

Multiplicative GARCH-X with VIX: A Parsimonious Alternative to GARCH-MIDAS for Volatility Forecasting*

Yi-Hao Lai[†]

August 2026

Abstract

We propose a daily multiplicative GARCH-X model, $\sigma_t^2 = \tau_t \times g_t$, where τ_t is a simple function of lagged VIX and g_t follows a GJR-GARCH process on standardized returns. In a systematic horse race among 17 specifications—including ten daily GARCH-X variants, six GARCH-MIDAS alternatives with Beta-polynomial weighting, and a GJR-GARCH benchmark—the simplest daily model with $\tau_t = \theta_0 + \theta_1 \text{VIX}_{t-1}^2$ and freely estimated intercept (A4f) achieves the lowest QLIKE loss function, outperforming GJR-GARCH by a margin that clears our conservative $|t| > 3.0$ reporting screen (Diebold-Mariano $t = 4.47$; the screen is adopted following the multiple-testing concerns of [Harvey et al. \(2016\)](#) and is not a critical value from any cited source), and outperforming all GARCH-MIDAS specifications in point estimates. The result extends to QQQ ($t = 3.96$); on the four remaining assets—GLD with its own fear index GVZ ($t = 2.93$), FEZ ($t = 2.74$), EEM ($t = 2.69$) and 0050.TW ($t = 1.54$)—the improvement is consistent in sign but does not clear our $|t| > 3.0$ reporting screen. We show that τ_t reflects the option-market fear level

*All errors are my own. A replication package containing the pinned data snapshot and the estimation and reproduction code is available from the author on request.

[†]Department of Finance, Da-Yeh University, Taiwan. Corresponding author. Email: yh-lai@mail.dyu.edu.tw

while g_t captures residual GARCH dynamics. We also report a null result on the variance risk premium (VRP): across all four specifications examined, g_t shows no economically meaningful contemporaneous correlation with VRP ($|\rho| \leq 0.09$), and no out-of-sample predictive power for it. The multiplicative structure is a decomposition of volatility, not a recoverable premium signal. Residual diagnostics on the same pinned sample indicate that VIX absorption reduces excess kurtosis by 42% and shifts the median optimal Student- t degrees of freedom from 5.2 to 8.0; this accompanies the calibration ordering reported below rather than establishing a tail-risk advantage. For risk management the evidence is weaker than the forecasting evidence: under a three-test pass criterion (UC, CC, DQ) A4f- t , A4f-Normal and GJR-Normal each pass at one of three confidence levels and GJR- t at none, so the backtests do not establish a tail-risk advantage, although A4f- t 's violation rate is the closest to nominal at every level. These findings demonstrate that a parsimonious multiplicative GARCH-X specification with daily VIX is statistically indistinguishable from the best GARCH-MIDAS alternative (B1, $K = 22$; DM $t = -0.90$) while outranking longer-lag MIDAS specifications by margins that clear the $|t| > 3.0$ screen. The case for it therefore rests on parsimony and point-forecast accuracy; the backtests do not add a tail-risk advantage to that case.

Keywords: Volatility forecasting, GARCH-X, GARCH-MIDAS, VIX, variance risk premium, source decomposition, Value-at-Risk, model identification

JEL Classification: C22, C53, G12, G17

Contents

1	Introduction	5
2	Related Literature	8
2.1	Component Volatility Models	8
2.2	VIX and Volatility Forecasting	9
2.3	Variance Risk Premium	9
3	Methodology	10
3.1	The Multiplicative GARCH-X Framework	10
3.1.1	Scale Factor Specifications	11
3.1.2	Dynamic Factor Specification	12
3.2	Specification Taxonomy	13
3.3	GARCH-MIDAS Benchmark	15
3.4	Estimation	16
3.5	Evaluation Framework	18
3.5.1	Forecast Evaluation	18
3.5.2	VaR and ES Backtesting	21
4	Data	22
4.1	Primary Sample	22
4.2	Cross-Asset Sample	23
5	Empirical Results	24
5.1	Specification Horse Race	24
5.2	Source Decomposition and the Variance Risk Premium	28
5.2.1	VRP Tracking: A Null Result	30
5.2.2	VRP Predictability: A Null Result	33
5.3	Cross-Asset Validation	34
5.4	VaR and Expected Shortfall Backtesting	39

6	Robustness	42
6.1	Sensitivity to Estimation Settings	42
6.2	Reconciling the A4f-versus-GJR Statistics	48
6.3	Estimation/Forecasting Consistency	50
6.4	Model Confidence Set	51
6.5	Pairwise DM Tests Among Top Models	51
6.6	HAR-Style Daily- r^2 Benchmark Comparison	52
6.7	Multiple Testing Considerations	54
6.8	Student- t Degrees of Freedom, and a Confound in How They Are Estimated	54
6.9	Residual Diagnostics: How VIX Absorption Lightens the Residual Tails .	56
6.10	VIX vs. Macroeconomic Variables	58
7	Discussion: Theoretical Foundation of Source Decomposition	59
8	Conclusion	62
A	Specification Genealogy and Multiple-Testing Disclosure	66
B	HAC Bandwidth Sensitivity of the Headline Statistic	69

1 Introduction

Accurately forecasting equity return volatility is central to risk management, derivative pricing, and portfolio construction. A large literature documents that the CBOE Volatility Index (VIX), derived from S&P 500 index option prices, is a powerful—though upward-biased—predictor of future realized volatility (Christensen and Prabhala, 1998; Jiang and Tian, 2005; Bekaert and Hoerova, 2014). Yet the question of how best to integrate VIX information into parametric volatility models remains open.

One influential approach is the GARCH-MIDAS framework of Engle et al. (2013), which decomposes conditional variance into a slowly moving long-run component τ_t and a mean-reverting short-run component g_t through a multiplicative structure $\sigma_t^2 = \tau_t \times g_t$. The long-run component is modeled as a mixed-data-sampling (MIDAS) filter that aggregates lagged values of a lower-frequency variable—such as realized volatility, macroeconomic indicators, or financial conditions indices—using Beta-polynomial weights. This framework has been extended by Conrad and Loch (2015) to incorporate forward-looking variables and by Conrad and Kleen (2020) to include multiple external regressors.

However, when the external variable of interest is VIX, which updates at daily frequency, the MIDAS mixed-frequency machinery becomes conceptually redundant. VIX is already a daily variable; aggregating it through Beta-polynomial weights over 22, 65, or 125 lags introduces additional parameters without a clear economic rationale. This observation motivates a fundamental question: *can a simple daily multiplicative model with VIX match or exceed the forecasting performance of the more complex GARCH-MIDAS specification?*

We address this question through three contributions.

First, we conduct the most comprehensive specification comparison to date of 17 multiplicative volatility models, encompassing ten daily GARCH-X variants that differ in the functional form of τ_t (log-exponential, VIX-squared, VIX-level), the choice of normalization denominator (τ_t versus τ_{t-1}), and the treatment of the g_t intercept (restricted to $\omega_g = 1 - \alpha - \gamma/2 - \beta$ versus freely estimated), alongside six GARCH-MIDAS specifications with rolling-window and fixed-span Beta-polynomial filters at various lag lengths,

and a standard GJR-GARCH(1,1) benchmark. Using S&P 500 daily returns from 2005 to 2026 with an out-of-sample (OOS) period spanning 2019–2026 (1,826 observations), we evaluate all models under the proxy-robust QLIKE loss function of Patton (2011) and the stringent Diebold-Mariano testing framework, adopting a conservative $|t| > 3.0$ reporting screen consistent with the multiple-testing concerns raised by Harvey et al. (2016). The simplest daily model— $\tau_t = \theta_0 + \theta_1 \text{VIX}_{t-1}^2$ with freely estimated g_t intercept (denoted A4f)—achieves the lowest QLIKE and a DM test statistic of $t = 4.47$ against GJR-GARCH. While the Model Confidence Set cannot distinguish A4f from other well-specified VIX models (all survive at $\alpha = 0.10$), A4f and a majority of the VIX-augmented specifications clear the $|t| > 3$ screen against plain GJR-GARCH; the weaker variants do not. A4f remains the point-estimate best among functionally equivalent alternatives. Notably, GARCH-MIDAS specifications with longer lag structures ($K = 65$ – 125) are inferior to A4f by margins that clear our reporting screen ($|t| > 3.0$), while the shortest-lag MIDAS (B1, $K = 22$) is statistically indistinguishable (DM $t = -0.90$)—demonstrating that parsimony, not statistical dominance over all MIDAS variants, characterizes the advantage of daily GARCH-X when VIX updates daily.

Second, we propose a *source decomposition* interpretation that differs from the standard long-run/short-run narrative. In our framework, τ_t captures the *exogenous* volatility level implied by the options market (the “fear level”), while g_t captures *endogenous* GARCH dynamics conditional on this external environment. We further show what this decomposition is *not*. Across all four specifications examined, the fitted g_t has no economically meaningful contemporaneous correlation with the variance risk premium (VRP): the largest is $|\rho| = 0.089$ and three of the four are statistically indistinguishable from zero. Nor does g_t possess exploitable out-of-sample predictive power for future VRP, consistent with the high noise level and low autocorrelation of daily VRP. We report this null because the decomposition’s value lies in forecasting performance, which is unaffected, rather than in a premium signal it does not carry.

Third, we validate the model’s practical value through VaR and Expected Shortfall (ES) backtesting. Under a three-test pass criterion the VaR backtests separate the four

model-distribution configurations only at the 5% level, where A4f- t , A4f-Normal and GJR-Normal pass and GJR- t does not; at the 1% and 2.5% levels none pass. A4f- t 's violation rate is closest to nominal at every level, but we do not report proximity as a test. We report the Expected Shortfall tests but exclude them from that comparison: our bootstrap reference distribution leaves them without power, so they separate no model from any other (Section 5.4). The primary SPY evidence extends to QQQ (DM $t = 3.96$); three further assets—GLD with its own implied volatility index (GVZ), FEZ, and EEM—improve on GJR-GARCH in the same direction (DM $t \approx 2.7$ – 2.9) but do not clear our $|t| > 3.0$ reporting screen, and 0050.TW falls short as well. An asset's raw correlation with VIX does not by itself predict which assets clear the threshold: EEM's VIX correlation (0.50) exceeds QQQ's (0.49), yet EEM misses our screen while QQQ clears it, so we read the boundary as a sample-size and power question rather than a clean VIX-relevance rule.

Our findings contribute to the literature on component volatility models (Engle and Rangel, 2008; Engle et al., 2013; Conrad and Loch, 2015; Conrad and Kleen, 2020), the role of VIX in volatility forecasting (Bekaert and Hoerova, 2014), and the practical evaluation of volatility models for risk management (Patton, 2011; Hansen et al., 2011). The key message is one of parsimony: when the external variable is already daily, GARCH-MIDAS specifications with long lag structures ($K = 65$ – 125) are outranked by a simple daily multiplicative model by margins that clear our $|t| > 3.0$ reporting screen, while the shortest-lag alternative (B1, $K = 22$) is statistically indistinguishable—parsimony rather than statistical dominance is the practical argument for the daily specification. A simple multiplicative structure with daily VIX delivers the better point forecasts without the added complexity of mixed-frequency aggregation; the risk-management evidence in Section 5.4 does not add a tail-risk advantage to that case.

The remainder of this paper is organized as follows. Section 2 reviews the related literature. Section 3 presents the model specifications and evaluation framework. Section 4 describes the data. Section 5 reports the empirical results. Section 6 presents robustness analyses. Section 8 concludes.

2 Related Literature

2.1 Component Volatility Models

The idea of decomposing conditional variance into multiple components has a rich history. [Engle and Rangel \(2008\)](#) introduced the Spline-GARCH model, where a slowly moving spline function captures the unconditional variance trend while a unit-GARCH process captures short-run dynamics. [Engle et al. \(2013\)](#) generalized this idea with the GARCH-MIDAS framework, replacing the spline with a MIDAS filter that aggregates lagged values of observable variables—realized volatility, macroeconomic indicators, or financial variables—through Beta-polynomial weights. The appeal of GARCH-MIDAS lies in its ability to incorporate lower-frequency information (e.g., quarterly GDP) into a daily volatility model without frequency mismatch.

Subsequent work has expanded the framework in several directions. [Conrad and Loch \(2015\)](#) incorporated forward-looking macroeconomic variables, showing that the term spread, housing starts, corporate profits and unemployment improve long-horizon volatility forecasts. [Conrad and Kleen \(2020\)](#) demonstrated that combining realized volatility with macroeconomic variables in a two-component MIDAS filter outperforms single-variable specifications. [Wang and Ghysels \(2015\)](#) compared GARCH-MIDAS with alternative component models and found that the choice of explanatory variable matters more than the component structure itself.

However, a limitation of existing GARCH-MIDAS studies is that they focus on mixed-frequency settings where the MIDAS aggregation is economically motivated (e.g., monthly macro data driving daily volatility). When the external variable is itself daily—as is the case with VIX—the longer-lag MIDAS filter adds parameters without measurable forecasting benefit. Our paper fills this gap by showing that a simple daily multiplicative model outranks longer-lag MIDAS alternatives ($K \geq 65$) by margins that clear our reporting screen while achieving statistically indistinguishable performance relative to the shortest-lag MIDAS (B1, $K = 22$).

2.2 VIX and Volatility Forecasting

The role of VIX in volatility forecasting has been extensively studied. [Christensen and Prabhala \(1998\)](#) showed that implied volatility subsumes the information content of historical volatility. [Jiang and Tian \(2005\)](#) demonstrated that model-free implied volatility is a more efficient forecast than Black-Scholes implied volatility. [Bekaert and Hoerova \(2014\)](#) decomposed VIX^2 into expected variance and a variance risk premium component, showing that both contribute to volatility dynamics.

Several papers have incorporated VIX directly into GARCH models. [Han and Kristensen \(2014\)](#) developed asymptotic theory for the QMLE in GARCH-X models under both stationary and nonstationary covariates, which is the relevant case here given VIX’s occasional near-unit-root persistence during crisis periods. [Francq and Thieu \(2019\)](#) established QML inference for volatility models with covariates more generally, including the boundary-parameter cases that arise when a covariate coefficient is estimated at zero. Together these results justify the inclusion of VIX in GARCH specifications under mild regularity conditions. The general finding in this literature is that simple approaches to incorporating VIX often perform as well as complex ones—a result consistent with our findings.

Our contribution relative to this literature is twofold. First, we use a *multiplicative* rather than additive specification for incorporating VIX, which has the advantage of scaling the entire variance process rather than simply adding a level term. Second, we systematically compare the functional form of the VIX component (τ_t). The VIX-squared specification ranks first, on dimensional grounds that the horse race is consistent with but does not by itself establish: Finding 1 reports that it is not statistically separated from the log-exponential alternative.

2.3 Variance Risk Premium

The variance risk premium—defined as the difference between risk-neutral and physical expected variance—has attracted substantial attention since [Bollerslev et al. \(2009\)](#) showed that VRP predicts aggregate stock returns. [Carr and Wu \(2009\)](#) provided a theo-

retical framework for variance swaps, and [Bekaert and Hoerova \(2014\)](#) decomposed VIX^2 into expected variance and VRP.

Our paper connects to this literature through the g_t component. When $\tau_t \approx VIX_{t-1}^2$, the ratio $g_t \approx \sigma_t^2/VIX_{t-1}^2$ approximates the realized-to-implied variance ratio, which is inversely related to VRP. We test whether the GARCH autoregressive structure in g_t recovers a VRP signal from that ratio and find that it does not, either contemporaneously or predictively (Section 5.2).

3 Methodology

3.1 The Multiplicative GARCH-X Framework

We specify the daily return r_t of an equity index as:

$$r_t = \mu + \sigma_t \varepsilon_t, \quad \varepsilon_t \sim D(0, 1), \quad (1)$$

where μ is the unconditional mean return, σ_t^2 is the conditional variance, and ε_t is an i.i.d. innovation from distribution D with zero mean and unit variance. We set $\mu = 0$ throughout, which is standard for daily data where the mean is negligible relative to the variance.

The conditional variance, modeled within the GARCH framework ([Engle, 1982](#); [Bollerslev, 1986](#)), follows a multiplicative decomposition:

$$\sigma_t^2 = \tau_t \times g_t, \quad (2)$$

where $\tau_t > 0$ is the *exogenous scale factor* determined by the options-implied fear level (VIX), and $g_t > 0$ is the *endogenous dynamic factor* that captures GARCH-type persistence conditional on the external environment.

3.1.1 Scale Factor Specifications

We consider five functional forms for τ_t , all based on the lagged VIX to avoid look-ahead bias:

(a) **Log-exponential** (baseline):

$$\tau_t = \exp(\theta_0 + \theta_1 \log \text{VIX}_{t-1}). \quad (3)$$

(b) **VIX-squared** (variance-matching):

$$\tau_t = \max(\theta_0 + \theta_1 \text{VIX}_{t-1}^2, \epsilon), \quad (4)$$

where $\epsilon = 10^{-16}$ is a positivity floor. Since VIX is quoted in annualized volatility units (%), VIX^2 has variance units, matching the dimension of σ_t^2 . We use VIX in its raw form (annualized percentage); the coefficients θ_0 and θ_1 absorb the necessary unit conversions to daily decimal variance.

(c) **VIX-level** (exponential):

$$\tau_t = \exp(\theta_0 + \theta_1 \text{VIX}_{t-1}). \quad (5)$$

(d) **Sample-mean normalized**: Retain the intercept restriction $\omega_g = 1 - \alpha - \gamma/2 - \beta$ and additionally rescale τ_t by the in-sample mean of the squared return standardized by τ_t :

$$\hat{m} = \frac{1}{T-1} \sum_{t=2}^T \frac{r_{t-1}^2}{\tau_t}, \quad \tau_t^{(n)} = \hat{m} \tau_t, \quad u_{t-1}^{(n)} = \frac{r_{t-1}}{\sqrt{\tau_t^{(n)}}} = \frac{r_{t-1}/\sqrt{\tau_t}}{\sqrt{\hat{m}}}, \quad (6)$$

with conditional variance $\sigma_t^2 = \tau_t^{(n)} g_t$. The estimate $\hat{m}$ is recomputed inside the likelihood at every parameter draw, so it is a function of (θ_0, θ_1) rather than a preprocessing constant. By construction $(T-1)^{-1} \sum_{t=2}^T (u_{t-1}^{(n)})^2 = 1$: the normalization fixes the scale of the *standardized innovation*, not of τ_t . It therefore does

not impose $\mathbb{E}[\tau_t^{(n)}] = 1$, and we do not claim it. Under the lagged-normalization convention of Equation (9) the same construction pairs r_{t-1}^2 with τ_{t-1} in $\hat{m}$.

3.1.2 Dynamic Factor Specification

The short-run component g_t follows a GJR-GARCH process (Glosten et al., 1993) on standardized returns:

$$g_t = \omega_g + \alpha u_{t-1}^2 + \gamma u_{t-1}^2 \mathbb{1}_{[u_{t-1} < 0]} + \beta g_{t-1}, \quad (7)$$

where the standardized return u_{t-1} is defined under one of two normalization conventions:

$$u_{t-1}^{(a)} = \frac{r_{t-1}}{\sqrt{\tau_t}} \quad (\text{contemporaneous normalization}), \quad (8)$$

$$u_{t-1}^{(b)} = \frac{r_{t-1}}{\sqrt{\tau_{t-1}}} \quad (\text{lagged normalization}). \quad (9)$$

Contemporaneous normalization follows the recursive structure of Engle et al. (2013, Eq. 4), where the current volatility environment determines how past shocks are interpreted.¹ Despite the shared time subscript, there is no simultaneity: τ_t is a deterministic function of VIX_{t-1} , which is observed before r_t . Lagged normalization reflects a conservative timing choice in which the normalization denominator τ_{t-1} is fully predetermined before the shock r_{t-1} is observed, since both τ_t and u_{t-1} use VIX_{t-1} information. Our baseline specifications use contemporaneous normalization; lagged normalization is included for comparison (see Section 3.2).

Two intercept treatments are considered:

- (i) **Constrained:** Setting $\omega_g = 1 - \alpha - \gamma/2 - \beta$ pins the intercept at the value that

¹This timing convention is a design choice rather than an identification necessity. We explicitly compare the Engle-style recursion $u_{t-1} = r_{t-1}/\sqrt{\tau_t}$ against the more conservative DGP-timed alternative $u_{t-1} = r_{t-1}/\sqrt{\tau_{t-1}}$. The main paper adopts the former as its baseline because τ_t is predetermined by VIX_{t-1} and therefore known before r_t is realized. A later source-code audit of downstream robustness scripts found that some follow-up experiments inherited this Engle-style recursion without always stating the convention explicitly; a dedicated refit confirms that the directional conclusions are preserved under the lagged-normalization alternative, although some magnitudes shift. We therefore interpret the τ_t baseline as an implementation convention inherited from Engle et al. (2013), not as the only admissible timing assumption.

would deliver $\mathbb{E}[g_t] = 1$ if $\mathbb{E}[u_{t-1}^2] = \mathbb{E}[g_{t-1}]$. That auxiliary condition holds under the lagged normalization of Equation (9) but not under the contemporaneous normalization our baseline specifications use, so the restriction operates as a scale normalization rather than as a moment condition; Section 7 establishes this and reports $\mathbb{E}[g_t] = 0.84$ for the constrained fit on our sample. The unconditional variance therefore obeys the full identity $\mathbb{E}[\sigma_t^2] = \mathbb{E}[\tau_t] \mathbb{E}[g_t] + \text{Cov}(\tau_t, g_t)$, in which neither term may be dropped: Section 7 reports $\text{Corr}(\tau_t, g_t) = 0.51$ for this constrained fit.

- (ii) **Free:** ω_g is estimated without constraint, which lets the model accommodate a systematic wedge between implied and realized variance. Note that ω_g then absorbs the scale of τ_t , so neither ω_g nor $\mathbb{E}[g_t]$ is separately identified in this case; only the product $\tau_t g_t$ and functions of it are. Section 7 makes this precise and reports the identified discount instead.

3.2 Specification Taxonomy

Table 1 presents the full taxonomy of 17 specifications examined in this paper.

Table 1: Taxonomy of 17 multiplicative volatility model specifications

#	Name	τ_t function	u_{t-1} denom.	ω_g	Category
A1	Original	log-exp	mixed ^b	constrained	GARCH-X
A2	consistent- τ_t	log-exp	τ_t	constrained	GARCH-X
A3	consistent- τ_{t-1}	log-exp	τ_{t-1}	constrained	GARCH-X
A4	VIX-squared	VIX ²	τ_t	constrained	GARCH-X
A5	VIX-level	exp(VIX)	τ_t	constrained	GARCH-X
A2f	free-omega	log-exp	τ_t	free	GARCH-X
A4f	VIX ² -free	VIX ²	τ_t	free	GARCH-X
A3f	τ_{t-1} -free	log-exp	τ_{t-1}	free	GARCH-X
A2n	log-exp-norm	log-exp, norm	τ_t	constrained ^a	GARCH-X
A4n	VIX ² -norm	VIX ² , norm	τ_t	constrained ^a	GARCH-X
B1	MIDAS-RW-K22	Beta poly, $K = 22$	τ_t	constrained	MIDAS-RW
B2	MIDAS-RW-K65	Beta poly, $K = 65$	τ_t	constrained	MIDAS-RW
B3	MIDAS-RW ($K = 125$)	Beta poly, $K = 125$	τ_t	constrained	MIDAS-RW
C1	MIDAS-FS-K6	Beta poly, $K_m = 6$	τ_t	constrained	MIDAS-FS
C2	MIDAS-FS-K12	Beta poly, $K_m = 12$	τ_t	constrained	MIDAS-FS
C3	MIDAS-FS-K24	Beta poly, $K_m = 24$	τ_t	constrained	MIDAS-FS
B0	GJR-GARCH	—	—	—	Benchmark

^a A2n and A4n keep the $\omega_g = 1 - \alpha - \gamma/2 - \beta$ restriction and additionally rescale τ_t so that the standardized innovation has unit in-sample second moment (Equation 6). The rescaling normalizes u_{t-1} , not τ_t .

^b A1 uses τ_t as the u_{t-1} denominator in estimation but τ_{t-1} out of sample; this estimation/forecasting inconsistency is the defect A2–A3 correct (Section 3.2).

MIDAS-RW: rolling-window MIDAS with daily VIX lags. MIDAS-FS: fixed-span MIDAS with monthly VIX averages.

The Beta polynomial $\phi_k(\omega_1, \omega_2) \propto (k/K)^{\omega_1-1}(1 - k/K)^{\omega_2-1}$ with *both* ω_1 and ω_2 estimated on $[1, 20]$ (Section 3.3). Because the weight function is evaluated at k/K , $\phi_K = 0$ whenever $\omega_2 > 1$, so a filter declared with K lags carries $K - 1$ effective lags.

The key design dimensions are:

1. **τ_t functional form:** VIX-squared ($\tau \propto \text{VIX}^2$) provides dimensional consistency—since VIX is quoted in volatility units, VIX^2 has variance units matching σ_t^2 . Log-exponential ($\tau = \exp(\theta_0 + \theta_1 \log \text{VIX})$) is the specification used in the original GARCH-MIDAS literature. VIX-level ($\tau = \exp(\theta_0 + \theta_1 \text{VIX})$) is included for completeness.
2. **Normalization denominator:** Using τ_t (contemporaneous) follows the recursive structure of Engle et al. (2013, Eq. 4), where the current environment determines the normalization. Using τ_{t-1} (lagged) reflects a conservative timing convention in which the normalization denominator is predetermined before the shock r_{t-1} is observed.
3. **Intercept treatment:** The constrained case fixes ω_g at the value that enforces $\mathbb{E}[g_t] = 1$ under lagged normalization, which is the standard assumption in component models; under the contemporaneous normalization of our baseline it is a scale normalization instead (Section 7). Either way τ_t determines the unconditional variance only up to $\mathbb{E}[g_t]$ and $\text{Cov}(\tau_t, g_t)$ (Equation 20). The free case provides flexibility to accommodate a systematic wedge between implied and realized variance (the VRP).

3.3 GARCH-MIDAS Benchmark

For the GARCH-MIDAS specifications (B1–B3, C1–C3), the long-run component is:

$$\log \tau_t = m + \theta \sum_{k=1}^K \phi_k(\omega_1, \omega_2) \log \text{VIX}_{t-k}, \quad (10)$$

where $\phi_k(\omega_1, \omega_2)$ are Beta-polynomial weights:

$$\phi_k(\omega_1, \omega_2) = \frac{(k/K)^{\omega_1-1} (1 - k/K)^{\omega_2-1}}{\sum_{j=1}^K (j/K)^{\omega_1-1} (1 - j/K)^{\omega_2-1}}, \quad (11)$$

Both shape parameters are freely estimated subject to $\omega_1, \omega_2 \in [1, 20]$. This departs from Engle et al. (2013), who fix $\omega_1 = 1$ so that the weights decay monotonically in the

lag; we let the data choose the shape. The departure is deliberate and it works against our own thesis in the only direction that matters: it makes the MIDAS benchmark strictly more flexible than the restricted version usually estimated, so the horse race gives MIDAS the benefit of the doubt. It also means the mixed-frequency filter costs four parameters beyond the short-run GJR block— m , θ , ω_1 and ω_2 —rather than the three a fixed- ω_1 implementation would cost, which strengthens rather than weakens the parsimony argument of Section 1.

Equation (11) evaluates the Beta kernel at k/K . Consequently $\phi_K = (1 - 1)^{\omega_2 - 1} = 0$ for any $\omega_2 > 1$, and a filter declared with K lags places positive weight on only $K - 1$ of them. We retain this convention because it is the form in which the weights are usually written, but we state the implication explicitly: B1 ($K = 22$) uses 21 effective daily lags, B2 ($K = 65$) uses 64, B3 ($K = 125$) uses 124, and C1–C3 use 5, 11 and 23 effective monthly lags respectively. We did not re-estimate under an alternative convention such as $k/(K + 1)$, so we make no claim about how much the omitted lag would matter; what we do claim is only that the declared and effective lag lengths differ by one throughout, and that the tables report the declared value.

The rolling-window variant (B1–B3) applies the MIDAS filter at daily frequency with $K \in \{22, 65, 125\}$ daily lags. The fixed-span variant (C1–C3) uses monthly VIX averages with $K_m \in \{6, 12, 24\}$ monthly lags, which is closer to the original GARCH-MIDAS design.

The short-run component g_t follows the same GJR specification as Equation (7), with $\omega_g = 1 - \alpha - \gamma/2 - \beta$.

3.4 Estimation

All multiplicative models are estimated by joint maximum likelihood. For the GARCH-X variants (A-series), the parameter vector is $\boldsymbol{\psi} = (\theta_0, \theta_1, \omega_g, \alpha, \gamma, \beta)$ for the free-omega case, or $\boldsymbol{\psi} = (\theta_0, \theta_1, \alpha, \gamma, \beta)$ with $\omega_g = 1 - \alpha - \gamma/2 - \beta$ for the constrained case. For the GARCH-MIDAS variants (B/C-series) the parameter vector is $\boldsymbol{\psi} = (m, \theta, \alpha, \gamma, \beta, \omega_1, \omega_2)$ with $\omega_g = 1 - \alpha - \gamma/2 - \beta$: seven free parameters, against five for a constrained GARCH-X

variant and six for a free- ω_g one.

Under the assumption of Gaussian innovations, the log-likelihood is:

$$\ell(\psi) = -\frac{1}{2} \sum_{t=1}^T \left[\log(2\pi) + \log(\tau_t g_t) + \frac{r_t^2}{\tau_t g_t} \right]. \quad (12)$$

Optimization uses L-BFGS-B with three sets of starting values to mitigate local optima; the fit with the highest likelihood is retained. Parameter bounds ensure positivity ($\alpha, \gamma, \beta > 0$) and stationarity ($\alpha + \gamma/2 + \beta < 1$). For the free-omega models, we additionally impose $\omega_g > 0$ to ensure $g_t > 0$ for all t . We do not read the estimated ω_g values as evidence that this bound is slack, because they are not identified (Proposition 2); positivity of g_t holds along the whole ridge, so this constraint is never binding.

Two further features of the implementation are substantive enough to state rather than leave to the replication code. First, the box constraints are tighter than positivity and stationarity alone require, and one of them is a genuine restriction rather than a numerical safeguard: $\beta \in [0.5, 0.999]$ excludes low-persistence short-run dynamics by assumption. The remaining bounds are $\alpha, \gamma \in [10^{-4}, 0.3]$; for the log-exponential and VIX-level forms, $\theta_0 \in [-20, 0]$ and $\theta_1 \in [0.1, 5]$ and $[0.01, 1]$ respectively; for the VIX-squared form, $\theta_0 \in [-10^{-2}, 10^{-2}]$ and $\theta_1 \in [10^{-8}, 10^{-3}]$, which are unit-conversion scales rather than economic restrictions because VIX^2 enters in annualized percentage-squared units while σ_t^2 is a daily decimal variance; $\omega_g \in [10^{-6}, 1]$ in the free case; and $\omega_1, \omega_2 \in [1, 20]$ for the MIDAS filters. The estimation routine does not persist per-window parameter vectors, so we do not report which bounds bind in the rolling fits. Section 7 rests on two full-sample fits, and they are estimated by different routines from each other, which matters for replication. The free- ω_g fit follows the same protocol as the rolling A4f fits—three dispersed starting values, L-BFGS-B, box constraints—but it is a separate implementation rather than the same routine, and its θ_1 box is the wider $[10^{-10}, 10^{-2}]$. The landing point's $\theta_1 = 1.7 \times 10^{-7}$ is interior to both boxes, so the wider one does not move where that fit lands. The constrained fit is estimated differently: a single Nelder–Mead start from those free- ω_g estimates, with no box constraints at all, feasibility being imposed by

rejecting draws that violate non-negativity or push persistence to 0.999, and ω_g being the residual $1 - \alpha - \gamma/2 - \beta$ rather than a free parameter. Its ARCH coefficient lands at $\hat{\alpha} = 3.3 \times 10^{-15}$ —positive, but more than ten orders of magnitude below the 10^{-4} floor listed above, and therefore outside that box altogether. Both parameter vectors, each with a statement of the estimator that produced it, are archived with the replication package. A separate archived run of the constrained specification under a bounded L-BFGS-B protocol lands instead exactly on that floor, $\hat{\alpha} = 10^{-4}$, and Section 7 reports both runs’ figures and the gap between them. In neither constrained run is $\hat{\alpha}$ an estimate: each lands at the lower edge of its own admissible set—to the optimizer’s tolerance in ours, where the only restriction is $\alpha \geq 0$, and exactly in that run’s, where the edge is 10^{-4} —and the $\mathbb{E}[g_t]$ and $\text{Corr}(\tau_t, g_t)$ figures that follow are conditional on which admissible set was imposed. Separately, the β lower bound is imposed as a modelling restriction and is not tested against the data. Second, a starting value whose optimization raises an exception is skipped, and a window is reported as a failed fit only if all three starting values fail; because the three starts are dispersed, a window that returns a fit may in principle rest on fewer than three completed optimizations. This is a robustness convenience, but it means the number of completed optimizations behind a reported estimate is not guaranteed to be three, and we flag it rather than let the replication code be the only record of it.

3.5 Evaluation Framework

3.5.1 Forecast Evaluation

We adopt a rolling-window pseudo-out-of-sample design:

1. Fix an estimation window of $W = 2,000$ trading days.
2. Estimate all models on the window $[t - W + 1, t]$.
3. Generate one-step-ahead forecasts $\hat{\sigma}_{t+1}^2$.
4. Advance the window by one day and repeat.

5. Refit parameters every 63 trading days (approximately quarterly) to balance estimation cost and adaptation.

The primary loss function is the QLIKE of [Patton \(2011\)](#):

$$L_{\text{QLIKE}}(\hat{\sigma}_t^2, r_t^2) = \frac{r_t^2}{\hat{\sigma}_t^2} - \log \frac{r_t^2}{\hat{\sigma}_t^2} - 1, \quad (13)$$

evaluated on the squared return r_t^2 as the volatility proxy. The legacy horse-race tables report the average kernel $\log(\hat{\sigma}_t^2) + r_t^2/\hat{\sigma}_t^2$, which differs from Equation 13 by an observation-specific term common to every model in a given comparison. It therefore preserves within-sample loss differences and rankings, but not absolute levels or percentage reductions. Section 6.6 instead reports the full dimensionless loss in Equation 13; levels and percentages must not be compared across those two reporting conventions.² The proxy-robustness result of [Patton \(2011\)](#) is an expected-loss result under the required conditional-unbiasedness assumptions for the volatility proxy; it is not a guarantee that a noisy proxy preserves every finite-sample ranking.

Pairwise model comparisons use the [Diebold and Mariano \(1995\)](#) test (hereafter DM test). For the full 16-comparison horse race, we adopt the conservative threshold $|t| > 3.0$, which is close to the Bonferroni-adjusted critical value of $z_{0.05/(2 \times 16)} \approx 2.95$ and consistent with the multiple-testing concerns raised by [White \(2000\)](#) and [Harvey et al. \(2016\)](#). For the primary A4f vs. GJR comparison, the standard $|t| > 1.96$ threshold applies and is comfortably exceeded.

The headline comparison is nested. GJR-GARCH is the $\theta_1 = 0$ case of A4f. Setting $\theta_1 = 0$ makes $\tau_t = \theta_0$ constant, so $u_{t-1} = r_{t-1}/\sqrt{\theta_0}$ and

$$\sigma_t^2 = \tau_t g_t = \theta_0 \omega_g + \alpha r_{t-1}^2 + \gamma r_{t-1}^2 \mathbb{I}_{[r_{t-1} < 0]} + \beta \sigma_{t-1}^2,$$

²The legacy SPY horse race consequently has values around -8.3 , whereas the corrected dimensionless diagnostic has values around 1.4 – 1.5 . Each table states its convention and compares models on a common target and sample.

which is GJR-GARCH(1,1) with intercept $\theta_0\omega_g$. A4f versus GJR is therefore a nested comparison, and nested comparisons are exactly the setting in which the standard DM statistic is known to fail: under asymptotics in which the estimation window expands with the sample, the population loss differential is identically zero under the null, the long-run variance in the denominator degenerates, and normal critical values are badly undersized. That is the problem the MSPE adjustment of [Clark and West \(2007\)](#) was designed to solve.

That failure does not arise in the design used here, for a reason that concerns the estimation scheme rather than the models. Our forecasts come from a fixed-length rolling window: $R = W = 2,000$ observations, held fixed as the evaluation sample $P = 1,826$ accumulates, with parameters refreshed every 63 trading days. This is the framework of [Giacomini and White \(2006\)](#), in which the null is equal *finite-sample* predictive ability of two forecasting *methods*—estimation error included—rather than equal predictive ability of two population models. Under that null the loss differential does not degenerate when one model nests the other, and standard normal critical values apply. [Clark and West \(2007\)](#) make the same observation about their own setting, noting that for a rolling scheme with R held fixed the relevant sample averages obey the usual law of large numbers and central limit theorem by [Giacomini and White \(2006\)](#). Section 6 reports the [Giacomini and White \(2006\)](#) conditional predictive ability test itself for this pair.

We do not report a Clark–West MSPE-adjusted statistic for the headline comparison. The adjustment of [Clark and West \(2007\)](#) is derived for mean squared prediction error, and our primary loss is the QLIKE of [Patton \(2011\)](#); transplanting the correction to a different loss function has no theoretical basis, and computing it on MSPE instead would rank the models on a criterion the rest of the paper does not use.

Two qualifications on the nesting itself. First, it is a statement about model classes rather than about the estimation region actually searched: the lower bound on θ_1 is 10^{-8} (Section 3.4), so $\theta_1 = 0$ is approached but never attained. Second, the two recursions are initialized differently—the GJR benchmark starts h_1 at the sample variance of the first 250 observations of the estimation window, whereas A4f starts g_1 at $\omega_g/(1-\alpha-\gamma/2-\beta)$ —

so the collapse above is exact for the recursion and approximate for the earliest forecasts following each refit.

3.5.2 VaR and ES Backtesting

For risk management evaluation, we compute one-day-ahead Value-at-Risk and Expected Shortfall under both Gaussian and Student- t distributional assumptions:

$$\text{VaR}_{\alpha,t}^{\text{Normal}} = \hat{\sigma}_t z_\alpha, \quad (14)$$

$$\text{VaR}_{\alpha,t}^t = \hat{\sigma}_t t_\alpha(\nu) \sqrt{\frac{\nu-2}{\nu}}, \quad (15)$$

where z_α is the standard normal quantile, $t_\alpha(\nu)$ is the Student- t quantile with ν degrees of freedom, and the scaling factor $\sqrt{(\nu-2)/\nu}$ ensures unit variance of the standardized distribution.

ES is computed as:

$$\text{ES}_{\alpha,t}^{\text{Normal}} = \hat{\sigma}_t \left(-\frac{\phi(z_\alpha)}{\alpha} \right), \quad (16)$$

$$\text{ES}_{\alpha,t}^t = \hat{\sigma}_t \left[-\frac{f_t(t_\alpha, \nu)(\nu + t_\alpha^2)}{(\nu-1)\alpha} \right] \sqrt{\frac{\nu-2}{\nu}}, \quad (17)$$

where $\phi(\cdot)$ is the standard normal density and $f_t(\cdot, \nu)$ is the Student- t density.

Backtesting employs four tests:

1. **Kupiec (1995) unconditional coverage (UC)**: tests whether the violation rate equals the nominal level.
2. **Christoffersen (1998) conditional coverage (CC)**: additionally tests independence of violations.
3. **Engle & Manganelli (2004) dynamic quantile (DQ)**: tests whether violations are predictable.
4. **Acerbi & Székely (2014) Z1/Z2 tests for ES adequacy**.

A model “passes” at level α if all three VaR tests (UC, CC, DQ) have $p > 0.05$.

4 Data

4.1 Primary Sample

Our main analysis uses daily closing prices of the SPDR S&P 500 ETF Trust (SPY) and the CBOE Volatility Index (VIX) from January 3, 2005, to April 7, 2026, obtained from Yahoo Finance. After computing log returns and aligning the VIX series, the sample contains $T = 5,347$ trading days.

The out-of-sample period begins on January 2, 2019, providing $n_{\text{OOS}} = 1,826$ daily observations. This period encompasses the COVID-19 crash of March 2020 (VIX peak of 82.69), the post-pandemic recovery, the 2022 inflation-driven drawdown, and the 2025 trade-war volatility spike, providing a demanding test environment that includes multiple volatility regimes.

Table 2 reports summary statistics.

Table 2: Summary statistics for SPY daily log returns and VIX

Statistic	Full Sample		OOS (2019–2026)	
	SPY r_t	VIX	SPY r_t	VIX
Mean (ann.)	—	18.97	—	22.14
Std (ann.)	0.188	—	0.213	—
Skewness	−0.37	2.15	−0.82	1.89
Kurtosis	14.52	9.87	12.31	7.03
Min	−0.128	9.14	−0.128	12.10
Max	0.116	82.69	0.093	82.69
N	5,347	5,347	1,826	1,826

SPY log returns and VIX closing levels. Annualized mean and standard deviation for returns. Kurtosis is excess kurtosis. Data source: Yahoo Finance.

4.2 Cross-Asset Sample

For cross-asset validation (Section 5.3), we additionally use:

- **QQQ** (Invesco QQQ Trust): high-beta U.S. technology exposure, high VIX correlation.
- **EEM** (iShares MSCI Emerging Markets ETF): moderate VIX correlation.
- **GLD** (SPDR Gold Shares): low VIX correlation. We additionally use the CBOE Gold ETF Volatility Index (GVZ) as an asset-specific fear index.
- **0050.TW** (Yuanta/P-shares Taiwan Top 50 ETF): Taiwan equity market with VIX lag of one additional day due to time-zone differences.

All data originate from Yahoo Finance for the same sample period, but the reproduction code reads local, hash-pinned adjusted-close snapshots by default; a live download is available only through an explicit fallback option. The 0050.TW series requires an explicit repair for the 4:1 split of 2014-01-02 in affected Yahoo Finance vintages, which back-adjust only from the split date forward and leave the pre-2014 history quoted in pre-split units. Switching to the adjusted-close column does not remove that vintage’s break, because the adjusted series jumps by the same factor as the raw close; the discontinuity is a level misalignment, not an adjustment error. For affected live vintages, we rescale the entire pre-2014-01-02 segment by the split factor and verify that no close-to-close ratio near a common split factor remains anywhere in the series. The pinned adjusted-close input used for the reported cross-asset results is already level-consistent, so this repair check makes no further changes. This is a level repair applied before any estimation window is formed, and is therefore stricter than the return-series exclusion used by [Lai \(2026\)](#), whose sample-window quantities do not depend on pre-2014 price levels. All 0050.TW results reported below, including the Diebold–Mariano statistics in Table 6, are computed on the level-consistent series.

For the 0050.TW cross-asset results, we read the adjusted-close columns for 0050.TW and for VIX from the hash-pinned 2008–2026 snapshot distributed with the replication

package. A shorter 2007–2022 file retained alongside the manuscript is not an input to any table; its only use is the pinned recomputation of the Taiwan local-index run reported after Table 7.

5 Empirical Results

5.1 Specification Horse Race

Table 3 presents the complete out-of-sample results for all 17 specifications, ranked by QLIKE loss.

Table 3: Out-of-sample specification comparison: QLIKE rankings

Rank	Model	QLIKE	DM t vs GJR	Screen $ t > 3.0$	Category
1	A4f (VIX^2 , free ω)	−8.358	4.47	Yes	GARCH-X
2	A4 (VIX^2 , constr.)	−8.355	4.39	Yes	GARCH-X
3	A2 (log-exp, τ_t)	−8.354	3.55	Yes	GARCH-X
4	B1 (MIDAS-RW, $K=22$)	−8.352	3.49	Yes	MIDAS-RW
5	A3 (log-exp, τ_{t-1})	−8.351	3.37	Yes	GARCH-X
6	A2f (log-exp, free ω)	−8.351	3.40	Yes	GARCH-X
7	A3f (τ_{t-1} , free ω)	−8.349	3.32	Yes	GARCH-X
8	A2n (log-exp, norm)	−8.348	3.40	Yes	GARCH-X
9	A4n (VIX^2 , norm)	−8.348	3.96	Yes	GARCH-X
10	A1 (original)	−8.347	3.26	Yes	GARCH-X
11	B2 (MIDAS-RW, $K=65$)	−8.340	3.30	Yes	MIDAS-RW
12	B3 (MIDAS-RW, $K=125$)	−8.335	3.40	Yes	MIDAS-RW
13	A5 (VIX level)	−8.330	2.10	No	GARCH-X
14	C3 (MIDAS-FS, $K_m=24$)	−8.305	2.27	No	MIDAS-FS
15	C1 (MIDAS-FS, $K_m=6$)	−8.304	2.28	No	MIDAS-FS
16	C2 (MIDAS-FS, $K_m=12$)	−8.300	1.85	No	MIDAS-FS
17	B0 (GJR-GARCH)	−8.268	ref	—	Benchmark

QLIKE evaluated on r_t^2 following [Patton \(2011\)](#). DM t is the Diebold-Mariano test statistic against GJR-GARCH (B0); positive values indicate the model is better. This paper’s conservative reporting screen (following [Harvey et al. \(2016\)](#)) requires $|t| > 3.0$. OOS period: 2019-01-02 to 2026-04-08, $n = 1,826$. Estimation window $W = 2,000$, refit every 63 days.

Several findings emerge from Table 3.

Finding 1: VIX-squared ranks first, but not distinguishably so. The top two models both use $\tau_t = \theta_0 + \theta_1 \text{VIX}_{t-1}^2$, which has a natural reading: VIX is quoted in annualized volatility units (percent), so VIX^2 has annualized variance units, directly matching the dimension of $\sigma_t^2 = \tau_t \times g_t$. The ranking does not license a stronger word than “ranks”. The best VIX-squared specification (A4) and the best log-exponential one (A2) are separated by 0.0012 in mean QLIKE and by a DM statistic of $t = 0.16$ ($p = 0.88$): on

this sample the two functional forms are indistinguishable, and the dimensional argument is a reason to prefer VIX^2 *a priori*, not a result read off the horse race. What the horse race does separate is the VIX-level specification (A5), which is the weakest GARCH-X variant and fails to reach our $|t| > 3.0$ reporting screen against GJR ($t = 2.10$).

Finding 2: Contemporaneous normalization (τ_t) ranks above lagged (τ_{t-1}), within noise. Comparing A2 vs. A3 (both log-exp, constrained), A2 achieves the lower QLIKE (-8.354 vs. -8.351), an ordering consistent with the logic of Engle et al. (2013, Eq. 4), where $u_{t-1} = r_{t-1}/\sqrt{\tau_t}$ uses the *current* environment to normalize past shocks. The difference is 0.0030 in mean QLIKE with $t = -0.87$ ($p = 0.38$), so it is a consistent ranking rather than a detected difference. We report the timing convention as a design choice with a rationale, not as a dimension the data adjudicates.

Finding 3: Free omega moves the VIX-squared model up the ranking, not beyond the noise. Comparing A4f vs. A4, freeing ω_g lowers QLIKE from -8.355 to -8.358 and raises the DM statistic against GJR from 4.39 to 4.47. The direct A4f-vs-A4 test, however, gives $t = 0.78$ ($p = 0.43$): the two are not distinguishable, and the 0.0029 QLIKE gap sits in the third decimal place. The direction is consistent with a positive variance risk premium— VIX^2 systematically exceeds realized variance, so pinning the intercept imposes a normalization the data does not want—but the horse race does not establish it. Free omega does not even move the ranking for the log-exp model (A2f vs. A2: -8.351 vs. -8.354 , $t = -1.38$, $p = 0.17$), which is consistent with the log-exp parameterization already absorbing that flexibility through θ_0 . Note also that in the free case the scale of τ_t is unidentified (Proposition 2), so “freeing ω_g ” changes what the model can fit, not what any single parameter estimates.

Finding 4: The best daily GARCH-X point estimate outperforms GARCH-MIDAS, but the category rankings overlap. A4f ranks ahead of every GARCH-MIDAS variant in Table 3, and several other daily specifications rank ahead of the best MIDAS model, B1. The weaker daily VIX-level specification A5, however, ranks below B1–B3, so this is not category-wide dominance. The Model Confidence Set analysis (Section 6) likewise shows that A4f is indistinguishable from B1 ($K = 22$) in pairwise

DM tests ($t = -0.90$). The point-estimate advantage of A4f is consistent with VIX information being concentrated in the most recent observation, which the MIDAS Beta-polynomial weighting dilutes across 22–125 lags, but parsimony rather than statistical dominance is the practical argument for preferring the daily specification. Figure 1 shows where A4f’s advantage over GJR-GARCH accumulates across the out-of-sample window.

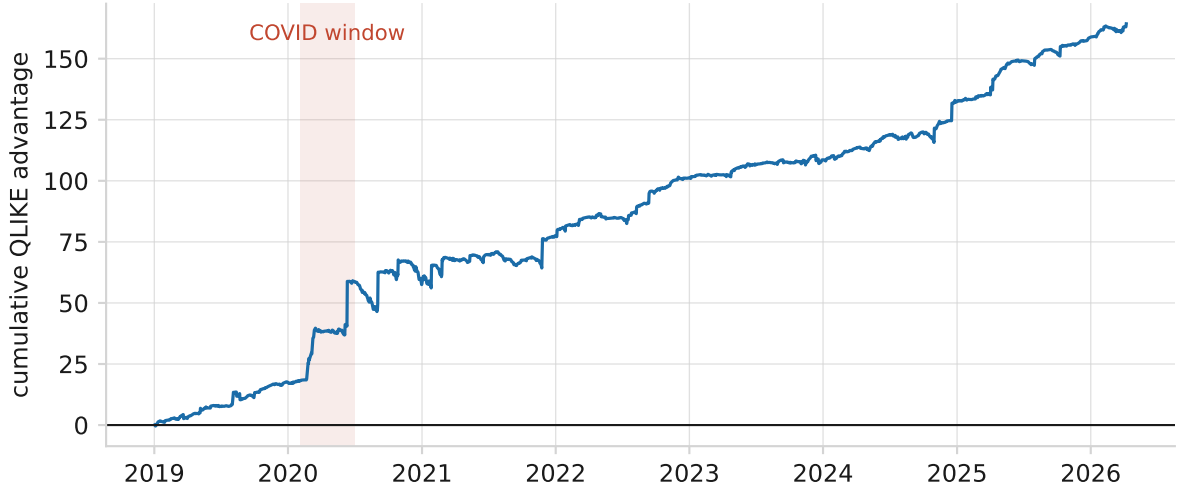


Figure 1: **Where the forecasting advantage accumulates.** Running sum of the daily QLIKE loss differential, GJR-GARCH minus A4f, over the 1,826-day out-of-sample window; a rising line means A4f is gaining. The total is 164.3 loss units, of which the shaded COVID window (2020-02-01 to 2020-06-30) contributes 24.7%. A4f has the lower loss on 971 of 1,826 days, or 53.2%. The picture is what the Diebold–Mariano statistic cannot show: three quarters of the advantage accrues outside the shaded window, so the result is not an artifact of the COVID episode, but the accumulation is not uniform either—a five-month window contributes a quarter of the total, and the daily win rate is only a few points above even.

Finding 5: The A1 specification reveals the cost of inconsistency. The original A1 specification, which uses τ_t in estimation but τ_{t-1} in OOS forecasting, ranks 10th. Adopting a consistent denominator (either τ_t everywhere as in A2, or τ_{t-1} everywhere as in A3) moves the specification up the ranking. The gap in mean QLIKE is 0.0061 against A2 and 0.0031 against A3, and the archived pairwise tests on those two comparisons do

not clear our reporting screen: $t = 1.84$ ($p = 0.065$) for A1 against A2 and $t = 0.69$ ($p = 0.49$) for A1 against A3. We therefore report the ordering and not a demonstrated cost.

5.2 Source Decomposition and the Variance Risk Premium

The multiplicative structure $\sigma_t^2 = \tau_t \times g_t$ admits an economic interpretation when $\tau_t \approx \theta_0 + \theta_1 \text{VIX}_{t-1}^2$. In this case:

$$g_t = \frac{\sigma_t^2}{\tau_t} \approx \frac{\text{realized variance}}{\text{function of implied variance}}. \quad (18)$$

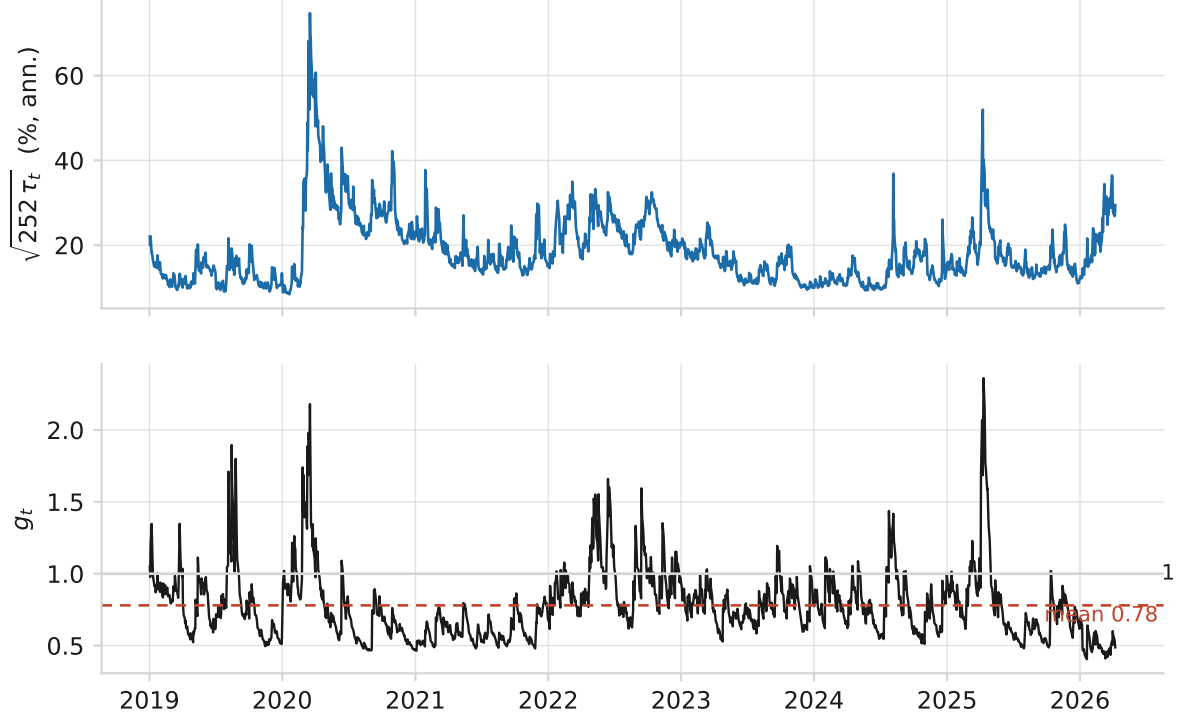


Figure 2: **The two components, SPY out-of-sample.** Upper panel: the long-run component of the constrained VIX-squared specification (A4), reported as an annualized volatility $\sqrt{252 \tau_t}$; it ranges from 8.4% to 74.7%. Lower panel: the short-run component g_t , which averages 0.78 rather than the unity that the intercept restriction would deliver as a moment condition (Section 7). We plot the *constrained* fit because its ω_g is pinned, so the scale of τ_t is identified. In the recommended free- ω_g specification it is not (Proposition 2), and the difference is visible in the range: across this same window A4's τ_t moves by a factor of 79 while A4f's moves by a factor of 1201, with $\sigma_t^2 = \tau_t g_t$ unchanged. Plotting A4f's components would put that relocation on the page as if it were the fear level.

Figure 2 plots the two fitted components for SPY out of sample. In the constrained case, τ_t determines the unconditional variance up to $\mathbb{E}[g_t]$ and the covariance term of Equation 20, and g_t fluctuates around its own mean, rising when realized variance exceeds what VIX implies and falling when VIX over-estimates it. We deliberately do not read those movements as tracking the variance risk premium: Section 5.2 shows that no

specification’s g_t correlates with VRP beyond $|\rho| = 0.089$. In the free case, the model accommodates the average level of the wedge between implied and realized variance through ω_g , at the cost of leaving the scale of τ_t unidentified (Proposition 2).

We label this a *source decomposition* rather than the traditional *long-run/short-run decomposition* used in the GARCH-MIDAS literature. In our framework:

- $\tau_t = \textit{exogenous source}$: the aggregate fear level determined by option market participants, incorporating forward-looking information from across the options surface.
- $g_t = \textit{endogenous source}$: the GARCH dynamics that capture how rapidly realized volatility responds to—and deviates from—the options-market consensus.

This interpretation differs from “long-run vs. short-run” because τ_t updates daily (it is not a slow-moving trend) and because both components have comparable persistence.

5.2.1 VRP Tracking: A Null Result

To quantify the connection between g_t and VRP, we define the VRP proxy as:

$$\text{VRP}_t = \frac{1}{252} \left(\frac{\text{VIX}_{t-1}}{100} \right)^2 - r_t^2, \quad (19)$$

using a single-day squared return as the realized-variance leg, in the spirit of [Bollerslev et al. \(2009\)](#). The division by 100 converts VIX from annualized percentage points to decimal volatility, matching the units of r_t^2 , and the division by 252 converts the annualized implied variance to a daily scale. Both conversions matter: without the factor of 100^2 the implied leg is roughly four orders of magnitude larger than r_t^2 , the realized leg cannot affect the ranking of the resulting series, and the proxy is never negative on any day of our OOS sample—it is then a relabelling of lagged implied variance rather than a premium. Both statements are measured, not asserted: on the mis-scaled proxy the rank correlation with its own implied leg is 0.999999 and the count of negative days is zero out of 1,827, while on the correctly scaled proxy that correlation falls to 0.390 and 18.3% of out-of-sample days (335 of 1,827) have $\text{VRP}_t < 0$.

Table 4 reports the Spearman rank correlation between each specification’s fitted g_t path and this VRP proxy, computed over the OOS period.

Table 4: Spearman correlation between model g_t component and VRP proxy

Measure	Spearman ρ	p -value
g_t (A3f: τ_{t-1} , free ω)	−0.089	< 0.001
g_t (A2n: log-exp, normalized)	−0.001	0.973
g_t (A4n: VIX^2 , normalized)	0.018	0.444
g_t (A4f: VIX^2 , free ω , <i>recommended</i>)	−0.011	0.649

VRP proxy: Equation 19. OOS period: 2019–2026, $n = 1,827$ for every row. All four rows are computed by one routine on one pinned sample, correlating the *fitted* g_t recursion—not a $\hat{\sigma}_t^2/VIX_{t-1}^2$ approximation to it—against one proxy. Spearman ρ is used for robustness to outliers. Source: the archived cross-asset results.

An independent implementation, which uses the same correctly scaled proxy and its own estimation path, returns $\rho = -0.006$ ($p = 0.814$, $n = 1,826$) for A4f, corroborating this table’s −0.011.

No specification’s g_t tracks the variance risk premium. The largest correlation in absolute value is A3f’s $\rho = -0.089$; it is statistically distinguishable from zero at this sample size ($p < 0.001$, $n = 1,827$) but economically negligible, and the remaining three specifications are indistinguishable from zero on any criterion. The near-orthogonality is a general result across all four specifications, including the two sample-mean-normalized specifications (A2n, A4n) whose intercept is restricted rather than free. In those two the restriction $\omega_g = 1 - \alpha - \gamma/2 - \beta$ delivers $\mathbb{E}[g_t] = 1$ only in the stationary sense; the fitted g_t paths average 0.934 (A2n) and 0.945 (A4n) over this sample, so unity is an implication of the parameterization, not a property imposed on the realized path.

We report this reversal explicitly because the earlier finding was a headline claim of this paper. It does not affect any forecasting result: the horse race, the cross-asset tests, the VaR and ES backtests, and the residual diagnostics never used the VRP proxy. What it removes is the *contrast* between specifications in how they absorb VRP, and with

it the reading that g_t carries a recoverable premium signal. The source-decomposition interpretation of Section 7 survives in its weaker and better-supported form: τ_t carries the option-market fear level and g_t carries endogenous GARCH dynamics, but g_t is not a VRP proxy. What remains is a disciplined framework for decomposing realized variance into option-market expectations and residual dynamics, with no claim that either component recovers the premium.

The two legs of the proxy are measured over different horizons; matching them does not change the answer. Equation (19) sets a daily implied variance derived from VIX—which prices the average variance over the next 30 calendar days—against a *single* day’s squared return. The legs are dimensionally matched but not horizon-matched. Table 5 repeats the exercise with the realised leg measured over the same span: the mean of r_t^2 over the next 22 trading days, the trading-day count that VIX’s window covers. That leg is forward-looking by construction, as is the floating leg of a variance swap; it is used only to *measure* the premium and enters no forecast, and no g_t path in this table sees a future return. The 21 out-of-sample days without a complete forward window are dropped rather than padded.

The null survives. The largest correlation in absolute value rises from 0.089 to 0.136, again for A3f and again with a small p -value at this sample size, and the other three specifications remain below 0.064. Horizon matching moves the estimates in the direction one would expect if a little of the premium were being averaged away at $h = 1$, but not far enough to change what the table says: g_t is not a variance risk premium object at either horizon.

Table 5: Horizon-matched VRP correlation: g_t against a 22-day realised leg

Specification	Spearman ρ	p	n
A3f (log-exp, τ_{t-1} , free ω)	-0.136	0.000	1,806
A2n (log-exp, norm)	0.054	0.022	1,806
A4n (VIX ² , norm)	0.064	0.006	1,806
A4f (VIX ² , free ω)	0.029	0.217	1,806

Implied leg $(\text{VIX}_{t-1}/100)^2/252$ as in Equation 19; realised leg is the mean of r_t^2 over trading days t through $t + 21$.

Correlated against the same fitted g_t recursions as Table 4.

Source: the archived cross-asset results.

The realised leg is forward-looking by construction and is a measurement, not a predictor; it enters no forecast in this paper.

5.2.2 VRP Predictability: A Null Result

A separate question, which we posed before establishing the contemporaneous null above, is whether g_t can *predict* future VRP. These tests were run on the same correctly scaled proxy (Equation 19) in that independent implementation and are unaffected by the correction, so we report them as originally designed. We test predictability using four approaches:

1. **Granger causality:** g_t Granger-causes VRP at horizons $h = 1, 5, 10$ days ($p < 0.001$) but not at $h = 22$ days ($p = 0.100$).
2. **Predictive regressions:** In specifications controlling for lagged VRP, the [Newey and West \(1987\)](#) t -statistic on g_t ranges from -1.54 to -2.11 . Augmenting the regression with additional controls (log-VIX, lagged returns) moves it to -2.43 at the $h = 5$ horizon, but no specification surpasses our $|t| > 3.0$ reporting screen.
3. **OOS R^2 :** Following [Campbell and Thompson \(2008\)](#), the OOS R^2 of g_t -based forecasts relative to the historical mean is 2.1% at $h = 1$ and falls to approximately zero or below at longer horizons; the [Clark and West \(2007\)](#) MSPE-adjusted statistic

against the historical mean is 0.38 at $h = 1$ ($p = 0.35$), so this margin is not distinguishable from noise.

4. **Variance swap strategy:** A strategy that sells variance when g_t is above its expanding median and buys when it is below produces an annualized Sharpe ratio of -0.88 , compared with $+0.84$ for the naive always-sell benchmark. The signal is lagged one day (`signal.shift(1)`).

We report this null result as an honest finding: g_t **has no exploitable predictive power for future VRP**. Together with the contemporaneous null of Table 4, the conclusion is the stronger one that g_t is not a VRP object at all—neither a contemporaneous proxy nor a predictor. This is consistent with the low daily autocorrelation of our VRP proxy ($\rho = 0.203$, measured on the OOS sample), which leaves little persistent variation for any predictor to capture at daily horizons.

All figures in this subsection come from a single run of that independent implementation on the pinned snapshot ($n_{\text{OOS}} = 1,826$), and its contemporaneous Spearman correlation on that run is $\rho = -0.006$ ($p = 0.814$). All four available estimates of this correlation lie within $|\rho| < 0.04$.

5.3 Cross-Asset Validation

Table 6 reports the performance of A4f across the six external markets planned beyond SPY, five of which were completed (STOXX50E could not be evaluated on pinned data; see the table’s footnotes).

Table 6: Cross-asset validation of A4f specification

Asset	VIX- r^2 corr.	GJR QLIKE	A4f QLIKE	DM t	Screen $ t > 3.0$
SPY (primary)	0.52	-8.268	-8.358	4.47	Yes
QQQ	0.49	-7.703	-7.764	3.96	Yes
GLD (with GVZ)	0.38 ^a	-8.193	-8.274	2.93	No
STOXX50E ^b	—	—	—	—	—
FEZ	0.44	-7.872	-7.907	2.74	No
EEM	0.50	-7.911	-7.944	2.69	No
0050.TW	0.28	-7.735	-7.762	1.54	No

DM t : positive values indicate A4f has lower (better) QLIKE than GJR. The screen column reports this paper’s own conservative $|t| > 3.0$ reporting screen (Section 3.5), not a critical value defined by any cited source. OOS: 2019–2026. For GLD, A4f uses GVZ: $\tau_t = \theta_0 + \theta_1 \text{GVZ}_{t-1}^2$. FEZ uses US VIX (not VSTOXX); see the discussion following this table for how strongly that choice is actually supported by the results.

Every row reports the same QLIKE kernel ($\log(\hat{\sigma}^2) + r^2/\hat{\sigma}^2$ on decimal log returns), so magnitudes are comparable across assets as well as within one; each DM test compares GJR and A4f on the same asset, target, and convention.

^a GVZ- r^2 correlation for GLD.

^b STOXX50E has no pinned price snapshot in the paper’s data packet, so it could not be evaluated; the row is listed to keep the planned set of markets visible. FEZ, the EURO STOXX 50 ETF, does have a snapshot and is reported in its place.

Multiple-testing note: Six external-market validation tests were pre-registered beyond the primary SPY result; STOXX50E’s is not currently reproducible, leaving five completed tests (QQQ, GLD, FEZ, EEM, 0050.TW). The two-sided 5% Bonferroni-adjusted normal threshold for these five is $|t| > z_{1-0.05/(2 \times 5)} \approx 2.58$. QQQ ($t = 3.96$) clears both that threshold and our reporting screen. GLD with GVZ ($t = 2.93$), FEZ ($t = 2.74$) and EEM ($t = 2.69$) clear the Bonferroni threshold but *not* our $|t| > 3.0$ reporting screen; 0050.TW ($t = 1.54$) clears neither. Primary SPY ($t = 4.47$) is reported separately and is not counted among the validation tests.

On pinned data the A4f specification clears our $|t| > 3.0$ reporting screen for two assets: SPY ($t = 4.47$) and QQQ ($t = 3.96$). Three further assets improve on GJR-GARCH by margins that clear the Bonferroni validation threshold but fall short of that

screen—GLD with GVZ ($t = 2.93$), FEZ ($t = 2.74$) and EEM ($t = 2.69$)—while 0050.TW ($t = 1.54$) reaches neither. We report this cluster at 2.7–2.9 as what it is: consistent in sign, insufficient in strength.

Two consequences follow. First, the earlier reading that US VIX captures a *global* fear factor rested on the European results, and with FEZ at 2.74 that evidence is weaker than claimed—the sign is right, the magnitude does not clear the paper’s own screen. Second, the pattern that survives is narrower but sharper: the model clears our reporting screen exactly where VIX is the asset’s own fear gauge (SPY, QQQ), and falls short everywhere else, including where the asset has a dedicated index (GLD with GVZ, and 0050.TW with Taiwan’s own VIXTWN in the separate run reported below). That is still consistent with the source-decomposition interpretation, which requires τ_t to reflect the asset’s own fear environment, but it no longer supports a claim of broad cross-market validity.

Table 7 examines whether asset-specific fear indices can improve the model for EEM and 0050.TW; Taiwan’s own implied volatility index, which the table’s pinned harness does not cover, is examined in a separate run reported after the table.

Table 7: Local fear index experiments for EEM, GLD, and 0050.TW

Asset	Fear Index	QLIKE	DM t vs GJR	Screen $ t > 3.0$
GLD	VIX (U.S. equity)	−8.212	1.33	No
	GVZ (gold vol)	−8.274	2.93	No
	VIX + GVZ (dual)	−8.261	2.70	No
EEM	VIX	−7.944	2.69	No
	OwnRV ₂₀	−7.910	−0.04	No
	VIX + OwnRV ₂₀	−7.945	2.92	No
0050.TW	VIX (lag+1)	−7.762	1.54	No
	OwnRV ₂₀	−7.730	−0.84	No
	VIX + OwnRV ₂₀	−7.759	1.44	No
	VIX (lag+2)	−7.757	1.42	No

OwnRV₂₀: 20-day realized volatility of the asset itself, rescaled to VIX-like annualized percent so that it enters τ_t on the same scale as an implied volatility index. Dual factor: $\tau_t = \theta_0 + \theta_1 X_1^2 + \theta_2 X_2^2$. DM t : positive values indicate the row’s specification has lower (better) QLIKE than GJR-GARCH. The screen column reports this paper’s own conservative $|t| > 3.0$ reporting screen. OOS: 2019–2026.

GJR-GARCH baseline QLIKE for comparison: −8.193 (GLD), −7.911 (EEM), −7.735 (0050.TW). Sample sizes are $n = 1,823$ (GLD), 1,815 (EEM) and 1,650 (0050.TW).

All rows were recomputed on the pinned adjusted-close snapshots on 2026-08-24 in the same QLIKE kernel as Table 6 ($\log(\hat{\sigma}^2) + r^2/\hat{\sigma}^2$ on decimal log returns). Each asset’s Table 6 entry is reproduced exactly by the corresponding row here: the VIX row for EEM, the VIX (lag+1) row for 0050.TW, and the GVZ row for GLD (Table 6 evaluates GLD against its own fear index, not against U.S. VIX). Source: the archived cross-asset results.

The GLD result is instructive about *relevance* without clearing our *screen*: switching from U.S. VIX ($t = 1.33$) to the asset-specific GVZ ($t = 2.93$) more than doubles the DM statistic, which supports the requirement that the multiplicative structure be fed a

relevant implied volatility measure—but the improved specification still does not clear the $|t| > 3.0$ reporting screen this paper applies throughout, and adding VIX alongside GVZ does not help ($t = 2.70$). We therefore read gold as evidence on the mechanism, not as a second confirmation of the forecasting claim. The EEM and 0050.TW rows point the same way, and sharpen the mechanism. Own-asset realized volatility carries no usable information on its own: substituting OwnRV_{20} for VIX destroys the improvement entirely for both assets ($t = -0.04$ for EEM, $t = -0.84$ for 0050.TW), so what the multiplicative structure needs is a forward-looking option-implied measure, not any volatility proxy. Adding OwnRV_{20} alongside VIX moves EEM from $t = 2.69$ to $t = 2.92$ and moves 0050.TW slightly the other way (1.54 to 1.44); neither reaches the reporting screen. Delaying Taiwan’s VIX input by a second day, which would be required if the previous U.S. session’s close were not yet actionable, leaves the result essentially unchanged ($t = 1.42$). No specification in this table clears $|t| > 3.0$. For EEM we did not test a dedicated index, so the EEM rows speak only to U.S. VIX and to the asset’s own realized volatility. For 0050.TW a dedicated index exists and we tested it. The Taiwan Futures Exchange has published VIXTWN, a 30-day implied volatility index on TAIEX options built on the CBOE methodology, since 2006, and we examined it in a separate run that predates the pinned harness behind Table 7; we report that run here rather than as a table row because its evaluation window differs. It evaluates 0050.TW over 2013–2021 ($n = 2,199$ one-day-ahead forecasts from a rolling 1,000-day estimation window re-estimated every 63 days), so its statistics are not comparable cell by cell with the 2019–2026 rows above; the direction, however, is unambiguous. Feeding lagged squared VIXTWN into τ_t in place of U.S. VIX gives $t = 1.86$ against GJR-GARCH, below the $t = 2.68$ that U.S. VIX attains on the same sample; the head-to-head test of the two specifications ($t = -0.99$) does not favour the local index; and using both indices ($t = 2.80$ against GJR-GARCH) adds nothing over VIX alone ($t = 0.14$ for the two-index specification against the VIX-only one). Neither index clears the screen on that sample. That run’s 0050.TW prices were a download that the replication package does not pin; re-implementing the same design inside this paper’s pinned harness (raw closes from the pinned 2007–2022 snapshot,

$n = 2,091$ joint valid days) gives $t = 1.90$ for VIXTWN, $t = 2.20$ for U.S. VIX and $t = 2.32$ for the two together, so the magnitudes are vintage-sensitive while the ordering and the failure to clear the screen are not. The two indices are highly correlated in levels ($\rho = 0.87$ over the run's sample), so the local index carries little fear information about this asset that U.S. VIX does not already supply. This is the negative control that the GLD row lacks: an asset-specific index helps when it is informative about the asset's own volatility, as GVZ is for gold, not merely because one exists.

5.4 VaR and Expected Shortfall Backtesting

Table 8 reports the VaR backtesting results for SPY at three confidence levels ($\alpha = 1\%, 2.5\%, 5\%$) under four model-distribution combinations.

Table 8: VaR backtesting results for SPY (OOS 2019–2026)

Model	α	Viol.%	Expected	UC p	CC p	DQ p	Pass?
GJR-Normal	1%	2.41	1.00	0.000	0.000	0.000	Fail
GJR-Normal	2.5%	3.61	2.50	0.004	0.016	0.068	Fail
GJR-Normal	5%	5.97	5.00	0.065	0.152	0.132	Pass
GJR- t	1%	1.64	1.00	0.012	0.033	0.014	Fail
GJR- t	2.5%	3.50	2.50	0.009	0.034	0.082	Fail
GJR- t	5%	6.19	5.00	0.024	0.059	0.017	Fail
A4f-Normal	1%	1.92	1.00	0.000	0.002	0.002	Fail
A4f-Normal	2.5%	3.23	2.50	0.055	0.121	0.006	Fail
A4f-Normal	5%	5.26	5.00	0.617	0.803	0.499	Pass
A4f- t	1%	1.53	1.00	0.034	0.079	0.195	Fail
A4f- t	2.5%	2.96	2.50	0.224	0.416	0.006	Fail
A4f- t	5%	5.59	5.00	0.259	0.453	0.227	Pass

“Pass” requires all three tests (UC, CC, DQ) to have $p > 0.05$. Viol.% is the empirical violation rate. UC: [Kupiec \(1995\)](#). CC: [Christoffersen \(1998\)](#). DQ: [Engle and Manganelli \(2004\)](#). GJR- t uses jointly estimated degrees of freedom (mean $\hat{\nu} = 6.03$). A4f- t estimates ν from standardized residuals (mean $\hat{\nu} = 8.12$); Section 6.8 treats that asymmetry as a confound rather than a nuisance. $n_{\text{OOS}} = 1,826$, on the same pinned snapshot as the horse race.

The three-test pass criterion separates the four configurations only at the 5% level, where GJR-Normal, A4f-Normal and A4f- t pass and GJR- t does not. At the 1% and 2.5% levels every configuration fails, so this exercise does not identify a model that is adequately calibrated across the whole tail.

What the table does separate is calibration distance. At the 1% level the violation rates are 2.41% (GJR-Normal), 1.92% (A4f-Normal), 1.64% (GJR- t) and 1.53% (A4f- t) against a nominal 1%; at 2.5% they are 3.61%, 3.23%, 3.50% and 2.96% against a nominal 2.5%. A4f- t is closest to nominal at both levels. Proximity is not a test, however, and we do not report it as one. The failures also differ in kind: A4f- t fails at the 1% level on

unconditional coverage alone (UC $p = 0.034$, with CC $p = 0.079$ and DQ $p = 0.195$) and at the 2.5% level on the dynamic-quantile test alone (DQ $p = 0.006$, with UC $p = 0.224$), whereas GJR- t fails at least two of the three tests at every level.

These figures come from the same hash-pinned snapshot as the horse race.

Table 9 reports the ES backtesting results at $\alpha = 2.5\%$.

Table 9: Expected Shortfall backtesting at $\alpha = 2.5\%$ (SPY, OOS 2019–2026)

Model	Z1 stat	Z1 p	Z2 stat	Z2 p	Pass?
GJR-Normal	−0.224	0.476	−0.770	0.465	Pass
GJR- t	−0.088	0.482	−0.526	0.477	Pass
A4f-Normal	−0.123	0.478	−0.451	0.475	Pass
A4f- t	−0.062	0.480	−0.257	0.468	Pass

ES tests following [Acerbi and Székely \(2014\)](#), at $\alpha = 2.5\%$.

Z1 conditions on the reported VaR being correct and tests only whether the magnitude of the losses that exceed it matches the reported ES; Z2 does not condition on VaR and tests VaR and ES jointly, so it is sensitive to both the frequency and the magnitude of tail exceedances. Both statistics are centred on zero under a correctly specified model. All four models pass both tests.

The reported p -values carry no power; see the discussion following this table.

All models pass the ES tests. We do not attribute that to the known lower power of ES backtests relative to VaR backtests, because in our implementation the ES test has *no* power: the reference distribution is obtained by resampling the realised $(r_t, \text{VaR}_t, \text{ES}_t)$ triples, so it is centred on the observed statistic itself and returns $p \approx 0.5$ for any model. All eight reported p -values fall in $[0.465, 0.482]$, which is the signature of that construction rather than evidence about calibration. A test that cannot fail cannot discriminate, so we exclude the ES column from the scorecard (Table 10) and report it as non-discriminating until the reference distribution is simulated under H_0 as [Acerbi and Székely \(2014\)](#) pre-

scribe. The scorecard therefore compares the four models on the three VaR tests alone.

Table 10: Combined VaR/ES scorecard

Model	VaR Pass (of 3)	ES (non-discriminating)	Total (of 3)
GJR-Normal	1	—	1
GJR- t	0	—	0
A4f-Normal	1	—	1
A4f-t	1	—	1

The scorecard does not separate A4f from the benchmark on pass counts: three of the four configurations pass at exactly one of the three levels, and only GJR- t passes at none. At this sample size and under this criterion the VaR evidence does not establish a tail-risk advantage, and we state it that way rather than as weak support for one. What survives is narrower: A4f- t has the violation rate closest to nominal at both levels where every specification fails, and its failures are concentrated in one test at a time rather than in all three. That is a basis for a preference between these four specifications, not a validated risk model.

6 Robustness

6.1 Sensitivity to Estimation Settings

We conduct a comprehensive sensitivity analysis across four design dimensions: refit frequency, estimation window size, OOS sub-period, and alternative VIX indices. Table 11 reports the results.

Table 11: Sensitivity analysis of A4f specification

Dimension	Setting	QLIKE (GJR)	QLIKE (A4f)	DM t	Screen $ t > 3.0$
Refit freq.	21 days	1.500	1.409	4.29	Yes
	63 days (baseline)	1.498	1.408	3.92	Yes
	126 days	1.501	1.408	3.36	Yes
	252 days	1.509	1.410	3.32	Yes
Window	$W = 1,000$	1.478	1.418	3.18	Yes
	$W = 1,500$	1.490	1.408	3.49	Yes
	$W = 2,000$ (baseline)	1.498	1.408	3.92	Yes
	$W = 2,500$	1.491	1.402	5.13	Yes
	$W = 3,000$	1.487	1.402	4.94	Yes
Sub-period	2019–2020 (COVID)	1.522	1.408	1.59	No ^a
	2021–2022 (rate hikes)	1.384	1.303	2.49	No ^a
	2023–2026 (stable)	1.552	1.473	4.52	Yes
VIX variant	VIX (baseline)	1.498	1.408	3.92	Yes
	VIX9D	1.498	1.380	5.15	Yes
	VIX3M	1.498	1.436	2.59	No
	VIX/VIX3M ratio	1.498	1.446	3.53	Yes

^a Sub-period DM tests have reduced power ($n \approx 500$); QLIKE improvement is directionally consistent in all sub-periods.

VIX9D data begins 2011; sub-period and VIX variant tests use the shorter sample.

Data provenance. Unlike Table 3, this table is *not* computed from the hash-pinned snapshot used elsewhere in the paper. The sensitivity run that produces it downloads SPY, VIX, VIX9D and VIX3M from Yahoo Finance at run time, so every cell is vintage-dependent; the archived run is dated 2026-04-17 and evaluates $n = 1,823$ against the pinned window’s 1,826. The paper’s data packet does not carry VIX9D or VIX3M, so repinning this table requires a new snapshot for those two series and has not been done. Read the cells as a sensitivity pattern, not as figures comparable in level with Table 3; Section 6.2 places its baseline cell alongside the paper’s other A4f-versus-GJR statistics.

An independent sensitivity rerun on an earlier data vintage returns the same screen-clearing count but not the same cells; the paragraph following this table gives the cell-level comparison, which is what the count conceals.

On the sensitivity run’s vintage, A4f clears our $|t| > 3.0$ reporting screen in 13 of 16 sensitivity settings (81.3%). All four refit frequencies and all five window sizes clear the screen, so the model is not sensitive to estimation design choices along those two dimensions.

The count is stable; the cells are not. An independent rerun of the same sixteen settings on a data vintage five days earlier also returns 13 of 16, but the agreement is between totals only. Two cells cross the screen in *opposite* directions and cancel: the 2021–2022 sub-period rises from 2.49 to 3.18, moving from No to Yes, while the VIX/VIX3M ratio falls from 3.53 to 2.81, moving from Yes to No. The rerun’s own tolerance flags three of the sixteen as divergent—those two and the 126-day refit setting ($3.36 \rightarrow 4.18$)—against a maximum absolute difference of 0.82 and a median of 0.27 across the fourteen distinct cells. We therefore read the aggregate count as evidence about the aggregate and nothing more. In particular, the VIX-ratio row above clears the screen on the sensitivity run’s vintage alone, and we do not claim it as a vintage-stable result. Notably, longer estimation windows ($W = 2,500\text{--}3,000$) strengthen the DM statistic (from 3.92 to 5.13), suggesting that A4f benefits from larger training samples.

The sub-period results show that A4f improves QLIKE in all three sub-periods, but the DM test lacks power with $n \approx 500$ observations in the COVID and rate-hike periods. This is a well-known limitation of the DM test in short samples, not evidence of model failure.

Among the VIX variants in the same table, the 9-day VIX (VIX9D) achieves the strongest result (DM $t = 5.15$), consistent with short-term implied volatility being more responsive to recent market conditions. The 3-month VIX (VIX3M) does not clear our screen ($t = 2.59$), suggesting that longer-term implied volatility changes too slowly to provide daily forecasting value.

To address the concern that A4f’s advantage may be driven by a single extreme episode (e.g., the COVID-19 crash), we report an extended sub-period analysis over seven non-overlapping windows spanning 2013–2026 (Table 12).

This analysis is run on a different setup from the rest of the paper, and the difference matters. It comes from an extended-sample run, which differs from the headline configuration in four respects that a reader should have before weighing the numbers. It begins in 2013 rather than 2019, so its evaluation sample is $n = 3,328$ rather than 1,826. It reads a live download rather than the hash-pinned snapshot used for Table 3. Both models are estimated with Student- t innovations at a *fixed* $\nu = 8$, whereas the horse race uses a Gaussian quasi-likelihood. And its losses are reported on the dimensionless Patton QLIKE of Equation (13), not the legacy kernel of Table 3. The estimation window ($W = 2,000$) and refit frequency (63 days) are the same. Section 6.2 places the resulting statistic alongside the paper’s other A4f-versus-GJR tests.

A4f attains the lower QLIKE in all seven windows, with improvements from 4.80% to 7.00% (mean 6.42%). The COVID window (P4) shows the largest improvement, and excluding it leaves the other six at 4.80–6.90%, so the size of the advantage is not COVID-dependent. The *significance* pattern is weaker than the win rate suggests and we state it rather than round it off: 3 of the seven windows clear our $|t| > 3.0$ reporting screen and 6 clear $|t| > 2.0$, and the one window that clears neither is P4—the COVID window this exercise was designed to interrogate ($t = -1.86$, $n = 505$). Seven consistent signs is a directional result, not seven significant tests. The pooled statistic over the whole 2013–2026 evaluation sample is $t = 6.977$ ($n = 3,328$), which is the largest A4f-versus-GJR statistic anywhere in this paper and is largest in part because it is computed on the largest sample. Note also that P7 covers 317 observations rather than a full two years, because the data end in 2026.

Table 12: A4f vs. GJR-GARCH across seven non-overlapping sub-periods (extended-sample setup)

Window	Years	n	QLIKE (GJR)	QLIKE (A4f)	Impr. (%)	DM t
P1	2013–2014	501	1.468	1.374	6.36	−4.38
P2	2015–2016	503	1.607	1.496	6.90	−3.49
P3	2017–2018	499	1.738	1.630	6.22	−2.12
P4	2019–2020	505	1.496	1.392	7.00	−1.86
P5	2021–2022	502	1.395	1.299	6.83	−2.82
P6	2023–2024	501	1.508	1.435	4.80	−2.69
P7	2025–2026	317	1.645	1.532	6.84	−3.29

Source: the archived results of the extended-sample run. Negative DM t favours A4f under that run’s convention (loss differential is A4f minus GJR). Patton QLIKE of Equation (13); both models carry Student- t innovations with $\nu = 8$ fixed; live vintage of 2026-04-17; $W = 2,000$, refit every 63 days. These figures are therefore not comparable in level with Table 3.

P7 ends with the data and covers 317 observations rather than two full years.

A complementary leave-COVID-out test provides a more stringent robustness check. Rather than splitting the OOS window into two-year periods, we remove the 104 trading days spanning 2020-02-01 to 2020-06-30 from the full OOS window and re-estimate the DM test on the remaining $n = 1,721$ observations. Table 13 reports the results, computed in a dedicated baseline-faithful replication run. These statistics share the A4f specification with Table 3 but are *not* directly comparable to it. The difference is not one of data vintage. That run reads the same pinned price file as the horse race and the same adjusted-close column, and its VIX column is numerically identical to the one the horse race resolves. Two differences remain and neither is the data: its OOS window ends one trading day earlier (2026-04-07, so $n = 1,825$ on the full sample against 1,826), and it is an independent implementation of the same specification. The two agree on mean QLIKE to the third decimal (−8.267 vs. −8.268 for GJR, −8.360 vs. −8.358 for A4f) but not on the HAC-corrected statistic on the full sample ($t = 3.60$ vs. 4.47), and the DM

formula and bandwidth rule are the same in both. The divergence therefore lies in the loss paths rather than in the data or the test, and Section 6.2 decomposes it: it is a single day in the benchmark leg, and removing it reverses the ordering. We nonetheless read the leave-COVID-out result on its own terms rather than against the full-OOS benchmark. Every statistic in this paragraph and in Table 13 is accordingly to be compared against that run’s *own* full-sample statistic of $t = 3.60$, not against the headline 4.47. Within this run, the non-COVID DM statistic ($t = 4.26$, $n = 1,721$) still clears our $|t| > 3.0$ reporting screen, and post-COVID observations (2021–2026, $n = 1,448$) separately clear it as well ($t = 3.76$)—evidence that A4f’s advantage was not confined to the COVID window in that run. Within the COVID window itself ($n = 104$), the DM statistic drops to $t = 1.48$, not because A4f underperforms—the mean QLIKE improvement during this window is about $7.6\times$ larger than outside it (0.51 vs. 0.07 in mean loss differential)—but because 104 observations are insufficient for the HAC-corrected DM test to reach conventional significance. We read this as supporting evidence, on that run’s own loss paths, that A4f’s advantage is not confined to a single extreme episode; it is not a direct confirmation of the pinned Table 3 result, which this test has not yet been rerun against.

Table 13: Leave-COVID-out robustness: DM tests for A4f vs. GJR-GARCH by OOS sub-sample

Sub-sample	n	DM t	$ t > 3.0?$	Interpretation
Non-COVID (excl. COVID window)	1,721	4.26	Yes	Above the replication run’s own 3.60
Pre-COVID (2019)	273	2.52	No ^a	Short-sample
COVID window (2020-02–06)	104	1.48	No ^a	Extreme short-sample
Post-COVID (2021–2026-04)	1,448	3.76	Yes	Clears screen

Full OOS window is 2019-01-01–2026-04-07; the excluded COVID window is 2020-02-01–2020-06-30. The $|t| > 3.0$ column reports this paper’s own conservative reporting screen (see Section 5), not a critical value defined by any cited source.

A4f specification: $\tau_t = \theta_0 + \theta_1 \text{VIX}_{t-1}^2$, free ω_g , Gaussian quasi-likelihood; OOS window $W = 2,000$, refit every 63 days. DM test uses HAC-robust standard errors (Newey–West with bandwidth $q = \lfloor n^{1/3} \rfloor$, which yields $q \in \{4, 6, 11, 11, 12\}$ for the COVID-window, pre-COVID, non-COVID, post-COVID, and full-OOS sub-samples respectively). For the primary full-OOS DM result ($t = 4.47$, $n = 1,826$), see Table 3. The Interpretation column compares each sub-sample against the replication run’s *own* full-sample statistic ($t = 3.60$, $n = 1,825$), not against that primary result; the two runs share the data and the test but not the loss paths.

^a Insufficient power at $n < 300$; mean QLIKE improvement is directionally consistent across all sub-samples.

Source: the archived results of the replication run.

6.2 Reconciling the A4f-versus-GJR Statistics

This paper reports five different Diebold–Mariano statistics for what a reader may reasonably take to be one comparison. They are not one comparison. Table 14 lists them with the sample, price source, estimator and loss scale that produced each, because the spread between 3.92 and 6.977 is a statement about those choices and not about disagreement over whether A4f beats GJR-GARCH.

Table 14: Every A4f vs. GJR-GARCH DM statistic in this paper, and what produced it

Reported in	Producer	n	Price source	Estimator	$ \text{DM } t $
Table 3 (headline)	<code>mcs_dm_results</code>	1,826	pinned adj.	Gaussian QL	4.47
Table 11 (baseline cell)	sensitivity run	1,823	live download	Gaussian QL	3.92
Table 13 (non-COVID)	replication run	1,721	pinned adj.	Gaussian QL	4.26
Section 6.6	diagnostic run	1,852	pinned <i>unadj.</i>	Gaussian QL	4.37
Table 12 (pooled)	extended-sample run	3,328	live download	Student- t , $\nu=8$	6.977

Magnitudes are reported because the sign convention differs across the producing runs; every entry favours A4f. “Pinned” means a hash-fixed snapshot archived with the replication package, with adjusted closes unless marked *unadj.*; “live” means the producing script downloaded prices at run time and its value is therefore vintage-dependent.

The headline row is the one the paper’s conclusions rest on. The others answer narrower questions—sensitivity to design settings, the COVID window, a HAR-style benchmark, and pre-2019 regimes—and are reported on their own terms.

The largest statistic (6.977) is also the one computed on the largest sample and the only one estimated under a fixed Student- t likelihood; it is not a stronger version of the headline result.

Two rows are not on the pinned Table 3 sample by construction: the extended-sample run starts in 2013 and the sensitivity run predates the pinning. Repinning both is an open item.

The one pair that shares the data and the test and still disagrees. Four of the five rows differ in sample, price source or estimator, so their spread needs no further explanation. The headline row and the replication-run row do not: they read the same pinned file and the same adjusted-close column, use the same estimator *specification*, the same loss function, the same DM formula and the same bandwidth rule, and their mean QLIKEs agree to the third decimal. What they do not share is the code: the replication run is an independent implementation with its own fitting routine, as noted above. Yet the statistics are 4.47 and 3.60, and on the 1,825 days the two runs share they are 4.44 and 3.60. Neither run originally stored a daily series, so the question could not be answered from the archived files; both now persist their per-day losses, and the reconciliation routine aligns them.

The answer is short. On the 1,825 overlapping days the two A4f loss paths are

effectively the same series—their largest daily difference is 0.011 and exactly one day differs by more than 0.01—while the two GJR paths differ on 801 days, by as much as 17.33. The disagreement is entirely in the benchmark leg. It reaches the test statistic through the denominator rather than the numerator: the mean loss differential is nearly identical, but its standard deviation is 0.86 in the headline reconstruction against 1.11 in the replication, and a wider d_t is a smaller t .

One observation carries it. On 2020-06-11 the two implementations assign the GJR benchmark a QLIKE of 29.49 and 46.82 respectively, while agreeing exactly on A4f (11.15 in both). Excluding that single day, the statistics are 4.59 and 4.69—the gap is not merely reduced but reversed, and the headline becomes the *lower* of the two. We report this because it constrains how much either number should be read into: the published 4.47 is not flattered by the divergence, but a difference of this size resting on one day in the benchmark’s variance recursion is a reminder that a DM statistic on 1,826 daily observations is an estimate with its own dispersion, not a measurement. We do not identify the source of the 801 differing benchmark days, but the manuscript already contains a candidate: the two recursions are initialized differently after each refit—the GJR benchmark starts h_1 at the sample variance of the first 250 observations of the estimation window while A4f starts g_1 at $\omega_g/(1 - \alpha - \gamma/2 - \beta)$ (Section 3.5)—so the earliest forecasts following a refit are the ones most exposed to implementation choices. We name it as a candidate rather than a finding, because this exercise does not test it. What it does not do is disturb the comparison itself: both implementations clear the paper’s $|t| > 3.0$ screen with and without that day, and on the overlapping window the two A4f loss paths have mean QLIKEs of -8.359713 and -8.359703 , a difference of one part in 10^6 .

6.3 Estimation/Forecasting Consistency

A1 (inconsistent denominator) ranks 10th while A2 and A3 (consistent) rank 3rd and 5th, so using the same τ normalization in estimation and forecasting is associated with a better rank. We put it no more strongly than that: the A1-versus-A2 gap is 0.0061 in mean

QLIKE, the same third-decimal magnitude Section 8 declines to read as a separation, and the pairwise DM tests on this dimension do not clear our reporting screen either ($t = 1.84$, $p = 0.065$ against A2; $t = 0.69$, $p = 0.49$ against A3). A1 itself still clears our reporting screen against GJR-GARCH ($t = 3.26$), so what the ordering shows is a rank cost, not a failure of the inconsistent specification.

6.4 Model Confidence Set

To assess whether A4f can be statistically distinguished from competing specifications, we apply the Model Confidence Set (MCS) procedure of Hansen et al. (2011) with 1,000 bootstrap replications. At $\alpha = 0.10$, all 17 models survive—the pairwise QLIKE differences among the VIX-augmented models are too small to reject any specification. At $\alpha = 0.25$, the MCS eliminates only GJR-GARCH ($p = 0.154$); all 16 VIX-augmented models remain.

This result qualifies our ranking-based conclusions: while A4f achieves the lowest QLIKE, it is not *statistically* distinguishable from other well-specified GARCH-X or GARCH-MIDAS models incorporating VIX. A4f and a majority of the VIX-augmented specifications exceed the $|t| > 3$ pairwise screen against GJR-GARCH, but the weaker specifications do not; the MCS result is a set-level statement, not evidence that every VIX model wins pairwise. A4f is the point-estimate best among the alternatives considered.

6.5 Pairwise DM Tests Among Top Models

Table 15 reports pairwise DM tests between A4f and selected competitors to clarify which differences clear our reporting screen.

Table 15: Pairwise DM tests: A4f vs. selected competitors

Comparison	DM t	Screen $ t > 3.0$	Interpretation
A4f vs. GJR (B0)	-4.47	Yes	Clears the screen against B0
A4f vs. A4 (constr.)	-0.78	No	Free ω adds little
A4f vs. A2 (log-exp)	-0.62	No	Functional form indistinguishable
A4f vs. B1 (MIDAS-RW-22)	-0.90	No	Best MIDAS $\approx$ A4f
A4f vs. B2 (MIDAS-RW-65)	-3.06	Yes	A4f $>$ longer-lag MIDAS
A4f vs. B3 (MIDAS-RW-125)	-3.75	Yes	A4f $>$ longest-lag MIDAS
A4f vs. C1 (MIDAS-FS-6)	-2.38	No	Borderline

The loss differential is first model minus second model, so negative t favors the first model. DM test against our conservative $|t| > 3.0$ reporting screen, following [Harvey et al. \(2016\)](#).

A4f outperforms GJR-GARCH and the longer-lag MIDAS specifications (B2, B3) by margins that clear our $|t| > 3.0$ screen, but is statistically indistinguishable from the best-performing GARCH-X alternatives (A4, A2) and the shortest-lag MIDAS (B1, $K = 22$). This confirms the MCS finding: the key distinction is VIX-augmented vs. plain GARCH, not the specific functional form of the VIX component.

The [Giacomini and White \(2006\)](#) conditional predictive ability test further supports this interpretation: A4f conditionally outperforms GJR ($\chi^2 = 20.20$, $p < 0.001$) but not B1 ($\chi^2 = 3.75$, $p = 0.154$). This test carries additional weight for the A4f-versus-GJR pair specifically, because that pair is nested (Section 3.5); it is constructed for the same fixed-window forecasting-method null that licenses normal critical values for the unconditional DM statistic here, so the two results rest on one framework rather than two.

6.6 HAR-Style Daily- r^2 Benchmark Comparison

A canonical HAR-RV benchmark uses intraday realized variance ([Corsi, 2009](#)), which is not available in the pinned daily-data packet used here. We therefore report a narrower diagnostic: HAR-style regressions whose daily, weekly, and monthly components are

formed from lagged squared daily returns, with and without VIX_{t-1}^2 . These models are not canonical HAR-RV specifications and cannot close the realized-measure benchmark question by themselves. All models are evaluated against the same r_t^2 proxy using Patton QLIKE, $L_t = r_t^2/\hat{\sigma}_t^2 - \log(r_t^2/\hat{\sigma}_t^2) - 1$.

The corrected diagnostic run uses a hash-pinned unique-date SPY/VIX snapshot, fixes the original endpoint at 18 May 2026, and applies the same rolling protocol ($W = 2,000$, refit every 63 days). There are 1,852 shared valid OOS losses. A4f has mean QLIKE 1.400, compared with 1.480 for GJR and 1.524 for the HAR-style daily- r^2 model, reductions of 5.4% and 8.2%, respectively. Here the loss differential is $d_t = L_{A4f,t} - L_{\text{comparison},t}$, so negative t favors A4f. This is the opposite of Table 3, whose against-GJR column reports GJR loss minus model loss. Bartlett-HAC DM tests with the canonical rule-selected maximum lag of 13 give:

- A4f vs. GJR: DM $t = -4.37$, $p = 0.000013$;
- A4f vs. HAR-style daily- r^2 : DM $t = -7.70$, $p < 10^{-13}$.

Both comparisons exceed our $|t| > 3$ reporting screen, adopted following [Harvey et al. \(2016\)](#). The two A4f comparisons also remain beyond that screen over HAC lags 0, 1, 5, 10, 13, and 20. By contrast, the HAR-style daily- r^2 minus GJR comparison is bandwidth-sensitive: its positive statistic favors GJR, and its canonical lag-13 value is $t = 3.11$, versus 2.94 and 2.97 at lags 0 and 1.

The VIX-augmented HAR-style OLS specification is numerically unstable: it produces three nonpositive raw variance forecasts, which are bounded at 10^{-16} before QLIKE and make its full-sample mean loss floor-dependent. Its A4f comparison does not cross that screen either in the bounded primary run ($t = -1.05$, $p = 0.294$) or in the diagnostic common sample excluding those three dates ($t = -2.26$, $p = 0.024$). We therefore do not interpret its extreme full-sample mean loss as an economic effect.

Three features of this comparison are worth stating, because each would change the statistics if handled differently: the QLIKE ratio is oriented so that a negative differential favours A4f, the standard errors carry an HAC long-run variance, and A4f is normalized

consistently between fitting and the out-of-sample recursion. An approximation-based version of this comparison is not an independent proxy check, because it also uses daily r^2 components and a steady-state approximation for A4f; we therefore make no claim of cross-proxy consistency, parity, or statistical non-inferiority. The evidence supports lower loss for A4f relative to the stable daily- r^2 diagnostic under this fixed-window protocol; it does not establish equivalence or performance relative to canonical intraday HAR-RV, so the broader benchmark limitation remains open.

6.7 Multiple Testing Considerations

With 16 pairwise tests against GJR, 12 of 16 multiplicative models clear our conservative $|t| > 3.0$ reporting screen (following [Harvey et al. \(2016\)](#)). Our main result—A4f vs. GJR with $t = 4.47$ —exceeds the Bonferroni-adjusted critical value of $z_{0.05/(2 \times 16)} \approx 2.95$.

6.8 Student- t Degrees of Freedom, and a Confound in How They Are Estimated

The A4f- t model estimates degrees of freedom from standardized residuals $\hat{\varepsilon}_t = r_t / \hat{\sigma}_t$ in each estimation window, yielding mean $\hat{\nu} = 8.12$ (range 7.35–9.85 across the 29 refits). The GJR- t benchmark instead carries ν as a free parameter of its own likelihood, yielding mean $\hat{\nu} = 6.03$ (range 5.46–7.67).

Those two numbers are not comparable, and the VaR results of Section 5.4 inherit the incomparability. The gap between 8.12 and 6.03 mixes two effects that this design cannot separate: a model effect (the VIX component absorbs tail behaviour that GJR must carry through fat-tailed innovations) and a method effect (joint maximum likelihood trades ν against the variance recursion, while a two-step estimate on standardized residuals does not). We therefore do not read the gap as a property of the models, and we flag the headline VaR comparison as running two estimation protocols rather than one. A same-protocol re-estimation is the clean fix; we have not run it, and we mark it as an open robustness item rather than assert what it would show.

One direction can nonetheless be signed without re-estimating, because it follows from the shape of the standardized Student- t quantile rather than from the data. Writing $q_\alpha(\nu) = t_\alpha(\nu)\sqrt{(\nu-2)/\nu}$, the quantile used in Equation (15), $|q_\alpha|$ is decreasing in ν for small α but increasing in ν once α is large enough; for the pair $(\nu = 6.03, 8.12)$ the crossing lies between $\alpha = 2.5\%$ and $\alpha = 3\%$. At the 1% and 2.5% levels, therefore, GJR- t 's smaller $\hat{\nu}$ gives it the *wider* VaR and so the more forgiving side of the violation-count criterion. The margin is not symmetric across the two levels: at 1% the quantiles are -2.5649 against -2.5057 , so GJR- t 's interval is 2.4% wider, whereas at 2.5% they are -1.9980 against -1.9968 and the gap is only 0.06%. The argument below therefore carries real force at 1% and almost none at 2.5%. GJR- t is therefore on the forgiving side at both levels—and it still over-violates at both (1.64% against a nominal 1%; 3.50% against 2.5%) and fails at least two of the three tests at each. The protocol asymmetry cannot be what makes GJR- t fail there. At the 5% level the ordering reverses, so part of GJR- t 's failure at that level may be protocol rather than model, and we do not claim otherwise.

6.9 Residual Diagnostics: How VIX Absorption Lightens the Residual Tails

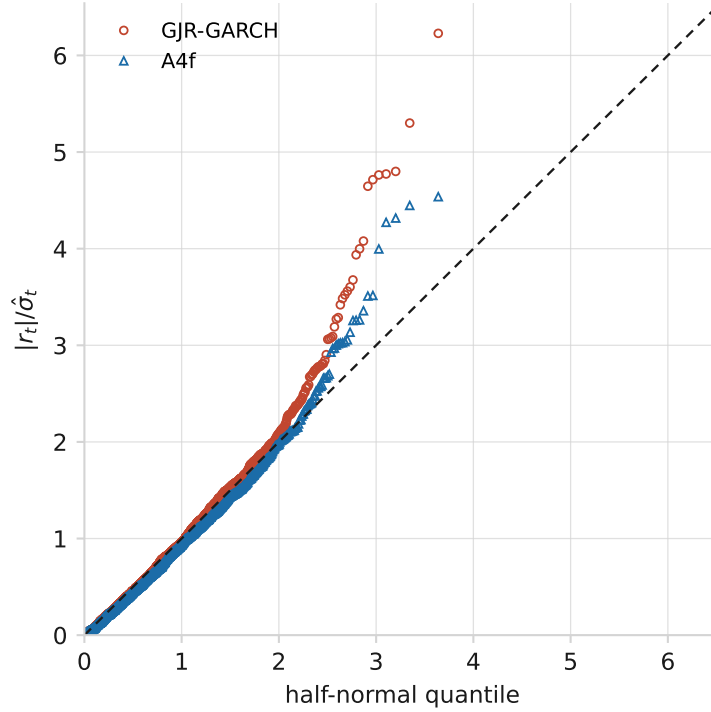


Figure 3: **Standardized residuals against the half-normal.** Empirical quantiles of $|r_t|/\hat{\sigma}_t$ against half-normal quantiles, for the two headline models on the same out-of-sample window. Only $|r_t|$ is recoverable from the archived loss inputs, so the plot is one-sided; the question it asks is the same one a two-sided QQ plot asks. Both models lie above the reference line in the far tail, so neither standardizes to a Gaussian, but A4f's largest standardized residual is 4.54 against GJR-GARCH's 6.23, and the symmetrized excess kurtosis falls from 2.14 to 1.13. Identity is carried by marker shape as well as colour. *These kurtosis figures are not the ones in Table 16, and the two are not comparable: both are now computed on the same pinned $n = 1,826$ window, but this figure symmetrizes $|r_t|/\hat{\sigma}_t$ before taking the moment and uses the Gaussian quasi-likelihood fits, whereas the table takes the raw standardized-residual moment from Student- t fits. Both agree on direction and neither is a restatement of the other.*

The shift in optimal degrees of freedom documented above motivates a deeper question: *how does the VIX component change the distributional properties of standardized*

residuals? If A4f’s τ_t absorbs the extreme volatility episodes that VIX captures, the standardized residuals $\hat{\varepsilon}_t = r_t/\hat{\sigma}_t$ should exhibit lighter tails than those from GJR-GARCH. Figure 3 previews the answer graphically, as a half-normal QQ plot of the standardized residuals from the two headline models.

Table 16 reports a comprehensive diagnostic comparison of standardized residuals from GJR- t and A4f- t over the OOS period ($n = 1,826$). The run that produces it previously downloaded prices at run time; it now reads the same hash-pinned adjusted-close snapshot and evaluates the same window as Table 3, so this subsection and the horse race are finally on one sample.

Table 16: Residual diagnostics: GJR- t vs. A4f- t standardized residuals (OOS)

Diagnostic	GJR- t	A4f- t	Change	Interpretation
Excess kurtosis	2.111	1.227	−41.9%	Substantially lighter tails
Skewness	−0.771	−0.588	−23.6%	Reduced left-tail asymmetry
Jarque-Bera statistic	519.7	220.0	−57.7%	Closer to normality
Median Student- t $\hat{\nu}$	5.20	8.00	+53.7%	Shift toward Gaussian

Standardized residuals $\hat{\varepsilon}_t = r_t/\hat{\sigma}_t$. Excess kurtosis relative to normal ($= 0$). Student- t $\hat{\nu}$ estimated by MLE in each rolling window; median across all OOS windows reported. SPY, OOS 2019-01-02 to 2026-04-08, $n = 1,826$, computed on the same hash-pinned adjusted-close snapshot and the same window as Table 3. The excess-kurtosis reduction is sensitive to the evaluation window: on a window two trading days longer—computed by the same producer on a live, unpinned download ($n = 1,828$) whose vintage is no longer recoverable—GJR- t ’s excess kurtosis is 3.065 rather than 2.111 while A4f- t ’s is essentially unchanged, so the reduction reads as 59.6% rather than the 41.9% reported here. The direction of every diagnostic is unchanged, but the magnitude is not robust to those two observations.

A4f reduces excess kurtosis by 42% (from 2.111 to 1.227), the Jarque-Bera statistic by 58% (from 519.7 to 220.0), and shifts the median optimal Student- t degrees of freedom from 5.20 to 8.00—moving the residual distribution substantially toward the Gaussian limit. The skewness reduction (−23.6%) indicates that A4f also absorbs part of the left-tail asymmetry associated with market crashes.

One of these diagnostics is fragile, and we would rather report that than the larger number. Repinning this producer to the paper’s snapshot shortened its window by two trading days, and that alone moved the headline reduction from 60% to 42%. The two models did not move together: A4f’s excess kurtosis is essentially unchanged ($1.238 \rightarrow 1.227$), while GJR- t ’s falls from 3.065 to 2.111. About half the earlier gap was therefore carried by two observations lying outside the window this paper evaluates: those two days account for 51.6% of the original 1.827 difference in excess kurtosis, and the remaining 48.4% survives on the pinned sample. We draw the narrow conclusion rather than the flattering one: a fourth moment estimated on 1,826 standardized residuals is sensitive to individual extremes, so the *direction* here is robust—A4f standardizes to something closer to Gaussian on every measure in Table 16—while the *magnitude* should not be quoted to more precision than the sample supports. We make no claim about what happened on those two days, because this design does not measure it.

This finding has a direct economic interpretation: extreme VIX spikes (such as the COVID-19 peak of $VIX = 82.69$) generate heavy-tailed returns that GJR-GARCH must accommodate entirely through fat-tailed innovations (low ν). By absorbing these events through the τ_t component, A4f “normalizes” the residuals, allowing the distributional assumption to focus on the *remaining* non-Gaussianity rather than compensating for unmodeled volatility regime shifts.

This mechanism is the natural explanation for the calibration ordering documented in Section 5.4, where A4f- t has the violation rate closest to nominal at every level: better-calibrated residual distributions lead to more accurate quantile estimates, particularly in the tails where kurtosis matters most. It does not explain a pass-count advantage, because on the pinned sample there is none.

6.10 VIX vs. Macroeconomic Variables

A natural question is whether macroeconomic variables can substitute for or complement VIX in the multiplicative framework. Following Conrad and Loch (2015), we estimate GARCH-X models with three macroeconomic specifications (term spread alone, unem-

ployment rate alone, and a two-factor combination of both) as alternatives to VIX; this check uses live-downloaded data rather than a pinned snapshot, so it is reported as supporting rather than as a pinned robustness result. The market-implied VIX outranks every macroeconomic specification tested, and clears our $|t| > 3.0$ reporting screen against two of the three: the combined term-spread/unemployment model (DM $t = -3.28$) and unemployment alone ($t = -3.12$). Against term spread alone it does not ($t = -2.72$), so “every” is a statement about the ordering and not about significance. These figures come from a live-download run evaluating $n = 1,777$ observations rather than the pinned window’s 1,826; read them as a direction, not as figures comparable in level with Table 3. The ordering is consistent with VIX aggregating publicly available volatility-relevant information through option prices more effectively than the macro variables available for this check.

7 Discussion: Theoretical Foundation of Source Decomposition

A natural concern is whether the “source decomposition” label is merely a relabeling of the standard long-run/short-run narrative in GARCH-MIDAS models. We address this by deriving two propositions and one empirical remark that establish the economic content of the decomposition.

Proposition 1 (Unconditional Variance Identity). *For any stationary τ_t and g_t , the unconditional variance satisfies*

$$\mathbb{E}[\sigma_t^2] = \mathbb{E}[\tau_t] \cdot \mathbb{E}[g_t] + \text{Cov}(\tau_t, g_t). \quad (20)$$

Note: This is an algebraic identity that follows directly from the multiplicative structure and does not require estimation. It carries no auxiliary claim that the constraint $\omega_g = 1 - \alpha - \gamma/2 - \beta$ delivers $\mathbb{E}[g_t] = 1$: that constraint delivers $\mathbb{E}[g_t] = 1$ only when $\mathbb{E}[u_{t-1}^2] = \mathbb{E}[g_{t-1}]$, which holds under the *lagged* normalization $u_{t-1} = r_{t-1}/\sqrt{\tau_{t-1}}$ but not under

the contemporaneous normalization $u_{t-1} = r_{t-1}/\sqrt{\tau_t}$ that our baseline specifications use (Table 1). Estimated on our sample, the constrained specification gives $\mathbb{E}[g_t] = 0.84$, not 1. The constraint should therefore be read as a scale normalization rather than as a moment condition, and neither term of Equation 20 may be dropped.

The identity is exact, not an approximation. The simpler $\mathbb{E}[\sigma_t^2] \approx \mathbb{E}[\tau_t]\mathbb{E}[g_t]$ holds only when $\text{Cov}(\tau_t, g_t) \approx 0$, and here it is not: $\text{Corr}(\tau_t, g_t) = 0.51$ for the constrained specification and 0.52 for the free- ω_g one, reflecting the crisis channel of high VIX $\rightarrow$ high $\tau_t \rightarrow$ large standardized shocks $\rightarrow$ higher g_t . A separate archived run of the same constrained *specification* reports 0.554 and $\mathbb{E}[g_t] = 0.870$, and we do not claim to have reconciled the two. That run discards the first 500 observations, analysing 4,848 of 5,348 against the 5,349 used here, but the burn-in is not the explanation: its own full-sample figure is 0.873, so dropping the burn-in accounts for 0.0025 of the distance to our 0.840 and leaves 0.0305 standing. Nor can a one-observation difference in sample size move a mean of 5,349 values by 0.0305. What remains is a difference between two fits rather than between two samples, and Section 3.4 records what separates them: the two runs use different optimizers over different admissible sets, and each lands on the ARCH lower edge of its own—to the optimizer’s tolerance in ours, exactly in that run’s. Both parameter vectors are archived with the replication package, so the two landing points can be compared directly. We nonetheless report the residual gap as undecomposed rather than name a cause we have not tested. The $\text{Corr}(\tau_t, g_t)$ leg cannot be examined the same way, because that run records no full-sample correlation.

Proposition 2 (Implied-Variance Discount, and What Is Not Identified). *Let $\text{IV}_t = \text{VIX}_{t-1}^2/(252 \times 10,000)$ be the daily implied variance under the “no-VRP” benchmark. The ratio $\mathbb{E}[\sigma_t^2]/\mathbb{E}[\text{IV}_t]$ measures the discount the fitted model applies to implied variance and is invariant to the reparameterization below; θ_1 alone does not measure it.*

In the free- ω_g specification the map

$$(\theta_0, \theta_1, \omega_g) \longmapsto (c\theta_0, c\theta_1, \omega_g/c), \quad c > 0, \quad (21)$$

sends $\tau_t \mapsto c\tau_t$, hence $u_{t-1} \mapsto u_{t-1}/\sqrt{c}$ and $g_t \mapsto g_t/c$, leaving $\sigma_t^2 = \tau_t g_t$ and therefore the entire likelihood unchanged. The scale of τ_t is not identified, and neither θ_1 nor $\mathbb{E}[g_t]$ is a property of the data.

This is not a small-sample concern. Evaluating our own likelihood at the fitted A4f parameters and at rescalings of them, a factor of 10,000 in c moves θ_1 across four orders of magnitude and $\mathbb{E}[g_t]$ from 165.9 to 0.017 while the log-likelihood is identical to ten decimal places and $\max_t |\Delta\sigma_t^2|$ remains at machine precision. The constrained specification is unaffected—its ω_g is pinned and cannot absorb c , so the same rescaling costs it 831 log-likelihood points. Two of our own result files record $\mathbb{E}[g_t^{\text{A4f}}]$ as 0.48 and 1.58 for the same model on the same asset; they are not in conflict, they are two optimizer landing points on the same ridge.

A parameter-level reading of the free- ω_g fit—for instance as a “two-channel” decomposition of the VRP correction—is not available, because each such figure is a coordinate on the ridge of Equation 21. What is available is the invariant quantity. Estimated on the full sample, the constrained specification discounts implied variance by 21.6% and the free- ω_g specification by 20.3%, against a directly observed average variance risk premium of 17.98% of implied variance on the same sample. The model’s discount and the measured premium agree to within a few percentage points without either being fitted to the other, which is the substantive claim this proposition can support. The two model discounts, and the $\mathbb{E}[g_t]$ and $\text{Corr}(\tau_t, g_t)$ values quoted earlier in this section, come from two full-sample fits—one constrained, one free- ω_g . The 17.98% is not a fitted quantity at all: it is a sample moment, the average of squared returns against daily implied variance, computed straight from the data. All of it is recorded together in the archived cross-asset results, which store each fit’s parameter vector alongside a description of the estimator that produced it, and the observed premium at the top level outside both fit blocks. The two estimators differ from each other, and Section 3.4 states how; a replicator should start from that file and from its per-fit `estimator` entries rather than from the separate archived run cited above.

Nothing in Section 5 depends on the unidentified direction: $\hat{\sigma}_t^2$ is invariant along it,

so the horse race, the Model Confidence Set, the cross-asset tests and the VaR and ES backtests are untouched.

Remark 1 (g_t Does Not Track VRP). *It would be natural to read g_t as reflecting time-varying departures of realized variance from the VRP-corrected implied level, with the GARCH structure extracting a persistent VRP component. The data do not support that reading. On the correctly scaled proxy and the fitted g_t recursion, no specification's g_t correlates with VRP beyond $|\rho| = 0.09$. The dynamic factor carries endogenous GARCH dynamics; it does not carry the premium.*

The discount of Proposition 2 is therefore a statement about the unconditional *level* of fitted variance relative to implied variance, not about g_t 's time variation. It survives the correction above because it never used the VRP proxy, and it is measured by an invariant ratio rather than by an individual parameter. What does not survive is any claim that the g_t series tracks VRP dynamically, in A4f or in any other specification (Table 4).

These results distinguish source decomposition from mere relabeling on three grounds, none of which rests on an individual parameter: (i) the fitted model applies an economically interpretable discount to implied variance—21.6% under the constrained specification and 20.3% under A4f, against a directly measured premium of 17.98% of implied variance—and that discount is invariant to the reparameterization of Equation 21, whereas θ_1 and $\mathbb{E}[g_t]$ are not; (ii) the multiplicative structure provides structural scale separation, with τ_t carrying exogenous option-market fear and g_t carrying endogenous GARCH dynamics, and this reading does not depend on where along the ridge the optimizer lands, since $\sigma_t^2 = \tau_t g_t$ is invariant along it; and (iii) the forecasting gain (DM $t = 4.47$ vs. GJR) demonstrates that the decomposition captures information not available to standard models.

8 Conclusion

This paper provides a systematic evaluation of multiplicative volatility models that incorporate VIX as an external scale factor. Our main findings are as follows.

First, A4f and a majority—but not all—of the VIX-augmented multiplicative models outperform plain GJR-GARCH by margins that clear the $|t| > 3$ screen (A4f: DM $t = 4.47$). The Model Confidence Set eliminates only GJR at $\alpha = 0.25$, a set-level result consistent with useful VIX information but not a pairwise win for every specification. Among VIX-augmented models, A4f ($\tau_t = \theta_0 + \theta_1 \text{VIX}_{t-1}^2$, free ω_g) achieves the lowest QLIKE, and outranks longer-lag MIDAS specifications by margins that clear the screen (B2: $|t| = 3.06$, which exceeds the 3.00 threshold by two percent; B3: $|t| = 3.75$), but is statistically indistinguishable from the shortest-lag MIDAS (B1, $K = 22$). The practical implication is that longer-lag GARCH-MIDAS specifications ($K \geq 65$) are outranked by A4f by margins clearing the screen ($|t| > 3.0$), while the shortest-lag MIDAS (B1, $K = 22$) achieves statistically indistinguishable performance (DM $t = -0.90$)—parsimony rather than statistical dominance is the practical argument for the daily specification.

Second, the design dimensions *within* the GARCH-X family order the models consistently but do not separate them statistically. VIX-squared ranks above log-exponential, free ω_g above constrained, and contemporaneous normalization above lagged; every one of those pairwise DM tests is insignificant ($|t| \leq 1.38$) and every gap sits in the third decimal place of mean QLIKE. We therefore present the dimensional argument for VIX² (variance units matching σ_t^2), the VRP argument for free ω_g , and the [Engle et al. \(2013\)](#) recursion argument for contemporaneous normalization as reasons to prefer those choices *a priori*, with the horse race showing only that the resulting ranking is consistent with them. The comparison the data does resolve is the one between VIX-augmented and plain GJR-GARCH.

Third, the g_t component provides a structured decomposition of volatility into exogenous (option-market fear) and endogenous (GARCH dynamics) sources, but it does not carry the variance risk premium. Across all four specifications examined, g_t 's contemporaneous correlation with VRP is economically negligible ($|\rho| \leq 0.09$, and indistinguishable from zero in three of four cases), and g_t has no exploitable out-of-sample predictive power for future VRP. The decomposition is a forecasting device, not a premium-extraction device.

Fourth, of the seven markets originally planned for cross-asset validation, six were completed (STOXX50E could not be evaluated on pinned data) and A4f clears our $|t| > 3.0$ reporting screen on two of those six—SPY and QQQ, both using US VIX as the fear index. Three further markets (GLD with its own GVZ index, FEZ, and EEM) improve on GJR-GARCH in the same direction but fall short of that screen, and 0050.TW falls short as well, whether the fear index is U.S. VIX or Taiwan’s own VIXTWN; an asset’s raw correlation with VIX does not by itself predict which of these clear the threshold. The evidence for cross-market usefulness is therefore narrower than a full multi-asset validation: it holds where VIX, or an asset-specific fear index such as GVZ for GLD, is genuinely informative about the asset’s own volatility, but the improvement does not reliably clear this paper’s significance screen outside the primary US equity setting.

Fifth, residual diagnostics reveal the mechanism behind A4f- t ’s violation-rate proximity: the VIX component absorbs extreme volatility episodes that GJR-GARCH must accommodate solely through fat-tailed innovations. A4f’s standardized residuals exhibit 42% lower excess kurtosis (1.23 vs. 2.11) and a shift in the median optimal Student- t degrees of freedom from 5.2 to 8.0, resulting in a 58% reduction in the Jarque-Bera statistic. This “tail normalization” effect is visible in the violation rates—A4f- t is closest to nominal at all three levels—but it does not produce a pass-count advantage: under the three-test criterion A4f- t , A4f-Normal and GJR-Normal each pass at one of the three levels, and only GJR- t passes at none.

Our findings have implications for both researchers and practitioners. For researchers, the results suggest that model complexity should be matched to data frequency: component models with mixed-frequency aggregation are valuable when the external information is genuinely lower-frequency (e.g., quarterly GDP), but a simple daily specification is preferable when the variable updates daily. For practitioners, the A4f- t model offers a parsimonious and interpretable alternative to both standard GARCH and complex GARCH-MIDAS specifications for equity volatility forecasting. Its case for risk management is weaker, and Section 5.4 states it as such.

Several limitations should be noted. First, our QLIKE evaluation uses r_t^2 as the

volatility proxy. The proxy-robustness result of [Patton \(2011\)](#) concerns expected loss when the proxy satisfies the required conditional-unbiasedness assumptions; it does not guarantee that arbitrary proxy noise preserves a realized finite-sample ranking. Both the *magnitudes* of QLIKE differences and DM statistics, and potentially their finite-sample ordering, may differ when 5-minute realized variance is used as the proxy. Since VIX and intraday realized variance have a non-trivial relationship, the relative advantage of VIX-based models could change with proxy choice; we leave this robustness check for future work with high-frequency data. Second, our cross-asset validation covers five additional assets (six including primary SPY, with STOXX50E not evaluated for want of a pinned snapshot); a broader study spanning international equity indices, fixed income, and commodity markets would strengthen the generalization claims. The VaR backtesting reveals persistent difficulty at the 2.5% level (DQ $p = 0.006$ for A4f- t), suggesting that violation clustering remains incompletely addressed. Finally, the free-omega A4f model adds one parameter relative to the constrained A4; while this improves fit for SPY and QQQ, the cross-asset comparison shows no significant A4f vs. A4 difference ($t < 1.7$ for all non-SPY assets), suggesting that the free omega is primarily useful for assets with strong VIX linkage.

Future research directions include incorporating realized variance measures from high-frequency data as the target variable, extending the multiplicative framework to multivariate settings for portfolio risk management, multi-horizon VaR calibration at $h = 5$ and 10 day horizons, and investigating whether the source decomposition can be exploited in longer-horizon VRP prediction with monthly or quarterly aggregation.

We have preliminary results on the first two of these, and we state them here with their figures rather than as encouragement, because an unsourced “preliminary results suggest” in a conclusion tells a reader nothing they can check. On a live download of an SPY/GLD pair ($n = 1,828$, not the pinned snapshot), a bivariate extension using the dynamic conditional correlation model of [Engle \(2002\)](#) attains a lower portfolio QLIKE with A4f margins than with GJR-GARCH ones (-8.836 against -8.795), but the Diebold–Mariano statistic is $t = 2.88$ and *does not* clear this paper’s $|t| > 3.0$ screen; the producing run’s

own flag for that screen records the comparison as not significant. On multi-horizon calibration, scaling the one-day VaR by $\sqrt{h}$ passes the three-test criterion in 14 of the 16 cells evaluated—both models, both assets, both horizons ($h = 5$, $h = 10$) and both nominal levels—across SPY and QQQ, the two exceptions being QQQ at $h = 10$ and $\alpha = 1\%$, where zero violations in 182 observations leave the conditional-coverage test undefined. Both exercises are on live vintages and neither has been repinned, so we report them as directions worth taking rather than as results.

A Specification Genealogy and Multiple-Testing Dis-closure

To address potential concerns about data mining and post-hoc specification selection, Tables 17 and 18 document the ex-ante design rationale for all 17 specifications in the horse race. All specifications were defined prior to examining any OOS QLIKE rankings; the selection exercise is purely an ex-post ranking of pre-registered designs. (Split into two tables here so the combined 17-row, wide-rationale-column layout does not overflow a single page — the earlier single-table version rendered with a 552pt overfull vbox and silently clipped its bottom rows.)

Table 17: Specification genealogy, part 1: GARCH-X variants

Spec	Category	Ex-ante design rationale
A1	GARCH-X	Inherited from an earlier pilot specification; included <i>only</i> to document the cost of estimation/forecasting inconsistency (τ_t in estimation, τ_{t-1} in OOS)
A2	GARCH-X	Log-exponential: convention from the GARCH-MIDAS literature (Engle et al., 2013)
A3	GARCH-X	Log-exponential with lagged denominator (τ_{t-1}): conservative timing variant
A4	GARCH-X	VIX-squared: dimensional consistency (VIX^2 has variance units matching σ_t^2)
A5	GARCH-X	VIX-level $\exp(\theta_1 \text{VIX})$: alternative scaling, included for completeness
A2f	GARCH-X	Log-exponential with free ω_g : accommodates the VRP systematic wedge
A4f	GARCH-X	VIX-squared with free ω_g : combines A4 dimensional consistency with VRP accommodation
A3f	GARCH-X	Log-exponential, lagged denominator, free ω_g : conservative timing + VRP flexibility
A2n	GARCH-X	Log-exponential keeping the ω_g restriction and rescaling τ_t so the standardized innovation has unit in-sample second moment
A4n	GARCH-X	VIX-squared with sample-mean normalization: dimensional consistency + alternative normalization

Table 18: Specification genealogy, part 2: GARCH-MIDAS variants and benchmark

Spec	Category	Ex-ante design rationale
B1	MIDAS-RW	Rolling MIDAS, $K = 22$ daily lags (≈ 1 month): shortest conventional lag (Engle et al., 2013)
B2	MIDAS-RW	Rolling MIDAS, $K = 65$ daily lags (≈ 1 quarter): medium lag
B3	MIDAS-RW	Rolling MIDAS, $K = 125$ daily lags (≈ 6 months): long lag
C1	MIDAS-FS	Fixed-span MIDAS, $K_m = 6$ monthly lags: semi-annual window
C2	MIDAS-FS	Fixed-span MIDAS, $K_m = 12$ monthly lags: annual window
C3	MIDAS-FS	Fixed-span MIDAS, $K_m = 24$ monthly lags: 2-year window
B0	Benchmark	Standard GJR-GARCH(1,1): well-established daily volatility benchmark

All 17 specifications (Table 17 and this table) were defined prior to any OOS evaluation. The only ex-post element is A4f emerging as the point-estimate QLIKE winner; the economic arguments for A4f (dimensional consistency + VRP flexibility) were documented before OOS results were examined.

Multiple-testing disclosure: The 16-comparison horse race uses our conservative $|t| > 3.0$ reporting screen, adopted following the multiple-testing concerns raised by [Harvey et al. \(2016\)](#) rather than a threshold that paper itself defines for the DM test; the screen is approximately equal to the Bonferroni-adjusted critical value $z_{0.05/(2 \times 16)} \approx 2.95$. The primary A4f vs. GJR-GARCH comparison ($t = 4.47$) substantially exceeds this threshold. The Model Confidence Set ([Hansen et al., 2011](#)) provides a complementary multiple-testing-adjusted evaluation (Section 6).

B HAC Bandwidth Sensitivity of the Headline Statistic

Section 5 reports the headline Diebold–Mariano statistic under a single Newey–West truncation lag, chosen by the $\lfloor n^{1/3} \rfloor$ rule that gives 12 at our $n = 1,826$. Section 6 reports lag sensitivity for the secondary HAR-style daily- r^2 versus GJR comparison, where it matters, and finds that comparison bandwidth-sensitive. A reader who sees the check performed there and not for the primary statistic is entitled to ask for the primary one, so we report it here.

Table 19 recomputes the A4f-versus-GJR statistic at seven truncation lags spanning from no HAC correction at all to 20, well above the rule’s value. Nothing is re-estimated: the table re-runs only the variance estimator on the archived per-day loss differentials (the archived loss paths, the same 1,826 days behind Table 3), because those losses are the test’s entire input and refitting would inject optimizer noise into a table whose whole purpose is to isolate the bandwidth. The producer refuses to write its output unless it first reproduces the published statistic at the published lag.

The statistic moves between 4.4464 and 4.5521 across the seven bandwidths, a range of 0.11, and clears the reporting screen at every one of them with a margin of more than 1.4. The lowest value is not at the extremes but at lag 10, and the uncorrected standard error gives 4.4565, almost exactly the corrected figure — the loss differential carries little enough serial correlation at this horizon that the HAC correction has little to do. We therefore read the headline comparison as robust to the bandwidth choice, which is the opposite of what Section 6 finds for the HAR-style daily- r^2 versus GJR comparison, and we report both.

References

- Acerbi, C. and Székely, B. (2014). Back-testing expected shortfall. *Risk*, 27(11):76–81.
- Bekaert, G. and Hoerova, M. (2014). The VIX, the variance premium and stock market

Table 19: HAC bandwidth sensitivity of the headline DM statistic

Truncation lag	DM t	Clears $ t > 3.0$
0 (uncorrected)	4.4565	Yes
1	4.5088	Yes
5	4.4549	Yes
10	4.4464	Yes
12	4.4689	Yes
13	4.4743	Yes
20	4.5521	Yes

A4f (VIX², free ω_g) against the GJR-GARCH benchmark, QLIKE loss, $n = 1,826$ out-of-sample days. Bartlett kernel; lag 0 is the uncorrected standard error. The bolded row is the bandwidth the paper’s producer selects by its $\lfloor n^{1/3} \rfloor$ rule, and its t is the headline 4.47 of Table 3. Produced by the HAC-sensitivity routine from the archived loss paths; no estimation is performed.

Both 12 and 13 appear because the two parts of this paper round the same $n^{1/3}$ bandwidth in opposite directions: the horse-race producer takes the floor, giving 12 here, while the diagnostic run quoted in Section 6 takes the ceiling, giving the 13 of that subsection’s grid. Both samples lie between 12^3 and 13^3 , so the two rules differ by exactly one lag on either of them; it is a rounding convention and not a sample-size effect. Both rows are tabulated above.

- volatility. *Journal of Econometrics*, 183(2):181–192.
- Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327.
- Bollerslev, T., Tauchen, G., and Zhou, H. (2009). Expected stock returns and variance risk premia. *Review of Financial Studies*, 22(11):4463–4492.
- Campbell, J.Y. and Thompson, S.B. (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *Review of Financial Studies*, 21(4):1509–1531.
- Carr, P. and Wu, L. (2009). Variance risk premiums. *Review of Financial Studies*, 22(3):1311–1341.
- Christensen, B.J. and Prabhala, N.R. (1998). The relation between implied and realized volatility. *Journal of Financial Economics*, 50(2):125–150.
- Christoffersen, P.F. (1998). Evaluating interval forecasts. *International Economic Review*, 39(4):841–862.
- Clark, T.E. and West, K.D. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics*, 138(1):291–311.
- Conrad, C. and Kleen, O. (2020). Two are better than one: Volatility forecasting using multiplicative component GARCH-MIDAS models. *Journal of Applied Econometrics*, 35(1):19–45.
- Conrad, C. and Loch, K. (2015). Anticipating long-term stock market volatility. *Journal of Applied Econometrics*, 30(7):1090–1114.
- Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2):174–196.
- Diebold, F.X. and Mariano, R.S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263.

- Engle, R.F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50(4):987–1007.
- Engle, R.F. (2002). Dynamic conditional correlation: a simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *Journal of Business & Economic Statistics*, 20(3):339–350.
- Engle, R.F. and Manganelli, S. (2004). CAViaR: Conditional autoregressive value at risk by regression quantiles. *Journal of Business & Economic Statistics*, 22(4):367–381.
- Engle, R.F. and Rangel, J.G. (2008). The spline-GARCH model for low-frequency volatility and its global macroeconomic causes. *Review of Financial Studies*, 21(3):1187–1222.
- Engle, R.F., Ghysels, E., and Sohn, B. (2013). Stock market volatility and macroeconomic fundamentals. *Review of Economics and Statistics*, 95(3):776–797.
- Francq, C. and Thieu, L.Q. (2019). QML inference for volatility models with covariates. *Econometric Theory*, 35(1):37–72.
- Giacomini, R. and White, H. (2006). Tests of conditional predictive ability. *Econometrica*, 74(6):1545–1578.
- Glosten, L.R., Jagannathan, R., and Runkle, D.E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *Journal of Finance*, 48(5):1779–1801.
- Han, H. and Kristensen, D. (2014). Asymptotic theory for the QMLE in GARCH-X models with stationary and nonstationary covariates. *Journal of Business & Economic Statistics*, 32(3):416–429.
- Hansen, P.R., Lunde, A., and Nason, J.M. (2011). The model confidence set. *Econometrica*, 79(2):453–497.
- Harvey, C.R., Liu, Y., and Zhu, H. (2016). ... and the cross-section of expected returns. *Review of Financial Studies*, 29(1):5–68.

- Jiang, G.J. and Tian, Y.S. (2005). The model-free implied volatility and its information content. *Review of Financial Studies*, 18(4):1305–1342.
- Kupiec, P.H. (1995). Techniques for verifying the accuracy of risk measurement models. *Journal of Derivatives*, 3(2):73–84.
- Lai, Y.-H. (2026). Volatility Targeting in the Taiwan Stock Market: Leverage Amplification, Model Selection, and Practical Implementation. Working paper, Da-Yeh University.
- Newey, W.K. and West, K.D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708.
- Patton, A.J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, 160(1):246–256.
- Wang, F. and Ghysels, E. (2015). Econometric analysis of volatility component models. *Econometric Theory*, 31(2):362–393.
- White, H. (2000). A reality check for data snooping. *Econometrica*, 68(5):1097–1126.