Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Physical-Measure Merton Jump-Diffusion Calibration Surrogates and Direct Realized Volatility Forecasting: An Empirical and Numerical Audit on S&P 500 ETF (SPY) Returns

Rajdeep Purohit
Indian Institute of Technology Roorkee
Mayank Goswami
mayankfph@iitr.ac.in

Indian Institute of Technology Roorkee https://orcid.org/0000-0001-8829-9980

Research Article

Keywords: Merton Jump-Diffusion, Compound-Poisson Gaussian Mixture, Maximum Likelihood Estimation, Kiefer Singularity, Realized Volatility Forecasting, Ridge Regression, Long Short-Term Memory (LSTM), Out-of-Sample, Temporal Data Leakage Audit

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11729845/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Physical-Measure Merton Jump-Diffusion Calibration Surrogates and Direct Realized Volatility Forecasting: An Empirical and Numerical Audit on S&P 500 ETF (SPY) Returns

Rajdeep Purohit[1], and Mayank Goswami [1, 2, #]

[1] Department of Physics, IIT Roorkee
[2]*Mehta Family School of Data Science and Artificial Intelligence*, IIT Roorkee,
Roorkee, Uttarakhand-247667, India.
[#]mayank.goswami@ph.iitr.ac.in

Abstract

Continuous-time Geometric Brownian Motion (GBM) fails to capture the leptokurtic tails and sudden price discontinuities observed in equity index returns. While Merton's (1976) Jump-Diffusion (MJD) model accounts for sudden shocks by superimposing a compound Poisson jump process onto Brownian diffusion, rolling physical-measure (P) Maximum Likelihood Estimation (MLE) suffers from severe finite-sample weak identification, Kiefer (1978) mixture singularities, and iterative computational overhead. In this study, we design, audit, and evaluate a rigorous, leakage-free dual-task quantitative pipeline on daily S&P 500 ETF (SPY) returns ($N = 1{,}237$ validated sessions, 2019–2023). We explicitly separate two research questions that are frequently conflated in student and practitioner implementations: (Task 1) whether Ridge regression and hybrid sequence-summary Long Short-Term Memory (LSTM) networks can serve as fast calibration surrogates for rolling 126-session multistart compound-Poisson MLE, and (Task 2) whether direct supervised volatility models or calibrated total-volatility proxies can forecast forward 21-session realized volatility better than classical historical persistence and RiskMetrics EWMA ($\lambda_{ewma} = 0.94$) baselines. Across 211 chronological calibration windows (134 Fit, 31 Validation, 46 Test) separated by an explicit $H = 21$-session forward-label purge, raw five-parameter MJD fits hit box boundaries in $165/211$ windows (78.2%), while a conservative Bayesian Information Criterion (BIC) parsimony rule against an exact nested two-parameter Gaussian submodel ($\lambda = 0$) selects jumps in $69/211$ windows (32.7%, concentrated entirely around the 2020 COVID-19 crash). Online surrogate inference achieves a $63.9\times$ wall-clock speedup ($0.359\ ms$ for Mean, Ridge, and LSTM combined versus $22.939\ ms$ median multistart MLE), but 5-session parameter persistence ($\hat{\theta}_t = \hat{\theta}_{t-5}$) outperforms ML surrogates on rolling 126-session labels (validation standardized MSE 0.0268 vs. 1.8052 for Ridge and 2.2362 for LSTM) due to 96% window overlap. On the held-out 2023 test split ($n = 46$), the direct hybrid LSTM achieves the lowest future 21-session realized volatility RMSE (0.026553, MAE 0.023191, Campbell–Thompson $R^2_{oos} = +16.95\%$, $r = 0.5049$), followed by direct Ridge ($RMSE = 0.028088$, $R^2_{oos} = +7.07\%$) and EWMA ($RMSE = 0.028179$, $R^2_{oos} = +6.46\%$), outperforming 21-session historical volatility ($RMSE = 0.029137$). However, because chronological validation during the high-volatility 2022 bear market selected the 126-session calibrated total-volatility proxy (mle_total_vol, validation $RMSE = 0.056528$), which subsequently lagged the 2023 volatility decline ($test\ RMSE = 0.050488$, $R^2_{oos} = -200.26\%$), freezing model selection prior to test evaluation prevents claiming consistent out-of-sample forecasting superiority without a fresh post-2023 holdout.

Keywords: Merton Jump-Diffusion; Compound-Poisson Gaussian Mixture; Maximum Likelihood Estimation; Kiefer Singularity; Realized Volatility Forecasting; Ridge Regression; Long Short-Term Memory (LSTM); Out-of-Sample R^2; Temporal Data Leakage Audit.

1 Introduction

Financial asset returns routinely violate the IID Gaussian assumptions of classical continuous-time diffusion models (**BlackScholes1973?**). Empirical daily equity log returns exhibit pronounced excess kurtosis, negative skewness, volatility clustering, and sudden discontinuous price gaps triggered by macroeconomic announcements, geopolitical shocks, or liquidity crises. To accommodate discontinuous sample paths, Merton (1976) (**Merton1976?**) introduced the jump-diffusion framework, combining a continuous Wiener diffusion with a compound Poisson jump process whose jump magnitudes follow a normal distribution in log-price space.

Despite its theoretical elegance, estimating the five parameters of the Merton Jump-Diffusion (MJD) model—diffusive log drift μ_{log}, diffusive volatility σ, Poisson jump intensity λ, mean log jump size μ_j, and log jump standard deviation σ_j—from discrete historical return series poses formidable statistical and computational challenges:

1. **Mixture Likelihood Unboundedness and Boundary Spikes:** As proven by Kiefer (1978) (**Kiefer1978?**) and demonstrated specifically for jump-diffusion models by Honoré (1998) (**Honore1998?**), the unconditional density of discrete returns under MJD is an infinite Poisson-weighted mixture of normal distributions with unequal variances ($\sigma^2 + k\sigma_j^2$). In unconstrained Maximum Likelihood Estimation (MLE), centring a jump component on a single return observation while letting $\sigma_j \to 0$ drives the likelihood toward infinity, producing spurious boundary spikes.

2. **Weak Finite-Sample Identification and Nonregular Zero-Jump Testing:** Disentangling continuous diffusive variance σ^2 from frequent small jumps requires either high-frequency infill asymptotics ($\Delta t \to 0$) (**AitSahalia2004?**) or long stationary samples. In short rolling windows (e.g., 21 trading days), five parameters are severely weakly identified. Moreover, testing the null hypothesis of zero jumps ($\lambda = 0$) places λ on the boundary of the parameter space where the nuisance parameters (μ_j, σ_j) become mathematically unidentified, invalidating standard χ^2 Wilks likelihood-ratio asymptotics (**McLachlanPeel2000?**).

3. **Iterative Calibration Cost vs. Real-Time Risk Evaluation:** Rolling multistart constrained optimization (L-BFGS-B) across hundreds of assets and rolling dates requires repeated evaluations of truncated Poisson mixture densities and score gradients. Recent quantitative finance literature (**Hernandez2017?**; **Liu2019?**; **Horvath2021?**) has therefore explored replacing slow iterative numerical calibration with trained machine-learning (ML) and deep-learning (DL) surrogates.

However, a major methodological flaw pervades many introductory implementations of "Deep Merton" calibration surrogates (including the instructional guide supplied alongside our raw dataset): they conflate *imitating historical MLE parameter labels* with *forecasting future realized risk*, use one-jump Bernoulli approximations that fail when λ is non-negligible, clip invalid predictions post-hoc, evaluate training loss with dropout active against validation loss with dropout disabled, and allow forward target windows to leak across train/validation/test boundaries.

In this paper, we present a complete scientific reconstruction, mathematical remediation, and empirical audit of the SPY Deep Merton project (finance_project_completed, version 2.0.0). Every claim, equation, table, and diagnostic reported in this manuscript is backed by deterministic Python 3.12 code (src/finance_project/), 26 automated pytest unit and integration tests, and 11 independent artifact verification checks (src/verify_project.py).

2 Problem Statement

Let $\{P_0, P_1, \dots, P_N\}$ denote observed daily closing prices of the SPDR S&P 500 ETF Trust (SPY) over $N + 1$ consecutive trading sessions, with exact single-session logarithmic returns defined for $t \in \{1, \dots, N\}$ as:

$$r_t = ln\left(\frac{P_t}{P_{t-1}}\right).$$

At a sequence of chronological forecast origins $t \in T$ spaced at stride $\Delta = 5$ sessions, only historical returns $\{r_1, \dots, r_t\}$ available strictly after market close on date t may enter feature construction or historical calibration.

We formulate and separate two distinct quantitative problems:

- **Problem 1 (Historical Physical-Measure Calibration Approximation):** Given a trailing window of $W = 126$ observed returns $r_{t-W+1:t} = (r_{t-125}, \dots, r_t)^\top$, classical multistart constrained MLE combined with a Bayesian Information Criterion (BIC) parsimony check produces a reference parameter vector $\hat{\theta}_t = (\hat{\mu}_{log,t}, \hat{\sigma}_t, \hat{\lambda}_t, \hat{\mu}_{j,t}, \hat{\sigma}_{j,t})^\top \in \Theta \subset R^5$. Can supervised regression models $F_\phi(X_t, s_t)$—where $X_t \in R^{S\times 2}$ $(S = 21)$ is a bivariate sequence of returns and absolute returns and $s_t \in R^{13}$ is a vector of trailing summary statistics—accurately and rapidly approximate the mapping $(X_t, s_t) \mapsto \hat{\theta}_t$ out of sample?

- **Problem 2 (Forward H-Session Realized Volatility Forecasting):** Let the future realized volatility over the next $H = 21$ trading sessions $(r_{t+1}, \dots, r_{t+H})$ be defined by the annualized demeaned sample standard deviation $(ddof = 1)$:

$$y_t^{vol} = \sqrt{A} \cdot \sqrt{\frac{1}{H-1}\sum_{\tau=1}^{H} (r_{t+\tau} - r^-_{t+1:t+H})^2}, \qquad r^-_{t+1:t+H} = \frac{1}{H}\sum_{\tau=1}^{H} r_{t+\tau}, \qquad A = 252.$$

  Can either (a) indirect total-volatility proxies $\hat{\sigma}_{total,ann}(\hat{\theta}_t) = \sqrt{A(\hat{\sigma}_t^2 + \hat{\lambda}_t(\hat{\mu}_{j,t}^2 + \hat{\sigma}_{j,t}^2))}$ derived from calibrated or surrogate parameters, or (b) direct supervised volatility models $G_\psi(X_t, s_t)$ trained directly against y_t^{vol}, achieve lower out-of-sample Root Mean Squared Error (RMSE) than simple 21-session historical volatility persistence $(Hist_{21d})$ and RiskMetrics Exponentially Weighted Moving Average $(EWMA_{0.94})$?

3 Research Objectives

We distinguish our scientific investigation objectives, software engineering objectives, and quantitative evaluation criteria:

1. **Primary Scientific Objectives:**
    - Formulate an exact, numerically stable physical-measure (P) compound-Poisson normal mixture log-likelihood using adaptive Poisson tail truncation $(\tau = 10^{-12})$, xlogy, gammaln, and logsumexp, paired with exact analytical score gradients.

- Resolve small-sample MLE boundary pathologies by expanding the calibration window from 21 to 126 sessions, evaluating an exact nested 2-parameter Gaussian submodel ($\lambda = \mu_j = \sigma_j = 0$), and applying a conservative BIC selection rule while retaining all raw 5-parameter fits and optimizer diagnostics.
- Decouple historical parameter imitation (Task 1) from forward 21-session realized volatility forecasting (Task 2), evaluating Ridge regression (L_2 shrinkage) and hybrid sequence-summary LSTM networks on both tasks against rigorous domain baselines.

2. **Secondary Engineering & Methodological Objectives:**
   - Eliminate post-hoc prediction clipping by introducing smooth parameter link transformations (ParameterTransform: $ln\sigma$ and $\sqrt{\lambda}, \sqrt{\sigma_j}$ on encoding; exp and $(\cdot)^2$ folding on decoding) and a log-volatility link (lny_t^{vol}) for direct volatility models.
   - Enforce strict chronological boundary purging so that every training origin's 21-session forward target finishes strictly before the first validation origin ($fit[-1] + 21 < val[0]$), and every validation origin's forward target finishes strictly before the first test origin ($val[-1] + 21 < test[0]$).
   - Eliminate dropout-induced training-versus-validation loss distortion by evaluating both training MSE and validation MSE in deterministic inference mode (net.eval(), dropout disabled) in identical standardized latent space immediately after each gradient update.

4 Theoretical Background

4.1 Continuous-Time Merton Jump-Diffusion Dynamics

Under the physical probability measure P, let $(\Omega, F, \{F_t\}_{t\geq 0}, P)$ be a filtered probability space supporting a standard one-dimensional Wiener process W_t, a time-homogeneous Poisson counting process N_t with arrival intensity $\lambda \geq 0$, and an IID sequence of normal log-jump variables $\{Y_i\}_{i\geq 1}$ with $Y_i \sim N(\mu_j, \sigma_j^2)$, where W_t, N_t, and $\{Y_i\}$ are mutually independent. In Merton's (1976) (**Merton1976?**) compensated arithmetic-drift SDE specification, the asset price S_t evolves as:

$$\frac{dS_t}{S_{t^-}} = (\mu - \lambda\kappa)\, dt + \sigma\, dW_t + (e^{Y_t} - 1)\, dN_t,$$

where $\mu \in R$ is the instantaneous expected arithmetic return per unit time, $\sigma > 0$ is the continuous Brownian diffusion volatility, and $\kappa = E[e^{Y_i} - 1] = exp\left(\mu_j + \frac{1}{2}\sigma_j^2\right) - 1$ is the expected relative jump size.

Applying Itô's Lemma for jump-diffusion semimartingales to $X_t = lnS_t$ separates the continuous quadratic variation $d[X^c, X^c]_t = \sigma^2\, dt$ from discrete jumps $\Delta X_t = ln(S_t/S_{t^-}) = Y_{N_t}$:

$$dlnS_t = \left(\mu - \frac{1}{2}\sigma^2 - \lambda\kappa\right) dt + \sigma\, dW_t + Y_t\, dN_t.$$

Integrating Eq. [eq:ito_log_sde] over one observed trading session ($\Delta t = 1$) yields the exact discrete-time single-session physical return model implemented in MODELING.md and merton.py:

$$r_t = \mu_{log} + \sigma\varepsilon_t + \sum_{i=1}^{N_t} Y_i, \qquad \varepsilon_t \sim N(0,1), \quad N_t \sim Poisson(\lambda), \quad Y_i \sim^{iid} N(\mu_j, \sigma_j^2),$$

where $\mu_{log} \equiv \mu - \frac{1}{2}\sigma^2 - \lambda\kappa$ represents the uncompensated diffusive log drift per session. Because $\mu_{log} \in R$ is a free location parameter in physical-measure return estimation, estimating $\theta = (\mu_{log}, \sigma, \lambda, \mu_j, \sigma_j)^\top$ directly is an exact one-to-one reparameterization of $(\mu, \sigma, \lambda, \mu_j, \sigma_j)^\top$ without requiring risk-neutral (Q) option-pricing assumptions.

4.2 Law of Total Variance and Annualized Total Volatility

By Wald's first identity, the unconditional expected log return per session is $E[r_t] = \mu_{log} + \lambda\mu_j$. To obtain the unconditional single-session return variance $V_{daily} = Var(r_t)$, we condition on the random jump count N_t and apply the Law of Total Variance:

$$Var\left(\sum_{i=1}^{N_t} Y_i\right) = E[Var(N_t)] + Var(E[N_t]) = E[N_t\sigma_j^2] + Var(N_t\mu_j) = \lambda\sigma_j^2 + \lambda\mu_j^2 = \lambda(\mu_j^2 + \sigma_j^2).$$

By independence of the Brownian increment $\sigma\varepsilon_t$ and the compound Poisson sum, the total daily variance and annualized total volatility ($A = 252$ sessions/year) implemented in src/finance_project/units.py are:

$$V_{daily}(\theta) = \sigma^2 + \lambda\left(\mu_j^2 + \sigma_j^2\right), \qquad \sigma_{total,annual}(\theta) = \sqrt{252 \cdot V_{daily}(\theta)} = \sqrt{252\left(\sigma^2 + \lambda(\mu_j^2 + \sigma_j^2)\right)}.$$

Critical Dimensional Invariance (units.py): When converting session parameters to annual units, $\mu_{log,ann} = A\mu_{log}$, $\sigma_{ann} = \sqrt{A}\sigma$, and $\lambda_{ann} = A\lambda$, whereas the jump-size distribution parameters μ_j and σ_j are dimensionless per-jump quantities and remain **unchanged**. Consequently, $\sigma_{ann}^2 + \lambda_{ann}(\mu_j^2 + \sigma_j^2) = A\sigma^2 + A\lambda(\mu_j^2 + \sigma_j^2) = AV_{daily}$, confirming scale consistency (test_units_and_horizon_scaling).

5 Literature Review

Our methodology synthesizes four distinct streams of quantitative finance, statistical mixture theory, and machine learning literature:

1. **Continuous Diffusion and Jump-Diffusion Asset Pricing:** Black and Scholes (1973) (**BlackScholes1973?**) established the continuous-time GBM benchmark ($\lambda = 0$), while Merton (1976) (**Merton1976?**) derived the Poisson-Gaussian mixture option pricing formula under diversifiable jump risk. Subsequent extensions by Bates (1996) (**Bates1996?**) combined Heston stochastic volatility with jumps (SVJ), and Kou (2002) (**Kou2002?**) replaced Gaussian jumps with asymmetric double-exponential jumps. Crucially, Bates (1996) documented the tension between risk-neutral (Q) option-implied jump parameters and physical (P) time-series estimates.

2. **Mixture Likelihood Singularities and Nonregular Boundary Selection:** Kiefer (1978) (**Kiefer1978?**) proved that while normal mixtures with unequal variances have globally unbounded likelihoods at the boundary ($v_k \to 0$), a consistent, asymptotically efficient interior root exists. Honoré (1998) (**Honore1998?**) demonstrated that unconstrained jump-diffusion MLE degenerates at single observations and proved that the popular Bernoulli one-jump approximation $(1 - \lambda)\phi_0 + \lambda\phi_1$ biases jump intensity estimates by over 200%. Aït-Sahalia (2004) (**AitSahalia2004?**) proved that disentangling diffusive volatility σ^2 from Lévy jumps relies on high-frequency sampling

($\Delta t \to 0$). For model selection, Schwarz (1978) (**Schwarz1978?**) derived the Bayesian Information Criterion ($BIC = 2NLL + klnn$), while McLachlan and Peel (2000) (**McLachlanPeel2000?**) established that testing zero mixture weight ($\lambda = 0$) violates Laplace regularity because (μ_j, σ_j) vanish under the null, making BIC a conservative parsimony heuristic rather than an exact χ^2 test.

3. **Realized Volatility Modeling and Multi-Horizon Cascades:** Andersen, Bollerslev, Diebold, and Labys (2001, 2003) (**Andersen2001?**; **Andersen2003?**) demonstrated that realized volatility (RV) is heavily right-skewed whereas logarithmic realized volatility ($lnRV$) is approximately Gaussian, motivating our logarithmic transformation $z_t = ln(y_t^{vol})$. Corsi (2009) (**Corsi2009?**) introduced the Heterogeneous Autoregressive Realized Volatility (HAR-RV) model, showing that combining multi-scale trailing volatility summaries $(1d, 5d, 22d)$ captures long-memory volatility persistence, directly motivating our multi-horizon summary feature vector $s_t \in R^{13}$.

4. **Regularized Regression, Sequence Learning, and Calibration Surrogates:** Hoerl and Kennard (1970) (**HoerlKennard1970?**) proved the Ridge Existence Theorem, showing that L_2 shrinkage $\hat{\beta}_{ridge} = (X^\top X + \alpha I)^{-1} X^\top Y$ reduces estimation variance under multicollinearity. Hochreiter and Schmidhuber (1997) (**Hochreiter1997?**) introduced the Long Short-Term Memory (LSTM) architecture with a linear Constant Error Carousel to prevent vanishing gradients, and Loshchilov and Hutter (2019) (**Loshchilov2019?**) proved that decoupled weight decay (AdamW) improves generalization over L_2-regularized Adam. In quantitative finance, Hernandez (2017) (**Hernandez2017?**), Liu et al. (2019) (**Liu2019?**), and Horvath, Muguruza, and Tomas (2021) (**Horvath2021?**) demonstrated $100\times$–$1000\times$ calibration speedups using neural surrogates, while warning that neural surrogates degrade when market regimes shift outside the training distribution. Finally, Campbell and Thompson (2008) (**CampbellThompson2008?**) formalized the out-of-sample R^2_{oos} metric relative to recursive historical benchmarks.

6 Proposed Methodology and Mathematical Formulation

6.1 Exact Truncated Compound-Poisson Mixture Log-Likelihood

Conditional on $N_t = k \in \{0,1,2,\dots\}$ jumps occurring during session t, the return r_t is normally distributed with conditional mean $\mu_k = \mu_{log} + k\mu_j$ and conditional variance $v_k = \sigma^2 + k\sigma_j^2$. Let $w_k(\lambda) = P(N_t = k) = e^{-\lambda}\lambda^k/k!$. To evaluate $lnw_k(\lambda)$ stably for all $\lambda \geq 0$ (including exact $\lambda = 0$, where $0ln0 \equiv 0$), merton.py uses SciPy's xlogy and gammaln:

$$lnw_k(\lambda) = -\lambda + xlogy(k, \lambda) - ln\Gamma(k+1), \qquad k = 0,1,\dots,K-1.$$

Given the configured upper bound $\lambda_{max} = 1.0\ session^{-1}$ and tail tolerance $\tau = 10^{-12}$, mixture_terms selects the smallest integer $K \geq 1$ such that the discarded Poisson survival mass satisfies $\sum_{k=K}^{\infty} w_k(\lambda_{max}) \leq \tau$, yielding $K = 15$ terms across all optimization steps. Define the log-component matrix $\ell_{t,k}(\theta)$ for observation r_t and jump count $k \in \{0,\dots,K-1\}$:

$$\ell_{t,k}(\theta) = lnw_k(\lambda) - \frac{1}{2}ln(2\pi v_k) - \frac{\delta_{t,k}^2}{2v_k}, \qquad \delta_{t,k} = r_t - \mu_{log} - k\mu_j, \qquad v_k = \sigma^2 + k\sigma_j^2.$$

Using the logsumexp identity $ln\sum_{k=0}^{K-1} e^{\ell_{t,k}} = m_t + ln\sum_{k=0}^{K-1} e^{\ell_{t,k}-m_t}$ with $m_t = max_k \ell_{t,k}$, the Negative Log-Likelihood (NLL) over a window of $W = 126$ returns is:

$$L_{NLL}(\theta) = -\sum_{t=1}^{W} \quad lnf(r_t \mid \theta) = -\sum_{t=1}^{W} \quad logsumexp_{k=0}^{K-1}\big(\ell_{t,k}(\theta)\big).$$

6.2 Analytical Score Gradients and Multistart Scaled L-BFGS-B

Let $\gamma_{t,k}(\theta) = P(N_t = k \mid r_t, \theta) = exp\big(\ell_{t,k}(\theta) - lnf(r_t \mid \theta)\big)$ denote the posterior responsibility of jump count k for observation t, satisfying $\sum_{k=0}^{K-1} \quad \gamma_{t,k} = 1$. Differentiating Eq. [eq:nll] analytically (objective_gradient in merton.py) with $g_{t,k}^{mean} = \delta_{t,k}/v_k$ and $g_{t,k}^{var} = \delta_{t,k}^2/v_k^2 - 1/v_k$ yields the exact partial derivatives of $L_{NLL}(\theta)$:

$$\frac{\partial L_{NLL}}{\partial \mu_{log}} = -\sum_{t=1}^{W} \quad \sum_{k=0}^{K-1} \quad \gamma_{t,k} \, \frac{\delta_{t,k}}{v_k}, \frac{\partial L_{NLL}}{\partial \sigma} = -\sum_{t=1}^{W} \quad \sum_{k=0}^{K-1} \quad \gamma_{t,k} \, \sigma\left(\frac{\delta_{t,k}^2}{v_k^2} - \frac{1}{v_k}\right), \frac{\partial L_{NLL}}{\partial \lambda}$$
$$= -\sum_{t=1}^{W} \quad \sum_{k=0}^{K-1} \quad \gamma_{t,k} \left(\frac{k}{\lambda} - 1\right), \frac{\partial L_{NLL}}{\partial \mu_j} = -\sum_{t=1}^{W} \quad \sum_{k=0}^{K-1} \quad \gamma_{t,k} \, k \, \frac{\delta_{t,k}}{v_k}, \frac{\partial L_{NLL}}{\partial \sigma_j}$$
$$= -\sum_{t=1}^{W} \quad \sum_{k=0}^{K-1} \quad \gamma_{t,k} \, k\sigma_j \left(\frac{\delta_{t,k}^2}{v_k^2} - \frac{1}{v_k}\right).$$

Because the natural magnitudes of $(\mu_{log}, \sigma, \lambda, \mu_j, \sigma_j)$ differ by orders of magnitude, optimization is performed on scaled coordinates $q = \theta \oslash s$ with scale vector $s = (0.01, 0.01, 0.1, 0.02, 0.02)^\top$ and chain-rule gradient $\nabla_q L_{NLL} = \nabla_\theta L_{NLL} \odot s$. For jump fits, λ has an interior lower bound $\lambda \geq 10^{-8}$ solely so k/λ is finite, while the exact $\lambda = 0$ boundary is evaluated separately via the closed-form bounded Gaussian MLE $\hat{\theta}_{Gauss} = (clip(r^-, -0.03, 0.03), clip(s_r, 10^{-4}, 0.15), 0, 0, 0)^\top$. Three deterministic heuristic starts—$(f_\sigma, \lambda_0, sign) \in \{(0.8, 0.03, -1), (0.8, 0.03, +1), (0.6, 0.20, 0)\}$—plus a warm start from the preceding window's full MLE fit are optimized via L-BFGS-B. The full MLE is guaranteed never to be worse than the nested Gaussian: $L_{NLL}^{full} = min(L_{NLL}^{MJD}, L_{NLL}^{Gauss})$. Finally, BIC compares the 5-parameter and 2-parameter models on $W = 126$ observations:

$$BIC_{MJD} = 2L_{NLL}^{full} + 5ln(126), \qquad BIC_{Gauss} = 2L_{NLL}^{Gauss} + 2ln(126),$$

selecting the jump model if and only if $BIC_{MJD} < BIC_{Gauss}$ (i.e., when the 3 jump parameters improve NLL by more than $\frac{3}{2} ln(126) \approx 7.2544 \, nats$).

6.3 Constrained Output Links, Ridge Regression, and Hybrid LSTM

To ensure that surrogate predictions always satisfy $\hat{\sigma} > 0$, $\hat{\lambda} \geq 0$, and $\hat{\sigma}_j \geq 0$ without post-hoc clipping, ParameterTransform maps θ to latent coordinates $z = g(\theta)$:

$$z = g(\theta) = \big(\mu_{log}, \; ln\sigma, \; \sqrt{\lambda}, \; \mu_j, \; \sqrt{\sigma_j}\big)^\top, \qquad g^{-1}(z) = \big(z_1, \; e^{z_2}, \; z_3^2, \; z_4, \; z_5^2\big)^\top.$$

Notice that squaring z_3 and z_5 is a smooth, differentiable many-to-one link that admits exact zero ($\hat{\lambda} = 0, \hat{\sigma}_j = 0$) and folds negative latent predictions smoothly rather than truncating them at an arbitrary threshold. For Task 2 (direct volatility), targets are encoded via $z^{vol} = ln(y^{vol})$ and decoded via $\hat{y}^{vol} = exp(\hat{z}^{vol}) > 0$.

All sequence channels, summary features, and encoded targets are standardized using StandardScaler objects fit strictly on the $n_{fit} = 134$ training origins.

- **Ridge Regression:** Concatenates the $S = 21$ standardized signed return lags with the 13 standardized summary statistics into a design matrix $\tilde{X} \in R^{n\times 34}$ and minimizes $\| \tilde{Y} - (\tilde{X}W + 1b^{\top}) \|_F^2 + \alpha \| W \|_F^2$ over $\alpha \in \{0.1,1.0,10.0,100.0,1000.0\}$.
- **Hybrid Sequence-Summary LSTM (Net):** Processes the standardized sequence tensor $\tilde{X}^{seq} \in R^{B\times 21\times 2}$ through a single-layer PyTorch nn.LSTM(2, hidden, batch_first=True) with hidden size $H_{lstm} \in \{8,16\}$. The terminal hidden state $h_{21} \in R^{B\times H_{lstm}}$ is concatenated with the standardized summary statistics $\tilde{s} \in R^{B\times 13}$, regularized via nn.Dropout(p=0.15), and projected linearly via nn.Linear(H_lstm + 13, d_out). Training uses full-batch AdamW ($lr = 0.003$, $weight_decay = 0.001$), gradient norm clipping at 1.0, max 200 epochs, and early stopping with patience 25.

7 Experimental Setup and Implementation Details

7.1 Data Provenance and Chronological Purged Partitions

Table 1 summarizes the raw dataset audit (results/data_audit/data_audit.json) and chronological split boundaries (run_manifest.json). Out of 990 raw training rows (2019-01-31 to 2023-01-04) and 248 raw test rows (2023-01-05 to 2023-12-29), only the initial row (2019-01-31) lacks a preceding closing price and is excluded from the return table, leaving $N = 1{,}237$ validated sessions. Recomputed log returns match the supplied returns column to within a maximum absolute discrepancy of 1.1701×10^{-8} (serialization rounding), verified against a strict 10^{-6} (0.01 bps) audit tolerance.

Chronological Forecast Partitions and Forward-Target Purge Boundaries ($W = 126$, $S = 21$, $H = 21$, $\Delta = 5$).

Partition	Origins (n)	First Origin	Last Origin	Last Target End	Boundary Purge Verification
Fit (Train)	134	2019-08-01	2022-03-22	2022-04-21	Precedes Val origin 1 (2022-04-27) by 4 sessions
Validation	31	2022-04-27	2022-11-30	2022-12-30	Precedes Test origin 1 (2023-01-05) by 3 sessions
Test (Holdout)	46	2023-01-05	2023-11-28	2023-12-28	4 final origins without $H = 21$ future sessions omitted

7.2 Modular Software Architecture and Verification Harness

The software (src/finance_project/) is organized into 9 single-responsibility modules: data.py (CSV validation and purged partitions), units.py (variance identities and ParameterTransform), merton.py

(adaptive Poisson mixture log-PDF, analytic score gradients, and multistart L-BFGS-B), features.py (EWMA, 21×2 sequences, and 13 trailing statistics), models.py (fit-only Preprocessor, Ridge, and hybrid Net), evaluation.py (date-keyed 1-to-1 joins and Campbell–Thompson R^2_{oos}), pipeline.py (chronological execution locking choices before test evaluation), plotting.py (5 diagnostic figures), and verification.py (11 independent post-run audits).

8 Empirical Results

8.1 MLE Calibration and Identification Diagnostics

Across all 211 current calibration windows (134 Fit, 31 Validation, 46 Test) plus 2 persistence context windows (213 total calibrations), multistart L-BFGS-B achieved 100% convergence (0 failed starts) and kept omitted Poisson tail mass below $1.11 \times 10^{-16} \leq 10^{-12}$. However, raw five-parameter MJD fits touched box boundaries in $165/211$ current windows (78.2%): specifically, σ_j hit a boundary in 110 windows, (λ, σ_j) jointly hit boundaries in 44 windows, and λ alone hit its upper bound (1.0) in 11 windows. When compared against the exact 2-parameter Gaussian submodel via BIC, the jump component was selected in $69/211$ windows (32.7%), all 69 of which occurred during the Fit partition ($69/134 = 51.5\%$, spanning the 2020 COVID-19 crash and aftermath), whereas BIC selected the Gaussian submodel ($\hat{\lambda} = \hat{\mu}_j = \hat{\sigma}_j = 0$) in $31/31$ Validation windows and $46/46$ Test windows.

8.2 Hyperparameter Tuning and Model Selection on Validation

Table 2 reports the complete validation hyperparameter grid search from results/metrics/model_candidates.csv. On Task 1 (parameters), Ridge selected $\alpha^* = 1.0$ (validation standardized MSE 1.805169) and LSTM selected $H^* = 8$ hidden units at epoch 183 (validation MSE 2.236159), whereas 16 hidden units overfit immediately after epoch 1 (validation MSE 3.544200). However, 5-session parameter persistence ($\hat{\theta}_t = \hat{\theta}_{t-5}$) achieved a validation standardized MSE of 0.026775, winning overall Task 1 selection (selection.json). On Task 2 (volatility), direct Ridge selected heavy shrinkage $\alpha^* = 100.0$ (validation native volatility MSE 0.008303, $RMSE = 0.091123$) and direct LSTM selected $H^* = 16$ hidden units at epoch 50 (validation MSE 0.006356, $RMSE = 0.079724$). Across all 10 candidate volatility forecasters on validation, mle_total_vol achieved the lowest validation RMSE (0.056528) and was locked as the overall validation-selected forecast model before test evaluation.

Hyperparameter Tuning Grid and Validation Scores (results/metrics/model_candidates.csv).

Task	Model Family	Ridge α	LSTM Hidden	Best Epoch	Validation Selection Score
Parameters	Ridge Regression	0.1	—	—	1.864830 (std. latent MSE)
Parameters	Ridge Regression	**1.0**	—	—	1.805169 **(Selected Ridge)**
Parameters	Ridge Regression	10.0 / 100 / 1000	—	—	2.122854 / 2.750867 / 3.168554

Task	Model Family	Ridge α	LSTM Hidden	Best Epoch	Validation Selection Score
Parameters	Hybrid LSTM	—	**8**	**183**	2.236159 **(Selected LSTM)**
Parameters	Hybrid LSTM	—	16	1	3.544200 (early overfitting)
Volatility	Direct Ridge	0.1 / 1.0 / 10.0	—	—	0.026015 / 0.014963 / 0.010920 (native MSE)
Volatility	Direct Ridge	**100.0**	—	—	0.008303 ($RMSE =$ 0.091123, **Selected Ridge**)
Volatility	Direct Ridge	1000.0	—	—	0.009635 ($RMSE =$ 0.098156)
Volatility	Direct LSTM	—	8	35	0.006734 ($RMSE =$ 0.082061)
Volatility	Direct LSTM	—	**16**	**50**	0.006356 ($RMSE =$ 0.079724, **Selected LSTM**)

8.3 Task 1 Results: Parameter Imitation and Reconstructed Total Volatility

Table [tab:param_metrics] presents the held-out 2023 test split ($n = 46$) parameter approximation errors (results/metrics/parameter_metrics.csv) alongside reconstructed historical total-volatility errors (results/metrics/reconstructed_volatility.csv). Because adjacent 126-session windows spaced by $\Delta = 5$ sessions share 121 of 126 observations (96.0% overlap), 5-session persistence achieves $R^2_{oos} = 0.9496$ on μ_{log} and $R^2_{oos} = 0.9941$ on σ ($r = 0.9963$). Among the machine-learning surrogates that do not require running MLE at $t - 5$, the hybrid LSTM ($H = 8$) outperforms Ridge ($\alpha = 1.0$) on diffusive volatility σ ($RMSE = 0.003212, R^2_{oos} = +0.2599, r = 0.3111$ vs. Ridge $RMSE = 0.004095, R^2_{oos} = -0.2033$) and on reconstructed historical total volatility ($RMSE = 0.048670$, $r = 0.5502$ vs. Ridge $RMSE = 0.052367$).

8.4 Task 2 Results: Forward 21-Session Realized Volatility Forecasting

Table [tab:vol_metrics] reports the complete validation ($n = 31$), test ($n = 46$), and nonoverlapping test subsequence ($n = 10$, every 5th origin) performance across all 10 evaluated models (future_volatility.csv and nonoverlapping_volatility.csv).

9 Discussion and Comparison with Existing Work

Three scientific insights emerge from Table [tab:vol_metrics]:

1. **Direct Supervision Beats Indirect Parameter-Surrogate Volatility:** Training a hybrid LSTM directly on log forward realized volatility (direct_lstm, test $RMSE = 0.026553$, $R^2_{oos} = +16.95\%$; nonoverlapping $RMSE = 0.022446$, $R^2_{oos} = +25.73\%$, $r = 0.8760$) substantially outperforms reconstructing total volatility from surrogate parameter predictions (lstm_total_vol, test $RMSE = 0.029806$; ridge_total_vol, test $RMSE = 0.034349$). This confirms that optimizing a multi-output loss over five noisy historical MLE parameters does not align with minimizing forward 21-session volatility forecast error.

2. **Why the 2022 Validation Winner (mle_total_vol) Lagged in 2023:** During the 2022 validation split (a sustained high-volatility bear market with mean realized volatility 24.7%), the 126-session trailing MLE total volatility was nearly unbiased ($bias = -0.0024$, $RMSE = 0.056528$), whereas models trained on the 2019–early 2022 fit split (direct_lstm, $bias = -0.0546$) underpredicted the elevated 2022 level. Conversely, when market volatility decayed sharply in early 2023 (mean realized volatility 12.8%), the 126-session trailing MLE window carried heavy late-2022 variance forward, resulting in a positive bias of $+0.036432$ ($+3.64$ annualized percentage points) and $RMSE = 0.050488$, even though its rank/linear correlation with future volatility remained high ($r = 0.6060$).

3. **Comparison with Literature:** Consistent with Honoré (1998) (**Honore1998?**) and Aït-Sahalia (2004) (**AitSahalia2004?**), daily equity returns over finite windows exhibit pervasive boundary hits ($165/211$) when fitting 5-parameter jump-diffusions. Consistent with Hernandez (2017) (**Hernandez2017?**), regime shifts between calibration/validation and test periods dominate surrogate generalization error.

10 Verified Scientific and Operational Limitations

In accordance with strict scientific transparency, we separate verified empirical limitations from operational scope boundaries:

1. **Prior Test-Set Exposure:** The supplied 2023 test CSV (spy_test_data.csv) was already inspected in the pre-repair project. Although our pipeline strictly excludes test data from scaling, training, and hyperparameter selection, 2023 represents a *fixed rerun holdout*, not a newly unseen holdout.

2. **Small Sample Size and Overlapping Horizons:** With $n_{fit} = 134$, $n_{val} = 31$, and $n_{test} = 46$ origins at stride $\Delta = 5$ and horizon $H = 21$, adjacent forecast errors share 16 common sessions. Consequently, metrics are reported as descriptive sample statistics without asymptotic IID p-values or confidence intervals.

3. **Noisy Latent Labels and Nonregular BIC:** Reference MLE parameters are finite-sample estimates, not ground-truth states. Canonical zeros assigned when BIC selects the Gaussian submodel ($\lambda = \mu_j = \sigma_j = 0$) reflect parsimony selection at a nonregular boundary, not proof that jumps cannot occur.

4. **Physical Measure Only (No Option Pricing or Trading Claims):** Estimation uses historical P-measure returns whose dividend/split adjustment provenance in the raw CSVs is unverified. No risk-neutral (Q) option-chain calibration or trading strategy profitability is claimed.

11 Conclusion and Future Scope

We have reconstructed, mathematically verified, and empirically audited a complete physical-measure Merton jump-diffusion calibration and realized volatility forecasting system on SPY. Our software resolves all numerical defects of the original prototype—replacing 1-jump Bernoulli clipping with an adaptive 15-term logsumexp Poisson mixture, analytical score gradients, nested Gaussian BIC selection, bijective/folding output links, comparable inference-mode loss tracking, and $H = 21$-purged chronological splits. While neural and Ridge surrogates deliver a $63.9\times$ online speedup over multistart MLE and the direct hybrid LSTM achieves the lowest 2023 test volatility RMSE (0.026553, $R^2_{oos} = +16.95\%$), regime shifts between the 2022 validation period and 2023 test period caused validation model selection to favor mle_total_vol, demonstrating that model-selection stability across volatility regimes remains the primary bottleneck.

Future engineering extensions suitable for follow-up research include: (1) evaluating on a genuinely fresh post-2024 out-of-sample holdout; (2) incorporating intraday 5-minute realized variance and realized bipower variation (**Andersen2003?**) to separate diffusive and jump variation nonparametrically; (3) replacing constant-parameter windows with state-dependent jump intensities or Bates (1996) (**Bates1996?**) stochastic volatility; and (4) computing Newey–West HAC standard errors and Model Confidence Sets across multi-asset panels.

Resources :

References:

1. Merton, R. C. (1976). Option pricing when underlying stock returns are discontinuous. *Journal of Financial Economics*, 3(1–2), 125–144. DOI: 10.1016/0304-405X(76)90022-2.
2. Black, F., & Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy*, 81(3), 637–654. DOI: 10.1086/260062.
3. Kiefer, N. M. (1978). Discrete parameter variation: Efficient estimation of a switching regression model. *Econometrica*, 46(2), 427–434. DOI: 10.2307/1913910.
4. Honoré, P. (1998). Pitfalls in estimating jump-diffusion models. *Working Paper Series No. 18*, Department of Finance, The Aarhus School of Business, pp. 1–30. URL: [Aarhus WP Full-Text PDF].
5. Aït-Sahalia, Y. (2004). Disentangling diffusion from jumps. *Journal of Financial Economics*, 74(3), 487–528. DOI: 10.1016/j.jfineco.2003.09.005.
6. Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461–464. DOI: 10.1214/aos/1176344136.
7. McLachlan, G., & Peel, D. (2000). *Finite Mixture Models*. Wiley Series in Probability and Statistics, John Wiley & Sons, New York. DOI: 10.1002/0471721182.
8. Bates, D. S. (1996). Jumps and stochastic volatility: Exchange rate processes implicit in Deutsche Mark options. *The Review of Financial Studies*, 9(1), 69–107. DOI: 10.1093/rfs/9.1.69.
9. Kou, S. G. (2002). A jump-diffusion model for option pricing. *Management Science*, 48(8), 1086–1101. DOI: 10.1287/mnsc.48.8.1086.166.
10. Andersen, T. G., Bollerslev, T., Diebold, F. X., & Labys, P. (2001). The distribution of realized exchange rate volatility. *Journal of the American Statistical Association*, 96(453), 42–55. DOI: 10.1198/016214501750332965.
11. Andersen, T. G., Bollerslev, T., Diebold, F. X., & Labys, P. (2003). Modeling and forecasting realized volatility. *Econometrica*, 71(2), 579–625. DOI: 10.1111/1468-0262.00418.

12. Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2), 174–196. DOI: 10.1093/jjfinec/nbp001.
13. Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55–67. DOI: 10.1080/00401706.1970.10488634.
14. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.
15. Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*. DOI: 10.48550/arXiv.1711.05101.
16. Hernandez, A. (2016/2017). Model calibration with neural networks. *Risk* (June 2017), 1–5; SSRN Working Paper 2812140. DOI: 10.2139/ssrn.2812140.
17. Liu, S., Borovykh, A., Grzelak, L. A., & Oosterlee, C. W. (2019). A neural network-based framework for financial model calibration. *Journal of Mathematics in Industry*, 9(1), Article 9. DOI: 10.1186/s13362-019-0066-7.
18. Horvath, B., Muguruza, A., & Tomas, M. (2021). Deep learning volatility: A deep neural network perspective on pricing and calibration in (rough) volatility models. *Quantitative Finance*, 21(1), 11–27. DOI: 10.1080/14697688.2020.1817974.
19. Campbell, J. Y., & Thompson, S. B. (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *The Review of Financial Studies*, 21(4), 1509–1531. DOI: 10.1093/rfs/hhm055.