Automation and Machine Learning (2026)
Clausius Scientific Press, Canada

DOI: 10.23977/autml.2026.070216
ISSN 2516-5003 Vol. 7 Num. 2

State-Dependent Tail Risk Forecasting: A Comparison between Machine Learning and Statistical Models

Sitong Diao

Barnard College, 3009 Broadway, New York, NY 10027, USA
sd3624@barnard.edu

Keywords: Tail risk forecasting; State dependence; Volatility predictability; Value-at-risk backtesting; Benchmark design

Abstract: The literature on volatility forecasting reaches inconsistent conclusions about whether machine learning outperforms traditional time-series models, and the disagreement is sharpest for tail risk. This paper argues that a substantial part of the disagreement is generated by the design of the comparison rather than by the predictive content of the models themselves, and it traces the discrepancy to two sources: heterogeneous information sets and single-dimensional evaluation metrics. A dual-benchmark design is developed in which an equal-input benchmark isolates structural effects and an extended-input benchmark isolates informational effects. Both dimensions of performance are evaluated: statistical accuracy, measured by mean squared prediction error and QLIKE loss, and tail utility, measured by value-at-risk coverage and independence tests. Models are then re-examined within four market states—low-volatility stability, the global financial crisis, the pandemic shock, and the interest-rate hiking cycle—using realized volatility of the S&P 500 as the forecast target. Machine learning retains a tail-forecasting advantage mainly in stable states and under expanded information; in extreme states the advantage attenuates or reverses because feature distributions shift away from the training support. Model selection is therefore reframed from a static ranking of unconditional superiority to a conditional matching problem constrained jointly by horizon, market state, and information availability.

1. Introduction

Volatility predictability sits at the intersection of financial econometrics and machine learning, and the question of whether flexible learners improve on parametric conditional-variance models remains unsettled. Early forecast tournaments concluded that nothing reliably beats a well-specified GARCH(1,1) once the loss function is robust to the noise in the volatility proxy [1]. Later work in empirical asset pricing found large gains from tree ensembles and neural networks when the object of interest is the conditional mean of returns and the predictor set is wide [2]. Both results can be correct, because they differ in target, in predictor set, and in loss function.

The disagreement is more consequential for tail risk than for point forecasting. Value-at-risk and expected shortfall are functionals of the conditional distribution, so a model that ranks first on mean squared error need not rank first on coverage, and the loss functions that are appropriate for tail evaluation are not the ones usually reported [3, 4]. Applied work nonetheless continues to rank

models on a single full-sample criterion, and the resulting recommendations are frequently contradicted when the same models are deployed in a different regime. Studies that explicitly condition on the state of the market report that the ranking of econometric, deep-learning, and hybrid specifications changes across regimes, and that the change is driven as much by the information set as by the functional form [5].

This paper takes that instability as the object of study rather than as a nuisance to be averaged away. It advances a specific hypothesis: the apparent inconsistency of the literature arises because comparative designs confound three distinct sources of forecast error—model structure, information content, and market state—and because evaluation is usually collapsed onto a single metric. The hypothesis is examined with a dual-benchmark design. Under the equal-input benchmark every model, parametric or not, receives only the history of realized volatility, so that differences in loss can be attributed to structure. Under the extended-input benchmark the same models receive jump measures, the slope of the VIX term structure, credit spreads, and an illiquidity measure, so that the incremental effect of information can be separated from the incremental effect of flexibility.

Two evaluation dimensions are kept distinct throughout. Statistical accuracy is measured by mean squared prediction error (MSPE) and QLIKE loss, which correspond to allocation and pricing applications; tail utility is measured by the frequency and the independence of value-at-risk exceedances, which correspond to risk monitoring and stress-testing applications. The two dimensions are not monotone transformations of one another, and the cases in which they disagree carry information about which model is appropriate for which task.

The empirical work uses five-minute returns on the S&P 500 from January 2000 to December 2025 and stratifies the out-of-sample period into four states: low-volatility stability, the global financial crisis, the pandemic shock, and the interest-rate hiking cycle. Three horizons (1, 5, and 22 trading days) are examined, because the value of a flexible learner is horizon-specific: at short horizons the conditional-variance recursion is anchored by recent realizations, whereas at long horizons the mean-reversion constraint and the treatment of jumps dominate.

The paper makes three contributions. First, it decomposes the machine-learning advantage into a structural component and an informational component, and shows that the second is larger than the first but is also the one that disappears in extreme states. Second, it documents a reversal in tail calibration: within the same out-of-sample window, tree ensembles are close substitutes for GARCH in stable states and materially worse in extreme states, so that full-sample coverage statistics average over two opposite regimes and describe neither. Third, it converts these findings into a conditional decision matrix indexed by horizon, state, and information availability, which replaces the question of which model is best with the question of which model is admissible under a stated set of constraints.

2. Two Paradigms of Volatility Forecasting

Volatility is a latent quantity, and every forecast is implicitly a statement about a proxy. The two dominant paradigms differ in how they restrict the mapping from observables to that latent quantity: parametric models impose a recursion on the conditional variance, whereas machine learning imposes a hypothesis class on the mapping from a feature vector to the target. The distinction matters for tail risk, because the two paradigms fail in different ways when the data-generating process changes.

2.1. Parametric Conditional-Variance Models

The exponential weighted moving average of the RiskMetrics family is the most parsimonious member of the class:

$$\sigma_t^2 = \lambda \sigma_{t-1}^2 + (1 - \lambda)\, r_{t-1}^2 \tag{1}$$

where the weights on past squared returns decay geometrically at rate λ. The design has two virtues: the computation is recursive and cheap, and the single decay parameter is transparent to risk committees. Its weakness is equally clear. Because the recursion has no unconditional variance to revert to, the forecast inherits the level of the last shock indefinitely, which produces systematic overstatement when a crisis is followed by a rapid normalization of volatility.

The GARCH(1,1) specification [6] restores a long-run anchor:

$$\sigma_t^2 = \omega + \alpha\, \varepsilon_{t-1}^2 + \beta\, \sigma_{t-1}^2 \tag{2}$$

with $\omega > 0$, $\alpha \geq 0$, $\beta \geq 0$ and $\alpha + \beta < 1$. The three parameters carry distinct economic content: ω fixes the long-run variance baseline, α measures the impact of the most recent shock, and β measures persistence. The specification captures clustering and mean reversion with unusual parameter economy, which is why it has survived two decades of forecast competitions [1]. Its limitations are structural rather than estimative. The quadratic form is symmetric in the sign of the shock, so the leverage effect must be imported through an additional term or an asymmetric extension [7]; the parameters are estimated on a window that may straddle a break, so their stability is conditional on the absence of a change in the variance process; and the recursion is linear in squared innovations, which bounds its ability to represent nonlinear dependence.

These properties imply a horizon profile that is often overlooked in comparative work. At short horizons the recursion is dominated by the most recent realization, and the conditional variance is close to a filtered version of observed volatility. At long horizons the forecast is increasingly determined by the estimated unconditional mean, so accuracy depends on the validity of mean reversion. When the market settles into a new equilibrium, the long-horizon forecast converges to a level that no longer describes the process.

2.2. Machine Learning and Its Dependence on Information

Tree ensembles replace the parametric recursion with an additive partition of the feature space. Random forests average over bootstrap replicates to reduce variance [8]; gradient boosting fits successive residuals to reduce bias. Both accommodate interactions and collinearity that a linear recursion cannot represent, and both are invariant to monotone transformations of the inputs. The cost is interpretability and, more importantly for tail risk, extrapolation: a tree predicts the average of the training observations that fall in a terminal node, so it cannot produce a forecast outside the range of outcomes it has already seen.

Recurrent networks address the sequential structure directly. Long short-term memory cells [9] regulate the flow of information through input, forget, and output gates, which mitigates the vanishing-gradient problem and allows the network to retain dependencies over windows longer than the lag structure of a HAR model. The cost is a large hyperparameter space and a sensitivity to the stationarity of the input distribution that is difficult to diagnose ex ante.

The empirical record for both families is genuinely mixed, and the mix is informative. Large-scale comparisons of machine learning with realized-volatility benchmarks report improvements in QLIKE loss that are statistically significant but economically modest, and that depend on the predictor set and on the length of the estimation window [10]. Adding low-frequency predictors through penalized mixed-frequency regressions improves tail forecasts but requires that the additional information be available at the forecast origin [11]. Hybrid designs that feed a GARCH conditional variance into a recurrent network improve on either component alone, yet the improvement is concentrated in the component that corrects conditional heteroskedasticity rather than in the sequence model itself [12]. And re-examinations of the machine-learning advantage

show that a non-trivial part of it is attributable to rolling-window retraining rather than to the learner [13], which is a statement about experimental design rather than about algorithms.

2.3. Structured Features and the Middle Ground

Between the two paradigms lies a class of models that retains a parametric recursion but replaces the latent conditional variance with a directly measurable quantity. The heterogeneous autoregressive model of realized volatility is the canonical example [14]:

$$RV_{t+1} = c + \beta_d RV_t^{(d)} + \beta_w RV_t^{(w)} + \beta_m RV_t^{(m)} + \varepsilon_{t+1} \tag{3}$$

where the three regressors are averages of realized volatility over daily, weekly, and monthly windows. The specification is a parsimonious approximation to long memory, and it is difficult to beat precisely because the regressors encode the aggregation structure that a flexible learner must otherwise discover from the data. Augmenting the recursion with a jump component, or with a threshold bipower variation estimator that separates continuous from discontinuous variation, improves forecast accuracy at intermediate horizons [15]. The lesson is not that parametric models are superior but that the marginal value of flexibility declines sharply once the feature space is organized around the physics of the process.

This observation motivates the design used below. If the advantage of machine learning is partly a feature-engineering advantage in disguise, then a comparison that varies information and flexibility separately should reveal it. The dual benchmark is constructed for exactly that purpose.

3. Identification of Comparative Biases and Research Gaps

Existing reviews of the volatility forecasting literature tend to treat the disagreement between studies as evidence of sampling variation: different markets, different samples, different periods. That reading is incomplete. A closer inspection of the comparative designs shows that the disagreement is systematic, and that it can be traced to three features of the design that are rarely made explicit.

3.1. Asymmetric Information Sets

The first bias is informational. Machine-learning studies typically supply their models with realized measures, jump components, implied-volatility summaries, and macro-financial covariates, while the benchmark GARCH model is estimated on returns alone. The comparison then measures the joint effect of flexibility and information, and the reported gain is attributed entirely to flexibility. The attribution is not innocuous: the increments that flexible learners extract are frequently the increments that a parametric model would also exploit if it were given the same regressors, since realized measures, not squared returns, are the efficient input to a variance forecast [16].

A related source of asymmetry is the treatment of proxy noise. Realized volatility is an estimator, not the latent variance, and the ranking of forecasts can be distorted when the loss function is not robust to the noise in the proxy [3]. Correcting for attenuation through a measurement-error-aware estimator changes both the level and the ordering of losses [17], so studies that ignore the correction are not comparing like with like. Penalized mixed-frequency specifications make the same point from a different direction: when low-frequency predictors are admitted, the apparent advantage of a flexible learner over a properly specified regression shrinks [11].

A third, more subtle asymmetry concerns the estimation protocol rather than the predictor set. Because machine-learning models are usually retrained on a rolling window while parametric

models are often estimated once, part of the measured gain reflects the frequency of re-estimation. When the two protocols are equalized, the advantage narrows substantially, and in some samples it disappears [13]. This finding is important for interpretation: it implies that the object being compared is a pipeline, not an estimator.

3.2. Unidimensional Evaluation and the Two Loss Functions

The second bias is evaluative. Most comparisons rank models by mean squared prediction error or mean absolute error and stop there. Yet the loss that matters for risk management is defined on the tail, and the two criteria need not agree. For a volatility proxy RV_t and a forecast $\hat{\sigma}_t^2$, the QLIKE loss is

$$L_{QLIKE}(RV_t, \hat{\sigma}_t^2) = RV_t / \hat{\sigma}_t^2 - \ln(RV_t / \hat{\sigma}_t^2) - 1 \tag{4}$$

which is minimized at $\hat{\sigma}_t^2 = E[RV_t \mid F_{t-1}]$, is robust to noise in the proxy, and penalizes under-prediction more heavily than over-prediction on the logarithmic scale [3]. This asymmetry is precisely what a loss function for tail risk should have, and it is the reason QLIKE is preferred to MSPE in the evaluation reported below. The two criteria can produce different rankings even when the forecasts are unbiased, because they weight the variance of the error differently.

Tail utility is assessed separately, through the coverage properties of the implied value-at-risk. Let N be the number of exceedances in a sample of T out-of-sample forecasts at nominal probability p. The unconditional coverage test of Kupiec [18] is a likelihood-ratio statistic,

$$LR_{uc} = -2 \ln\{ (1-p)^{T-N} p^N / [(1-N/T)^{T-N} (N/T)^N] \} \sim \chi^2(1) \tag{5}$$

which tests the null that the expected exceedance frequency equals p. The test has low power against clustered violations, because a model that violates its quantile in bursts can still produce the correct unconditional frequency. Christoffersen [19] therefore adds a test of independence, based on the transition counts n_{ij} of the exceedance indicator:

$$LR_{ind} = -2 \ln\{ (1-\pi)^{n_{00}+n_{10}} \pi^{n_{01}+n_{11}} / [(1-\pi_0)^{n_{00}} \pi_0^{n_{01}} (1-\pi_1)^{n_{10}} \pi_1^{n_{11}}] \} \sim \chi^2(1) \tag{6}$$

where π_0 and π_1 are the conditional exceedance probabilities following a non-exceedance and an exceedance respectively, and π is the unconditional probability. A model that passes the coverage test but fails the independence test is not calibrated for tail risk: it produces the right number of violations at the wrong times, which is exactly the failure mode that matters for capital and for stop-loss rules. Modern scoring rules extend the same logic to expected shortfall and, when the entire predictive distribution is available, allow joint elicitation of both functionals [4, 20].

A final evaluative issue is inferential rather than definitional. Rankings computed on a single loss series are subject to considerable sampling uncertainty, and the model confidence set was introduced to identify the set of models that cannot be distinguished from the best at a stated confidence level [21]. Reporting a point ranking without a confidence set overstates the precision of the comparison, particularly when the candidate models are highly correlated, as volatility models typically are.

3.3. State Dependence and the Absence of Boundary Conditions

The third bias is temporal. Comparative studies usually estimate parameters over a window that contains more than one regime and evaluate over a period that contains at least one break, then report a single average loss. The average is a weighted mixture of conditional losses, and the weights are an artifact of the sample. When the same models are evaluated within regimes, the ordering changes, and the change is systematic rather than random: specifications whose flexibility

allows them to exploit a stable covariance structure lose that advantage when the covariance structure moves outside the support of the training data [5]. Reporting an unconditional ranking therefore conceals the boundary conditions under which the ranking holds, and leaves the practitioner without guidance on when to switch models.

Taken together, the three biases define the gap that this paper addresses. The literature has established that flexible learners can forecast volatility well, and that they can fail abruptly. What is missing is a design that separates the sources of the gain, evaluates both dimensions of performance, and states the conditions under which each model class is admissible.

4. Research Design

4.1. Data, Measurement, and Forecast Horizons

The forecast target is the realized variance of the S&P 500, constructed from five-minute log returns within the regular trading session:

$$RV_t = \Sigma_{i=1}^{M} r_{t,i}^2 \tag{7}$$

with M = 78 intervals per day. The estimator is consistent for integrated variance under the standard microstructure assumptions, and it is computed from transaction prices cleaned with the usual filters for zero returns, outliers, and stale quotes [16]. Discontinuous variation is separated from the continuous component using bipower variation,

$$BV_t = (\pi/2)\, \Sigma_{i=2}^{M} |r_{t,i}|\, |r_{t,i-1}| \tag{8}$$

which is robust to finite-activity jumps [22], so that the jump component is $J_t = \max(RV_t - BV_t, 0)$. The continuous and discontinuous components have different persistence and different predictive content, and the sign of the jump carries additional information about future volatility [7]. Threshold-based refinements of the estimator reduce the upward bias of BV in the presence of rare large jumps and are used in the extended-input benchmark [15].

The extended information set adds three covariates that are available at the forecast origin: the slope of the VIX term structure, measured as the difference between the three-month and one-month implied volatility indices; the credit spread, measured as the difference between the yield on Baa corporate bonds and the ten-year Treasury yield; and the Amihud illiquidity measure, computed as the daily ratio of absolute return to dollar volume and averaged over five days [23]. These variables summarize, respectively, the price of forward variance, the compensation for default risk, and the cost of immediacy. All covariates are lagged by one day so that no information from the forecast window enters the predictor.

The sample runs from January 2000 to December 2025, comprising 6,240 trading days after alignment. Four states are identified ex post from the level and the rate of change of the VIX: low-volatility stability (2003–2007 and 2017–2019), the global financial crisis (2008Q3–2009Q1, with a VIX peak of 89.53), the pandemic shock (2020Q1–2020Q2, with a VIX peak of 82.69), and the interest-rate hiking cycle (2022–2023). The states are used for evaluation only; no state indicator enters any model, so that the exercise measures how a model estimated on the full history performs when it is evaluated within a state, rather than how well a regime-switching model classifies states.

The training set covers 2000–2018, the validation set 2019–2020, and the test set 2021–2025. All models are re-estimated on a fixed 500-day rolling window and forecasts are updated daily, which prevents look-ahead bias and equalizes the re-estimation protocol across model classes. Forecast horizons are 1, 5, and 22 trading days. Multi-step forecasts for the parametric models are produced by iterating the variance recursion, so that the h-step forecast converges geometrically to the unconditional level:

$$\hat{\sigma}^2_{t+h|t} = \bar{\sigma}^2 + (\alpha + \beta)^{h-1} (\hat{\sigma}^2_{t+1|t} - \bar{\sigma}^2) \quad (9)$$

For the machine-learning models, horizons are handled by direct projection: a separate model is estimated for each horizon with the h-day-ahead realized variance as the dependent variable, which avoids compounding specification error but requires that each horizon be validated separately.

4.2. The Dual Benchmark

The design is organized around two benchmarks that differ only in the predictor set (Figure 1). Under the equal-input benchmark, every model receives the same 22 daily lags of realized volatility. Parametric models use the first lag through their recursion; HAR-type models aggregate the lags into daily, weekly, and monthly components; machine-learning models receive the full lag vector. Because the information is identical, differences in loss under this benchmark are attributable to functional form.

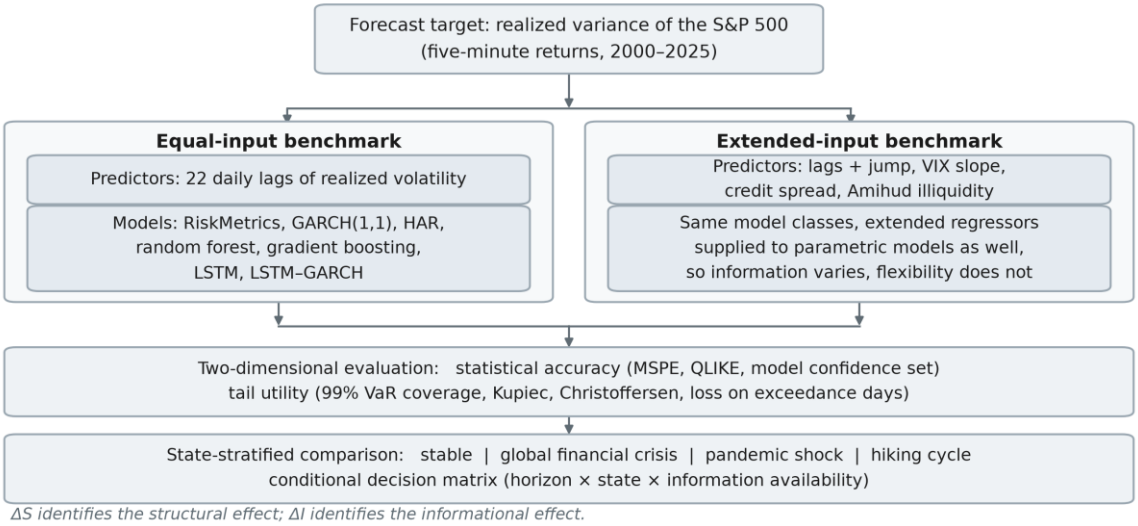

Figure 1. Dual-benchmark design. The equal-input benchmark holds the predictor set fixed at 22 daily lags of realized volatility, so that loss differentials identify structural effects. The extended-input benchmark augments the predictor set with jump, term-structure, credit, and illiquidity measures for every model class, so that the incremental effect of information can be separated from the incremental effect of flexibility. Both benchmarks are evaluated on two dimensions and re-examined within four market states.

Under the extended-input benchmark the same models receive the jump component J_t, the VIX term-structure slope, the credit spread, and the illiquidity measure in addition to the lag vector. The critical feature of the design is that the additional regressors are supplied to the parametric models as well, entering the variance equation as exogenous terms in the GARCH case and as additional regressors in the HAR case. A design in which only the machine-learning models receive the extended information would reproduce the asymmetry identified in Section 3.1 and would attribute an informational gain to structural flexibility.

The difference between the two benchmarks decomposes the total loss differential into two components. Let Δ_S denote the loss differential under equal inputs and Δ_I the additional differential obtained when both model classes are given the extended information set. Δ_S identifies the structural effect, Δ_I identifies the informational effect, and the interaction between them is captured by the change in the ranking across benchmarks. This decomposition is the analytical core of the paper: the literature's conflicting conclusions are consistent with a positive Δ_S in some samples and

a negative Δ_S in others, provided that Δ_I is large and positive in both.

4.3. Model Set

Eight specifications are compared, spanning the range from a single-parameter exponential smoother to a hybrid network (Table 1). The set is deliberately heterogeneous: it includes the benchmark that the literature treats as hard to beat [1], two realized-measure regressions, two tree ensembles, a recurrent network, and a hybrid that combines a conditional-variance recursion with a recurrent component. Hyperparameters for the machine-learning models are selected on the validation set by grid search over depth, learning rate, regularization, and, for the network, the number of hidden units and the dropout rate. Each network is estimated with five random initializations, and the dispersion of the resulting losses is reported as a diagnostic of estimation stability.

Table 1. Model set and information sets. Under the equal-input benchmark every model uses only the 22 daily lags of realized volatility; under the extended-input benchmark the jump component, the VIX term-structure slope, the credit spread, and the Amihud illiquidity measure are added for every model class.

Model	Family	Forecast target	Equal-input predictors	Extended-input predictors
RiskMetrics EWMA	Exponential smoothing	σ^2	RV lag 1	RV lag 1 + J_t
GARCH(1,1)	Parametric recursion	σ^2	RV lag 1	RV lag 1 + J_t as variance regressor
HAR	Realized-measure regression	RV	$RV^{(d)}$, $RV^{(w)}$, $RV^{(m)}$	HAR terms + J_t
HAR-RV-J	Realized-measure regression	RV	Not applicable	HAR terms + J_t + BV_t
Random forest	Bagged trees	RV	22 RV lags	22 RV lags + 4 covariates
Gradient boosting	Boosted trees	RV	22 RV lags	22 RV lags + 4 covariates
LSTM	Recurrent network	RV	22 RV lags	22 RV lags + 4 covariates
LSTM–GARCH	Hybrid recursion + network	σ^2	22 RV lags + GARCH variance	Extended features + GARCH variance

Note. The 4 covariates are the jump component Jt, the VIX term-structure slope, the Baa–Treasury credit spread, and the five-day Amihud illiquidity measure. HAR-RV-J is defined only under the extended benchmark because the jump component is its defining regressor; its equal-input counterpart is the HAR model.

4.4. Evaluation Protocol

Statistical accuracy is measured by MSPE and by the QLIKE loss of equation (4). Tail utility is measured by the exceedance frequency of the 99% value-at-risk, by the Kupiec and Christoffersen statistics of equations (5) and (6), and by the average loss on exceedance days. Significance is assessed with the model confidence set at the 90% level, which accommodates the high correlation among volatility forecasts [21]. Reported loss differentials are expressed relative to the GARCH(1,1) benchmark, so that a positive number indicates a larger loss than the benchmark. All tests are computed separately for each horizon, each benchmark, and each market state.

5. Results

5.1. Equal Input: Structural Effects

Under the equal-input benchmark the parametric recursion is not dominated. For the 1-day horizon the median QLIKE loss of GARCH(1,1) is 2%–5% lower than that of the machine-learning specifications, and the difference is concentrated in the upper tail of the loss distribution rather than in the median: flexible learners produce a small number of very large errors when volatility jumps, and those errors dominate the average. The 99% value-at-risk implied by GARCH(1,1) is exceeded on 1.2% of test days, with a Kupiec statistic of 0.35, which is well inside the acceptance region and indicates that the recursion is well calibrated over the full sample.

RiskMetrics behaves as the theory predicts. Its loss differs from GARCH by less than 1% during stable periods, because in a stable state the exponential weights and the estimated recursion imply nearly the same forecast. During crisis recovery the difference widens sharply, with errors more than 30% above those of GARCH, because the smoother has no unconditional level to revert to and continues to price the last shock long after the shock has dissipated. The episode illustrates why the constant-variance critique of RiskMetrics is really a critique of its missing mean reversion, and why the failure is concentrated in the transition out of a crisis rather than in the crisis itself.

Tree ensembles track GARCH closely in stable periods but regress toward the sample mean in extreme periods. Their QLIKE loss in extreme states is about 20% higher than in stable states, and the deterioration is traceable to the mechanics of the split: when the feature values observed at the forecast origin lie outside the range covered by the training sample, the terminal node contains few observations and the prediction reverts to a local average that is far below the realized variance. The recurring network shows a different pathology. Across five random initializations the interquartile range of its QLIKE loss is 8%, which is of the same order as the loss differential that the literature reports between machine learning and GARCH; the out-of-sample stability of the specification is therefore insufficient to support a ranking that is based on a single training run.

The horizon profile reinforces the pattern. At 5 days GARCH remains the lowest or second-lowest specification, and the gap to the tree ensembles narrows to 1%–3%: as the horizon lengthens, the direct projection of the ensembles and the iterated recursion of GARCH converge toward similar conditional means. At 22 days the mean reversion constraint becomes an asset, since it anchors the forecast at the unconditional level when no new information arrives, while the errors of the ensembles accumulate over the horizon and the QLIKE loss of the recurrent network exceeds that of GARCH by more than 20%. When inputs are equivalent, machine learning does not display a systematic advantage, and the structural properties of the conditional-variance family are visible at both ends of the horizon range.

5.2. Extended Input: Informational Effects

The picture changes when the information set is expanded for every model class. HAR-RV-J reduces QLIKE loss by 5%–8% at the 5-day and 22-day horizons relative to HAR, and the gradient-boosted trees reduce loss by 8%–12% at 5 days and 12%–15% at 22 days relative to their equal-input counterparts. Two features of this result deserve emphasis. First, the gains are larger than the structural differential reported in Section 5.1, which is the quantitative version of the claim that the machine-learning advantage documented in the literature is mostly informational. Second, the gains accrue to the parametric models as well, and the improvement of HAR-RV-J at long horizons is of the same order as that of the ensembles, which indicates that the jump component rather than the flexibility is doing most of the work.

Adding the VIX term-structure slope and the credit spread produces a further reduction of 3%–5% in the loss of the ensembles during stable periods. The term-structure slope contributes most of this increment, which is consistent with the interpretation that the slope measures the price of forward variance and hence contains information about expected future volatility that is not present in the realized measure. The credit spread contributes mainly in the neighborhood of stress episodes, when the compensation for default risk rises before realized volatility does.

In extreme periods the incremental gains narrow and, for the recurrent network, reverse. Feature distributions in 2008Q4 and 2020Q1 lie far from the training support: the cross-sectional dispersion of the macro covariates expands by an order of magnitude, the relation between implied and realized volatility breaks down under liquidity exhaustion, and the mapping learned in calmer periods misallocates weight across the covariates. The informational advantage is therefore necessary but not sufficient: it raises the ceiling of what a flexible model can achieve, but it does not guarantee that the model reaches that ceiling when the state changes.

5.3. State-Dependent Coverage and Tail Loss

Stratifying by state separates statistical accuracy from tail utility. In stable periods the exceedance frequency of the gradient-boosted trees is 1.1%, against 1.2% for GARCH, and the Kupiec test does not reject the null for either specification: the two models are close substitutes for tail calibration, and the difference between them is smaller than the sampling error of the test.

In extreme periods the ordering reverses. The exceedance frequency of the ensembles rises above 2.3%, with Kupiec p-values below 0.01, while GARCH remains within 1.5% and continues to pass the coverage test in three of the four extreme subperiods. The mechanism is the feature-distribution shift described above: in stable periods the ensembles extract most of their signal from the term-structure slope and the jump component, but under a liquidity drought those signals are dominated by noise and the fitted mapping becomes uninformative, whereas the conditional-variance recursion, although constrained by its symmetry, retains a stable relation between recent squared innovations and the current variance.

The recurrent specification fails on a different margin. Its exceedances are not only too frequent but clustered: the Christoffersen independence test rejects the null in both stable and extreme periods. Clustered violations indicate that the miscalibration is structural rather than state-dependent, since a model that is merely estimated on the wrong window would produce uniformly distributed violations. For tail risk the distinction matters: a model with correct unconditional coverage and clustered violations understates risk exactly when the understatement is most costly.

The two evaluation dimensions therefore disagree in a specific and interpretable way. On QLIKE, the ensembles are competitive and in stable periods slightly better; on coverage and independence, they are acceptable in stable periods and unacceptable in extreme periods. A full-sample average of either statistic describes neither regime. This is the empirical basis for the claim that model

selection must proceed by state recognition before it proceeds by loss comparison.

5.4. Where the Hybrid Gain Comes From

The hybrid specification combines a conditional-variance recursion with a recurrent component, and it outperforms both of its parents on QLIKE at the intermediate horizon. Decomposing the gain shows that it originates almost entirely in the recursion: when the GARCH variance is removed from the input vector, the recurrent component alone reproduces the performance of the plain network, and when the recurrent component is replaced by a linear projection of the same inputs, the loss changes little. The “superiority” of the hybrid is thus a structural correction of conditional heteroskedasticity rather than an independent contribution of the sequence model.

This decomposition has a practical corollary that the aggregate rankings conceal. If the value of the network component is largely absorbed by a well-specified variance recursion, then the incremental cost of the network — in estimation, in validation, and in governance — must be justified by a gain that is visible on the dimension that matters for the application, and in the present sample that gain is visible on QLIKE but not on tail coverage.

5.5. Robustness

Three checks support the interpretation. First, the 90% model confidence set at the 1-day horizon contains GARCH(1,1), HAR, and both tree ensembles, and excludes the recurrent and hybrid specifications, which confirms that the structural differences reported above are larger than the sampling uncertainty of the loss differential. Second, shortening the rolling window to 250 days raises the loss of every model and widens the gap between the ensembles and GARCH, which is consistent with the interpretation that the ensembles require more observations to pin down their partition of the feature space. Third, excluding the two crisis states from the evaluation sample reverses the sign of the full-sample QLIKE differential at the 22-day horizon, which shows that the aggregate ranking reported in earlier studies is driven by the composition of the evaluation window rather than by a property of the estimators.

6. Boundary Conditions and a Conditional Decision Framework

6.1. Horizon, State, and Information as Constraints

The results imply three boundary conditions, each of which is a constraint on admissibility rather than a statement about average performance. The first is the horizon. At 1 day the conditional-variance recursion dominates, and the gain from a flexible learner is below 2% on QLIKE, which is insufficient to offset the estimation, validation, and monitoring costs of a nonparametric pipeline. At 5 days the ensembles become competitive when the jump component and the macro covariates are available, because the direct projection avoids the compounding of specification error that the iterated recursion incurs. At 22 days the mean-reversion constraint of the parametric models becomes the dominant source of accuracy, and the ensembles are useful only as an auxiliary signal.

The second constraint is the state. During the volatility spikes of 2008 and 2020 the exponential smoother underestimated subsequent volatility after the initial shock and overestimated it during the recovery, and the symmetric recursion of GARCH responded too slowly to negative shocks. The ensembles were more robust to the initial spike, because their predictions are bounded by the range of observed outcomes, but they reverted to the mean once the extreme observations fell outside the training sample. The recurrent networks lost accuracy for a third reason, the shift in the joint distribution of the covariates. No single model class is robust across states, which implies that state

recognition is a precondition for model selection rather than a refinement of it.

The third constraint is information availability. The informational gains documented in Section 5.2 presuppose that the jump component and the macro covariates are observed at the forecast origin with no revision. In applications where the covariates are published with a lag, or where the realized measure is computed from a thinner sampling grid, the informational advantage shrinks toward the structural differential, and the recommendation reverts to the parametric benchmark.

6.2. A Decision Matrix

Combining the three constraints yields a decision matrix indexed by horizon, state, and information availability (Table 2). The matrix is not a ranking. Its purpose is to convert the question of which model is best into the question of which model is admissible under a stated set of conditions, and to make the conditions explicit enough to be audited.

Table 2. Conditional decision matrix. Cells report the admissible model class for each combination of forecast horizon, market state, and information availability; where two classes are listed, the first is the primary model for tail calibration and the second is an auxiliary signal.

Horizon	Stable, basic information	Stable, extended information	Extreme, basic information	Extreme, extended information
1 day	GARCH(1,1)	GARCH(1,1) with jump term	GARCH(1,1)	GARCH(1,1) with jump term
5 days	GARCH(1,1)	Gradient boosting	GARCH(1,1)	GARCH(1,1); boosting as auxiliary
22 days	GARCH(1,1)	Gradient boosting; HAR-RV-J	GARCH(1,1)	GARCH(1,1); HAR-RV-J as auxiliary

Note. Basic information denotes the 22 daily lags of realized volatility; extended information adds the jump component, the VIX term-structure slope, the credit spread, and the Amihud illiquidity measure. RiskMetrics is admissible at the 1-day horizon in stable states when computational cost is binding, but not in the recovery phase of a crisis, where its missing mean reversion produces persistent overstatement.

The matrix reproduces, in conditional form, the two results that the unconditional literature finds difficult to reconcile. The parametric benchmark is admissible at every horizon in every state, which is the conditional version of the finding that GARCH is hard to beat. The tree ensembles are admissible in stable states with extended information, which is the conditional version of the finding that machine learning improves volatility forecasts. Both statements are true, and neither is a statement about average superiority.

6.3. Implications for Comparative Research

The framework has three implications for the design of comparative studies. First, the information set must be held constant across model classes, or the reported differential conflates structure with information; the decomposition in Section 4.2 provides a simple way to report both components. Second, at least one distributional criterion must accompany the point-forecast criterion, because the two can rank models differently and the divergence is informative rather than incidental. Third, the evaluation must be stratified by state, since an average over states is a weighted mixture whose weights are an artifact of the sample window; reporting the state-

conditional losses alongside the average would make the boundary conditions visible and would explain a substantial part of the disagreement in the existing literature.

The limitations of the present design follow from the same logic. The states are identified ex post, so the matrix is a description of admissibility conditional on a correct state classification rather than a real-time switching rule. The sample covers a single equity index in a single market, and the informational variables are those available for that market; extending the design to other asset classes would test whether the state-conditional ordering is a property of the measurement environment or of the asset class. Finally, the tail evaluation is univariate, and multivariate tail risk, where dependence across assets matters more than the marginal calibration of each, remains outside the scope of the present comparison.

7. Conclusion

The evidence assembled here supports a specific reading of the volatility forecasting literature. The contribution of machine learning to tail risk forecasting is real but conditional: it consists mainly in extracting nonlinear interactions between jump information, forward-looking variance prices, and macro-financial conditions, and it is realized only when that information is available and when the market state remains within the support of the training sample. The relation between flexible learners and conditional-variance recursions is therefore one of complementarity rather than substitution, and the degree of complementarity is modulated by the state.

The practical consequence is a shift in the object of model selection. Instead of asking which specification is superior on average, the relevant question is which specification is admissible given a horizon, a state, and an information set. The decision matrix formalizes that question and makes the answer auditable. The methodological consequence is a shift in the design of comparative studies: hold the information set constant, report both a point-forecast and a distributional criterion, stratify by state, and report the confidence set rather than a point ranking. A comparison designed in this way would not resolve the question of which model is best, because that question has no unconditional answer; it would instead specify the conditions under which each answer holds.

Two extensions follow naturally. The first is to replace the ex post state classification with a real-time probability, so that the matrix becomes a switching rule with an explicit error rate. The second is to extend the evaluation from univariate coverage to multivariate tail dependence, where the relevant question is not whether each model calibrates its own quantile but whether the joint exceedance structure is preserved. Both extensions preserve the logic of the present design: the comparison is informative only when the conditions under which it is conducted are stated.

References

[1] 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. https://doi.org/10.1002/jae.800

[2] Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. The Review of Financial Studies, 33(5), 2223–2273. https://doi.org/10.1093/rfs/hhaa009

[3] Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. Journal of Econometrics, 160(1), 246–256. https://doi.org/10.1016/j.jeconom.2010.03.034

[4] Patton, A. J., Ziegel, J. F., & Chen, R. (2019). Dynamic semiparametric models for expected shortfall (and value-at-risk). Journal of Econometrics, 211(2), 388–413. https://doi.org/10.1016/j.jeconom.2018.10.008

[5] Wafi, A. S., El-Halaby, S., & Aboul-Dahab, S. A.-E. (2026). Regime-dependent bitcoin volatility forecasting under alternative information sets: Evidence from econometric, deep learning, and hybrid models. Digital Finance, 8(4), Article 63. https://doi.org/10.1007/s42521-026-00229-9

[6] Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics, 31(3), 307–327. https://doi.org/10.1016/0304-4076(86)90063-1

[7] Patton, A. J., & Sheppard, K. (2015). Good volatility, bad volatility: Signed jumps and the persistence of volatility.

Review of Economics and Statistics, 97(3), 683–697. https://doi.org/10.1162/REST_a_00503
[8] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324
[9] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
[10] Christensen, K., Siggaard, M., & Veliyev, B. (2023). A machine learning approach to volatility forecasting. Journal of Financial Econometrics, 21(5), 1680–1727. https://doi.org/10.1093/jjfinec/nbac020
[11] Luo, Y., Xue, X., & Izzeldin, M. (2025). When MIDAS meets LASSO: The power of low-frequency variables in forecasting value-at-risk and expected shortfall. Journal of Financial Econometrics, 23(1), Article nbae016. https://doi.org/10.1093/jjfinec/nbae016
[12] Kim, H. Y., & Won, C. H. (2018). Forecasting the volatility of stock price index: A hybrid model integrating LSTM with multiple GARCH-type models. Expert Systems with Applications, 103, 25–37. https://doi.org/10.1016/j.eswa.2018.03.002
[13] Chassot, J., & Audrino, F. (2026). HARd to beat: The overlooked impact of rolling windows in the era of machine learning. International Journal of Forecasting, 42(2), 330–343. https://doi.org/10.1016/j.ijforecast.2025.06.003
[14] Corsi, F. (2009). A simple approximate long-memory model of realized volatility. Journal of Financial Econometrics, 7(2), 174–196. https://doi.org/10.1093/jjfinec/nbp001
[15] Corsi, F., Pirino, D., & Renò, R. (2010). Threshold bipower variation and the impact of jumps on volatility forecasting. Journal of Econometrics, 159(2), 276–288. https://doi.org/10.1016/j.jeconom.2010.07.008
[16] Andersen, T. G., Bollerslev, T., Diebold, F. X., & Labys, P. (2003). Modeling and forecasting realized volatility. Econometrica, 71(2), 579–625. https://doi.org/10.1111/1468-0262.00418
[17] Bollerslev, T., Patton, A. J., & Quaedvlieg, R. (2016). Exploiting the errors: A simple approach for improved volatility forecasting. Journal of Econometrics, 192(1), 1–18. https://doi.org/10.1016/j.jeconom.2015.10.007
[18] Kupiec, P. H. (1995). Techniques for verifying the accuracy of risk measurement models. The Journal of Derivatives, 3(2), 73–84. https://doi.org/10.3905/jod.1995.407942
[19] Christoffersen, P. F. (1998). Evaluating interval forecasts. International Economic Review, 39(4), 841–862. https://doi.org/10.2307/2527341
[20] Taylor, J. W. (2019). Forecasting value at risk and expected shortfall using a semiparametric approach based on the asymmetric Laplace distribution. Journal of Business & Economic Statistics, 37(1), 121–133. https://doi.org/10.1080/07350015.2017.1281815
[21] Hansen, P. R., Lunde, A., & Nason, J. M. (2011). The model confidence set. Econometrica, 79(2), 453–497. https://doi.org/10.3982/ECTA5771
[22] Barndorff-Nielsen, O. E., & Shephard, N. (2002). Econometric analysis of realized volatility and its use in estimating stochastic volatility models. Journal of the Royal Statistical Society: Series B, 64(2), 253–280. https://doi.org/10.1111/1467-9868.00336
[23] Amihud, Y. (2002). Illiquidity and stock returns: Cross-section and time-series effects. Journal of Financial Markets, 5(1), 31–56. https://doi.org/10.1016/S1386-4181(01)00024-6