Journal of Mathematics and Modeling in Finance (JMMF)
Vol. 7, No. 1, Winter & Spring 2027
Research paper

When Does Regime Information Improve Volatility Forecasting? Evidence from TGARCH-X, Transformer, and Combined Models

Daniel Ndirangu Mwaniki[1], **Christine Nanjala Simiyu**[2], **Jacob Otieno Ong'ala**[3]

[1] School of Business, KCA University, Nairobi, Kenya.
danimwaniki@gmail.com

[2] Dean, School of Business and Associate Professor of Economics, KCA University, Nairobi, Kenya.
cresearch@kcau.ac.ke

[3] Associate Professor of Statistics, Faculty of Biological and Physical Sciences, Tom Mboya University, Kenya.

jongala@tmu.ac.ke

Abstract:
This study examines whether latent regime information improves volatility forecasts consistently across econometric, neural, and combined models. Using daily observations for 52 firms across 11 Nairobi Securities Exchange sectors, the study evaluates forecasts from August 2016 to July 2026. The study compares matched baseline and regime-informed models within a leakage-safe chronological design, with regime-informed models using real-time filtered regime probabilities. Regimes are estimated using a zero-neutral two-state Student-t hidden Markov model that separates exact zero returns from state-dependent non-zero return variation. The forecasting set includes GARCH-family models, GRU, Transformer, PatchTST, and static and regime-dependent forecast combinations. Market-capitalisation-weighted QLIKE is the primary loss measure, supported by dependence-robust predictive-accuracy tests, Model Confidence Sets, alternative volatility targets, and specification checks. Simple econometric benchmarks remain difficult to beat on the primary squared-return target. TGARCH-X differences remain below their 80% minimum detectable effects, while equivalence testing supports a practically small overall regime-calibration effect. The probability-aware Transformer and probability-dependent combination perform worse than their respective matched baselines under the primary target. Parkinson variance strengthens the TGARCH-X result, but a Parkinson-based HMM does not outperform the primary return-based regime extraction. The economic-significance analysis finds no reliable portfolio advantage from regime calibration, despite its approximately 3.12% reduction in turnover. Overall, regime information provides selective forecasting value across models, market states, volatility targets, and specifications rather than a universal predictive benefit.

Keywords: Volatility Forecasting, Regime-Aware Forecasting, TGARCH-X, Hidden Markov Model, Transformer, Forecast Combinations.
Classification: C45; C53; C58.

[1]Corresponding author
Received: 19/08/2026 Accepted: 25/10/2026
https://doi.org/10.22054/jmmf.2026.95229.1314

1 Introduction

Accurate volatility forecasts matter for portfolio allocation, risk management, derivative valuation, and capital planning, yet financial volatility rarely follows one stable data-generating process. Volatility clustering, asymmetry, heavy tails, and structural change can alter relative model performance across market conditions. This instability makes volatility forecasting a natural setting in which to ask whether latent regime information improves predictions beyond what established econometric and modern machine-learning models already capture [19, 25]. The forecasting literature offers increasingly sophisticated answers to this problem, but greater complexity does not guarantee better out-of-sample performance. GARCH and asymmetric extensions remain difficult benchmarks because they parsimoniously capture persistent and sign-dependent volatility dynamics [3, 33]. Deep learning models can represent richer nonlinear temporal relationships, while Transformer-based architectures extend this flexibility through attention and patch-based sequence representations [18, 34]. Recent evidence, however, shows that additional architectural complexity may yield small, unstable, or target-dependent gains rather than uniform improvements [21, 22, 26]. Regime-dependent forecasting offers an alternative route to improved prediction. Markov-switching GARCH models allow volatility dynamics to vary across persistent latent states, while newer hybrid frameworks feed estimated state information into statistical or neural forecasts [10, 15, 30]. Recent work also favours probabilistic state information over rigid ex post labels because filtered probabilities preserve uncertainty about the prevailing market condition [29]. These developments suggest that regime information can matter, but they do not establish that its incremental value is the same across forecasting architectures. That distinction is central to this study. Existing regime-aware applications typically evaluate a single regime model, a single hybrid architecture, or several models that embed state information differently. Such comparisons make it difficult to determine whether forecast gains arise from the regime signal itself or from the surrounding model design. Forecast-combination research raises a related concern because combinations can help when forecast errors contain complementary information, while estimated weights may become unstable as predictive relationships change [2, 31]. A matched design using the same regime signal across econometric, neural, and combined forecasts can therefore isolate its incremental predictive value. A further complication arises in thinly traded equity markets. Repeated exact-zero returns are consistent with trading frictions and stale price adjustment, although they do not directly prove either mechanism [1, 6, 16]. In a conventional continuous-emission hidden Markov model, concentrated exact-zero returns can mechanically define the low-volatility state. The resulting classification may then reflect inactivity as much as underlying volatility. This problem is particularly relevant when regime probabilities are later treated as forecasting inputs, because a contaminated state variable can propagate directly into every regime-informed model. The study examines these issues using

daily data for 52 firms across 11 sectors of the Nairobi Securities Exchange from August 1, 2016, to July 31, 2026. The design compares established volatility benchmarks with TGARCH-X, GRU, Transformer, PatchTST, and forecast combinations under a common chronological evaluation framework. To prevent zero returns from mechanically identifying the low-volatility state, the study estimates a zero-neutral two-state Student-t hidden Markov model. Exact zero returns receive a common state-independent zero-mass contribution, while state-specific Student-t scales identify low- and high-volatility conditions from non-zero observations. The models use only filtered probabilities available before each forecast origin. The study examines when regime information improves volatility forecasts and whether its value depends on the model through which it enters. The analysis evaluates three matched comparisons: regime-calibrated and baseline TGARCH-X; probability-aware and baseline Transformer; and probability-dependent and static forecast combinations. The same filtered high-volatility probability serves as the incremental signal across all three comparisons. The study contributes by isolating the channel through which regime information affects forecasts, adapting regime extraction to repeated zero returns, and applying leakage-safe chronological evaluation with dependence-robust predictive-accuracy tests. Results favour a conditional interpretation. Regime calibration slightly improves TGARCH-X on the primary target, but the differences remain below their minimum detectable effects. Equivalence testing nevertheless supports a small overall effect in practice. The probability-aware Transformer and probability-dependent combination perform worse than their baselines, while Parkinson variance strengthens the TGARCH-X result. These findings indicate that regime information has selective rather than universal forecasting value.

2 Literature Review

Volatility forecasting has developed along several partly overlapping lines. Econometric models emphasise conditional heteroskedasticity and asymmetry, while machine-learning models emphasise flexible nonlinear representation. Regime-based models allow predictive relationships to vary with latent market conditions. A fourth strand examines forecast combination and conditional predictive ability when no single model uniformly dominates. The present study connects these four strands because its central question concerns the incremental value of the same regime information across different forecasting architectures.

2.1 Econometric Volatility Forecasting

GARCH models remain a natural benchmark for short-horizon volatility forecasting because they parsimoniously represent volatility clustering through lagged shocks and conditional variance [3]. Extensions that distinguish positive and negative shocks address the well-established asymmetry of financial volatility. [9] incorporated

asymmetric shock effects within a GARCH framework, while [33] developed a threshold specification directly for conditional scale. These models remain relevant even as forecasting research adopts more flexible nonlinear methods. Recent evidence continues to treat conventional GARCH-family forecasts as demanding comparators rather than obsolete baselines. [19] compare Markov-switching multifractal volatility forecasts with established alternatives in commodity, currency, and stock markets. [27] similarly benchmark statistical, machine-learning, LSTM, and Transformer approaches within a common volatility-forecasting framework. A recent review by [17] shows that realised-volatility forecasting has expanded substantially, yet model rankings remain sensitive to the target, horizon, information set, and market under study. Forecast evaluation adds another complication because volatility is latent. [25] shows that forecast comparison can remain valid with imperfect volatility proxies when the loss function satisfies appropriate robustness conditions. This insight supports QLIKE as a primary loss measure and motivates sensitivity analysis using alternative observable proxies. In this setting, squared returns provide a common daily target, while the Parkinson range estimator offers an alternative proxy when intraday highs and lows are available.

2.2 Deep Learning and Modern Time-Series Architectures

Deep-learning methods broaden the set of forecast information by learning nonlinear temporal representations without requiring a fully specified parametric volatility recursion. Recurrent architectures such as LSTM and GRU can retain information across sequences, while attention-based models can represent longer-range dependence more flexibly. [18] demonstrated the forecasting value of attention and gating in the Temporal Fusion Transformer. More recently, [34] adapted multi-scale patching and channel-independent attention to model volatile financial time series. Financial applications increasingly combine neural architectures with volatility-specific inputs or econometric components. [21] developed DeepVol to forecast day-ahead volatility directly from high-frequency observations. [20] compared multi-Transformer, LSTM, and GARCH-based architectures across several financial markets. [32] used adaptive integration of convolutional, attention, and recurrent components for stock forecasting. Together, these studies illustrate a broader movement towards data-dependent model fusion. However, recent work also cautions against equating architectural complexity with forecasting superiority. [26] uses leakage-aware deep-learning evaluation and ablation analysis and finds that an additional Transformer pathway does not consistently improve forecasting. [22] likewise reports that relative model performance varies across assets and loss functions, with no single learning algorithm dominating uniformly. These findings make strong econometric and recurrent benchmarks essential when evaluating a Transformer-based volatility model.

2.3 Regime-Dependent Forecasting

Markov-switching and hidden Markov models capture persistent latent states through state-dependent distributions and transition probabilities. [15] and [10] link such states directly to conditional volatility, while more recent work extends regime information to multifractal, statistical, and neural forecasts ([19], [30]). [29] further show that probabilistic state assignments can enter forecasts without treating regimes as perfectly observed. In thinly traded markets, however, repeated zero returns may cause continuous-emission HMMs to overidentify the low-volatility state. This study therefore neutralises exact zeros in state discrimination and uses filtered probabilities for forecasting.

2.4 Forecast Combination and Conditional Predictive Value

Forecast combination provides a complementary way to exploit heterogeneous models. [2] show that combining imperfect forecasts can improve predictive performance when their errors contain different information. In addition, [31] similarly explains why simple combinations may not outperform their components, as gains depend on error dependence, weight estimation, and stability. This conditional perspective also shapes how the study evaluates forecasts. [5] test equal predictive accuracy using loss differentials, while [8] shift attention towards conditional predictive ability. [7] further show that relative forecast performance can vary in unstable environments. [13] addressed model multiplicity through the Model Confidence Set, which identifies a set of statistically indistinguishable superior models rather than mechanically selecting a single winner. Recent forecasting papers increasingly adopt this restrained interpretation. [22] combines pairwise predictive accuracy tests with model-set comparisons and reports heterogeneous performance across assets and objectives. [30] evaluate regime-aware forecasting across markets rather than assuming that state information helps uniformly. Together, this literature suggests that the relevant question is not whether regime information is universally useful, but when it changes predictive accuracy and through which modelling channel.

2.5 Research Gap and Contribution

The literature leaves three related gaps. First, regime-dependent studies typically evaluate a particular regime model or hybrid architecture. As a result, they provide limited evidence on whether the same latent-state information consistently adds value across econometric, neural, and combined forecasts. Second, deep-learning volatility studies often compare architectures on average accuracy without isolating the incremental contribution of an externally estimated regime probability. Third, repeated zero returns can distort regime identification in thinly traded markets, yet the regime-forecasting literature gives limited attention to this issue. This study addresses these gaps through matched baseline and regime-informed forecasts. It

evaluates regime-calibrated TGARCH-X against TGARCH-X, a probability-aware Transformer against the same Transformer without regime probabilities, and a probability-dependent combination against a static combination. The common filtered high-volatility probability therefore acts as the incremental information variable across model classes rather than as a different regime definition for each model. The study also adapts a zero-neutral Student-t HMM specification to this forecasting context. Exact zero returns receive a common state-independent emission contribution, while non-zero observations identify low- and high-volatility states through state-specific Student-t scales. Filtered rather than smoothed probabilities preserve the real-time information set. The contribution lies in adapting regime extraction to a thinly traded equity environment and then testing whether that information improves out-of-sample forecasts. Finally, the empirical design treats conditionality as an outcome rather than an assumption. The analysis evaluates forecasts using strong econometric and neural benchmarks, dependence-robust predictive-accuracy tests, and Model Confidence Sets. The study further assesses robustness across alternative volatility proxies, sectors, weighting schemes, regime specifications, neural sequence lengths, and TGARCH-X sensitivity analyses. Together, this literature shifts the question from whether regime information is universally useful to when it changes predictive accuracy and through which modelling channel.

3 Materials and Methodology

The empirical design separates data construction, model estimation, forecast combination, and final evaluation. This structure preserves a strict chronological information set and evaluates every focal forecast on the same untouched FinalTest period.

3.1 Sample and Sector Portfolio Construction

The sample contains daily observations for 52 firms listed on the Nairobi Securities Exchange from August 1, 2016, to July 31, 2026. The 11 sectors are Agriculture, Automobiles and Accessories, Banking, Commercial and Services, Construction and Allied, Energy and Petroleum, Insurance, Investment, Investment Services, Manufacturing and Allied, and Telecommunication and Technology. Firm returns are one-trading-day log differences in closing prices. The construction excludes price changes that span non-consecutive market trading dates rather than treating them as ordinary one-day returns. Sector returns are then defined as lagged market-capitalisation-weighted averages of available firm returns. Using lagged capitalisation prevents contemporaneous returns from affecting the portfolio weights used to construct them.

$$r_{s,t} = \sum_{i \in s} w_{i,s,t-1} r_{i,t}, \qquad w_{i,s,t-1} = \frac{MC_{i,t-1}}{\sum_{j \in s} MC_{j,t-1}}. \tag{1}$$

where $r_{s,t}$ is the return of sector s on trading day t; $w_{i,s,t-1}$ is the market-capitalisation weight of firm i in sector s, determined on the preceding trading day $t-1$; $r_{i,t}$ is the one-trading-day log return of firm i on day t; and $MC_{i,t-1}$ is the market capitalisation of firm i on day $t-1$. Here, i denotes the firm, s the sector, t the trading day, and j indexes the firms included in the sector-level market-capitalisation denominator. The source panel contains occasional sector-date gaps and several thinly traded sectors. The analysis neither fabricates missing returns nor replaces missing prices with zeros. When a sector skips a market-wide trading date, the rolling forecast refers to its next observed date. The data audit records these gaps and contributing-stock counts by sector-date.

3.2 Volatility Targets and Trading-Activity Predictors

The primary one-day volatility proxy is the squared sector return, bounded below by 10^{-8} for numerical stability. Although noisy, it provides a common observable sector-day target. To assess target sensitivity, the analysis also evaluates the fixed forecasts against the range-based Parkinson estimator ([24, 25]).

The primary volatility proxy is defined as follows:

$$RV_{s,t} = \max\left(r_{s,t}^2,\, 10^{-8}\right). \tag{2}$$

where $RV_{s,t}$ is the one-day volatility proxy for sector s on trading day t; $r_{s,t}$ is the sector return; and 10^{-8} is the lower bound imposed for numerical stability.

The firm-level Parkinson variance is defined as follows:

$$PV_{i,t} = \frac{\left[\ln\left(\frac{H_{i,t}}{L_{i,t}}\right)\right]^2}{4\ln(2)}. \tag{3}$$

where $PV_{i,t}$ is the firm-level Parkinson variance for firm i on trading day t; $H_{i,t}$ and $L_{i,t}$ are, respectively, the highest and lowest observed prices of firm i on day t; and ln denotes the natural logarithm.

The analysis aggregates firm-level Parkinson variances using lagged market-capitalisation weights when valid high–low ranges cover at least 80% of contributing firms. Forecasts remain fixed across evaluation proxies, isolating target sensitivity from model re-estimation. The focal TGARCH-X exogenous variable is the one-period lag of the change in log sector turnover. Sector turnover equals traded value divided by lagged sector market capitalisation, with traded value defined as price multiplied by volume. The transformation applies $\ln(1 + \text{Turnover})$ before differencing. Volume enters only when valid observations cover at least 80% of contributing firms.

The lagged turnover-change predictor is defined as follows:

$$x_{s,t-1} = \Delta \ln\left(1 + \text{Turnover}_{s,t-1}\right). \tag{4}$$

where $x_{s,t-1}$ is the lagged change in log turnover for sector s; Δ is the one-period first-difference operator; and $\text{Turnover}_{s,t-1}$ is sector turnover measured on the preceding trading day $t-1$.

Within each rolling TGARCH-X window, the analysis standardises the exogenous regressor using only in-window observations. Missing values are replaced by the corresponding training-window median, with a separate missingness indicator included in the variance-regressor matrix. No future observation enters regressor imputation or scaling.

3.3 Chronological Forecasting Design and Leakage Control

All neural models and regime-dependent procedures follow a common chronological split that separates hyperparameter selection, forecast calibration, and final evaluation. Table 1 summarises the four segments.

Table 1: Chronological Forecasting Design

Split	Period	Dates	Primary purpose
Train	22 May 2017–10 Mar. 2021	950	Neural tuning; tuning HMM fit
Validation	11 Mar. 2021–18 Feb. 2022	238	Hyperparameter selection; final HMM boundary
StackTrain	21 Feb. 2022–9 May 2023	298	Calibration and forecast-combination estimation
FinalTest	10 May 2023–31 July 2026	801	Untouched out-of-sample evaluation

Note. The sample begins before the Train segment because rolling econometric models and neural sequences require presample history. No FinalTest outcome enters model tuning, HMM estimation, calibration, or combination weights.

Neural hyperparameters are selected using Train observations and chronological Validation loss. Final neural models are then refitted through validation and begin retained forecasting at StackTrain. StackTrain estimates calibration factors and combination weights, while FinalTest remains untouched until model selection is complete. Feature scaling follows the same chronology. Tuning uses means, standard deviations, and medians estimated only from Train, while final neural fits recompute them through validation. The analysis never shuffles observations, and each forecast uses a 20-day sequence ending before the forecast date. This design follows leakage-control principles for latent-state and deep-learning forecasts ([26, 30]).

3.4 Econometric Benchmark Forecasts and TGARCH-X

The econometric benchmarks include historical variance, EWMA, GARCH(1,1), GJR-GARCH(1,1), TGARCH(1,1), and TGARCH-X. EWMA uses $\lambda = 0.94$. The analysis re-estimates GARCH-family models in 200-observation rolling windows and generates one-step-ahead forecasts using Student-t innovations as the primary

specification. The TGARCH specifications use the conditional-scale parameterisation implemented by the fGARCH-TGARCH model rather than a squared-variance GJR recursion. For one lag, the focal TGARCH-X scale evolves as follows.

$$\sigma_t = \omega + \alpha_1 \left(|\varepsilon_{t-1}| - \eta_1 \varepsilon_{t-1} \right) + \beta_1 \sigma_{t-1} + \delta_1 x_{t-1} + \delta_2 m_{t-1}. \tag{5}$$

where σ_t is the conditional volatility scale on trading day t; ω is the constant term; α_1 measures the response to the magnitude of the lagged shock. ε_{t-1} is the innovation from the conditional mean equation on day $t-1$; η_1 is the asymmetry parameter, allowing negative and positive shocks of equal magnitude to have different effects. β_1 is the persistence coefficient on the lagged conditional scale σ_{t-1}; δ_1 denotes the coefficient on x_{t-1}, the standardised lagged turnover-change signal, and δ_2 represents the coefficient on m_{t-1}, the corresponding lagged missingness indicator. The vertical bars around ε_{t-1} denote its absolute value. Removing both external regressors yields the TGARCH benchmark, with a constant conditional mean and no ARMA terms. The scale-based threshold form follows Zakoian's asymmetric-volatility formulation and distinguishes TGARCH-X from a regime-switching GARCH model [33]. GARCH-family models remain demanding volatility benchmarks, and added regime or neural complexity need not dominate simpler specifications ([19, 22]).

3.5 Neural Forecasting Benchmarks

Three neural architectures provide complementary benchmarks: GRU as a recurrent model, Transformer as an attention-based sequence model, and PatchTST as a patch-based time-series model. All forecast the same positive one-day variance target and minimise QLIKE during training. Inputs comprise seven lagged variables: sector return, realised variance, log sector volume, volume return, and 5-, 10-, and 20-day rolling return volatility. Each continuous feature has a missingness indicator, and the primary sequence contains 20 trading observations. The analysis replaces missing values with the applicable training-sample median before standardisation. The GRU uses one recurrent layer followed by a dense GELU layer. The Transformer combines latent projection, positional encoding, multi-head self-attention, global average pooling, and a positive output layer. PatchTST forms five-observation channel patches and applies attention across them. These architectures follow established attention-based and patch-based forecasting designs ([18, 23]). The analysis selects hyperparameters from two compact candidate configurations per architecture. Training uses at most 40 epochs, batch size 64, and early stopping after six validation epochs without improvement. Final forecasts average predictions across five fixed seeds (11, 23, 42, 71, and 101). The Supporting Information reports parameter counts and five-seed forecast stability.

3.6 Zero-Neutral Student-t Hidden Markov Model

The study estimates regime information from sector returns rather than forecast errors or the squared-return volatility target. To prevent repeated exact-zero returns from mechanically identifying low volatility, the primary HMM separates the zero mass from the continuous state-dependent return distribution. Let $S_t \in \{L, H\}$ denote the latent low- or high-volatility state. Exact zero returns receive a common state-independent zero-mass contribution π_0, while non-zero standardised returns follow state-specific Student-t emissions, ordered such that $\sigma_H > \sigma_L$.

$$P(r_t = 0 \mid S_t = s) = \pi_0, \qquad s \in \{L, H\}. \tag{6}$$

where r_t is the sector return on trading day t; S_t is the latent volatility state on day t; s denotes a generic state; L and H denote the low- and high-volatility states, respectively; and π_0 is the common state-independent probability mass assigned to an exact zero return. $P(\cdot)$ denotes probability, and the vertical bar denotes conditioning. The state-specific distribution of non-zero standardised returns is specified as follows:

$$z_t \mid (r_t \neq 0, S_t = s) \sim t_\nu(0, \sigma_s), \qquad \sigma_H > \sigma_L. \tag{7}$$

where z_t is the demeaned and standardised non-zero return on trading day t; t_ν denotes a Student-t distribution with ν degrees of freedom; and σ_s is the scale parameter for state s, with σ_H and σ_L denoting the high- and low-volatility state scales, respectively. The analysis represents the transition probabilities between the low- and high-volatility states as follows:

$$P = \begin{bmatrix} p_{LL} & 1 - p_{LL} \\ 1 - p_{HH} & p_{HH} \end{bmatrix}. \tag{8}$$

where P is the two-state transition-probability matrix. p_{LL} is the probability of remaining in the low-volatility state from one trading day to the next, p_{HH} is the probability of remaining in the high-volatility state, and $1 - p_{LL}$ and $1 - p_{HH}$ are the corresponding probabilities of switching between the two states. The analysis demeans and standardises non-zero returns using only the relevant HMM fitting sample. Estimation uses deterministic multi-start optimisation, Student-t degrees of freedom above 2.05, and filtered rather than smoothed probabilities. Accepted solutions must converge, retain at least 3% of fitting observations per state, achieve a high-to-low realised-variance ratio of at least 1.20, and avoid boundary transition probabilities. A separate validity screen tests whether the low state merely reproduces zero-return days using zero dominance, zero-return capture, and zero-state association criteria. These thresholds are diagnostic safeguards rather than universal cut-offs, and the model permits no fallback classification when the primary model fails. The Supporting Information reports the detailed validity rules and filtering timeline. Regime estimation follows the forecasting chronology. The

study fits the tuning HMM through the Train period and filters validation forward. It then re-estimates the final HMM through validation and filters StackTrain and FinalTest using fixed parameters. The forecast-time conditioning variable is the lagged filtered high-volatility probability:

$$p_{t-1} = P(S_{t-1} = H \mid \mathcal{F}_{t-1}). \tag{9}$$

where p_{t-1} is the lagged filtered probability of the high-volatility state. $P(\cdot)$ denotes conditional probability; S_{t-1} is the latent volatility state at trading day $t-1$, H denotes the high-volatility state; and $\mathcal{F}_{t-1}$ is the information set available through trading day $t-1$. This timing ensures that forecasts dated t use only regime information available before t. Hard states are retained for descriptive tables, while the focal regime-informed models use the continuous filtered probability, consistent with probabilistic regime-aware forecasting designs [11, 30].

3.7 Regime-Informed Forecasts and Forecast Combinations

The analysis evaluates regime information through three matched channels, each with a directly comparable baseline. First, regime-calibrated TGARCH-X rescales the fixed baseline forecast rather than estimating a regime-switching GARCH model. StackTrain estimates low- and high-state calibration factors from realised-to-forecast variance ratios, shrinks them toward the unconditional factor, and bounds them between 0.25 and 4.00. The forecast-time factor interpolates continuously with the filtered high-state probability. The forecast-time calibration factor is defined as follows:

$$c_t = (1 - p_{t-1})c_L + p_{t-1}c_H. \tag{10}$$

where c_t is the forecast-time calibration factor, p_{t-1} is the lagged filtered high-state probability from the previous equation, and c_L and c_H are the low- and high-state calibration factors. The regime-calibrated TGARCH-X forecast is then:

$$\widehat{v}_t^{(RC)} = c_t\, \widehat{v}_t^{(\text{TGARCH-X})}. \tag{11}$$

where $\widehat{v}_t^{(RC)}$ is the regime-calibrated forecast, $\widehat{v}_t^{(\text{TGARCH-X})}$ is the baseline TGARCH-X forecast, and c_t is defined in the previous equation. Second, the probability-aware Transformer adds the filtered high-state probability to the same 20-day feature sequence used by the baseline Transformer, with p_{t-1} as the latest regime-probability observation. Architecture, tuning, QLIKE loss, five-seed ensembling, and fitting horizon otherwise remain unchanged. Third, the static combined forecast uses sector-specific convex weights on baseline Transformer and TGARCH-X forecasts, estimated on StackTrain by chronological cross-validation and fixed in FinalTest. Equal-weight and unconstrained linear combinations provide robustness checks ([2, 31]). The regime-aware combination allows the Transformer weight to vary smoothly with the high-state probability, with both parameters estimated only on

StackTrain:

$$w_t = \frac{\exp(a + b\, p_{t-1})}{1 + \exp(a + b\, p_{t-1})}. \tag{12}$$

where w_t is the Transformer weight, a is the intercept, b governs its response to p_{t-1}, and the logistic transformation constrains w_t to $(0, 1)$. The regime-aware combined forecast is:

$$\widehat{v}_t^{(RCOMB)} = w_t \widehat{v}_t^{(RTRANS)} + (1 - w_t)\widehat{v}_t^{(RC)}. \tag{13}$$

where $\widehat{v}_t^{(RCOMB)}$ combines the regime-aware Transformer forecast $\widehat{v}_t^{(RTRANS)}$ and regime-calibrated TGARCH-X forecast $\widehat{v}_t^{(RC)}$ using weights w_t and $1 - w_t$, respectively. The logistic constraint keeps the combination convex at every forecast origin; constant- and equal-weight versions provide robustness checks.

3.8 Forecast Evaluation and Statistical Inference

Forecast evaluation uses QLIKE as the primary loss for latent volatility, with RMSE and MAE as complementary scale-dependent measures [25]. QLIKE is defined as follows:

$$\text{QLIKE}_t = \frac{y_t}{f_t} - \ln\left(\frac{y_t}{f_t}\right) - 1. \tag{14}$$

where QLIKE_t is the QLIKE loss on trading day t, y_t is the observed variance proxy on day t, f_t is the corresponding variance forecast for day t, and ln denotes the natural logarithm. Primary aggregate loss uses daily lagged market-capitalisation-weighted averages of sector-level losses. Equal-sector and 25%-capped market-capitalisation weights assess concentration sensitivity, while sector-level results assess cross-sectional heterogeneity. Matched comparisons use Diebold–Mariano tests with HAC long-run variance estimates and the Harvey–Leybourne–Newbold finite-sample correction ([5, 14]). Holm adjustment controls family-wise error within each volatility condition, while 95% moving-block bootstrap intervals use 5,000 replications and a 10-day block. The same framework evaluates fixed forecasts under Parkinson variance. Minimum detectable effects assess whether non-significant differences could reflect limited detectability under the observed sample and dependence structure. The analysis computes 80% MDEs using the matched-test HAC variance and HLN adjustment. Conventional MDEs use a two-sided 5% level, while family-wise MDEs apply 5%/3 across the three comparisons within each condition. For TGARCH-X, TOST evaluates practical equivalence at ±0.5%, ±1%, ±2%, and ±5% of baseline QLIKE. Equivalence requires both one-sided tests to reject at 5%, with corresponding 90% HAC-HLN confidence intervals reported. MCS and SPA procedures use 5,000 stationary-bootstrap replications, with the MCS evaluated at 90% confidence ([12, 13]). Giacomini–White tests use the lagged loss differential and filtered high-state probability to assess conditional predictive ability, while Giacomini–Rossi statistics assess local instability ([7, 8]). Mincer–Zarnowitz regressions assess calibration at one- and five-day horizons. A secondary moderation regression models $\log(1 +$

QLIKE) using model indicators, continuous forecast-time high-state probability, their interactions, realised variance, lagged volume return, and sector fixed effects. Standard errors are clustered by sector and date following [4]. Given the cross-section of 11 sectors, interaction inference also uses a Webb-weight wild-cluster bootstrap with 999 replications ([28]). Date fixed effects and hard-state interactions provide sensitivity checks, with Holm correction applied across the model-by-regime interaction family.

3.9 Robustness and Reproducibility

Robustness checks vary the evaluation target, portfolio weighting, regime structure, neural sequence length, split classification, and TGARCH-X specification. Regime checks include zero-neutral Gaussian, conventional Student-t, and three-state zero-neutral Student-t HMMs, while 10-, 20-, and 40-day neural sequences are fully re-estimated. TGARCH-X sensitivity covers 125-, 200-, and 250-observation windows, Gaussian, Student-t, and GED innovations, and alternative exogenous-variable definitions and lags. Comparisons use common FinalTest support, paired QLIKE inference, 5,000 bootstrap replications, and Holm adjustment. The pre-declared primary specification remains the 200-observation Student-t model with turnover-change lag 1. A separate two-state Gaussian HMM tests whether Parkinson sensitivity depends on regime extraction. It uses standardised log positive Parkinson variance, treats zero or missing ranges equally across states, and orders the high state by the larger emission mean. The model follows the primary HMM chronology and uses forward-filtered probabilities without smoothing. Each sector requires at least 120 valid fitting observations and must pass the pre-specified occupancy, separation, and transition checks. No fallback classification is applied, and analysis compares only accepted sectors on matched common support. This variant changes regime extraction while retaining the squared-return calibration target. An illustrative volatility-targeting exercise assesses economic significance. Each model targets 10% annual volatility with exposure capped at two. The analysis combines model exposure with lagged market-capitalisation sector weights and assigns the residual position to cash or leverage. Matched Sharpe-ratio differences use 5,000 paired moving-block bootstrap replications with a 10-day block. One-way turnover equals one-half of the sum of absolute sector-weight changes and the absolute change in the cash/leverage position. Transaction-cost sensitivity applies 0, 5, 10, 25, and 50 basis points per unit of turnover before recomputing net Sharpe ratios, assuming a zero risk-free rate. The complete Python analysis was regenerated from scratch in one session using fixed seeds and no prior checkpoints. The reproducibility bundle stores exact inputs, software environment, run configuration, SHA256 hashes, and the file manifest. The primary and sensitivity R stages retain corresponding manifests and session information, while the Supporting Information reports detailed diagnostic and robustness results.

4 Results

This section evaluates out-of-sample forecasting performance over the untouched FinalTest period from May 10, 2023, to July 31, 2026. Market-capitalisation-weighted QLIKE is the primary loss measure, while RMSE and MAE provide complementary evidence. The analysis first compares overall forecast performance and then isolates the incremental value of regime information. It next examines statistical significance, practical equivalence, volatility-target sensitivity, sector heterogeneity, robustness, and economic significance.

4.1 Overall Out-of-Sample Forecast Performance

Overall, the analysis evaluates out-of-sample performance across the benchmark, baseline, and regime-informed models. Table 2 compares the benchmark, baseline, and regime-informed models using aggregate FinalTest QLIKE, RMSE, and MAE.

Table 2: Overall FinalTest Forecast Accuracy

Model	QLIKE	RMSE	MAE	Role
EWMA	2.0889	0.000701	0.000285	Econometric benchmark
GARCH	2.1182	0.000872	0.000358	Econometric benchmark
TGARCH	2.1322	0.001059	0.000396	Econometric benchmark
Regime-Calibrated TGARCH-X	2.1364	0.009969	0.000607	Focal regime model
Baseline TGARCH-X	2.1372	0.010071	0.000613	Focal baseline
Historical Variance	2.1469	0.000714	0.000296	Econometric benchmark
GJR-GARCH	2.1604	0.000822	0.000330	Econometric benchmark
PatchTST	2.3971	0.000819	0.000432	Neural benchmark
Static Combined Forecast	3.7632	0.045016	0.021196	Focal baseline
Regime-Aware Combined Forecast	3.9988	0.043831	0.022085	Focal regime model
Baseline Transformer	4.4178	0.166517	0.078300	Focal baseline
Regime-Aware Transformer	4.6912	0.173429	0.087494	Focal regime model
GRU	6.4091	1.686747	1.101387	Neural benchmark

Note. Results use daily market-capitalisation weights over 801 FinalTest dates. Lower values indicate better forecast accuracy. QLIKE is the primary ranking criterion.

EWMA records the lowest QLIKE (2.0889), followed by GARCH (2.1182) and TGARCH (2.1322). Regime-calibrated TGARCH-X (2.1364) improves only slightly over baseline TGARCH-X (2.1372). Historical variance and GJR-GARCH remain competitive, while PatchTST is the strongest neural benchmark but does not surpass the leading econometric models. The Transformer, combined forecasts, and GRU perform materially worse, indicating that greater architectural complexity does not

automatically reduce out-of-sample volatility loss. Supporting Information Tables S1–S4 provide the evidence crosswalk, forecast-coverage and imputation audits, and neural tuning diagnostics.

4.2 Incremental Forecast Value of Regime Information

Before comparing regime-informed forecasts, the primary zero-neutral Student-t HMM passed the pre-specified regime-validity screen in all 11 sectors. No sector failed the zero-return dominance rule. Automobiles and Accessories generated a rare-state warning because its low-volatility state covered 4.67% of the fitting observations, below the 5% warning threshold but above the 3% minimum occupancy requirement. Table 3 reports the incremental QLIKE effects of regime information for the three matched model pairs across the full, low-, and high-volatility samples.

Table 3: Incremental QLIKE Effect of Regime Information

Matched model	Condition	Baseline QLIKE	Regime-informed QLIKE	Difference
TGARCH-X	All	2.1372	2.1364	-0.0009
TGARCH-X	Low	1.9653	1.9637	-0.0016
TGARCH-X	High	2.2625	2.2521	-0.0104
Transformer	All	4.4178	4.6912	+0.2734
Transformer	Low	4.4091	4.5620	+0.1529
Transformer	High	3.7808	3.9199	+0.1390
Combined forecast	All	3.7632	3.9988	+0.2356
Combined forecast	Low	3.7847	3.8511	+0.0663
Combined forecast	High	3.1267	3.3921	+0.2653

Note. Difference equals regime-informed QLIKE minus baseline QLIKE. Negative values favour the regime-informed forecast. Low and High conditions use the forecast-time regime classification derived from the filtered high-volatility probability.

Regime calibration lowers TGARCH-X QLIKE overall and in both states, most under high volatility, but the overall gain remains small. Regime probabilities instead raise Transformer QLIKE in both states, while the probability-dependent combination also underperforms its static counterpart, especially under high volatility. Section 4.3 tests whether these differences are statistically and practically meaningful. Further HMM diagnostics, regime-calibration results, and implementation details are reported in Supporting Information Tables S5–S10 and S24.

4.3 Statistical Significance, Detectable Effects, and Practical Equivalence

The analysis next examines whether regime information produces statistically reliable forecast differences. Table 4 tests matched baseline and regime-informed forecast differences across the full sample and both volatility states.

Table 4: Predictive-Accuracy Tests for Matched Baseline and Regime-Informed Models

Matched model	Condition	Mean Δ QLIKE	HLN p	Holm p	95% block-bootstrap CI
TGARCH-X	All	-0.0009	0.878	0.878	[-0.0132, 0.0085]
TGARCH-X	Low	-0.0016	0.809	0.809	[-0.0151, 0.0109]
TGARCH-X	High	-0.0104	0.155	0.155	[-0.0249, 0.0042]
Transformer	All	+0.2734	< 0.001	< 0.001	[0.1867, 0.3632]
Transformer	Low	+0.1529	< 0.001	< 0.001	[0.0624, 0.2391]
Transformer	High	+0.1390	0.024	0.048	[0.0120, 0.2776]
Combined forecast	All	+0.2356	< 0.001	< 0.001	[0.1833, 0.2913]
Combined forecast	Low	+0.0663	0.006	0.013	[0.0116, 0.1177]
Combined forecast	High	+0.2653	< 0.001	< 0.001	[0.1718, 0.3677]

Note. Mean Δ QLIKE equals regime-informed loss minus baseline loss, so negative values favour regime information. HLN denotes the Harvey–Leybourne–Newbold correction to the Diebold–Mariano statistic. The analysis constructs confidence intervals using 5,000 moving-block bootstrap replications.

Formal inference yields different conclusions across the three model classes. The TGARCH-X loss reductions are not statistically distinguishable from zero overall or within either volatility state. The high-state reduction is directionally larger, but its HLN p-value remains 0.155, and the bootstrap interval includes zero. In contrast, the regime-aware Transformer and combined forecast produce significantly higher loss than their matched baselines after Holm adjustment. Supporting Information Table S13 reports the complete matched predictive-accuracy inference.

Non-significance does not establish a zero effect. Table 5 therefore reports 80% minimum detectable effects and TGARCH-X equivalence tests to distinguish limited detectability from practical negligibility.

The absolute TGARCH-X differences are smaller than their 80% MDEs across the full sample and both volatility states. The non-significant predictive-accuracy tests therefore cannot, by themselves, establish an exactly zero effect. Equivalence testing provides complementary evidence. The full-sample difference satisfies the stringent ±0.5% baseline-QLIKE margin (TOST $p = 0.0383$), while the low- and high-state differences satisfy the wider ±1% margin. These results indicate that the overall TGARCH-X effect is practically small, although the state-specific effects are less tightly bounded. Supporting Information Table S25 reports the detailed MDE and equivalence results.

The 90% Model Confidence Set retains only baseline and regime-calibrated TGARCH-X among the six focal models across all three conditions. The Giacomini–White test does not reject equal conditional predictive ability for the TGARCH-X pair ($p = 0.116$), whereas it rejects equality for the Transformer and combined-forecast pairs. Giacomini–Rossi statistics likewise indicate temporal instability in the latter two comparisons but not in TGARCH-X. Model-by-regime interactions provide weaker evidence of state dependence, with none remaining significant after Holm correction. Supporting Information Tables S14–S16 report the corresponding

model-set, conditional predictive-ability, forecast-instability, and moderation results. Figures S1–S3 show the corresponding rolling matched QLIKE differentials.

Table 5: Detectable Effects and Practical Equivalence

Panel A. Minimum detectable effects at 80% power

Matched model	Condition	Observed Δ QLIKE	80% MDE	MDE (% baseline)	Family-wise 80% MDE	Family-wise MDE (% baseline)
TGARCH-X	All	-0.0009	0.0155	0.73%	0.0180	0.84%
Transformer	All	+0.2734	0.1340	3.03%	0.1549	3.51%
Combined forecast	All	+0.2356	0.0810	2.15%	0.0936	2.49%
TGARCH-X	Low	-0.0016	0.0185	0.94%	0.0214	1.09%
Transformer	Low	+0.1529	0.1319	2.99%	0.1524	3.46%
Combined forecast	Low	+0.0663	0.0775	2.05%	0.0895	2.36%
TGARCH-X	High	-0.0104	0.0208	0.92%	0.0240	1.06%
Transformer	High	+0.1390	0.1926	5.09%	0.2226	5.89%
Combined forecast	High	+0.2653	0.1447	4.63%	0.1672	5.35%

Panel B. TGARCH-X practical-equivalence sensitivity

Condition	Observed Δ QLIKE	Smallest supported margin	Absolute margin	TOST p	90% CI
All	-0.0009	±0.5%	±0.0107	0.0383	[-0.0100, 0.0083]
Low	-0.0016	±1.0%	±0.0197	0.0031	[-0.0124, 0.0093]
High	-0.0104	±1.0%	±0.0226	0.0495	[-0.0226, 0.0018]

Note. MDE is the minimum absolute QLIKE difference detectable with 80% power under HAC inference. Family-wise MDE incorporates the multiple-testing adjustment across the matched-comparison family. The equivalence margin is the smallest tested percentage of baseline QLIKE (0.5%, 1%, 2%, or 5%) satisfying TOST at the 5% level, with HAC standard errors and HLN adjustment.

4.4 Volatility-Target Sensitivity and Alternative Regime Extraction

The matched TGARCH-X conclusion changes materially when Parkinson variance replaces squared returns as the evaluation target. Table 6 applies the same dependence-robust comparison framework to the fixed FinalTest forecasts evaluated against Parkinson variance.

The analysis holds the forecasts fixed, so differences from the primary results come only from changing the evaluation target. Under Parkinson variance, regime calibration reduces TGARCH-X QLIKE from 2.8255 to 2.5694 overall (Holm $p < 0.001$) and by 0.3678 in the low-volatility state ($p < 0.001$). The high-volatility reduction is not statistically significant. Supporting Information Tables S18 and S26 report the broader Parkinson-target results and complete matched predictive-accuracy inference.

Table 6: Matched Forecast Comparisons under Parkinson Variance

Matched model	Condition	Baseline QLIKE	Regime-informed QLIKE	Δ QLIKE	Holm p	95% bootstrap CI
TGARCH-X	All	2.8255	2.5694	-0.2561	< 0.001	[-0.2925, -0.2262]
Transformer	All	3.0355	3.2416	+0.2061	< 0.001	[0.1139, 0.3114]
Combined forecast	All	2.6622	2.8001	+0.1379	< 0.001	[0.0775, 0.2014]
TGARCH-X	Low	3.5680	3.2001	-0.3678	< 0.001	[-0.4265, -0.3195]
Transformer	Low	3.1882	3.3148	+0.1266	0.022	[0.0310, 0.2186]
Combined forecast	Low	3.1765	3.1189	-0.0576	0.093	[-0.1252, 0.0043]
TGARCH-X	High	1.5225	1.4945	-0.0281	0.362	[-0.0693, 0.0142]
Transformer	High	2.5637	2.5842	+0.0205	0.778	[-0.1001, 0.1683]
Combined forecast	High	1.8184	1.9897	+0.1713	0.006	[0.0714, 0.2800]

Note. Negative Δ QLIKE values indicate that regime information improves forecast accuracy. The Holm adjustment controls the three matched comparisons within each volatility condition, and 95% CIs use 5,000 moving-block bootstrap replications.

Regime information remains architecture-dependent. The Transformer performs worse overall and in the low-volatility state, while the combined forecast performs worse overall and in the high-volatility state. The stronger TGARCH-X result may reflect the information captured by the high–low range. Squared returns use close-to-close changes and can miss substantial intraday movement when closing prices change little, whereas Parkinson variance captures intraday price dispersion. This distinction may matter in a less-liquid market, although range-based measures remain sensitive to sparse trading and incomplete price discovery.

A Parkinson-based HMM provides a further robustness check under the same chronological constraints. The analysis excludes Automobiles and Accessories because its fitting sample contains only 61 valid Parkinson observations, leaving matched support across 10 sectors without fallback classification. Under Parkinson evaluation, this calibration reduces QLIKE relative to baseline TGARCH-X (2.5823 versus 2.8243), but the primary return-HMM remains slightly better at 2.5683. The difference between the two HMM calibrations is not significant after Holm adjustment ($p = 0.170$). The Parkinson result therefore demonstrates target sensitivity but does not support replacing the primary return-based HMM. Supporting Information Tables S27–S29 report the Parkinson-HMM acceptance, common-support, and target-aligned sensitivity analyses.

4.5 Sector Heterogeneity

Sector-level results reveal substantial variation in forecast rankings. Figure 1 visualises these cross-sector differences.

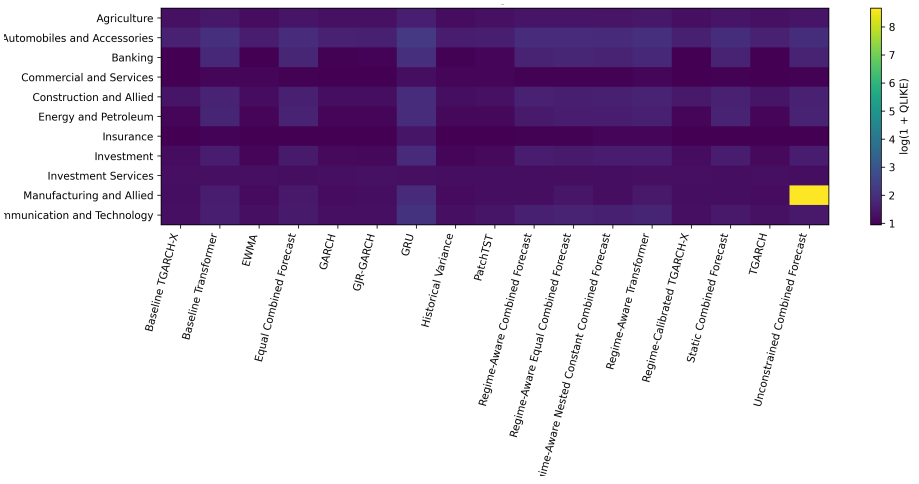


Figure 1: Sector-Level QLIKE Across Forecasting Models

Note. The heatmap displays $\log(1 + \text{QLIKE})$ to improve visual comparability across models. Lower values indicate better forecast accuracy.

Forecast rankings vary substantially across sectors, with no specification consistently dominating. Econometric models remain competitive, regime effects differ across models and sectors, and neither Transformer specification consistently outperforms. This heterogeneity reinforces the aggregate finding that regime information has selective rather than uniform value. Supporting Information Tables S11–S12 and Figure S4 provide the corresponding sector-level results.

4.6 Robustness and Economic Significance

The analysis also assesses the stability of the main findings across alternative specifications. Table 7 summarises remaining robustness checks for portfolio weighting, regime structure, neural sequence length, TGARCH-X specification, and trading-activity construction. Section 4.4 reports Parkinson separately.

Equal-sector and capped-capitalisation weighting produce only small reversals in the TGARCH-X calibration difference. At the same time, the three-state HMM improves QLIKE in the low and high states but not the medium state. The 20-day Transformer remains best among the re-estimated sequence lengths. Sector-volume lag 1 leaves TGARCH-X QLIKE virtually unchanged. Gaussian innovations improve QLIKE after global Holm correction, GED worsens it, and the 250-observation improvement is not significant after family-wise adjustment. Supporting Information Tables S17 and S19–S22 report the detailed weighting, regime-state, sequence, TGARCH-X sensitivity, and residual-diagnostic results.

Table 7: Robustness of the Main Forecasting Evidence

Robustness check	Comparison	QLIKE result	Difference	Inference
Equal-sector weights	TGARCH-X vs Reg-Cal TGARCH-X	2.3582 vs 2.3638	+0.0056	Small reversal
Capped-cap weights	TGARCH-X vs Reg-Cal TGARCH-X	2.1494 vs 2.1523	+0.0029	Small reversal
3-state HMM	TGARCH-X vs Reg-Cal TGARCH-X	Low/Med/High	−0.0119/ + 0.0099/ − 0.0120	State-dependent
Transformer sequence	10 / 20 / 40 days	4.4757 / 4.4178 / 4.6624	—	20-day sequence remains best
Gaussian innovations	vs Student-t anchor	2.0785 vs 2.1372	-0.0587	Global Holm p = 0.042
GED innovations	vs Student-t anchor	2.7665 vs 2.1372	+0.6293	Global Holm p < 0.001
125-observation window	vs 200 anchor	2.4025 vs 2.1372	+0.2653	Global Holm p = 0.373
250-observation window	vs 200 anchor	2.0731 vs 2.1372	-0.0642	Global Holm p = 0.076
Sector-volume change lag 1	vs turnover change lag 1	2.1371 vs 2.1372	-0.0002	Global Holm p = 0.988
Turnover change lag 2	vs lag 1	2.2233 vs 2.1372	+0.0861	Global Holm p = 0.250

Note. TGARCH-X sensitivity comparisons use common FinalTest support and 5,000 moving-block-bootstrap replications. Global Holm p-values adjust across all six alternative TGARCH-X specifications.

The economic-significance analysis finds no reliable portfolio advantage from regime information. None of the matched Sharpe differences has a 95% block-bootstrap interval excluding zero. For TGARCH-X, the difference is −0.0181 [−0.0441, 0.0103]. Regime calibration reduces mean daily one-way turnover by about 3.12%, but this does not reverse the risk-adjusted ranking after costs. At 10 basis points, net Sharpe is 1.854 for baseline versus 1.839 for regime-calibrated TGARCH-X. At 5 basis points, the corresponding values are 1.991 and 1.974. Lower forecast loss therefore does not necessarily translate into superior portfolio outcomes. Supporting Information Tables S23, S30, and S31 report the gross volatility-targeting diagnostics, matched Sharpe-ratio inference, and turnover and transaction-cost sensitivity.

5 Discussion

The central finding is that regime information has selective, not universal, forecasting value. Applying the same filtered high-volatility probability across three model classes produces different outcomes. Regime calibration yields a small TGARCH-X improvement on the primary squared-return target, while the probability-aware Transformer and probability-dependent combination underperform their baselines. The TGARCH-X differences remain below their 80% minimum detectable effects,

so non-significance alone does not establish a zero effect. However, equivalence testing bounds the full-sample difference within ±0.5% of baseline QLIKE and the state-specific differences within ±1%, supporting a small overall effect in practice. The TGARCH-X results suggest that latent-state information is more useful as a calibration signal than as a source of additional model flexibility. This interpretation is consistent with [19], who report market- and horizon-dependent performance, and [30], who find that latent-state contributions vary across markets and model configurations. The present study extends this evidence by holding the regime signal constant across matched econometric, neural, and combined forecasts. Moderation results remain qualified, as no model-by-regime interaction survives family-wise adjustment. The aggregate rankings also reinforce the value of parsimony. EWMA records the lowest market-capitalisation-weighted QLIKE, while GARCH and TGARCH remain competitive and PatchTST is the strongest neural benchmark. The weak performance of the Transformer specifications is consistent with [26], who finds no consistent forecasting gain from an additional Transformer pathway. In contrast, [20] report improved volatility forecasts from architectures combining Transformer components with other models. These contrasting results indicate that greater architectural complexity does not guarantee superior forecast accuracy across data environments. The dynamic forecast combination provides similar evidence. Allowing the combination weight to vary with the filtered regime probability increases QLIKE relative to the static combination, while the optimiser reaches its parameter bounds in five sectors. With only 298 StackTrain dates, the additional state-dependent flexibility may be difficult to estimate reliably. The result therefore supports a more restrained use of regime-dependent weighting when the calibration sample is limited. The regime diagnostics provide a separate methodological insight for less-liquid markets. Conventional continuous Gaussian and Student-t HMMs assign exact zero returns disproportionately to the low-volatility state in several sectors, potentially conflating low volatility with unchanged prices. The zero-neutral Student-t HMM removes exact zeros from state discrimination and uses non-zero returns to identify the continuous state scales. This treatment is relevant because zero-return incidence contains information about trading frictions in less-liquid markets [16]. The resulting filtered probability remains available before each forecast origin and enters each regime-informed model without replacing its underlying forecasting structure. Robustness results qualify rather than overturn the central interpretation. Regime calibration improves TGARCH-X materially under Parkinson variance, unlike squared returns. The high-low range captures intraday price dispersion and may therefore reflect latent state changes that close-to-close squared returns miss, although sparse trading can also weaken range-based measures. A Parkinson-based HMM does not significantly outperform the primary return-based HMM, so the sensitivity does not support replacing the primary regime extraction. Innovation distribution also affects TGARCH-X accuracy, while alternative trading-activity measures do not. The economic-significance results remain weak because

the TGARCH-X Sharpe difference has a 95% interval containing zero. Regime calibration lowers turnover by about 3.12% but does not reverse the risk-adjusted ranking after transaction costs. Improved forecast loss therefore does not necessarily imply a superior portfolio outcome. Overall, the findings support a conditional view of regime-aware volatility forecasting. Regime information does not automatically improve forecasts and appears most useful as a probabilistic calibration signal for an already competitive econometric model. Its contribution depends on the forecasting channel, market state, volatility target, and specification. The study therefore identifies the conditions under which a common regime signal adds predictive value rather than assuming that regime awareness is universally beneficial.

6 Limitations

Several limitations warrant consideration when interpreting the findings. Squared sector returns provide a noisy daily volatility target, although the Parkinson robustness analysis partly addresses this limitation. The study also focuses on one less-liquid equity market with uneven sector histories, which may constrain generalisability. The primary HMM assumes two latent states and applies study-specific zero-return validity thresholds, while the dynamic forecast combination relies on a relatively short StackTrain sample. Future research could extend the design to other markets, longer horizons, richer realised-volatility measures, and larger calibration samples.

7 Conclusion

This study examines when regime information improves volatility forecasting and whether its value depends on the model through which it enters. Matched econometric, neural, and combined forecasts show that regime information provides selective rather than universal predictive value. Simple econometric benchmarks remain difficult to outperform on the primary squared-return target, while regime calibration produces only a small TGARCH-X improvement. Although MDE analysis prevents interpreting non-significance as proof of a zero effect, equivalence testing supports a small overall TGARCH-X effect in practice. By contrast, adding the filtered regime probability to the Transformer and allowing combination weights to vary with it increases forecast loss. The zero-neutral Student-t HMM also shows why regime construction matters in thinly traded markets by preventing exact zeros from mechanically defining the low-volatility state. Robustness analysis qualifies the primary conclusion because TGARCH-X improves materially under Parkinson variance, although a Parkinson-based HMM does not outperform the primary return-based regime extraction. Forecast accuracy also varies with innovation distribution and estimation window, while the economic-significance analysis finds no reliable portfolio advantage from regime calibration. Overall, the evidence supports using regime

information selectively as a probabilistic calibration signal rather than assuming that additional regime-aware complexity improves every forecasting architecture.

Disclosure Statement

The authors declare no conflict of interest in relation to this manuscript.

Funding

The authors received no funding for this study.

Data Availability Statement

The data, code, model outputs, and reproducibility materials supporting the findings of this study are openly available at https://github.com/MWANIKID/When-Does-Regime-Information-Improve-Volatility-Forecasting.

Ethics Statement

Not applicable. The study uses secondary financial market data and does not involve human participants or animals.

Bibliography

[1] Y. Amihud, *Illiquidity and stock returns: Cross-section and time-series effects*, Journal of Financial Markets, (2002) vol. 5 no. 1 pp. 31–56. https://doi.org/10.1016/S1386-4181(01)00024-6

[2] J. M. Bates, and C. W. J. Granger, *The combination of forecasts*, Operational Research Quarterly, (1969) vol. 20 no. 4 pp. 451–468. https://doi.org/10.2307/3008764

[3] T. Bollerslev, *Generalized autoregressive conditional heteroskedasticity*, Journal of Econometrics, (1986) vol. 31 no. 3 pp. 307–327. https://doi.org/10.1016/0304-4076(86)90063-1

[4] A. C. Cameron, J. B. Gelbach, and D. L. Miller, *Robust inference with multiway clustering*, Journal of Business & Economic Statistics, (2011) vol. 29 no. 2 pp. 238–249. https://doi.org/10.1198/jbes.2010.07136

[5] F. X. Diebold, and R. S. Mariano, *Comparing predictive accuracy*, Journal of Business & Economic Statistics, (1995) vol. 13 no. 3 pp. 253–263. https://doi.org/10.1080/07350015.1995.10524599

[6] K. Y. L. Fong, C. W. Holden, and C. A. Trzcinka, *What are the best liquidity proxies for global research?*, Review of Finance, (2017) vol. 21 no. 4 pp. 1355–1401. https://doi.org/10.1093/rof/rfx003

[7] R. Giacomini, and B. Rossi, *Forecast comparisons in unstable environments*, Journal of Applied Econometrics, (2010) vol. 25 no. 4 pp. 595–620. https://doi.org/10.1002/jae.1177

[8] R. Giacomini, and H. White, *Tests of conditional predictive ability*, Econometrica, (2006) vol. 74 no. 6 pp. 1545–1578. https://doi.org/10.1111/j.1468-0262.2006.00718.x

[9] L. R. GLOSTEN, R. JAGANNATHAN, AND D. E. RUNKLE, *On the relation between the expected value and the volatility of the nominal excess return on stocks*, THE JOURNAL OF FINANCE, (1993) vol. 48 no. 5 pp. 1779–1801. https://doi.org/10.1111/j.1540-6261.1993.tb05128.x

[10] M. HAAS, S. MITTNIK, AND M. S. PAOLELLA, *A new approach to Markov-switching GARCH models*, JOURNAL OF FINANCIAL ECONOMETRICS, (2004) vol. 2 no. 4 pp. 493–530. https://doi.org/10.1093/jjfinec/nbh020

[11] J. D. HAMILTON, *A new approach to the economic analysis of nonstationary time series and the business cycle*, ECONOMETRICA, (1989) vol. 57 no. 2 pp. 357–384. https://doi.org/10.2307/1912559

[12] P. R. HANSEN, AND A. LUNDE, *A forecast comparison of volatility models: Does anything beat a GARCH(1,1)?*, JOURNAL OF APPLIED ECONOMETRICS, (2005) vol. 20 no. 7 pp. 873–889. https://doi.org/10.1002/jae.800

[13] P. R. HANSEN, A. LUNDE, AND J. M. NASON, *The model confidence set*, ECONOMETRICA, (2011) vol. 79 no. 2 pp. 453–497. https://doi.org/10.3982/ECTA5771

[14] D. HARVEY, S. LEYBOURNE, AND P. NEWBOLD, *Testing the equality of prediction mean squared errors*, INTERNATIONAL JOURNAL OF FORECASTING, (1997) vol. 13 no. 2 pp. 281–291. https://doi.org/10.1016/S0169-2070(96)00719-4

[15] F. KLAASSEN, *Improving GARCH volatility forecasts with regime-switching GARCH*, In J. D. Hamilton and B. Raj (eds.), *Advances in Markov-Switching Models*, PHYSICA-VERLAG HD, (2002) pp. 223–254. https://doi.org/10.1007/978-3-642-51182-0_10

[16] D. A. LESMOND, J. P. OGDEN, AND C. A. TRZCINKA, *A new estimate of transaction costs*, THE REVIEW OF FINANCIAL STUDIES, (1999) vol. 12 no. 5 pp. 1113–1141. https://doi.org/10.1093/rfs/12.5.1113

[17] R. M. LEUSHUIS, AND N. PETKOV, *Advances in forecasting realized volatility: A review of methodologies*, FINANCIAL INNOVATION, (2026) vol. 12 no. 1 Article 14. https://doi.org/10.1186/s40854-025-00809-5

[18] B. LIM, S. Ö. ARIK, N. LOEFF, AND T. PFISTER, *Temporal Fusion Transformers for interpretable multi-horizon time series forecasting*, INTERNATIONAL JOURNAL OF FORECASTING, (2021) vol. 37 no. 4 pp. 1748–1764. https://doi.org/10.1016/j.ijforecast.2021.03.012

[19] R. LIU, M. SEGNON, O. CEPNI, AND R. GUPTA, *Forecasting volatility of commodity, currency, and stock markets: Evidence from Markov-switching multifractal models*, JOURNAL OF FORECASTING, (2026) vol. 45 no. 6 pp. 2905–2941. https://doi.org/10.1002/for.70145

[20] A. K. MISHRA, J. RENGANATHAN, AND A. GUPTA, *Volatility forecasting and assessing risk of financial markets using multi-transformer neural network based architecture*, ENGINEERING APPLICATIONS OF ARTIFICIAL INTELLIGENCE, (2024) vol. 133 108223. https://doi.org/10.1016/j.engappai.2024.108223

[21] F. MORENO-PINO, AND S. ZOHREN, *DeepVol: Volatility forecasting from high-frequency data with dilated causal convolutions*, QUANTITATIVE FINANCE, (2024) vol. 24 no. 8 pp. 1105–1127. https://doi.org/10.1080/14697688.2024.2387222

[22] N. NAIFAR, *Do supply-chain stress and geopolitical risk predict strategic commodity and clean energy market returns? Evidence from explainable machine learning*, FORECASTING, (2026) vol. 8 no. 4 Article 59. https://doi.org/10.3390/forecast8040059

[23] Y. NIE, N. H. NGUYEN, P. SINTHONG, AND J. KALAGNANAM, *A time series is worth 64 words: Long-term forecasting with Transformers*, THE ELEVENTH INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS, (2023). https://openreview.net/forum?id=Jbdc0vTOcol

[24] M. PARKINSON, *The extreme value method for estimating the variance of the rate of return*, THE JOURNAL OF BUSINESS, (1980) vol. 53 no. 1 pp. 61–65.

[25] A. J. PATTON, *Volatility forecast comparison using imperfect volatility proxies*, JOURNAL OF ECONOMETRICS, (2011) vol. 160 no. 1 pp. 246–256. https://doi.org/10.1016/j.jeconom.2010.03.034

[26] M. Pinitjitsamut, *Multi-scale forecasting of natural rubber prices using VMD-augmented BiLSTM: A hybrid architecture ablation study*, Forecasting, (2026) vol. 8 no. 3 Article 43. https://doi.org/10.3390/forecast8030043

[27] Z. Qiu, C. Kownatzki, F. Scalzo, and E. S. Cha, *Historical perspectives in volatility forecasting methods with machine learning*, Risks, (2025) vol. 13 no. 5 Article 98. https://doi.org/10.3390/risks13050098

[28] D. Roodman, M. Ø. Nielsen, J. G. MacKinnon, and M. D. Webb, *Fast and wild: Bootstrap inference in Stata using boottest*, The Stata Journal: Promoting Communications on Statistics and Stata, (2019) vol. 19 no. 1 pp. 4–60. https://doi.org/10.1177/1536867X19830877

[29] M. Saidane, and E. Ben Mohamed, *Forecasting regime-dependent tail risk in digital asset portfolios with a mixture of hidden Markov factor analyzers*, Journal of Forecasting, (2026) vol. 45 no. 6 pp. 2735–2759. https://doi.org/10.1002/for.70159

[30] C. Tian, R. Tansuchat, and S. Sriboonchitta, *Regime-aware stock index forecasting under latent market states: A hybrid statistical learning framework with cross-market validation*, Forecasting, (2026) vol. 8 no. 3 Article 50. https://doi.org/10.3390/forecast8030050

[31] A. Timmermann, *Forecast combinations*, In G. Elliott, C. W. J. Granger and A. Timmermann (eds.), *Handbook of Economic Forecasting*, Elsevier, (2006) vol. 1 pp. 135–196. https://doi.org/10.1016/1574-0706(05)01004-9

[32] Y. Xiao, C. He, M. Yi, and Y. Hu, *Adaptive model integration for stock price forecasting with CNN-ATTN-BiLSTM*, Journal of Forecasting, (2026) vol. 45 no. 4 pp. 1331–1349. https://doi.org/10.1002/for.70098

[33] J.-M. Zakoian, *Threshold heteroskedastic models*, Journal of Economic Dynamics and Control, (1994) vol. 18 no. 5 pp. 931–955. https://doi.org/10.1016/0165-1889(94)90039-6

[34] H. Zhao, M. Crane, and M. Bezbradica, *MSPCIFormer: A multi-scale patching channel-independent Transformer for cryptocurrency price forecasting*, Expert Systems, (2026) vol. 43 no. 9 e70343. https://doi.org/10.1111/exsy.70343

How to Cite: Daniel Ndirangu Mwaniki[1], Christine Nanjala Simiyu[2], Jacob Otieno Ong'ala[3], *When Does Regime Information Improve Volatility Forecasting? Evidence from TGARCH-X, Transformer, and Combined Models*, Journal of Mathematics and Modeling in Finance (JMMF), Vol. 7, No. 1, Pages:327–351, (2027).

The Journal of Mathematics and Modeling in Finance (JMMF) is licensed under a Creative Commons Attribution NonCommercial 4.0 International License.