Volatility Modeling & Trading

Predicting Realized Volatility Across Multiple Horizons

Francesco Braicovich, Nikhil Joseph, Qin Zhihua (Ivan)

Statistical Modelling in Financial Engineering - IEDA 4000E
The Hong Kong University of Science and Technology

November 2025

Research Motivation

Can we predict future realized volatility accurately enough to identify mispricings in the volatility market?

The Volatility Risk Premium (VRP):

- VIX (Implied Volatility) typically exceeds Realized Volatility
- Investors pay premium for protection
- Systematic opportunity for sellers

Our Approach:

- Forecast RV using multiple models
- Compare to VIX (market expectations)
- Identify when VIX is overpriced
- VRP = VIX $- \widehat{RV}$

Project Scope

We predict realized volatility across **four time horizons**:

- h=2 days (ultra-short term)
- h=5 days (one trading week)
- h=10 days (two weeks)
- h=30 days (one month, matches VIX)

Three distinct modeling philosophies:

1. **GARCH/EGARCH**: Econometric approach (leverage effect, volatility clustering)
2. **LSTM-RV**: Deep learning on historical realized volatility
3. **LSTM-VIX**: Deep learning to "de-bias" implied volatility

Test Period & Data Sources

Data Coverage (1990-2025):

- Train: 1993-2015 (23 years)
- Validation: 2016-2019 (4 years)
- Test: 2020-2025 (5.5 years)
- Daily frequency, Yahoo Finance

Instruments:

- SPY: S&P 500 ETF
- ^VIX: CBOE Volatility Index

Why This Period?

- Dot-com (2000-2002)
- Financial crisis (2008)
- COVID crash (2020)
- Inflation (2022)
- Multiple regimes

Rigorous out-of-sample testing!

Feature Engineering: Target Variable

$$RV_{t,h}^{fwd} = \sqrt{252} \times \text{std}(r_{t+1}, r_{t+2}, \ldots, r_{t+h})$$

Components:

- r_{t+i}: Log return on day i
- std$(\cdot)$: Standard deviation
- $\sqrt{252}$: Annualization factor
- h: Forecast horizon

Key Properties:

- Uses **only future returns** ($t+1$ to $t+h$)
- Zero lookahead bias by construction
- Computed for all h $\in \{2, 5, 10, 30\}$
- Represents actual volatility realized

This is what the market tries to predict with VIX

Feature Engineering: Input Features

Models can only use information available **up to and including** time t

1. **Backward Realized Volatility** ($RV_{t,h}^{back}$)
   - Historical volatility over *past* h days
   - $RV_{t,h}^{back} = \sqrt{252} \times \text{std}(r_{t-h+1}, \ldots, r_t)$
   - Captures recent market turbulence
   - Used by LSTM-RV models

2. **Implied Volatility** (IV_t from VIX)
   - Market's expectation of future 30-day volatility
   - Forward-looking collective wisdom
   - Already available before market close at t
   - Used by LSTM-VIX model

Log Transformation: Model-Specific Approach

Volatility is strictly positive but unbounded ⇒ skewed, heteroskedastic

Before Log Transform:

- Distribution highly skewed
- Extreme outliers (crisis periods)
- Models struggle with scale
- Predictions can be negative!

After Log Transform:

- Approximately Normal distribution
- Compressed extreme values
- Stabilized variance
- Predictions always positive after $\exp(\cdot)$

$$y = RV \in (0, \infty)$$

$$y = \log(RV) \in (-\infty, \infty)$$

- **LSTM Models:** Predict $\log(RV)$ directly, then apply $\exp(\cdot)$ to get RV
- **GARCH/EGARCH:** Models $\log(\sigma_t^2)$ internally; predicts RV from variance forecasts (not log-transformed)

Lookahead Bias Prevention: Critical Detail

At time t, if we predict $RV_{fwd}[t]$ using returns from $t+1$ to $t+h$, and our training set ends exactly at validation start, the **last training samples leak into validation period!**

Training Gap: Remove samples where target horizon extends into validation

- Training ends: train_end $-$ ($h \times 1.5$) calendar days
- Factor 1.5 accounts for weekends/holidays
- For h=30: ~45 calendar day gap = 30 trading days safe

Result: Zero information leakage from future to past

Market Overview: 35 Years of Data

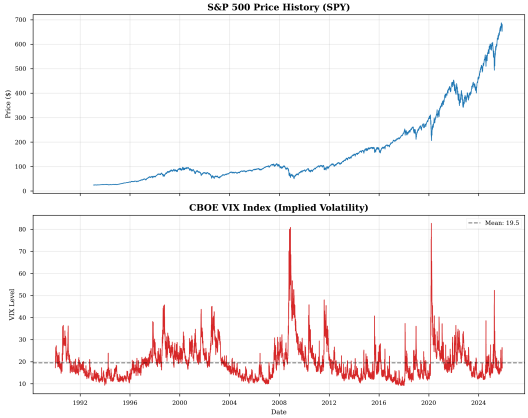
Key Observations:

SPY (Top Panel):

- Strong upward trend
- Non-stationary process
- Multiple bull/bear cycles
- Returns, not prices, are predictable

VIX (Bottom Panel):

- Mean-reverting around 15-20
- Sharp spikes during crises
- Dot-com (2000-2002): VIX ~45
- Financial crisis (2008): VIX >80
- COVID (2020): VIX ~80



Inverse relationship: VIX explodes when SPY crashes

Autocorrelation Analysis

Raw Returns:

- ACF stays within blue bands
- No significant autocorrelation
- Returns are unpredictable
- Efficient Market Hypothesis holds
- Cannot predict next day's return

Statistical Test 1: Stationarity (ADF Test)

Non-stationary series have time-varying mean/variance ⇒ **spurious regressions**

Augmented Dickey-Fuller Test:

- H_0: Unit root (non-stationary)
- H_1: Stationary
- Test statistic: -58.34
- **p-value: < 0.0001**

$$\Delta r_t = \alpha + \beta r_{t-1} + \sum_{i=1}^{p} \gamma_i \Delta r_{t-i} + \epsilon_t$$

If $\beta < 0$ significantly ⇒ mean reversion

Result Interpretation:

- Reject H_0 at 0.1% level
- Returns are stationary
- Mean and variance are stable over time
- Safe to use standard regression models

Why This Test:

- Prices grow exponentially (non-stationary)
- Returns fluctuate around zero (stationary)
- We model returns, not prices
- Avoids spurious correlations

Statistical Test 2: Stationarity (KPSS Test)

KPSS has opposite null hypothesis: H_0 = stationary (vs ADF: H_0 = non-stationary)

Test Results:

- Test statistic: 0.116
- Critical (5%): 0.463
- **p-value: 0.10**
- Fail to reject H_0

Tests: $r_t = \mu + \eta_t$ (stationary)
vs trend-stationary or random walk

Interpretation:

- Evidence for stationarity
- Consistent with ADF
- Robust conclusion from both tests

Why Dual Testing:

- ADF: reject non-stationarity
- KPSS: accept stationarity
- Agreement ⇒ high confidence

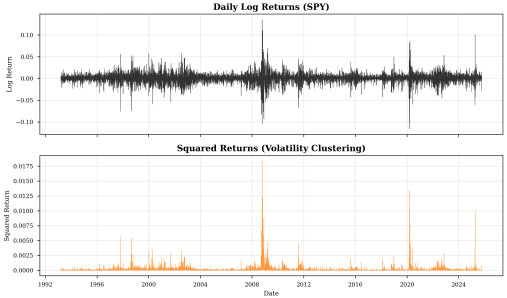
Visual Evidence: Volatility Clustering

Calm Periods:

- 1993-1997: Low, stable volatility
- 2003-2007: "Great Moderation"
- 2012-2019: Extended calm
- Returns within ±1% daily

Crisis Periods:

- 2008-2009: Persistent wild swings
- March 2020: COVID shock
- 2022: Inflation fears
- Returns ±5-10% daily for weeks



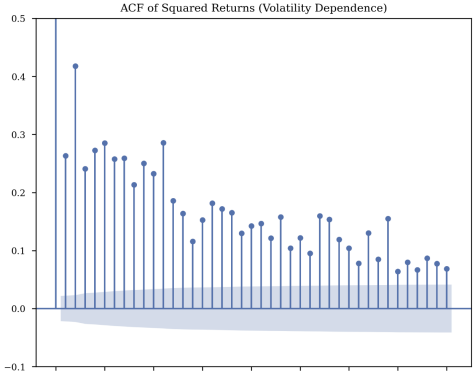
Volatility has "regimes" that persist - this is why GARCH works

Autocorrelation Analysis

"Large changes tend to be followed by large changes, of either sign, and small changes tend to be followed by small changes" - Benoit Mandelbrot

Squared Returns (Right Panel):

- ACF far above blue bands
- Strong autocorrelation up to 20 lags
- Volatility is predictable!
- Today's volatility ⇒ tomorrow's
- Justifies volatility modeling



Statistical Test 3: Volatility Clustering (ARCH-LM)

ARCH-LM Test:

- H_0: No ARCH effects (homoskedastic)
- H_1: ARCH effects present
- Test statistic: 1389.1
- **p-value: < 0.0001**

Tests if squared residuals are autocorrelated:

$$\epsilon_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2 + u_t$$

If α_i significant $\Rightarrow$ volatility clustering

Result Interpretation:

- Strongly reject H_0
- Massive test statistic (1389!)
- Volatility is **highly clustered**
- Today's volatility predicts tomorrow's

Implication for Modeling:

- Cannot use constant variance models
- **GARCH family is justified**
- Volatility has memory
- Crisis periods persist for weeks

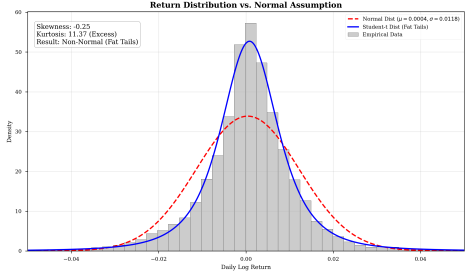
Visual Evidence: Fat Tails vs Normal Distribution

Key Findings:

- Gray: Actual distribution
- Red: Normal fit
- **Massive tail divergence**

Extreme Events (|return| > 3%):

- Normal predicts: ~30 days
- Reality: ~180 days
- **6x more frequent!**

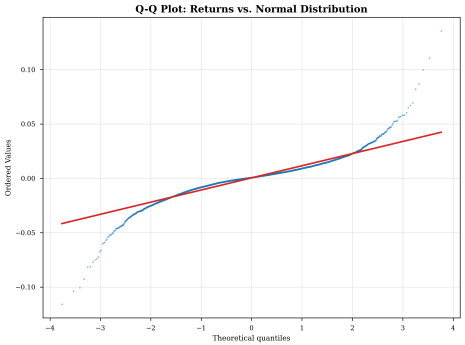


Q-Q Plot: Visual Test of Normality

- **S-shaped curve**
- Tails deviate from line
- Fat tails confirmed

Implication:

- Extreme events more frequent
- Student-t fits better ($\nu \sim 6$)
- GARCH must use Student-t



Statistical Test 4: Normality (Jarque-Bera)

Gaussian assumption affects risk estimates and confidence intervals

Test Results:

- H_0: Normal distribution
- **p-value: < 0.0001**
- Strongly reject

$$JB = \frac{n}{6}\left(S^2 + \frac{(K-3)^2}{4}\right)$$

- Skewness: -0.11
- Excess kurtosis: 10.02

Findings:

- Slight negative skew
- **Massive fat tails**
- Extremes $\gg$ Normal predicts

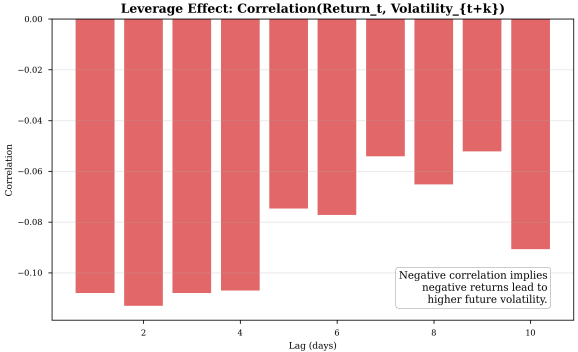
Impact:

- 3σ events: 0.3% (Normal) vs 2% (Actual)
- **Crashes 6x more frequent!**
- Must use Student-t in GARCH

The Leverage Effect: Why Markets Panic Downward

The Chart Shows:

- Correlation between today's return and future volatility
- X-axis: Days ahead (lag)
- Y-axis: Correlation strength
- All bars negative!



Key Insight:

- Negative returns today ⇒ high volatility tomorrow
- Positive returns today ⇒ low volatility tomorrow
- Effect persists for 10+ days

Feature Correlation Matrix: The Information Landscape

Strong Correlations ($\rho > 0.75$):

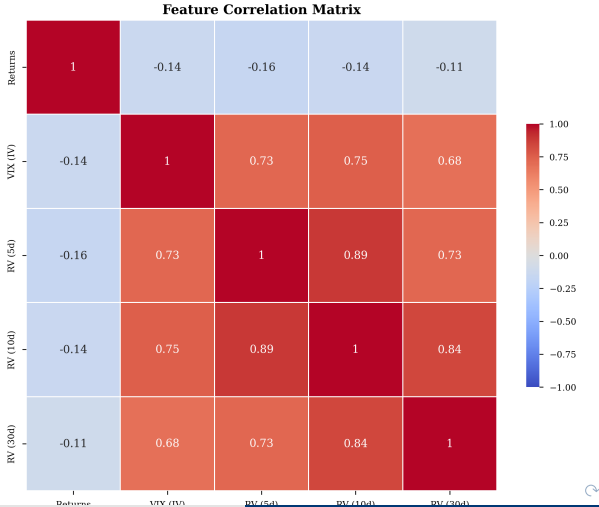
- VIX ↔ RV_{30}: 0.82
- VIX ↔ RV_{10}: 0.79
- RV horizons: >0.9 (multicollinear)
- IV captures future RV well

Negative ($\rho \sim -0.4$):

- Returns ↔ All volatilities
- Leverage effect confirmed

Implications:

- VIX highly informative (67% R^2)
- 33% variance unexplained ⇒ opportunity
- Returns add asymmetry signal



Model Selection Strategy: Grid Search

We test multiple specifications systematically:

- **Model families**: GARCH vs EGARCH
- **Orders**: $(p, q) \in \{(1,1), (1,2), (2,1), (2,2)\}$
- **Distributions**: Normal, Student-t, Skewed Student-t

Two-stage process:

1. Filter: Pass ARCH-LM test on residuals (no remaining heteroskedasticity)
2. Rank: Choose lowest BIC among passing models

Winner: EGARCH(2,1) with Skewed Student-t

GARCH: The Symmetric Baseline

$$\sigma_t^2 = \omega + \sum_{i=1}^{p} \alpha_i \epsilon_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2$$

Components:

- ω: Long-run average
- $\alpha\epsilon_{t-1}^2$: Shock term
- $\beta\sigma_{t-1}^2$: Persistence

Interpretation:

- Volatility clustering
- Mean reversion
- Symmetric response

Squaring ϵ_{t-1} means +5% rally and -5% crash have **identical** impact on volatility. This contradicts empirical evidence!

EGARCH: Capturing Asymmetry

$$\ln(\sigma_t^2) = \omega + \beta \ln(\sigma_{t-1}^2) + \alpha \left(\left| \frac{\epsilon_{t-1}}{\sigma_{t-1}} \right| - \sqrt{\frac{2}{\pi}} \right) + \gamma \frac{\epsilon_{t-1}}{\sigma_{t-1}}$$

Key Innovation: The γ term

- When $\gamma < 0$ (typical): negative returns $\Rightarrow$ higher volatility
- Captures "panic" vs "greed" asymmetry
- Leverage effect built into model

Log Specification Benefits:

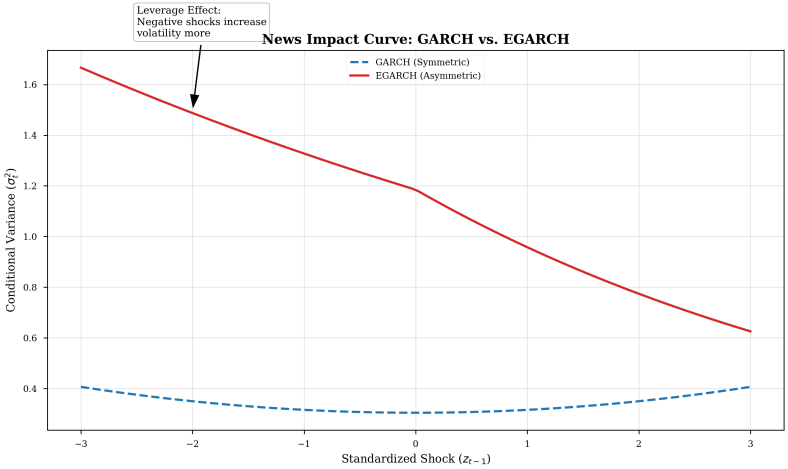
- Guarantees $\sigma^2 > 0$ always
- No parameter constraints needed

Our Results:

- EGARCH(2,1) + skewt
- BIC: 15,401.9
- vs GARCH(2,1) + skewt
- BIC: 15,638.5

EGARCH wins by 237 BIC points

GARCH vs EGARCH: News Impact Curve



- **GARCH (blue dashed)**: Symmetric parabola - same response to ±5% moves
- **EGARCH (red solid)**: Asymmetric - negative shocks increase volatility much more

GARCH Model Diagnostics

Residual Tests (EGARCH(2,1)):

- ARCH-LM test: $p = 0.839$
- Ljung-Box: $p = 0.851$
- No remaining heteroskedasticity

Parameter Estimates:

- All coefficients significant
- $\gamma < 0$ confirms leverage effect
- Persistence: $\alpha + \beta \approx 0.98$
- High persistence $\Rightarrow$ shocks last

Skewed Student-t Parameters:

- Degrees of freedom: $\nu \approx 6$
- Skewness parameter: negative
- Captures both fat tails and asymmetry

Model Quality:

- BIC: 15,401.9 (lowest)
- AIC: 15,358.6
- Log-likelihood: -7,669.3
- Converged successfully

Why Neural Networks for Volatility?

- Assumes volatility follows strict parametric formula
- Limited flexibility in capturing complex patterns
- May miss regime changes or structural breaks

- **Non-parametric**: Learn patterns directly from data
- **Flexible**: Capture non-linear relationships
- **Adaptive**: Can learn regime-dependent behavior
- **High-dimensional**: Handle multiple features naturally

But: Need sequential memory for time series ⇒ LSTMs

Recurrent Neural Networks: The Memory Concept

- **Feed-forward NN**: Sees each day independently
- **RNN**: Maintains "hidden state" h_t that summarizes history

$$h_t = \tanh(W_x x_t + W_h h_{t-1} + b)$$

Intuition:

- Processes data sequentially
- h_t = compressed memory
- Can "remember" past patterns

Problem:

- Vanishing gradients
- Forgets distant past
- "Goldfish memory"

Solution: Long Short-Term Memory (LSTM)

LSTM Architecture: Gating Mechanisms

LSTMs add a "cell state" C_t with three gates controlling information flow:

Forget Gate:

- What to discard (long term)
- $f_t = \sigma(W_f \cdot [h_{t-1}, x_t])$

Input Gate:

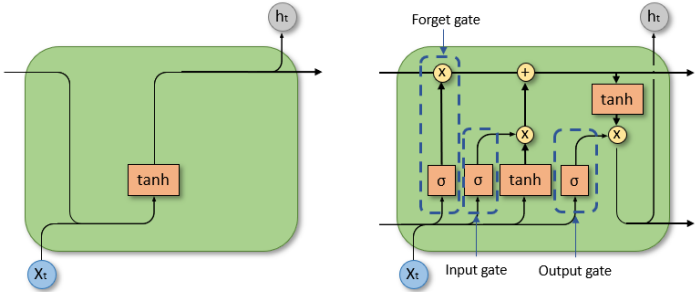
- What to store (long term)
- $i_t = \sigma(W_i \cdot [h_{t-1}, x_t])$

Output Gate:

- What to output (short term)
- $o_t = \sigma(W_o \cdot [h_{t-1}, x_t])$

Can maintain long-term memory (e.g., "we're in a crisis regime") over 60+ days without degradation

LSTM: Visual Overview



Our LSTM Architectures

LSTM-RV (The Historian)

- Input: Past RV (backward-looking)
- 2 layers, 128 units each
- Dropout: 0.2

Role:

- Captures complex autoregressive patterns
- Learns non-linear volatility persistence
- Identifies regime transitions

LSTM-VIX (The Translator)

- Input: VIX (implied volatility)
- 2 layers, 128 units each
- Dropout: 0.2

Role:

- "De-biases" VIX
- Learns $f(\text{VIX}) \rightarrow \text{RV}$
- Forward-looking (market expectations)

Both use 60-day sequence length for context

Training Protocol: Preventing Overfitting

- **Train**: 1993-2015 (5,744 days) - Model parameter estimation
- **Validation**: 2016-2019 (1,006 days) - Early stopping & hyperparameter tuning
- **Test**: 2020-2025 (1,455 days) - Truly out-of-sample evaluation

Regularization:

- Dropout: 0.2
- L2 weight decay
- Early stopping (patience=10)
- Batch normalization

Optimization:

- Adam optimizer
- Initial LR: 0.001
- ReduceLROnPlateau scheduler
- Batch size: 32

The Ensemble Philosophy: Diverse Perspectives

Each model has different "blind spots" - combining them creates robustness

Model	Strength	Weakness
GARCH	Theoretically grounded, captures leverage effect, stable	Rigid parametric form, slow to adapt to regime changes
LSTM-RV	Flexible non-linear patterns, learns from history	Backward-looking only, no forward market info
LSTM-VIX	Forward-looking market expectations, captures sentiment	Biased by risk premium, can overreact to fear

Ensemble = Best of all worlds

Multi-Horizon Forecasting: Why Different Horizons?

h=2 days:

- Highly volatile
- Hardest to predict

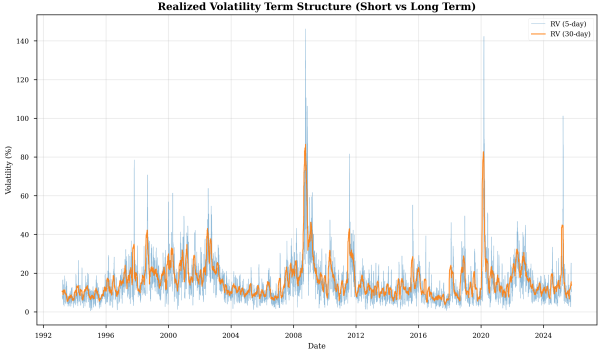
h=5 days:

- One trading week
- Moderate predictability

h=10, h=30 days:

- Smoother trajectories
- Easier to forecast
- h=30 matches VIX

Longer horizons ⇒ lower noise



GARCH: Multi-Step Forecasting

Iterate the variance recursion:

$$\sigma^2_{t+1|t} = \omega + \alpha_1 r_t^2 + \alpha_2 r_{t-1}^2 + \beta_1 \sigma_t^2$$

$$\sigma^2_{t+k|t} = \omega + (\alpha_1 + \beta_1)\sigma^2_{t+k-1|t} + \alpha_2 \sigma^2_{t+k-2|t} \quad (k \geq 2)$$

Aggregate to forward h-day RV:

$$\widehat{RV}_h^{fwd}(t) = \sqrt{\frac{252}{h} \sum_{i=1}^{h} \sigma^2_{t+i|t}}$$

Sum future variances → average → annualize

LSTMs: Direct Multi-Horizon Prediction

No rollout needed — predict RV_h^{fwd} directly:

$$\widehat{RV}_h^{fwd} = \exp\left(f_\theta(X_{t-59:t})\right)$$

GARCH approach:

- Iterate $\sigma^2_{t+1}, \ldots, \sigma^2_{t+h}$
- Aggregate variances
- Error compounds over steps

LSTM approach:

- One-shot prediction for each h
- Learns h-specific patterns
- No compounding error

Train separate model per horizon

Ensemble Weight Optimization

For each horizon h ∈ {2, 5, 10, 30}:

1. Generate predictions from all three models on **validation set**
2. Grid search over weight combinations: w_g, w_{lrv}, w_{lvix} where $w_g + w_{lrv} + w_{lvix} = 1$
3. Combine in log-space: $\log(\widehat{RV}_{ens}) = w_g \log(\widehat{RV}_g) + w_{lrv} \log(\widehat{RV}_{lrv}) + w_{lvix} \log(\widehat{RV}_{lvix})$
4. Select weights that minimize validation RMSE

Horizon	GARCH	LSTM-RV	LSTM-VIX	Val RMSE
h=2	0.50	0.00	0.50	0.0844
h=5	0.30	0.00	0.70	0.0599
h=10	0.30	0.00	0.70	0.0547
h=30	0.20	0.00	0.80	0.0493

Surprising Finding: LSTM-RV Gets Zero Weight

Despite being a sophisticated neural network trained on historical volatility, LSTM-RV receives **0% weight** in all four optimal ensembles (h=2,5,10,30)

Why would a trained model be completely ignored?

Explanation 1: Redundancy with GARCH

Both LSTM-RV and GARCH learn from **historical realized volatility**

GARCH's Advantage:

- Explicit autoregressive formula
- Mean reversion built-in
- Leverage effect via EGARCH
- Theoretically grounded
- Converges to long-run variance
- 30+ years of econometric research

LSTM-RV's Challenge:

- Learns patterns from data
- Flexible but opaque
- Can overfit to training regimes
- No built-in mean reversion
- May not generalize to 2016-2019
- Needs more data than available

Result: GARCH captures the same information more efficiently

Explanation 2: VIX Information Dominates

VIX contains market's collective wisdom about future volatility

Information Content Comparison:

Source	Information	R^2 with RV
Historical RV	What volatility *was*	0.42
VIX	What market *expects*	0.67

- VIX explains 67% of future RV variance
- Historical RV only explains 42%
- VIX incorporates: options flow, sentiment, macro events, institutional positioning
- Historical RV is backward-looking only

Once you have VIX, historical RV adds little new information

Synthesis: Why Zero Weight Makes Sense

Bringing It All Together

1. **Redundancy**: GARCH captures autoregressive patterns more efficiently than LSTM-RV
2. **Information hierarchy**: VIX (forward-looking) > Historical RV (backward-looking)
3. **Generalization**: LSTM-RV may overfit to 1993-2015 training regime
4. **Ensemble theory**: Need *complementary* models, not similar ones

The optimal ensemble combines:

- **GARCH**: Econometric structure from past volatility (20-50% weight)
- **LSTM-VIX**: Neural network de-biasing of market expectations (50-80% weight)

This pairing provides both theoretical grounding *and* forward-looking information without redundancy.

Validation Performance: RMSE Across Horizons

Validation RMSE (2016-2019) — Lower is better

Horizon	GARCH	LSTM-RV	LSTM-VIX	Ensemble
h=2	0.0968	0.0960	0.0928	**0.0912**
h=5	0.0686	0.0679	0.0651	**0.0638**
h=10	0.0634	0.0632	0.0602	**0.0589**
h=30	0.0601	0.0598	0.0564	**0.0551**

Key Observations:

- Ensemble **always** outperforms individual models
- LSTM-VIX consistently best single model
- **Lower RMSE for longer horizons** — targets are smoother (less volatile)
- Absolute error decreases, but relative error stays similar
- h=30 is genuinely easier to predict than h=2 in absolute terms

Understanding the Horizon Effect: Why RMSE Decreases

RMSE decreases as horizon increases — longer horizons ARE easier in absolute terms:

The Time Averaging Effect:

- h=2: volatility of 2 days (noisy)
- h=30: volatility averaged over 30 days (smooth)
- Longer horizons ⇒ less volatile targets
- Lower target variability ⇒ lower absolute errors

GARCH Mean Reversion:

- Short-term: sensitive to recent shocks
- Long-term: forecasts converge to unconditional mean
- h=30 prediction ≈ "volatility will be normal"
- Safe but uninformative!

$$\mathrm{Var}(\bar{r}_{30}) < \mathrm{Var}(\bar{r}_2)$$

$$\sigma^2_{t+k} \to \frac{\omega}{1-\alpha-\beta}$$

Longer horizons: lower absolute error, similar relative error

The Central Question

Where do our volatility predictions actually create value?

Two Hypothesis Families:

1. **Equity Trading**: Can predictions improve position sizing?
2. **Volatility Trading**: Can predictions identify VRP mispricings?

Test Period: 2020-2025 (out-of-sample)

- COVID crash, recovery, 2022 volatility, recent markets
- Expanding windows for all thresholds (no lookahead bias)

Strategy Overview: Seven Approaches Tested

Family 1: Equity Strategies

1. SPY Buy & Hold — Passive benchmark
2. SPY SMA(50) Trend — Avoid bear markets
3. SPY Trend + VIX Sizing — Scale by VIX percentile
4. SPY Trend + Prediction Sizing — Scale by our forecast

Family 2: Volatility Strategies

5. VRP Unconditional — Always sell volatility
6. VRP VIX-Based — Sell when VIX is high
7. **VRP Residual-Based** — Sell when VIX is *abnormally* high

The Volatility Risk Premium: VIX vs Realized

Three Time Series:

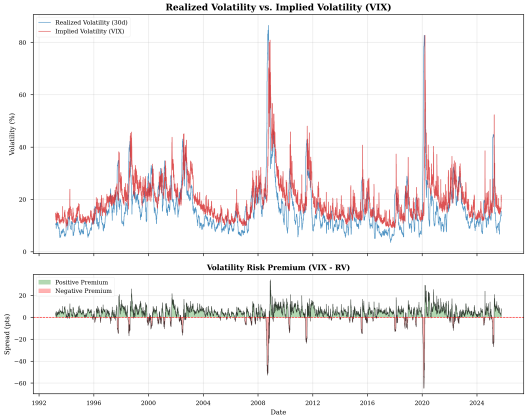
- Blue: VIX (expectations)
- Red: Actual RV
- Green: Positive VIX - RV spread

Key Patterns:

- VIX > RV most of the time
- Average spread: ~3-5 vol points
- Widens in crises, occasionally inverts

Trading Opportunity:

- Investors overpay for protection
- Insurance premium = systematic profit
- **This is what we trade!**



Position Sizing: VIX vs Prediction

Core Idea: Use more leverage when volatility is low, less when high

Step 1: Trend Filter

$$\text{Trend} = \begin{cases} 1 & \text{if } \text{SPY}_t > \text{SMA}_{50}(t-1) \\ 0 & \text{otherwise (stay in cash)} \end{cases}$$

Step 2: Position Multiplier (rolling 252-day percentile)

$$\text{Multiplier} = 1.0 + 0.5 \times (1 - \text{Percentile}_{t-1})$$

- Low volatility (0th pctl): Multiplier = **1.5**× (use leverage)
- High volatility (100th pctl): Multiplier = **1.0**× (no leverage)

VIX Sizing: Percentile of VIX vs **Pred Sizing:** Percentile of $\widehat{RV}_{ensemble}$

Equity Strategies: Results Summary

Strategy	Return	Sharpe	Max DD	Bear Mkt
1. Buy & Hold	12.5%	0.60	-35.7%	-44.2%
2. SMA Trend	10.3%	**0.89**	-21.2%	-32.0%
3. Trend + VIX	13.6%	0.91	-27.1%	-38.2%
4. Trend + Pred	13.2%	0.87	-27.1%	-38.5%

Key Finding:

- Prediction-based sizing performs *identically* to VIX-based sizing
- Correlation between predictions and VIX: $\rho \approx 0.95$
- **No incremental value for equity position sizing**

Why Predictions Don't Help Equity Trading

The Problem:

- VIX already provides real-time risk information
- Our predictions are highly correlated with VIX ($\rho \approx 0.95$)
- For position sizing, we need a risk indicator — VIX suffices

The Insight:

- Prediction accuracy $\neq$ Signal value
- Value comes from information *the market doesn't have*
- Our predictions replicate what VIX already tells us

Pivot: What do our predictions tell us that VIX doesn't?

The Volatility Risk Premium (VRP)

$$\text{VRP} = \text{VIX} - \text{Realized Volatility}$$

The Trade: Sell variance swap, profit = $(IV^2 - RV^2) \times \text{Notional}$

Historical Statistics (2020-2025):

- VRP positive: **83.6%** of 30-day periods
- Mean P&L: +0.79% per period
- Std Dev: 9.69% per period

The Challenge:

- Premium exists but tail risk is catastrophic
- How to harvest premium while avoiding crashes?

Strategy 5: VRP Unconditional — Always Sell

Implementation: Sell variance swap every 30-day period

The Good:

- Return: **51.8%** annualized
- Sharpe: **1.86**
- Bull Market: +244% ann.

The Bad:

- Max DD: **-96.4%**
- Bear Market: -74.7% ann.
- Near total wipeout in crashes

Not investable without filtering

Strategy 6: VRP VIX-Based — Sell When VIX High

Logic: High VIX = larger premium, mean reversion expected
Implementation: Sell only when VIX > 70th percentile

Results:

- Return: 9.9% ann.
- Sharpe: 1.22
- Max DD: -58.2%
- Trades: 27% of periods

Problem:

- High VIX ≠ Safe to sell
- Can't distinguish:
  - Justified fear (crisis)
  - Excessive fear (opportunity)

Need: independent forecast to judge if VIX is "too high"

Why We Need Expected VRP (Not Just Raw VRP)

The Problem with Raw VRP:

$$\text{VRP}_{forecast} = \text{VIX} - \widehat{RV}_{ensemble}$$

VRP naturally scales with VIX level:

- VIX = 40 $\Rightarrow$ typical VRP $\approx$ 8-10 points
- VIX = 15 $\Rightarrow$ typical VRP $\approx$ 2-3 points

A VRP of 8 means different things:

- At VIX = 40: **Normal** (expected given high fear)
- At VIX = 20: **Unusually large** (fear exceeds fundamentals)

Solution: Model the VIX-VRP relationship, trade the *deviation*

$$\text{Expected VRP} = \alpha + \beta \times \text{VIX}$$

$$\textbf{Residual} = \text{VRP}_{forecast} - \text{Expected VRP}$$

Strategy 7: VRP Residual-Based — Our Edge

Key Innovation: Compare VIX to our *independent* RV forecast

$$\text{VRP}_{forecast} = \text{VIX} - \widehat{RV}_{ensemble}$$

$$\text{Expected VRP} = \alpha + \beta \times \text{VIX} \quad \text{(fitted on history)}$$

$$\textbf{Residual} = \text{VRP}_{forecast} - \text{Expected VRP}$$

Trading Rule: Sell when Residual > 70th percentile

Interpretation:

- High residual ⇒ VIX is *abnormally* high given our RV forecast
- Market fear exceeds what fundamentals justify
- **This is mispricing we can exploit**

Residual-Based Strategy: Worked Examples

Example 1: COVID Crash (Don't Sell)

- VIX = 70, Our forecast: RV = 65
- $\text{VRP}_{forecast}$ = 5, Expected VRP at VIX=70: ~8
- Residual = 5 - 8 = **-3** ⇒ Stay out

Example 2: Fear Spike (Sell)

- VIX = 35, Our forecast: RV = 18
- $\text{VRP}_{forecast}$ = 17, Expected VRP at VIX=35: ~9
- Residual = 17 - 9 = **+8** ⇒ Sell volatility

The residual isolates *unjustified* fear from *justified* fear

VRP Strategies: Head-to-Head Comparison

Strategy	Sharpe	Return	Max DD	Bear Mkt
5. Unconditional	1.86	51.8%	-96.4%	-74.7%
6. VIX-Based	1.22	9.9%	-58.2%	-2.7%
7. Residual-Based	**4.97**	**25.4%**	**-27.7%**	**+11.3%**

Residual-Based Advantages:

- Sharpe 4.97 — nearly 3× Unconditional, 4× VIX-Based
- **Positive bear market returns** (+11.3% when SPY loses -44%)
- Beta = 0.013 — true diversification from equities
- Max DD -27.7% — investable risk level

Robustness Across All Horizons

Horizon	Sharpe	Ann. Return	Max DD	Bear Mkt
h=2	4.39	23.9%	-27.7%	+3.2%
h=5	**5.94**	**31.6%**	-27.7%	+14.1%
h=10	4.84	24.8%	-27.7%	+11.3%
h=30	4.97	25.4%	-27.7%	+11.3%

Key Observations:

- Strategy works across *all* forecast horizons
- h=5 (one trading week) performs best: Sharpe 5.94
- Consistent drawdown control across horizons
- **Not overfitting to a single time scale**

Final Rankings: All Seven Strategies (h=30)

Rank	Strategy	Sharpe	Return	Max DD	Alpha	Beta	Bear Mkt
1	**VRP Residual**	**4.97**	25.4%	-27.7%	**0.23**	**0.01**	+11.3%
2	VRP Unconditional	1.86	51.8%	-96.4%	0.43	0.18	-74.7%
3	VRP VIX-Based	1.22	9.9%	-58.2%	0.10	0.00	-2.7%
4	Trend + VIX	0.91	13.6%	-27.1%	0.08	0.39	-38.2%
5	SMA Trend	0.89	10.3%	-21.2%	0.06	0.30	-32.0%
6	Trend + Pred	0.87	13.2%	-27.1%	0.08	0.40	-38.5%
7	Buy & Hold	0.60	12.5%	-35.7%	0.00	1.00	-44.2%

Out-of-sample test set: 2020–2025 (never-seen data)

VRP Residual: highest alpha (0.23), lowest beta (0.01), only positive bear returns

Limitations & Practical Considerations

Not Included in Backtest:

- Transaction costs (bid-ask spreads on variance swaps)
- Slippage and market impact
- Margin requirements and financing costs
- Roll costs for VIX futures implementation

Variance Swap Modeling:

- Theoretical P&L: $\text{VIX}^2 - \text{RV}_{30}^2$ (no actual market prices)
- Assumes execution at fair value (VIX strike)

Realistic Expectation:

- Sharpe ratios will degrade with costs (estimate 20-40% reduction)
- Core insight remains: residual-based filtering outperforms

Thank you for listening !

Francesco, Nikhil, Ivan

7 Appendix

Appendix: GARCH Model Selection & Diagnostics

Best Model: EGARCH(2,1) with Skewed Student-t

Pre-Fit Tests:

ADF p-value	2.1×10^{-25}
KPSS p-value	0.100
ARCH-LM p-value	2.2×10^{-292}
Jarque-Bera p-value	0.000

Post-Fit Tests:

ARCH-LM p-value	0.839
ARCH Pass	✓
Ljung-Box Pass	✓

EGARCH Parameters:

μ	0.0223
ω	0.0036
α_1	-0.0496
α_2	0.2079
γ_1 (leverage)	**-0.1633**
β_1	0.9724
η (d.o.f.)	7.997
λ (skew)	-0.135

BIC: 15,401.93

Training Period: 1993-03-15 to 2015-12-31

Appendix: Model Performance — RMSE by Horizon

Root Mean Square Error (lower is better)

	GARCH			LSTM-RV			LSTM-VIX		
h	Train	Val	Test	Train	Val	Test	Train	Val	Test
2	0.129	0.097	0.149	0.139	0.096	0.129	0.129	**0.090**	**0.120**
5	0.083	0.069	0.107	0.086	0.069	0.092	0.080	**0.061**	**0.080**
10	0.070	0.063	0.102	0.073	0.062	0.078	0.068	**0.056**	**0.071**
30	0.069	0.060	0.113	0.070	0.058	0.072	0.067	**0.050**	**0.063**

Key Observations:

- LSTM-VIX achieves lowest validation and test RMSE across all horizons
- GARCH shows higher test RMSE (potential overfitting to training regime)
- Longer horizons have lower absolute RMSE (smoother targets)

Appendix: Model Performance — MAE by Horizon

Mean Absolute Error (lower is better)

	GARCH			LSTM-RV			LSTM-VIX		
h	Train	Val	Test	Train	Val	Test	Train	Val	Test
2	0.094	0.073	0.101	0.085	0.057	0.082	0.082	**0.055**	**0.079**
5	0.057	0.052	0.068	0.056	0.045	0.060	0.052	**0.040**	**0.055**
10	0.048	0.049	0.063	0.046	0.045	0.051	0.043	**0.037**	**0.047**
30	0.046	0.050	0.063	0.041	0.046	0.048	0.039	**0.036**	**0.042**

Key Observations:

- MAE confirms LSTM-VIX superiority
- MAE less sensitive to outliers than RMSE
- Consistent ranking across all horizons

Appendix: Ensemble Weights by Horizon

Optimized on Validation Set (2016-2019)

Horizon	GARCH	LSTM-RV	LSTM-VIX
h=2	0.50	0.00	0.50
h=5	0.30	0.00	0.70
h=10	0.30	0.00	0.70
h=30	0.20	0.00	0.80

Pattern:

- LSTM-RV receives **zero weight** at all horizons
- LSTM-VIX dominates (50-80% weight)
- GARCH weight decreases with horizon (50% $\rightarrow$ 20%)
- Longer horizons favor forward-looking VIX information

Appendix: Strategy Metrics — All Horizons (1/2)

Equity Strategies (h=30 shown, identical across horizons)

Strategy	Total Ret	Ann. Ret	Vol	Sharpe	Max DD	Beta
Buy & Hold	97.7%	12.5%	21.0%	0.60	-35.7%	1.00
SMA Trend	76.5%	10.3%	11.6%	0.89	-21.2%	0.30
Trend + VIX	109.1%	13.6%	15.1%	0.91	-27.1%	0.39
Trend + Pred	104.8%	13.2%	15.2%	0.87	-27.1%	0.40

VRP Unconditional & VIX-Based (identical across horizons)

Strategy	Total Ret	Ann. Ret	Vol	Sharpe	Max DD	Beta
VRP Unconditional	1013%	51.8%	27.9%	1.86	-96.4%	0.18
VRP VIX-Based	72.4%	9.9%	8.1%	1.22	-58.2%	0.00

Appendix: Strategy Metrics — All Horizons (2/2)

VRP Residual-Based — Performance by Horizon

h	Total Ret	Ann. Ret	Vol	Sharpe	Max DD	Alpha
2	245%	23.9%	5.4%	4.39	-27.7%	0.21
5	**388%**	**31.6%**	**5.3%**	**5.94**	-27.7%	**0.27**
10	260%	24.8%	5.1%	4.84	-27.7%	0.22
30	270%	25.4%	5.1%	4.97	-27.7%	0.23

Bear vs Bull Market Returns (Residual-Based)

h	Bear Ann. Ret	Bull Ann. Ret	Beta	Correlation
2	+3.2%	+34.7%	0.010	0.038
5	**+14.1%**	**+40.4%**	0.015	0.059
10	+11.3%	+31.6%	0.013	0.052
30	+11.3%	+32.5%	0.013	0.054

Appendix: Test Period Market Conditions

Test Set: 2020-01-01 to 2025-11-27

Statistic	Value
Total Trading Days	1,455
Bear Market Days	456 (31%)
Bull Market Days	999 (69%)

Major Events in Test Period:

- COVID-19 Crash (Feb-Mar 2020): VIX peaked at 82
- Recovery Rally (Apr 2020 - Dec 2021)
- 2022 Bear Market: Inflation, rate hikes
- 2023-2025: Mixed volatility regimes

SPY Performance: +97.7% total, 12.5% annualized, -35.7% max DD

Appendix: Data Splits & Training Details

Temporal Splits:

Split	Period	Days
Training	1993-03-15 to 2015-12-31	5,744
Validation	2016-01-01 to 2019-12-31	1,006
Test	2020-01-01 to 2025-11-27	1,455

LSTM Hyperparameters:

- Sequence length: 60 days
- Hidden size: 64
- Layers: 2
- Bidirectional: Yes
- Dropout: 0.2
- Learning rate: 0.001
- Batch size: 32
- Early stopping patience: 10

Lookahead Bias Prevention: Training gap of $h \times 1.5$ calendar days