

Filtered Extreme-Value Theory Meets Asymmetric GARCH: Out-of-Sample Volatility Forecasting and 99% Value-at-Risk Backtesting for U.S. Equity Indices, 2004–2018

Arin Gopalan Yadav

arin.yadav2022@vitstudent.ac.in

Vellore Institute of Technology University <https://orcid.org/0009-0008-6832-8957>

Research Article

Keywords: GARCH, EGARCH, GJR-GARCH, skewed Student-t, extreme value theory, peaks over threshold, Value-at-Risk, expected shortfall, Kupiec test, Christoffersen test, backtesting

Posted Date: October 20th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11419530/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.

Filtered Extreme-Value Theory Meets Asymmetric GARCH: Out-of-Sample Volatility Forecasting and 99% Value-at-Risk Backtesting for U.S. Equity Indices, 2004–2018

Arin Yadav

IIT Madras

arin.technovate@outlook.com

October 2026

Abstract

Accurate one-day volatility forecasts and well-calibrated tail-risk estimates are central to market-risk management, yet the separate contributions of asymmetric volatility dynamics and explicit tail modeling to Value-at-Risk (VaR) performance are rarely isolated. Using 5,030 daily log returns of the S&P 500 and the NASDAQ Composite (5 January 1999 to 31 December 2018) and a strictly out-of-sample design (3,530 daily forecasts from 22 December 2004; rolling 1,500-day window; 57 re-estimations), we compare GARCH(1,1), EGARCH(1,1,1) and GJR-GARCH(1,1,1) with skewed Student- t innovations for volatility forecasting, and a filtered extreme-value model (GARCH-EVT, peaks over threshold) for VaR and expected shortfall (ES) against RiskMetrics EWMA, 500-day historical simulation and a GJR-normal model. Both asymmetric models improve on GARCH(1,1): EGARCH lowers variance-forecast RMSE by 2.2% (S&P 500) and 1.2% (NASDAQ) (Diebold–Mariano $p = 0.017$ and 0.007) and GJR attains the lowest QLIKE on both indices. At the 99% level GARCH-EVT records 34 and 44 breaches against 35.3 expected and is not rejected by the Kupiec or Christoffersen tests (all $p > 0.15$), whereas RiskMetrics (86 and 80 breaches), historical simulation (63 and 68) and GJR-normal (70 and 69) are all rejected ($p < 0.001$). Holding the volatility filter fixed and replacing the normal quantile with an EVT tail estimate halves S&P 500 breaches (70 to 34), suggesting that, at 99%, the tail model matters more than the volatility filter. In the September 2008 to June 2009 stress window GARCH-EVT records 2 and 3 breaches in 209 days versus 16 for historical simulation, and mean predicted ES is within 3% of the realized loss on breach days. We discuss limitations, including two highly correlated indices, informal ES validation and a single confidence level.

JEL classification: C22, C52, C53, G17, G32

Keywords: GARCH, EGARCH, GJR-GARCH, skewed Student- t , extreme value theory, peaks over threshold, Value-at-Risk, expected shortfall, Kupiec test, Christoffersen test, backtesting

1 Introduction

Value-at-Risk remains the standard quantile summary of market risk, and the Basel framework has moved toward expected shortfall for internal-model capital (Basel Committee on Banking Supervision, 2019). Whatever the risk measure, its quality rests on two modeling choices that are often bundled together in practice: how conditional volatility is filtered, and how the tail of the standardized return distribution is represented. Conditional-volatility models capture

volatility clustering (Engle, 1982; Bollerslev, 1986) and the leverage effect (Black, 1976; Nelson, 1991; Glosten et al., 1993), while extreme value theory (EVT) provides a principled way to extrapolate into the tail (Balkema and de Haan, 1974; Pickands, 1975; Embrechts et al., 1997). McNeil and Frey (2000) combine the two in a two-step filtered EVT approach. We revisit this combination with modern asymmetric filters and a transparent ablation design.

We address three research questions.

RQ1 (volatility forecasting). Do asymmetric GARCH specifications (EGARCH, GJR) improve one-day-ahead out-of-sample volatility forecasts relative to symmetric GARCH(1,1), a benchmark known to be hard to beat (Hansen and Lunde, 2005)?

RQ2 (tail risk). Does a filtered GARCH–EVT model deliver correctly calibrated 99% VaR, as assessed by the unconditional-coverage test of Kupiec (1995) and the conditional-coverage test of Christoffersen (1998), relative to RiskMetrics, historical simulation and a normal-quantile GARCH model?

RQ3 (attribution). At the 99% level, is VaR calibration driven mainly by the volatility filter or by the tail model?

Our contributions are (i) a look-ahead-free rolling evaluation of three volatility specifications under a common innovation distribution, (ii) a four-way VaR comparison that includes a GJR-normal model constructed to isolate the tail-model effect, and (iii) a stress-window analysis for 2008–09. The pipeline is fully reproducible from a single script.

2 Related Literature

Volatility models. The ARCH model (Engle, 1982) and GARCH (Bollerslev, 1986) are the foundation. EGARCH (Nelson, 1991) models log-variance and allows sign-dependent responses without positivity constraints; GJR (Glosten et al., 1993) adds a threshold term to the variance recursion. Hansen and Lunde (2005) compare hundreds of volatility models and find little evidence that anything beats GARCH(1,1) for exchange rates, but clear evidence that models with leverage effects outperform it for equity returns, which motivates our asymmetric comparison.

Innovation distributions. Daily index returns are leptokurtic and mildly skewed (Cont, 2001). We use Hansen’s (1994) skewed Student- t for all three models so that differences across specifications reflect the variance recursion rather than the error distribution.

Extreme value theory in risk management. The Pickands–Balkema–de Haan theorem implies that exceedances over a high threshold are approximately generalized Pareto distributed (GPD). McNeil and Frey (2000) apply the peaks-over-threshold method to GARCH-filtered residuals; McNeil et al. (2005) give a textbook treatment.

Backtesting. Kupiec (1995) tests the unconditional coverage of VaR exceedances and Christoffersen (1998) adds an independence component. Expected shortfall is a coherent risk measure (Artzner et al., 1999) but is not elicitable on its own (Gneiting, 2011); dedicated ES backtests exist (Acerbi and Székely, 2014), and we use only an informal check.

Forecast comparison. Because latent variance is proxied by squared returns, loss functions must be robust to proxy noise. Patton (2011) shows that squared error and QLIKE are the robust choices, and we compare forecasts with the Diebold and Mariano (1995) test using Newey and West (1987) standard errors.

3 Data

We use daily closing levels of the S&P 500 and the NASDAQ Composite as distributed with the `arch` Python package (Sheppard, n.d.), from 4 January 1999 to 31 December 2018, and compute percentage log returns $r_t = 100 \cdot \ln(P_t/P_{t-1})$. This yields $T = 5,030$ returns per index (5 January 1999 to 31 December 2018). The first 1,500 returns form the initial estimation window, leaving 3,530 out-of-sample forecasts from 22 December 2004 to 31 December 2018.

Table 1: Descriptive statistics of daily log returns (%).

Statistic	S&P 500	NASDAQ
Observations	5,030	5,030
Mean	0.0142	0.0219
Standard deviation	1.2038	1.5932
Skewness	−0.205	−0.015
Excess kurtosis	8.169	5.427
Minimum	−9.47	−10.17
Maximum	10.96	13.25
Annualized volatility, out-of-sample	18.8%	20.4%
Annualized volatility, 1 Sep 2008–30 Jun 2009	49.0%	48.7%

The correlation between the two return series is 0.887. The Jarque–Bera test rejects normality for both series ($p < 0.001$), consistent with the large excess kurtosis. The stress window (1 September 2008 to 30 June 2009) contains 209 out-of-sample trading days, and realized volatility in it is roughly 2.6 times the out-of-sample average.

4 Methodology

4.1 Volatility models

All models share a constant-mean specification and Hansen skewed Student- t innovations:

$$r_t = \mu + \varepsilon_t, \quad \varepsilon_t = \sigma_t z_t, \quad z_t \stackrel{\text{i.i.d.}}{\sim} \text{skewed-}t(\nu, \lambda), \quad \mathbb{E}[z_t] = 0, \quad \text{Var}(z_t) = 1. \quad (1)$$

GARCH(1,1):

$$\sigma_t^2 = \omega + \alpha \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2. \quad (2)$$

GJR-GARCH(1,1,1):

$$\sigma_t^2 = \omega + (\alpha + \gamma \mathbf{1}\{\varepsilon_{t-1} < 0\}) \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2. \quad (3)$$

EGARCH(1,1,1):

$$\ln \sigma_t^2 = \omega + \alpha(|e_{t-1}| - \kappa) + \gamma e_{t-1} + \beta \ln \sigma_{t-1}^2, \quad e_t = \varepsilon_t / \sigma_t, \quad (4)$$

where κ is a centering constant. In (3) a leverage effect corresponds to $\gamma > 0$; in (4), where the sign term multiplies the standardized residual, it corresponds to $\gamma < 0$. All parameters are estimated by maximum likelihood.

4.2 Rolling out-of-sample protocol

At each re-estimation date t_0 (every 63 trading days, 57 re-estimations in total; the last block contains only 2 days), the three models are estimated on the window $[t_0 - 1500, t_0)$. The estimated parameters are then held fixed and the conditional variance recursion is run forward over the next block of up to 63 days. The one-step-ahead conditional standard deviation σ_t used for day t depends only on returns through $t - 1$ and on parameters estimated before t_0 , so no forecast uses future information.

4.3 Volatility forecast evaluation

Latent variance is proxied by squared returns r_t^2 . We report three criteria over the 3,530 forecasts:

$$\text{RMSE}_{\text{var}} = \sqrt{\text{mean}[(\sigma_t^2 - r_t^2)^2]}, \quad (5)$$

$$\text{RMSE}_{\text{vol}} = \sqrt{\text{mean}[(\sigma_t - |r_t|)^2]}, \quad (6)$$

$$\text{QLIKE} = \text{mean}[\ln \sigma_t^2 + r_t^2 / \sigma_t^2]. \quad (7)$$

Squared error on variance (5) and QLIKE (7) are robust to proxy noise (Patton, 2011); (6) is reported for interpretability, since $|r_t|$ is a biased proxy for σ_t . To test whether EGARCH improves on GARCH(1,1), we regress the loss differential $d_t = \text{loss}_t^{\text{GARCH}} - \text{loss}_t^{\text{EGARCH}}$ on a constant with Newey–West standard errors (HAC, 10 lags); a positive t -statistic favors EGARCH. We use squared-error loss on variance and QLIKE as the two loss functions.

4.4 Filtered GARCH–EVT Value-at-Risk and expected shortfall

The VaR model follows McNeil and Frey (2000) with a GJR skewed- t filter. Within each estimation window, standardized residuals are $z_t = (r_t - \mu) / \sigma_t$. Writing losses as $L_t = -z_t$, we set the threshold u at the 90th percentile of L_t (about 150 exceedances per window) and fit a generalized Pareto distribution with shape ξ and scale β to the excesses $L_t - u$ by maximum likelihood, with location fixed at 0. With n residuals and N_u exceedances, the tail quantile and ES of the standardized loss at tail probability p are

$$q_p = u + \frac{\beta}{\xi} \left[\left(\frac{n}{N_u} p \right)^{-\xi} - 1 \right], \quad (8)$$

$$\text{ES}_p = \frac{q_p + \beta - \xi u}{1 - \xi}, \quad \xi < 1. \quad (9)$$

The one-day 99% VaR forecast (positive number, loss convention) for day t is $\text{VaR}_t = \sigma_t q_{0.01} - \mu$, and the corresponding ES forecast is $\sigma_t \text{ES}_{0.01} - \mu$. The GPD parameters are re-estimated in each window, with σ_t filtered forward with fixed parameters as in Section 4.2.

4.5 Benchmarks

RiskMetrics. EWMA variance with $\lambda = 0.94$ (initialized with the variance of the first 250 returns) and a normal quantile, $\text{VaR}_t = z_{0.99} \sigma_t$.

Historical simulation (HS). The empirical 1% quantile of the previous 500 returns.

GJR-normal. The same GJR skewed- t conditional mean and volatility as GARCH–EVT, but with the normal 99% quantile (2.326) in place of the EVT quantile. This model differs from GARCH–EVT only in the tail quantile, which isolates the contribution of the tail model (RQ3).

4.6 Backtesting

Let $h_t = \mathbf{1}\{\text{loss}_t > \text{VaR}_t\}$ be the breach indicator. With $n = 3,530$ forecasts and tail probability 1%, the expected number of breaches is 35.3. The Kupiec (1995) proportion-of-failures likelihood ratio, LR_{uc} , tests $\mathbb{E}[h_t] = 1\%$ and is $\chi^2(1)$ under the null. The Christoffersen (1998) conditional-coverage statistic $LR_{cc} = LR_{uc} + LR_{ind}$ adds a first-order Markov independence test and is $\chi^2(2)$. For ES we make an informal check, comparing the mean realized loss on breach days with the mean predicted ES on those days.

5 Empirical Results

5.1 Stationarity and ARCH effects

Table 2: Unit-root and ARCH tests on log prices and log returns.

Test	S&P 500	NASDAQ
ADF, log price (p -value)	0.915	0.908
ADF, log return (p -value)	3.4×10^{-29}	1.7×10^{-28}
KPSS, log return (p -value)	≥ 0.10	≥ 0.10
ARCH-LM, 10 lags (p -value)	$< 10^{-200}$	$< 10^{-200}$

Prices contain a unit root, returns are stationary (ADF rejects, KPSS does not), and the ARCH-LM test overwhelmingly rejects the absence of conditional heteroskedasticity, confirming volatility clustering and justifying the GARCH family.

5.2 Leverage effect (full-sample estimates)

Table 3: Full-sample asymmetry coefficients (skewed- t innovations).

Parameter	S&P 500	NASDAQ
EGARCH γ (p -value)	$-0.160 (2.4 \times 10^{-32})$	$-0.110 (1.3 \times 10^{-21})$
GJR γ (p -value)	$0.190 (4.2 \times 10^{-17})$	$0.138 (1.1 \times 10^{-12})$

Both asymmetry coefficients are strongly significant with the signs implying that negative shocks raise future volatility more than positive shocks of the same size. These full-sample estimates are in-sample evidence of the leverage effect; the out-of-sample value of asymmetry is assessed next.

5.3 Out-of-sample volatility forecasts (RQ1)

Relative to GARCH(1,1), EGARCH lowers the variance-forecast RMSE by 2.2% for the S&P 500 and 1.2% for the NASDAQ, and GJR by 2.2% and 1.2% respectively; volatility-forecast RMSE falls by 3.9% (S&P 500) and 2.0% (NASDAQ) for EGARCH and by 1.4% and 0.4% for GJR. QLIKE ranks the models identically on both indices, $\text{GJR} < \text{EGARCH} < \text{GARCH}$.

EGARCH improves significantly on GARCH(1,1) under both loss functions for both indices (all $p < 0.05$). We did not test GJR against GARCH or EGARCH, so the finer ranking between the two asymmetric models (GJR best on QLIKE, essentially tied with EGARCH on variance RMSE) is descriptive rather than inferential. The economic size of the gain is modest, in line

Table 4: Out-of-sample volatility forecast accuracy, 3,530 days.

Index	Model	RMSE variance	RMSE volatility	QLIKE
S&P 500	GARCH(1,1)	4.5156	0.8206	0.7626
S&P 500	EGARCH	4.4143	0.7890	0.7292
S&P 500	GJR-GARCH	4.4181	0.8091	0.7166
NASDAQ	GARCH(1,1)	4.5367	0.8782	1.0960
NASDAQ	EGARCH	4.4823	0.8607	1.0647
NASDAQ	GJR-GARCH	4.4823	0.8744	1.0585

Table 5: Diebold–Mariano tests, EGARCH versus GARCH(1,1) (Newey–West, 10 lags).

Loss function	S&P 500: $t(p)$	NASDAQ: $t(p)$
Squared error on variance	2.40 (0.017)	2.70 (0.007)
QLIKE	2.26 (0.024)	2.88 (0.004)

with the difficulty of beating GARCH(1,1) emphasized by Hansen and Lunde (2005), but it is consistent across indices, losses and models.

5.4 99% VaR backtests (RQ2)

Expected breaches: 35.3 of 3,530 forecasts (1.00%).

Table 6: 99% one-day VaR backtests, 3,530 out-of-sample forecasts.

Index	Model	Breaches	Breach rate	Kupiec p	Christoffersen p
S&P 500	GARCH–EVT	34	0.96%	0.825	0.701
S&P 500	RiskMetrics	86	2.44%	4.3×10^{-13}	1.9×10^{-12}
S&P 500	Historical simulation	63	1.78%	2.4×10^{-5}	2.9×10^{-6}
S&P 500	GJR-normal	70	1.98%	2.3×10^{-7}	1.4×10^{-6}
NASDAQ	GARCH–EVT	44	1.25%	0.156	0.210
NASDAQ	RiskMetrics	80	2.27%	8.8×10^{-11}	5.2×10^{-10}
NASDAQ	Historical simulation	68	1.93%	9.3×10^{-7}	9.0×10^{-7}
NASDAQ	GJR-normal	69	1.95%	4.6×10^{-7}	2.9×10^{-6}

GARCH–EVT is the only model that is not rejected by either test on either index. On the S&P 500 its breach count (34) is almost exactly the expected 35.3. On the NASDAQ it records 44 breaches (1.25%), a mild excess that is not statistically significant (Kupiec $p = 0.156$); given the limited power of coverage tests at the 1% level with 3,530 observations, this should be read as absence of evidence of miscalibration rather than proof of correct calibration. The three benchmarks overshoot the nominal rate by a factor of 1.8 to 2.4, and all are rejected at conventional levels.

5.5 What drives VaR calibration? (RQ3)

GJR-normal and GARCH-EVT share the same conditional mean and volatility; they differ only in the quantile of the standardized loss. Replacing the normal quantile with the EVT quantile reduces breaches from 70 to 34 on the S&P 500 and from 69 to 44 on the NASDAQ. By contrast, filtering volatility but keeping a normal quantile (70 and 69 breaches) does no better at the 99% level than the unfiltered 500-day empirical quantile (63 and 68 breaches). Within this study, therefore, the representation of the tail, not the volatility filter, accounts for most of the improvement in 99% VaR calibration. This attribution is based on a single filter and a single confidence level, and should not be generalized without further testing.

5.6 Stress window, September 2008 to June 2009

At a 1% tail probability, the expected number of breaches in the 209-day window is 2.09.

Table 7: 99% VaR breaches in the stress window (209 days).

Model	S&P 500 breaches	NASDAQ breaches
GARCH-EVT	2	3
RiskMetrics	6	5
GJR-normal	7	6
Historical simulation	16	16

GARCH-EVT stays close to the nominal count in the stress window. Historical simulation, which reacts to the surge in volatility only slowly because it weights all 500 prior days equally, breaches 16 times (7.7% of days), roughly 7.7 times the expected number. The volatility-filtered models (RiskMetrics, GJR-normal) fall in between, which suggests that the volatility filter helps with timing in turbulent regimes even though it does not by itself fix the level of tail calibration.

5.7 Expected shortfall (informal check)

Table 8: Predicted versus realized loss on 99% VaR breach days.

Index	Mean realized loss	Mean predicted ES	Difference
S&P 500	2.825%	2.746%	−2.8%
NASDAQ	3.108%	3.095%	−0.4%

On days when the 99% VaR is breached, the model’s predicted ES is within 3% of the realized loss on the S&P 500 and within 0.5% on the NASDAQ, with a slight underestimate in both cases. This is a conditional-on-breach comparison for a single model and is not a formal ES backtest.

6 Discussion

Asymmetry helps, modestly. The significant leverage coefficients (Section 5.2) translate into out-of-sample forecast gains of one to two percent in RMSE terms. Under QLIKE, which penalizes under-prediction of variance more heavily than over-prediction, GJR is best on both

indices. A plausible explanation is that GJR’s threshold term reacts more sharply to large negative shocks than a log-variance specification, but we do not test this.

Why EVT works here. The 90th-percentile threshold leaves about 150 exceedances per window, enough for stable GPD estimation, and the 1% quantile lies only a factor of ten beyond the threshold in tail probability, so the extrapolation is modest. The Student- t family used for the filter can describe the bulk of the distribution, but the quantile of the standardized residual at 99% is estimated directly from the data, which may account for the improved calibration relative to the normal quantile. This is a hypothesis, not a tested result.

Unconditional versus conditional methods. Historical simulation cannot adapt to changing volatility and breaches in clusters during stress; RiskMetrics adapts quickly but uses a fixed decay and a normal quantile and is the worst performer overall, with breach rates of 2.4% and 2.3% at a nominal 1%.

7 Limitations and Threats to Validity

1. **Two highly correlated indices.** The return correlation is 0.887, so the two indices are not independent replications, and both are U.S. equity price indices over one sample period.
2. **Single confidence level and horizon.** We evaluate 99% VaR at a one-day horizon only. Results at 97.5% or at multi-day horizons may differ.
3. **Informal ES validation.** The ES check is conditional-on-breach and descriptive; formal tests (for example Acerbi–Székely) are not performed, and ES is not elicitable on its own (Gneiting, 2011).
4. **Limited inference on model ranking.** The Diebold–Mariano test is applied only to EGARCH versus GARCH(1,1). GJR versus EGARCH is not tested, and no Model Confidence Set (Hansen et al., 2011) is computed. The volatility RMSE in (6) uses a biased proxy.
5. **Fixed design parameters.** The window (1,500), re-estimation frequency (63), EVT threshold (90th percentile), RiskMetrics λ (0.94) and HS window (500) are not varied.
6. **Model scope.** All models use a constant mean. Neither ARMA mean dynamics nor GARCH-in-mean are estimated, and skewed- t shape parameters and GPD parameters are not reported.
7. **Coverage-test power.** With 3,530 forecasts and a 1% tail probability, coverage tests have limited power against moderate misspecification, so non-rejection is weak evidence.
8. **Sample.** The sample ends in December 2018 and includes the 2008–09 crisis; results are conditional on this history and exclude, for example, the 2020 volatility shock.
9. **Data.** Price indices exclude dividends, which is irrelevant for volatility and VaR of price changes but matters if the VaR is interpreted for total-return positions.

8 Conclusion

On 3,530 out-of-sample forecasts for the S&P 500 and the NASDAQ Composite, asymmetric GARCH models improve modestly but consistently on GARCH(1,1) (EGARCH variance RMSE -2.2% and -1.2% , Diebold–Mariano $p = 0.017$ and 0.007), and a filtered GARCH–EVT model is the only one of four VaR approaches that is not rejected by Kupiec and Christoffersen tests at the 99% level (34 and 44 breaches against 35.3 expected). An ablation that replaces only the quantile of a GJR filter shows that the tail model accounts for most of the calibration gain (70 to

34 breaches on the S&P 500), and in the 2008–09 stress window GARCH–EVT records 2 and 3 breaches against 16 for historical simulation. The practical implication is that volatility filtering and tail modeling address different failure modes, timing and level, and that a normal-quantile GARCH model is not a substitute for explicit tail estimation at 99%.

Natural extensions, in order of priority, are (i) formal ES backtests and a 97.5% ES evaluation in line with current regulatory practice, (ii) Model Confidence Set comparison of all volatility models, (iii) a sensitivity grid over threshold, window and re-estimation frequency, (iv) a wider cross-section of assets and an out-of-sample extension through 2020–22, and (v) alternative filters such as realized-volatility models.

Reproducibility

All results are produced by a single Python script (`project1_garch_evt_var.py`) using `arch` 8.0.0 (Sheppard, n.d.) (volatility models and data), `SciPy` (generalized Pareto estimation and χ^2 tests) and `statsmodels` (ADF, KPSS, ARCH-LM and HAC inference). Output is saved to `project1_results.json`. All design parameters are stated in Sections 4.2 to 4.5.

References

- Acerbi, C., and Székely, B. (2014). Backtesting expected shortfall. *Risk*, December, 76–81.
- Artzner, P., Delbaen, F., Eber, J.-M., and Heath, D. (1999). Coherent measures of risk. *Mathematical Finance*, 9(3), 203–228.
- Balkema, A. A., and de Haan, L. (1974). Residual life time at great age. *Annals of Probability*, 2(5), 792–804.
- Basel Committee on Banking Supervision (2019). *Minimum capital requirements for market risk*. Bank for International Settlements.
- Black, F. (1976). Studies of stock price volatility changes. *Proceedings of the 1976 Meetings of the American Statistical Association, Business and Economic Statistics Section*, 177–181.
- Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3), 307–327.
- Christoffersen, P. F. (1998). Evaluating interval forecasts. *International Economic Review*, 39(4), 841–862.
- Cont, R. (2001). Empirical properties of asset returns: Stylized facts and statistical issues. *Quantitative Finance*, 1(2), 223–236.
- Diebold, F. X., and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3), 253–263.
- Embrechts, P., Klüppelberg, C., and Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*. Springer.
- Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50(4), 987–1007.
- Glosten, L. R., Jagannathan, R., and Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *Journal of Finance*, 48(5), 1779–1801.
- Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494), 746–762.
- Hansen, B. E. (1994). Autoregressive conditional density estimation. *International Economic Review*, 35(3), 705–730.

- Hansen, P. R., and Lunde, A. (2005). A forecast comparison of volatility models: Does anything beat a GARCH(1,1)? *Journal of Applied Econometrics*, 20(7), 873–889.
- Hansen, P. R., Lunde, A., and Nason, J. M. (2011). The model confidence set. *Econometrica*, 79(2), 453–497.
- Kupiec, P. H. (1995). Techniques for verifying the accuracy of risk measurement models. *Journal of Derivatives*, 3(2), 73–84.
- McNeil, A. J., and Frey, R. (2000). Estimation of tail-related risk measures for heteroscedastic financial time series: An extreme value approach. *Journal of Empirical Finance*, 7(3–4), 271–300.
- McNeil, A. J., Frey, R., and Embrechts, P. (2005). *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton University Press.
- Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica*, 59(2), 347–370.
- Newey, W. K., and West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3), 703–708.
- Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, 160(1), 246–256.
- Pickands, J. (1975). Statistical inference using extreme order statistics. *Annals of Statistics*, 3(1), 119–131.
- Sheppard, K. (n.d.). **arch**: ARCH models in Python (software package).