

Time Series Analysis of Google Realized Volatility

February 2026

Abstract

Stock prices are driven by a mix of economic fundamentals, investor sentiment, and external shocks, making them inherently difficult to predict. While forecasting price direction remains elusive, volatility exhibits strong temporal persistence that time series models can exploit. We study the weekly log realized volatility (Log RV) of Google (GOOG) from 2019 to 2025 using ARIMA and SARIMAX frameworks. An extensive AIC-based grid search identifies ARIMA(3,0,3) as the best univariate model. Incorporating the VIX index as an exogenous regressor yields SARIMAX(3,0,3)(0,0,1)₄ with substantially lower AIC. One-step-ahead rolling forecasts on a held-out test set confirm that the SARIMAX model outperforms the ARIMA baseline on RMSE, MAE, and MAPE, demonstrating that market-wide fear captured by VIX provides genuine predictive power for individual stock volatility.

1 Introduction

Every trading day, financial markets produce a stream of prices, volumes, and derived indicators that unfold in sequence. Because this data is naturally ordered by time, time series analysis offers a natural lens for studying it. ARIMA models, for instance, describe how a variable's recent values and recent forecast errors carry forward into the future, while SARIMAX models build on this by allowing for seasonal rhythms and the influence of outside variables [1; 2]. These frameworks fit financial data well because many market quantities tend to be persistent and mean-reverting, patterns that methods ignoring temporal order would overlook.

Predicting whether a stock will go up or down is famously difficult. The efficient market hypothesis holds that prices already reflect available information, so returns behave roughly like a random walk [3]. Volatility tells a different story. It clusters, where calm periods tend to follow calm periods, and turbulent ones tend to persist. A trader who can anticipate a spike in volatility can hedge a portfolio with options, whose prices are tied directly to expected volatility, or scale down positions before the storm hits. Conversely, when implied volatility rose above what actually realized, strategies like selling straddles can profit from the gap. Realized volatility (RV), computed from observed returns, measures how much a stock actually moved, and forecasting it well is critical to option pricing, risk budgeting, and portfolio construction.

We focus on Google (Alphabet, ticker GOOG), one of the most liquid large-cap equities globally. Its volatility is shaped by both company-specific events like earnings reports and regulatory actions and by broader market forces. The question we set out to answer is: **can market-wide volatility, as measured by the VIX index, improve forecasts of Google's realized volatility beyond what the stock's own history provides?**

The VIX is often called the “investor fear gauge” because it captures the market’s expectation of near-term S&P 500 volatility as implied by option prices [4]. Prior work has shown that realized volatility exhibits long memory and strong persistence [5; 6], which motivates fitting ARMA-type models to log-transformed RV. We take this a step further by adding VIX as an exogenous regressor within a SARIMAX framework and testing whether it actually improves out-of-sample forecasts.

Our data runs from January 2019 through December 2025, a window that covers the COVID crash, the post-pandemic recovery, and the 2022 rate-hike cycle, giving us a wide range of volatility regimes to work with.

2 Data and Preprocessing

2.1 Realized volatility construction

We download daily GOOG closing prices via the `yfinance` Python API and compute daily log returns $r_t = \log(P_t/P_{t-1})$. Realized volatility over a 5-day (weekly) window is given by

$$RV_t = \hat{\sigma}_t \sqrt{252}, \quad \hat{\sigma}_t = \text{std}(r_{t-4}, \dots, r_t),$$

where the $\sqrt{252}$ annualizes the estimate. We log-transform to obtain $\text{Log } RV_t = \log(RV_t)$, which produces a more symmetric, approximately stationary series [5].

To avoid mechanical autocorrelation from overlapping windows, we retain every 5th observation, yielding 351 non-overlapping weekly Log RV values.

2.2 Exogenous variables

We also download the VIX index and extract GOOG daily trading volume over the same period. After merging onto the non-overlapping dates (forward-filling VIX for any missing days), each observation carries three predictors: lagged Log RV values, the VIX level, and trading volume.

Figure 1 shows the resulting Log RV series. The COVID spike in early 2020 is the dominant feature, followed by elevated volatility during the 2022 tightening cycle. Figure 2 displays the VIX and volume series for context.

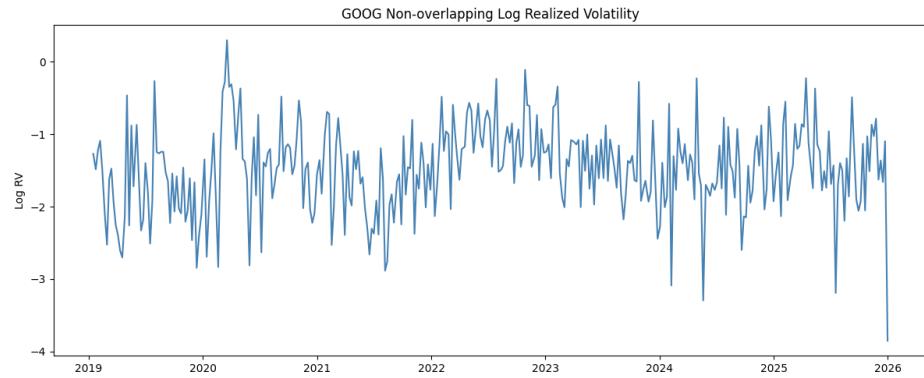


Figure 1: Non-overlapping weekly Log Realized Volatility of GOOG, January 2019 to December 2025.

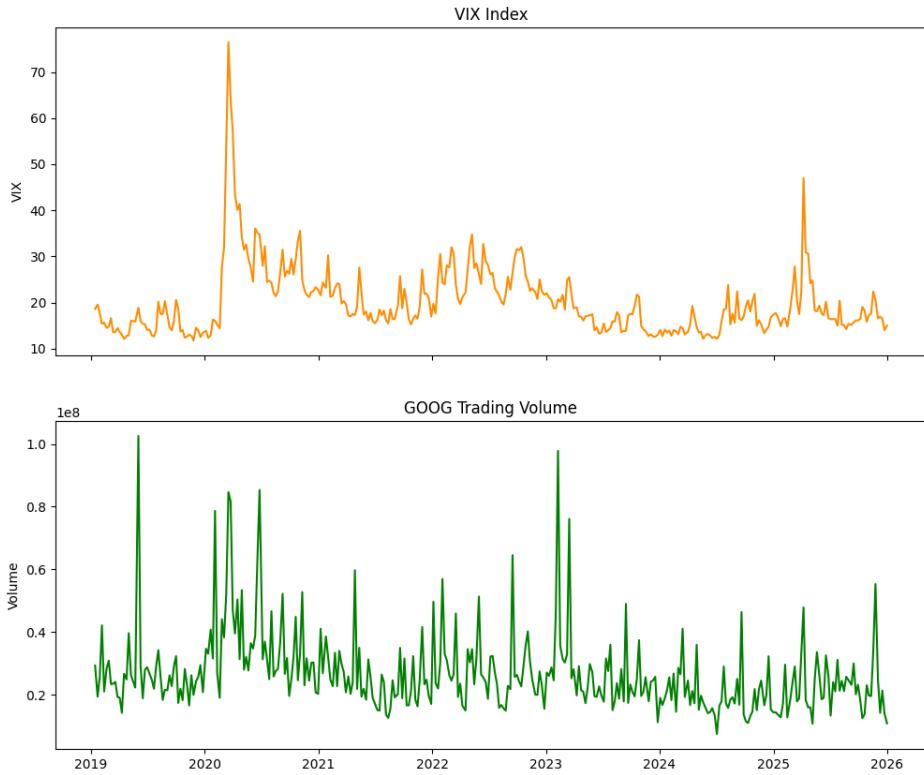


Figure 2: VIX index and GOOG daily trading volume over the sample period.

2.3 Train-test split

We split the data chronologically: the first 80% (280 observations, Jan 2019 – Aug 2024) form the training set, and the remaining 20% (71 observations, Aug 2024 – Dec 2025) form the test set. All model fitting and selection use only the training set; the test set is reserved for out-of-sample evaluation. Figure 3 shows the partition.

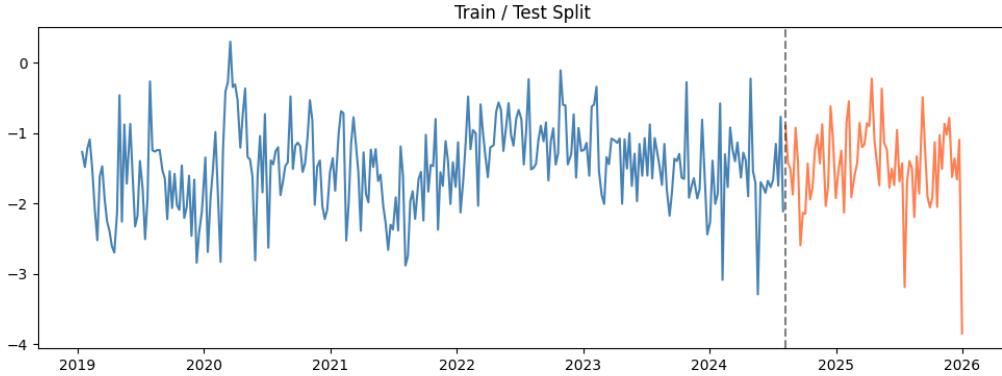


Figure 3: Train/test split of the Log RV series. The dashed line marks the boundary.

3 Exploratory Analysis

3.1 Stationarity

We apply two complementary tests to assess stationarity. The Augmented Dickey-Fuller (ADF) test fits the regression

$$\Delta Y_t = \alpha + \delta Y_{t-1} + \sum_{j=1}^k \gamma_j \Delta Y_{t-j} + u_t,$$

and tests $H_0: \delta = 0$ (unit root present) against $H_1: \delta < 0$ (stationary). The test statistic is compared to the Dickey-Fuller critical values rather than the standard normal, because the distribution is non-standard under the null [2]. Our training Log RV yields an ADF statistic of -5.32 ($p < 10^{-5}$), which falls well below the 1% critical value of -3.45 , so we reject the unit root null.

However, rejecting a unit root does not by itself establish stationarity, but it only rules out one form of non-stationarity. To approach the question from the opposite direction, we apply the KPSS test [7], which decomposes the series as $Y_t = \xi t + r_t + \epsilon_t$, where $r_t = r_{t-1} + u_t$ is a random walk component. The test statistic is

$$\eta = \frac{1}{n^2} \sum_{t=1}^n \frac{S_t^2}{\hat{\sigma}_\epsilon^2}, \quad S_t = \sum_{i=1}^t \hat{e}_i,$$

where \hat{e}_i are the residuals from a regression of Y_t on a constant (or constant plus trend) and $\hat{\sigma}_\epsilon^2$ is a consistent estimator of the long-run variance. Under H_0 (stationarity), η is small; we reject stationarity when η exceeds the KPSS critical value. Our KPSS test returns $p = 0.309$, so we fail to

reject stationarity at any conventional level ($\alpha = 0.01, 0.05, 0.10$). Since the ADF rejects the unit root and the KPSS does not reject stationarity, both tests agree that the Log RV series is stationary and no differencing is needed ($d = 0$).

3.2 Autocorrelation structure

Figure 4 shows the ACF and PACF of training Log RV. The ACF is generally noisy, but there is a slight decreasing trend, consistent with the well-documented long-memory behavior of volatility [5; 6]. The PACF shows significant spikes at lags 1-2, suggesting AR components of order at least 2. Both functions show mild structure around lag 4, motivating a seasonal component with period $s = 4$ (roughly monthly in weekly data).

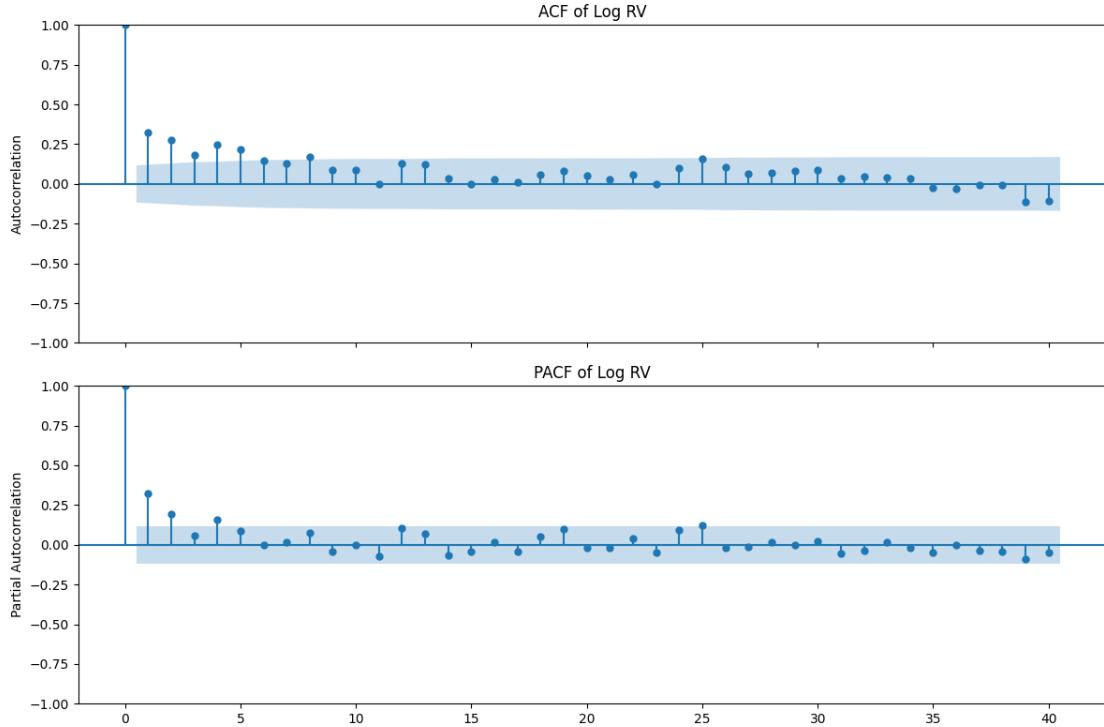


Figure 4: ACF and PACF of training Log RV.

3.3 Spectral analysis

Figure 5 presents the raw periodogram and a smoothed spectral density estimate. The raw periodogram, computed via the discrete Fourier transform, is an inconsistent estimator of the true spectral density because its variance does not decrease with sample size [1]. To obtain a more stable estimate, we apply Welch's method [8; 9], which divides the series into overlapping segments, computes a periodogram for each segment, and averages them. This trading of frequency resolution for reduced variance aligns with our dataset since there are few, dominant periodicities.

The raw periodogram shows elevated power near frequency zero, but this should not be interpreted as a genuine long-period cycle. With only about 280 training observations, the lowest resolvable frequencies correspond to periods of 140 observations (weeks), meaning the series contains too few complete cycles at those frequencies for the estimate to be reliable. This kind of low-frequency inflation is a well-known artifact in short financial time series and typically reflects the slow mean-reversion and persistence of volatility rather than any true periodicity [6].

The more meaningful feature is a peak near frequency 0.24 cycles per observation, corresponding to a period of approximately 4.2 weeks, which we round to 4 weeks. This aligns with the seasonal period $s = 4$ that we adopt in our SARIMA specifications, and it likely reflects the monthly cadence of earnings-related news flow and options expiration cycles that periodically elevate stock volatility.

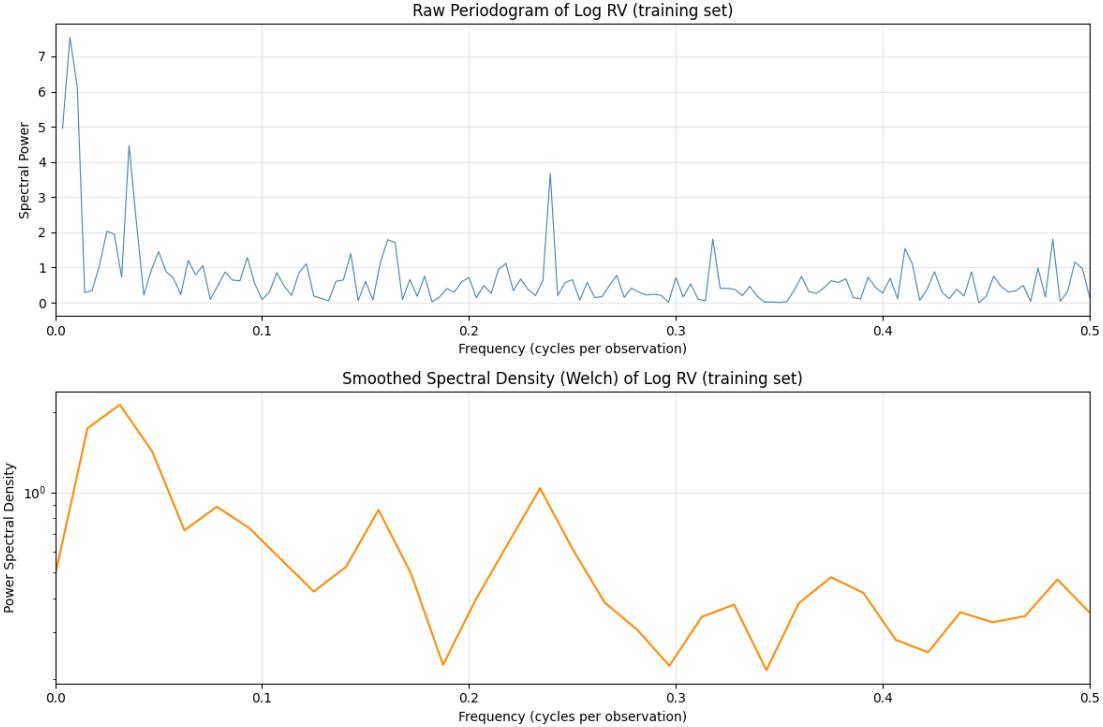


Figure 5: Raw periodogram (top) and Welch-smoothed spectral density (bottom).

4 Model Selection

We search over three model families, all fit on the training set with AIC as the selection criterion.

4.1 ARIMA and SARIMA

An ARIMA(p, d, q) model writes the differenced series as a linear function of its own past values and past white-noise errors [8]:

$$\phi(B)(1 - B)^d Y_n = \mu + \theta(B)\epsilon_n, \quad (1)$$

where B is the backshift operator ($BY_n = Y_{n-1}$), $\epsilon_n \sim WN(0, \sigma^2)$ is white noise with mean zero and variance σ^2 , and

$$\begin{aligned}\phi(B) &= 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p, \\ \theta(B) &= 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q\end{aligned}$$

are the AR and MA polynomials respectively. Since our Log RV series is stationary ($d = 0$), this simplifies to an ARMA(p, q).

A SARIMA(p, d, q)(P, D, Q) _{s} extends this by including seasonal polynomials that operate at lag s [8]:

$$\phi(B) \Phi(B^s) (1 - B)^d (1 - B^s)^D Y_n = \mu + \theta(B) \Theta(B^s) \epsilon_n, \quad (2)$$

where $\epsilon_n \sim WN(0, \sigma^2)$ and

$$\begin{aligned}\Phi(B^s) &= 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps}, \\ \Theta(B^s) &= 1 + \Theta_1 B^s + \Theta_2 B^{2s} + \cdots + \Theta_Q B^{Qs}\end{aligned}$$

are the seasonal AR and MA polynomials that capture periodic patterns at the seasonal period s .

We fit ARIMA($p, 0, q$) for $p, q \in \{0, \dots, 5\}$ (36 models) and SARIMA($p, 0, q$)($P, 0, Q$)₄ with $p, q \in \{0, \dots, 3\}$ and $P, Q \in \{0, 1\}$ (64 additional models). Table 1 shows the top models from both families. ARIMA(3,0,3) achieves the lowest AIC at 459.31, consistent with the ACF/PACF evidence for three AR and three MA lags. The seasonal extensions do not improve AIC enough to justify their extra parameters.

Table 1: Top (S)ARIMA models by AIC.

Model	AIC
ARIMA(3,0,3)	459.31
SARIMA(1,0,3)(0,0,1) ₄	463.20
SARIMA(3,0,3)(0,0,1) ₄	463.34
ARIMA(4,0,4)	463.92
SARIMA(1,0,3)(1,0,1) ₄	464.16
SARIMA(3,0,3)(1,0,0) ₄	464.28

4.2 SARIMAX

A SARIMAX model augments the SARIMA specification (Equation 2) with exogenous regressors X_n [1]:

$$\phi(B) \Phi(B^s) (1 - B)^d (1 - B^s)^D Y_n = \mu + \beta X_n + \theta(B) \Theta(B^s) \epsilon_n, \quad (3)$$

where $\epsilon_n \sim WN(0, \sigma^2)$ and β captures the contemporaneous effect of the external variable on the response. All polynomial terms $\phi(B)$, $\theta(B)$, $\Phi(B^s)$, $\Theta(B^s)$ follow the same definitions as in Equation 1 and Equation 2.

For each SARIMA specification above, we include three exogenous configurations: VIX only, Volume only, and VIX + Volume, yielding $64 \times 3 = 192$ additional fits. Table 2 shows the top models.

Table 2: Top SARIMAX models by AIC. All top entries use VIX as the sole exogenous variable.

Model	Exogenous	AIC
SARIMAX(3,0,3)(0,0,1) ₄	VIX	403.28
SARIMAX(3,0,3)(1,0,0) ₄	VIX	404.23
SARIMAX(3,0,3)(0,0,0) ₄	VIX	407.83
SARIMAX(1,0,1)(0,0,1) ₄	VIX	409.92
SARIMAX(1,0,3)(0,0,1) ₄	VIX	410.81
SARIMAX(1,0,1)(1,0,0) ₄	VIX	410.86

VIX-only models uniformly dominate. Volume-only configurations fare poorly (best AIC ≈ 721), and adding both VIX and Volume actually worsens AIC relative to VIX alone (best ≈ 621), suggesting Volume introduces noise when VIX is already present. VIX provides a clean, direct measure of market-wide volatility expectations, which explains its dominance.

4.3 Selected models

We carry forward two models for diagnostics and forecasting:

- **Model A:** ARIMA(3,0,3) — best pure time series model (AIC = 459.31)
- **Model B:** SARIMAX(3,0,3)(0,0,1)₄ with VIX — best exogenous model (AIC = 403.28)

The AIC gap of ≈ 56 strongly favors Model B.

5 Model Interpretation

5.1 ARIMA(3,0,3)

This model says the current Log RV depends on the past 3 values (AR terms) and the past 3 forecast errors (MA terms). The AR(3) component captures the short-term persistence: if volatility was elevated recently, it tends to remain so. The MA(3) component allows the model to correct for recent surprises.

5.2 SARIMAX(3,0,3)(0,0,1)₄ [VIX]

This model adds two new variables. First, VIX is included as a linear regressor with coefficient $\hat{\beta} \approx 0.041$ ($p < 0.001$), meaning each 1-point increase in VIX raises predicted Log RV by about 0.04 — a sensible relationship since elevated market fear corresponds to higher realized stock volatility. Secondly, the seasonal MA(1) at lag 4 captures monthly-scale corrections: if the model's forecast error from 4 weeks ago was large, it still influences today's prediction.

6 Diagnostics

6.1 Causality and invertibility

A fitted ARMA model is useful only if it satisfies two structural conditions [1; 8]. The first is causality: the roots of the AR polynomial $\phi(z)$ must all lie outside the unit circle ($|z| > 1$). When this holds, we can express the current value of the series purely in terms of past and present noise, so the model only looks backward in time. The second is invertibility: the roots of the MA polynomial $\theta(z)$ must also lie outside the unit circle. This ensures we can recover the noise terms from the observed data, which is what makes the fitted residuals meaningful. If either condition breaks down, there may be a different set of parameters that produces the same statistical behavior, and the estimates we obtained would not be uniquely interpretable.

We check these conditions by computing the polynomial roots and plotting them against the unit circle. For the ARIMA(3,0,3) model (Figure 6), the three AR roots have moduli of 1.073, 1.073, and 1.132, and the MA roots have moduli of 1.002, 1.002, and 1.434, all outside the circle.

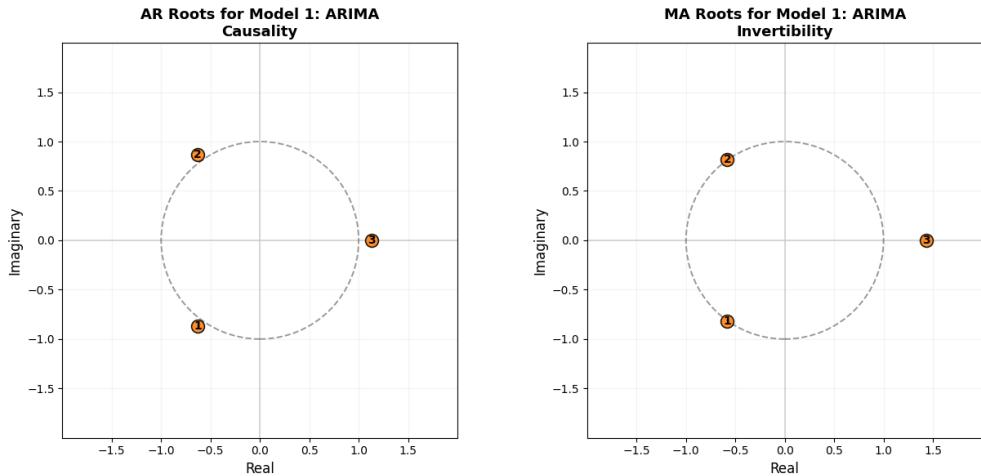


Figure 6: AR and MA roots for ARIMA(3,0,3).

The SARIMAX(3,0,3)(0,0,1)₄ model (Figure 7) has more roots to check because the seasonal MA factor $1 + \Theta_1 B^4$ multiplies the non-seasonal MA, producing 7 MA roots in total. The four outermost roots in Figure 7 (moduli ≈ 2.25) come from that seasonal term and are outside the circle. The three non-seasonal AR roots have moduli of 1.001, 1.427, and 1.427. One of them is extremely close to the boundary ($|z| = 1.001$), reflecting the strong persistence of volatility, but causality still holds. The MA roots range from 1.045 to 2.250, all safely outside. Both models pass these checks, so the fitted representations are valid.

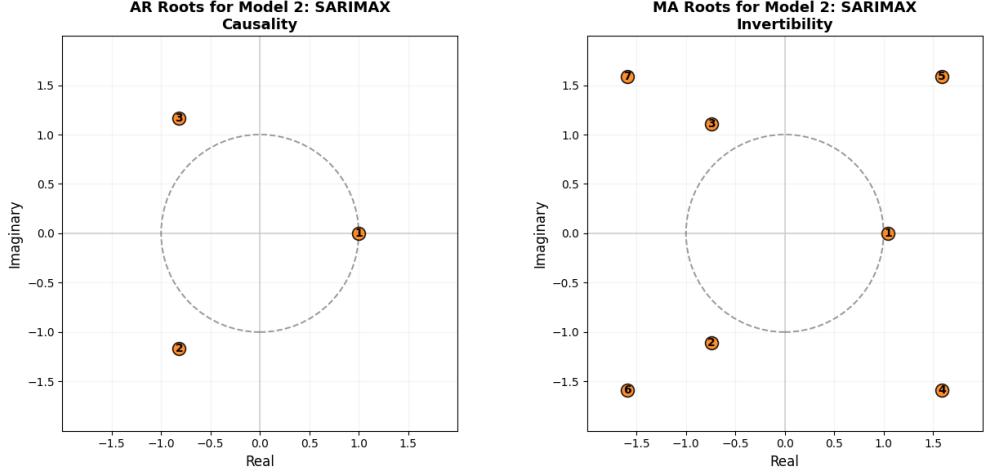


Figure 7: AR and MA roots for SARIMAX(3,0,3)(0,0,1)₄.

6.2 Residual tests

Table 3 summarizes the diagnostic tests on both fitted models.

Table 3: Residual diagnostic tests for both models.

Test	ARIMA(3,0,3)	SARIMAX(3,0,3)(0,0,1) ₄ [VIX]
Ljung-Box (lag 10)	$Q = 3.10, p = 0.979$	$Q = 6.09, p = 0.807$
Jarque-Bera	$JB = 2.93, p = 0.231$	$JB = 26.20, p < 0.001$
ARCH LM (5 lags)	$LM = 1.97, p = 0.854$	$LM = 0.96, p = 0.966$

The Ljung-Box test [10] evaluates whether the first h residual autocorrelations are jointly zero. The test statistic is

$$Q(h) = n(n + 2) \sum_{k=1}^h \frac{\hat{\rho}_k^2}{n - k},$$

where $\hat{\rho}_k$ is the sample autocorrelation at lag k and n is the sample size. Under H_0 (no autocorrelation), $Q(h) \sim \chi_h^2$, and we reject when $p < 0.05$.

The Jarque-Bera test [11] assesses whether the residuals follow a normal distribution by measuring departures in skewness \hat{S} and kurtosis \hat{K} from their Gaussian values ($S = 0, K = 3$):

$$JB = \frac{n}{6} \left(\hat{S}^2 + \frac{(\hat{K} - 3)^2}{4} \right).$$

Under normality, $JB \sim \chi_2^2$, and we reject when $p < 0.05$.

Both models pass the Ljung-Box test with large p -values (0.979 and 0.807), confirming no significant residual autocorrelation [10]. Neither shows ARCH effects (both $p > 0.85$), suggesting that residual heteroskedasticity is not a major concern. The ARIMA model also passes the Jarque-Bera normality test ($p = 0.231$), while the SARIMAX model shows slight non-normality ($p < 0.001$, excess kurtosis of 3.87), likely driven by a few large residuals associated with the VIX regressor during extreme market events.

6.3 Residual plots

Figure 8 and Figure 9 present three diagnostic plots for each model. The residual ACF assesses whether any linear dependence remains unexplained. Significant spikes beyond the confidence bands would suggest the model has not fully captured the autocorrelation structure of the series. Nearly all ACF values fall within the 95% confidence bands for both models. There is one lag in the range of 15–20 that barely touches the boundary, but a single marginal exceedance out of 20 lags is expected by chance alone at the 5% level and does not indicate a systematic problem.

The QQ plot compares the empirical quantiles of the residuals to those of a standard normal distribution. Residuals that follow the diagonal closely are consistent with normality. The ARIMA residuals adhere well to the reference line, while the SARIMAX residuals deviate at both tails, indicating heavier-than-normal extremes. This is consistent with the Jarque-Bera rejection noted in Table 3 and likely reflects a small number of large residuals driven by abrupt VIX movements that the linear regressor does not fully accommodate.

The histograms confirm that both sets of residuals are approximately symmetric and centered near zero. The SARIMAX distribution demonstrates slightly heavier tails, in agreement with the QQ plot. Although statistically significant, this degree of non-normality does not compromise the validity of the model, as maximum likelihood parameter estimates remain consistent under moderate departures from the Gaussian assumption [2].

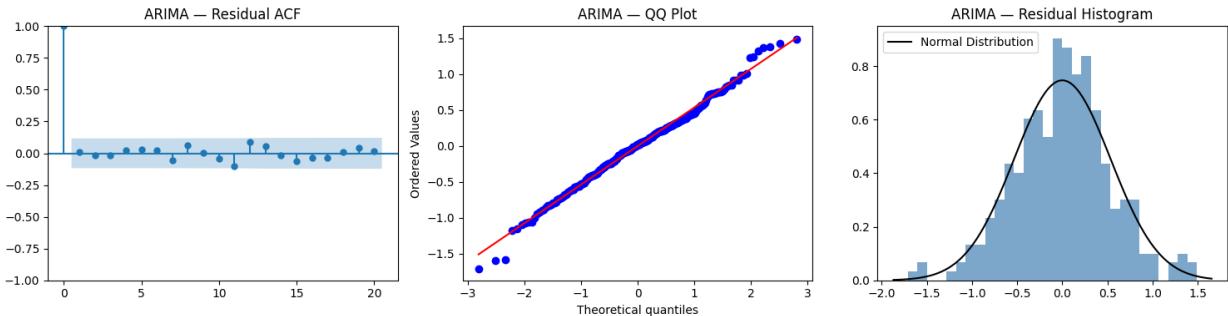


Figure 8: Residual diagnostics for ARIMA(3,0,3): ACF, QQ plot, and histogram.

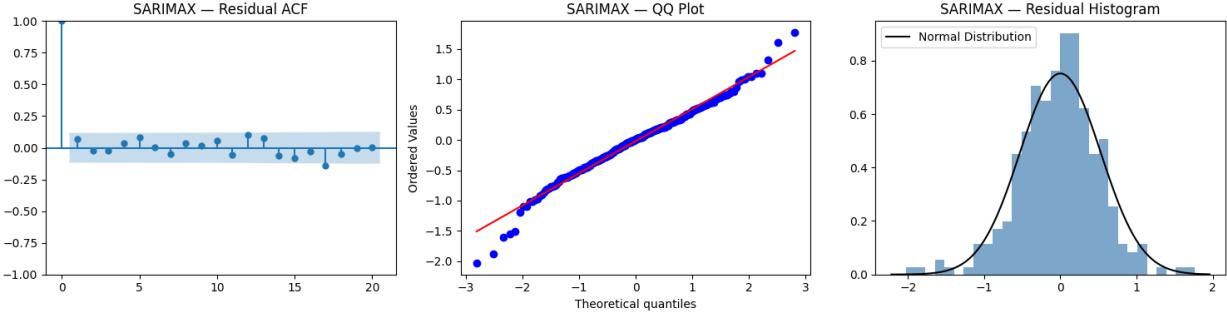


Figure 9: Residual diagnostics for SARIMAX(3,0,3)(0,0,1)₄ with VIX.

7 Out-of-Sample Evaluation

We produce one-step-ahead rolling forecasts on the 71-observation test set. At each step, the model is refit on all data observed so far (training + previously revealed test observations) and a single forecast is issued for the next period.

7.1 Forecast accuracy

Table 4 compares the two models on the test set.

Table 4: Out-of-sample forecast metrics.

Metric	ARIMA(3,0,3)	SARIMAX [VIX]
RMSE	0.6222	0.5723
MAE	0.4662	0.4238
MAPE (%)	43.96	39.67
Test R^2	-0.093	0.076
Train R^2	0.199	0.213

The SARIMAX model reduces RMSE by 8% and MAE by 9% relative to the pure ARIMA. The ARIMA model achieves a negative test R^2 , meaning it performs worse than simply predicting the test-set mean, while the SARIMAX model achieves a small but positive test R^2 of 0.076. These low R^2 values are not an anomaly but rather a direct implication of standard volatility models, as Andersen and Bollerslev [12] demonstrated. Realized volatility contains a large unpredictable component driven by news arrivals, earnings surprises, and sudden shifts in sentiment, so even well-specified models typically achieve out-of-sample R^2 in the range of 0.05 to 0.15 at weekly horizons. In that context, the SARIMAX result are reasonably within the expected range and confirms that VIX provides out-of-sample predictive content.

7.2 Forecast plots

Figure 10 overlays the actual and predicted Log RV on the test period for both models.

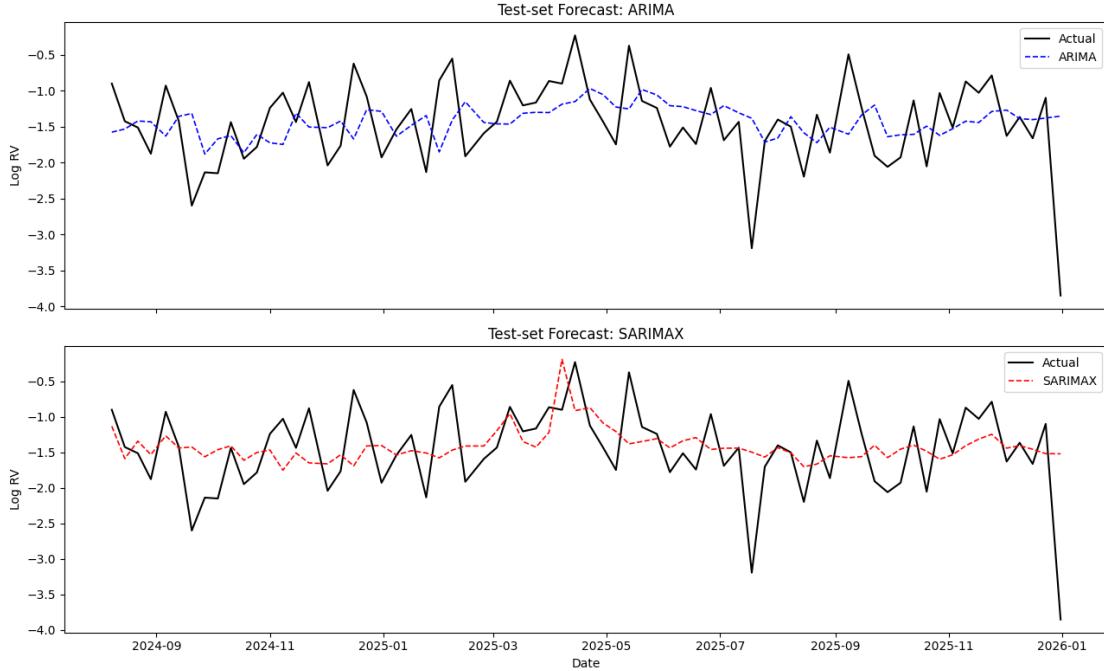


Figure 10: One-step-ahead rolling forecasts on the test set.

Both models are weak (as expected from the R^2 values), but the SARIMAX model is visibly more responsive to sharp movements (visible in the March-May of 2025 range). This is because VIX provides a contemporaneous signal of market stress that the ARIMA model, relying solely on past Log RV, cannot access.

8 Conclusions

We draw three main conclusions from this analysis.

First, **ARIMA(3,0,3)** provides a reasonable baseline for weekly GOOG Log RV. The ACF/PACF structure, AIC selection, and clean residual diagnostics all support this specification. However, its out-of-sample performance is weak: it barely improves upon a naive mean forecast.

Second, **VIX is a powerful exogenous predictor**. Among all exogenous configurations tested, VIX-only models consistently achieved the lowest AIC by a wide margin while volume provided negligible additional information.

Third, **the SARIMAX(3,0,3)(0,0,1)₄ model with VIX is our best overall model**, achieving the lowest AIC (403.28), the best test-set RMSE (0.572), and the only positive out-of-sample R^2 .

The seasonal MA term at lag 4 contributes a modest monthly correction. The VIX coefficient of ≈ 0.04 is highly significant and economically interpretable.

These results suggest that for volatility forecasting of individual stocks, incorporating a market-wide volatility measure like VIX is more valuable than increasing the complexity of the univariate time series model. Future work could explore GARCH-type models for the conditional variance of residuals, or test whether intraday realized volatility measures improve upon our daily-close-based estimates.

Acknowledgments

Previous projects explored similar results (notably the S&P 500 Volatility project from W25). Our project differed from other groups that analyzed stocks because we looked at realized volatility rather than price. The rationale behind this is that we wanted to do analyses that would be seen at a trading firm (as they trade volatility rather than direction in options trading). Thus, we are most similar to the S&P 500 Volatility analysis in that sense. However, we differ in our collection of data. They analyzed 30-day volatility of the S&P 500 whereas we specifically analyzed weekly realized volatility of Google. We wanted to analyze a symbol with more volatility and hopefully come up with more interesting results but due to honing in on GOOG, we had less data and could not do monthly RVs as it would shrink our dataset too much. Additionally, we log-transformed our RVS as they are strictly positive values reduces the variance.

While reading their paper, we learned lots of ways to model the data and attempted to improve on them. For instance, they dove into analyses of ARIMA, SARIMA, and ARMA+GARCH. Rather than using computation power in all of these families of models, we scored ARIMA, SARIMA, and SARIMAX with AIC values and only analyzed the top scoring models from each family ((S)ARIMA, (S)ARIMAX) in-depth.

Lastly, all market data (GOOG prices, VIX index, trading volume) were obtained through the `yfinance` Python API. AI (Claude, Anthropic) was used for debugging, formatting, and editing in this report.

Bibliography

- [1] Robert H Shumway and David S Stoffer. *Time Series Analysis and Its Applications: With R Examples*. Springer, 4th edition, 2017.
- [2] James D Hamilton. *Time Series Analysis*. Princeton University Press, 1994.
- [3] Eugene F Fama. Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 1970.
- [4] Robert E Whaley. The investor fear gauge. *The Journal of Portfolio Management*, 2000.
- [5] Torben G Andersen, Tim Bollerslev, Francis X Diebold, and Paul Labys. Modeling and forecasting realized volatility. *Econometrica*, 2003.
- [6] Clive WJ Granger and Zhuanxin Ding. Varieties of long memory models. *Journal of Econometrics*, 1996.

- [7] Denis Kwiatkowski, Peter CB Phillips, Peter Schmidt, and Yongcheol Shin. Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics*, 1992.
- [8] Edward L. Ionides. Stats 531 (winter 2026) course notes: Analysis of time series. <https://ionides.github.io/531w26/>, 2026.
- [9] Peter Welch. The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio and Electroacoustics*, 1967.
- [10] Greta M Ljung and George EP Box. On a measure of lack of fit in time series models. *Biometrika*, 1978.
- [11] Carlos M Jarque and Anil K Bera. A test for normality of observations and regression residuals. *International Statistical Review*, 1987.
- [12] Torben G Andersen and Tim Bollerslev. Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International Economic Review*, 1998.