Proceedings of ICFTBA 2026 Symposium: Digital Government, Finance, Accountability, and Sustainable Economic Development
DOI: 10.54254/2754-1169/2026.AB39272

Out-of-Sample Volatility Forecasting for the SPDR S&P Retail ETF: A Comparison of GARCH and GJR-GARCH Models

Jingyi Yang

Department of Mathematics, Imperial College London, London, UK
toby.yang26@imperial.ac.uk

Abstract. This paper compares the out-of-sample volatility forecasting performance of several GARCH-family specifications for the SPDR S&P Retail ETF (XRT). The analysis uses daily adjusted closing prices from June 2006 to September 2026 to construct percentage log returns. Observations through December 2022 form the training sample, leaving 939 trading days for out-of-sample evaluation. The comparison covers GARCH(1,1) and GJR-GARCH(1,1) models under both normal and standardized Student-t innovations. One-step-ahead variance forecasts are produced with an expanding estimation window, with parameters re-estimated every 20 trading days. A 22-day historical-volatility measure serves as the benchmark. Forecast performance is evaluated mainly by QLIKE, RMSE and MAE, with QLIKE-based Diebold–Mariano tests used as supplementary evidence. GJR-GARCH with Student-t innovations records the lowest forecast loss across QLIKE, RMSE and MAE. The positive asymmetric coefficient also indicates a stronger conditional-variance response to negative innovations than to positive innovations of the same magnitude. These gains are small, however, and none of the pairwise QLIKE differences remains significant after Holm adjustment. Replacing the ARMA(0,1) mean specification with ARMA(0,0) does not change the model ranking. Overall, the evidence points to some benefit from allowing for asymmetry and heavy tails, but the differences in predictive performance remain limited.

Keywords: Volatility forecasting, GARCH, GJR-GARCH, Student-t innovations, exchange-traded fund

1. Introduction

Volatility is central to financial risk because it measures how widely asset returns fluctuate over time. It is also persistent rather than constant: quiet periods tend to be followed by further quiet observations, while turbulent periods often cluster together. Financial returns may also display heavy tails and asymmetric responses to positive and negative innovations. These features make constant-variance models a poor description of the data and motivate the conditional heteroskedasticity models examined in this paper.

This paper asks whether allowing for asymmetric volatility dynamics and heavy-tailed innovations improves out-of-sample forecasts for XRT. To examine this, GARCH(1,1) and GJR-

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2754-1169/2026.AB39272

GARCH(1,1) are estimated under both normal and standardized Student-t innovations. The GJR specification allows volatility to react differently to positive and negative innovations of the same magnitude, while the Student-t distribution accommodates heavier tails in the return distribution.

Daily adjusted closing prices from June 2006 to September 2026 are converted into percentage log returns. The sample is split chronologically into training and test periods so that model estimation is separated from forecast evaluation. An ARMA specification is used for the conditional mean, and the four GARCH-family models are then estimated jointly with this mean equation. Out-of-sample forecasts are generated recursively with an expanding window, with parameters re-estimated every 20 trading days. A 22-day historical-volatility measure is used as a simple benchmark.

Forecast performance is compared using QLIKE, RMSE and MAE. QLIKE-based Diebold–Mariano tests provide supplementary evidence on whether the observed loss differences are statistically meaningful, with Holm adjustment applied to the pairwise comparisons. Because ARMA(0,1) is only marginally preferred to ARMA(0,0) by BIC and some residual autocorrelation remains, the forecasting exercise is also repeated under an ARMA(0,0) mean specification as a robustness check.

The out-of-sample results favour GJR-GARCH with Student-t innovations in numerical terms, as it produces the lowest QLIKE, RMSE and MAE. The estimated asymmetric coefficient is also positive and statistically significant, indicating a stronger variance response to negative innovations. The differences in forecast loss are nevertheless small, and the pairwise QLIKE comparisons do not remain significant after Holm adjustment. Repeating the exercise with an ARMA(0,0) mean specification leaves the model ranking unchanged. The evidence therefore suggests a modest forecasting advantage from allowing for asymmetry and heavy tails, rather than clear statistical dominance.

The remainder of the study is organised as follows. Section 2 provides the relevant background and literature. Section 3 describes the data, model specifications, forecasting design and evaluation methods. Section 4 presents the empirical results and robustness analysis. Section 5 concludes.

2. Literature review

2.1. GARCH models and volatility persistence

Volatility clustering is a common feature of financial returns: large price movements tend to occur in groups, while quieter periods can also persist. A constant-variance model cannot represent this pattern effectively. Engle [1] addressed this issue through the ARCH framework, in which conditional variance changes over time in response to past squared innovations.

Bollerslev [2] extended ARCH by allowing conditional variance to depend not only on past squared innovations but also on its own previous values. This leads to the GARCH framework, which captures persistent volatility dynamics with relatively few parameters. The GARCH(1,1) model is especially useful as a benchmark because its persistence can be summarized by the combined effects of recent shocks and past conditional variance. Greater persistence implies that the effect of a volatility shock fades more slowly over time.

For this study, GARCH provides a natural benchmark for modelling persistent volatility in XRT returns. Its main limitation is the symmetry restriction: positive and negative innovations of the same magnitude have the same effect on the next period's conditional variance. This may be too restrictive if adverse shocks generate a stronger volatility response. The possibility of such asymmetry motivates the GJR-GARCH specification considered next.

DOI: 10.54254/2754-1169/2026.AB39272

2.2. Asymmetric and heavy-tailed volatility

To allow for asymmetric volatility responses, Glosten, Jagannathan, and Runkle [3] extended the standard GARCH framework so that the effect of an innovation can depend on its sign. In the resulting GJR-GARCH model, a negative innovation may affect conditional variance differently from a positive innovation of the same magnitude. This makes the model suitable for examining whether adverse shocks are associated with a stronger volatility response in XRT returns.

The choice of conditional distribution is also important. Normal innovations provide a useful benchmark, but financial returns often contain more extreme observations than a normal distribution would imply. A standardized Student-t distribution allows for heavier tails and can therefore accommodate large innovations more flexibly while preserving the conditional-variance interpretation.

These two modelling choices lead naturally to the four specifications used in this study: GARCH-Normal, GJR-Normal, GARCH-Student-t and GJR-Student-t. Comparing them allows the effects of asymmetry and heavy-tailed innovations to be examined separately. This comparison is especially relevant for XRT because the return data show a strong departure from unconditional normality, although this alone does not determine the correct conditional innovation distribution.

2.3. Volatility forecast evaluation

Evaluating volatility forecasts is difficult because conditional variance is not directly observable. A proxy is therefore needed for empirical comparison. In this study, squared daily returns are used because higher-frequency realised-volatility measures are unavailable. However, squared returns are a noisy proxy, so differences in forecast performance should be interpreted with caution.

The choice of loss function also matters when comparing volatility forecasts. Patton [4] shows that forecast rankings can depend on both the volatility proxy and the loss function used. For this reason, QLIKE is treated as the primary evaluation criterion, while RMSE and MAE are reported as complementary measures. Considering all three helps show whether the model ranking is robust to the way forecast errors are measured.

Lower forecast loss does not necessarily imply a statistically meaningful improvement in predictive accuracy. Diebold and Mariano [5] provide a framework for testing whether the average loss differential between two forecasts is different from zero. In this study, QLIKE-based Diebold–Mariano tests are therefore used as supplementary evidence. Newey and West [6] standard errors account for serial dependence and heteroskedasticity in the loss differentials, while the Holm [7] procedure adjusts for the multiple pairwise comparisons.

Taken together, these choices provide the evaluation framework for the empirical analysis. GARCH serves as the baseline model, GJR-GARCH allows for asymmetric volatility responses, and the alternative innovation distributions capture possible heavy tails. Their out-of-sample forecasts are compared using several loss measures, with statistical tests used to assess whether observed differences are likely to reflect genuine predictive improvements rather than sampling variation.

3. Research design

3.1. Data source and variable construction

This study uses daily adjusted closing prices for the SPDR S&P Retail ETF (XRT) obtained from Yahoo Finance. Adjusted prices are used to reduce the effect of distributions and other price

DOI: 10.54254/2754-1169/2026.AB39272

adjustments on calculated returns. The resulting return sample covers 23 June 2006 to 30 September 2026 and contains 5,099 trading-day observations.

The sample is divided chronologically into training and test periods. The training sample runs from 23 June 2006 to 30 December 2022 and contains 4,160 observations, while the test sample covers 3 January 2023 to 30 September 2026 and contains 939 observations. Test-period observations are excluded from the initial model selection and estimation, but once an observation has occurred, it may enter later estimation windows under the predetermined expanding-window procedure.

Let P_t denote the adjusted closing price on trading day t. Daily logarithmic returns are defined in equation (1):

$$r_t = 100ln(\frac{P_t}{P_{t-1}}) \tag{1}$$

Returns are measured in percentage units and conditional variances in squared percentage units. Since daily conditional variance is unobservable, squared returns, $y_t = r_t^2$, serve as a volatility proxy. For conditional mean μ_t and variance h_t, equation (2) gives:

$$E(r_t^2 \mid F_{t-1}) = h_t + \mu_t^2 \tag{2}$$

Evaluating conditional variance against squared returns therefore requires the conditional-mean contribution to be relatively small. This study also examines the relative magnitude of squared mean forecasts and evaluates $\widehat{h}_t + \widehat{\mu}_t^2$ against squared returns as a conditional second-moment sensitivity analysis. Squared returns are not treated as direct observations of true volatility.

3.2. Mean equation and preliminary tests

The conditional mean is represented by an ARMA equation in (3):

$$r_t = c + \sum_{i=1}^{p} \phi_i r_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t \tag{3}$$

where ε_t is the innovation relative to the conditional mean. The mean equation accounts for possible linear dependence; return point forecasting is not a separate research objective.

The candidate set consists of ARMA(0,0), ARMA(1,0), ARMA(0,1), ARMA(1,1), ARMA(1,2), ARMA(2,1), and ARMA(2,2). All candidates are estimated on the training sample. BIC, following Schwarz [8], is the primary selection criterion, with AIC reported as supplementary evidence. The selected order is applied consistently across the GARCH models and remains fixed throughout the test period.

Preliminary analysis includes descriptive statistics, the normality test of Jarque and Bera [9], and an augmented unit-root test based on Dickey and Fuller [10], with a constant and linear trend in the test regression. Stationary returns are modelled without further differencing. The test of Ljung and Box [11] examines ARMA residual dependence, and the ARCH-LM test associated with Engle [1] assesses conditional heteroskedasticity. Conventional results are reported at lags 12 and 20. Information criteria provide relative model comparisons but do not replace residual diagnostics.

Because the BIC difference between ARMA(0,1) and ARMA(0,0) is very small, the evidence favouring the MA(1) term is weak. Therefore, ARMA(0,0), including an estimated constant mean, is considered as an alternative mean specification for robustness analysis. The four GARCH

DOI: 10.54254/2754-1169/2026.AB39272

specifications are re-estimated under ARMA(0,0) using the same training and test samples, expanding-window forecasting procedure, and 20-trading-day refitting schedule.

3.3. GARCH model specifications

Following Bollerslev [2], innovations are written as $\varepsilon_t = \sqrt{h_t} z_t$, where $E(z_t) = 0$ and $\mathrm{Var}(z_t) = 1$. The GARCH(1,1) variance equation in (4) is:

$$h_t = \omega + \alpha \varepsilon_{t-1}^2 + \beta h_{t-1} \tag{4}$$

The parameters α and β capture the effects of the previous innovation and previous conditional variance. Under the usual non-negativity restrictions, $\alpha + \beta < 1$ corresponds to stationarity with finite unconditional variance. Values nearer one indicate stronger persistence.

The GJR-GARCH(1,1) model of Glosten, Jagannathan, and Runkle [3] allows an asymmetric response, as shown in equation (5):

$$h_t = \omega + \alpha \varepsilon_{t-1}^2 + \gamma I(\varepsilon_{t-1} < 0) \varepsilon_{t-1}^2 + \beta h_{t-1} \tag{5}$$

The shock coefficients for positive and negative innovations are α and $\alpha + \gamma$, respectively. A significantly positive γ indicates a larger conditional variance increment following a negative innovation of equal absolute magnitude. Innovation signs are defined relative to the conditional mean, rather than simply by observed return signs. Under a symmetric innovation distribution with unit variance, overall GJR persistence is given by equation (6):

$$\rho = \alpha + \beta + \frac{\gamma}{2} \tag{6}$$

Standard normal innovations provide the baseline. Standardized Student-t innovations with unit variance and degrees of freedom $\nu > 2$ provide the robustness specification. The four models are GARCH-Normal, GJR-Normal, GARCH-Student-t, and GJR-Student-t.

Mean and variance parameters are estimated jointly after order selection. Inference primarily uses robust standard errors. Model comparisons report log-likelihood and total, rather than per-observation, AIC and BIC. Diagnostics examine standardized residuals, their squares, and remaining ARCH effects. Conventional Ljung–Box and ARCH-LM diagnostics are supplemented by the native weighted Ljung–Box and weighted ARCH-LM tests reported by rugarch, drawing on the weighted portmanteau framework of Fisher and Gallagher [12]. Because the weighted tests use different weighting schemes and reference distributions, their p-values are interpreted separately from the conventional diagnostics. Computation uses R 4.6.1 and rugarch 1.5-6.

3.4. Out-of-sample forecasting and benchmark design

An expanding window generates daily one-step-ahead forecasts. For target day t, only information available through $t-1$ is used to obtain $\widehat{\mu}_{t|t-1}$ and $\widehat{h}_{t|t-1}$. Parameters are re-estimated every 20 trading days to reduce repeated optimization costs. Between refits, new daily returns update conditional means, innovations, and conditional variances. This remains a sequence of one-step forecasts, rather than a single 20-step forecast path or a constant variance forecast within each block.

All four models use identical initial training data, forecast dates, and refit schedules. The 939 test dates require 47 estimation blocks per model, or 188 rolling estimations in total. Every comparison

DOI: 10.54254/2754-1169/2026.AB39272

uses the same test observations.

The historical benchmark is the mean squared return over the preceding 22 trading days, as defined in equation (7):

$$\widehat{h}_{t|t-1}^{HV22} = \frac{1}{22}\sum_{j=1}^{22} r_{t-j}^2 \tag{7}$$

It excludes the target-day return and follows the squared-return convention of the main evaluation.

3.5. Forecast evaluation and statistical testing

Let T be the test-sample size, $y_t = r_t^2$, and $\widehat{h}_t$ the one-step variance forecast. Forecast accuracy is measured by QLIKE, RMSE and MAE, as defined in equations (8)-(10):

$$QLIKE = \frac{1}{T}\sum_{t=1}^{T}(ln\widehat{h}_t + \frac{y_t}{\widehat{h}_t}) \tag{8}$$

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(y_t - \widehat{h}_t)^2} \tag{9}$$

$$MAE = \frac{1}{T}\sum_{t=1}^{T}|y_t - \widehat{h}_t| \tag{10}$$

RMSE and MAE are expressed in squared percentage units (%²), while QLIKE is reported as a unitless loss score. Lower values indicate smaller losses. QLIKE is the primary criterion; RMSE and MAE provide complementary evidence. Patton [4] establishes expected-loss ranking properties for QLIKE and squared loss under conditions including conditional unbiasedness of the proxy. These properties do not eliminate proxy noise or finite-sample uncertainty.

The framework of Diebold and Mariano [5] compares average predictive losses. The loss differential is defined in equation (11):

$$d_t = L(y_t, \widehat{h}_{A,t}) - L(y_t, \widehat{h}_{B,t}),\ H_0 : E(d_t) = 0 \tag{11}$$

A negative average differential favours model A. The heteroskedasticity and autocorrelation consistent covariance estimator of Newey and West [6] is used for two-sided asymptotic tests, with no prewhitening and a finite-sample adjustment. The lag rule $\left\lfloor 4(T/100)^{2/9} \right\rfloor$ gives six lags. Diebold–Mariano tests based on QLIKE loss differentials are reported as supplementary statistical evidence rather than as a definitive test of model superiority. Five forecasting methods generate ten pairwise QLIKE comparisons, and the corresponding p-values are adjusted using the Holm [7] procedure.

4. Empirical results and analysis

4.1. Descriptive statistics and preliminary evidence

Table 1 reports the distributional characteristics of returns. Training-sample mean daily return is 0.0336%, with a standard deviation of 1.7459%, a minimum of −13.3095%, and a maximum of

DOI: 10.54254/2754-1169/2026.AB39272

11.3640%. Variability is large relative to the mean. Skewness is −0.2890 and Pearson kurtosis is 9.0496, above the normal benchmark of three.

Table 1. Return statistics and preliminary tests

Statistic	Training	Test	Full
Observations	4160	939	5099
Mean	0.0336	0.0376	0.0344
Standard deviation	1.7459	1.4323	1.6924
Minimum	-13.3095	-8.4076	-13.3095
Maximum	11.3640	8.9718	11.3640
Skewness	-0.2890	0.2092	-0.2359
Pearson kurtosis	9.0496	5.9197	8.9203
JB statistic	6,401.5717	340.3853	7,493.9046
JB p-value	<0.001	<0.001	<0.001
ADF statistic	-14.9758	-9.9777	-16.4204
ADF lags	16	9	17
ADF p-value bound	≤0.01	≤0.01	≤0.01

Note: Returns are in percentage units. Pearson kurtosis has a normal benchmark of 3. ADF p-values are bounded by the software output; no p-value is interpreted as exactly zero.
Source: calculations using Yahoo Finance data.

Figure 1 shows clusters of large return movements separated by relatively quiet periods. This is consistent with volatility clustering, subject to formal examination using residual ARCH tests.

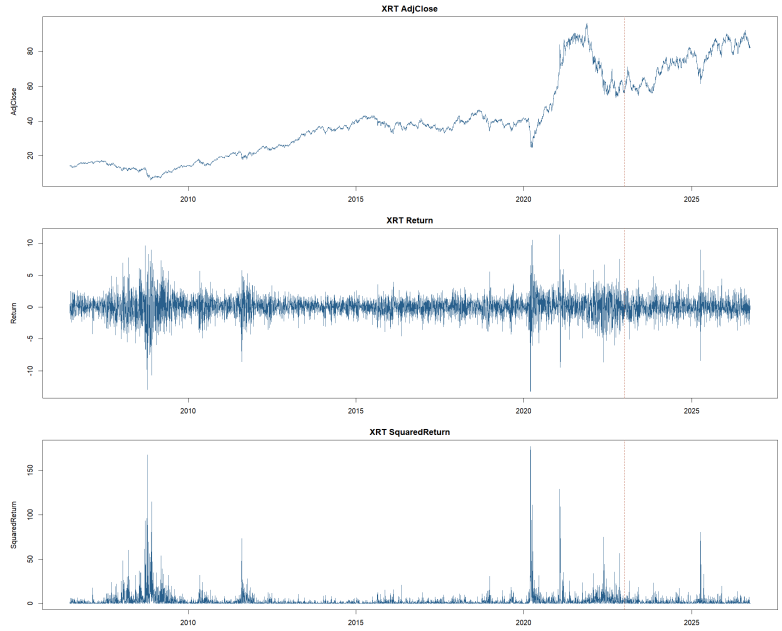

Figure 1. Adjusted prices, daily returns and squared returns

Source: author calculations using Yahoo Finance data. Panels show adjusted closing prices (USD), daily logarithmic returns (%) and squared returns (%²). The dashed line marks the start of the test period.

DOI: 10.54254/2754-1169/2026.AB39272

4.2. Mean equation estimation and residual diagnostics

All seven ARMA candidate models converge successfully. Table 2 identifies ARMA(0,1) as the BIC-selected specification, with a BIC of 16,457.7932 and an AIC of 16,438.7934, compared with a BIC of 16,457.8132 for ARMA(0,0). The BIC advantage over ARMA(0,0) is therefore only about 0.02, indicating weak evidence in favour of the MA(1) term. Although ARMA(1,1) has a lower AIC, its BIC is higher; because BIC is the predetermined primary criterion, ARMA(0,1) is retained as the baseline mean specification.

Table 2. ARMA selection and standalone coefficients

Model	AIC	BIC	Converged
ARMA(0,1)	16,438.7934	16,457.7932	Yes
ARMA(0,0)	16,445.1467	16,457.8132	Yes
ARMA(1,0)	16,439.0004	16,458.0002	Yes
ARMA(1,1)	16,437.6303	16,462.9634	Yes
ARMA(2,1)	16,439.5554	16,471.2218	Yes
ARMA(1,2)	16,439.5684	16,471.2348	Yes
ARMA(2,2)	16,441.5760	16,479.5756	Yes
Selected ARMA(0,1) coefficients			
ma1	-0.0454	0.0157	0.0038
intercept	0.0336	0.0258	0.1928

Note: These coefficients are from standalone ARMA estimation, not joint GARCH estimation.

The standalone MA coefficient is −0.0454, with p = 0.0038. The mean estimate is 0.0336, with p = 0.1928. The former suggests weak short-run negative dependence. These standalone estimates should not be confused with mean parameters from the subsequent joint GARCH estimations.

Table 3 reports residual diagnostic p-values for the selected ARMA(0,1) specification. The Ljung–Box p-values at lags 12 and 20 are approximately 0.000074 and 0.000094, rejecting the null of no residual autocorrelation at both lags. ARCH-LM statistics of 839.5543 and 936.4648, both with $p < 0.001$, provide strong evidence of conditional heteroskedasticity. Although ARMA(0,1) is retained as the parsimonious BIC-selected mean specification, it does not fully remove serial dependence from the conditional mean. This remaining autocorrelation motivates the ARMA(0,0) robustness exercise reported in Section 4.5.2.

Table 3. ARMA residual diagnostic p-values

Model	Lag	LB	LB squared	ARCH-LM
ARMA	12	<0.001	<0.001	<0.001
ARMA	20	<0.001	<0.001	<0.001

Note: ARMA diagnostics use ordinary residuals; GARCH diagnostics use standardized residuals. Conventional post-estimation tests are interpreted alongside native weighted diagnostics.

The squared-residual dependence in Figure 2 further indicates that a constant-variance ARMA specification is inadequate for the second-moment dynamics.

DOI: 10.54254/2754-1169/2026.AB39272

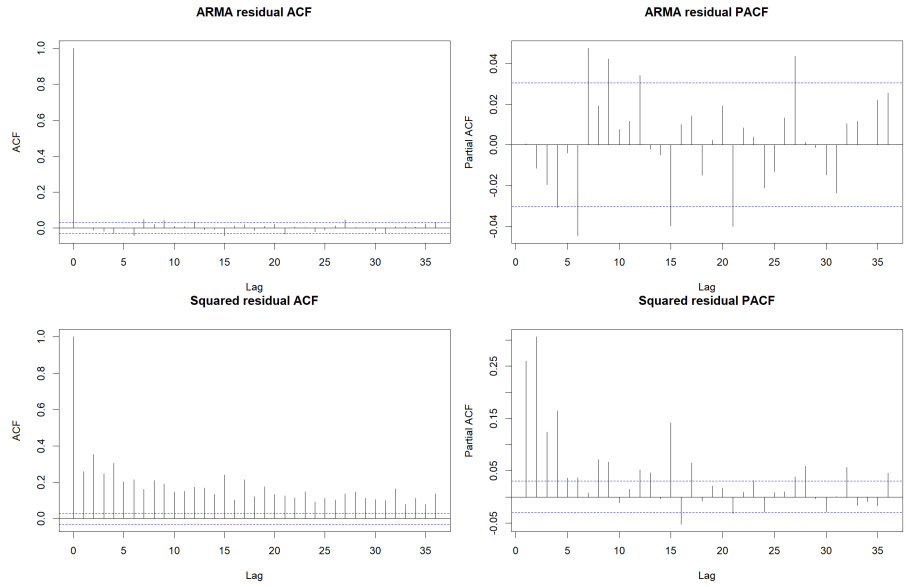

Figure 2. ARMA residual and squared-residual correlograms

Source: author calculations using Yahoo Finance data. The figure shows correlograms of the ARMA(0,1) residuals and squared residuals used to assess remaining serial dependence.

4.3. GARCH estimation and volatility characteristics

Table 4 reports joint estimates for all four specifications; this subsection first considers normal innovations. GARCH-Normal has $\alpha = 0.0912$ and $\beta = 0.8963$, implying persistence of 0.9875 and a model-implied volatility half-life of approximately 55 trading days. Shocks therefore decay slowly within this specification.

GJR-Normal has $\alpha = 0.0528$, $\beta = 0.8942$, and $\gamma = 0.0777$, with a robust asymmetry p-value of 0.0007. The coefficients associated with positive and negative innovations are 0.0528 and 0.1305, respectively. The coefficient associated with a negative innovation is approximately 2.47 times that associated with a positive innovation. Holding other conditions constant, a negative innovation of equal magnitude therefore produces a larger conditional variance increment. This is not a causal identification of the effects of particular news events.

Table 4. Joint parameter estimates

Parameter	GARCH-N	GJR-N	GARCH-t	GJR-t
mu	0.0606*** (0.0181)	0.0341* (0.0193)	0.0700*** (0.0172)	0.0472*** (0.0182)
ma1	-0.0163 (0.0156)	-0.0170 (0.0158)	-0.0227 (0.0148)	-0.0236 (0.0149)
omega	0.0352*** (0.0093)	0.0378*** (0.0092)	0.0272*** (0.0070)	0.0281*** (0.0072)
alpha1	0.0912*** (0.0172)	0.0528** (0.0219)	0.0830*** (0.0140)	0.0392** (0.0162)

DOI: 10.54254/2754-1169/2026.AB39272

Table 4. (continued)

beta1	0.8963*** (0.0174)	0.8942*** (0.0185)	0.9078*** (0.0138)	0.9099*** (0.0150)
gamma1	—	0.0777*** (0.0230)	—	0.0801*** (0.0194)
shape	—	—	11.8467*** (1.9163)	12.2806*** (2.2211)

Note: Robust standard errors in parentheses. ***, **, and * indicate robust p < 0.01, 0.05, and 0.10, respectively. N = normal; t = standardized Student-t. shape is the degrees-of-freedom estimate.

Table 5 summarizes model fit and persistence across the four specifications. GJR-Normal persistence is 0.9858, with a model-implied half-life of approximately 49 trading days. Its AIC and BIC are 14,840.8591 and 14,878.8587, below the corresponding GARCH-Normal values of 14,878.3662 and 14,910.0326. The asymmetric specification fits better after penalizing its additional parameter, although this does not establish superior out-of-sample accuracy.

Table 5. Model fit and persistence

Measure	GARCH-N	GJR-N	GARCH-t	GJR-t
Log-likelihood	-7,434.1831	-7,414.4296	-7,406.1949	-7,387.0442
AIC	14,878.3662	14,840.8591	14,824.3898	14,788.0885
BIC	14,910.0326	14,878.8587	14,862.3894	14,832.4214
Persistence	0.9875	0.9858	0.9908	0.9891
Half-life (days)	55.01	48.59	75.03	63.46

Note: All models use 4,160 training observations. AIC and BIC are totals; they are not the per-observation values shown in native rugarch reports.

Table 6 shows that conventional standardized-residual Ljung–Box and ARCH-LM tests at lags 12 and 20 do not reject their respective nulls at 5% for either normal specification. For example, lag-12 ARCH-LM p-values are 0.2621 and 0.4406. Nevertheless, some native weighted Ljung–Box tests at shorter lags remain significant. Overall, the GARCH-family models improve the representation of dynamic dependence, although some diagnostic concerns remain.

Table 6. Standardized GARCH residual diagnostic p-values

Model	Lag	LB	LB squared	ARCH-LM
GARCH-N	12	0.1922	0.2324	0.2621
GARCH-N	20	0.2915	0.4897	0.5142
GJR-N	12	0.1171	0.4242	0.4406
GJR-N	20	0.2217	0.6170	0.6459
GARCH-t	12	0.1803	0.0909	0.1092
GARCH-t	20	0.2676	0.2533	0.2795
GJR-t	12	0.0964	0.0120	0.0172
GJR-t	20	0.1895	0.0400	0.0597

DOI: 10.54254/2754-1169/2026.AB39272

Note: ARMA diagnostics use ordinary residuals; GARCH diagnostics use standardized residuals. Conventional post-estimation tests are interpreted alongside native weighted diagnostics.

Table 7 reports the native weighted diagnostics for the GJR-Student-t specification. Several weighted Ljung–Box tests are significant at the reported lags, whereas the weighted ARCH-LM tests are not. Taken together, the weighted diagnostics indicate that some residual dependence remains in the GJR-Student-t specification.

Table 7. Native weighted diagnostics for GJR-Student-t

Model	Test	Lag	Statistic	p-value
GJR-t	Weighted LB: z	1	0.4689	0.493491
GJR-t	Weighted LB: z	2	3.8945	0.005422
GJR-t	Weighted LB: z	5	6.2160	0.038652
GJR-t	Weighted LB: z^2	1	11.23	8.053e-04
GJR-t	Weighted LB: z^2	5	18.75	4.277e-05
GJR-t	Weighted LB: z^2	9	20.44	1.543e-04
GJR-t	Weighted ARCH-LM	3	1.108	0.2924
GJR-t	Weighted ARCH-LM	5	2.202	0.4283

Note: z denotes the standardized residual and z^2 its square. Values and lag settings are reproduced from the native rugarch reports. The weighted tests use different weighting and reference distributions from the conventional diagnostics and are therefore interpreted separately.

Figure 3 plots training-sample conditional standard deviations. Persistent episodes of elevated volatility are consistent with the estimated persistence parameters.

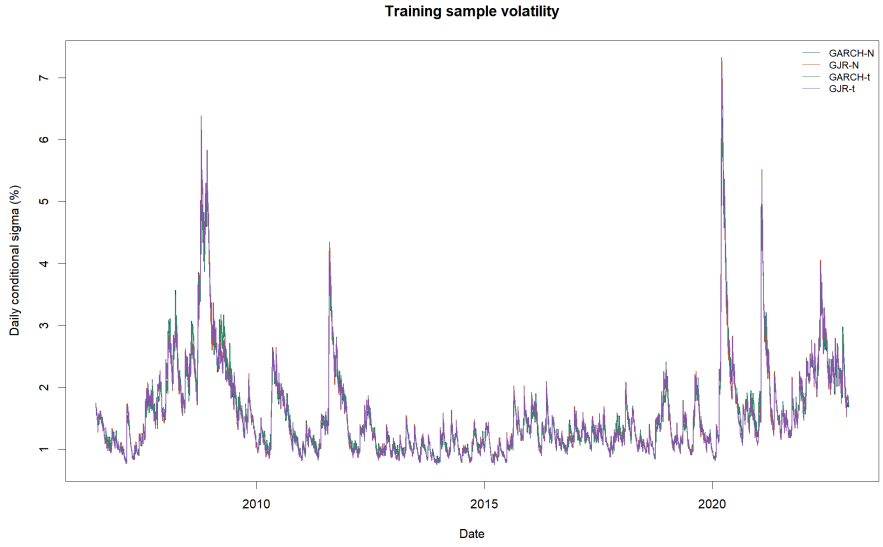

Figure 3. Training-sample conditional standard deviations

Source: author calculations using Yahoo Finance data. The lines show fitted daily conditional standard deviations (%) from the four training-sample ARMA(0,1)-GARCH specifications.

Figure 4 uses full-sample re-estimation solely for ex post description. The financial-crisis window covers 15 September 2008 to 9 March 2009, the COVID-19 window covers 20 February to

DOI: 10.54254/2754-1169/2026.AB39272

30 April 2020, and the 2022 window covers 3 January to 30 December. Volatility levels and patterns vary across these windows. The plots do not enter forecast evaluation or identify causal macroeconomic effects.

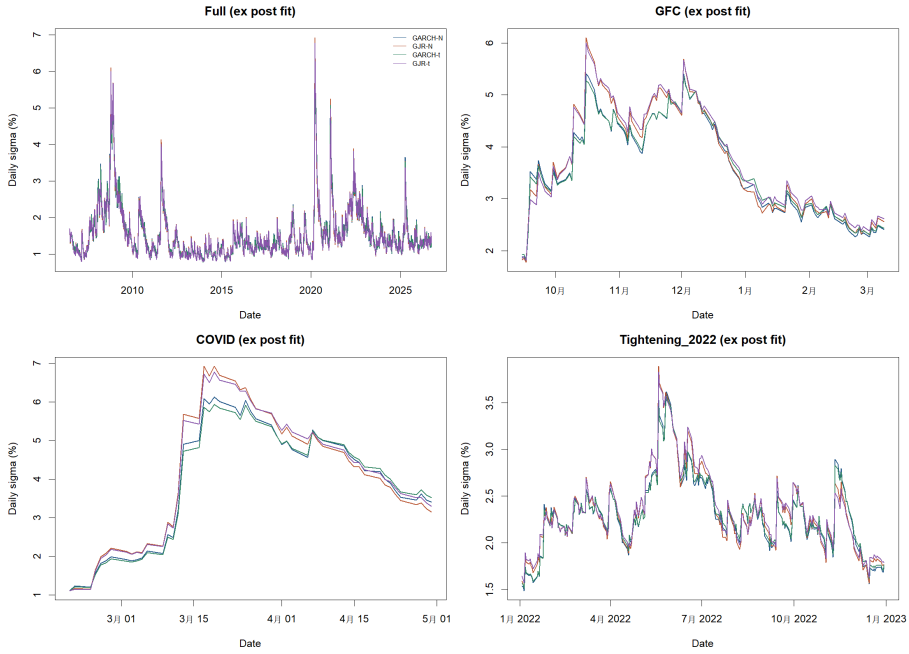

Figure 4. Ex post full-sample and crisis-window conditional standard deviations

Source: author calculations using Yahoo Finance data. Daily conditional standard deviations (%) are obtained by re-estimating the four models on the full sample. Panels cover the full sample, the financial crisis, COVID-19 and 2022. These are ex post descriptions, not out-of-sample forecasts.

4.4. Out-of-sample forecast performance

Table 8 reports losses over 939 common test dates. GARCH-Normal has QLIKE, RMSE, and MAE of 1.699152, 4.544614, and 2.240186. GJR-Normal has values of 1.697894, 4.506040, and 2.211188. HV22 has values of 1.723383, 4.592742, and 2.231532. The table also includes Student-t results, discussed in the following subsection.

Table 8. Out-of-sample forecast losses

Model	QLIKE	RMSE(%²)	MAE(%²)
GARCH-N	1.699152	4.544614	2.240186
GJR-N	1.697894	4.506040	2.211188
GARCH-t	1.697775	4.542244	2.237470
GJR-t	1.695478	4.499113	2.200817
HV22	1.723383	4.592742	2.231532

Note: All methods use 939 identical test dates. QLIKE is the primary criterion. RMSE and MAE are expressed in squared percentage units (%²), because both the volatility proxy and variance forecasts are measured in squared returns. QLIKE is unitless. Lower values indicate smaller forecast losses.

Both normal models have lower QLIKE and RMSE than HV22, but only GJR-Normal has a lower MAE. GARCH-Normal has slightly higher MAE than the benchmark. Relative performance

DOI: 10.54254/2754-1169/2026.AB39272

therefore depends on the loss function, and the results do not establish universal dominance by GARCH models.

GJR-Normal has lower losses than GARCH-Normal under all three measures, but numerical improvements are not necessarily significant. Table 9 gives unadjusted QLIKE p-values of 0.0643 and 0.0779 for the two normal models against HV22. Neither meets the 5% threshold. The normal-model comparison has p = 0.8244. These differences remain insignificant after Holm adjustment. The appropriate conclusion is numerical improvement under some criteria, with limited statistical evidence.

The DM tests provide supplementary evidence on whether the observed QLIKE loss differences are statistically distinguishable from zero.

Table 9. QLIKE comparisons with HAC standard errors

Model A	Model B	A minus B	Raw p	Holm p
GARCH-N	GJR-N	0.001259	0.8244	1.0000
GARCH-N	GARCH-t	0.001377	0.2621	1.0000
GARCH-N	GJR-t	0.003675	0.5772	1.0000
GARCH-N	HV22	-0.024230	0.0643	0.5144
GJR-N	GARCH-t	0.000118	0.9842	1.0000
GJR-N	GJR-t	0.002416	0.1989	1.0000
GJR-N	HV22	-0.025489	0.0779	0.5450
GARCH-t	GJR-t	0.002297	0.7269	1.0000
GARCH-t	HV22	-0.025608	0.0399	0.3985
GJR-t	HV22	-0.027905	0.0461	0.4152

Note: Negative loss differences favour A. Two-sided asymptotic tests use Newey–West standard errors with six lags. Holm adjustment covers all ten QLIKE comparisons.

Figure 5 compares the noisy squared-return proxy with daily one-step variance forecasts. Pointwise agreement is not expected because squared returns themselves contain substantial noise.

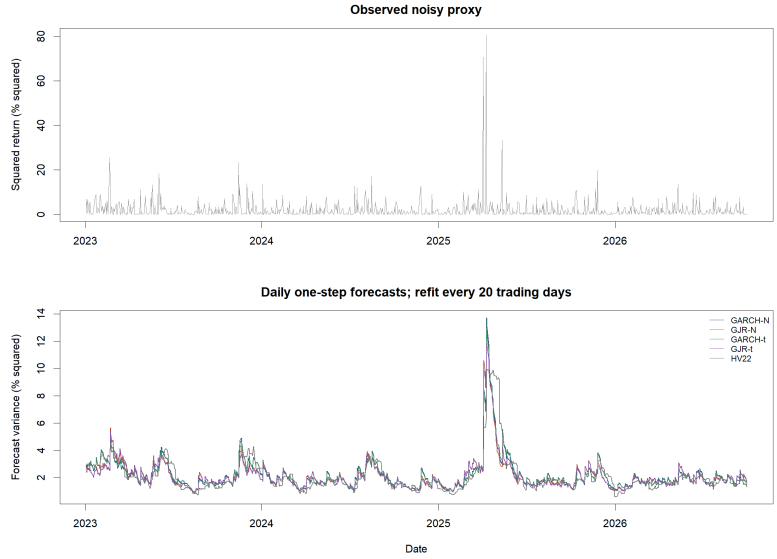

Figure 5. Out-of-sample squared returns and one-step variance forecasts

DOI: 10.54254/2754-1169/2026.AB39272

Source: author calculations using Yahoo Finance data. The figure compares test-period squared returns with daily one-step-ahead variance forecasts; both are expressed in squared percentage units (%²). Forecasts use the ARMA(0,1) mean specification with parameters re-estimated every 20 trading days.

Figure 6 shows cumulative QLIKE loss differentials relative to HV22. Values below zero indicate lower cumulative model loss through that date. The paths help identify periods of relative gains or losses but do not replace the DM tests.

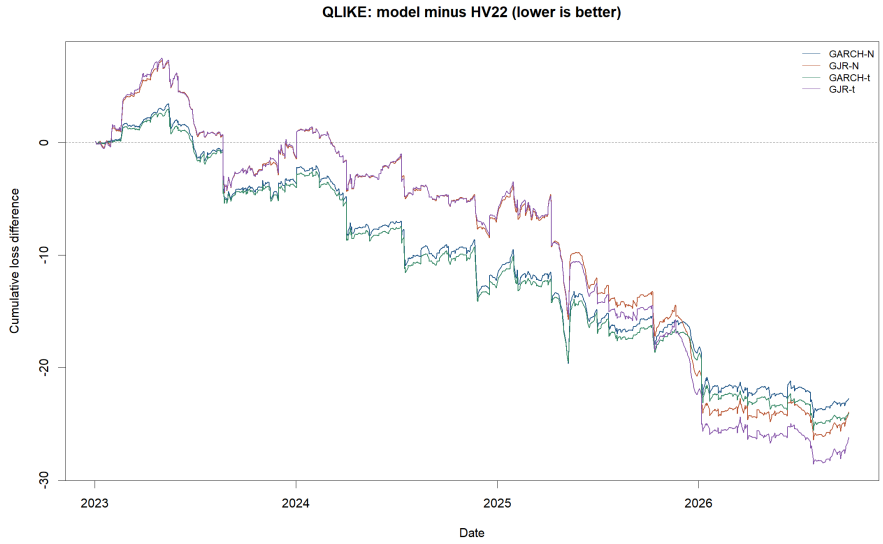

Figure 6. Cumulative QLIKE loss relative to HV22

Source: author calculations using Yahoo Finance data. Each line is the cumulative sum of daily QLIKE(model) minus QLIKE(HV22). The loss difference is unitless. Values below zero indicate lower cumulative QLIKE than HV22; these paths do not themselves establish statistical significance.

Overall, the loss measures indicate modest numerical improvements for the GARCH-family forecasts, particularly GJR-Student-t. However, none of the pairwise QLIKE differences remains statistically significant after Holm adjustment.

4.5. Robustness analysis

4.5.1. Robustness under Student-t innovations

The estimated Student-t degrees of freedom are 11.8467 for GARCH and 12.2806 for GJR, allowing heavier conditional tails than a normal distribution. Both Student-t specifications have lower AIC and BIC than their normal counterparts. GJR-Student-t has the lowest AIC and BIC, at 14,788.0885 and 14,832.4214, respectively.

Student-t GARCH and GJR persistence estimates are 0.9908 and 0.9891. The GJR asymmetry coefficient is 0.0801, with a robust p-value of approximately 0.000035. Its sign and significance agree with the normal specification, supporting robustness of the asymmetric response across the two innovation distributions.

GJR-Student-t achieves the lowest QLIKE, RMSE, and MAE among the five methods: 1.695478, 4.499113%², and 2.200817%². Relative to HV22, RMSE decreases by approximately 2.04% and MAE by 1.38%. GARCH-Student-t has slightly lower QLIKE than GJR-Normal but higher RMSE and MAE, showing that rankings still depend on the evaluation criterion.

DOI: 10.54254/2754-1169/2026.AB39272

The QLIKE comparisons of Student-t GARCH and GJR against HV22 have unadjusted p-values of 0.0399 and 0.0461. Although these meet the unadjusted 5% threshold, Holm-adjusted p-values are 0.3985 and 0.4152. The evidence therefore does not establish statistically significant superiority of the Student-t models over the HV22 benchmark after adjustment for multiple comparisons.

For GJR-Student-t, conventional squared standardized-residual Ljung–Box p-values are 0.0120 and 0.0400 at lags 12 and 20, while the corresponding ARCH-LM p-values are 0.0172 and 0.0597. The native weighted diagnostics in Table 7 provide mixed evidence: the weighted squared-residual Ljung–Box tests reject at the reported lags, whereas the weighted ARCH-LM tests do not. These results indicate that low predictive loss does not imply complete model adequacy, and some residual dependence remains.

The ratio of average squared conditional mean forecasts to average conditional variance forecasts ranges from approximately 0.085% to 0.255%. Evaluating $\widehat{h}_t + \widehat{\mu}_t^2$ against squared returns leaves rankings under all three measures unchanged; GJR-Student-t has QLIKE of 1.695465 under this alternative convention. The ordering is stable to the conditional-mean adjustment, although volatility-proxy noise remains.

The evidence supports substantial persistence and asymmetric responses within the estimated GJR models. GJR-Student-t has the lowest out-of-sample losses, but improvements are modest, no pairwise QLIKE comparison remains statistically significant after multiple-comparison adjustment, and some residual diagnostics indicate scope for further refinement.

4.5.2. Robustness to the mean specification

Given the weak BIC advantage of ARMA(0,1) and the remaining residual autocorrelation documented in Section 4.2, the forecasting exercise is repeated using ARMA(0,0) with an estimated constant mean. Table 10 reports the resulting forecast losses under the baseline ARMA(0,1) and alternative ARMA(0,0) mean specifications.

Table 10. Forecast losses under alternative mean specifications

Model	Mean model	QLIKE	RMSE (%²)	MAE (%²)
GARCH-N	ARMA(0,1)	1.699152	4.544614	2.240186
GARCH-N	ARMA(0,0)	1.699681	4.547142	2.243327
GJR-N	ARMA(0,1)	1.697894	4.506040	2.211188
GJR-N	ARMA(0,0)	1.698645	4.508406	2.214410
GARCH-t	ARMA(0,1)	1.697775	4.542244	2.237470
GARCH-t	ARMA(0,0)	1.698535	4.546019	2.241917
GJR-t	ARMA(0,1)	1.695478	4.499113	2.200817
GJR-t	ARMA(0,0)	1.696601	4.502449	2.205451
HV22	—	1.723383	4.592742	2.231532

Note: ARMA(0,0) includes an estimated constant mean. Each specification uses the same 939 test observations, expanding-window forecasting procedure, 20-trading-day refitting schedule, and daily one-step-ahead forecasts. RMSE and MAE are expressed in squared percentage units (%²), while QLIKE is unitless.

Replacing ARMA(0,1) with ARMA(0,0) produces only small changes in forecast losses. The ranking of the five forecasting methods remains unchanged under QLIKE, RMSE and MAE. GJR-Student-t continues to achieve the lowest loss among the competing models, with QLIKE changing

DOI: 10.54254/2754-1169/2026.AB39272

from 1.695478 to 1.696601, RMSE from 4.499113 to 4.502449 %², and MAE from 2.200817 to 2.205451 %². The main forecast ranking is therefore robust to this alternative mean specification. However, this robustness check does not establish that either mean specification fully removes residual dependence.

5. Conclusion

This paper compares four GARCH-family specifications for forecasting the volatility of the SPDR S&P Retail ETF (XRT). The comparison covers GARCH(1,1) and GJR-GARCH(1,1) models under normal and standardized Student-t innovations, with a 22-day historical-volatility measure used as the benchmark. Out-of-sample performance is evaluated using QLIKE, RMSE and MAE, while QLIKE-based Diebold–Mariano tests provide supplementary statistical evidence.

Among the models considered, GJR-GARCH with Student-t innovations gives the lowest QLIKE, RMSE and MAE. The estimated asymmetric coefficient is positive and statistically significant, indicating that negative innovations are associated with a larger increase in conditional variance than positive innovations of the same magnitude. Taken together, these results suggest that allowing for both asymmetry and heavy-tailed innovations is useful when modelling XRT volatility.

The forecasting advantage should nevertheless be interpreted cautiously. Differences in loss are small, and none of the pairwise QLIKE comparisons remains statistically significant after Holm adjustment. Some residual dependence also remains after model estimation. In addition, ARMA(0,1) is only marginally preferred to ARMA(0,0) by BIC, although replacing it with ARMA(0,0) does not change the forecast ranking. The evidence therefore supports a modest numerical advantage and a robust ranking, rather than clear statistical superiority.

These findings should also be considered in light of several limitations. Squared daily returns are only a noisy proxy for latent conditional variance, the analysis focuses on a single ETF, and the comparison is restricted to a relatively small set of GARCH-family models and one historical benchmark. Future work could use higher-frequency realised-volatility measures, extend the analysis to additional assets, and consider a broader range of volatility models. Within the present setting, GJR-GARCH with Student-t innovations delivers the strongest numerical forecasting performance, but the remaining diagnostic and statistical uncertainty calls for cautious interpretation.

References

[1] Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica, 50*(4), 987–1007. https://doi.org/10.2307/1912773

[2] Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics, 31*(3), 307–327. https://doi.org/10.1016/0304-4076(86)90063-1

[3] Glosten, L. R., Jagannathan, R., & Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *The Journal of Finance, 48*(5), 1779–1801. https://doi.org/10.1111/j.1540-6261.1993.tb05128.x

[4] Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics, 160*(1), 246–256. https://doi.org/10.1016/j.jeconom.2010.03.034

[5] Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics, 13*(3), 253–263. https://doi.org/10.1080/07350015.1995.10524599

[6] Newey, W. K., & West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica, 55*(3), 703–708. https://doi.org/10.2307/1913610

[7] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*(2), 65–70. https://www.jstor.org/stable/4615733

DOI: 10.54254/2754-1169/2026.AB39272

[8] Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

[9] Jarque, C. M., & Bera, A. K. (1987). A test for normality of observations and regression residuals. *International Statistical Review, 55*(2), 163–172. https://doi.org/10.2307/1403192

[10] Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association, 74*(366a), 427–431. https://doi.org/10.1080/01621459.1979.10482531

[11] Ljung, G. M., & Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika, 65*(2), 297–303. https://doi.org/10.1093/biomet/65.2.297

[12] Fisher, T. J., & Gallagher, C. M. (2012). New weighted portmanteau statistics for time series goodness of fit testing. *Journal of the American Statistical Association, 107*(498), 777–787. https://doi.org/10.1080/01621459.2012.688465