKE among the evaluated routing and fixed strategies without significantly outperforming fixed CGARCH.

The results during Stress period lead to a similar conclusion. Most reported Dual-EWMA loss differentials favor routing numerically, but none is statistically significant at the 5% level relative to the corresponding fixed benchmark. Appendix Table B6 reports matched-panel losses and Stress-subset comparisons.

5.3.2. GAS tree parameter adaptation

Under the prespecified AIC rule, Trees are activated for AGG, DBC, IYR, and SPY. AGG selects a VIX-percentile cutoff of 0.90, while DBC, IYR, and SPY select 0.60. GLD remains unsplit. However, BIC favors the Standard GAS model without a split for all five assets. Therefore, the support for parameter localization depends on how the additional model complexity is evaluated.

Table 6. GAS tree performance across volatility targets

Panel A. Squared-Return Evaluation					
Asset	Locked VIX cutoff	Tree QLIKE	Tree vs GAS p-value	Selected Fixed Comparator	Tree vs fixed p-value
AGG	0.90	-1.835	0.0208	CGARCH	0.0198
DBC	0.60	1.030	0.00731	CGARCH (1.028)	0.596
IYR	0.60	1.001	0.000470	CGARCH (0.998)	0.540
SPY	0.60	0.712	1.07e-07	EGARCH (0.697)	0.521
Panel B. GK QLIKE on 3,901 matched dates					
Asset	GAS Tree	Standard GAS	Tree-GAS	HAC p-value	

DOI: 10.54254/2754-1169/2026.38677

Table 6. (continued)

AGG	−2.3324	−2.4065	+0.0741	2.50e−09
DBC	0.5076	0.4046	+0.1030	3.04e−109
IYR	0.7679	0.7413	+0.0266	7.88e−06
SPY	0.3612	0.2743	+0.0869	1.35e−13

Notes: Both panels use the same 3,901 matched forecast dates. Panel A evaluates forecasts using squared returns, while Panel B evaluates the same forecasts using the Garman–Klass volatility proxy. Lower QLIKE indicates better performance. In Panel B, positive Tree-minus-GAS differentials favor Standard GAS. Reported p-values use five-lag Newey–West HAC inference.

Under the squared-return target, the GAS Tree significantly improves Standard GAS for all four activated assets. For AGG, the Tree significantly outperforms CGARCH overall. For DBC, IYR, and SPY, the Tree improves on Standard GAS, but it does not significantly outperform the strongest fixed GARCH-family model.

The ranking changes when the same forecasts are evaluated using the Garman–Klass proxy: Standard GAS significantly outperforms GAS Tree for all four activated assets. Therefore, the evidence for parameter adaptation is strongly target-sensitive: localization improves the GAS baseline under squared-return evaluation, but this ranking reverses under Garman–Klass. Appendix Table B7 reports initial Tree-selection results and AIC/BIC comparisons.

5.3.3. Machine-learning-assisted model selection

Developmental absolute-loss regressions achieve much higher correlations between predicted and realized losses, whereas these correlations fall sharply when the average loss shared across the four models is removed from the target. This suggests that the machine-learning models can more easily identify common forecasting difficulty than differences in model suitability.

The forecast-selection results provide little evidence of gains over the fixed benchmarks. Among the reported XGBoost selectors, DBC has slightly lower mean QLIKE than fixed CGARCH, but the difference is not significant. The GLD and IYR selectors have higher mean losses than their benchmarks. Random Forest has higher mean loss for all five assets; the difference is significant at 5% for AGG and at 10%, but not 5%, for DBC. Appendix Table B8 reports prediction-correlation diagnostics. These results remain exploratory because target and feature development used the same test period.

Table 7. Machine-learning selector performance relative to fixed benchmarks

Algorithm / scope	Asset	Benchmark	Selector − fixed	HAC SE	p-value
XGBoost / asset-specific	DBC	CGARCH	−0.000114	0.003891	0.9767
XGBoost / asset-specific	GLD	EGARCH	+0.010192	0.010995	0.3540
XGBoost / asset-specific	IYR	EGARCH	+0.010148	0.012179	0.4047
Random Forest / pooled	AGG	EGARCH	+0.007427**	0.003311	0.0249
Random Forest / pooled	DBC	CGARCH	+0.009327*	0.005633	0.0978
Random Forest / pooled	GLD	EGARCH	+0.007800	0.007006	0.2656
Random Forest / pooled	IYR	EGARCH	+0.008867	0.010626	0.4040
Random Forest / pooled	SPY	EGARCH	+0.002860	0.006373	0.6536

DOI: 10.54254/2754-1169/2026.38677

Notes: Negative selector-minus-fixed QLIKE differentials favor the selector. Inference uses five-lag Newey–West HAC standard errors. *, **, and *** indicate significance at 10%, 5%, and 1%. Comparisons use the same 893 test observations per asset. Benchmarks are selected within each machine-learning evaluation framework and may differ from those on the 4,152- or 4,030-date panels.

5.4. Integrated asset-class interpretation

Table 8 summarizes the asset-specific evidence, model implications, and limits of the forecasting comparisons.

Table 8. Portfolio-oriented interpretation within the examined framework

Asset	Main evidence	Relative model implication	Boundary
SPY	Stable negative asymmetry; repeated EGARCH preference	EGARCH shows recurring support, particularly under Common Normal conditions.	Overall lead is not always significant; Stress rankings depend on sample
AGG	Low ordinary volatility, extreme tails, changing fixed ranks	GAS Tree offers selective gains under the squared-return target; fixed-family rankings vary	GK reverses the Tree–GAS ranking.
DBC	Persistent fractional/component dynamics; amplified Stress gaps	CGARCH is the main full-panel fixed-model benchmark, with selected relative advantages increasing under Stress	Routing and GAS Tree adaptation do not reliably improve on the strongest fixed benchmark
GLD	Window-sensitive historical model evidence; persistent regime behavior; weak common-state overlap	Several fixed specifications remain plausible; no activated Tree	Small ranking gaps; no reliable ML selection gain
IYR	Negative asymmetry, persistence, window-sensitive historical regime behavior	Asymmetric/component specifications remain plausible	Tree improves on Standard GAS but does not significantly outperform the selected CGARCH comparator.

Notes: The table synthesizes evidence from the historical, one-step-ahead, state-dependent, routing, GAS Tree, and machine-learning analyses.

6. Discussion, limitations, and future research

The results show why model suitability and forecast improvement need to be considered separately. SPY shows recurring support for EGARCH, while AGG's fixed-model rankings change with its

DOI: 10.54254/2754-1169/2026.38677

volatility state. These differences help describe the assets, but they do not by themselves tell a forecaster which model to select at the next date.

The adaptive forecasting results support this conclusion. In the forecasting exercises, routing and machine-learning selection generally fail to outperform the fixed benchmarks reliably. GAS Tree gives more encouraging results for AGG under squared-return evaluation, but its advantage over Standard GAS reverses under the Garman–Klass proxy. The question is therefore not simply whether a model uses state information, but how that information is used and which volatility target it is evaluated against.

Several limitations define the scope of these findings. First, model rankings apply only to the specifications and forecasting design examined in this study. Each asset class is represented by a single ETF and conditional variances are modeled separately. Thus, the study does not evaluate portfolio diversification, cross-asset correlations, or optimal allocation.

Second, volatility is not directly observed, so the results depend on how it is measured. Squared close-to-close returns provide a noisy proxy that includes overnight movements, whereas the implemented Garman–Klass measure captures within-day variation. The reversal of the GAS Tree ranking therefore reflects sensitivity to both volatility measurement and target definition. Future research could use high-frequency realized volatility or other range-based measures to examine whether the rankings persist.

Other limitations arise from the empirical design. The historical windows differ in both length and endpoints. Some FIGARCH and higher-state MS-GARCH estimations face boundary or convergence issues, and state-dependent tests are unadjusted for multiple comparisons. Adaptive methods are evaluated against fixed benchmarks selected retrospectively on each evaluation sample, not choices available in advance. These limitations mean that isolated significant results should be interpreted cautiously.

Finally, because machine-learning specifications were developed using the same test period, future work should pre-specify targets, features, and selection procedures before testing on an untouched sample. Extensions could include multiple instruments per asset class, multivariate correlation dynamics, and forecast combinations. Small loss differences among candidate models motivate comparing combinations with hard selection.

7. Conclusions

Across the five assets, volatility-model suitability varies with asset characteristics, observable states, and the evaluation target. SPY shows recurring support for EGARCH; AGG combines low overall volatility with heavy tails and state sensitivity; DBC exhibits fractional and component-type persistence; and GLD and IYR provide less clear support for one specification.

Selected model pairs exhibit significant state-dependent ranking reversals or changes in existing advantages. However, numerical rank changes alone do not establish statistical shifts or gains from switching.

Routing and machine-learning selectors provide no consistent improvements over strong fixed benchmarks. GAS Tree offers selective gains, particularly for AGG, but its advantage over Standard GAS reverses under the Garman–Klass proxy.

Within this study, more flexible methods are not consistently more accurate. Model choice still depends on the asset, market state, and volatility measure used for evaluation.

DOI: 10.54254/2754-1169/2026.38677

References

[1] Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics, 31*(3), 307–327.

[2] Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica: Journal of the Econometric Society*, 987–1007.

[3] Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica: Journal of the Econometric Society*, 347–370.

[4] Baillie, R. T., Bollerslev, T., & Mikkelsen, H. O. (1996). Fractionally integrated generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics, 74*(1), 3–30.

[5] Lee, G. G. J., & Engle, R. F. (1993). A permanent and transitory component model of stock return volatility. *Available at SSRN 5848*.

[6] Hansen, P. R., & Lunde, A. (2005). A forecast comparison of volatility models: Does anything beat a GARCH (1, 1)? *Journal of Applied Econometrics, 20*(7), 873–889.

[7] Haas, M., Mittnik, S., & Paolella, M. S. (2004). A new approach to Markov-switching GARCH models. *Journal of Financial Econometrics, 2*(4), 493–530.

[8] Creal, D., Koopman, S. J., & Lucas, A. (2013). Generalized autoregressive score models with applications. *Journal of Applied Econometrics, 28*(5), 777–795.

[9] Patton, A. J., & Simsek, Y. (2026). Generalized autoregressive score trees and forests. *Journal of Business & Economic Statistics*, just-accepted, 1–25.

[10] Ardia, D., et al. (2019). Markov-switching GARCH models in R: The MSGARCH package. *Journal of Statistical Software, 91*, 1–38.

[11] Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics, 160*(1), 246–256.

[12] Newey, W. K., & West, K. D. (1986). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix.

[13] Garman, M. B., & Klass, M. J. (1980). On the estimation of security price volatilities from historical data. *Journal of Business*, 67–78.

[14] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

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

DOI: 10.54254/2754-1169/2026.38677

Appendix A. Model Specifications and Implementation

A.1Fixed-Model Specifications

The four fixed GARCH-family specifications use a common constant-mean specification with standardized Student-t innovations. For each asset, returns are represented as

$$r_t = \mu + \varepsilon_t, \quad \varepsilon_t = \sqrt{h_t} z_t \tag{A1}$$

where μ is the conditional mean, h_t is the conditional variance, and z_t follows a standardized Student-t distribution with estimated degrees of freedom. Asset subscripts are suppressed for readability.

GARCH(1,1)

$$h_t = \omega + \alpha \varepsilon_{t-1}^2 + \beta h_{t-1} \tag{A2}$$

where $\omega > 0$, $\alpha, \beta \geq 0$. Parameter α and β capture responses to squared innovations and variance persistence, respectively. Covariance stationarity requires $\alpha + \beta < 1$.

EGARCH(1,1)

$$\log h_t = \omega + \beta \log h_{t-1} + \alpha(|z_{t-1}| - E|z_{t-1}|) + \gamma z_{t-1} \tag{A3}$$

α captures innovation-magnitude effects and γ captures signed asymmetry. A negative γ implies a stronger variance response to negative than positive innovations of equal magnitude, while a positive value implies the reverse.

FIGARCH

Using the lag operator (L), the FIGARCH process is represented as:

$$\phi(L)(1-L)^d \varepsilon_t^2 = \omega + [1 - \beta(L)](\varepsilon_t^2 - h_t) \tag{A4}$$

where d governs fractional persistence and $\phi(L)$ and $\beta(L)$ are lag polynomials. Estimates near or beyond conventional parameter bounds are treated as boundary-sensitive rather than as evidence of stronger long memory.

Component GARCH

Let q_t denote the long-run variance component. Conditional variance is represented as:

$$h_t = q_t + \alpha(\varepsilon_{t-1}^2 - q_{t-1}) + \beta(h_{t-1} - q_{t-1}) \tag{A5}$$

with the long-run component evolving according to

$$q_t = \omega + \rho q_{t-1} + \varphi\left(\varepsilon_{t-1}^2 - h_{t-1}\right) \tag{A6}$$

DOI: 10.54254/2754-1169/2026.38677

ρ and φ govern long-run persistence and responses to variance innovations, while α and β govern short-run deviations from the long-run component.

Estimates exhibiting convergence or boundary problems are retained with the corresponding estimation qualification rather than re-parameterized.

A.2Historical regimes and observable states

Historical MS-GARCH regimes are distinct from the observable forecasting states. The historical model uses demeaned returns, regime-specific sGARCH(1,1) dynamics, Student-t innovations, and a first-order Markov transition matrix P. Observations are classified by their highest smoothed probability; states are then labeled Low or High Volatility according to the standard deviation of assigned returns. Hard-classification shares, stationary probabilities, and transition-implied durations are distinct quantities. Three-state specifications are supplementary and considered only when valid estimates are available.

Observable forecasting states

The market signal is the VIX level; the asset-specific signal is the right-aligned 21-day rolling standard deviation of daily returns. Incomplete rolling windows remain missing and are not imputed.

For a valid signal x_t, its expanding empirical percentile is defined as

$$\widehat{F}_t(x_t) = \frac{1}{N_t} \sum_{s \le t,\, x_s \text{ valid}} 1\{x_s \le x_t\} \tag{A7}$$

where N_t is the number of valid, nonmissing observations available through date t. Ties are handled using the weak inequality $x_s \le x_t$.

Common Stress and Asset Stress are defined when the corresponding expanding signal percentile strictly exceeds 0.80; other valid observations are Normal. Forecasting states are lagged by one trading day. Combining the Common and Asset labels gives four Dual States: Normal/Normal, Normal/Stress, Stress/Normal, and Stress/Stress. Descriptive overlap statistics use separate contemporaneous samples rather than the out-of-sample forecasting panels.

A.3Forecast evaluation and routing implementation

Volatility proxies

QLIKE, defined in Equation (1) of the main text, is evaluated using raw squared percentage returns, without demeaning. For robustness, the same forecasts are evaluated using the Garman–Klass [13] range-based variance estimator:

$$GK_t = \frac{1}{2}\left[\log\left(\frac{H_t}{L_t}\right)\right]^2 - (2\log 2 - 1)\left[\log\left(\frac{C_t}{O_t}\right)\right]^2 \tag{A8}$$

Open, High, and Low prices are scaled by the Adjusted Close/Close ratio, with Adjusted Close used as the closing price. The GK estimate is multiplied by 100^2 to match percentage-return variance units. Existing GAS and GAS Tree forecasts are rescored without re-estimation, split reselection, or retuning. Unlike squared close-to-close returns, the implemented GK proxy measures within-day OHLC variation and excludes overnight price changes.

DOI: 10.54254/2754-1169/2026.38677

Forecast-comparison inference

For models m and j, pairwise comparisons use the QLIKE loss differential

$$d_t = L_{m,t} - L_{j,t} \tag{A9}$$

Negative mean differentials favor model m. Inference uses five-lag Newey–West HAC standard errors, with no pre-whitening and with finite-sample adjustment. Tests are two-sided and unadjusted for multiple testing. State-dependent regressions follow Equation (2) in the main text. For state-specific subsets, HAC lags follow the retained observation sequence rather than necessarily consecutive trading days.

Routing implementation

All routers select among the four precomputed forecasts using only information available before realization. Rolling-Best uses the lowest mean QLIKE over the preceding 63 trading observations. Common-State, Asset-State, and Dual-State use cumulative mean QLIKE by asset, model, and forecast-origin state, requiring at least 20 prior observations in that state.

Dual-EWMA replaces the cumulative state-specific mean with an exponentially weighted loss score. Let $\tau_{j(s)}$ denote the date of the j -th occurrence of Dual State (s), and let $E_{m,s,j}$ denote the score available for model m at that occurrence. The score is initialized from the first realized loss observed in the state and subsequently updated as:

$$E_{m,s,1} = L_{m,\tau_1(s)}$$

$$E_{m,s,j} = \lambda E_{m,s,j-1} + (1-\lambda) L_{m,\tau_{j-1}(s)}, \quad j \geq 2 \tag{A10}$$

The decay parameter is $\lambda = 0.95$. Scores update only when the same state occurs, with the current realized loss added after the selection decision. Initialization and eligibility are separate: selection requires at least 20 prior observations in that state. The fixed benchmark is selected retrospectively as the lowest-mean-QLIKE specification on the same matched panel, whereas routing decisions remain ex-ante.

A.4GAS and GAS tree estimation

Standard GAS

Returns are modeled using a constant conditional mean and a standardized Student-t conditional distribution,

$$r_t = \mu + \varepsilon_t, \quad \varepsilon_t = \sqrt{h_t} z_t, \tag{A11}$$

where h_t is the conditional variance and z_t follows a standardized Student-t distribution with degrees-of-freedom parameter $\upsilon > 2$. Let $f_t = \log h_t$, so that $h_t = \exp(f_t)$. The conditional log-likelihood contribution is denoted by

DOI: 10.54254/2754-1169/2026.38677

$$l_t = -\frac{1}{2} f_t - \frac{v+1}{2} \log\left(1 + \frac{\varepsilon_t^2}{(v-2)\exp(f_t)}\right) + c(v) \tag{A12}$$

Where

$$c(v) = \log\Gamma\left(\frac{v+1}{2}\right) - \log\Gamma\left(\frac{v}{2}\right) - \frac{1}{2}\log[(v-2)\pi] \tag{A13}$$

The implementation uses unit scaling, so the updating term is the score with respect to log variance $s_t = \nabla_t = \frac{\partial l_t}{\partial f_t}$. The Standard GAS recursion is

$$f_{t+1} = \omega + \beta f_t + \alpha s_t \tag{A14}$$

GAS tree localization

The GAS Tree uses one chronological variance filter with leaf-specific coefficients. Let x_t denote the partition signal for date t, based on information through $t-1$, and let c denote the selected threshold. Leaf assignment is defined as:

$$R_t = \begin{cases} L, & x_t \le c, \\ R, & x_t > c. \end{cases} \tag{A15}$$

Conditional on the assigned leaf, the variance recursion becomes

$$f_t = \omega_{R_t} + \beta_{R_t} f_{t-1} + \alpha_{R_t} s_{t-1} \tag{A16}$$

The conditional mean μ and degrees-of-freedom parameter v remain common across leaves.

Candidate partition signals are the lagged expanding VIX percentile and lagged expanding asset-volatility percentile. Candidate thresholds are $c \in \{0.60, 0.70, 0.80, 0.90\}$. Tree selection is conducted on the first 1,000 observations for which both candidate signals are available, and each admissible split must allocate at least 250 observations to each leaf.

The unsplit Standard GAS model contains five estimated parameters, $(\mu, \upsilon, \omega, \alpha, \beta)$, while the two-leaf Tree contains eight, $(\mu, \upsilon, \omega_L, \alpha_L, \beta_L, \omega_R, \alpha_R, \beta_R,)$. Candidate Trees are estimated by maximizing the Student-t conditional likelihood. Information criteria are calculated as:

$$AIC = -2\mathscr{L}(\hat{\theta}) + 2k \tag{A17}$$

$$BIC = -2\mathscr{L}(\hat{\theta}) + k\log T \tag{A18}$$

where $\mathscr{L}$ is the maximized log-likelihood, k is the number of estimated parameters, and T is the number of observations used for estimation.

A Tree is activated only if its AIC is lower than Standard GAS; BIC is reported but does not determine activation. The selected partition variable and threshold remain fixed, while coefficients are re-estimated every 20 trading days using an expanding sample.

DOI: 10.54254/2754-1169/2026.38677

Table A1. Parameter bounds for GAS and GAS tree estimation

Parameter	Lower	Upper
μ	−1	1
ν	2.1	30
ω_L, ω_R	−3	3
α_L, α_R	0.0001	2
β_L, β_R	0.10	0.999

Notes: The bounds apply to the conditional mean, Student-t degrees of freedom, and the log-variance intercept, score-response, and autoregressive parameters used in Standard GAS and the corresponding Tree leaves.

Numerical implementation

Each likelihood or filtering pass initializes log variance from the sample variance of its available return history. Optimization uses bounded L-BFGS-B, with nlminb as fallback. If both fail, the last valid coefficients are retained until the next scheduled refit. This occurs once, at AGG refit 190, affecting 20 forecasts; excluding them leaves the principal inference unchanged.

A.5Machine-Learning Implementation

Four candidate-specific regressions predict the future-five relative QLIKE defined in Equation (3) of the main text. The model with the lowest prediction is selected and evaluated using its realized future-five QLIKE. The target averages five one-step-ahead losses, not an aggregated five-day volatility realization.

Predictors and encoding

Predictors include lagged returns, absolute and squared returns, 21- and 63-day rolling volatility, expanding asset-volatility and VIX percentiles, Common and Asset States, and asset identity. All information is available at the forecast origin. Historical 21- and 63-day model-loss features are excluded, but the development-stage common feature panel is retained rather than expanded after their removal.

Pooled XGBoost uses dummy encoding for asset identity and states, with continuous predictors entered directly, and estimates one squared-error regression per candidate model. Asset-specific variants are also estimated for DBC, GLD, and IYR. Random Forest uses categorical factors for asset identity and states in ranger.

Table A2. Retained machine-learning settings

Setting	XGBoost	Random Forest
Regression objective	Squared error	Regression forest
Maximum tree depth	4	Not specified in summary
Learning rate	0.03	Not applicable
Rounds / trees	At most 1,000 rounds	500 trees
Observation subsampling	0.80	0.80
Feature sampling	0.80	mtry = 3

DOI: 10.54254/2754-1169/2026.38677

Table A2. (continued)

Minimum terminal-node size	Not specified in summary	10
Early stopping	40 validation rounds	Not applicable
Importance measure	Not specified in summary	Permutation
Random seed	123	123

Notes: Unspecified settings were not recorded in the retained implementation summary; no software defaults are assumed.

Training and evaluation

Both algorithms are fitted on training data and applied to the test sample. XGBoost uses validation data for early stopping but is not refitted on combined training and validation data. Training targets end in 2019, validation covers 2020–2022, and testing begins in 2023. Targets crossing subsample boundaries are excluded so that each five-date window lies wholly within one subsample. Because model development referenced the same test period, results are exploratory rather than based on a fully untouched holdout sample.

A.6Sample Accounting and Implementation Record

Table A3 supplements Table 1 with sample coverage and observation exclusions.

Table A3. Sample coverage and exclusion accounting

Panel	N per asset	Recorded coverage / accounting
Aligned price panel	5,153	6 Feb 2006–31 Jul 2026
Full return series	5,152	One initial price observation lost to differencing
Historical exploratory windows	2,658 / 3,916 / 5,152	10Y / 15Y / 20Y; historical endpoints are not perfectly aligned
Fixed-model OOS	4,152	28 Jan 2010–31 Jul 2026
ML complete feature panel	4,089	29 Apr 2010–31 Jul 2026
Complete future-five targets	4,085	Four terminal observations lack a complete $t, \ldots, t+4$ target
ML sample after boundary purging	4,077	2,432 training + 752 validation + 893 test
Boundary-overlap observations removed	8	Four observations removed at each of two sample boundaries

Notes: Routing and selected-Tree sample sizes are reported in Table 1. Historical windows use common observations within each window but have nonidentical endpoints, so comparisons reflect both sample length and historical coverage.

Historical coverage is not fully balanced: some archived SPY and VIX 20-year analyses include threshold/GJR specifications rather than uniform FIGARCH coverage. These results therefore describe structural and sample sensitivity, not a fully balanced cross-model forecasting comparison.

Implementation primarily uses R: rugarch for fixed models, MSGARCH for regimes, sandwich for HAC inference, xgboost and ranger for machine learning, and gasmodel plus custom likelihood code for GAS models. Historical analyses also use archived EViews and R output. Some historical package versions were not preserved.

DOI: 10.54254/2754-1169/2026.38677

Appendix B. Supplementary empirical evidence

B.1Preliminary diagnostic statistics

Table B1. ADF, ARCH-LM, and raw-return Ljung-Box Tests

Asset	Window	ADF statistic	ARCH-LM statistic	Ljung-Box statistic
SPY	10Y	-13.93901	797.0150***	213.55852***
SPY	15Y	-16.28049	1155.4366***	165.09549***
SPY	20Y	-17.67601	1456.5644***	94.19897***
AGG	10Y	-14.07412	1031.5586***	53.23081***
AGG	15Y	-15.92173	1502.3806***	61.94190***
AGG	20Y	-16.06681	670.1103***	87.70312***
DBC	10Y	-13.71053	242.0967***	13.17497
DBC	15Y	-15.04872	312.1485***	7.40092
DBC	20Y	-16.07262	674.7087***	9.38539
GLD	10Y	-13.85287	293.6489***	20.32632*
GLD	15Y	-16.29774	290.5852***	10.07019
GLD	20Y	-17.53278	570.4034***	14.56997
IYR	10Y	-14.09690	842.1373***	110.27132***
IYR	15Y	-16.04568	1149.2403***	89.51947***
IYR	20Y	-17.03187	1477.5464***	156.59787***
VIX	10Y	-16.16868	117.6091***	40.15162***
VIX	15Y	-18.79568	200.6483***	51.55856***
VIX	20Y	-19.32663	261.7829***	80.90950***

Notes: *, **, and *** indicate significance at the 10%, 5%, and 1% levels. ADF tests use the default tseries::adf.test() specification; all reject the unit-root null at 1%. ARCH-LM and raw-return Ljung–Box tests use 12 lags and test the nulls of no ARCH effects and no serial correlation, respectively.

B.2 Historical model assessments and selected coefficients

Table B2. Exploratory historical model assessments by asset and window

Asset	10-year	15-year	20-year	Qualification
SPY	EGARCH	EGARCH	EGARCH	Stable signed asymmetry across windows.
AGG	GARCH	FIGARCH†	No clean winner	10Y FIGARCH failure; 15Y/20Y boundary sensitivity; 20Y residual ARCH.
DBC	FIGARCH	EGARCH	FIGARCH	Window-sensitive rankings; recurring fractional persistence.

DOI: 10.54254/2754-1169/2026.38677

Table B2. (continued)

GLD	EGARCH	No clean winner	FIGARCH by IC	15Y: AIC favors EGARCH, BIC favors GARCH; residual ARCH in all four models. 20Y: residual-adequacy caveat.
IYR	EGARCH	EGARCH by IC	EGARCH	15Y: component/fractional models have cleaner residuals despite the IC ranking.
VIX	EGARCH	EGARCH	EGARCH	Exploratory; positive signed asymmetry across windows.

Notes: Exploratory assessments combine information criteria, parameter behavior, and residual diagnostics within compatible estimation settings; they are not a uniform selection rule or final out-of-sample rankings. Historical model coverage is not fully balanced (Appendix A.6). †The AGG 15-year FIGARCH result is boundary-sensitive because the estimated fractional parameter lies outside the conventional interpretation range.

Table B3. Selected cross-window coefficient evidence

Series/ parameter	10-year	15-year	20-year	Interpretation in source
SPY: EGARCH γ	≈ −0.170	≈ −0.193	≈ −0.184	Stable negative shock asymmetry
VIX: EGARCH γ	≈ +0.245	≈ +0.248	≈ +0.235	Positive signed-shock asymmetry
DBC: FIGARCH d	≈ 0.370	≈ 0.400	≈ 0.428	Persistent fractional dynamics
IYR: FIGARCH d	≈ 0.385	≈ 0.455	≈ 0.756	Strong window sensitivity
AGG: FIGARCH d	Estimation failed	≈ 1.02	≈ 1.01	Boundary-sensitive estimates

Notes: Approximate values retain archived precision and come from separately estimated historical windows with nonidentical endpoints. No standard errors or p-values are reported; these are descriptive estimates, not formal parameter-stability tests. AGG FIGARCH estimates are boundary-sensitive.

B.3 Historical MS-GARCH regime evidence

Table B4. Historical high-state occupancy and transition-implied durations

Asset	Window	Hard high share (%)	Stationary high probability (%)	Low-state duration (days)	High-state duration (days)
SPY	10Y	99.25+	94.44	1.23	20.83
SPY	15Y	20.94	22.24	≈435	≈122
SPY	20Y	89.13+	88.59	181.82	1428.57
AGG	10Y	11.74	10.57	2000.0	243.9
AGG	15Y	17.16	17.40	≈714	≈152
AGG	20Y	11.59	13.02	833.3	122.0
DBC	10Y	4.70	15.79	65.8	12.3
DBC	15Y	46.91	46.54	≈61	≈53
DBC	20Y	44.00	47.82	≈556	≈500
GLD	10Y	46.12	49.05	227.27	217.39
GLD	15Y	21.07	26.45	≈159	≈57

DOI: 10.54254/2754-1169/2026.38677

Table B4. (continued)

GLD	20Y	32.55	36.50	153.8	88.5
IYR	10Y	29.95	37.61	70.42	42.37
IYR	15Y	5.67	21.63	28.49	7.87
IYR	20Y	4.35	21.27	22.22	6.01
VIX	10Y	4.48	14.73	13.95	2.41
VIX	15Y	0.97	3.23	≈32.5	≈1.1
VIX	20Y	96.39+	81.60	7.26	32.15

Notes: Hard high-state shares use the maximum-smoothed-probability classification and return-dispersion labeling in Appendix A.2. Stationary probabilities and expected durations are transition-implied, not realized sample shares; $D_i = \frac{1}{1-p_{ii}}$. Approximation signs preserve archived precision. +SPY 10Y, SPY 20Y, and VIX 20Y classifications are sensitive to labeling or occupancy interpretation and are not treated as crisis frequencies. Transition probabilities are not reconstructed where the original matrix is unavailable.

B.4 Supplementary MS-GARCH model-count check

Within compatible asset-window settings, AIC favors three states only for SPY 10Y and 20Y, DBC 15Y, and VIX 20Y; BIC favors two states in every valid comparison. The AGG 20-year three-state fit is not interpretable. The two-state specification therefore remains primary.

B.5 Fixed-model forecast performance

Table B5. Fixed-model forecast performance and selected pairwise comparisons

Panel A. Overall Numerical Winner and Selected Competitor				
Asset	Best fixed model	Best fixed QLIKE	Selected competitor	Competitor QLIKE
SPY	EGARCH	0.713	CGARCH	0.734
AGG	GARCH	−1.82194	CGARCH	−1.82190
DBC	CGARCH	1.050	GARCH	1.055
GLD	CGARCH	0.945	GARCH	0.946
IYR	CGARCH	1.045	EGARCH	1.046
Panel B. Selected Pairwise QLIKE Comparisons				
Asset	Evaluation state	Favored model	Comparator	p-value
SPY	Common Normal	EGARCH	FIGARCH	0.00019
SPY	Common Normal	EGARCH	CGARCH	0.00025
SPY	Common Normal	EGARCH	GARCH	0.00002
AGG	Asset Normal	GARCH	CGARCH	0.0227
AGG	Asset Normal	GARCH	FIGARCH	0.0030
AGG	Asset Stress	CGARCH	GARCH	0.0784
DBC	Overall	CGARCH	FIGARCH	0.033
DBC	Asset Stress	CGARCH	GARCH	4.29e−05
DBC	Asset Stress	CGARCH	FIGARCH	5.95e−08

DOI: 10.54254/2754-1169/2026.38677

Notes: Panel A reports numerical winners and selected close competitors on the 4,152-date panel. Panel B reports selected pairwise comparisons using five-lag Newey–West HAC inference; the favored model has lower mean QLIKE. Tests are two-sided and unadjusted for multiple testing, and assess pairwise loss differences rather than the Stress coefficients in Table 4. Panel B is not exhaustive.

B.6 Routing performance

Table B6. Fixed-model and routing performance

Panel A. Fixed GARCH-family specifications				
Asset	GARCH	EGARCH	FIGARCH	CGARCH
SPY	0.718031	0.690131	0.713540	0.711039
AGG	−1.818695	−1.811212	−1.813886	−1.817845
DBC	1.037413	1.038643	1.039153	1.032429
GLD	0.928613	0.937347	0.935926	0.927912
IYR	1.014351	1.014149	1.016086	1.012547

Panel B. Routing strategies					
Asset	Rolling-Best	Common-State	Asset-State	Dual-State	Dual-EWMA
SPY	0.685330	0.681063	0.700983	0.684618	0.682443
AGG	−1.821228	−1.816639	−1.820302	−1.817217	−1.823610
DBC	1.033244	1.030070	1.033546	1.031273	1.029768
GLD	0.931829	0.927053	0.937333	0.922764	0.927646
IYR	1.018574	1.014302	1.018864	1.011960	1.013970

Panel C. Dual-EWMA Stress-Subset Loss Differentials Relative to Fixed Benchmarks		
Asset	Common Stress loss differential	Asset Stress loss differential
SPY	−0.1239	+0.0069
AGG	−0.0429	−0.0276
DBC	−0.0017	+0.0023
GLD	−0.0118	−0.0160
IYR	−0.0142	−0.0107

Notes: Panels A and B use the same 4,030 matched dates; lower QLIKE is better. Bold entries mark the lowest routing loss for each asset in Panel B. The fixed benchmark is selected retrospectively on this panel, while routing decisions remain ex-ante. Panel C reports Dual-EWMA-minus-fixed-benchmark QLIKE; negative values favor Dual-EWMA. None of the Stress-subset differences is significant at the 10% level, so numerical gains alone do not establish superiority.

B.7 GAS tree selection

Table B7. Initial-Sample GAS Tree Selection

Asset	Signal	Cutoff	LL gain	ΔAIC	ΔBIC	Tree activated?
AGG	VIX percentile	0.90	4.89	−3.77	+10.95	Yes
DBC	VIX percentile	0.60	4.42	−2.84	+11.88	Yes
GLD	Own-volatility percentile	0.60	2.10	+1.80	+16.52	No
IYR	VIX percentile	0.60	4.14	−2.27	+12.45	Yes

DOI: 10.54254/2754-1169/2026.38677

Table B7. (continued)

SPY	VIX percentile	0.60	5.85	−5.70	+9.03	Yes

Notes: LL gain is the Tree log-likelihood minus the Standard GAS log-likelihood. ΔAIC and ΔBIC are the Tree-minus-Standard-GAS differences in the respective information criteria.

B.8 Machine-learning prediction diagnostics

Table B8. Machine-learning prediction diagnostics

Panel A. Pooled relative-loss prediction correlations		
Target specification	XGBoost	Random Forest
GARCH	0.058	0.117
EGARCH	0.078	0.070
FIGARCH	0.097	0.099
CGARCH	0.004	0.031
Panel B. Developmental Absolute-Loss Prediction Correlations		
Feature set	Reported average correlation	
Conditions only	0.676	
Historical model performance	0.658	
Full feature set	0.640	

Notes: Panel A reports prediction–realization correlations for the retained future-five relative-loss models; Panel B reports developmental absolute-loss regressions. They illustrate overall forecasting difficulty versus relative model suitability and should not be used to rank algorithms directly. Higher absolute-loss correlations may partly reflect persistent cross-asset differences captured by asset identity.