Modeling and Forecasting Gold Volatility

Cy COldiron

Abstract

This project investigates the underlying time-series dynamics of gold volatility. Using daily XAU/USD data from 1990 to the present, I identify volatility regimes, estimate competing ARCH models, and evaluate their out-of-sample forecast performance. My results suggest the mean of gold log returns is white noise while the highest performing variance forecasting model is a rolling EGARCH(1,1) with a regime dummy. My results suggest the presence of both volatility clustering (high variance is followed by high variance) and a reverse leverage effect: gold volatility responds more strongly to positive shocks than to negative ones. Considering the recent proliferation of gold prices, these findings are especially important for investors, portfolio managers, and central bankers in understanding the fundamental dynamics — and risks — posed by Gold.

1.1 Introduction

Since the end of the Bretton Woods system in 1971, Gold has played a markedly different role in global financial markets — going from a fixed-price and tied to the U.S dollar to a freely traded asset used to hedge uncertainty and inflation. With Gold prices rising over 50% YTD, and 130% since 2020 (greatly outperforming the S&P 500). Moreover, in a rapidly changing financial ecosystem characterized by de-globalization, tariffs, and rising geopolitical tensions — Gold is now at the forefront of growing discussions on the commodities role in portfolio allocation, risk mitigation, and central bank balance sheets. Understanding the statistical properties underlying this Gold's behavior is therefore both timely and important.

By investigating these properties, we hope to answer the following:

1. How has gold's volatility evolved across distinct historical regimes such as the Global Financial Crisis and the COVID-19 period?
2. Which EGARCH specification best captures and forecasts conditional variance?

3. What do the model parameters reveal about gold's sensitivity to positive versus negative price movements and its co-movement with uncertainty indices?

A core component of this analysis deals with the two concepts central to modeling volatility in financial time series: clustering and asymmetric leverage. Clustering is the tendency for periods of high variance to follow high variance, while asymmetric leverage refers to volatility reacting differently to positive and negative price shocks. Most financial time series data, such as stocks and equities, primarily exhibit a leverage effect — negative shocks raise volatility more than positive ones. While our results suggest the presence of volatility clustering, it also indicates a reverse leverage effect: gold volatility responds more strongly to positive shocks than to negative ones.

Grasping the nuances of these two characteristics—volatility clustering and asymmetry—is crucial for portfolio managers, investors, and central bankers. By integrating an understanding of volatility regimes and Gold's unusual sensitivity to positive shocks, these stakeholders can more effectively assess the risks, impacts, and opportunities this asset class presents for their organizations and decision-makers.

2 Data

2.1 Gold Data

We use daily gold price data (XAU/USD) obtained from Stooq, covering the period January 1990 to October 2025. All prices represent daily closing values quoted in U.S. dollars per troy ounce. As the data only contains official trading days (ie, weekends are not included), the dataset contains 9,191 price observations. This timeframe — which spans multiple macro financial cycles and crises — provides sufficient variation for both volatility and structural analysis.

2.2 Data Transformation

For our analysis, Gold prices were converted to log returns as

$$

r_t = 100 \times \big(\ln P_t - \ln P_{t-1}\big).

$$

where PtPt denotes the closing price at time tt.
Scaling by 100 yields approximate percentage log-returns, consistent with econometrics best practices.

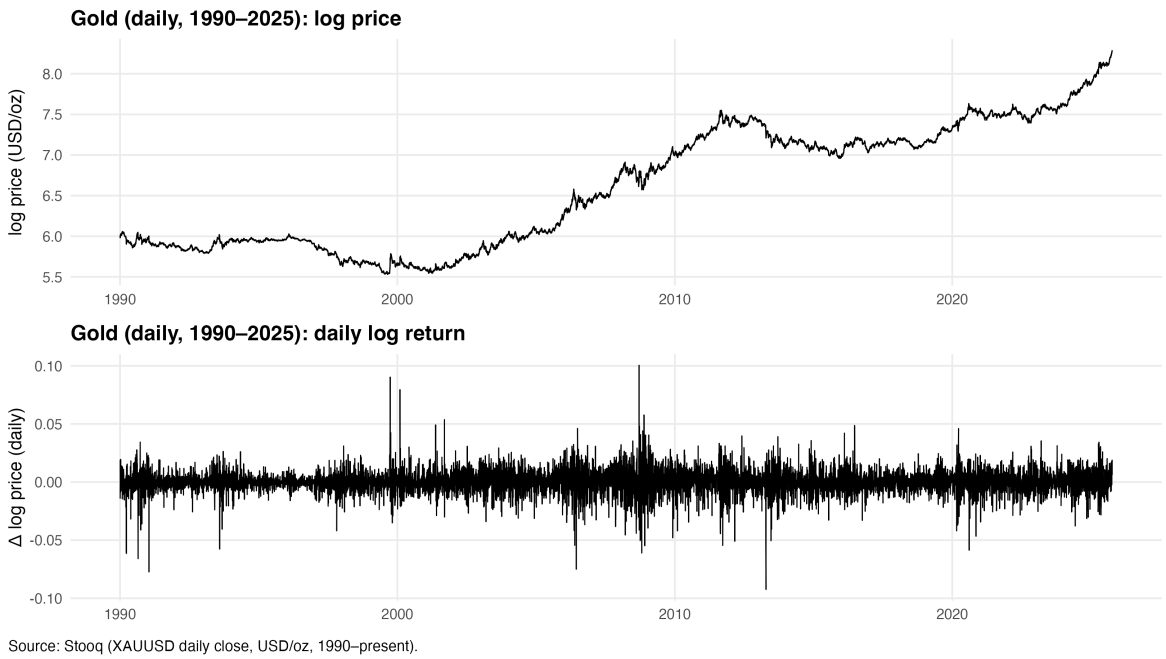

Figure 1: Gold time-series overview

Although explored in greater detail in following section — we transform prices to log returns to obtain an approximately stationary process. This allows us to use ARMA/GARCH methods for mean and variance modeling which would otherwise not be possible using prices which, as Figure 1 demonstrate, is clearly non stationary.

2.3 Other Data (Uncertainty Proxies)

For the later descriptive comparison (Section 9), we merge the daily gold return series with two external uncertainty measures:

(i) the CBOE Volatility Index (VIX), representing the S&P 500 30-day expected volatility (Data source: FRED).

(ii) the Economic Policy Uncertainty (EPU) developed by Baker, Bloom, and Davis (2016), quantifies policy-related economic uncertainty based on newspaper coverage, tax code provisions, and forecaster disagreement. (Data source: Economic Policy Uncertainty project.)

Both of these indices serve as external benchmarks for periods of elevated uncertainty and are not used in our primary model estimation.

3 Methodology

Throughout this paper, to identify the most appropriate model we make general use of the Box and Jenkins (1979) method. This framework for model identification consists three stages:

(1) Identification of an appropriate functional form for the model.

(2) Estimation of the model parameters (typically using MLE or OLS).

(3) Diagnostic checking to assess the fit of the model (i.e, checking residuals autocorrelation)

Our framework will also employ various tests — primarily Diebold and Mariano (1995) — to test the predictive capability of various models.

3.1 Model Identification

3.1.1 ACF / PACF plots

To confirm our beliefs regarding the underlying data generating processes for both log prices and returns — we make use of the autocorrelation and partial autocorrelation function.

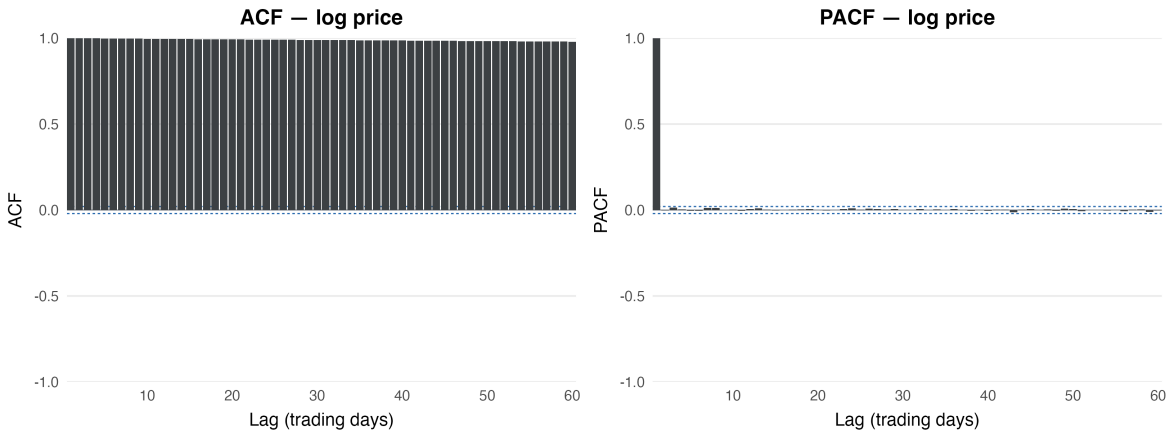

Figure 2: ACF / PACF charts for log prices

Here, the autocorrelation function measures the linear relationship between gold prices and returns and lagged values versions of itself measured by trading days. Figure 2 demonstrates a slowly decaying ACF and PACF with a sharp cut-off at lag 1. In other words, gold price today is highly correlated to values in the past. Yet, once you condition for intermediate values (which the PACF does), there is no longer a significant relationship left. Taken together, these results suggest that log prices are a random walk with a drift.

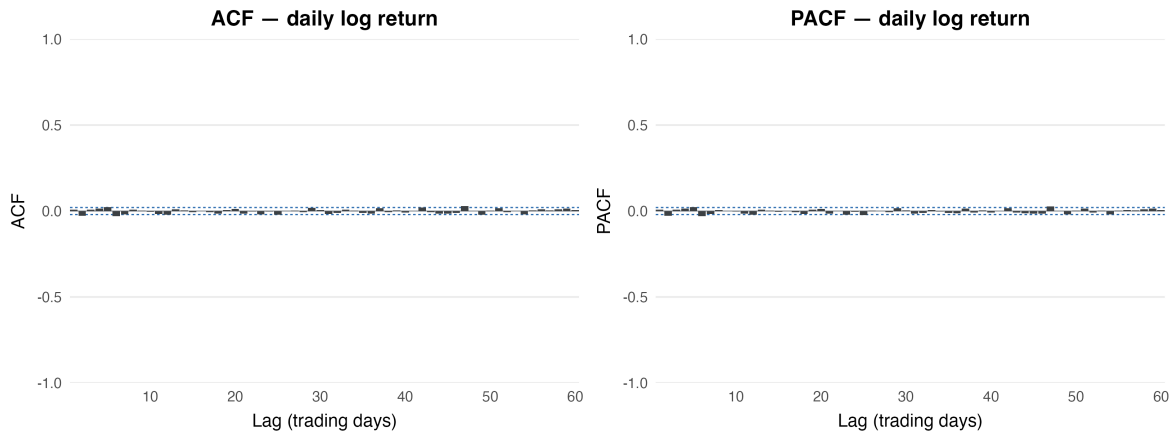

Figure 3: ACF / PACF charts for log returns

In contrast, Figure 3 highlights no significant autocorrelation: the rate of return on gold today is uncorrelated to previous days return values. Such results suggest that the underlying data generating process is gausian white noise.

3.1.2 Stationary — Augmented Dickey Fuller Test

While the ACF plots suggest the presence of a unit root (non-stationary) in log prices and a white noise process (stationary) in log returns, it is useful to formally test this assumption.

To do so we make use of the Augmented Dickey Fuller (ADF) test.

Let the ADF regression be

$$\Delta y_t \;=\; \alpha \;+\; \beta t \;+\; \gamma\, y_{t-1} \;+\; \sum_{i=1}^{p} \delta_i\, \Delta y_{t-i} \;+\; \varepsilon_t.$$

Null and alternative hypotheses

$$\begin{aligned} H_0 &: \; \gamma = 0 \quad \text{(unit root; non-stationary)} \\ H_1 &: \; \gamma < 0 \quad \text{(stationary)} \end{aligned}$$

(The deterministic terms α and βt may be included or excluded depending on the test variant—none, drift, or trend—but the hypothesis is always on γ.)

Because of the likely persistence in both processes (i.e, autocorrelation in the residuals) — we employ the augmented test (Dickey & Fuller, 1981) which incorporates lagged versions of delta yt. The number of included lags is chosen based on information criteria (AIC)

Stationarity summary — **Gold prices vs returns**[1]				
	Deterministic	*ADF stat*	Lag k	Decision @ 5%
Gold: Daily Log Return				
Jan 1990 – Oct 2025	Drift	p=0.01	20.00	Reject unit root
Jan 2010 – Oct 2025	Drift	p=0.01	15.00	Reject unit root
Jan 2018 – Oct 2025	Drift	p=0.01	12.00	Reject unit root
Oct 2024 – Oct 2025	Drift	p=0.01	6.00	Reject unit root
Gold: Log Price				
Jan 1990 – Oct 2025	Trend	p=0.69	20.00	Fail to reject
Jan 2010 – Oct 2025	Trend	p=0.99	15.00	Fail to reject
Jan 2018 – Oct 2025	Trend	p=0.99	12.00	Fail to reject
Oct 2024 – Oct 2025	Trend	p=0.62	6.00	Fail to reject

[1] Stat shows **τ** (urca) or **p-value** (tseries) when urca fails. Lags: Schwert cap with AIC for urca; k≈T^(1/3) for tseries.

Figure 4: Stationarity Tests: Augmented Dickey Fuller

Figure 4 shows the ADF test statistics, along with corresponding p-values, for gold log prices and returns in four time regimes encompassing thirty-five, fifteen year, seven, and one year. As expected, the optimal number of autoregressive lags (k) is non-zero suggesting the presence of autocorrelation in the residuals and confirming the appropriateness of an Augmented Dickey Fuller Test.

Figure 4 suggests the presence of a unit root as demonstrated by consistently large p-values, suggesting a random walk process. In other words, prices are a function of the previous days value plus a shock. Unsuprinsgly, the small p-values for log returns indicate a stationary process.

3.1.3 - White Noise Test

Given the low degree of persistence in log returns — as shown in Figure 3 — Box and Jenkins suggest using Q-statistics to test if a group of autocorrelations is different from zero (i.e, process is not white noise).

White-noise tests (daily log returns)

		Ljung–Box Q statistic			p-value		
	N obs.	Stat (L=10)	Stat (L=20)	Stat (L=40)	Pvalue (L=10)	Pvalue (L=20)	Pvalue (L=40)
1990-Pres	9190	27.2	39.9	64.9	0.002	0.005	0.008
2010-Pres	4069	3.6	14.1	28.4	0.965	0.824	0.916
2018-Pres	2005	10.1	24.5	45.3	0.431	0.222	0.260

H0: no autocorrelation up to lag L. Cells with p < 0.05 reject H0.

Figure 5: White Noise Test

The Ljung–Box Q-statistics in Figure 5 test the joint null hypothesis that autocorrelation up to lag L is zero (i.e, process is white noise). For the full 1990–present sample, the null is rejected at the 1% level across all lag lengths (L = 10, 20, 40). However, Box and Jenkins (2016) note that in large samples — such as ours (>9 000 obs) — the Q-statistic can reject the null for trivially small correlations. Knowing this, we included smaller sample sizes to see if the detected autocorrelation in the 1990 sample can be attributable to its large sample size or sysematic autocorrelation. As shown in rows two and three, once the sample is restricted to post-2010 or post-2018 periods, p-values grow substantially, indicating no significant autocorrelation. In other words, it is likely that the mild autocorrelation seen in the full sample is due to the large sample size — not a consistent linear relationship in residuals. This result is consistent with prior evidence that asset returns are approximately white noise in the mean and that most dependence — if any — arises from conditional heteroskedasticity (Tsay, 2010).

3.2 Model Estimation

While our Q-statistics suggest that returns are approximately white noise in the mean, we want to formalize this assumption by testing the goodness-of-fit of a white noise process compared to other candidate models.

Mean Model Comparison — Daily Gold Log Returns

N = 9190

	AIC	BIC
ARMA(0,0)	−58,953.34	**−58,939.09**
AR(1)	−58,952.08	−58,930.70
MA(1)	−58,952.12	−58,930.74
ARMA(1,1)	**−58,954.88**	−58,926.37

Source: Stooq (XAUUSD).

Figure 6: Mean Model Testing

Figure 6 compares simple mean specifications using information criteria that balances goodness-of-fit with complexity (i.e, penalty for added parameters). AIC slightly prefers ARMA(1,1) wihile BIC (which penalizes complexity more heavily) suggests the white noise mean as the best model.

Likelihood Ratio Test — ARMA(1,1) vs ARMA(0,0)

N = 9190 | H_0: extra parameters = 0

Model 1	Model 0	LR ($2\Delta LL$)	df	p-value	Decision
ARMA(1,1)	ARMA(0,0)	5.53	2	0.063	**Fail to reject H_0 — No significant improvement**

Gaussian ML; χ^2(df) reference.

Figure 7: LR Test: ARMA(1,1) v.s White Noise

Thus, we use a Liklihood Ratio test of ARMA(1,1) v.s ARMA(0,0) determine if the extra

AR and MA terms add explanatory power. Figure 7 suggests that the added parameters are insignificant, providing sufficient evidence to treat gold returns as white noise in the mean.

3.3 Diagnostics

Because the mean process is modeled as white noise, the Ljung–Box Q-tests in Figure 5 effectively serve as diagnostic checks on the residuals. The results therefore confirm that the mean specification is appropriate, and any remaining dependence likely arises from conditional heteroskedasticity — which we will explore in section XYZ — rather than serial correlation in the mean.

4 Structural Breaks (Variance Focus)

Having specified the mean of gold returns as white noise, we now want to ensure a consistent data generating process over our specified time period. To check this assumption, we make use of Bai-Perron (1998) structural break tests — a procedure that finds the optimal quantity and location of all possible breakpoints in a time series.

Bai–Perron mean-break test on gold log returns

Windows as defined in outline. **h** = minimal segment length (observations) used in the search.

Window	n (obs)	h (min seg, obs)	supF p	Num. breaks	Decision
1990–Pres	9,190	1,531	0.087	0	no break
2010–Pres	4,069	678	0.078	0	no break
2018–Pres	2,005	334	0.071	0	no break

Num. breaks selected by BIC.

Source: Stooq (XAUUSD daily close, USD/oz). Bai–Perron via strucchange.

Figure 8: Structural Break Test on Mean of Gold Returns

Figure 8 shows Bai–Perron multiple mean-break tests for gold log returns across three sample windows (1990–present, 2010–present, and 2018–present). The large p-values (0.07–0.09) esuggest zero breaks, indicating that the mean of gold returns is stable over time—consistent with returns behaving as a stationary, white-noise process.

Having established the mean as both a white-noise process with no structural breaks — our attention will now shift towards exclusively to the variance, and thus volatility, of Gold returns. To identify variance regimes — periods of significant change in the underlying data generating process for the second moment — we employ a multiple change point detection approach using Pruned Exact Linear Time (PELT). This approach — proposed by Prop R. Killick, Fearnhead,

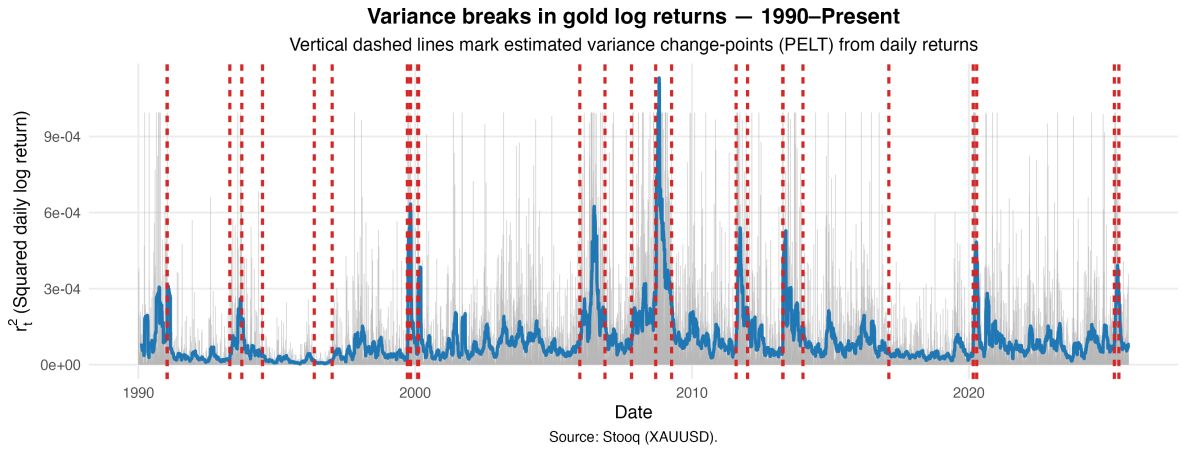

Figure 9: PELT Changepoint Tests

and Eckley (2012) —uses a dynamic programming algorthim to find the optimal segmentation for all admissible changepoints.

Figure 9 plots the variance of gold from 1990 to the present, with red dashed lines marking estimated variance change-points identified by the PELT algorithm. Because our process is white noise (i.e, E[yt] = 0) the variance (y-axis) is equal to the squared daily return. The data indicate twenty-seven distinct variance regimes over the twenty–five year period and illustrates a clear relationship between financial uncertainty and structural breaks. Notable economic shocks — such as the dot-com bubble (2001), great recession (2007-09) and COVID-19 (2019-20) — are all associated with step-changes in gold's variance. This is consistent with prior literature which shows volatility of gold prices tends to intensify during global crises such as the 1987 stock market crash and the COVID-19 pandemic (Lamouchi & Badkook, 2020)

Moreover, periods of relative economic certainty and sustained growth — such as the great moderation (1980-2007) and great expansion (2009-2020) — are linked to lower volatility and fewer structural breaks. Volatility has been relatively nascent since COVID-19 — experiencing a five-year uninterrupted regime broken by the surge in volatility (and nominal price) in mid-2025, likely driven rising trade tensions, falling consumer sentiment and rising inflation expectations (University of Michigan, Surveys of Consumers, 2025).

Figure 9 confirms two important points. First, the relationship between economic uncertainty with both distinct, and elevated, volatility regimes. Second, the number and spacing of breaks implies that a single, static variance model is unlikely to be stable over decades. In other words, gold returns exhibit heteroskedasticity (volatility clustering), that must be incorporated properly using a more complex variance model.

5 Constructing a Volatility Model

The number and spacing of breaks — as seen in Figure 9 — implies that a single, static variance model is unlikely to be stable over decades. In other words, gold returns likely exhibit heteroskedasticity (volatility clustering), that must be incorporated properly using a variance model that accounts for such patterns.

To build a parsimonious volatility model, we follow four standard steps in volatility modeling (Kotzé, 2025):

1. Specify a mean equation after testing for serial dependence in the data.
2. Make use of the residuals from the mean equation to test for ARCH eects.
3. Specify a volatility equation. If ARCH effects are statistically significant then perform a joint estimation of the mean and volatility equations.
4. Check the fitted model carefully and refine it where necessary.

5.1 Testing for ARCH effects (Heteroskedasticty)

Having already specified the mean equation (white-noise log returns), we next test whether the variance of gold returns is conditionally heteroskedastic. Put simply, is volatility today a function of its past self? Such a relationship would imply time-varying volatility rather than constant variance.

To test for volatility clustering we examine whether the squared residuals exhibit serial correlation using the ARCH-LM test (Engle 1982).

Given the following linear regression, which regresses the current squared residual on lagged versions of itself

$$a_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + \alpha_2 a_{t-2}^2 + \cdots + \alpha_m a_{t-m}^2 + \varepsilon_t$$

We test the null hypothesis of homoskedasticty:

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_m = 0 \quad \text{(no ARCH effects; homoskedastic variance)}$$

Rejection of

$$H_0$$

implies the presence of some degree of conditional heteroskedastity (ie, today's variance is correlated to lagged versions of itself)

Engle ARCH–LM Test (Returns)

H_0: no ARCH effects

Lags	F-Statistic	p-value
6	405.548	< 0.001
12	482.653	< 0.001
24	619.293	< 0.001

Figure 10: ARCH-LM test for conditional heteroskedasticity

Figure 10 reports the F-statistics and p-values for lags 6, 12, and 24. The small p-values across all lags suggest the presence of homoskedasticty, providing strong evidence of volatility clustering in gold returns— i.e, large shocks tend to be followed by large shocks.

Figure 11 visualizes this dependency — plotting the ACF of squared returns across various lags (trading days). The consistently significant autocorrealations confirms the need to fit an ARCH-type volatility model — which incorporates the conditional dependency in the second moment.

5.2 Specifying an ARCH Model

Given the presence of volatility clustering we proceed with to identifying the most parsimonious conditional variance model.

To identify the degree of persistence (ARCH order) in potential candidate models, we make use of the PACF of squared returns.

Figure 12 shows a moderately decaying PACF, with significant autocorrelations up to lag 10. Nonetheless, the sharp initial spike and gradual decay suggest that a lower order ARCH process (q = 1-3) provides a constructive starting point. If higher order processes are deemed more parsimonious than further exploration of higher-order processes may be necessary.

For our preliminary model testing, we will consider five iterations of conditional heteroskedastic (ARCH) models:

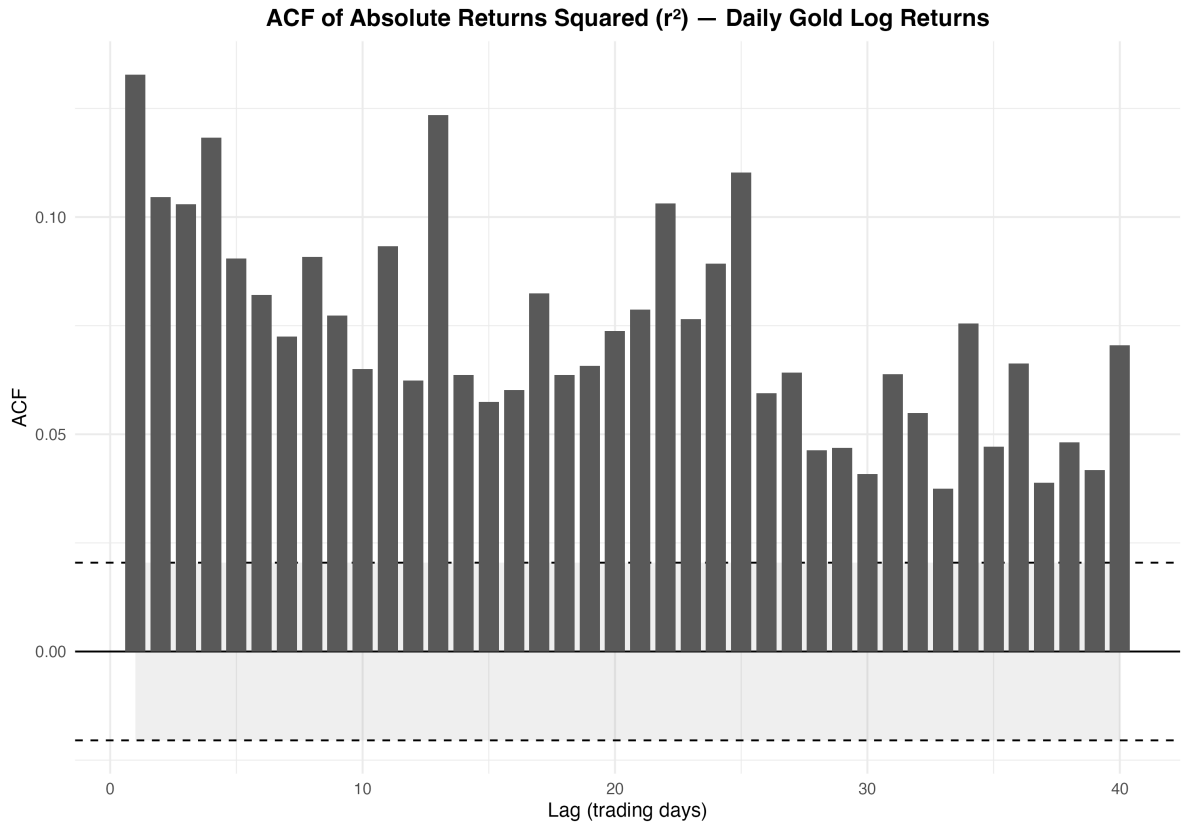

Figure 11: ACF of Squared Returns

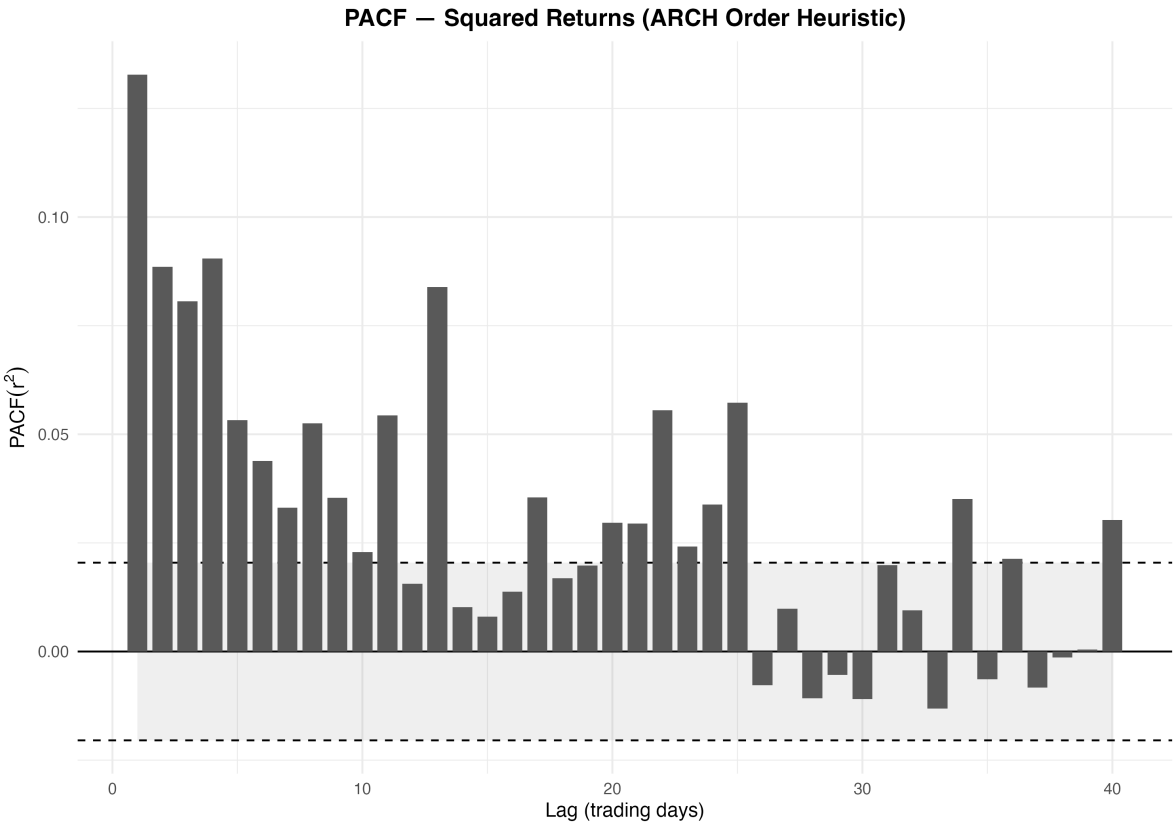

Figure 12: PACF of Squared Returns

1. ARCH(q): Volatility today is a weighted sum of the last q squared shocks.
2. GARCH(p,q): Adds lagged variance terms to ARCH (longer memory in volatility).
3. IGARCH(1,1):GARCH process with a unit root (shocks have permanent effects).
4. GJR-GARCH(1,1): GARCH prcoess with a asymmetry term to allow for uneven volatility responses to positive or negative shocks.
5. EGARCH(1,1): GARCH process that models the log variance, while having an asymmetry term.

Variance Model Comparison — Baselines & ARCH/GARCH Family

	AIC	BIC	Converged
ARCH(1) — t	-60897.83	-60869.32	✓
ARCH(2) — t	-61119.59	-61083.96	✓
ARCH(3) — t	-61268.98	-61226.23	✓
GARCH(1,1) — t	-61995.54	-61959.91	✓
GARCH(1,1) — skew-t	-61997.16	-61954.40	✓
GARCH(1,2) — t	-61993.41	-61950.65	✓
GARCH(2,1) — t	-61994.89	-61952.13	✓
IGARCH(1,1) — t	-13615.61	-13579.98	✓
GJR(1,1) — t	-62037.06	-61994.30	✓
GJR(1,1) — skew-t	-62038.01	-61988.13	✓
EGARCH(1,1) — t	-62064.55	**-62021.80**	✓
EGARCH(1,1) — skew-t	-62066.34	-62016.46	✓

Figure 13: Baseline Variance Model Testing

Figure 13 shows the information criteria of various candidate models suggesting that EGARCH(1,1) -t is the best fitting specification.

To understand the significance — or broader implications — of an exponential GARCH model (EGARCH) being the best performing model, we need to examine its two key attributes

1) Guaranteed Positivity

- By modeling log variance, we ensure that the variance is always positive.

2) Asymmetric responses

- The gamma term allows for variance to react differently to positive and negative shocks.

In other words, EGARCH predicts how log variance evolves over time — making it smoother (less-sensitive to shocks), always positive, and able to capture the fact that volatility often reacts differently to positive and negative shocks.

EGARCH(1,1)-t Model — Estimated Parameters

Robust standard errors and significance

Parameter	Estimate	Robust SE	t-value	p-value
μ	0.0002	0.0001	3.03	0.00245
ω	−0.0659	0.0021	−31.82	0.00000
α_1	0.0368	0.0065	5.63	0.00000
β_1	0.9930	0.0002	5,470.78	0.00000
γ_1	0.1157	0.0040	28.95	0.00000
ν (shape)	4.3946	0.2321	18.93	0.00000

Figure 14: Baseline Model — Parameter Summary

?@fig-egarch-paramsn shows the maximum likelihood estimates for the six models parameters — all of which are statistically significant. The average return of gold (mu) shows a tiny coefficient (0.0002) — implying that gold returns geometrically compounds (roughly 5% a year). It also indicates that our white noise mean model has a minuscule positive drift.

The long-run level (omega) highlights the baseline level of volatility. Its negative value implies, without shocks, log variance tends to be below zero and relatively moderate. Moreover, the short-run reaction coeffient (alpha 1 = 0.368) — also described as the ARCH effect — shows that conditional volatility changes moderately to new shocks.

The GARCH effect, (B1 = 0.993), indicates that volatility is highly persistent. In other words, high volatility yesterday tends to be followed by high volatility today.

The leverage effect (gamma) measures the degree to which volatility reacts differently to positive v.s negative returns. The positive value (gamma = .12) indicates that positive shocks (ie, higher gold returns) are associated with greater changes in volatility than negative ones. This dynamic — which can be characterized as a "reverse leverage effect" — is antithetical to most equities which tend to exhibit greater volatility in response to negative returns. Such an effect suggests gold's volatility rises during bullish periods — consistent with the narrative of being a safe-haven asset during periods of economic uncertainty. This is an important finding that we will explore in more depth in SECTION XYZ.

Finally, the shape parameter f the Student-t distribution (v =4.39) suggests the presence of fat tails. In other words, significant deviations are far more pronounced, and likely, than they would be under a normal distribution. Such an effect illustrates that Gold has a tendency for large irregular swings in its volatility — likely due to its correlation with macro and geopolitical shocks.

5.3 Diagnostics on Candidate Model

Information criteria in the preceding section suggests (EGARCH(1,1) is the best performing candidate model. However, conventional volatility modeling framework (Kotzé, 2025) suggests performing diagnostic checks on the residuals.

To do so, we make use of two tests:

1. Ljung–Box Q-test on squared standardized residuals(H : no autocorrelation in r^2—i.e., no ARCH effect.
2. EngleARCH–LM test on standardized residuals (H : no ARCH effects up to lag m).

Large test statistics (small p-values) would suggest that our model is insufficient in describing ARCH effects: ie, there is leftover conditional dependence in the variance.

To do so

Residual Diagnostics — EGARCH(1,1), Student-t				
Test	df / Lags	Statistic	p-value	max |ACF(r²)| (1–5)
Full sample				
LB on squared standardized residuals	10	66.586	< 0.001	0.076
ARCH–LM on standardized residuals	12	67.680	< 0.001	
Post-2010				
LB on squared standardized residuals	10	49.290	< 0.001	0.103
ARCH–LM on standardized residuals	12	48.885	< 0.001	
Post-2018				
LB on squared standardized residuals	10	8.356	0.594	0.045
ARCH–LM on standardized residuals	12	7.772	0.803	

Figure 15: EGARCH(1,1) - T: Diagnostics

5.4 Refining Model

5 Evidence of Conditional Heteroskedastcity

6 Variance Modeling

7 Break-AwareVariance Modeling

8 Forecast Design & Evaluation