

Regime-Gated Tail Transfer: Reliable Cross-Industry Machine Learning for Value-at-Risk and Expected Shortfall

Rakshay Pawar¹

Abstract

Pooling financial series gives machine-learning tail models more extreme observations, but it can also transfer the wrong tail dynamics across assets and regimes. Regime-gated tail transfer (RGTT) combines Student-t GARCH, filtered historical simulation, industry-specific quantile boosting, and pooled cross-industry quantile boosting. FZ0-scored weights depend on nine lagged volatility–breadth states, with shrinkage toward global performance. A coverage controller (RGTT-C) uses past exceedances and preserves $ES \leq VaR$. Hyperparameters are selected on 1995–1999 and frozen before a 2000–July 2026 test across ten U.S. industry portfolios (66,840 industry-days). RGTT significantly improves on equal weighting and a state-agnostic gate. RGTT-C attains mean FZ0 -3.4887 and a 2.491% breach rate at the 2.5% target, with significant loss advantages over filtered historical simulation and both quantile-ML experts. Its advantage over Student-t GARCH is not significant at 5% (two-sided $p = 0.072$). The pooled expert’s mean weight falls from 22.5% to 13.5% between low-volatility/low-breadth and high-volatility/high-breadth states. The framework makes cross-industry transfer state dependent while separating expert weighting from an order-preserving coverage controller.

JEL classification numbers: C22, C45, C53, G17, G32.

Keywords: Expected shortfall, Value-at-risk, Forecast combination, Quantile machine learning; negative transfer, Regime adaptation, Market risk.

¹ Codeflow Capital LLC, United States.

1. Introduction

Expected Shortfall (ES) has become central to market-risk measurement because it describes the average loss beyond a Value-at-Risk (VaR) threshold rather than only the threshold itself. Yet tail forecasting remains a small-data problem in disguise. A 2.5% daily tail supplies roughly six observations per trading year for any single asset. Pooling related assets can multiply the effective number of tail observations and expose common market structure, but the same pooling can blur industry-specific leverage, drawdown, and jump behavior. The statistical gain from transfer is therefore inseparable from the risk of negative transfer.

This tension is especially acute in financial machine learning. Flexible learners can exploit nonlinear interactions among volatility, trend, breadth, and cross-sectional dispersion, but a model trained on a broad panel may treat a shared market shock as if it had the same tail consequence in energy, technology, utilities, and consumer sectors. Conversely, a separate learner for each series protects local structure but is estimated from very few extreme observations. A useful risk system should not choose permanently between pooling and separation. It should learn, using only information available at the forecast date, when transfer is reliable.

We propose regime-gated tail transfer (RGTT), an online mixture-of-experts architecture for joint VaR and ES. It combines two econometric experts and two quantile-machine-learning experts. The first machine-learning expert is local to each industry; the second is pooled across industries. An online gate scores the experts with the strictly consistent FZ0 joint VaR–ES loss and adapts the weights separately in nine states defined by lagged market volatility and downside breadth. Sparse state evidence is shrunk toward a global performance estimate, so the gate does not overreact to a handful of tail observations. Because the combined forecasts are convex averages of coherent expert pairs, the ES–VaR ordering is retained by construction.

We also introduce RGTT-C, a causal coverage-feedback layer. A single positive scale factor is updated after each trading day from the preceding cross-sectional exceedance rate. Multiplying both forecasts by that factor preserves coherence while correcting slow common under- or overcoverage. The state gate and coverage controller solve different problems: the former reallocates probability-model expertise; the latter regulates long-run calibration.

The paper makes four contributions.

- It formulates cross-industry pooling as a state-dependent transfer problem rather than an unconditional modeling choice.
- It develops a shrinkage-stabilized, proper-score gate for coherent VaR–ES pairs and a separate order- preserving coverage controller.
- It uses a long, genuinely held-out evaluation spanning the dot-com decline, the global financial crisis, the euro-area crisis, the COVID-19 shock, and the 2022–2023 tightening cycle.

- It opens the mechanism to inspection: state weights, expert ablations, calibration tests, loss-difference inference, and a complete reproducibility workflow are reported.

The headline evidence is deliberately nuanced. The regime gate itself improves significantly on both equal weighting and a state-agnostic performance gate. The coverage-controlled version has the best mean FZ0 score and near-nominal aggregate coverage, but its advantage over Student-t GARCH is not significant at the conventional 5% two-sided level. The method should therefore be viewed as a reliable transfer-and-calibration architecture, not as proof that machine learning universally dominates econometrics.

2. Related Literature

2.1 Joint VaR–ES evaluation

Artzner et al. (1999) established the coherence rationale for tail-average risk measures. VaR alone is elicitable, whereas ES alone is not; Fissler and Ziegel (2016) showed that the VaR–ES pair is jointly elicitable. This permits statistically principled forecast comparison using strictly consistent joint loss functions. Nolde and Ziegel (2017) developed comparative backtests from this framework, and Patton, Ziegel and Chen (2019) used dynamic semiparametric models to forecast and evaluate VaR and ES jointly. Taylor (2020) and Storti and Wang (2022, 2023) further developed joint models and combinations. RGTT uses a member of the Fissler–Ziegel class not only for final evaluation but also as the causal signal that governs online model weights.

2.2 Econometric and machine-learning tail forecasts

Conditional volatility models remain difficult to displace in daily tail risk. The GARCH model of Bollerslev (1986) and the filtered historical simulation of McNeil and Frey (2000) combine parsimonious volatility dynamics with distributional or empirical tail assumptions. Engle and Manganelli (2004) model the conditional quantile directly. More recent work brings flexible nonlinear learners into volatility and tail forecasting. Zhang et al. (2024) document gains from machine learning and intraday commonality, while Rosenbaum and Zhang (2024) study universality in cross-sectional volatility prediction. Recent large and recurrent architectures for VaR–ES forecasting show the value of nonlinear representation learning (Wang et al., 2025; Zhang et al., 2026). Our objective is complementary: retain simple auditable experts, then make cross-sectional transfer conditional on observable market state.

2.3 Forecast combinations and negative transfer

Forecast combinations often outperform a single specification because model instability and misspecification vary over time (Wang et al., 2023). Joint VaR–ES model averaging has recently received direct attention (Jiao et al., 2026), and preprint evidence explores ES-specific averaging (Wu et al., 2026). Wang and

Wang (2025) instead use cross-sectional aggregation as a mathematical construction for long memory; that meaning is distinct from pooling an observed asset panel. Standard combinations usually learn one set of weights or a smooth global time variation. RGTT instead asks whether an expert’s recent proper-score advantage survives within the current volatility–breadth state. This distinction matters because a pooled learner can be informative in calm or narrow markets and hazardous during synchronized selloffs. The architecture makes that negative-transfer channel measurable through the weight assigned to the pooled expert.

3. Methodology

3.1 Forecast target and scoring rule

Let $y_{i,t}$ denote the daily simple return of industry i on date t . For $\alpha = 0.025$, $q_{i,t}$ is the conditional α -quantile and $e_{i,t}$ is the corresponding lower-tail expectation, with $e_{i,t} < q_{i,t} < 0$. Each forecast uses information in $\mathcal{F}_{t-1}$. We evaluate coherent pairs with the FZ0 loss

$$L_\alpha(y, q, e) = -\frac{1\{y \leq q\}(q - y)}{\alpha e} + \frac{q}{e} + \log(-e) - 1, \quad (1)$$

where lower values are better. The score is strictly consistent for the VaR–ES pair under standard regularity conditions. A common positive rescaling of returns and forecasts adds the same constant to every model’s loss, leaving loss differences and forecast rankings unchanged.

3.2 The four experts

The ensemble deliberately mixes distributional, resampling, local-learning, and transfer-learning views of the tail.

- Student-t GARCH (GARCH-t). For each industry, a GARCH(1,1) variance recursion with standardized Student-t innovations is estimated by maximum likelihood. The one-step VaR and ES are analytic functions of the forecast variance and fitted degrees of freedom.
- Exponentially weighted filtered historical simulation (EWMA-FHS). Returns are divided by a causal EWMA volatility estimate with decay parameter implied by an update rate of 0.03. VaR is the rolling 2.5% quantile of the preceding 1,000 standardized returns, rescaled by current conditional volatility. ES is approximated by the mean of five lower-tail quantiles.
- Local quantile machine learning (Local-QML). A separate histogram gradient-boosting quantile model is fit for each industry. The response is the next return divided by causal EWMA scale. Models estimate the 2.5% VaR quantile and five midpoint tail quantiles (Koenker and Bassett, 1978; Friedman, 2001; Pedregosa et al., 2011).
- Pooled quantile machine learning (Pooled-QML). The same learner is estimated on the stacked ten-industry panel, with industry identity treated as categorical. It

receives the local and common predictors available to Local-QML but can borrow nonlinear structure across industries.

For the two QML experts, the ES forecast is the midpoint quadrature approximation

$$\hat{e}_{i,t} = \frac{1}{5} \sum_{k=1}^5 \hat{Q}_{i,t}(u_k),$$

$$u = \{0.0025, 0.0075, 0.0125, 0.0175, 0.0225\}. \quad (2)$$

The five tail quantiles are sorted before averaging, and $\hat{e}$ is clipped below $\hat{q}$ by a numerical tolerance. Inputs include two own-return lags; own absolute return; 5- and 20-day cumulative returns; 5-, 20-, 60-, and 252-day volatility; 20- and 60-day downside volatility; 60-day skewness; 252-day drawdown; 60-day market beta; analogous market variables; lagged SMB and HML; cross-industry mean and dispersion; and one- and five-day downside breadth. Appendix A gives the complete specification.

3.3 Causal market states

The gate separates market conditions along two dimensions. The first is lagged 20-day market volatility. The second is lagged five-day downside breadth, the average fraction of industries with a negative return. Each variable is classified into a rolling low, middle, or high tercile using the preceding 1,260 observations, with at least 756 observations required. Combining the two classifications yields nine states $s_t \in \{0, \dots, 8\}$. Both state variables and their thresholds are lagged; the date- t state is therefore known before y_t is observed.

3.4 Shrinkage-stabilized performance gate

For expert m , let $\ell_{m,t}$ be its date- t FZ0 loss averaged across the ten industries. The global and state-specific discounted means are updated only after date t is forecast and realized. With decay $\rho = \exp[\log(0.5)/h]$, the effective state count and loss sum obey

$$N_{s,t} = \rho N_{s,t-1} + 1\{s_t = s\},$$

$$A_{s,m,t} = \rho A_{s,m,t-1} + 1\{s_t = s\} \ell_{m,t}. \quad (3)$$

Writing $\bar{L}_{m,t}$ for the analogous global discounted mean, the score used to forecast date $t+1$ in state s is

$$\tilde{L}_{s,m,t} = \frac{N_{s,t}(A_{s,m,t}/N_{s,t}) + \kappa \bar{L}_{m,t}}{N_{s,t} + \kappa}. \quad (4)$$

The parameter κ behaves like a prior effective sample size. Expert weights follow an exponential rule

$$w_{s,m,t+1} = \frac{\exp(-\eta \tilde{L}_{s,m,t})}{\sum_j \exp(-\eta \tilde{L}_{s,j,t})}. \quad (5)$$

The RGTT pair is $q^{\text{RGTT}} = \sum_m w_m q_m$ and $e^{\text{RGTT}} = \sum_m w_m e_m$. Non-negative weights sum to one, so $e^{\text{RGTT}} \leq q^{\text{RGTT}}$ whenever the inequality holds for every expert. The validation-selected values are $h = 504$ trading days, $\kappa = 90$, and $\eta = 5$.

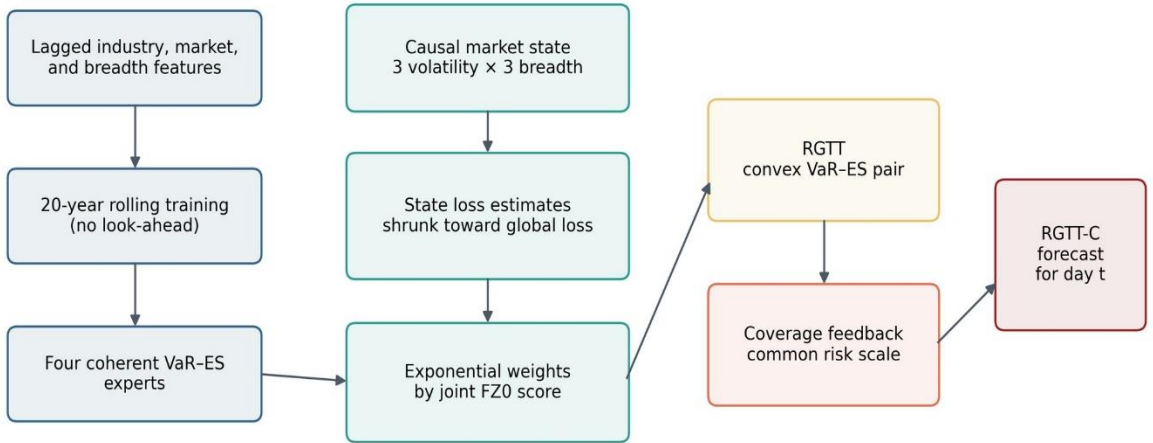
3.5 Coverage feedback without incoherence

A proper-score gate can still inherit persistent common undercoverage from its experts. RGTT-C applies a scale c_t to both negative forecasts. Let $\bar{H}_t$ be the fraction of industries whose realized return is below $c_t q_{i,t}^{\text{RGTT}}$. The next scale is

$$c_{t+1} = \text{clip}\{c_t \exp[\gamma(\bar{H}_t - \alpha)], 0.70, 2.00\}, \quad c_1 = 1. \quad (6)$$

If exceedances have just been too frequent, c rises and both negative risk forecasts become more conservative. The date- t scale depends only on hits through $t-1$. Because the same positive factor multiplies q and e , the ES–VaR order is preserved. Validation selects $\gamma = 0.02$.

Figure 1 summarizes the causal information flow.



After observing day t , only past loss and exceedance states are updated for day $t+1$.

Figure 1: RGTT information flow

All predictors, state assignments, expert scores, weights, and coverage scaling used at date t are functions of data observed no later than $t-1$. The realized return updates the gate and controller only for subsequent dates.

4. Data and Experimental Design

4.1 Data

The study uses daily value-weighted returns for the ten industry portfolios and daily Fama–French factors from the Kenneth R. French Data Library. The portfolios contain eligible NYSE, AMEX, and NASDAQ stocks assigned from SIC codes (Fama and French, 1997). The raw sample runs from 1 July 1926 through 31 July 2026. The ten portfolio labels are nondurables, durables, manufacturing, energy, high technology, telecommunications, shops, health, utilities, and other. Returns are converted from percent to decimal units. Values used by the provider as missing-data codes are removed.

The untouched test contains 6,684 dates for each of ten industries, or 66,840 industry-day forecasts, from January 2000 through July 2026. Over this period, daily volatility ranges from 0.95% for nondurables to 2.04% for durables. Empirical 2.5% tail means range from -2.83% for nondurables to -5.99% for durables, providing substantial cross-sectional heterogeneity for the transfer problem.

4.2 No-lookahead protocol

The forecast schedule is designed before the test is examined. Each expert is estimated on a rolling 20-year window and refit at five-year boundaries beginning in 1990. The 1990–1994 period initializes discounted gate histories. All gate parameters and the coverage gain are selected using only 1995–1999. The values are then frozen. The final test begins on 3 January 2000 and runs through 31 July 2026. Within the test, the gate and controller continue to update causally, but no hyperparameter is reselected. Histogram gradient boosting uses 180 maximum boosting iterations, learning rate 0.055, 15 leaf nodes, L2 regularization 1.0, and internal early stopping on a subset of the historical training window. Random seeds are fixed.

Table 1: Empirical design and predetermined choices

Component	Specification
Forecast target	One-day 2.5% VaR and ES
Universe	10 value-weighted U.S. industry portfolios
Raw data	1 Jul 1926–31 Jul 2026
Gate warm-start	1990–1994
Validation	1995–1999
Untouched test	3 Jan 2000–31 Jul 2026; 66,840 industry-days
Expert refits	Rolling 20 years; five-year blocks
Gate	9 states; $h = 504$; $\kappa = 90$; $\eta = 5$
Coverage controller	$\gamma = 0.02$; scale clipped to $[0.70, 2.00]$
Inference	Daily cross-sectional loss; NW(20); block bootstrap 20 days, 2,000 draws

Notes: All choices reported as selected were selected on the 1995–1999 validation interval. The 2000–2026 test was not used for tuning.

4.3 Evaluation and inference

Primary ranking uses mean FZ0 loss. For inference, losses are first averaged across industries each day. Pairwise loss differences (Diebold and Mariano, 1995) are evaluated with a Newey–West standard error (Newey and West, 1987) using 20 lags and with a 95% moving-block bootstrap confidence interval using 20-day blocks and 2,000 replications. Calibration is assessed industry by industry using the unconditional coverage test of Kupiec (1995), the first-order independence test of Christoffersen (1998), and a Newey–West test of the exceedance-residual moment $1\{y \leq q\}(y - e)$. The last diagnostic is a direct ES calibration moment and should not be interpreted as the full set of regression-based ES backtests.

5. Results

5.1 Overall forecasting performance

Table 2 ranks all methods over the untouched test. RGTT-C achieves the lowest mean FZ0 loss, -3.4887 . The uncorrected RGTT gate is second at -3.4815 , followed by the global gate, equal combination, and GARCH-t. The pooled QML expert ranks last and breaches VaR on 4.719% of industry-days—almost twice the nominal 2.5% rate. This is the negative-transfer failure that the gate is designed to contain.

Table 2: Out-of-sample VaR–ES performance, January 2000–July 2026

Model	Mean FZ0	Breach %	Kupiec pass	Indep. pass	ES-cal. pass
RGTT-C	-3.4887	2.491	10/10	6/10	10/10
RGTT	-3.4815	3.149	0/10	6/10	10/10
Global-gate	-3.4794	3.140	0/10	4/10	10/10
Equal-4	-3.4771	3.238	0/10	5/10	10/10
GARCH-t	-3.4762	3.069	2/10	5/10	10/10
Local-QML	-3.4480	3.583	0/10	7/10	9/10
EWMA-FHS	-3.4332	2.603	10/10	2/10	8/10
Pooled-QML	-3.3820	4.719	0/10	9/10	6/10

Notes: Lower FZ0 is better. Pass counts report industries with $p \geq 0.05$. The nominal breach rate is 2.5%. ES-cal. is the exceedance-residual moment test described in Section 4.3.

Coverage feedback makes a material calibration difference. RGTT without feedback breaches on 3.149% of industry-days and fails unconditional coverage in all ten industries. RGTT-C lowers the rate to 2.491% and passes Kupiec coverage in all ten, while retaining the best mean loss. Independence remains difficult for every method: RGTT-C passes in six industries, compared with two for EWMA-FHS and nine for the undercovered pooled model. Exact unconditional coverage is therefore not evidence that exceedances are serially ideal.

5.2 Statistical comparisons

Table 3: Daily loss differences: RGTT-C minus each benchmark

Benchmark	Mean Δ loss	NW SE	p-value	95% block CI
GARCH-t	-0.0125	0.0069	0.072	[-0.0257, 0.0020]
EWMA-FHS	-0.0555	0.0092	<0.001	[-0.0747, -0.0398]
Local-QML	-0.0407	0.0126	0.001	[-0.0642, -0.0167]
Pooled-QML	-0.1067	0.0227	<0.001	[-0.1526, -0.0621]
Equal-4	-0.0116	0.0081	0.151	[-0.0280, 0.0041]
Global-gate	-0.0093	0.0069	0.180	[-0.0231, 0.0039]
RGTT	-0.0072	0.0065	0.273	[-0.0203, 0.0055]

Notes: Negative differences favor RGTT-C. Newey–West uses 20 lags; moving-block bootstrap uses 20-day blocks and 2,000 draws. P-values are two-sided.

RGTT-C significantly outperforms EWMA-FHS ($p < 0.001$), Local-QML ($p = 0.001$), and Pooled-QML ($p < 0.001$). Its advantage over GARCH-t has $p = 0.072$ and a bootstrap interval that includes zero; its advantages over Equal-4, Global-gate, and base RGTT are also not significant at 5%. These results separate two findings. First, state-conditioned gating itself improves on Equal-4 by -0.00442 FZ0 units (Newey–West $p = 0.037$; 95% block interval $[-0.00870, -0.00035]$) and on the state-agnostic Global-gate by -0.00213 ($p = 0.021$; interval $[-0.00407, -0.00048]$). Second, the coverage layer improves calibration and the point estimate, but its incremental loss improvement is not statistically resolved over the test span. Figure 2 shows cumulative loss differences relative to Equal-4, and Figure 3 compares mean test losses.

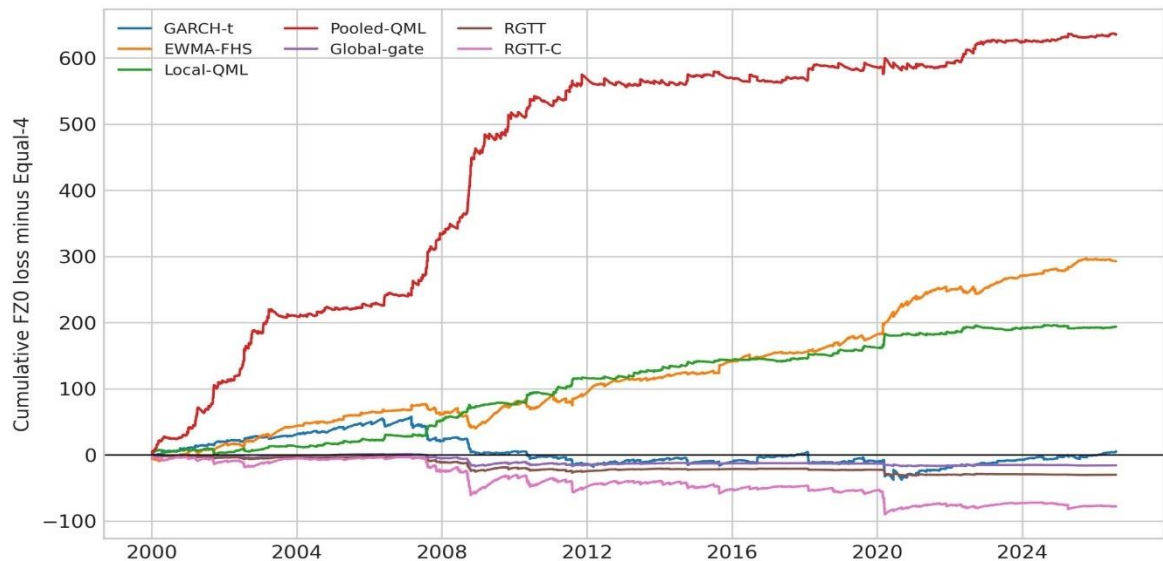


Figure 2: Cumulative daily FZ0 loss relative to the equal-weight combination

A downward path indicates lower cumulative loss than Equal-4. The final ordering, rather than every short subperiod, is the primary comparison.

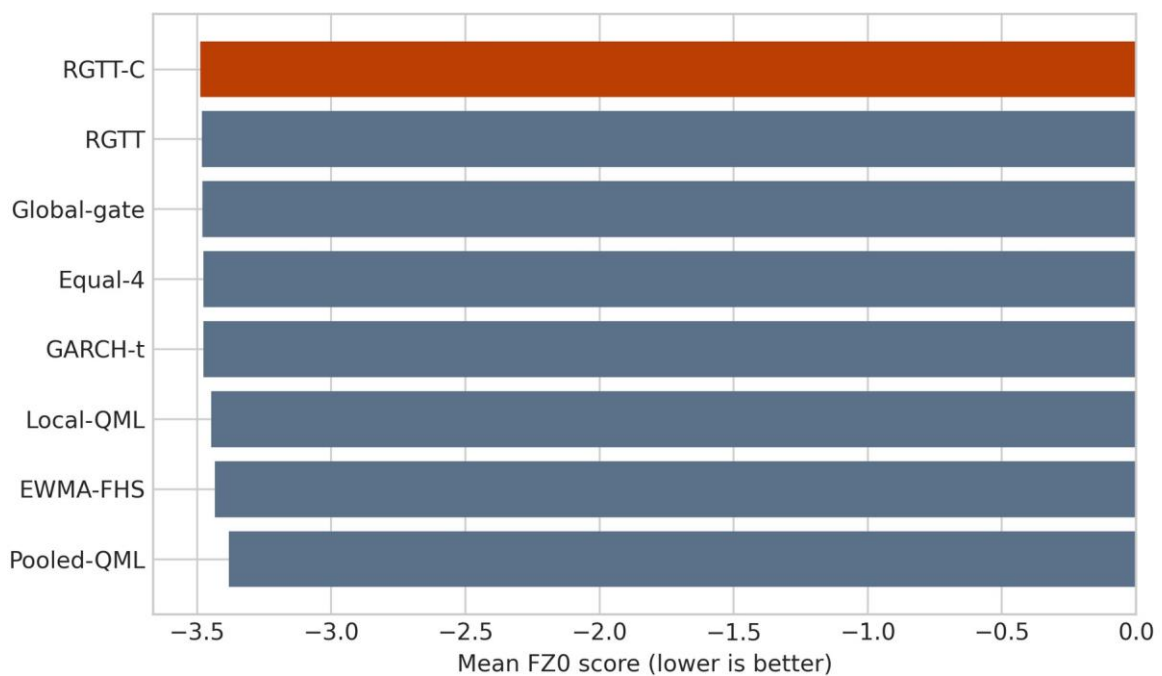


Figure 3: Mean FZ0 loss over the untouched test

Lower values are better; RGTT-C is highlighted.

5.3 Stress-period behavior

Table 4: Stress-period FZ0 loss and VaR breach rate

Period	GARCH-t	Pooled-QML	Equal-4	RGTT	RGTT-C
All test days	-3.476 / 3.07	-3.382 / 4.72	-3.477 / 3.24	-3.481 / 3.15	-3.489 / 2.49
Dot-com/9-11	-3.168 / 3.95	-2.951 / 6.47	-3.208 / 3.67	-3.215 / 3.53	-3.225 / 2.53
Global financial crisis	-3.014 / 4.44	-2.526 / 8.19	-2.934 / 5.14	-2.965 / 4.37	-2.998 / 3.02
Euro-area crisis	-3.212 / 3.06	-3.107 / 5.48	-3.183 / 3.77	-3.176 / 3.53	-3.200 / 2.70
COVID-19 shock	-1.966 / 6.54	-1.802 / 8.08	-1.828 / 7.50	-1.889 / 7.40	-2.094 / 6.06
Inflation tightening	-3.324 / 3.22	-3.272 / 4.26	-3.347 / 2.80	-3.346 / 2.91	-3.342 / 2.02

Notes: Each entry is mean FZ0 / breach percent. Periods are prespecified calendar windows. Lower FZ0 is better; the nominal breach rate is 2.5%.

RGTT-C has the best FZ0 score among the displayed models in the dot-com/9-11 window and the COVID-19 shock; GARCH-t narrowly leads during the global financial and euro-area crises, while Equal-4 leads during the tightening window. During COVID-19, every method substantially exceeds the 2.5% breach target, but RGTT-C records the smallest displayed breach rate (6.06%) and lowest loss (−2.094). No method is crisis-proof, and the results argue for monitoring conditional clustering even when long-run coverage is correct.

6. Mechanism, Robustness, and Interpretability

6.1 How the gate controls transfer

The gate's behavior is economically interpretable. In the low-volatility, low-downside-breadth state, mean weights are 26.0% for GARCH-t, 24.9% for EWMA-FHS, 26.6% for Local-QML, and 22.5% for Pooled-QML. In the high-volatility, high-breadth state, the GARCH-t weight rises to 36.0%, while the pooled weight falls to 13.5%. The pooled learner is not removed; its influence is reduced exactly where synchronized downside conditions make indiscriminate transfer most hazardous.

Figure 4 displays the mean expert weights across all nine states.

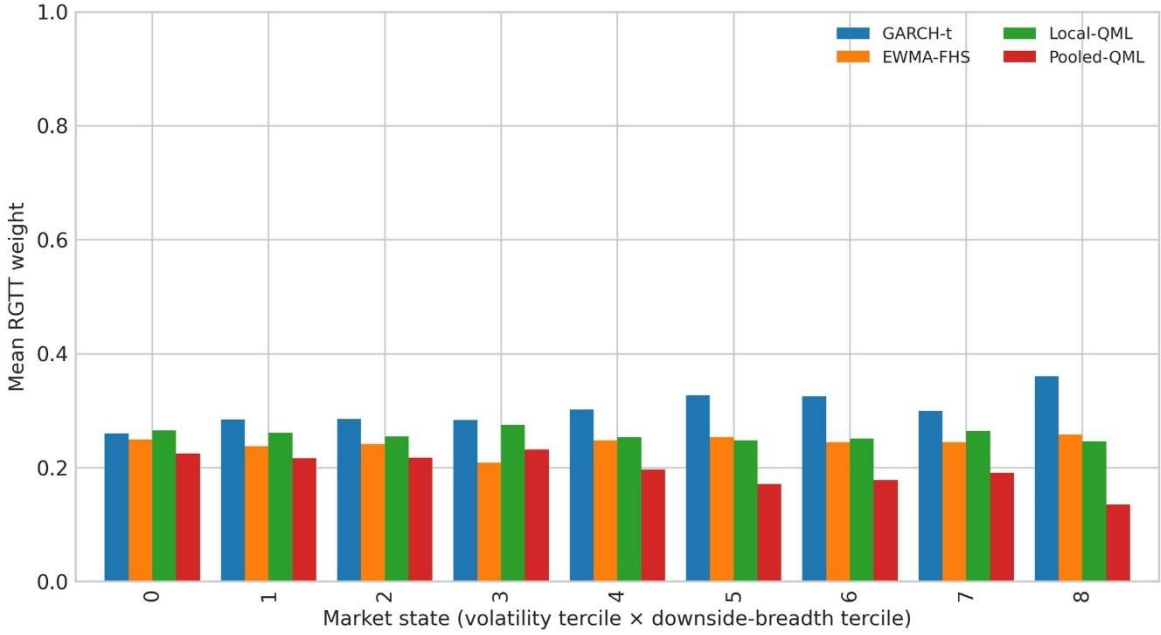


Figure 4: Mean RGTT expert weights by market state

State numbers map to volatility tercile $\times$ downside-breadth tercile: state 0 is low/low and state 8 is high/high.

6.2 Expert-set ablations

Table 5: Validation-retuned expert-set ablations

Expert set	Test FZ0	Breach %	$h/\kappa/\eta$
All four experts	-3.4815	3.149	504/90/5
Without pooled QML	-3.4750	2.965	504/90/2
Econometric only	-3.4724	2.787	504/10/10
QML only	-3.4521	3.794	126/10/10

Notes: Each reduced expert set is retuned only on 1995–1999 using the reduced grid in the reproducibility code. Results are for the base gate without the coverage controller.

Removing pooled QML worsens test FZ0 from -3.4815 to -3.4750 even though pooled QML is the weakest standalone expert. The econometric-only and QML-only gates are weaker still. This is a useful distinction between an expert’s marginal and unconditional value: a poor average model can contribute information if a gate can identify the states in which its errors are less damaging. The result also cautions against screening experts solely by unconditional rank.

6.3 Predictive content in the pooled learner

Permutation importance is computed for the pooled 2.5% quantile model within each five-year test block and then averaged. The top five variables are the market 20-day return, one-day cross-sectional return dispersion, the industry's one-day lagged return, the market five-day return, and the industry's 20-day return. Common variables therefore matter alongside local dynamics. Negative values for some correlated predictors do not establish harmful structural effects; permutation scores are conditional on the fitted model and can be unstable when features overlap.

Figure 5 reports the pooled learner's average permutation importance.

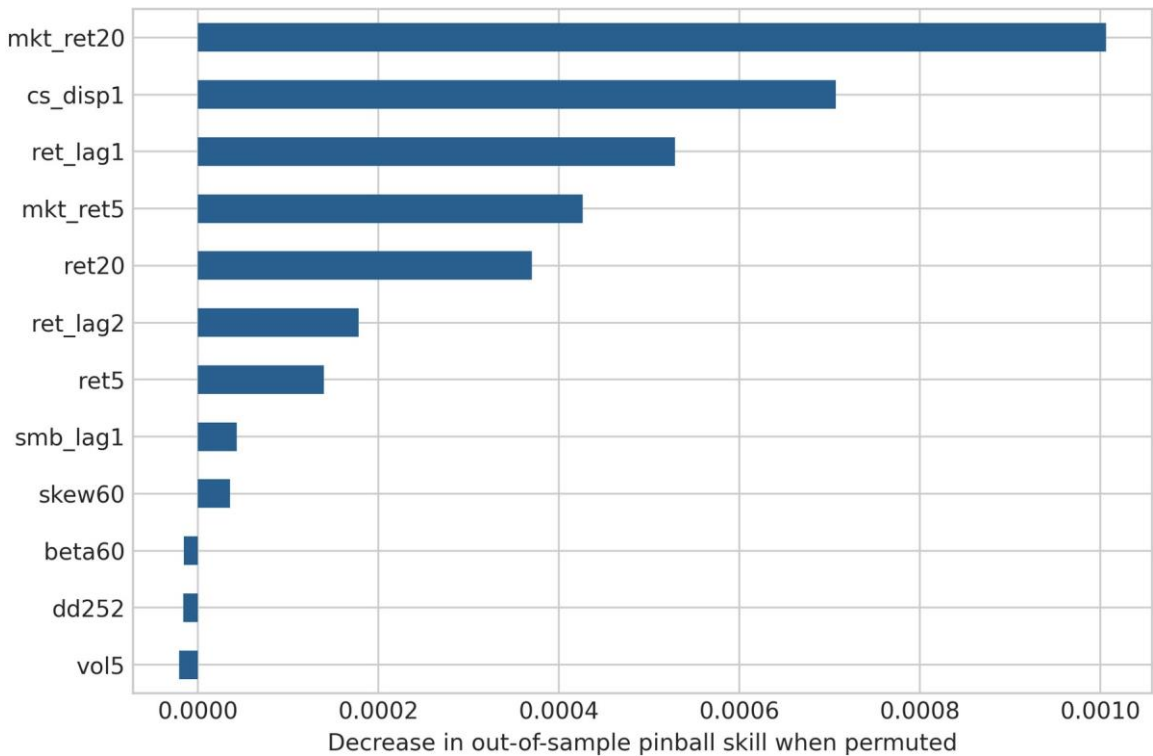


Figure 5: Average out-of-sample permutation importance for the pooled 2.5% quantile learner

Importance is the decrease in pinball-loss skill after permutation; correlated predictors can have near-zero or negative marginal values.

6.4 Parameter stability

The validation surface is flat near the selected gate. The three best combinations— (h, κ, η) equal to $(504, 90, 5)$, $(252, 30, 2)$, and $(504, 90, 2)$ —have validation FZ0 scores of -3.65582 , -3.65570 , and -3.65565 . Coverage-gain values from 0.005 to 0.05 also produce similar validation losses; $\gamma = 0.02$ is selected by the prespecified minimum. The absence of a sharp optimum reduces, but does not eliminate, concern that a single lucky hyperparameter drives the test result.

7. Practical Implications and Limitations

RGTT is intended for settings where several related books, assets, or sectors provide a common data reservoir but cannot safely be treated as exchangeable. It produces an audit trail: each forecast can be decomposed into expert pairs, state membership, gate weights, and a coverage scale. A risk manager can therefore distinguish a model change caused by rising volatility from one caused by recent miscalibration. The framework is also modular. Alternative experts can be added without changing the coherence argument, provided every expert supplies an ordered VaR–ES pair and is evaluated causally.

Several limitations matter. First, the evidence covers ten aggregate U.S. industry portfolios, not individual equities, options, foreign markets, or intraday horizons. Second, only a one-day 2.5% left tail is studied. Third, five-year expert refits are deliberately conservative and may miss faster structural change. Fourth, the aggregate coverage controller does not guarantee conditional coverage within every state and does not eliminate exceedance clustering. Fifth, the French Library series can be revised, so exact future replications should archive retrieval dates and checksums. Sixth, the study models returns rather than liquidity, market impact, or capital constraints. Finally, one long historical path cannot reveal every form of model risk. Deployment should include independent validation, stress scenarios, challenger models, and clear override governance.

A natural extension is hierarchical gating across firms, asset classes, and forecast horizons, with state-specific calibration constraints and transaction-aware capital objectives. Another is to replace discrete terciles with a continuously regularized gate while retaining the causal proper-score update. Those extensions should be evaluated on additional markets before broad claims of universality are made.

8. Conclusion

Cross-sectional pooling is neither inherently beneficial nor inherently dangerous. Its value depends on the state in which transferred information is used. RGTT operationalizes that idea with a small set of interpretable VaR–ES experts, a proper-score gate, shrinkage for rare states, and an order-preserving coverage controller. In a 26.6-year untouched test across ten U.S. industries, the regime gate significantly improves on equal and state-agnostic combinations. The controlled version delivers the best mean FZ0 score and a 2.491% aggregate breach rate, while the pooled learner's weight is sharply reduced in synchronized high-volatility selloffs. The evidence supports a disciplined middle ground: use pooled machine learning as a conditional expert, not as an unconditional replacement for robust tail models.

Data availability

The return and factor series are publicly available from the Kenneth R. French Data Library:

https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.

The exact source archives used were the daily 10 Industry Portfolios and daily Fama/French 3 Factors files retrieved in September 2026. Raw data are not redistributed in the reproducibility archive; the README gives the source URLs and SHA-256 checksums.

Code availability

A reproducibility archive accompanying the submission contains the Python analysis scripts, environment specification, machine-readable result tables, daily forecasts, and figure files. Running the primary script from a clean output directory after downloading the two source archives reproduces the reported tables and figures. The GARCH and quantile-model estimation can take substantial time on a standard workstation. The archive is available through this anonymous OSF reviewer link. [OSF | RGTT Supplementary Materials](#)

Funding

This research received no external funding.

Conflict of interest

The author declares no conflict of interest.

Use of Generative-AI tools declaration

The author declares that no AI was used for study-design brainstorming, literature discovery, or drafting and debugging the reproducible Python workflow. All numerical results reported in this article were generated by executing the code on the cited public data. The author reviewed and is responsible for the analysis, citations, interpretation, and final text.

References

- [1] Artzner, P., Delbaen, F., Eber, J.M. and Heath, D. (1999). Coherent measures of risk. *Mathematical Finance*, 9(3), pp. 203–228.
- [2] Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3), pp. 307–327.
[https://doi.org/10.1016/0304-4076\(86\)90063-1](https://doi.org/10.1016/0304-4076(86)90063-1)
- [3] Christoffersen, P.F. (1998). Evaluating interval forecasts. *International Economic Review*, 39(4), pp. 841–862. <https://doi.org/10.2307/2527341>
- [4] Diebold, F.X. and Mariano, R.S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3), pp. 253–263.
- [5] Engle, R.F. and Manganelli, S. (2004). CAViaR: Conditional autoregressive Value at Risk by regression quantiles. *Journal of Business & Economic Statistics*, 22(4), pp. 367–381.
- [6] Fama, E.F. and French, K.R. (1997). Industry costs of equity. *Journal of Financial Economics*, 43(2), pp. 153–193. [https://doi.org/10.1016/S0304-405X\(96\)00896-3](https://doi.org/10.1016/S0304-405X(96)00896-3)
- [7] Fissler, T. and Ziegel, J.F. (2016). Higher-order elicibility and Osband's principle. *Annals of Statistics*, 44(4), pp. 1680–1707.
<https://doi.org/10.1214/16-AOS1439>
- [8] Friedman, J.H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), pp. 1189–1232.
- [9] Jiao, S., Xu, W., Ye, W. and Zhang, X. (2026). Optimal model averaging for joint Value-at-Risk and Expected Shortfall regression. *Econometric Theory*, 42(1), pp. 101–134.
- [10] Koenker, R. and Bassett, G. Jr (1978). Regression quantiles. *Econometrica*, 46(1), pp. 33–50.
- [11] Kupiec, P.H. (1995). Techniques for verifying the accuracy of risk measurement models. *Journal of Derivatives*, 3(2), pp. 73–84.
<https://doi.org/10.3905/jod.1995.407942>
- [12] McNeil, A.J. and Frey, R. (2000). Estimation of tail-related risk measures for heteroscedastic financial time series: An extreme value approach. *Journal of Empirical Finance*, 7(3–4), pp. 271–300. [https://doi.org/10.1016/S0927-5398\(00\)00012-8](https://doi.org/10.1016/S0927-5398(00)00012-8)
- [13] Newey, W.K. and West, K.D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3), pp. 703–708.
- [14] Nolde, N. and Ziegel, J.F. (2017). Elicibility and backtesting: Perspectives for banking regulation. *Annals of Applied Statistics*, 11(4), pp. 1833–1874.
<https://doi.org/10.1214/17-AOAS1041>
- [15] Patton, A.J., Ziegel, J.F. and Chen, R. (2019). Dynamic semiparametric models for expected shortfall (and Value-at-Risk). *Journal of Econometrics*, 211(2), pp. 388–413. <https://doi.org/10.1016/j.jeconom.2018.10.008>

- [16] Pedregosa, F., Varoquaux, G., Gramfort, A. et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, pp. 2825–2830.
- [17] Rosenbaum, M. and Zhang, J. (2024). On universality in financial time series forecasting. *Frontiers of Mathematical Finance*, 3(1), pp. 106–126. <https://doi.org/10.3934/fmf.2024002>
- [18] Storti, G. and Wang, C. (2022). Nonparametric Expected Shortfall estimation incorporating weighted quantiles. *International Journal of Forecasting*, 38(1), pp. 224–239. <https://doi.org/10.1016/j.ijforecast.2021.04.004>
- [19] Storti, G. and Wang, C. (2023). Modelling uncertainty in financial tail risk: A forecast combination and weighted quantile approach. *Journal of Forecasting*, 42, pp. 1648–1663. <https://doi.org/10.1002/for.2972>
- [20] Taylor, J.W. (2020). Forecast combinations for Value at Risk and Expected Shortfall. *International Journal of Forecasting*, 36(2), pp. 428–441. <https://doi.org/10.1016/j.ijforecast.2019.05.014>
- [21] Wang, J. and Wang, Y. (2025). Forecasting Expected Shortfall and Value-at-Risk with cross-sectional aggregation. *Journal of Forecasting*, 44(2), pp. 391–423. <https://doi.org/10.1002/for.3195>
- [22] Wang, S., Wang, Q., Lu, H., Zhang, D., Xing, Q. and Wang, J. (2025). Learning about tail risk: A large model approach to dynamic Value-at-Risk and Expected Shortfall forecasting. *Omega*, 133, Article 103249. <https://doi.org/10.1016/j.omega.2024.103249>
- [23] Wang, X., Hyndman, R.J., Li, F. and Kang, Y. (2023). Forecast combinations: An over 50-year review. *International Journal of Forecasting*, 39(4), pp. 1518–1547.
- [24] Wu, X., Zhang, Y. and Zeng, B. (2026). Expected Shortfall model averaging. arXiv preprint, arXiv:2608.26805.
- [25] Zhang, C., Zhang, Y., Cucuringu, M. and Qian, Z. (2024). Volatility forecasting with machine learning and intraday commonality. *Journal of Financial Econometrics*, 22(2), pp. 492–530. <https://doi.org/10.1093/jjfinec/nbad005>
- [26] Zhang, Y., Zhu, K. and Zhu, Z. (2026). ReSGA: A recurrent score-guided architecture for joint Value-at-Risk and Expected Shortfall forecasting. arXiv preprint, arXiv:2606.04576.

Appendix A. Predictor specification

Table 6 presents the complete predictor specification.

Table 6: Predictors supplied to the quantile-machine-learning experts

Group	Predictors
Industry return	lags 1–2; lag 1 ; 5- and 20-day cumulative return
Industry risk	volatility 5/20/60/252; downside volatility 20/60; skewness 60; drawdown 252; market beta 60
Market	lag 1; lag 1 ; cumulative return 5/20; volatility 5/20/60/252; drawdown 252
Factors	lagged SMB and HML
Cross-section	one-day mean; one-day dispersion; downside breadth 1/5
Pooled only	industry code as a categorical feature

Notes: Every rolling statistic is shifted one day. The response is the next return divided by a causal EWMA scale.

Appendix B. Industry descriptive statistics

Table 7 presents the industry-level test-sample descriptive statistics.

Table 7: Test-sample return characteristics by industry (percent units)

Industry	N	Mean	SD	Skew	Min	q _{2.5}	Tail mean
Durbl	6684	0.052	2.042	0.052	-14.49	-4.15	-5.99
Enrgy	6684	0.050	1.788	-0.149	-19.68	-3.52	-5.15
HiTec	6684	0.047	1.664	0.255	-13.14	-3.54	-4.74
Hlth	6684	0.040	1.137	-0.114	-9.62	-2.34	-3.35
Manuf	6684	0.045	1.283	-0.202	-11.03	-2.59	-3.79
NoDur	6684	0.039	0.955	-0.199	-9.87	-1.89	-2.83
Other	6684	0.040	1.469	-0.107	-13.55	-2.97	-4.48
Shops	6684	0.046	1.230	-0.048	-10.66	-2.49	-3.50
Telcm	6684	0.017	1.312	0.076	-9.67	-2.70	-3.94
Utils	6684	0.044	1.187	0.071	-11.65	-2.36	-3.56

Notes: Test period is January 2000–July 2026. Tail mean is the average return at or below the empirical 2.5% quantile.

Appendix C. Validation grids

The gate grid is $h \in \{63, 126, 252, 504\}$, $\kappa \in \{10, 30, 90\}$, and $\eta \in \{2, 5, 10, 20\}$; the coverage grid is $\gamma \in \{0, 0.0025, 0.005, 0.01, 0.02, 0.05, 0.10, 0.20\}$. Full grids are supplied as CSV files. The selected gate has validation FZ0 -3.65582 . The selected coverage controller has validation FZ0 -3.65827 , validation breach rate 2.637%, and mean scale 1.028.