Time Series Modeling of Google Realized Volatility with Exogenous Market Signals

Huy Le and Tyler Yee

February 2026

Abstract

We model the weekly log realized volatility (Log RV) of Google (GOOG) stock from 2019 to 2025 using ARIMA and SARIMAX frameworks. After constructing non-overlapping 5-day realized volatility estimates and merging exogenous variables (VIX and trading volume), we conduct an extensive AIC-based model selection over ARIMA, SARIMA, and SARIMAX specifications. The best pure time series model is ARIMA(3,0,3), while incorporating the VIX index as an exogenous regressor yields a SARIMAX$(3,0,3)(0,0,1)_4$ model 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

Realized volatility (RV) is a key quantity in financial risk management. Unlike implied volatility, which is forward-looking and derived from option prices, RV is computed directly from observed returns and provides an ex-post measure of how much a stock actually moved. Accurate forecasting of RV is essential for portfolio allocation, option pricing, and risk budgeting.

Google (Alphabet, ticker GOOG) is one of the most liquid large-cap equities globally. Its volatility dynamics are shaped by both idiosyncratic events (earnings, regulation, product launches) and systematic market forces. We ask: **can market-wide volatility, as measured by the VIX index, improve forecasts of Google's realized volatility beyond what is achievable with the stock's own history alone?**

The VIX, often called the "investor fear gauge" (1), measures the market's expectation of 30-day S&P 500 volatility implied by option prices. Prior literature has established that realized volatility exhibits long memory and strong persistence (2; 3), motivating ARMA-type models on log-transformed RV. We extend this by incorporating VIX as an exogenous regressor in a SARIMAX framework and evaluating whether it improves out-of-sample prediction.

Our data spans January 2019 to December 2025, covering the COVID crash, the post-pandemic recovery, and the 2022 rate-hike cycle — a period with substantial variation in volatility regimes.

2 Data and Preprocessing

2.1 Realized volatility construction

We download daily GOOG closing prices and compute daily log returns $r_t = \log(P_t/P_{t-1})$. Realized volatility over a 5-day (weekly) window is then

$$\mathrm{RV}_t = \hat{\sigma}_t\sqrt{252}, \quad \hat{\sigma}_t = \mathrm{std}(r_{t-4}, \ldots, r_t),$$

where the $\sqrt{252}$ annualizes the estimate. We log-transform to obtain Log $\mathrm{RV}_t = \log(\mathrm{RV}_t)$, which produces a more symmetric, approximately stationary series (2).

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 potential 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.

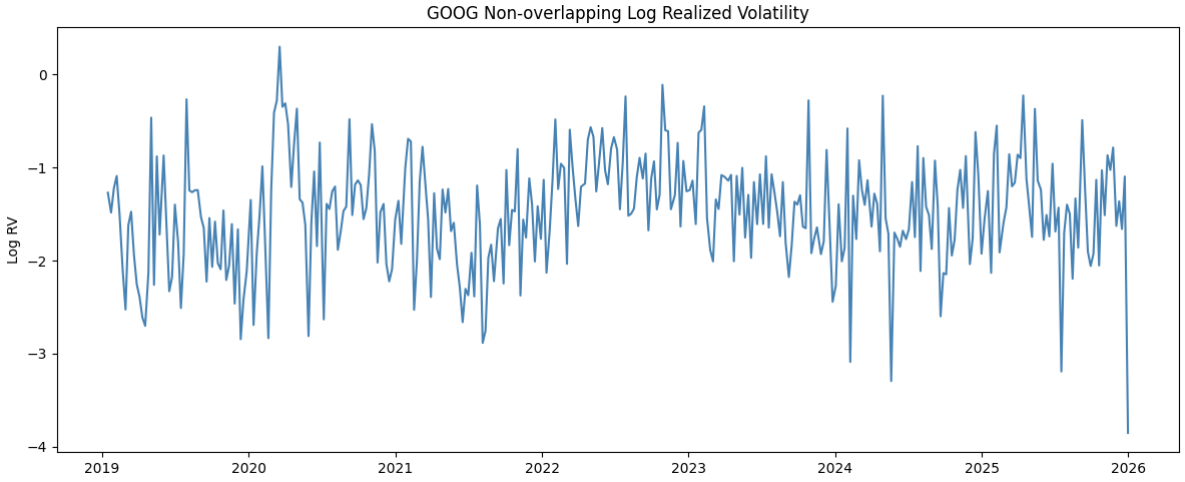

Figure 1: Non-overlapping weekly Log Realized Volatility of GOOG, January 2019 to December 2025.

Figure 2: Exogenous variables: 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.

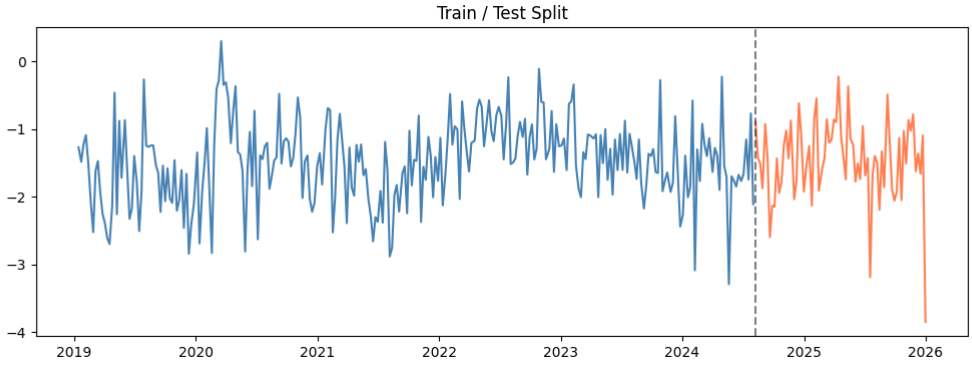

Figure 3: Train/test split of the Log RV series. The dashed line marks the boundary.

3 Exploratory Analysis

3.1 Stationarity

The Augmented Dickey-Fuller test on the training Log RV yields a test statistic of -5.32 ($p < 10^{-5}$), strongly rejecting the unit root null. The series is stationary, so no differencing is needed ($d = 0$).

3.2 Autocorrelation structure

Figure 4 shows the ACF and PACF of training Log RV. The ACF decays slowly, consistent with the well-documented long-memory behavior of volatility (2; 3). The PACF shows significant spikes at lags 1–3, suggesting AR components of order at least 3. Both functions show mild structure around lag 4, motivating a seasonal component with period $s = 4$ (roughly monthly in weekly data).

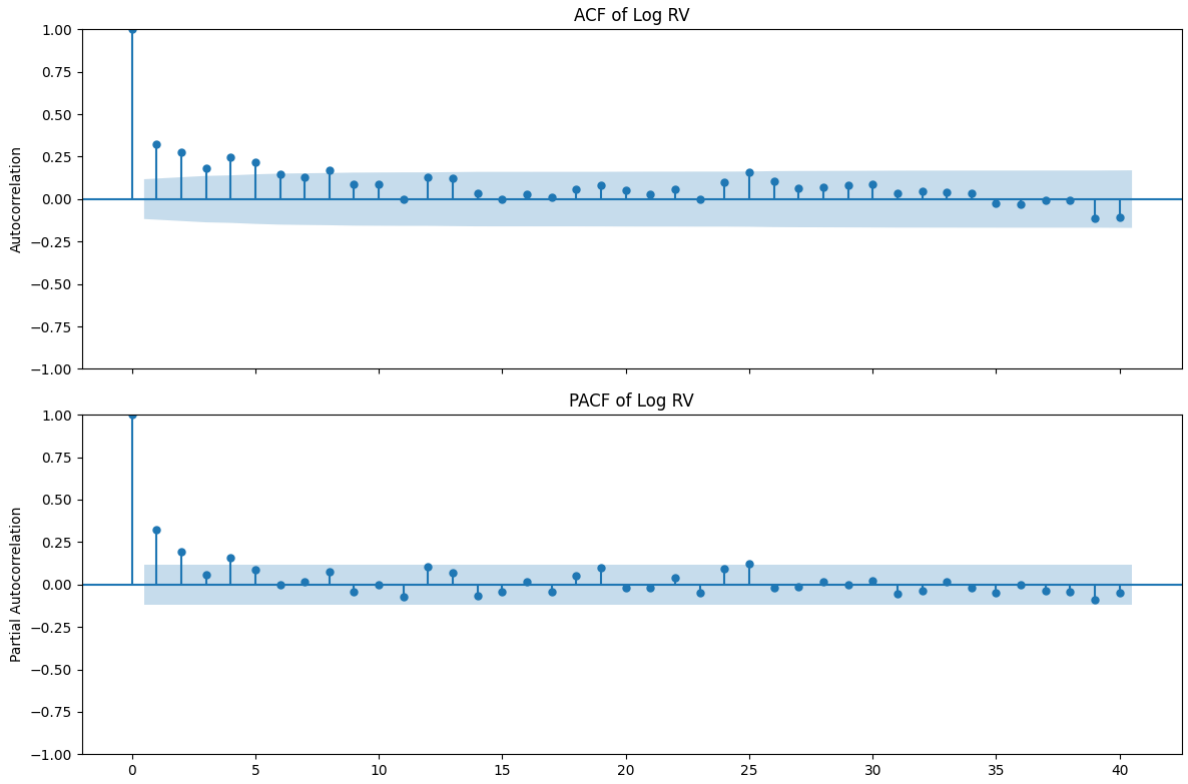

Figure 4: ACF and PACF of training Log RV. Significant autocorrelation persists well beyond lag 10; the PACF cuts off sharply after lag 3 with minor structure near lag 4.

3.3 Spectral analysis

Figure 5 presents the raw periodogram and the Welch-smoothed spectral density. The raw periodogram is noisy at low frequencies, which is expected since it is an inconsistent estimator of the spectral density (4). The high power near frequency zero reflects the long-memory persistence of volatility rather than a genuine long-period cycle (3).

The more interpretable feature is a broad peak near frequency 0.04 cycles per observation (period $\approx$ 25 weeks, or roughly 6 months), along with elevated power near 0.25 (period $\approx$ 4 weeks). The latter aligns with the seasonal period $s = 4$ used in our SARIMA specifications.

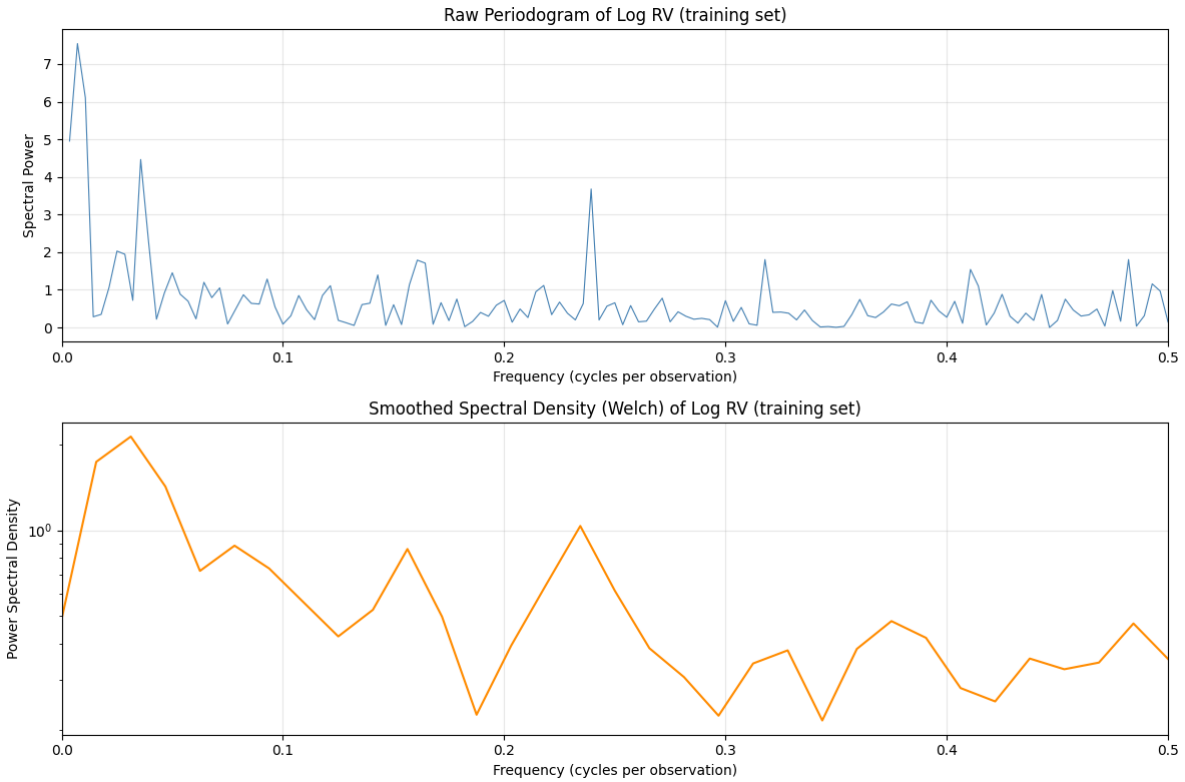

Figure 5: Raw periodogram (top) and Welch-smoothed spectral density (bottom) of training Log RV.

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

We fit ARIMA$(p,\ 0,\ q)$ for $p, q \in \{0, \ldots, 5\}$ (36 models) and SARIMA$(p,\ 0,\ q)(P,\ 0,\ Q)_4$ with $p, q \in \{0, \ldots, 3\}$ and $P, Q \in \{0, 1\}$ (64 additional models). Table 1 shows the top models from both families, ranked by AIC. 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. ARIMA(3,0,3) leads despite having no seasonal component.

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

Model	AIC
SARIMA(1,0,3)$(1,0,1)_4$	464.16
SARIMA(3,0,3)$(1,0,0)_4$	464.28

4.2 SARIMAX grid search

For each SARIMA specification, we additionally include three exogenous configurations: VIX only, Volume only, and VIX + Volume. This yields $64 \times 3 = 192$ additional fits. Table 2 summarizes 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 outperform Volume-only (best AIC ≈ 721) and VIX+Volume (best AIC ≈ 621) configurations. VIX directly measures market-wide volatility expectations, providing a clean signal. Volume is noisier and adds parameters without commensurate AIC improvement. Including both VIX and Volume worsens AIC relative to VIX alone, suggesting Volume introduces noise when VIX is already present.

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. All AR and MA coefficients except ar.L2 are significant at the 5% level.

5.2 SARIMAX(3,0,3)(0,0,1)$_4$ [VIX]

This model adds two ingredients. First, VIX enters 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. Second, 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 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)	$LM = 1.97$, $p = 0.854$	$LM = 0.96$, $p = 0.966$

Both models pass the Ljung-Box test, confirming no significant residual autocorrelation (5). Neither shows ARCH effects, suggesting that residual heteroskedasticity is not a major concern. The ARIMA model also passes the Jarque-Bera normality test, while the SARIMAX model shows mild non-normality (excess kurtosis of 3.87), likely driven by a few large residuals associated with the VIX regressor during extreme market events.

6.2 Residual plots

Figure 6 and Figure 7 show ACF, QQ, and histogram plots of the residuals. The ARIMA residuals are well-behaved. The SARIMAX residuals show slightly heavier tails in the QQ plot, consistent with the Jarque-Bera rejection, but the ACF is clean.

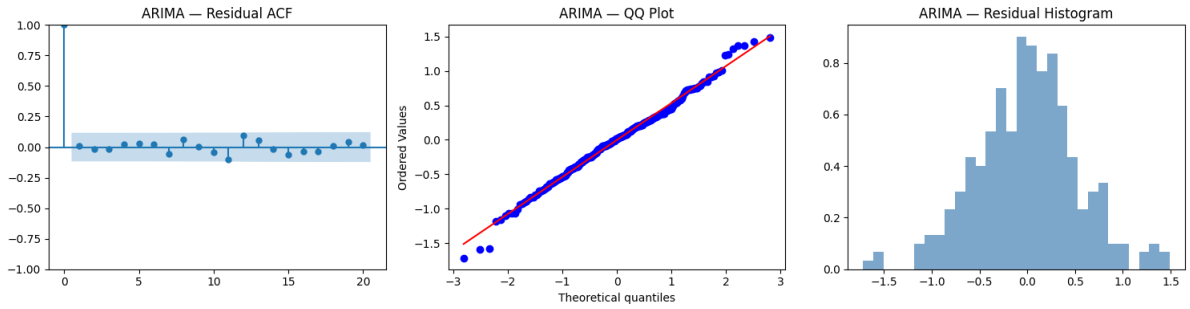

Figure 6: Residual diagnostics for ARIMA(3,0,3): ACF, QQ plot, and histogram.

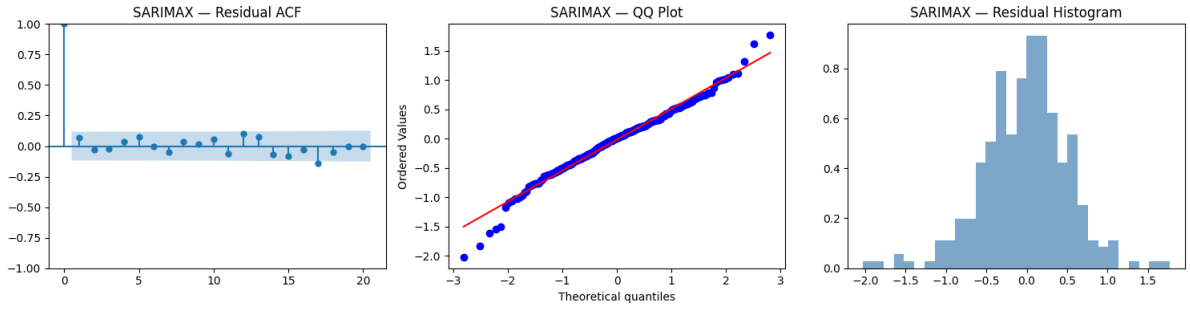

Figure 7: Residual diagnostics for SARIMAX(3,0,3)$(0,0,1)_4$ with VIX.

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. Bold indicates the better model.

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. Perhaps most strikingly, the ARIMA model achieves negative test R^2, meaning it performs worse than simply predicting the test-set mean. The SARIMAX model achieves a small but positive test R^2 of 0.076, demonstrating that VIX provides genuine out-of-sample predictive content.

7.2 Forecast plots

Figure 8 overlays the actual and predicted Log RV on the test period for both models.

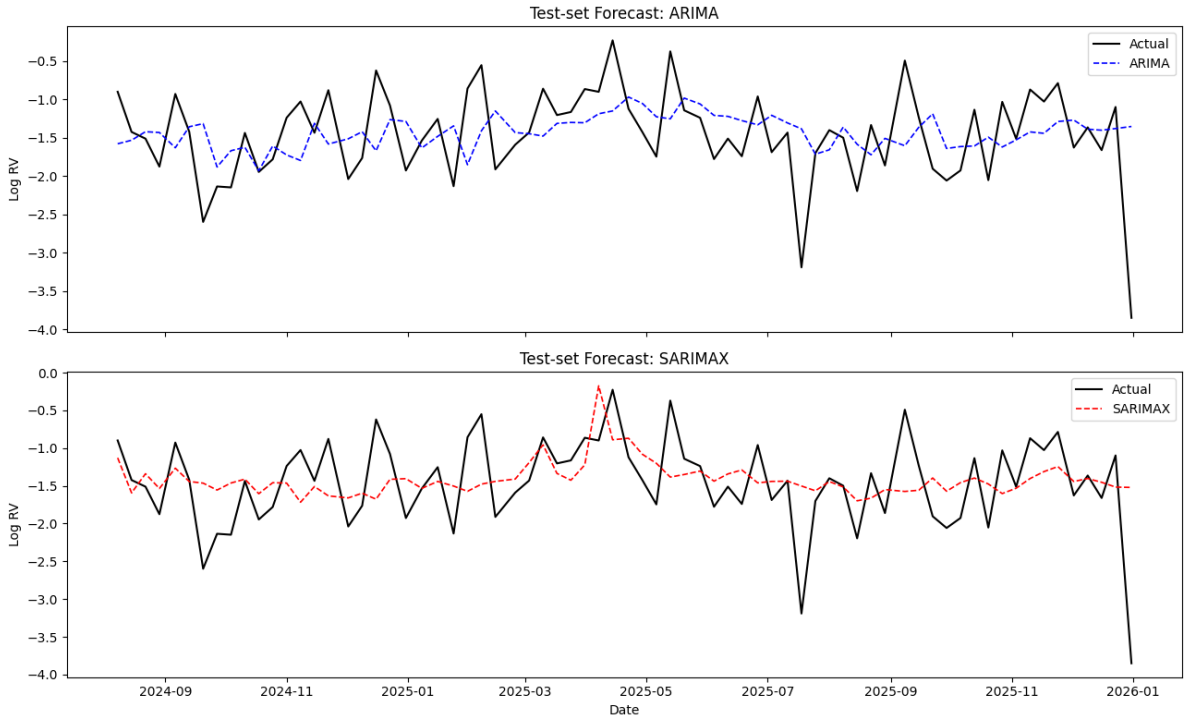

Figure 8: One-step-ahead rolling forecasts on the test set. Top: ARIMA(3,0,3). Bottom: SARIMAX(3,0,3)$(0,0,1)_4$ with VIX.

Both models track the general level and broad swings, but the SARIMAX model is visibly more responsive to sharp movements — precisely 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 — and by R^2 actually underperforms — 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. Volume provided negligible additional information. This aligns with the interpretation that market-wide fear, not idiosyncratic trading activity, drives the forecastable component of Google's volatility.

Third, **the SARIMAX(3,0,3)$(0,0,1)_4$ model with VIX is the 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 practical 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 (using 5-minute returns) improve upon our daily-close-based estimates.

Acknowledgments

[Previous project context to be added.]

AI (Claude, Anthropic) was used for debugging code, formatting assistance, and editing prose in this report. All statistical analysis was conducted by the author using the code available in `main.ipynb`.

Bibliography

[1] Robert E Whaley. The investor fear gauge. *The Journal of Portfolio Management*, 26(3):12–17, 2000.

[2] Torben G Andersen, Tim Bollerslev, Francis X Diebold, and Paul Labys. Modeling and forecasting realized volatility. *Econometrica*, 71(2):579–625, 2003.

[3] Clive WJ Granger and Zhuanxin Ding. Varieties of long memory models. *Journal of Econometrics*, 73(1):61–77, 1996.

[4] Robert H Shumway and David S Stoffer. *Time Series Analysis and Its Applications: With R Examples*. Springer, 4th edition, 2017.

[5] Greta M Ljung and George EP Box. On a measure of lack of fit in time series models. *Biometrika*, 65(2):297–303, 1978.

9 Supplementary Material

9.1 AIC grid search details

The full ARIMA grid searched 36 models ($p, q \in \{0, \dots, 5\}$); the SARIMA grid searched 64 models ($p, q \in \{0, \dots, 3\}$, $P, Q \in \{0, 1\}$, $s = 4$); and the SARIMAX grid searched 192 models (64 SARIMA specifications $\times$ 3 exogenous configurations). Models that failed to converge were excluded. All fitting used `statsmodels` SARIMAX with `enforce_stationarity=False` and `enforce_invertibility=False` to allow the optimizer full flexibility; root checks were performed post-hoc.

9.2 Causality and invertibility

For both selected models, we verified the AR and MA polynomial roots. The ARIMA(3,0,3) model has all AR roots with modulus > 1 (causal) and MA roots with modulus > 1 (invertible). The SARIMAX model's non-seasonal roots also satisfy these conditions. The ADF test on residuals of both models rejects the unit root null ($p < 10^{-5}$), confirming stationarity.