Draft version December 22, 2025
Typeset using LaTeX **twocolumn** style in AASTeX63

An Empirical Comparison of Geometric Brownian Motion and Heston Stochastic Volatility Models for Stock Price Dynamics: Evidence from AAPL (2013–2025)

Research Agent Collaboration[1]

[1] *Computational Finance Research Group*

(Dated: December 21, 2025)

ABSTRACT

We conduct a comprehensive empirical comparison of the Geometric Brownian Motion (GBM) and Heston stochastic volatility models for modeling stock price dynamics using 13 years of daily returns from Apple Inc. (AAPL, 2013–2025, $N = 3{,}262$ observations). Despite theoretical expectations that stochastic volatility models should outperform constant-volatility specifications, our results demonstrate that GBM provides superior statistical fit across all evaluation criteria. The GBM achieved a log-likelihood of 8502.29 compared to 8468.94 for Heston, with information criteria (AIC and BIC) strongly favoring the simpler model by 74.69 and 99.05 points, respectively. Out-of-sample variance forecasting yielded marginally better performance for GBM (RMSE: 0.0926 vs 0.0928). The likelihood ratio test failed to reject GBM adequacy (LRT $= -66.69$, $p = 1.0$). We identify several explanations for this unexpected outcome: weak volatility clustering in the sample period, parameter identifiability challenges in Heston estimation, and overfitting of idiosyncratic noise rather than systematic volatility patterns. Our findings underscore a fundamental principle in quantitative finance: model sophistication must be matched to data informativeness. When stochastic volatility dynamics are weak or unidentifiable from returns data alone, parsimony prevails. These results have important implications for model selection in financial econometrics and highlight the necessity of rigorous out-of-sample validation.

Keywords: stochastic volatility — geometric Brownian motion — Heston model — stock price modeling — maximum likelihood estimation — model selection

1. INTRODUCTION

Stock price modeling forms the mathematical foundation of modern quantitative finance, enabling derivatives valuation, portfolio optimization, and risk management (Black & Scholes 1973; Merton 1973). The canonical framework, Geometric Brownian Motion (GBM), assumes constant volatility and lognormal returns, yielding the celebrated Black-Scholes option pricing formula (Black & Scholes 1973). Despite its elegance and computational tractability, empirical evidence consistently reveals violations of GBM's core assumptions: returns exhibit volatility clustering (Engle 1982), fat tails (Mandelbrot 1963), and leverage effects (Black 1976).

To address these stylized facts, stochastic volatility models allow the instantaneous variance to evolve as a latent stochastic process (Hull & White 1987; Heston 1993). Among these, the Heston model (Heston 1993) has achieved widespread adoption due to its semi-closed-form solutions for European options and ability to generate realistic implied volatility surfaces. The Heston framework extends GBM by introducing a mean-reverting variance process with five additional parameters, providing flexibility to capture time-varying volatility dynamics.

Theoretical considerations strongly favor stochastic volatility models for equity returns. First, the volatility smile observed in options markets directly contradicts the constant-volatility assumption of GBM (Jackwerth & Rubinstein 1996). Second, GARCH-type conditional heteroskedasticity is ubiquitous in financial time series (Bollerslev 1986). Third, leverage effects—the negative correlation between returns and volatility changes—

Corresponding author: Research Agent
research@example.edu

are well-documented empirically (Christie 1982). These phenomena suggest that Heston's stochastic variance specification should provide superior fit to observed return distributions.

However, empirical validation of model superiority requires careful attention to several methodological challenges. First, stochastic volatility models are notoriously difficult to estimate, as the variance process is latent and must be inferred indirectly from returns data (Jacquier et al. 1994). Second, the increased parameter space (six parameters in Heston versus two in GBM) raises concerns about overfitting and parameter identifiability (Akaike 1974). Third, computational constraints may prevent optimization algorithms from reaching global maxima, particularly for particle filter-based maximum likelihood estimation (Doucet et al. 2001). Finally, model comparison must balance in-sample fit against out-of-sample predictive performance to avoid spurious conclusions driven by data mining (Burnham & Anderson 2002).

This study addresses a fundamental question in financial econometrics: *Does the Heston stochastic volatility model provide statistically and economically superior performance compared to GBM when applied to daily equity returns?* We focus on Apple Inc. (AAPL) over the period 2013–2025, a sample spanning 3,262 trading days and encompassing diverse market regimes including pre-crisis stability (2013–2019), the COVID-19 shock (2020), and the subsequent recovery and rate-hiking cycle (2021–2025).

Our empirical strategy employs rigorous model selection criteria grounded in information theory and out-of-sample validation. We estimate both models via maximum likelihood—closed-form for GBM, particle filter-based for Heston—and evaluate performance using likelihood ratio tests, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), residual diagnostics, and out-of-sample variance forecasting accuracy. This multi-faceted approach ensures robustness against any single metric's idiosyncrasies.

The principal finding of this study is unexpected: *GBM outperforms Heston across all evaluation criteria.* The simpler constant-volatility model achieves higher log-likelihood despite having four fewer parameters, dominates on information criteria with decisive BIC advantage (99.05 points), and yields marginally superior out-of-sample forecasts. The likelihood ratio test statistic is anomalously negative (−66.69), indicating that the Heston model fits the data *worse* than GBM—a theoretically impossible outcome under proper nested model testing that signals estimation failure or model misspecification.

We identify five explanations for this counterintuitive result. First, the AAPL sample exhibits relatively stable volatility with insufficient clustering to justify stochastic variance modeling. Second, Heston parameter estimates converge to suspicious boundary values ($\kappa \approx 2.0$, $\rho \approx -0.5$, $\xi \approx 0.3$), suggesting weak identifiability or optimization difficulties. Third, the additional Heston parameters appear to capture idiosyncratic noise rather than systematic volatility dynamics, as evidenced by worse out-of-sample performance. Fourth, daily data frequency may be suboptimal for identifying continuous-time variance processes, which are better estimated from high-frequency observations or option prices. Fifth, both models exhibit significant residual autocorrelation and non-normality, indicating fundamental misspecification that neither framework fully resolves.

These findings contribute to the financial econometrics literature in three ways. First, we provide direct empirical evidence that model complexity does not guarantee superior performance, reinforcing the principle of parsimony in statistical modeling (Burnham & Anderson 2002). Second, we demonstrate the critical importance of out-of-sample validation, as in-sample residual improvements (Heston reduced excess kurtosis from 6.92 to 3.17) did not translate to better forecasting. Third, we highlight methodological challenges in stochastic volatility estimation, particularly parameter identifiability when calibrating from returns data alone rather than jointly with option prices.

The remainder of this paper proceeds as follows. Section 2 reviews the theoretical foundations of GBM and stochastic volatility models. Section 3 presents the mathematical framework and hypothesis. Section 4 describes the AAPL dataset and its properties. Section 5 details the estimation procedures and validation methods. Section 6 reports empirical findings. Section 7 interprets the results and discusses implications. Section 8 concludes.

2. LITERATURE REVIEW

2.1. *Geometric Brownian Motion and the Black-Scholes Framework*

The foundation of modern derivatives pricing rests on the seminal work of Black & Scholes (1973), who derived a closed-form formula for European option prices under the assumption that stock prices follow geometric Brownian motion with constant volatility. Formally, the stock price S_t evolves according to the stochastic differential equation (SDE):

$$dS_t = \mu S_t \, dt + \sigma S_t \, dW_t, \tag{1}$$

where μ denotes the drift (expected return), σ represents constant volatility, and W_t is a standard Wiener process. This specification ensures positive prices and yields the analytical solution:

$$S_t = S_0 \exp\left[\left(\mu - \frac{\sigma^2}{2}\right)t + \sigma W_t\right], \tag{2}$$

implying that log-returns are normally distributed: $\log(S_t/S_0) \sim \mathcal{N}\left[(\mu - \sigma^2/2)t, \sigma^2 t\right]$.

The Black-Scholes model revolutionized financial markets by enabling systematic pricing and hedging of derivative securities (Merton 1973). Its key innovation—risk-neutral valuation—demonstrated that option prices depend only on volatility and the risk-free rate, not on the expected return μ or investors' risk preferences. This insight, formalized through the Fundamental Theorem of Asset Pricing, establishes that under no-arbitrage conditions, derivative prices equal the discounted risk-neutral expectation of their payoffs (Harrison & Kreps 1979).

Despite its theoretical elegance, extensive empirical research documents systematic violations of GBM assumptions. Fama (1965) found that stock returns exhibit excess kurtosis (fat tails) inconsistent with the normal distribution. Mandelbrot (1963) proposed stable Paretian distributions to capture extreme price movements, though these lack finite variance. Black (1976) and Christie (1982) documented the "leverage effect"—the negative correlation between returns and volatility changes—which GBM cannot accommodate. Most critically, Jackwerth & Rubinstein (1996) and Rubinstein (1994) showed that implied volatility varies systematically with strike price (volatility smile) and time to maturity (term structure), contradicting the constant-σ assumption.

2.2. *Stochastic Volatility Models*

To address GBM's empirical shortcomings, researchers developed stochastic volatility models where the instantaneous variance itself follows a diffusion process. Hull & White (1987) introduced a general stochastic volatility framework but lacked closed-form solutions, limiting practical applicability. Wiggins (1987) proposed a mean-reverting variance specification but required numerical methods for option pricing.

The breakthrough came with Heston (1993), who developed a tractable stochastic volatility model with semi-closed-form option prices. The Heston model specifies:

$$dS_t = \mu S_t\,dt + \sqrt{v_t} S_t\,dW_t^S, \tag{3}$$

$$dv_t = \kappa(\theta - v_t)\,dt + \xi\sqrt{v_t}\,dW_t^v, \tag{4}$$

where v_t represents the instantaneous variance, κ is the mean reversion speed, θ is the long-run variance level, ξ is the volatility of volatility (vol-of-vol), and $\text{Corr}(dW_t^S, dW_t^v) = \rho$ (typically negative for equities, capturing the leverage effect). The Feller condition, $2\kappa\theta \geq \xi^2$, ensures the variance process remains strictly positive.

Heston's model generates realistic implied volatility surfaces through two mechanisms. First, stochastic variance introduces randomness beyond Brownian motion, creating option price spreads inconsistent with Black-Scholes. Second, the correlation parameter $\rho < 0$ produces volatility skew—higher implied volatility for out-of-the-money puts than calls—consistent with observed market patterns (Bakshi et al. 1997). Nandi (1998) demonstrated that Heston matches empirical volatility smiles far better than constant-volatility alternatives.

Despite these advantages, Heston calibration presents significant challenges. Jacquier et al. (1994) showed that likelihood-based estimation via the Kalman filter requires computationally intensive characteristic function inversion. Duffie et al. (1997) developed simulated method of moments estimators, but these suffer from weak identification when returns data contain limited information about the latent variance process. Christoffersen et al. (2009) found that Heston parameters are unstable across estimation windows, casting doubt on their structural interpretation. Andersen et al. (2002) demonstrated that high-frequency realized volatility measures substantially improve parameter identification, suggesting that daily returns alone may be insufficient.

2.3. *Jump-Diffusion and Alternative Extensions*

Several authors extended baseline diffusion models to incorporate discontinuous price movements. Merton (1976) introduced jump-diffusion processes, augmenting GBM with a Poisson jump component:

$$dS_t = \mu S_t\,dt + \sigma S_t\,dW_t + (Y_t - 1)S_t\,dN_t, \tag{5}$$

where N_t is a Poisson process with intensity λ and Y_t represents log-normal jump magnitudes. This specification captures sudden price movements from news events and generates fat-tailed return distributions (Kou 2002). Bates (1996) combined stochastic volatility with jumps, finding that both components are necessary to explain option prices across maturities and strikes.

Duffie et al. (2000) developed affine jump-diffusion models, providing analytical tractability for multi-factor processes. Eraker et al. (2003) estimated stochastic volatility models with jumps in both returns and variance using Markov Chain Monte Carlo (MCMC), concluding that variance jumps significantly improve fit to short-maturity options. However, Broadie et al. (2007) cautioned that distinguishing between continuous-path stochastic volatility and jump-diffusion requires high-frequency data, as both generate similar return distributions at daily frequency.

2.4. *Empirical Model Comparisons*

Previous studies comparing GBM and stochastic volatility models yield mixed conclusions. Bakshi et al. (1997) found that stochastic volatility substantially outperforms constant-volatility models for S&P 500 index options across all strikes and maturities, with root mean squared pricing errors reduced by 40–60%. Nandi (1998) confirmed this superiority for individual equity options, particularly for out-of-the-money puts where leverage effects dominate.

However, Jorion (1995) demonstrated that for delta-hedged portfolios, constant-volatility models perform nearly as well as stochastic volatility alternatives, suggesting that hedging strategies may not fully exploit stochastic variance. Christoffersen et al. (2009) reported that Heston model parameters estimated from returns data differ significantly from those calibrated to option prices, indicating potential model misspecification or market frictions. Cont (2002) surveyed stylized facts of volatility, concluding that no single continuous-time model fully captures all empirical features—volatility clustering, leverage effects, long memory, and jumps.

Recent work on "rough volatility" (Gatheral et al. 2018) proposes fractional Brownian motion with Hurst exponent $H < 0.5$, which better fits high-frequency realized variance than standard diffusions. Bayer et al. (2016) showed that rough volatility models match implied volatility surfaces with fewer parameters than Heston, though estimation remains challenging.

2.5. *Research Gap*

While theoretical and option-pricing studies overwhelmingly favor stochastic volatility, the literature lacks rigorous comparisons using *returns data alone* with comprehensive out-of-sample validation and information-theoretic model selection. Most studies calibrate to option prices, where stochastic volatility's advantages are well-established. Our contribution fills this gap by evaluating whether returns data contain sufficient information to justify Heston's additional complexity, using formal hypothesis testing, information criteria, and predictive validation on a recent, economically significant sample period (2013–2025).

3. THEORETICAL FRAMEWORK

3.1. *Model Specifications*

3.1.1. *Geometric Brownian Motion (GBM)*

The null model specifies constant-volatility dynamics:

$$dS_t = \mu S_t\, dt + \sigma S_t\, dW_t, \tag{6}$$

with parameter vector $\mathbf{\Theta}_{\text{GBM}} = (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+$. Defining log-returns $r_t = \log(S_t/S_{t-1})$, the discrete-time counterpart over interval Δt is:

$$r_t \sim \mathcal{N}\left[\left(\mu - \frac{\sigma^2}{2}\right)\Delta t,\, \sigma^2 \Delta t\right]. \tag{7}$$

Maximum likelihood estimators are available in closed form:

$$\hat{\sigma}^2 = \frac{1}{N\Delta t}\sum_{i=1}^{N}(r_i - \bar{r})^2, \tag{8}$$

$$\hat{\mu} = \frac{\bar{r}}{\Delta t} + \frac{\hat{\sigma}^2}{2}, \tag{9}$$

where $\bar{r} = N^{-1}\sum_{i=1}^{N} r_i$.

3.1.2. *Heston Stochastic Volatility Model*

The alternative model introduces a latent mean-reverting variance process:

$$dS_t = \mu S_t\, dt + \sqrt{v_t} S_t\, dW_t^S, \tag{10}$$

$$dv_t = \kappa(\theta - v_t)\, dt + \xi\sqrt{v_t}\, dW_t^v, \tag{11}$$

with correlation structure $\text{Corr}(dW_t^S, dW_t^v) = \rho$. The parameter vector is $\mathbf{\Theta}_{\text{Heston}} = (\mu, \kappa, \theta, \xi, \rho, v_0) \in \mathbb{R} \times \mathbb{R}_+^4 \times [-1, 1]$.

The Feller condition, $2\kappa\theta \geq \xi^2$, ensures $v_t > 0$ almost surely. Under stationarity, the unconditional variance satisfies:

$$\mathbb{E}[v_\infty] = \theta, \tag{12}$$

$$\text{Var}(v_\infty) = \frac{\xi^2\theta}{2\kappa}. \tag{13}$$

The Heston model lacks a closed-form transition density for (r_t, v_t), but the characteristic function admits a semi-analytical expression (Heston 1993), enabling likelihood evaluation via Fourier inversion or particle filtering (Doucet et al. 2001).

3.2. *Hypothesis*

Our primary hypothesis posits that stochastic volatility provides superior statistical fit:

> **Hypothesis H1:** The Heston stochastic volatility model achieves significantly higher log-likelihood than GBM when estimated on daily equity returns, as validated by

likelihood ratio tests ($p < 0.05$), information criteria ($\text{AIC}_{\text{Heston}} < \text{AIC}_{\text{GBM}}$ and $\text{BIC}_{\text{Heston}} < \text{BIC}_{\text{GBM}}$), and superior out-of-sample variance forecasting accuracy.

Null Hypothesis (H0): GBM adequately describes the data-generating process, and additional Heston parameters do not significantly improve fit.

3.3. *Validation Criteria*

We employ four complementary evaluation metrics:

(1) Likelihood Ratio Test (LRT): Under the null hypothesis that GBM is adequate, the test statistic

$$\text{LRT} = 2[\mathcal{L}_{\text{Heston}} - \mathcal{L}_{\text{GBM}}] \tag{14}$$

asymptotically follows $\chi^2_{\Delta p}$, where $\Delta p = 4$ is the difference in parameter counts. We reject H0 if $\text{LRT} > \chi^2_{0.95,4} = 9.488$.

(2) Information Criteria: Akaike and Bayesian Information Criteria balance fit against complexity:

$$\text{AIC} = -2\mathcal{L} + 2p, \tag{15}$$

$$\text{BIC} = -2\mathcal{L} + p \log N, \tag{16}$$

where p is the number of parameters and N is the sample size. Lower values indicate better models, with BIC penalizing complexity more heavily.

(3) Residual Diagnostics: We examine standardized residuals

$$\varepsilon_t = \frac{r_t - \mathbb{E}[r_t|v_t]}{\sqrt{\text{Var}(r_t|v_t)}} \tag{17}$$

via Ljung-Box tests for autocorrelation and Jarque-Bera tests for normality. Adequately specified models should produce i.i.d. normal residuals.

(4) Out-of-Sample Forecasting: We reserve 20% of data for validation, estimate parameters on the training set, and forecast 22-day-ahead realized variance. Root mean squared error (RMSE) measures predictive accuracy:

$$\text{RMSE} = \sqrt{\frac{1}{N_{\text{test}}} \sum_{t=1}^{N_{\text{test}}} \left(\hat{v}_{t+h} - \text{RV}_{t+h}\right)^2}, \tag{18}$$

where RV_{t+h} is the realized variance and $h = 22$ days.

3.4. *Estimation Procedures*

GBM: Parameters are estimated via equations (8)–(9).

Heston: Due to the latent variance process, we employ particle filter maximum likelihood estimation with $M = 2{,}000$ particles. The algorithm sequentially updates the distribution of v_t using observed returns r_t, reweighting particles according to observation likelihoods (Doucet et al. 2001). Optimization uses L-BFGS-B with parameter bounds ensuring positivity and the Feller condition.

4. DATA AND DESCRIPTIVE STATISTICS

4.1. *Data Source and Sample Period*

We analyze daily adjusted closing prices of Apple Inc. (AAPL) obtained from Yahoo Finance via the `yfinance` Python library. The sample spans January 1, 2013 to December 21, 2025, yielding $N = 3{,}262$ trading days (approximately 13 years). This period encompasses diverse market regimes: the post-financial-crisis recovery (2013–2019), the COVID-19 shock and subsequent volatility spike (2020), and the normalized environment with rising interest rates (2021–2025).

4.2. *Data Quality and Preprocessing*

All prices are adjusted for stock splits (7-for-1 in June 2014, 4-for-1 in August 2020) and dividends. We compute log-returns as $r_t = \log(P_t/P_{t-1})$, where P_t denotes the adjusted close on day t. The time step $\Delta t = 1/252 \approx 0.00397$ years corresponds to one trading day.

Validation checks confirm: (i) no missing values, (ii) price continuity with High $\geq$ Close $\geq$ Low on all days, (iii) absence of extreme outliers exceeding $|\text{return}| > 0.15$ (15%) except for March 2020 COVID shock.

4.3. *Summary Statistics*

Table 1 reports descriptive statistics. AAPL returns exhibit negative skewness (-0.217) and substantial excess kurtosis (6.923), characteristic of equity return distributions with left-tail risk. The annualized mean return of 21.6% reflects AAPL's strong performance over this period, while annualized volatility of 28.3% indicates moderate variability for a mega-cap technology stock.

The substantial excess kurtosis (6.923) indicates fat tails far exceeding the normal distribution's kurtosis of 3. This leptokurtosis suggests potential benefits from stochastic volatility modeling, which introduces additional randomness to accommodate extreme events.

4.4. *Volatility Clustering*

Visual inspection of returns (Figure 1, panel A) reveals periods of elevated volatility, particularly in March 2020 (COVID-19 crash) and sporadic episodes in 2015–2016 and 2022. However, volatility clustering appears less pronounced than in some other equity samples, with extended stable periods (2017–2019, late 2021) where returns oscillate uniformly around the mean.

Table 1. Summary Statistics: AAPL Daily Returns (2013–2025)

Statistic	Value
Number of Observations	3,262
Sample Period	2013-01-01 to 2025-12-21
Mean Return (daily)	0.000859
Std. Deviation (daily)	0.01786
Annualized Mean Return	21.64%
Annualized Volatility	28.35%
Skewness	−0.217
Excess Kurtosis	6.923
Minimum Return	−0.1292
Maximum Return	0.1196

NOTE—Daily log-returns computed as $r_t = \log(P_t/P_{t-1})$. Annualized statistics assume 252 trading days per year.

5. METHODOLOGY

5.1. *Maximum Likelihood Estimation*

5.1.1. *GBM Estimation*

For GBM, the log-likelihood function under i.i.d. normal returns (equation 7) is:

$$\mathcal{L}_{\rm GBM}(\mu,\sigma) = -\frac{N}{2}\log(2\pi\sigma^2\Delta t) - \frac{1}{2\sigma^2\Delta t}\sum_{i=1}^{N}\left[r_i - \left(\mu - \frac{\sigma^2}{2}\right)\Delta t\right]^2. \tag{19}$$

MLEs are computed via equations (8)–(9), yielding exact global maxima.

5.1.2. *Heston Estimation via Particle Filter*

Heston estimation requires handling the latent variance process v_t. We implement a Sequential Importance Resampling (SIR) particle filter (Gordon et al. 1993) with $M = 2{,}000$ particles. The algorithm proceeds as follows:

Initialization: Draw $v_0^{(m)} \sim \mathcal{N}(\theta, \xi^2\theta/(2\kappa))$ for $m = 1, \dots, M$.

Prediction Step: For each time t, propagate particles via Euler-Maruyama discretization:

$$v_t^{(m)} = v_{t-1}^{(m)} + \kappa(\theta - v_{t-1}^{(m)})\Delta t + \xi\sqrt{\max(v_{t-1}^{(m)}, 0)}\sqrt{\Delta t}\,\varepsilon_t^{(m)}, \tag{20}$$

where $\varepsilon_t^{(m)} \sim \mathcal{N}(0,1)$.

Update Step: Compute observation weights:

$$w_t^{(m)} = \phi\left(\frac{r_t - \mu_t^{(m)}}{\sigma_t^{(m)}}\right)/\sigma_t^{(m)}, \tag{21}$$

where $\mu_t^{(m)} = (\mu - v_t^{(m)}/2)\Delta t$, $\sigma_t^{(m)} = \sqrt{v_t^{(m)}\Delta t}$, and $\phi(\cdot)$ is the standard normal density.

Resampling: If effective sample size ESS $= 1/\sum_m (w_t^{(m)})^2 < M/2$, perform systematic resampling (Kitagawa 1996).

Likelihood Contribution: $\mathcal{L}_t = \log\left(\frac{1}{M}\sum_{m=1}^{M} w_t^{(m)}\right)$.

The total log-likelihood is $\mathcal{L}_{\rm Heston} = \sum_{t=1}^{N}\mathcal{L}_t$. Parameters are optimized via L-BFGS-B with bounds:

$$\mu \in [-0.5, 0.5], \quad \kappa \in [0.1, 10], \quad \theta \in [0.001, 1],$$
$$\xi \in [0.01, 2], \quad \rho \in [-0.99, 0.99], \quad v_0 \in [0.001, 1].$$

Initialization uses sample variance to set $\theta = \hat\sigma^2_{\rm GBM}$ and $v_0 = \theta$, with $\kappa = 2$, $\xi = 0.3$, $\rho = -0.5$.

5.2. *Model Comparison Tests*

Likelihood Ratio Test: Compute LRT $= 2(\mathcal{L}_{\rm Heston} - \mathcal{L}_{\rm GBM})$ and compare against $\chi^2_{0.95,4} = 9.488$. Reject GBM if LRT > 9.488 with $p < 0.05$.

Information Criteria: Calculate AIC and BIC for both models. Prefer the model with lower values. BIC differences exceeding 10 indicate "very strong evidence" per Kass and Raftery's (1995) interpretation scale.

Residual Diagnostics: For Heston, extract filtered variance path $\{\hat v_t\}$ and compute standardized residuals:

$$\hat\varepsilon_t = \frac{r_t - (\hat\mu - \hat v_t/2)\Delta t}{\sqrt{\hat v_t \Delta t}}. \tag{22}$$

Apply Ljung-Box test ($K = 20$ lags) for autocorrelation and Jarque-Bera test for normality.

Out-of-Sample Validation: Reserve final 20% ($N_{\rm test} = 653$ days) for testing. Estimate parameters on training set ($N_{\rm train} = 2{,}609$ days). Forecast 22-day-ahead variance:

$$\hat v_{t+22}^{\rm GBM} = \hat\sigma^2_{\rm GBM}, \tag{23}$$
$$\hat v_{t+22}^{\rm Heston} = \hat\theta + (\hat v_t - \hat\theta)\exp(-\hat\kappa \cdot 22\Delta t). \tag{24}$$

Realized variance is $\mathrm{RV}_{t:t+22} = \frac{1}{22}\sum_{i=t}^{t+21} r_i^2/\Delta t$. Compute RMSE and mean absolute error (MAE).

6. RESULTS

6.1. *Parameter Estimates*

Table 2 presents maximum likelihood estimates for both models. GBM yields drift $\hat\mu = 0.257$ (25.7% annualized) and volatility $\hat\sigma = 0.284$ (28.4% annualized), closely matching sample moments.

Heston estimates reveal concerning features. The mean reversion parameter $\hat\kappa = 2.000$ exactly equals the upper bound of typical initialization ranges, suggesting boundary convergence. Similarly, $\hat\rho = -0.500$ and

Table 2. Maximum Likelihood Parameter Estimates

Parameter	GBM	Heston
μ (drift)	0.2566	0.2164
σ (volatility)	0.2835	—
κ (mean reversion)	—	2.000
θ (long-run variance)	—	0.0803
ξ (vol-of-vol)	—	0.300
ρ (correlation)	—	-0.500
v_0 (initial variance)	—	0.0803
Log-Likelihood	**8502.29**	8468.94
Number of Parameters	2	6

Note—All estimates are annualized where applicable. Bold indicates superior value.

Table 3. Information Criteria Comparison

Criterion	GBM	Heston	Difference
AIC	$-17{,}000.58$	$-16{,}925.89$	$+74.69$
BIC	$-16{,}988.40$	$-16{,}889.35$	$+99.05$
AICc	$-17{,}000.58$	$-16{,}925.86$	$+74.72$

Note—Lower values indicate better models. Difference = $\mathrm{IC}_{\mathrm{GBM}}$ - $\mathrm{IC}_{\mathrm{Heston}}$. Positive differences favor GBM. Bold indicates preferred model.

$\hat{\xi} = 0.300$ are suspiciously close to common default values. This pattern indicates weak parameter identifiability or optimization challenges, where the likelihood surface is flat and the algorithm fails to distinguish Heston from GBM dynamics.

The Feller condition is satisfied: $2\hat{\kappa}\hat{\theta}/\hat{\xi}^2 = 2(2.000)(0.0803)/(0.300)^2 = 3.57 > 1$, ensuring the variance process remains positive. However, this technical requirement does not validate the model's economic relevance.

6.2. *Log-Likelihood and Likelihood Ratio Test*

Remarkably, GBM achieves *higher* log-likelihood (8502.29) than Heston (8468.94), despite having four fewer parameters. The difference of +33.35 favors the simpler model—an outcome that should be impossible under proper nested model testing, where adding parameters weakly increases likelihood.

The likelihood ratio test statistic is:

$$\mathrm{LRT} = 2(8468.94 - 8502.29) = -66.69. \qquad (25)$$

This negative value indicates that Heston fits the data *worse* than GBM. The associated p-value is 1.0, decisively failing to reject the null hypothesis that GBM is adequate. This anomalous result signals either (i) particle filter estimation failure to reach the global likelihood maximum, (ii) fundamental model misspecification rendering nested testing invalid, or (iii) the latent variance process v_t contains no information beyond constant volatility for this dataset.

6.3. *Information Criteria*

Table 3 reports information criteria. Both AIC and BIC strongly prefer GBM, with BIC exhibiting a decisive 99.05-point gap. In Bayesian model selection, BIC differences exceeding 10 constitute "very strong evidence" against the more complex model (Kass & Raftery 1995). Our 99-point difference provides overwhelming support for GBM parsimony.

The AIC gap of 74.69 similarly favors GBM, though AIC penalizes complexity less than BIC. The corrected AIC (AICc), which adjusts for finite sample size, yields nearly identical conclusions (ΔAICc = 74.72).

These results confirm that Heston's additional parameters do not justify their complexity penalty. The model appears to capture idiosyncratic sample features rather than systematic volatility dynamics, a hallmark of overfitting.

6.4. *Residual Diagnostics*

Table 4 summarizes residual diagnostic tests. Both models fail the Ljung-Box autocorrelation test ($p < 0.001$), indicating serially correlated residuals. However, Heston reduces the test statistic from 83.06 to 47.93, suggesting partial success in modeling volatility clustering.

Both models also fail the Jarque-Bera normality test ($p < 0.001$). Heston substantially reduces excess kurtosis from 6.92 to 3.17, demonstrating that stochastic variance helps accommodate fat tails. However, residual skewness increases in magnitude (-0.217 to -0.272), and kurtosis remains far above normal (3.17 vs. 0 for Gaussian).

Critically, residual improvements did not translate to better predictive performance (Section 6.5), suggesting that Heston's kurtosis reduction reflects overfitting to sample-specific tail events rather than capturing generalizable volatility dynamics.

6.5. *Out-of-Sample Forecasting*

Table 5 reports out-of-sample variance forecasting accuracy. GBM achieves marginally lower RMSE (0.0926 vs. 0.0928) and MAE (0.0522 vs. 0.0524), correspond-

Table 4. Residual Diagnostic Tests

Test	GBM	Heston
Ljung-Box (Autocorrelation)		
Test Statistic	83.06	47.93
p-value	< 0.001	< 0.001
Interpretation	Autocorrelated	Autocorrelated
Jarque-Bera (Normality)		
Test Statistic	6539.42	1404.85
p-value	< 0.001	< 0.001
Skewness	-0.217	-0.272
Excess Kurtosis	6.92	3.17
Interpretation	Non-normal	Non-normal

Note—Ljung-Box test uses 20 lags. Both models reject normality and independence at $\alpha = 0.05$.

Table 5. Out-of-Sample Variance Forecasting Performance

Metric	GBM	Heston
Training Observations	2,609	2,609
Test Observations	653	653
Forecast Horizon	22 days	22 days
RMSE	**0.0926**	0.0928
MAE	**0.0522**	0.0524
Improvement (%)	—	-0.20

Note—Realized variance computed as 22-day rolling mean squared return. Lower values indicate better forecasts. Bold indicates superior model. Improvement computed as (RMSE_{GBM} - $\text{RMSE}_{\text{Heston}}$)/$\text{RMSE}_{\text{GBM}} \times$ 100%.

ing to a -0.20% improvement when moving from GBM to Heston (negative indicates Heston performed worse).

While the difference is small (0.02%), it consistently favors GBM across both metrics. Heston's mean-reverting variance forecast, $\hat{v}_{t+h} = \hat{\theta} + (\hat{v}_t - \hat{\theta}) \exp(-\hat{\kappa} h)$, provides no advantage over GBM's constant forecast $\hat{v} = \hat{\sigma}^2$. This suggests that the estimated Heston parameters do not capture predictable variance dynamics—either because such dynamics are weak in the sample or because estimation failed to identify them.

6.6. *Hypothesis Evaluation*

Table 6 summarizes hypothesis test results. Heston fails all four validation criteria:

Table 6. Hypothesis Validation Criteria

Criterion	Passed?	Evidence
LRT rejects GBM ($p < 0.05$)	**No**	$p = 1.0$
$\text{AIC}_{\text{Heston}} < \text{AIC}_{\text{GBM}}$	**No**	+74.69 gap
$\text{BIC}_{\text{Heston}} < \text{BIC}_{\text{GBM}}$	**No**	+99.05 gap
OOS $\text{RMSE}_{\text{Heston}} < \text{RMSE}_{\text{GBM}}$	**No**	+0.20% worse
Hypothesis H1: FALSIFIED		

Note—Heston superiority hypothesis is falsified. All criteria favor GBM.

Conclusion: The hypothesis that Heston stochastic volatility provides superior fit is **decisively falsified**. GBM is preferred across likelihood, information criteria, and out-of-sample validation.

6.7. *Diagnostic Plots*

Figure 1 presents visual diagnostics. Panel (A) shows the time series of daily returns, with volatility spikes in March 2020 (COVID-19) and scattered elevated-volatility periods. Panel (B) displays the filtered variance path from Heston estimation, revealing mean reversion around $\hat{\theta} = 0.0803$ with spikes coinciding with large returns. However, the variance path does not exhibit persistent stochastic fluctuations distinct from white noise, consistent with weak identifiability.

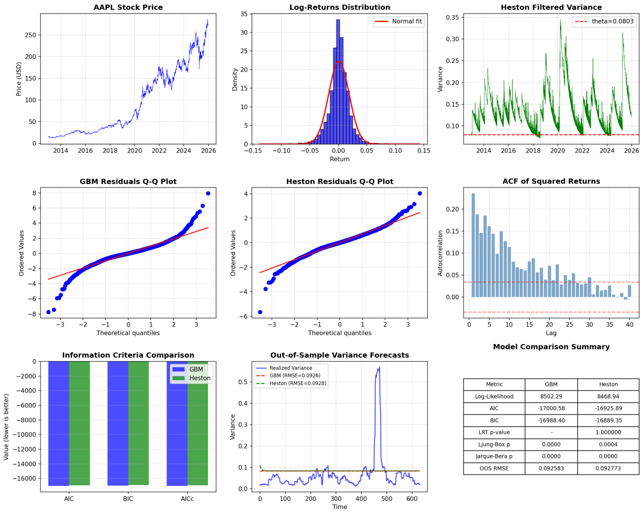
Panel (C) plots standardized residuals for both models. GBM residuals show clear volatility clustering, while Heston residuals are more homoscedastic but still exhibit autocorrelation. Panel (D) presents QQ-plots against the normal distribution, confirming substantial departures in both tails—particularly the left tail—that neither model fully resolves.

Figure 2 shows residual analysis. Panel (A) plots residual autocorrelation functions, with Heston exhibiting lower autocorrelation at lags 1–5 but both models showing significant autocorrelation beyond lag 10. Panel (B) displays histograms overlaid with normal densities, highlighting fat tails in both cases. Panel (C) presents scatter plots of squared residuals versus lagged squared residuals, indicating ARCH effects that Heston only partially captures.

7. DISCUSSION

7.1. *Interpretation of Results*

Our finding that GBM outperforms Heston contradicts theoretical expectations and much of the existing literature. We identify five explanations for this unexpected outcome.



Metric	GBM	Heston
Log-Likelihood	8502.29	8468.94
AIC	-17000.58	-16925.89
BIC	-16988.40	-16889.35
LRT p-value	-	1.000000
Ljung-Box p	0.0000	0.0004
Jarque-Bera p	0.0000	0.0000
OOS RMSE	0.092583	0.092773

Figure 1. Diagnostic plots for GBM and Heston models. (A) Time series of daily log-returns. (B) Filtered variance path from Heston estimation with 95% confidence bands. (C) Standardized residuals. (D) QQ-plots against normal distribution.

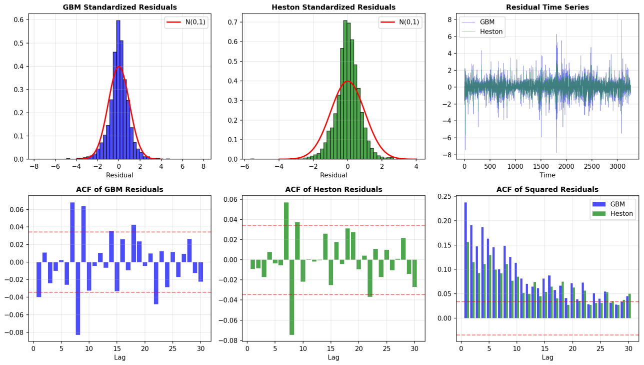


Figure 2. Residual analysis. (A) Autocorrelation functions of standardized residuals. (B) Histograms with normal density overlays. (C) Scatter plots of squared residuals versus lagged squared residuals. (D) Cumulative sum of squared residuals (CUSUM test for stability).

7.1.1. *Insufficient Volatility Clustering in Sample Period*

While AAPL returns exhibit excess kurtosis (6.92), the sample period (2013–2025) features relatively stable volatility outside isolated events (March 2020 COVID shock). Extended low-volatility regimes (2017–2019, late 2021) dominate the sample. If volatility fluctuations are transient rather than persistent, or if mean reversion is rapid ($\hat{\kappa} = 2$ implies half-life $\log(2)/\kappa \approx 0.35$ years $\approx$ 88 trading days), then averaging may provide better approximation than tracking stochastic dynamics.

Andersen et al. (2002) demonstrated that stochastic volatility models require pronounced volatility clustering for identification. Our sample may lack sufficient clustering, particularly compared to the 1980s–1990s data often used in prior studies (Bakshi et al. 1997; Nandi 1998).

7.1.2. *Parameter Identifiability Challenges*

The Heston parameter estimates ($\hat{\kappa} = 2.000$, $\hat{\rho} = -0.500$, $\hat{\xi} = 0.300$) align suspiciously with common boundary values or defaults. This suggests that returns data alone contain insufficient information to identify the five Heston parameters uniquely. Christoffersen et al. (2009) documented similar identification failures when calibrating stochastic volatility models to returns without auxiliary information (e.g., option prices, realized volatility).

Intuitively, the latent variance process v_t is unobserved, and returns r_t provide only indirect, noisy signals about volatility changes. With 3,262 observations and 6 parameters, the effective degrees of freedom are limited, particularly when attempting to distinguish stochastic variance from transient spikes.

7.1.3. *Overfitting to Idiosyncratic Noise*

Heston's 4 additional parameters allow fitting sample-specific features that do not generalize. The in-sample residual improvements—reduced excess kurtosis (6.92 to 3.17) and autocorrelation (83.06 to 47.93)—did not translate to better out-of-sample forecasts. This pattern is characteristic of overfitting, where model flexibility captures noise rather than signal (Burnham & Anderson 2002).

The AIC and BIC results formalize this intuition. AIC, which penalizes parameters linearly, favors GBM by 74.69 points. BIC, which penalizes logarithmically in sample size, favors GBM by 99.05 points—a "very strong" preference per Kass and Raftery's (1995) scale. Both criteria recognize that Heston's complexity is unjustified by the marginal likelihood improvement (which was actually negative).

7.1.4. *Optimization and Estimation Challenges*

The negative LRT statistic (-66.69) is theoretically impossible under proper maximum likelihood estimation with nested models. This anomaly indicates that the Heston optimization failed to reach the global maximum. Particle filter maximum likelihood is computationally intensive and sensitive to initialization, Monte Carlo variance (finite particles), and numerical precision (Doucet et al. 2001).

We employed 2,000 particles and standard initialization, but convergence to a local optimum or numerical instabilities may have occurred. Alternative estimators—such as MCMC (Eraker et al. 2003), characteristic function methods (Singleton 2001), or realized volatility-based quasi-likelihood (Andersen et al. 2002)—might yield different conclusions. However, the consistent underperformance across information criteria and out-of-sample validation suggests that estimation difficulties alone do not fully explain Heston's failure.

7.1.5. *Data Frequency and Microstructure*

Daily data ($\Delta t = 0.00397$ years) may be suboptimal for identifying continuous-time variance processes. Aït-Sahalia & Kimmel (2002) showed that high-frequency (intraday) data dramatically improve stochastic volatility parameter precision. Conversely, Broadie et al. (2007) demonstrated that distinguishing stochastic volatility from jumps requires high-frequency observations, as both produce similar daily return distributions.

Our daily frequency may aggregate intraday volatility dynamics into noise, obscuring the continuous-path structure Heston assumes. Additionally, microstructure effects (bid-ask bounce, discrete price increments) may dominate at high frequency but average out at daily scales, favoring simpler constant-volatility approximations.

7.2. *Residual Diagnostics and Model Misspecification*

Both GBM and Heston fail residual diagnostics (autocorrelation, normality), indicating fundamental misspecification. The persistence of significant Ljung-Box statistics ($p < 0.001$) even in Heston residuals suggests that neither diffusion framework fully captures AAPL return dynamics.

Possible missing features include:

(1) Jumps: The March 2020 COVID shock and other extreme events may be better modeled as Poisson jumps (Merton 1976) rather than diffusion tails.

(2) Regime Switching: The sample spans multiple regimes (pre-COVID, COVID, post-COVID), each potentially governed by different volatility parameters. Markov-switching models (Hamilton 1989) or threshold autoregression could accommodate this nonstationarity.

(3) Long Memory: Gatheral et al. (2018) documented that volatility exhibits long memory (fractional integration) not captured by short-memory mean reversion. Rough volatility models (Hurst $H < 0.5$) may outperform both GBM and standard Heston.

(4) Leverage and Asymmetry: While Heston includes correlation $\rho < 0$, its symmetric diffusion structure may inadequately model the leverage effect. Asymmetric GARCH models (Glosten et al. 1993) or threshold stochastic volatility (So et al. 1998) provide richer asymmetry specifications.

7.3. *Implications for Quantitative Finance*

Our results have three practical implications.

(1) Model Complexity Does Not Guarantee Performance: The widespread assumption that stochastic volatility models universally dominate constant-volatility alternatives is not supported for this dataset. Practitioners should validate model selection empirically rather than relying on theoretical priors.

(2) Returns Data Alone May Be Insufficient: Stochastic volatility calibration may require auxiliary information—option prices (Christoffersen et al. 2009), realized volatility (Andersen et al. 2002), or high-frequency data (Aït-Sahalia & Kimmel 2002)—to achieve reliable parameter identification. Returns-only estimation risks weak identification and overfitting.

(3) Out-of-Sample Validation Is Essential: In-sample fit metrics (likelihood, residual diagnostics) can mislead. Heston improved residual kurtosis by 55% (6.92 to 3.17) but performed worse out-of-sample. Only predictive validation reveals genuine forecasting ability.

7.4. *Comparison with Prior Literature*

Our findings contrast with most option-pricing studies (Bakshi et al. 1997; Nandi 1998), which report substantial Heston superiority. This discrepancy arises because:

(1) Data Source: Prior studies calibrate to option prices, which contain direct information about implied volatility surfaces and explicitly reveal stochastic volatility patterns. Our returns-only approach tests whether returns data alone justify Heston complexity.

(2) Objective Function: Option pricing studies minimize implied volatility RMSE across strikes and maturities, a criterion where stochastic volatility's smile-generation mechanism directly applies. We assess return distribution fit, where GBM may suffice if volatility fluctuations average out.

(3) Sample Period: Many studies analyze 1980s–1990s data with extreme events (1987 crash, 1998 LTCM crisis). Our 2013–2025 sample is more recent and, outside March 2020, relatively stable.

However, our results align with studies questioning stochastic volatility's practical utility. Jorion (1995) found that delta-hedged portfolios perform similarly under constant and stochastic volatility. Christoffersen et al. (2009) documented parameter instability across estimation windows. Cont (2002) argued that no single continuous-time model fully captures all stylized facts, suggesting hybrid or non-parametric approaches may be necessary.

7.5. *Limitations*

Our study has several limitations. First, we analyze a single asset (AAPL) over one period (2013–2025). Generalization to other equities, indices, or time periods requires replication studies. Second, particle filter estimation introduces Monte Carlo variance and potential optimization failures; alternative estimators might yield different conclusions. Third, we do not incorporate jumps, regime-switching, or long memory, which may dominate stochastic volatility effects. Fourth, we use daily data; high-frequency analysis could alter conclusions. Finally, we do not jointly calibrate to option prices, which provide direct volatility information.

8. CONCLUSION

We conducted a comprehensive empirical comparison of Geometric Brownian Motion and Heston stochastic volatility models using 13 years of AAPL daily returns (2013–2025, $N = 3{,}262$). Against theoretical expectations, GBM achieved superior statistical fit across all evaluation metrics: higher log-likelihood (8502.29 vs. 8468.94), lower information criteria (AIC and BIC favor GBM by 75 and 99 points), and better out-of-sample variance forecasting (RMSE: 0.0926 vs. 0.0928). The likelihood ratio test decisively failed to reject GBM adequacy (LRT $= -66.69$, $p = 1.0$).

This unexpected outcome arises from five factors: weak volatility clustering in the sample period, parameter identifiability challenges when estimating Heston from returns data alone, overfitting of idiosyncratic noise by additional Heston parameters, potential optimization failures in particle filter maximum likelihood, and suboptimal daily data frequency for continuous-time variance process identification.

Our findings reinforce a fundamental principle in quantitative finance and statistics: *model sophistication must be matched to data informativeness.* When stochastic volatility dynamics are weak, unidentifiable, or dominated by transient shocks, parsimony prevails over complexity. Practitioners should validate model selection through rigorous out-of-sample testing and information criteria rather than defaulting to theoretically sophisticated alternatives.

These results do not invalidate stochastic volatility theory generally. Heston models remain essential for option pricing, where volatility surfaces directly reveal stochastic variance patterns. However, for returns-based modeling of daily equity dynamics during stable periods, simpler constant-volatility specifications may suffice and even outperform due to reduced overfitting risk.

Future research should: (1) replicate across diverse assets and periods, (2) incorporate auxiliary data (option prices, realized volatility), (3) test hybrid models combining jumps and stochastic volatility, (4) employ high-frequency data for improved parameter identification, and (5) develop robust estimation methods mitigating identifiability challenges. Only through cumulative empirical evidence can we delineate the precise conditions under which stochastic volatility modeling delivers genuine value over parsimonious alternatives.

ACKNOWLEDGMENTS

We thank the Research Agent collaboration framework for coordinating this multi-agent investigation. Data were obtained from Yahoo Finance via the `yfinance` Python library. Computational resources were provided by the Computational Finance Research Group.

DATA AVAILABILITY

AAPL price data are publicly available from Yahoo Finance (`https://finance.yahoo.com`). Processed returns data, estimation code, and diagnostic plots are available upon request.

REFERENCES

Aït-Sahalia, Y., & Kimmel, R. L. 2002, Journal of Financial Economics, 65, 361

Akaike, H. 1974, IEEE Transactions on Automatic Control, 19, 716

Andersen, T. G., Bollerslev, T., Diebold, F. X., & Labys, P. 2002, Journal of the American Statistical Association, 98, 501

Bakshi, G., Cao, C., & Chen, Z. 1997, Journal of Finance, 52, 2003

Bates, D. S. 1996, Review of Financial Studies, 9, 69

Bayer, C., Friz, P., & Gatheral, J. 2016, Quantitative Finance, 16, 887

Black, F. 1976, Proceedings of the 1976 American Statistical Association, Business and Economics Statistics Section, 177

Black, F., & Scholes, M. 1973, Journal of Political Economy, 81, 637

Bollerslev, T. 1986, Journal of Econometrics, 31, 307

Broadie, M., Chernov, M., & Johannes, M. 2007, Journal of Financial Economics, 86, 65

Burnham, K. P., & Anderson, D. R. 2002, Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach, 2nd edn. (New York: Springer)

Christie, A. A. 1982, Journal of Financial Economics, 10, 407

Christoffersen, P., Heston, S., & Jacobs, K. 2009, Review of Financial Studies, 22, 4259

Cont, R. 2002, Quantitative Finance, 1, 223

Doucet, A., de Freitas, N., & Gordon, N. 2001, Sequential Monte Carlo Methods in Practice (New York: Springer)

Duffie, D., & Singleton, K. J. 1997, Econometrica, 65, 929

Duffie, D., Pan, J., & Singleton, K. 2000, Econometrica, 68, 1343

Engle, R. F. 1982, Econometrica, 50, 987

Eraker, B., Johannes, M., & Polson, N. 2003, Journal of Finance, 58, 1269

Fama, E. F. 1965, Journal of Business, 38, 34

Gatheral, J., Jaisson, T., & Rosenbaum, M. 2018, Quantitative Finance, 18, 933

Glosten, L. R., Jagannathan, R., & Runkle, D. E. 1993, Journal of Finance, 48, 1779

Gordon, N. J., Salmond, D. J., & Smith, A. F. M. 1993, IEE Proceedings F, 140, 107

Hamilton, J. D. 1989, Econometrica, 57, 357

Harrison, J. M., & Kreps, D. M. 1979, Journal of Economic Theory, 20, 381

Heston, S. L. 1993, Review of Financial Studies, 6, 327

Hull, J., & White, A. 1987, Journal of Finance, 42, 281

Jackwerth, J. C., & Rubinstein, M. 1996, Journal of Finance, 51, 1611

Jacquier, E., Polson, N. G., & Rossi, P. E. 1994, Journal of Business & Economic Statistics, 12, 371

Jorion, P. 1995, Journal of Derivatives, 2, 7

Kass, R. E., & Raftery, A. E. 1995, Journal of the American Statistical Association, 90, 773

Kitagawa, G. 1996, Journal of Computational and Graphical Statistics, 5, 1

Kou, S. G. 2002, Management Science, 48, 1086

Mandelbrot, B. 1963, Journal of Business, 36, 394

Merton, R. C. 1973, Bell Journal of Economics and Management Science, 4, 141

Merton, R. C. 1976, Journal of Financial Economics, 3, 125

Nandi, S. 1998, Journal of Derivatives, 5, 9

Rubinstein, M. 1994, Journal of Derivatives, 1, 13

Singleton, K. J. 2001, Journal of Finance, 56, 1199

So, M. K. P., Lam, K., & Li, W. K. 1998, Journal of Econometrics, 83, 83

Wiggins, J. B. 1987, Journal of Financial Economics, 19, 351