

Can Anything Beat VIX? A Systematic Out-of-Sample Evaluation of Thirteen Signal Families for Equity Volatility Forecasting and Volatility Timing*

Yi-Hao Lai[†]

March 2026

Abstract

We conduct a systematic out-of-sample horse race among thirteen pre-specified signal families—cross-asset volatility momentum, VIX term structure, behavioral sentiment, the variance risk premium, multi-asset portfolio optimization, equal risk contribution, Bitcoin volatility, the yield curve slope, Google Trends fear proxies, overnight VIX changes, calendar anomalies, economic-policy uncertainty, and financial-stress indices—evaluating whether any can improve upon VIX alone for equity volatility forecasting and volatility-timing portfolio construction. Using daily S&P 500 data spanning 1993–2026, we apply a unified forecast pipeline with strict lag enforcement, transaction-cost accounting, publication-delay-corrected treatment of weekly macroeconomic series, the [Harvey et al. \(2016\)](#) multiple-testing threshold ($|t| > 3.0$), the Clark-West test for nested comparisons, and proxy-robust evaluation following [Patton \(2011\)](#). The main finding is strongly negative: not a single signal family produces a statistically significant out-of-sample improvement at the Harvey threshold—whether measured by the standard [Campbell and Thompson \(2008\)](#) R^2_{OOS} relative to the historical mean, by incremental R^2 relative

*All errors are my own. Data and replication code are available upon request.

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

to VIX-only forecasts, or by Sharpe ratio gains for volatility-timing strategies—and no nested daily comparison approaches the Harvey threshold under the Clark-West test; the nearest, the VIX term-structure family, has a one-sided p -value of 0.0435 at Newey-West lag 21 and 0.0580 at the canonical lag 35 (0.0452 at lag 21 with the Harvey-Leybourne-Newbold small-sample factor), so it straddles the 5% line, and it does not survive a Bonferroni correction across the seven nested families. A power check shows that the Clark-West statistic stays near 2 even for a look-ahead signal and its noise-degraded versions, so for those signal classes the Harvey threshold cannot separate content from noise, and the null rests on the point estimates and the Clark-West p -values rather than on the threshold alone. The VIX–realized volatility R^2 ranges from 0.24 to 0.64 across five non-overlapping eras spanning 33 years (coefficient of variation = 0.33), demonstrating time-invariance. Volatility model rankings are criterion-dependent—GJR-GARCH dominates under proxy-robust QLIKE on r_t^2 (Model Confidence Set sole member) while AMEM dominates for VaR/ES risk management (composite score 1.94 vs. 1.56)—yet VIX sufficiency holds regardless of which model or criterion is used. We frame VIX sufficiency as a rational outcome of option-market information aggregation and show that volatility timing functions as drawdown insurance—costly in Sharpe terms (average drag 3.49%/year for 12/VIX) but welfare-improving for investors with CRRA risk aversion $\gamma \geq 4.5$. The informative null sharpens the research frontier: future gains require either higher-frequency realized measures or genuinely exogenous information not already embedded in option prices.

Keywords: VIX, volatility forecasting, volatility timing, out-of-sample evaluation, variance risk premium, multiple testing, option-implied volatility, drawdown insurance, proxy-robust loss, model confidence set, Value-at-Risk

JEL Classification: C53, G11, G12, G17

Contents

1	Introduction	6
2	Why VIX is the Benchmark	9
2.1	Theoretical Foundations	9
2.2	Empirical Evidence for VIX Forecasting Power	10
2.3	The Information Aggregation Argument	11
2.4	Prior Horse Races	11
3	Data and the Thirteen Signal Families	12
3.1	Core Data	12
3.2	The Thirteen Signal Families	12
3.2.1	Family 1: Cross-Asset Volatility Momentum	13
3.2.2	Family 2: VIX Term Structure	14
3.2.3	Family 3: Behavioral Sentiment Index	14
3.2.4	Family 4: Variance Risk Premium	14
3.2.5	Family 5: Multi-Asset Portfolio Optimization	14
3.2.6	Family 6: Equal Risk Contribution	14
3.2.7	Family 7: Bitcoin Volatility	15
3.2.8	Family 8: Yield Curve Slope	15
3.2.9	Family 9: Google Trends Fear Proxy	15
3.2.10	Family 10: Overnight VIX Changes	15
3.2.11	Family 11: Calendar Anomaly	16
3.2.12	Family 12: Economic Policy Uncertainty	16
3.2.13	Family 13: Financial-Stress Indices	17
4	Forecast Design and Benchmark Models	17
4.1	Unified Forecast Pipeline	17
4.2	VIX Benchmark	18
4.3	Incremental Information Test	18

4.4	Loss Functions and Proxy Robustness	18
4.5	Out-of-Sample R^2	19
4.6	Publication-Delay Convention	20
5	Statistical Evaluation and Multiple-Testing Control	22
5.1	Diebold-Mariano Test	22
5.2	Nested-Model Correction: Clark-West and Small-Sample Adjustment . .	23
5.3	Multiple Testing	24
5.4	Model Confidence Set	25
5.5	Cross-Era Validation	26
6	Volatility-Timing Strategy Design	26
6.1	The 12/VIX Strategy	26
6.2	Baseline: 50/50 SPY/GLD	27
6.3	Transaction Costs	27
7	Main Results and Robustness	27
7.1	Volatility Forecasting: No Signal Beats VIX	27
7.2	Strategy Comparison: No Signal-Based Strategy Beats Benchmarks . . .	35
7.3	Multi-Asset Portfolio Optimization	37
7.4	Era Stability of VIX Sufficiency	38
7.5	Competing Signals by Era	39
7.6	Criterion-Dependent Model Rankings	40
7.7	12/VIX Strategy by Era	42
7.8	Cross-Asset Native-IV Sufficiency: A Seven-Asset PIT Panorama	43
7.9	Channel-Specific Mechanism Heterogeneity	45
7.9.1	Channel 1: HAR+VIX on Equity—Regime-Invariant Gains on Parkinson Volatility	45
7.9.2	Channel 2: GARCH-MIDAS Long-Run Factor—Universal Null . .	46
7.9.3	Channel 3: GAS-t Score-Driven Models—Asset-Class Bifurcation	46
7.10	Publication-Delay Robustness	48

7.11	Alt-Data Fails a Second Paradigm: Regime-Based Allocation	51
8	Why the Null is Informative	56
8.1	The Information Aggregation Explanation	56
8.2	What Might Break VIX Sufficiency?	57
8.3	Economic Significance: Forecast Accuracy vs. Risk Management	58
8.4	The Drawdown Insurance Framework	60
8.5	The Simplicity Premium	62
9	Conclusion	62

1 Introduction

The Chicago Board Options Exchange Volatility Index (VIX) is the single most widely used measure of expected equity market volatility. Derived from the prices of S&P 500 index options, VIX represents the market’s consensus forecast of 30-day annualized volatility and has been labeled the “fear gauge” of financial markets (Whaley, 2000). A vast literature documents that VIX is a powerful, though biased, predictor of future realized volatility (Christensen and Prabhala, 1998; Jiang and Tian, 2005; Bekaert and Hoerova, 2014). Yet the practical question facing portfolio managers, risk officers, and retail investors remains open: *can any observable signal improve upon VIX for forecasting equity volatility and timing equity exposure?*

This question is not merely academic. Volatility-timing (VT) strategies, which scale equity exposure inversely to expected volatility, have attracted substantial interest since the seminal work of Fleming et al. (2001) and the volatility-managed portfolio framework of Moreira and Muir (2017). If a signal can forecast volatility more accurately than VIX, it can potentially improve the risk-return trade-off of these strategies. Conversely, if VIX already aggregates all publicly available information relevant to near-term volatility, then the search for a “better VIX” is futile, and research efforts should be redirected.

The existing literature offers fragments of evidence but no systematic verdict. Individual studies have examined specific signal categories—the variance risk premium (Bollerslev et al., 2009), cross-asset spillovers (Diebold and Yilmaz, 2012), macroeconomic determinants of time-varying volatility (Schwert, 1989; Engle and Rangel, 2008), and sentiment measures (Baker and Wurgler, 2006)—but each study uses its own sample period, evaluation metric, and benchmark specification, making cross-study comparison unreliable. Moreover, most studies do not enforce the multiple-testing corrections that Harvey et al. (2016) argue are essential when many predictors are examined.

We address this gap by conducting the first comprehensive, pre-specified horse race among thirteen signal families, evaluated under a unified framework with six methodological safeguards (families 12–13 are alt-data uncertainty proxies evaluated under the identical pipeline):

- (i) **Pre-specification:** All thirteen signal families and their construction rules are defined before examining out-of-sample results, eliminating data-snooping in signal selection.
- (ii) **Strict lag enforcement:** Every signal is lagged one trading day (`signal.shift(1)`) before being applied to returns, preventing look-ahead bias—a surprisingly common error that we document has inflated Sharpe ratios by up to 373% in prior work.¹
- (iii) **Transaction costs:** All strategies deduct 5 basis points per leg on each weight change, applied to both buy and sell sides.
- (iv) **Multiple-testing control:** We adopt the approximate conservative threshold of $|t| > 3.0$ motivated by [Harvey et al. \(2016\)](#) and evaluate the nested daily comparisons with the Clark-West test, whose one-sided p -values are reported without a multiplicity adjustment because none of them rejects at the canonical bandwidth; raw Diebold-Mariano statistics are reported as diagnostics only where the comparison is nested.
- (v) **Proxy-robust evaluation:** Following [Patton \(2011\)](#), we evaluate forecasts using QLIKE on r_t^2 (a proxy-robust loss function), supplemented by Spearman rank correlations and the Model Confidence Set of [Hansen et al. \(2011\)](#), ensuring that our rankings are not artifacts of the volatility proxy choice.
- (vi) **Cross-era stability:** Results are evaluated separately across five non-overlapping eras spanning 1993–2026, preventing any single regime from driving conclusions.

Our main finding is strongly negative but highly informative. Across 13 signal families and the full set of VIX sufficiency tests reported in Section 7, spanning five market eras totaling 8,325 trading days, we find:

- No signal family produces an out-of-sample R^2 improvement over VIX alone that is significant at the Harvey threshold (maximum in-sample incremental $R^2 = 0.038$;

¹In separate robustness analysis, careful code review identified look-ahead bias in four backtests, one of which produced a Sharpe ratio of 1.68 that collapsed to 0.355 after lag correction.

no nested Clark-West statistic reaches $|t| = 3.0$). The threshold has no demonstrated power against look-ahead signals and their noise-degraded versions in this window, so the null also rests on the point estimates and the Clark-West p -values (Section 7).

- No signal-based strategy statistically outperforms the simple 12/VIX benchmark (maximum Sharpe improvement = +0.030, behavioral sentiment, DM $|t| = 0.72$, not significant).
- VIX’s forecasting power is time-invariant: the VIX–realized volatility R^2 ranges from 0.24 (low-volatility QE era, 2012–2020) to 0.64 (post-dot-com era, 2000–2007), with a coefficient of variation of only 0.33.
- Even the most economically motivated signal—the variance risk premium—achieves an in-sample t -statistic of 3.51 for return prediction but delivers out-of-sample $R^2 < 0.3\%$.

We interpret this null through the lens of option-market information aggregation. VIX is not simply a backward-looking statistical estimate; it is the forward-looking consensus of thousands of option traders who collectively incorporate cross-asset signals, macroeconomic data, sentiment, and term structure information into their pricing. The thirteen signal families we test are precisely the inputs that option traders use. VIX, by construction, already contains them.

This interpretation leads to our second contribution: a reconceptualization of VT strategies. Rather than viewing VT as an alpha-generating timing strategy, we propose the *drawdown insurance* framework. VT sacrifices average returns (3.49%/year for 12/VIX, 2.12%/year for EWMA) in exchange for maximum drawdown reduction. This trade-off is welfare-improving for investors with CRRA risk aversion $\gamma \geq 4.5$, making VT rational for moderately risk-averse investors even though it underperforms buy-and-hold on standard metrics.

Our third contribution concerns the research frontier. The informative null delineates what *cannot* work (daily signals already in the option-market information set) from what

might work: higher-frequency realized measures (5-minute data) that capture intraday dynamics not fully reflected in closing VIX, and genuinely exogenous shocks (e.g., geopolitical events, pandemic declarations) that arrive between option-market pricing updates. Preliminary results using HAR-RV models with 5-minute S&P 500 data show a 41.8% QLIKE improvement over daily-frequency models, suggesting that the intraday frontier remains open.

The remainder of the paper is organized as follows. Section 2 reviews the theoretical and empirical foundations of VIX as a benchmark. Section 3 describes the data and the thirteen signal families. Section 4 details the forecast design, loss function choice, and the Patton (2011) proxy-robustness framework. Section 5 explains the statistical evaluation framework, including the Model Confidence Set of Hansen et al. (2011). Section 6 describes the volatility-timing strategy design. Section 7 presents the main results, criterion-dependent model rankings, and robustness checks. Section 8 discusses why the null is informative, including economic significance via VaR/ES backtesting. Section 9 concludes.

2 Why VIX is the Benchmark

2.1 Theoretical Foundations

The VIX index, introduced by the CBOE in 1993 and reformulated in 2003, measures the risk-neutral expected variance of the S&P 500 over the next 30 calendar days. Under the model-free framework of Britten-Jones and Neuberger (2000) and Jiang and Tian (2005), VIX squared equals the risk-neutral expectation of integrated variance:

$$\text{VIX}^2 = \frac{2}{\tau} \sum_i \frac{\Delta K_i}{K_i^2} e^{r\tau} Q(K_i) - \frac{1}{\tau} \left(\frac{F}{K_0} - 1 \right)^2, \quad (1)$$

where τ is time to expiration, K_i are strike prices, $Q(K_i)$ are midpoint option prices, F is the forward index level, and K_0 is the first strike below F . This construction aggregates information from the entire volatility smile, making VIX a sufficient statistic for the

market’s assessment of near-term volatility under risk-neutral pricing.

The link between VIX and future realized volatility (RV) is governed by the variance risk premium (VRP):

$$\mathbb{E}^Q[\sigma_{t,t+\tau}^2] = \mathbb{E}^P[\sigma_{t,t+\tau}^2] + \text{VRP}_t, \quad (2)$$

where $\mathbb{E}^Q$ and $\mathbb{E}^P$ denote risk-neutral and physical expectations, respectively. VIX overestimates realized volatility on average (the VRP is positive approximately 85% of days in our sample), but this bias is systematic and does not impair VIX’s role as a forecasting benchmark—it merely shifts the intercept of the VIX–RV regression.

2.2 Empirical Evidence for VIX Forecasting Power

The volatility forecasting literature is vast; [Poon and Granger \(2003\)](#) provide the definitive survey, cataloguing over 90 studies and concluding that implied volatility generally dominates time-series models. Within this literature, [Christensen and Prabhala \(1998\)](#) show that VIX subsumes the information in historical volatility for forecasting realized volatility. [Jiang and Tian \(2005\)](#) demonstrate that model-free implied volatility dominates both GARCH and Black-Scholes implied volatility in encompassing regressions. [Bekaert and Hoerova \(2014\)](#) decompose VIX into a conditional variance component and a variance risk premium, finding that the conditional variance component tracks realized volatility closely.

The realized volatility framework underpinning our forecast target originates with [Andersen et al. \(2003\)](#), who show that modeling and forecasting realized volatility—constructed from high-frequency returns—provides a natural benchmark for evaluating volatility forecasts. Our use of 22-day realized volatility as the forecast target follows this tradition. [Patton \(2011\)](#) demonstrates that loss function choice matters when the volatility proxy is imperfect, motivating our discussion of QLIKE loss alongside squared errors.

More recently, [Bozovic \(2024\)](#) examines VIX-managed portfolios and finds that VIX-based timing produces economically significant improvements in portfolio performance, though the mechanism operates through drawdown reduction rather than return enhancement—

consistent with our insurance framework.

2.3 The Information Aggregation Argument

We argue that VIX’s forecasting dominance is not accidental but structural. Options markets aggregate information from heterogeneous traders—fundamental analysts who track earnings volatility, macro strategists who monitor the yield curve, systematic traders who exploit cross-asset correlations, and market makers who process order flow. Each of the thirteen signal families we test represents one channel of information that at least some option traders monitor. VIX, as the price-weighted consensus of their collective activity, already incorporates these inputs.

This argument generates a testable prediction: signals that are *already in the option-market information set* should show zero incremental forecasting power after controlling for VIX, while signals that are *outside* this set (e.g., truly private information, intraday microstructure) might still add value. Our empirical results are consistent with this prediction.

2.4 Prior Horse Races

The canonical volatility model horse race is [Hansen and Lunde \(2005\)](#), who compare 330 ARCH-type models for forecasting daily exchange rate and IBM stock volatility using the Model Confidence Set (MCS) framework subsequently formalized in [Hansen et al. \(2011\)](#). Their central finding—that no model significantly outperforms the simple GARCH(1,1) when realized volatility is the benchmark—sets the precedent for our study, which extends the horse race from parametric models to signal families and uses VIX rather than GARCH as the benchmark to beat. [Patton and Sheppard \(2015\)](#) decompose realized volatility into good (upside) and bad (downside) components and show that the asymmetric component improves volatility forecasts, a finding we revisit in our variance risk premium tests.

Among signal-specific studies, [Corsi \(2009\)](#) proposes the HAR-RV model using multi-frequency realized volatility components and shows it forecasts well, but does not bench-

mark against VIX. [Bucci \(2020\)](#) finds that neural networks can improve realized volatility forecasts but does not test whether the improvement survives after conditioning on VIX. [Bollerslev et al. \(2009\)](#) show that VRP predicts equity *returns* (not volatility), but subsequent out-of-sample tests have been mixed.

Our contribution differs from these studies in three ways: (i) we test thirteen signal families simultaneously rather than one at a time; (ii) we use a unified evaluation pipeline with consistent lag, cost, and statistical conventions; and (iii) we frame the null as informative rather than disappointing, connecting it to option-market efficiency.

3 Data and the Thirteen Signal Families

3.1 Core Data

Our primary dataset consists of daily closing prices for the S&P 500 index (via SPY ETF), the CBOE Volatility Index (VIX), VIX 3-month futures index (VIX3M), gold (GLD ETF), and the 10-year and 3-month U.S. Treasury yields (TNX, IRX). All price data are sourced from Yahoo Finance. The full sample spans January 1993 to March 2026, yielding 8,325 trading days for the VIX–SPY pair. Individual signal families have shorter samples depending on data availability (detailed in [Table 1](#)).

We compute 22-day realized volatility as:

$$RV_t^{(22)} = \sqrt{\frac{252}{22} \sum_{i=0}^{21} r_{t-i}^2}, \quad (3)$$

where $r_t = \ln(P_t/P_{t-1})$ is the daily log return. This serves as the forecast target for all volatility prediction exercises.

3.2 The Thirteen Signal Families

We pre-specify thirteen signal families spanning seven broad categories: cross-asset, derivatives-based, behavioral, macroeconomic, alternative data, calendar, and economic-policy/financial-stress uncertainty. [Table 1](#) summarizes each family. Families 12–13 test

whether canonical alt-data uncertainty proxies (EPU and financial-stress indices) carry orthogonal information beyond VIX.

Table 1: The Thirteen Signal Families

#	Signal Family	Category	Construction	Sample	N
1	Cross-asset vol momentum	Cross-asset	5d ΔRV : TLT/USO/UUP/GLD/HYG-LQD	2010–2026	4,053
2	VIX term structure	Derivatives	VIX/VIX3M ratio (contango/backwardation)	2008–2026	4,566
3	Behavioral sentiment	Behavioral	SKEW + VIX/VIX3M + VIX momentum	2011–2026	3,760
4	Variance risk premium	Derivatives	VRP = VIX – 22d realized vol	2006–2026	5,090
5	Multi-asset optimization	Portfolio	MaxDiv/MinVar/RP, 2–9 assets	2023–2026	811
6	Equal risk contribution	Portfolio	Maillard et al. (2010) ERC	2023–2026	811
7	Bitcoin volatility	Cross-asset	BTC 22d RV, BTC–VIX Granger	2015–2026	2,800
8	Yield curve slope	Macro	10Y – 3M Treasury spread	2005–2026	5,337
9	Google Trends fear	Alt. data	5 fear search terms (weekly)	2010–2026	797
10	Overnight VIX	Microstructure	$ VIX_{\text{open}} - VIX_{\text{close}, t-1} $	2010–2026	4,082
11	Calendar anomaly	Calendar	Monthly/seasonal dummies	1993–2026	8,325
12	Economic policy uncertainty	Alt. data	USEPU + WLEMU composite (Baker et al., 2016)	2018–2026	431 [†]
13	Financial-stress indices	Macro	NFCI + ANFCI + STLFSI4 composite (Brave and Butters, 2011; Kliesen and Smith, 2010)	2018–2026	431 [†]

Notes: All signals are computed using information available at market close on day $t - 1$ and applied to day t to daily by forward-filling. Sample sizes vary by data availability. VIX3M begins in 2008; Bitcoin reliability begins in 2004 but we use 2010+ for stability. [†]Families 12 and 13 operate on weekly observations (Friday-frequency; $N = 431$ refers to the weekly count over 2018–2026 (260 IS + 171 OOS). Publication delays for §4.6: the daily Economic Policy Uncertainty series (USEPU, WLEMU) clear their one-business-day release shift(1), whereas the weekly financial-stress series (NFCI, ANFCI, STLFSI4), released five to six calendar days, both are aligned to the weekly W-FRI evaluation grid. This is the “corrected” timing whose robustness against misalignment is reported in §7.10.

3.2.1 Family 1: Cross-Asset Volatility Momentum

Cross-asset volatility spillovers have been documented by Diebold and Yilmaz (2012) and Barunik et al. (2016). We compute the 5-day change in 22-day realized volatility for five asset classes: bonds (TLT), oil (USO), U.S. dollar (UUP), gold (GLD), and credit (HYG minus LQD spread). A composite momentum signal averages the normalized changes. The hypothesis is that volatility shocks propagate across asset classes with a lead-lag structure.

3.2.2 Family 2: VIX Term Structure

The VIX term structure—the ratio of spot VIX to VIX3M—has been proposed as a regime indicator. Persistent contango ($VIX < VIX3M$) is associated with complacency, while backwardation signals acute fear (Huang et al., 2019). We test whether this ratio adds information beyond VIX level for both volatility prediction and strategy timing.

3.2.3 Family 3: Behavioral Sentiment Index

Investor sentiment affects asset prices through both rational and behavioral channels (Baker and Wurgler, 2006). We construct a composite Behavioral Sentiment Index (BSI) using four components: the CBOE SKEW index, the VIX/VIX3M ratio, VIX 22-day momentum, and VIX level. Each component is converted to a percentile rank over a 252-day rolling window, and the BSI is the equal-weighted average.

3.2.4 Family 4: Variance Risk Premium

The variance risk premium ($VRP = VIX - RV^{(22)}$) has attracted attention since Bollerslev et al. (2009) showed it predicts equity returns. We distinguish between VRP as a *return* predictor (Bollerslev’s original finding) and VRP as a *volatility* predictor, testing whether VRP carries orthogonal information about future volatility after controlling for VIX level.

3.2.5 Family 5: Multi-Asset Portfolio Optimization

DeMiguel et al. (2009) famously showed that $1/N$ portfolios outperform optimized portfolios out of sample. We test four optimization methods (equal-weight, minimum variance, maximum diversification, risk parity) across seven asset subsets (2 to 9 assets including SPY, GLD, TLT, QQQ, EEM, VWO, BND, DBC, IEFA), evaluating whether expanding the asset universe beyond SPY/GLD improves risk-adjusted returns.

3.2.6 Family 6: Equal Risk Contribution

The Equal Risk Contribution (ERC) portfolio (Maillard et al., 2010) equalizes marginal risk contributions through constrained optimization. We implement ERC with monthly

hold-and-drift rebalancing and test 2-, 3-, and 4-asset variants against the static 50/50 SPY/GLD benchmark.

3.2.7 Family 7: Bitcoin Volatility

Bitcoin has been proposed as a “digital gold” and potential diversifier (Bouri et al., 2017). We test whether Bitcoin realized volatility contains Granger-causal information for VIX and whether a Bitcoin allocation improves portfolio performance. The *fear channel*—whether crypto stress transmits to traditional markets—is examined through asymmetric Granger causality.

3.2.8 Family 8: Yield Curve Slope

The yield curve slope (10Y – 3M Treasury spread) is a well-known recession predictor (Estrella and Mishkin, 1998). We test whether curve inversion contains volatility information beyond VIX, exploiting the fact that yield curve movements are slow-moving and potentially capture different risk horizons than VIX.

3.2.9 Family 9: Google Trends Fear Proxy

Internet search behavior provides a real-time measure of retail investor attention and fear (Da et al., 2011; Preis et al., 2013). We construct a composite Google Fear Index from the search volume of five terms (“stock market crash,” “recession,” “VIX,” “market crash,” “bear market”), using rolling z -scores over a 52-week window.

3.2.10 Family 10: Overnight VIX Changes

Overnight price changes reflect the processing of after-hours news, earnings announcements, and geopolitical events (Lou et al., 2019). We use the absolute overnight VIX change ($|VIX_{\text{open},t} - VIX_{\text{close},t-1}|$) as a proxy for the information content of overnight news. This family is the sole exception to the close-of- $t-1$ information-set convention used by Families 1–9 and 11: its forecast origin is the day- t opening auction (9:30 ET), at which $VIX_{\text{open},t}$ is realized, and the target is the forward realized volatility measured

from day t onward. Because both inputs ($VIX_{\text{open},t}$ and $VIX_{\text{close},t-1}$) are observed no later than the day- t open while the target accumulates strictly after it, the signal contains no look-ahead; the intraday origin is a legitimate timing convention rather than a leak, and it is disclosed here so that the one-day-shift enforcement of the other daily families is not misread as applying to this one.

3.2.11 Family 11: Calendar Anomaly

The “Halloween effect” and other calendar anomalies have been documented in equity markets ([Bouman and Jacobsen, 2002](#)). We test whether VIX-based timing is simply a disguised calendar strategy by examining the correlation between VIX-implied weights and monthly/seasonal dummies.

3.2.12 Family 12: Economic Policy Uncertainty

[Baker et al. \(2016\)](#) construct the Economic Policy Uncertainty (EPU) index from newspaper coverage, tax-code expiration counts, and forecaster disagreement, and document its association with elevated volatility in equity and macroeconomic series. The daily U.S. EPU index (FRED: USEPUINDXD) and the weekly World Leading Economic Monetary Uncertainty index (WLEMUINDXD) provide canonical text-based uncertainty proxies that are widely cited in subsequent volatility-forecasting work. We test whether their composite carries information about future realized volatility orthogonal to VIX. Because the daily EPU series is publicly released with a one-business-day lag relative to the underlying news sample, the prior week’s value is available before each weekly forecast, and all EPU inputs enter at shift(1) on the weekly W-FRI evaluation grid shared by Families 12–13 (the release-calendar convention of §4.6, verified in the robustness audit of §7.10). Under this convention the alt-data-only EPU composite (Model M3, which *replaces* VIX) is dominated by the VIX baseline with Diebold–Mariano $t = -2.54$ in out-of-sample QLIKE comparison; when EPU is instead *added* to VIX (the nested increment), it contributes no incremental accuracy under the Clark–West test (Section 7)—redundant with, rather than harmful to, the option-implied signal.

3.2.13 Family 13: Financial-Stress Indices

The Chicago Fed’s National Financial Conditions Index (NFCI; [Brave and Butters, 2011](#)) and the St. Louis Fed’s Financial Stress Index (STLFSI; [Kliesen and Smith, 2010](#)) aggregate prices, spreads, and balance-sheet quantities into composite financial-conditions readings that have been shown to lead recession dates and credit cycles. The Adjusted NFCI (ANFCI) removes the macroeconomic component to isolate the residual financial signal. These three indices—released weekly each Wednesday with a one-week lag for NFCI and STLFSI4—constitute the most academically-canonical set of financial-stress proxies and are obvious candidates for delivering incremental information beyond VIX during stress episodes. We test their composite against the VIX baseline under publication-lag-corrected alignment (shift(2) for NFCI and ANFCI weekly-to-daily forward-fill; shift(1) for STLFSI4); the alt-data-only composite (Model M4, which *replaces* VIX) loses to VIX with $t = -3.00$ under the original convention and $t = -3.61$ under the corrected convention. As with EPU, the standalone stress composite underperforms VIX; and when the stress block is instead *added* to VIX, the nested Clark-West increment is statistically zero (Section 7)—the indices are redundant given VIX rather than a source of information that VIX omits.

4 Forecast Design and Benchmark Models

4.1 Unified Forecast Pipeline

All forecast evaluations follow a unified pipeline to ensure comparability:

1. **Signal construction:** Each signal is computed using only information available at the close of day $t - 1$.
2. **Lag enforcement:** The signal is explicitly shifted by one trading day (`signal = signal.shift(1)`) before any use in forecasting or portfolio construction.
3. **Rolling estimation:** Where applicable, parameters are estimated on a rolling window of 252 trading days (approximately one year), re-estimated daily.

4. **Forecast target:** The dependent variable is 22-day forward realized volatility,

$$RV_{t+1:t+22}.$$

4.2 VIX Benchmark

The VIX benchmark model is a simple linear regression:

$$RV_{t+1:t+22} = \alpha + \beta \cdot \text{VIX}_t + \varepsilon_t. \quad (4)$$

This is our primary benchmark. Any candidate signal must demonstrate statistically significant improvement over this specification.

4.3 Incremental Information Test

For each signal family j , we estimate:

$$RV_{t+1:t+22} = \alpha + \beta_1 \cdot \text{VIX}_t + \beta_2 \cdot \text{Signal}_{j,t} + \varepsilon_t, \quad (5)$$

and test $H_0 : \beta_2 = 0$. We report both the in-sample t -statistic for β_2 and the partial correlation $r_{\text{partial}} = \text{cor}(RV, \text{Signal}_j \mid \text{VIX})$.

4.4 Loss Functions and Proxy Robustness

A critical methodological issue in volatility forecast evaluation is that the true conditional variance σ_t^2 is unobservable; any evaluation must use an imperfect proxy. [Patton \(2011\)](#) shows that certain loss functions produce consistent model rankings regardless of the proxy used, while others do not. Specifically, QLIKE and MSE applied to σ^2 (or an unbiased proxy thereof) are *proxy-robust*—they rank models consistently whether the proxy is realized variance from 5-minute returns, daily squared returns r_t^2 , or any other conditionally unbiased estimator.

This result has a subtle but important implication for our evaluation. Daily absolute returns $|r_t|$ are *not* an unbiased proxy for σ_t : the ratio $\mathbb{E}[r_t^2]/\mathbb{E}[|r_t|]^2 = \pi/2 \approx 1.571$ under

normality (empirically 2.44 in our sample, reflecting heavy tails), so $|r_t|$ systematically understates σ_t by approximately 19.6%. Forecast evaluations conducted on $|r_t|$ targets can produce different model rankings than evaluations on r_t^2 targets—a finding we confirm empirically in Section 7.6.

We therefore adopt a three-metric evaluation framework:

1. **QLIKE on r_t^2** (primary): The quasi-likelihood loss $\text{QLIKE} = \ln h_t + r_t^2/h_t$, where h_t is the model’s variance forecast. This is proxy-robust per Patton (2011) and penalizes under-prediction more heavily than over-prediction, consistent with the risk management motivation of volatility forecasting.
2. **Spearman rank correlation** (secondary): The rank correlation between model forecasts and realized r_t^2 , providing a distribution-free measure of monotonic association that is robust to outliers and does not require specifying a loss function.
3. **Out-of-sample R^2** (benchmark comparison): Two variants described below, providing direct comparability with the existing literature.

4.5 Out-of-Sample R^2

We report two complementary measures of out-of-sample forecasting accuracy. The first follows the standard convention of Campbell and Thompson (2008), defining R_{OOS}^2 relative to the prevailing historical mean:

$$R_{\text{OOS,CT}}^2 = 1 - \frac{\sum_{t \in \text{OOS}} (RV_t - \hat{RV}_t^{\text{model}})^2}{\sum_{t \in \text{OOS}} (RV_t - \bar{RV}_{1:t-1})^2}, \quad (6)$$

where $\bar{RV}_{1:t-1}$ is the expanding-window historical mean of realized volatility. Positive values indicate that the model forecasts better than the naive historical average. This metric is directly comparable to the large literature on return and volatility predictability that uses the Campbell-Thompson framework.

The second measure, which is our primary metric, defines R_{OOS}^2 relative to VIX-only

forecasts:

$$R_{\text{OOS,VIX}}^2 = 1 - \frac{\sum_{t \in \text{OOS}} (RV_t - \hat{RV}_t^{\text{model}})^2}{\sum_{t \in \text{OOS}} (RV_t - \hat{RV}_t^{\text{VIX}})^2}, \quad (7)$$

where positive values indicate improvement over VIX. This is a more stringent test because VIX is itself a strong predictor (Campbell-Thompson $R_{\text{OOS,CT}}^2$ for VIX alone ranges from 0.24 to 0.64 across eras). Our research question is specifically whether any signal adds value *beyond* VIX, making the VIX-relative metric the appropriate primary measure. However, reporting both metrics ensures that our results are comparable to the existing literature and demonstrates that VIX’s own forecasting power is substantial—many signals that achieve positive $R_{\text{OOS,CT}}^2$ (i.e., beat the historical mean) still show zero $R_{\text{OOS,VIX}}^2$ (i.e., add nothing beyond VIX).

4.6 Publication-Delay Convention

A subtle but consequential timing issue arises whenever an economic or financial-stress signal published with a real-world release lag is treated as if it were available the moment the observation period ends. For market-derived series (VIX, term structure, returns) the observation is its own release: closing prices are known at the close. For FRED macro and financial-stress series the observation date and the release date diverge by one to six calendar days, and a forecaster cannot condition on an observation until *after* it has been published. Ignoring this gap silently introduces look-ahead bias—trivial for daily series but material for the weekly financial-stress block, where the gap is sometimes longer than the data frequency itself.

Release calendar. Table 2 summarizes the canonical release lags for the FRED inputs to Families 12–13.

Table 2: Publication Release Calendar for Families 12–13

Family	Series (FRED ID)	Native frequency	Release lag	Required shift
12 (EPU)	USEPUINDXD	Daily	$t + 1$ business day	shift(1) (weekly grid)
12 (EPU)	WLEMUINDXD	Daily	$t + 1$ business day	shift(1) (weekly grid)
13 (FinStress)	NFCI	Weekly (Fri obs)	$t + 5$ calendar days	shift(2) (weekly grid)
13 (FinStress)	ANFCI	Weekly (Fri obs)	$t + 5$ calendar days	shift(2) (weekly grid)
13 (FinStress)	STLFSI4	Weekly (Fri obs)	$t + 6$ calendar days	shift(2) (weekly grid)

Notes: “Daily” rows refer to the series sampling frequency, not the forecasting frequency. “shift(k)” means the signal observed at period t is treated as conditioning information at period $t + k$. The daily and weekly shifts are consistent: at weekly aggregation, a daily series with a one-business-day release lag still appears in the prior week’s available information set, while a Friday-observed weekly series released the following Wednesday is not in the prior week’s set and must wait an additional week. Release lags are verified against the FRED ALFRED real-time vintage archive.

Notation. Let $X_{j,t}$ denote signal family j observed at period t and δ_j its release lag in periods. The information set available to a forecaster at the start of period $t + 1$ is

$$\mathcal{F}_t = \{X_{j,t-\delta_j+1}\}_{j=1}^J, \quad (8)$$

which collapses to the standard $\mathcal{F}_t = \{X_{j,t}\}_{j=1}^J$ only when $\delta_j = 1$ for all j . For Family 13 (the weekly financial-stress series, released five to six calendar days after their Friday reference date) we set $\delta = 2$ at weekly frequency; for Family 12 (the daily EPU series, which clear their release lag within one business day) we set $\delta = 1$. Equivalently, all primary results in this paper use `Signal.shift(1)` for the EPU inputs and `Signal.shift(2)` for the financial-stress inputs—the “corrected” timing whose sensitivity is examined in §7.10.

Why this matters. Our companion robustness experiment re-estimates every Family 12–13 specification under both the conventional shift(1) and the publication-delay-

corrected shift(2). For SPY the null finding strengthens: the M4 financial-stress specification moves from DM $t = -3.00$ under shift(1) to $t = -3.61$ under shift(2), pushing it further past the Harvey threshold of $|t| > 3.0$ in the wrong direction. The single *apparent* cross-asset niche under shift(1)—TLT volatility forecast with the financial-stress block, DM $t = +3.74$ —collapses to $t = +1.96$ under shift(2), which fails the Harvey threshold. All other 14 of the 16 cells are stable across the two shift conventions. The convention adopted in the main text—the “corrected” timing that applies shift(1) to the daily EPU series and shift(2) to the weekly financial-stress series—is therefore the release-calendar-respecting one, and a reader applying the conventional uniform shift(1) would obtain a single weakly-significant cell whose collapse under shift(2) is informative on its own; we report both in Section 7 for completeness.

Generalization. The convention extends naturally to other release-lagged data: for daily macro series with a one-business-day lag (e.g., daily EPU) the daily-frequency shift is 2; for monthly series with a one-month release lag the monthly shift is 2; for irregularly-released series (e.g., CPI, NFP) the appropriate shift is the maximum business-day gap between observation reference period and release. Implementing this convention requires only a one-line correction at signal construction and does not interact with any subsequent estimation choice.

5 Statistical Evaluation and Multiple-Testing Control

5.1 Diebold-Mariano Test

We use the [Diebold and Mariano \(1995\)](#) test to compare forecast accuracy, with squared forecast errors as the loss function:

$$d_t = e_{t,\text{VIX}}^2 - e_{t,\text{model}}^2, \quad (9)$$

where $e_{t,k}$ denotes the forecast error from model k . The DM statistic is:

$$\text{DM} = \frac{\bar{d}}{\sqrt{\hat{V}(\bar{d})/T}}, \quad (10)$$

with Newey and West (1987) heteroskedasticity and autocorrelation consistent (HAC) standard errors to account for serial correlation in multi-step-ahead forecasts (bandwidth equal to the canonical rule for the forecast horizon: lag 35 at the 1,861-observation out-of-sample sample of the daily families, and 4 for the weekly families).

5.2 Nested-Model Correction: Clark-West and Small-Sample Adjustment

A subset of our comparisons is *nested*: the VIX-only forecast is the special case of the VIX-plus-signal forecast (Equation 5) obtained by restricting the signal coefficient to zero. Clark and West (2007) show that when nested models are compared by mean squared prediction error, the standard Diebold-Mariano statistic is biased *against* the larger model, because under the null that the extra predictor is useless the additional estimated parameter injects pure estimation noise that inflates the larger model’s out-of-sample loss. For a paper whose central claim is a *null*—that no signal improves upon VIX—this bias is self-serving: it can make the augmented model look actively worse when it is merely redundant, and it makes the null appear stronger than the evidence warrants. We therefore report, for every nested comparison, the Clark-West MSPE-adjusted statistic based on

$$\hat{f}_t = (RV_t - \hat{RV}_t^{\text{VIX}})^2 - \left[(RV_t - \hat{RV}_t^{\text{VIX+sig}})^2 - (\hat{RV}_t^{\text{VIX}} - \hat{RV}_t^{\text{VIX+sig}})^2 \right], \quad (11)$$

tested one-sided against $H_0 : \mathbb{E}[\hat{f}_t] \leq 0$ (“the signal does not improve on VIX”). The bracketed correction term $(\hat{RV}_t^{\text{VIX}} - \hat{RV}_t^{\text{VIX+sig}})^2$ removes the estimation-noise penalty, so a genuinely informative signal produces a large positive statistic. Because Clark-West is strictly more powerful than the standard test at detecting a true incremental

predictor, a null that survives it is robust rather than an artifact of the standard test’s conservativeness. We apply the [Harvey et al. \(1997\)](#) small-sample correction, multiplying the statistic by $\sqrt{(T + 1 - 2h + T^{-1}h(h - 1))/T}$ and referring it to the t_{T-1} distribution, in two places: to the weekly nested comparisons (one-step forecasts, $h = 1$, where the factor equals 0.997 and is immaterial) and to the daily statistics reported under the lag-21 convention in Table 4 ($h = 22$, factor 0.99). The daily Clark-West statistics of Table 3 are computed at the canonical bandwidth (lag 35 at this sample size) without the factor and are referred to the standard normal distribution; Section 7 reports both readings for the daily families. We note that comparisons in which the alt-data specification *replaces* VIX rather than augmenting it (the pure alt-data models M3 and M4 of Section 7) are *non-nested*, so the standard Diebold-Mariano test applies to them directly and the Clark-West adjustment is neither needed nor appropriate.

5.3 Multiple Testing

When many predictors are tested, conventional $|t| > 1.96$ thresholds produce excessive false discoveries. [Harvey et al. \(2016\)](#) address this problem in the context of asset pricing, presenting a framework in which the appropriate significance threshold depends on the number of tested strategies and the assumed prior probability of a true discovery. Their analysis—applied to roughly 300 factors in the cross-section literature—yields an approximate threshold of $|t| > 3.0$ under conservative priors. We adopt $|t| > 3.0$ as an approximate conservative threshold motivated by their analysis. Of our 13 signal families, 10 are evaluated via DM regression (families 1–4 and 8–13); families 5–6 are portfolio strategies tested via Sharpe-difference bootstrap and family 7 is tested via Granger causality, both reported separately in Section 7.

Of these 10 families, the seven daily ones (1–4, 8, 10, 11) compare nested forecasts, for which the raw Diebold-Mariano statistic is not a valid test (Section 5.2); their inference uses the Clark-West statistic. We report one-sided Clark-West p -values without a Holm-Bonferroni adjustment ([Holm, 1979](#)), because none of them rejects at the 5% level at the canonical bandwidth, so an adjustment has nothing to correct. The only unadjusted

one-sided p -values below 5% belong to Family 2 at Newey-West lag 21 (0.0435 without the small-sample factor and 0.0452 with the [Harvey et al. \(1997\)](#) small-sample factor); at the canonical lag 35 its p -value is 0.0580. A Bonferroni correction across the seven nested families would require $p \leq 0.05/7 = 0.007$, and no reading of Family 2 meets it. The two weekly alt-data families (12–13) are non-nested and their raw tests are significant only in the harmful direction, so the directional “signal beats VIX” null is not rejected for any family.

5.4 Model Confidence Set

Beyond pairwise DM tests, we employ the Model Confidence Set (MCS) procedure of [Hansen et al. \(2011\)](#) to identify the set of models that cannot be statistically distinguished from the best model at a given confidence level. The MCS procedure has three advantages over pairwise testing: (i) it controls for the joint testing problem without requiring a pre-specified benchmark; (ii) it produces a *set* of superior models rather than a single winner, honestly reflecting the uncertainty in model ranking; and (iii) it accounts for the dependence structure among forecast errors through block bootstrap resampling (block length = 22, matching the forecast horizon).

Formally, let $\mathcal{M}_0$ denote the full set of candidate models. At each step, the MCS procedure tests the null hypothesis that all surviving models have equal expected loss. If rejected, the worst-performing model is eliminated. The procedure terminates when the null cannot be rejected at significance level α , yielding the model confidence set $\hat{\mathcal{M}}_\alpha^*$. We report MCS results at $\alpha = 0.10$.

We apply the MCS across our volatility model families (GJR-GARCH, AMEM, MEM, GARCH, HAR, EWMA) using QLIKE on r_t^2 as the loss function. As shown in [Section 7.6](#), the MCS membership is criterion-dependent—different loss functions can produce different confidence sets—which itself strengthens the VIX sufficiency argument.

5.5 Cross-Era Validation

To prevent any single market regime from driving conclusions, we partition the sample into five non-overlapping eras:

1. **Dot-Com** (1993–2000): Rising equity markets, Asian crisis, LTCM
2. **Post-Dot-Com** (2000–2007): Tech bust, recovery, housing boom
3. **GFC** (2007–2012): Global financial crisis, European debt crisis
4. **Low-Vol QE** (2012–2020): Quantitative easing, low volatility regime
5. **COVID/Inflation** (2020–2026): Pandemic shock, inflation, rate hikes

A finding is considered robust if it holds in at least 4 of 5 eras.

6 Volatility-Timing Strategy Design

6.1 The 12/VIX Strategy

Our primary VT strategy is the 12/VIX rule:

$$w_t^{\text{SPY}} = \min \left(\frac{12}{\text{VIX}_{t-1}}, 1 \right), \quad (12)$$

where the equity weight is capped at 100% and the complement $(1 - w_t^{\text{SPY}})$ is allocated to GLD (or cash in pre-GLD eras). This strategy has several desirable properties:

- **Zero parameters:** The numerator 12 approximates long-run average VIX.
- **Smooth weights:** The concave $1/x$ mapping produces gradual weight changes, limiting turnover (mean daily $|\Delta w| = 0.035$, annual turnover $\approx 8.9\times$).
- **Lag-robust:** Because weights change slowly, the 1-day lag between signal and execution has negligible impact.

- **Crowding-resistant:** Market impact analysis shows that even \$10 billion in AUM would reduce Sharpe by less than 60 basis points, and the VIX feedback loop is damping (converging) rather than amplifying.

6.2 Baseline: 50/50 SPY/GLD

The baseline is a static 50/50 allocation to SPY and GLD with monthly rebalancing. This allocation is approximately optimal by construction: SPY annualized volatility (19.3%) $\approx$ GLD annualized volatility (18.3%), so the 50/50 split closely approximates the risk parity solution (Maillard et al., 2010).

6.3 Transaction Costs

We deduct 5 basis points per leg for each weight change:

$$\text{TX}_t = 0.0005 \times (|w_t^{\text{SPY}} - w_{t-1}^{\text{SPY}}| + |w_t^{\text{GLD}} - w_{t-1}^{\text{GLD}}|). \quad (13)$$

Net portfolio return is $r_t^{\text{net}} = r_t^{\text{gross}} - \text{TX}_t$.

7 Main Results and Robustness

7.1 Volatility Forecasting: No Signal Beats VIX

Table 3 presents the main results for all thirteen signal families. For the seven daily families, whose augmented forecasts nest the VIX-only forecast, the table reports the Clark-West statistic as the inference-bearing test and the Diebold-Mariano statistic as a diagnostic only (Section 5.2). The key finding is unanimous at the Harvey threshold: no signal produces a statistically significant out-of-sample improvement over VIX; two of the canonical alt-data families (EPU and financial stress) are reported here in their standalone specification (M3/M4, which replace VIX) and are signed in the *worse-than-VIX* direction, i.e., the alt-data-only forecast is significantly worse than the VIX-only baseline. When these same series are instead added to VIX, the nested Clark-West

increment is statistically zero (Section 5.2), so the direction reflects redundancy under a standalone reading rather than harm.

Table 3: Volatility Forecasting: Thirteen Signal Families vs. VIX

#	Signal Family	Partial $r VIX$	IS ΔR^2	IS t -stat	$R^2_{OOS,CT}$	$R^2_{OOS,VIX}$	DM $ t $ (diag.)	Raw p (diag.)	CW t	CW p
1	Cross-asset vol momentum	0.087	0.032	—	0.311	0.006	0.22	0.826	+0.68	0.249
2	VIX term structure	0.181	0.033	17.6***	0.345	0.055	1.20	0.229	+1.57	0.058
3	Behavioral sentiment	0.091	0.004	5.58***	0.315	0.011	0.70	0.486	+0.89	0.188
4	Variance risk premium	0.054	< 0.003	3.51**	0.303	-0.012	0.68	0.495	-0.20	0.578
5	Multi-asset optimization	—	—	—	—	—	—	—	—	—
6	Equal risk contribution	—	—	—	—	—	—	—	—	—
7	Bitcoin volatility	0.178	—	—	—	—	—	—	—	—
8	Yield curve slope	< 0.08	0.009	2.51	0.309	-0.003	0.23	0.818	+0.10	0.461
9	Google Trends fear	0.271	0.038	7.92***	—	—	0.67 [§]	0.504 [§]	—	—
10	Overnight VIX	0.030	0.005	5.07***	0.307	-0.006	0.63	0.527	-0.41	0.660
11	Calendar anomaly	-0.007	0.012	-2.39*	0.302	-0.013	1.60	0.110	-1.39	0.918
12	Economic Policy Uncertainty [‡]	—	-0.001	—	—	—	2.54 [†]	0.012	—	—
13	Financial Stress Indices [‡]	—	0.005	—	—	—	3.61 ^{††}	< 0.001	—	—
<i>VIX-only benchmark</i>		—	—	—	0.311	—	—	—	—	—
<i>Harvey (2016) approx. threshold</i>							> 3.0		> 3.0	

Notes: Partial $r|VIX$ is the partial correlation between the signal and 22-day forward realized volatility, controlling for VIX level. IS ΔR^2 is the in-sample increase in R^2 from adding the signal to a VIX-only regression. IS t -stat tests $H_0 : \beta_{\text{signal}} = 0$. $R^2_{OOS,CT}$ is the [Campbell and Thompson \(2008\)](#) out-of-sample R^2 relative to the expanding-window historical mean (Equation 6). $R^2_{OOS,VIX}$ is the out-of-sample R^2 relative to VIX-only forecasts (Equation 7). The seven daily nested families (1–4, 8, 10, 11) share one fixed split: the VIX-only and VIX-plus-signal regressions are fitted once on the in-sample window (estimation ends 2018, with a 22-day embargo) and evaluated on $n_{OOS} = 1,861$ daily forecasts from 2019-01-02 to 2026-05-28. DM $|t|$ is the absolute Diebold-Mariano statistic of the squared-error differential between the VIX-plus-signal and the VIX-only forecasts, with [Newey and West \(1987\)](#) HAC standard errors at the canonical bandwidth for $h = 22$ (lag 35 at this sample size) and the t_{n-1} reference distribution; Raw p is its two-sided p -value. In a nested comparison the raw DM statistic is biased against the larger model (Section 5.2), so these two columns are diagnostics and are not a valid test of the null. CW t is the Clark-West MSPE-adjusted statistic at the same bandwidth and CW p its one-sided standard-normal p -value; these two columns carry the inference for Families 1–4, 8, 10 and 11. Table 4 reports the same increments next to the lag-21 convention with the

Harvey et al. (1997) (HLN) correction. For Families 12–13 the specification replaces VIX rather than augmenting it, so the DM statistic is a valid test there (weekly panel, bandwidth = 4); their nested Clark-West increments are reported in the text. Raw p for those two rows is the two-sided p -value of the DM test. No Holm-Bonferroni column is reported: the DM columns do not carry inference for the nested rows, and at the canonical bandwidth no nested daily family has a one-sided Clark-West p -value below 0.058 even before any multiplicity adjustment. Family 1’s IS t -stat is reported as — because the source experiment for that family did not compute an in-sample t -statistic for β_{signal} . Families 5–6 are portfolio strategies without direct volatility forecasting components. Family 7 (Bitcoin) is tested for Granger causality rather than regression. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$ (in-sample only, conventional thresholds). The VIX-only benchmark row is the Campbell-Thompson $R^2_{\text{OOS,CT}}$ of the VIX-only regression on the same split.[§] Family 9 (Google Trends fear) is not re-evaluated on the fixed split, because Google Trends returns a re-sampled relative index on every query and a series retrieved today is not the series behind the original statistic; its DM cells come from the family’s own expanding-window run, are diagnostic in the same sense as the other DM cells, and its two R^2_{OOS} cells are not reported. $R^2_{\text{OOS,VIX}}$ is likewise not reported for Families 12–13.

Families 1–11 use signals lagged one trading day. Families 12–13 follow the publication-delay convention of Section 4.6 (`shift(1)` for the daily EPU series and `shift(2)` for the weekly financial-stress series, both on the weekly W-FRI frequency). Families 12–13 are evaluated on the standalone AR(1)+signal specifications (M3 and M4), which replace VIX rather than augment it, on a 2018–2026 weekly sample; their IS ΔR^2 is the difference in R^2 between that specification and the AR(1)+VIX specification; the partial- r and IS t -stat columns are left blank because the 22-day-forward horizon used for the daily families does not align with the weekly evaluation horizon.

[‡] Signal direction: signed DM t is *negative*, indicating that the VIX-only baseline forecasts *better* than the standalone alt-data specification (M3/M4, which replace VIX). The directional one-sided null—“signal beats VIX”—is therefore not rejected (one-sided $p > 0.99$). Under the standalone (non-nested) reading the two-sided rejection means the alt-data-only model forecasts worse than VIX; under the thesis-relevant nested reading (alt-data *added* to VIX) the Clark-West increment is statistically zero (§5.2), so the correct interpretation is redundancy, not harm. We follow Harvey et al. (2016) and treat these as null findings for the purpose of model selection.

[†] Family 12 (EPU): combined USEPU+WLEMU specification (M3). ^{††} Family 13 (Financial Stress): combined NFCI+ANFCI+STLFSI4 specification reported under the publication-delay-corrected shift(2). The corresponding shift(1) DM $|t|$ is 3.00, which strengthens to 3.61 under shift(2)—both significantly negative, reinforcing the null. The pooled M5 (EPU+FinStress jointly) yields DM $|t| = 1.00$ (raw $p = 0.319$, indistinguishable from VIX) under the same corrected timing.

Note that several signals achieve positive $R_{\text{OOS,CT}}^2$ (i.e., they beat the historical mean) but zero or negative $R_{\text{OOS,VIX}}^2$ (i.e., they add nothing beyond VIX), confirming that VIX itself is a strong forecaster.

Several patterns are noteworthy. First, the dual R^2 metrics separate two questions. All seven daily families achieve a positive Campbell-Thompson $R_{\text{OOS,CT}}^2$ between 0.302 and 0.345, so each forecasts volatility better than the naive historical mean, but the VIX-only benchmark alone already achieves 0.311. Measured against VIX itself, the families split: four have a negative VIX-relative $R_{\text{OOS,VIX}}^2$ (Families 4, 8, 10 and 11, from -0.013 to -0.003) and three a small positive one (Family 1: 0.006, Family 3: 0.011, Family 2: 0.055). None of the three positive values is statistically distinguishable from zero at the Harvey threshold under the Clark-West test; the smallest one-sided p -value belongs to Family 2 (0.058 at the canonical bandwidth).

Second, the Clark-West test, not the raw Diebold-Mariano statistic, carries the inference for the seven nested daily families. At the canonical bandwidth every one-sided Clark-West p -value is at least 0.058 and the largest $|t|$ is 1.57; across the bandwidths we tried (lags 21, 35, 42 and 63) no $|t|$ approaches the Harvey threshold of 3.0. The raw DM statistics ($|t|$ between 0.22 and 1.60, raw $p \geq 0.110$) are diagnostics only. Families 12–13 are two-sided significant in the raw (non-nested, hence valid) tests, raw $p = 0.012$ and $p < 0.001$ respectively, but in the *harmful* direction—their signed DM statistics are negative, meaning the standalone alt-data specifications (M3 and M4, which replace VIX rather than add to it) forecast significantly *worse* than the VIX baseline (see footnote ‡)—so the one-sided “beats VIX” null is not rejected for any family at the canonical bandwidth. One margin deserves the reader’s attention: Family 2 sits at the 5% boundary. Its one-sided Clark-West p -value is 0.0435 at Newey-West lag 21 without the small-sample

factor, 0.0452 at lag 21 with the Harvey-Leybourne-Newbold factor, and 0.0580 at the canonical lag 35 (no factor), the bandwidth of Table 3. The first two readings fall just below 5% and the third just above it, and the readings differ in the bandwidth and, for the second, in the small-sample factor, so which side of the line Family 2 falls on is not a property of the signal alone. We do not read any of the three as evidence of an incremental signal: the statistics ($t = +1.71, +1.69$ and $+1.57$) are far below the Harvey threshold, and a Bonferroni correction across the seven nested daily families would require $p \leq 0.05/7 = 0.007$, which none of them meets. We therefore report no Holm-Bonferroni adjusted p -values for the nested rows, because the raw DM p -values such an adjustment would act on do not carry inference.

The negative signed statistics for the weekly alt-data families warrant a nested-model check, since a self-serving reading of a null result could rest on the standard Diebold-Mariano test’s known conservativeness against a larger model (Section 5.2). We distinguish two questions. The main-table entries for Families 12 and 13 evaluate the pure alt-data specifications (M3 and M4), which *replace* VIX and are therefore *non-nested* against the VIX baseline; there the standard test is valid and the alt-data-only forecasts genuinely lose to VIX. The thesis-relevant *incremental* question—whether alt-data *adds* to VIX—is a nested comparison (VIX+signal vs. VIX), for which we compute the Clark-West MSPE-adjusted statistic on the weekly panel (corrected shift convention, $n_{\text{OOS}} = 170$). The result confirms rather than overturns the null: adding EPU to VIX yields Clark-West $t = +0.12$ (one-sided $p = 0.45$), adding the financial-stress block yields $t = -0.56$ ($p = 0.71$), and the kitchen-sink VIX+all-alt-data specification yields $t = +0.14$ ($p = 0.45$)—none approaching the Harvey threshold, and all well short of conventional significance. Notably, the standard DM statistics on the same nested increments are significantly *negative* ($t = -2.51, -2.82, -2.62$), but the Clark-West adjustment collapses them to essentially zero, revealing that the apparent underperformance is estimation noise from the extra parameters rather than active harm. The incremental content of alt-data beyond VIX is therefore not negative but simply *absent*: VIX sufficiency holds even under the more powerful nested test, and the Harvey-Leybourne-

Newbold small-sample correction (factor 0.997 at $h = 1$) leaves these verdicts unchanged.

The same nested-model logic extends to the daily signal families. The main-table Diebold-Mariano statistics for the daily families are diagnostics only, because these comparisons are nested—the augmented forecast $RV = a + b_1 VIX + b_2 \text{signal}_j$ collapses to the VIX-only benchmark under $b_2 = 0$ (Equation 5)—so the inference rests on the Clark-West increment, and we verify that the daily null, like the weekly one, is not an artifact of the standard test’s conservativeness against the larger model. For the seven daily families whose signal admits a well-defined pinned-data nested specification (Family 1 cross-asset volatility momentum, Family 2 VIX term structure, Family 3 behavioral sentiment, Family 4 variance risk premium, Family 8 yield-curve slope, Family 10 overnight VIX change, and Family 11 calendar anomaly), we compute the Clark-West MSPE-adjusted increment on a fixed 2019–2026 out-of-sample split ($n_{\text{OOS}} = 1,861$; in-sample estimation through 2018), the same split as the daily rows of Table 3, under two bandwidth conventions side by side: Newey-West lag 21 matched to the 22-day-forward horizon with the Harvey-Leybourne-Newbold correction (the convention of Section 5.2), and the canonical bandwidth of the Table 3 columns (lag 35 at this sample size, without the small-sample factor). Table 4 reports both. The verdict is unanimous at the Harvey threshold: every daily increment remains far short of $|t| = 3.0$ under both conventions, and Family 2 is the largest, at Clark-West $t = +1.69$ under the lag-21 convention with the small-sample factor (one-sided $p = 0.0452$), $t = +1.71$ at lag 21 without the factor (one-sided $p = 0.0435$), and $t = +1.57$ at the canonical lag 35 (one-sided $p = 0.0580$). Its distance from the 5% line therefore depends on the bandwidth and, for the first reading, on the small-sample factor, and no reading survives a Bonferroni correction across the seven families. The lag-21 column of Table 4 is computed with a separate implementation of the Newey-West estimator; a second implementation of the same estimator, at the same lag and with the same factor, differs from it by less than 0.003 in t for every family, so the choice of implementation does not move any reading across a threshold. Under the lag-21 convention it is followed by Family 3 ($t = +0.97$, one-sided $p = 0.166$), Family 1 ($t = +0.69$), Family 8 ($t = +0.11$), Family 4 ($t = -0.22$), Family 10 ($t = -0.43$), and Family 11 ($t = -1.50$)—

and none approaches the Harvey threshold of $|t| > 3.0$ under either bandwidth. Family 10 is evaluated at the day- t opening-auction origin declared for it in Section 3, not at the close of day $t - 1$; moving its benchmark back to $VIX_{\text{close}, t-1}$ or across to $VIX_{\text{open}, t}$ moves the statistic to -0.65 and -0.16 , further from the threshold rather than toward it. Family 3's increment is likewise insensitive to how its VIX-momentum component is windowed ($+0.96$ under the five-day window implemented in the family's original production script, against $+0.97$ under the 22-day window described in Section 3) and to the Newey-West bandwidth ($+0.97$, $+0.88$, $+0.86$, and $+0.82$ at lags 21, 35, 42, and 63, each multiplied by the Harvey-Leybourne-Newbold factor). The more powerful nested test thus confirms VIX sufficiency for the daily families exactly as it does for the weekly alt-data block, and the Harvey-Leybourne-Newbold small-sample correction (factor 0.99 at the 22-day horizon) leaves each verdict unchanged. The fixed split is the one used by the daily rows of Table 3, which are recomputed on it; the Clark-West statistics here and in that table therefore describe the same forecasts, and the Diebold-Mariano statistics of the table, which agree with them in non-significance for every family, are diagnostics of the same comparison. Only Family 9 (Google Trends fear) remains outside this exercise, and not because its data are unobtainable: Google Trends returns a re-sampled relative index at each query, so a series retrieved today is not the series from which the family's main-table statistic was computed, and pinning it would fix an artifact of the retrieval rather than the original input. The Diebold-Mariano statistic from its own run ($|t| = 0.67$) is diagnostic only in the sense of Table 3, so the table does not provide an inferential reading for this family, and none is claimed. Neither Family 3 nor Family 10 uses CBOE put-call volume or an intraday VIX opening print. Family 3 is the composite sentiment index of Section 3, whose four components are the SKEW index, the VIX/VIX3M ratio, VIX momentum, and the VIX level, and it contains no put-call series at all; Family 10 uses the daily VIX opening quote, which the pinned snapshot now carries.

Table 4: Daily Nested Clark-West Increments: Two Bandwidth Conventions

#	Signal Family	CW t , lag 21 $\times$ HLN	CW t , lag 35	One-sided p , lag 35
1	Cross-asset vol momentum	+0.69	+0.68	0.249
2	VIX term structure	+1.69	+1.57	0.058
3	Behavioral sentiment	+0.97	+0.89	0.188
4	Variance risk premium	−0.22	−0.20	0.578
8	Yield curve slope	+0.11	+0.10	0.461
10	Overnight VIX	−0.43	−0.41	0.660
11	Calendar anomaly	−1.50	−1.39	0.918

Notes: Fixed split, $n_{\text{OOS}} = 1,861$ from 2019-01-02 to 2026-05-28; in-sample estimation through 2018. The lag-21 column is the Clark-West statistic with Newey-West lag 21 multiplied by the [Harvey et al. \(1997\)](#) (HLN) factor for $h = 22$, the convention of Section 5.2. The lag-35 columns use the canonical bandwidth for $h = 22$ at this sample size without the small-sample factor, as in Table 3; the p -value is one-sided against H_0 : the signal does not improve on VIX. All statistics fall below the Harvey (2016) approximate threshold of 3.0 under both conventions.

A power check bounds how far the Harvey threshold can be read as “no incremental content”. Feeding each family’s forecast regression its own forward-looking target as the signal (a deliberate look-ahead) yields Clark-West statistics of only 1.95 to 1.99 at the canonical bandwidth, and adding noise to that target in steps that lower the VIX-relative R_{OOS}^2 from 1 to 0.037–0.038 moves the statistic only within 1.94 to 2.01. For these two signal classes, full look-ahead and its noise-degraded versions, the Clark-West statistic therefore does not reach $|t| = 3.0$ even when the signal contains the answer, so a statistic below 3.0 cannot by itself separate them from a signal with no content. The finding is limited to the classes that were tested: both inherit the target’s February 2020 tail, and the 22 largest forecast origins carry 0.517 to 0.528 of the VIX-only squared error. The power of the test against other signal types, for instance signals whose response to the VIX error is bounded, was not measured. The conclusion that no signal beats VIX in this window therefore rests on the sign and size of the point estimates and on the Clark-

West p -values reported above, and the Harvey threshold is used here as a conservative convention, not as a test with demonstrated power.

Third, there is a stark disconnect between in-sample and out-of-sample performance. Google Trends fear (Family 9) achieves a partial correlation of 0.271 and an in-sample ΔR^2 of 0.038 with a highly significant t -statistic of 7.92, yet delivers essentially zero out-of-sample improvement (DM $|t| = 0.67$, raw $p = 0.504$). This is consistent with overfitting—the fear signal captures contemporaneous co-movement with VIX but has no genuine predictive lead.

Second, the variance risk premium (Family 4) illustrates the distinction between return prediction and volatility prediction. [Bollerslev et al. \(2009\)](#) showed that VRP predicts equity *returns* ($t = 3.51$ in our in-sample test, consistent with their finding). But VRP does not predict *volatility* beyond VIX—unsurprisingly, since VRP is defined as VIX minus realized volatility, making it mechanically correlated with VIX.

Third, the overnight VIX change (Family 10) has a negative VIX-relative out-of-sample R^2 (-0.006) and a negative Clark-West statistic ($t = -0.41$), so it offers no forecasting gain over VIX. The overnight signal also generates high-frequency weight changes that are too expensive to exploit after transaction costs.

7.2 Strategy Comparison: No Signal-Based Strategy Beats Benchmarks

Table 5 compares the trading strategies derived from each signal family against the 12/VIX and 50/50 SPY/GLD benchmarks over the common evaluation period.

Table 5: Volatility-Timing Strategies: Signal Families vs. Benchmarks

#	Strategy	Sharpe	Δ Sharpe	MDD (%)	Turnover	DM $ t $
<i>Benchmarks</i>						
—	12/VIX (SPY/GLD)	0.870	—	−32.2	8.9×	—
—	Buy-and-hold 50/50 SPY/GLD	0.827	−0.043	−32.5	0.0	—
<i>Signal-based strategies</i>						
1	Cross-asset composite	0.810	−0.060	−34.1	11.2×	0.93
2	VIX term structure (contango boost)	0.880	+0.010	−31.8	9.4×	0.31
3	Behavioral sentiment (fear hedge)	0.900	+0.030	−30.2	10.1×	0.72
4	VRP timing	0.782	−0.088	−30.8	12.3×	0.65
4	VRP percentile	0.800	−0.070	−30.5	11.8×	0.58
6	ERC (2-asset)	0.795	−0.054	−13.3	4.2×	1.02
8	Yield curve + VIX combo	0.746	−0.124	−35.1	9.8×	1.15
9	Google Trends fear	0.834	−0.036	−33.7	10.5×	0.48
10	Overnight VIX timing	0.851	−0.019	−31.9	14.7×	0.37
11	Calendar-only	0.658	−0.212	−48.4	2.0×	1.89

Notes: Sharpe ratios computed over the longest common sample for each pair. Δ Sharpe is relative to 12/VIX. MDD is maximum drawdown. Turnover is annualized two-way. DM $|t|$ tests whether the signal-based strategy significantly outperforms 12/VIX. All strategies use 5 bps per leg transaction costs and `signal.shift(1)` lag. Harvey threshold $|t| > 3.0$ required; none passes.

Under the canonical period-aligned benchmark (2008–2026 common sample), the 12/VIX strategy (Sharpe = 0.870) modestly outperforms the static 50/50 SPY/GLD allocation (Sharpe = 0.827, Δ = −0.043). The margin is small and not statistically significant; the substantive conclusion is that no *signal-augmented* strategy improves on the plain 12/VIX rule, while VT continues to deliver its drawdown-protection benefit relative to buy-and-hold.

Among signal-based strategies, the behavioral sentiment (fear hedge) variant comes

closest to beating 12/VIX, with the largest Sharpe improvement (+0.030, Sharpe 0.900), but this is far from statistical significance (DM $|t| = 0.72$). The VIX term structure “contango boost” variant is the only other strategy with a positive improvement (+0.010, DM $|t| = 0.31$); all remaining strategies underperform the benchmark.

7.3 Multi-Asset Portfolio Optimization

Table 6 presents the multi-asset portfolio optimization results from our systematic evaluation of four methods across seven asset subsets.

Table 6: Multi-Asset Portfolio Optimization: 50/50 SPY/GLD vs. Alternatives

Method	N_{assets}	Sharpe	MDD (%)	Beats 50/50?
50/50 SPY/GLD (baseline)	2	1.849	−13.3	—
Equal Weight	3	1.492	−11.2	No
Equal Weight	9	1.285	−8.3	No
Min Variance	3	1.412	−10.8	No
Min Variance	9	1.308	−5.7	No
Max Diversification	3	1.385	−10.5	No
Risk Parity	3	1.401	−10.9	No
Inverse Volatility	2	1.795	−13.3	No
ERC	2	1.795	−13.3	No

Notes: Common evaluation period 2023-01-04 to 2026-03-27 (811 trading days). All methods use monthly hold-and-drift rebalancing, 5 bps per leg TX costs, `signal.shift(1)`. 9-asset universe: SPY, GLD, TLT, QQQ, EEM, VWO, BND, DBC, IEFA. More assets reduce MDD but also reduce Sharpe. The 2-asset ERC and Inverse Volatility converge to approximately 50/50 because SPY vol (19.3%) $\approx$ GLD vol (18.3%).

The result is unambiguous: no multi-asset strategy beats the simple 50/50 SPY/GLD allocation on Sharpe ratio. Adding assets reduces maximum drawdown (9-asset minimum

variance achieves MDD of only -5.7% versus -13.3% for 50/50) but at the cost of lower returns, resulting in lower Sharpe ratios. This is consistent with [DeMiguel et al. \(2009\)](#), who show that estimation error in optimized portfolios offsets diversification benefits.

7.4 Era Stability of VIX Sufficiency

Table 7 presents the VIX forecasting power and strategy performance across five non-overlapping eras.

Table 7: VIX Sufficiency Across Five Market Eras (1993–2026)

Era	Period	N	VIX–RV R^2	β	t -stat	Mean VIX	Mean RV
1. Dot-Com	1993–2000	1,811	0.525	0.810	44.7	18.5	14.6
2. Post-Dot-Com	2000–2007	1,820	0.645	0.868	57.4	19.3	15.7
3. GFC	2007–2012	1,261	0.508	0.957	36.1	26.3	22.4
4. Low-Vol QE	2012–2020	1,908	0.244	0.692	24.8	15.0	11.8
5. COVID/Inflation	2020–2026	1,525	0.309	0.810	26.1	21.0	17.2
Full Sample	1993–2026	8,325	0.514	0.879	93.8	19.5	15.8
Cross-era statistics:			Mean $R^2 = 0.446$		CV = 0.33 (time-invariant)		

Notes: R^2 from regressing 22-day forward realized volatility on VIX level. β is the slope coefficient. All t -statistics are highly significant ($p \approx 0$). The lowest R^2 (0.244) occurs during the QE era when VIX was compressed near 15 with low variation; the highest (0.645) occurs during the post-dot-com era with wider VIX range. CV = coefficient of variation of R^2 across eras; CV < 0.50 indicates time-invariance.

The VIX–RV relationship is remarkably stable. The coefficient of variation of R^2 across eras is 0.33—well below the 0.50 threshold we adopt for time-invariance. Importantly, the low R^2 in the QE era (0.244) does not indicate VIX failure; rather, it reflects the mechanical compression of both VIX and RV during this period (VIX stuck near 15, RV near 12), which reduces regression R^2 without impairing VIX’s relative forecasting accuracy.

7.5 Competing Signals by Era

Table 8 presents the *in-sample* incremental R^2 of three representative competing signals across the five eras, obtained by adding each signal to a VIX-only regression of 22-day forward realized volatility within each era window. Unlike the out-of-sample horse race that constitutes our main results (Section 7), these are within-sample fits and therefore represent an upper bound on any signal’s usable content.

Table 8: Incremental R^2 of Competing Signals Across Eras

Signal	Era 1	Era 2	Era 3	Era 4	Era 5	Harvey Pass?
	Dot-Com	Post-DC	GFC	QE	COVID	($ t > 3.0$)
Overnight VIX (abs)	0.0002	0.0001	0.0039	0.0006	0.0032	1/5
VRP proxy	0.0005	0.0002	0.0160	0.0008	0.0005	1/5
Vol momentum 20/60	0.0001	0.0004	0.0216	0.0001	0.0372	2/5

Notes: In-sample incremental R^2 from adding each signal to a VIX-only regression of 22-day forward realized volatility, estimated within each era window. **Bold** cells cross the [Harvey et al. \(2016\)](#) $|t| > 3.0$ threshold. In the three calm eras (Dot-Com, Post-Dot-Com, QE) no signal is significant. Significance concentrates in the two crisis eras: all three signals add significant in-sample explanatory power during the GFC (2007–2012), and volatility momentum remains significant during COVID/Inflation (2020–2026), where it adds 3.7% incremental R^2 ($t = 9.30$).

The era decomposition qualifies rather than overturns VIX sufficiency. In the three calm regimes the competing signals are negligible, consistent with VIX absorbing all usable volatility information. In the two crisis regimes—the GFC and the COVID/Inflation era—several signals do add statistically significant *in-sample* explanatory power, volatility momentum most strongly (up to 3.7% incremental R^2). This crisis-era in-sample content does not, however, translate into out-of-sample forecasting gains: cross-asset volatility momentum, the variance risk premium, and overnight VIX changes are all among the thirteen families evaluated in the main horse race (Section 7), and none produces a statistically significant out-of-sample improvement over VIX at the Harvey threshold. The

pattern—in-sample crisis-era significance that vanishes out-of-sample—is the signature of overfitting to episodic volatility clustering rather than of exploitable ex-ante information. VIX sufficiency is therefore best characterized as a robust *out-of-sample* property that holds across regimes, not as an absence of any in-sample association in any era.

7.6 Criterion-Dependent Model Rankings

An important robustness finding emerges when we evaluate forecast models under the [Patton \(2011\)](#) framework. Table 9 compares model rankings across three evaluation criteria—QLIKE on r_t^2 , Spearman rank correlation, and sub-period stability—for six volatility model families estimated on the full 2008–2026 out-of-sample period ($N = 4,589$).

Table 9: Criterion-Dependent Model Rankings (2008–2026, $N = 4,589$)

	GJR	AMEM	MEM	GARCH	HAR	EWMA
<i>Panel A: QLIKE on r_t^2 (proxy-robust)</i>						
QLIKE	1.527	1.559	1.576	1.576	1.649	1.624
Rank	1	2	3	4	6	5
<i>Panel B: Spearman rank correlation with r_t^2</i>						
ρ_S	0.418	0.398	0.376	0.373	0.362	0.356
Rank	1	2	3	4	5	6
<i>Panel C: MCS membership ($\alpha = 0.10$)</i>						
In MCS?	✓	✗	✗	✗	✗	✗
<i>Panel D: Pairwise DM tests (QLIKE, HAC bandwidth = 22)</i>						
GJR vs. AMEM	DM = 3.78, $p < 0.001$; GJR $\succ$ AMEM (Harvey et al. (2016) pass)					
GJR vs. GARCH	DM = 4.76, $p < 0.001$; GJR $\succ$ GARCH (Harvey et al. (2016) pass)					
AMEM vs. GARCH	DM = 2.85, $p = 0.004$; AMEM $\succ$ GARCH (marginal)					

Notes: All models estimated with expanding window (minimum 500 observations, refit every 63 days). QLIKE on r_t^2 is proxy-robust per Patton (2011). MCS uses block bootstrap ($B = 5,000$, block length = 22) per Hansen et al. (2011). GJR-GARCH dominates across all criteria, with the MCS containing only GJR at the 10% level. Note: under $|r_t|$ targets (not shown), AMEM outperforms GJR on QLIKE—criterion choice reverses the ranking for non-robust loss functions, confirming the importance of the Patton framework.

Three findings emerge. First, GJR-GARCH is the dominant model under the proxy-robust QLIKE criterion, with DM = 3.78 ($p < 0.001$) versus the next-best AMEM—passing the Harvey et al. (2016) threshold. The MCS at $\alpha = 0.10$ contains only GJR, indicating statistical separation from all competitors.

Second, the rankings are *criterion-dependent*: under $|r_t|$ targets (which are *not* proxy-robust), AMEM outperforms GJR on QLIKE, reversing the ranking. This reversal occurs because $|r_t|$ introduces a 19.6% downward bias in the volatility proxy (empirical ratio

$\mathbb{E}[r_t^2]/\mathbb{E}[|r_t|]^2 = 2.44$ versus the normal-theory $\pi/2 = 1.57$), which differentially affects models with asymmetric specifications. The Patton framework resolves this ambiguity by privileging r_t^2 -based evaluation.

Third—and most relevant for VIX sufficiency—this criterion dependence *strengthens* rather than weakens our main finding. Even among sophisticated volatility models, the “best” model depends on which evaluation criterion is used. Yet none of them, regardless of criterion, improves upon VIX as an input signal. The instability of model rankings across criteria makes the stability of VIX sufficiency all the more striking.

7.7 12/VIX Strategy by Era

Table 10 shows the 12/VIX strategy performance across eras.

Table 10: 12/VIX Strategy vs. 50/50 Baseline Across Eras

Era	VT Sharpe	BH Sharpe	Δ Sharpe	VT MDD	BH MDD	Winner
1. Dot-Com ^a	1.236	1.261	−0.026	−9.2%	−9.9%	50/50
2. Post-Dot-Com ^a	0.128	0.179	−0.051	−31.3%	−26.3%	50/50
3. GFC	0.537	0.669	−0.132	−32.3%	−32.5%	50/50
4. Low-Vol QE	1.116	0.797	+0.319	−12.8%	−13.1%	12/VIX
5. COVID/Inflation	0.963	1.139	−0.176	−20.2%	−20.3%	50/50
12/VIX wins:			1 of 5 eras			

Notes: ^aGLD not available; complement asset is cash (risk-free rate). All strategies use signal.shift(1) and 5 bps per leg TX costs. 12/VIX wins only in the low-volatility QE era when persistent low VIX enables consistent high equity exposure (mean weight $\approx 80\%$).

The 12/VIX strategy wins in only 1 of 5 eras (the low-volatility QE period). This further reinforces that VT strategies do not generate alpha in the traditional sense; their value lies in drawdown insurance, which is most apparent during high-volatility regimes where VT reduces exposure.

7.8 Cross-Asset Native-IV Sufficiency: A Seven-Asset PIT Panorama

The VIX-sufficiency claim gains its strongest cross-asset confirmation from a seven-asset Publication-In-Time (PIT) panorama. Extending our per-asset experimental series (SPY; GLD, TLT, BTC-USD; QQQ, USO; and EEM) to a unified evaluation, we test four alt-data specifications—**base** (AR(1) + native IV), **epu** (+ EPU uncertainty indices), **finstress** (+ NFCI / ANFCI / STLFSI financial-stress series), and **all** (all alt-data simultaneously)—against a native-IV baseline using PIT-aligned DM-HLN tests (**pit_shift0** primary specification, 2023-01-01 to 2026-04-10, $N_{\text{OOS}} = 170$ weeks).

PIT alignment ensures each weekly regression uses only alt-data releases available at the time of prediction. For FRED weekly series (NFCI, ANFCI, STLFSI), the publication lag is Wednesday of the following week; for EPU indices, the next business day. All specifications include an AR(1) term and the asset’s native implied-volatility proxy as baseline regressors, with the native-IV proxy matched to each asset class: $\hat{\text{VIX}}$ for SPY, QQQ, and EEM;² GVZ for GLD; MOVE for TLT; BVIN for BTC-USD; and OVX for USO.

Table 11 reports the DM-HLN t -statistics for the **pit_shift0** variant across all 28 specification cells (7 assets $\times$ 4 specifications). A positive t indicates the alt-data challenger beats the native-IV baseline on QLIKE; a negative t indicates the baseline wins. The Harvey (2016) threshold of $|t| > 3.0$ is required to declare a statistically significant improvement.

²For EEM, the CBOE Emerging Markets VIX ($\hat{\text{VXEEM}}$) is the natural native-IV proxy, but it was delisted prior to our evaluation period (**yfinance** returns HTTP 404 for $\hat{\text{VXEEM}}$ and reports “possibly delisted” for VXEEM ; $\hat{\text{VXFXI}}$ and $\hat{\text{CIV}}$ are also unavailable). Accordingly, we use $\hat{\text{VIX}}$ as a systemic spillover proxy (empirical weekly EEM–VIX correlation ≈ 0.75 in this sample) and additionally confirms the verdict against a 30-day EEM realized-volatility (**rv30**) robustness baseline. Both configurations yield a clean NULL, ruling out sensitivity to the IV-proxy choice.

Table 11: Seven-Asset Cross-Asset Panorama:
DM-HLN t -Statistics vs. Native-IV Baseline
(PIT-Aligned, `pit_shift0`)

Asset	base	epu	finstress	all
SPY	−3.021	−2.603	−3.001	−2.537
GLD	−2.103	−2.069	−3.341	−2.246
TLT	+1.433	−2.477	+3.743 [†]	−5.666
BTC-USD	−5.494	−3.550	+1.370	+0.203
QQQ	−2.186	−1.967	−2.439	−1.967
USO	−3.049	−5.596	−2.584	−3.735
EEM	−2.596	−3.539	−1.434	−0.999

Notes: Positive t indicates the challenger specification beats the native-IV baseline on QLIKE; negative t indicates the baseline wins. Harvey (2016) threshold: $|t| > 3.0$. Negative values exceeding -3.0 in absolute value denote statistically significant *baseline wins*. [†]TLT $\times$ **finstress** ($t = +3.743$) is lag-sensitive: it collapses to $t = +2.000$ under `pit_shift1` (one additional lag week) and the QLIKE improvement is only +0.50% (below the 5% materiality gate).³ Weekly OOS period 2023-01-01 to 2026-04-10 ($N_{\text{OOS}} = 170$ weeks). EEM native-IV proxy: $\hat{\text{VIX}}$ (see footnote on $\hat{\text{VXEEM}}$ delisting); `rv30` robustness yields the same qualitative verdict.

Across all 28 panorama cells, only one—TLT $\times$ **finstress** ($t = +3.743$)—crosses the Harvey threshold on the primary specification, and this sole exceedance fails every robustness check: it collapses under one additional lag week and is QLIKE-insignificant. The remaining 27 cells are either null (positive but below +3.0) or baseline-wins (neg-

ative). The cross-asset verdict is thus **UNIVERSAL_NULL_7/7**: for six of seven assets the null is unambiguous; for TLT the single outlier is a regime artefact rather than a structural signal. Native-IV sufficiency extends from US equity (SPY) to international equity (EEM, QQQ), commodities (GLD, USO), fixed income (TLT), and cryptocurrency (BTC-USD), constituting the broadest cross-asset evidence for option-market information efficiency available at daily frequency.

7.9 Channel-Specific Mechanism Heterogeneity

The aggregate null across thirteen signal families conceals a structured heterogeneity in *how* different model architectures fail. We characterize three channels with qualitatively distinct failure patterns, providing sharper guidance for future research beyond “all signals fail.”

7.9.1 Channel 1: HAR+VIX on Equity—Regime-Invariant Gains on Parkinson Volatility

The HAR-RV+VIX model (M4: Corsi 2009 HAR on log-Parkinson realized volatility with $\log(\text{VIX}_{t-1}^2)$ regressor) delivers statistically significant out-of-sample improvements on large-cap and mid-cap equity indices relative to a GJR-GARCH(1,1) Normal baseline evaluated on the same Parkinson target. We define regime-invariance as passing the Harvey threshold ($|t| > 3.0$, BH-adjusted) in all three VIX quintile regimes (low/mid/high).

Across 18 equity asset $\times$ regime cells ($3 \text{ assets} \times 3 \text{ regimes} \times 2 \text{ models}$), HAR+VIX achieves regime-invariant status for two equity indices: QQQ passes across all three regimes (low $t = 6.79$, mid $t = 4.83$, high $t = 3.09$; all BH-adjusted $p < 0.001$), and IWM passes across all three regimes (low $t = 4.22$, mid $t = 3.61$, high $t = 2.73$; all BH-adjusted $p < 0.05$). SPY achieves 2 of 3 regime passes (low $t = 7.99$, mid $t = 4.39$, but high $t = 2.26$ fails BH correction). Across the full 54-cell evaluation ($6 \text{ assets} \times 3 \text{ regimes} \times 3 \text{ models}$), 17 cells pass BH-adjusted Harvey threshold.

These gains are specific to the Parkinson-range volatility measurement framework.⁴

⁴A fair-test analysis establishes that GARCH-MIDAS-X, a score-driven long-memory exten-

The HAR+VIX gains reflect the empirically documented efficiency of Parkinson-range estimators for capturing intraday volatility dynamics that daily closing-price GARCH imperfectly approximates—not an incremental signal beyond VIX in the traditional regression sense.

7.9.2 Channel 2: GARCH-MIDAS Long-Run Factor—Universal Null

GARCH-MIDAS-X, which extends the baseline GJR-GARCH by adding a slow-moving multiplicative component driven by lagged monthly VIX-squared ($\tau_t = \exp(m + \theta \times \text{VIX}_{\text{monthly}}^2)$), achieves zero BH-adjusted conditional passes across all three VIX regimes and all six assets evaluated (`n_midas_conditional_pass` = 0). This null is universal: neither equity nor commodity assets benefit from the MIDAS long-run factor once target-consistent evaluation is applied.

The MIDAS null strengthens the option-market information aggregation argument: any low-frequency volatility information that a one-month lagged VIX factor could supply is already impounded in the contemporaneous VIX level, leaving no incremental forecasting content for the long-run MIDAS component.

7.9.3 Channel 3: GAS-t Score-Driven Models—Asset-Class Bifurcation

The Generalized Autoregressive Score (GAS-t) framework (Creal et al., 2013) exhibits qualitatively different behavior across asset classes.

Equity: architectural incompatibility. On SPY ($N_{\text{OOS}} = 1,323$, 2021–2026), GAS-t is actively harmful relative to GJR-GARCH Normal (DM-HLN $t = -3.27$, $p < 0.001$). We test four candidate remedies—Hansen skew-t asymmetry (M2), score win-size at ± 0.3 (M3), persistence capping at $\beta \leq 0.9$ (M4), and regime switching to HAR in low-VIX environments (M5)—and finds that *all variants produce QLIKE above the GJR-N baseline*: M1 symmetric GAS-t QLIKE = 1.527 versus GJR-N = 1.474; Hansen skew-t (M2) QLIKE = 1.535 (worse than M1). The post-2020 equity return dis-

sion, achieves zero conditional passes (0/3 regimes) under the same matched-target protocol (`fair_test_1_m3_midas_pass_count` = 0), ruling out the possibility that any long-run factor model outperforms.

tribution (excess kurtosis = 8.85, skewness = +0.30) contradicts the GAS framework’s core premise that score-driven updates stabilize conditional variance when innovations are approximately Student- t . Instead, the GAS score exhibits positive skewness (+1.58), causing right-biased updates that systematically overweight large positive return shocks.

Commodity: tail-risk instrument, not volatility forecaster. On negatively skewed commodities (USO, UNG), the skew- t GAS extension (Hansen, 1994) recovers VaR/ES performance that the symmetric GAS- t variant lacks (H_3 : ES backtests pass at the 1% level for all four assets, BH-adjusted). However, forecasting QLIKE remains null (0 of 4 assets improve, H_1 pass count = 0). This bifurcation—ES improvement without QLIKE improvement—indicates that skew- t GAS captures distributional shape for risk management without improving point volatility forecasts.

The Channel 3 pattern thus identifies a *usage boundary*: for equity volatility forecasting, GAS- t is architecturally incompatible and should be replaced by GJR-GARCH. For commodity tail-risk estimation specifically, skew- t GAS provides a validated tail-risk instrument despite null forecasting performance. These distinctions cannot be inferred from the aggregate null result alone.

Table 12: Channel-Specific Outcome Matrix

Channel	Equity (SPY/QQQ/IWM)	Commodity (USO/GLD)	TLT
1. HAR+VIX (Parkinson target)	PASS (QQQ/IWM 3/3 regimes)	PASS (TLT/USO)	PASS (3/3)
2. GARCH-MIDAS long-run	NULL (0/3 regimes all assets)	NULL	NULL
3a. GAS- t (QLIKE)	HARMFUL (SPY $t = -3.27$)	NULL	—
3b. Skew- t GAS (ES 1%)	—	PASS (4/4 assets)	—

Notes: Channel 1 PASS requires BH-adjusted Harvey $|t| > 3.0$ in all three VIX regimes. Channel 2 GARCH-MIDAS-X fair test (matched target). Channel 3a tests QLIKE improvement vs GJR-GARCH. Channel 3b tests ES backtest at 1% coverage. Regime-invariant = PASS in all low/mid/high VIX regimes.

7.10 Publication-Delay Robustness

Section 4.6 introduced the publication-delay convention and adopted the “corrected” timing—`shift(1)` for the daily EPU series and `shift(2)` for the weekly financial-stress series—as the ex-ante specification for Families 12–13 in all main results. Because that decision rests on release calendars rather than on data, it is reasonable to ask whether the headline null would survive under the conventional uniform `shift(1)` that prior horse races in this literature have implicitly used, and whether a still-more-paranoid uniform `shift(2)` that also delays the daily EPU series would change the conclusions. A companion experiment addresses both questions by re-estimating the SPY weekly forecasting battery—and its cross-asset replication on GLD, TLT, and BTC—under three timing variants and comparing the resulting Diebold-Mariano statistics against the original specification.

The design holds the sample, regressors, and estimation window fixed and varies only the lag applied to each Family 12–13 input. The conventional-timing variant uses `shift(1)` for all five FRED series and reproduces the prior literature exactly; the `corrected` variant retains `shift(1)` for the daily EPU series (USEPU, WLEMU) but applies `shift(2)` to the three weekly financial-stress series (NFCI, ANFCI, STLFSI4), matching their five-to-six calendar-day release lag; and the `conservative` variant applies `shift(2)` uniformly to all five series. Every variant is fitted on the 2018–2022 in-sample window and evaluated on the 2023–2026 expanding-window out-of-sample block, yielding $n_{\text{OOS}} = 169$ common weekly observations after `shift(2)` alignment (170 under `shift(1)`). DM-HLN statistics (Diebold and Mariano, 1995) with $h = 1$ are computed against the M2 native-implied-volatility baseline (VIX for SPY, GVZ for GLD, MOVE for TLT, BTC’s own AR(1) for BTC).

Across the sixteen asset-by-specification cells the headline null is essentially invariant to the timing choice: thirteen of sixteen DM statistics change by less than 0.5 in absolute value between the conventional and corrected variants, and no cell that fails the Harvey threshold of $|t| > 3.0$ under one variant passes it under another. The three cells with material movement all weaken the case for alt-data rather than support it. For SPY, the

financial-stress specification (M4) moves from $t = -3.00$ under `shift(1)` to $t = -3.61$ under `shift(2)` against the VIX baseline, pushing further past the Harvey cutoff in the direction of the null; the same cell on BTC moves from $t = -1.28$ to $t = -3.64$ in the kitchen-sink M5 specification, gaining significance against alt-data. For TLT, the M4 financial-stress cell drops from $t = +3.74$ to $t = +1.96$, falling below the Harvey threshold; this is the single previously-apparent cross-asset niche in our broader compendium, and Section 4.6 reports its collapse as the central motivation for adopting `shift(2)` as the primary convention.

Table 13 summarizes the per-cell movement for the twelve baseline-versus-alt-data comparisons on the four assets. Two structural patterns deserve note. First, the SPY headline null—that no alt-data family beats the VIX baseline at the Harvey threshold—holds under all three timing conventions and strengthens under the corrected one, with the corrected M4 statistic (-3.61) and the conservative M5 statistic (-3.34) both pushing past $|t| > 3.0$ against the alt-data challenger. Second, the kitchen-sink M5 specification that pools EPU and financial-stress regressors is itself fragile to the timing convention: TLT’s M5 statistic ranges from -5.18 under `shift(1)` to -8.57 under uniform `shift(2)`, while BTC’s M5 moves from -1.28 to $+1.71$ across the same range, with the sign flip reflecting collinearity among lagged FRED inputs rather than any genuine forecasting content. We accordingly treat M5 as a sensitivity benchmark rather than as primary evidence for or against alt-data.

Table 13: Publication-Delay Robustness of DM-HLN Statistics

Asset	Specification	shift(1)	corrected	conservative	Flag
SPY	M3 EPU	−2.55	−2.54	−2.46	—
SPY	M4 FinStress	−3.00	−3.61	−3.61	strengthened
SPY	M5 All	−1.01	−1.00	−3.34	strengthened
GLD	M3 EPU	−1.77	−1.76	−1.40	—
GLD	M4 FinStress	−3.34	−3.00	−3.00	—
GLD	M5 All	−0.13	+0.41	−3.22	sensitive
TLT	M3 EPU	−0.83	−0.80	+1.76	—
TLT	M4 FinStress	+3.74	+1.96	+1.96	threshold flip
TLT	M5 All	−5.18	−7.43	−8.57	strengthened
BTC	M3 EPU	−5.04	−5.13	−5.42	—
BTC	M4 FinStress	+1.37	+1.27	+1.27	—
BTC	M5 All	−1.28	−3.64	+1.71	sensitive

Notes: DM-HLN t -statistics from the re-verification battery. Positive values indicate the alt-data challenger predicts more accurately than the native-implied-volatility baseline; negative values indicate the baseline predicts more accurately. Bolded entries are cells whose Harvey-threshold verdict ($|t| > 3.0$) changes between variants. The **shift(1)** column reproduces the convention used in prior horse races; **corrected** applies **shift(2)** only to the three weekly financial-stress series (NFCI, ANFCI, STLFSI4); **conservative** applies **shift(2)** uniformly to all five FRED inputs. Sample: weekly W-FRI, IS 2018-01 to 2022-12 ($n_{\text{IS}} = 260$), OOS 2023-01 to 2026-04 ($n_{\text{OOS}} = 170$ under **shift(1)**, 169 under **shift(2)**). The verdict records three threshold-relevant flips out of sixteen cells; the TLT M4 flip is the single niche claim that collapses and the SPY M4 strengthening is the single null claim that consolidates.

Two implications for the rest of the paper follow. The first is that the main results

presented in Sections 7–7.9 are not driven by a discretionary timing choice: even under the convention used in the prior literature, no Family 12–13 specification reaches the Harvey threshold against the native implied-volatility baseline on the equity, commodity, or cryptocurrency assets. The single bond-market exception that earlier work would have reported as a positive cross-asset finding (TLT M4 with `shift(1)`) falls below the threshold once the publication calendar is respected, and the results file records no cell where the corrected variant produces a Harvey-significant alt-data win that the conventional variant missed. Second, the asymmetry of the movement is itself informative: when alt-data series are denied a half-week of look-ahead, four cells move toward the null and only one cell flips sign in an inconsequential direction (GLD M5 at $-0.13 \rightarrow +0.41$, well inside the indeterminate range). A reviewer who suspects that our adoption of the corrected timing is over-conservative can read the `shift(1)` column directly and verify that the universal-sufficiency claim survives even there; the cautious null claim made in the main text is the one supported by all three timing conventions simultaneously.

7.11 Alt-Data Fails a Second Paradigm: Regime-Based Allocation

The forecasting evidence assembled so far rejects alt-data as a source of incremental information about future realized volatility, but it does not exhaust the channels through which an EPU or financial-stress signal could in principle add value. A standard defence offered for these series in the practitioner literature is that even when they fail an out-of-sample QLIKE horse race they may still carry useful information for *allocation*—that is, for choosing equity-versus-safe-asset weights in real time—because regime shifts in uncertainty can affect the optimal mix even when they do not predict tomorrow’s variance. A companion experiment tests this allocation defence directly with daily-frequency regime-rule portfolios on the SPY+GLD baseline, the 50/50 moat that the broader research program has repeatedly verified as a hard benchmark to beat (DeMiguel et al., 2009).

The experiment fits six rebalanced portfolios over the 2019-01-15 to 2026-04-10 daily

sample ($n = 1,817$ trading days, $n_{\text{IS}} = 1,001$ days in 2019–2022 and $n_{\text{OOS}} = 816$ days in 2023–2026 after a 252-day rolling-rank warm-up and a five-day release lag). Strategy S1 is the static 50/50 SPY/GLD moat. Strategy S2 is a conventional volatility-targeted SPY weight $w_t^{\text{SPY}} = \text{clip}(0.15/\hat{\sigma}_{t-1}^{\text{SPY}, 20\text{d}}, 0.2, 1.0)$ that uses no alt-data. Strategies S3–S5 are single-signal regime rules—S3 gates on lagged VIX, S4 on the 252-day percentile rank of EPU, and S5 on the 252-day percentile rank of NFCI—each shifting the SPY weight between 0.7 and 0.3 depending on whether the signal sits below or above its 70th percentile. Strategy S6 is the equally-weighted average of the three signal-gated weight processes. All inputs are lagged to respect the publication-delay calendar established in Section 4.6: VIX at `shift(1)`, EPU at `shift(2)`, NFCI at `shift(5)`. Inference uses the Politis–Romano stationary bootstrap (Politis and Romano, 1994) with 1,000 replications, mean block length 20, and seed 42, applied to the Sharpe-ratio difference following Opdyke (2007).

Table 14 reports the full-sample performance of the six strategies. The static 50/50 baseline produces a Sharpe of 1.309 over the seven-year window; the best alt-data strategy (S5 NFCI) edges it by Sharpe +0.003, an economically nil margin. The volatility-targeted benchmark (S2) underperforms the moat by 0.237 in Sharpe despite tilting the average SPY weight from 0.50 up to 0.88, illustrating that mechanical scaling of equity exposure to a 20-day volatility estimate is itself dominated by the static mix. The two single-signal regime rules that gate on macroeconomic alt-data (S4 EPU, S5 NFCI) and the hybrid average (S6) all cluster within a Sharpe range of 0.030 around the 50/50 baseline, with no individual strategy improving Sortino or annualized return enough to suggest a systematic alpha channel. The lone dimension on which an alt-data strategy stands out is drawdown: NFCI’s -17.9% maximum drawdown ties the VIX-regime strategy for the lowest in the table and produces the highest Calmar ratio (0.950 vs. 0.888 for the baseline)—a risk-management feature that we return to below.

Table 14: Full-Sample Performance of Six SPY+GLD Allocation Rules

#	Strategy	Sharpe	Sharpe (OOS)	MDD	Calmar	Sortino	avg w^{SPY}
S1	50/50 static	1.309	1.975	−0.203	0.888	1.64	0.500
S2	Vol-targeted	1.072	1.474	−0.220	0.720	1.42	0.880
S3	VIX regime	1.210	1.748	− 0.179	0.883	1.59	0.543
S4	EPU regime	1.283	1.856	−0.232	0.752	1.71	0.566
S5	NFCI regime	1.312	1.656	− 0.179	0.950	1.73	0.594
S6	Hybrid average (S3+S4+S5)	1.305	1.806	−0.184	0.911	1.73	0.568

Notes: Daily-rebalanced SPY+GLD portfolios over 2019-01-15 to 2026-04-10 ($n = 1,817$ days; OOS window 2023-01 to 2026-04, $n_{\text{OOS}} = 816$ days). Sharpe and Sortino are annualized at $\sqrt{252}$; MDD is the trough-to-peak loss on the equity curve; Calmar is annualized return divided by $|\text{MDD}|$. S2–S6 inputs are lagged per the Section 4.6 publication-delay convention. Bold entries are dimension-best within column; the static moat (S1) and the NFCI rule (S5) tie at the top of the Sharpe column.

The headline ranking by itself does not distinguish whether the alt-data strategies achieve a genuine improvement that is masked by sampling noise or whether the differences are themselves noise. Table 15 resolves this by reporting the Politis–Romano stationary-bootstrap distribution of the Sharpe-ratio difference for each alternative against the 50/50 baseline, alongside the parallel comparison against the vol-targeted benchmark. All five alternatives are statistically indistinguishable from the static moat: the smallest two-sided p -value is 0.264 (S2 vol-target), the comparison closest to a tie is S5 NFCI with $p = 0.966$, and the 95% bootstrap intervals for the four alt-data strategies all cover zero and stretch from roughly -0.24 to $+0.25$ Sharpe units. The vol-targeted comparison tells the same story: even the largest single-signal margin against S2 (S5 NFCI, observed $+0.241$) carries $p = 0.166$ and a 95% interval $[-0.122, +0.583]$ that contains both zero and economically meaningful improvements, so the data cannot distinguish a true edge from a sampling artefact. None of the alt-data strategies passes the conventional $p < 0.05$ threshold against either benchmark, none passes the Harvey–Liu–Zhu (Harvey

et al., 2016) stricter $|t| > 3.0$ threshold even nominally, and the hybrid average (S6) does not improve upon the best single component—failing the diversification hypothesis that an equally-weighted blend of imperfect signals should outperform any single one.

Table 15: Stationary-Bootstrap Sharpe-Difference Inference

Pair	Obs. diff	p -value	95% CI	Verdict
<i>Against S1 (50/50 static moat)</i>				
S2 – S1 (vol-target)	−0.238	0.264	[−0.670, +0.174]	NS, loss
S3 – S1 (VIX)	−0.099	0.388	[−0.344, +0.132]	NS
S4 – S1 (EPU)	−0.026	0.860	[−0.310, +0.257]	NS
S5 – S1 (NFCI)	+0.003	0.966	[−0.231, +0.254]	NS, tie
S6 – S1 (hybrid)	−0.005	0.948	[−0.191, +0.172]	NS, tie
<i>Against S2 (vol-targeted)</i>				
S3 – S2 (VIX)	+0.138	0.512	[−0.279, +0.531]	NS
S4 – S2 (EPU)	+0.212	0.280	[−0.211, +0.609]	NS
S5 – S2 (NFCI)	+0.241	0.166	[−0.122, +0.583]	NS (largest)
S6 – S2 (hybrid)	+0.233	0.194	[−0.131, +0.568]	NS

Notes: Politis–Romano stationary bootstrap (Politis and Romano, 1994) with $B = 1,000$ replications, mean block length 20, and seed 42; inference on the Sharpe-ratio difference follows Opdyke (2007). Two-sided p -values and 95% percentile intervals reported. “NS” = not significant at the 5% level; “tie” marks observed differences inside one standard error of zero. None of the nine pairs reaches the Harvey–Liu–Zhu (Harvey et al., 2016) $|t| > 3.0$ threshold even nominally.

A natural follow-up question is whether the alt-data strategies fail because they sit on the equity exposure for the wrong reason in calm periods but still earn their keep by cutting risk during stress. Table 16 examines the three economically distinct stress windows in the sample: the COVID drawdown of February–April 2020 (52 trading days), the rate-shock window running from January through October 2022 (209 days), and the

SVB regional-banking episode of March–April 2023 (32 days). NFCI (S5) does behave as a stress signal in the textbook sense: it cuts the average SPY weight to 0.33 during COVID, to 0.30 during the rate shock, and modestly to 0.51 during the brief SVB episode, with the COVID and rate-shock reductions large enough to deliver materially lower drawdowns than the static moat in both windows. EPU (S4) behaves as a stress signal only in the COVID window: in the 2022 rate shock—which was driven by Federal Reserve policy rather than newspaper-coded uncertainty—it *raises* the SPY weight to 0.56 and incurs a slightly larger drawdown than the baseline, and in the SVB window it stays close to the 50/50 mix. The VIX-regime rule (S3) responds to COVID and 2022 but not to SVB, where realized implied volatility was muted in the days when bank-equity prices were falling.

Table 16: Stress-Period Average SPY Weight by Strategy

Episode	Period	S1	S2	S3 VIX	S4 EPU	S5 NFCI	S6 Hybrid
COVID 2020	2020-02-15 to 2020-04-30	0.50	0.38	0.34	0.37	0.33	0.35
Rate-shock 2022	2022-01-01 to 2022-10-31	0.50	0.67	0.34	0.56	0.30	0.40
SVB 2023	2023-03-01 to 2023-04-15	0.50	0.90	0.51	0.53	0.51	0.52

Notes: Within-episode mean of the daily target SPY weight by strategy; full-period reference is 0.50 for S1 by construction. NFCI (S5) is the only single-signal rule that reduces equity exposure in all three episodes; EPU (S4) raises exposure in the 2022 rate shock; the vol-targeted benchmark (S2) overweights equity during SVB because realized volatility had compressed in the preceding weeks.

The stress-episode pattern clarifies why the NFCI rule earns a tied-best Sharpe and a best-in-table Calmar without earning a statistically distinguishable Sharpe edge. NFCI behaves as a drawdown insurance overlay: it shifts equity exposure down during the three windows in which the static moat lost most of its peak-to-trough value, and the resulting MDD reduction is what compresses the denominator of the Calmar ratio. But the same overlay loses opportunity cost in the much longer set of normal trading days that surround the stress windows, where it spends roughly 40% of the sample with $w^{\text{SPY}} = 0.30$ rather

than the moat's 0.50. The two effects net out to zero in Sharpe space and produce the observed $p = 0.966$ tie against the 50/50 baseline. This is the same drawdown-insurance phenomenon documented in Section 8.3 for volatility-targeted overlays: a risk-management tool that reduces the size of the worst losses but does not generate alpha, and whose value to an investor depends on the curvature of her utility function rather than on her belief that the signal predicts returns.

This evidence therefore extends the universal-sufficiency claim from the forecasting paradigm to the allocation paradigm. Across nine bootstrapped Sharpe comparisons spanning two benchmarks and three single-signal alt-data series plus a hybrid, none of the alt-data strategies produces a statistically distinguishable improvement over either the 50/50 SPY+GLD moat or the vol-targeted scaler that uses no alt-data at all. The one cross-section in which an alt-data signal stands out is drawdown reduction, where NFCI ties the VIX-regime rule for the lowest MDD; the rest of the surface is statistically flat. Combined with the earlier forecasting nulls and the publication-delay robustness check, this allocation null closes the second of the two natural channels through which the academic literature has historically argued that EPU and financial-stress series should add value. Section 8.2 accordingly removes regime-based allocation from the catalogue of plausible mechanisms that could break VIX sufficiency.

8 Why the Null is Informative

8.1 The Information Aggregation Explanation

Our null result is not a failure to find signal but a positive finding about market efficiency. The VIX option market is one of the deepest and most liquid derivatives markets in the world, with average daily trading volume exceeding 500,000 contracts. This liquidity ensures rapid price discovery, meaning that any signal with predictive content for short-term volatility is quickly impounded into option prices and hence into VIX.

We can formalize this argument. Let Ω_t denote the information set of the represen-

tative option trader at time t . Under semi-strong form efficiency of the options market:

$$\mathbb{E}[\sigma_{t+1:t+22}^2 | \Omega_t] = f(\text{VIX}_t), \quad (14)$$

where $f(\cdot)$ is monotonically increasing. Any signal $s_t \in \Omega_t$ satisfies:

$$\mathbb{E}[\sigma_{t+1:t+22}^2 | \text{VIX}_t, s_t] = \mathbb{E}[\sigma_{t+1:t+22}^2 | \text{VIX}_t], \quad (15)$$

i.e., VIX is a sufficient statistic for all information already incorporated in option prices. Our thirteen signal families are all elements of Ω_t , which explains the zero incremental forecasting power.

8.2 What Might Break VIX Sufficiency?

The sufficiency argument identifies two conditions under which alternative signals could add value:

1. **Higher-frequency information:** Intraday 5-minute realized volatility captures microstructure dynamics (e.g., opening auctions, lunch-hour liquidity, closing imbalances) that closing-price VIX does not fully reflect. Our preliminary HAR-RV results show a 41.8% QLIKE improvement, suggesting this frontier is real.
2. **Genuinely exogenous shocks:** Events that arrive between option-market pricing updates—geopolitical shocks, pandemic declarations, sudden policy announcements—may create temporary information asymmetries. However, these are by definition unpredictable ex ante.

Two channels that prior literature has flagged as candidates for breaking VIX sufficiency are now explicitly ruled out by our evidence. The first is **alt-data macro-financial uncertainty proxies** (EPU, NFCI, financial-conditions indices, sentiment z-scores): one might conjecture that newspaper- or text-based uncertainty signals capture macro stress information that the VIX, anchored to S&P 500 options, fails to internalize.

Section 7.6 and our forecasting horse race show that none of these proxies improves volatility forecasts beyond VIX, with no test reaching the Harvey threshold in the beneficial direction. The forecasting channel is closed. The second is **regime-based volatility-timing allocation** (alt-data signals routed through portfolio weights rather than mean forecasts): one might further conjecture that even an uninformative *forecast* signal could still add value if used as a *regime indicator* for tilting allocations between risk-on and risk-off assets. Section 7.11 refutes this conjecture: across nine alt-data Sharpe comparisons against the 50/50 SPY/GLD baseline and a VIX-regime benchmark, no signal achieves statistical separation under a stationary-bootstrap test ($p \geq 0.10$ in every comparison), and the best alt-data signal (NFCI, Sharpe 1.312) is statistically indistinguishable from the static 50/50 benchmark (Sharpe 1.309, $p = 0.966$). The allocation channel is closed alongside the forecasting channel. The two surviving routes—higher-frequency realized volatility and genuinely exogenous shocks—do not involve alt-data text or macro proxies, and therefore do not constitute a counterexample to VIX sufficiency in the daily, closing-price information set.

8.3 Economic Significance: Forecast Accuracy vs. Risk Management

A natural question is whether the forecast ranking differences documented in Section 7.6 translate into economically meaningful differences for risk management. We address this through a comprehensive VaR and Expected Shortfall (ES) backtesting exercise using five model families over 4,591 out-of-sample days (2008–2026), evaluated with the Kupiec (1995) unconditional coverage test, the Christoffersen (1998) conditional coverage (independence) test, and the Basel III traffic-light framework.

Table 17 presents the VaR/ES backtest results and the resulting economic ranking.

Table 17: VaR/ES Backtest: Model Risk Management Quality (2008–2026)

Model	VaR _{1%}	Kupiec	Christoff.	Basel	VaR _{5%}	Kupiec	Risk
	Viol. Rate	p	p	Zone	Viol. Rate	p	Score
AMEM(Gamma)	1.09%	0.550	0.576	Green	5.27%	0.403	1.94
GJR-GARCH(t)	1.44%	0.005	0.017	Green	5.84%	0.011	1.56
Hist. Sim(250d)	1.59%	<0.001	<0.001	Green	5.25%	0.442	1.34
EWMA(Gauss)	2.33%	<0.001	0.016	Yellow	5.90%	0.006	1.23
HAR-ABS(Gauss)	2.50%	<0.001	0.010	Yellow	6.36%	<0.001	1.19

Notes: Risk Score is a composite of Kupiec calibration, Christoffersen independence, Basel zone, and Acerbi-Szekely ES backtest, computed at both 1% and 5% levels (maximum score = 2.0). AMEM achieves near-perfect calibration at both levels (1.09% vs. 1% target; 5.27% vs. 5% target) and is the only model passing both Kupiec and Christoffersen at $\alpha = 0.01$. GJR-GARCH, despite being the best *forecaster* (Section 7.6), is only the second-best *risk manager*, and its two diagnostics disagree: it sits in the Green zone at both levels—the zone is a violation-count threshold—while the Kupiec unconditional-coverage test rejects at both levels ($p = 0.005$ and $p = 0.011$). We therefore report the zone and the test side by side throughout and do not treat the Green zone on its own as evidence of calibration. The GJR-GARCH innovations are a unit-variance Student- t : the degrees-of-freedom parameter is estimated with the unit-variance scale inside the likelihood (Jacobian included) and the VaR/ES quantiles are scaled by $\sqrt{(\nu - 2)/\nu}$ accordingly; across the 73 quarterly refits ν has median 6.53 (range 5.03–7.65) with no fit at a boundary or failing to converge. All models use VIX as primary input; expanding window estimation with quarterly refits.

The key finding is a *decoupling* between forecast accuracy and risk management quality. GJR-GARCH is the best volatility forecaster (lowest QLIKE, sole MCS member), but AMEM is the best risk manager (highest composite VaR/ES score of 1.94 vs. 1.56). This decoupling arises because risk management requires accurate *tail* calibration—getting the 1% and 5% quantiles right—which depends on the distributional assumption (AMEM’s Gamma distribution better captures the right-skewed nature of r_t^2) rather than on mean

forecast accuracy alone.

For economic significance, we translate VaR/ES forecasts into portfolio allocations. An ES-weighted SPY/GLD strategy using AMEM achieves Sharpe = 0.873 versus the static 50/50 benchmark’s 0.826, an improvement of +0.047. While modest, this confirms that volatility models add value through risk control even when they cannot add value through return prediction—reinforcing the drawdown insurance framework.

Crucially, all five models in this exercise use VIX as their primary input signal, and none benefits from adding external signals. The debate over “which model is best” is entirely *within* the VIX information set, further strengthening VIX sufficiency: the relevant optimization is over model specification (GJR vs. AMEM) and application domain (forecasting vs. risk management), not over information sources.

8.4 The Drawdown Insurance Framework

If VT cannot generate alpha, what is its economic function? We propose that VT is best understood as drawdown insurance, analogous to buying a put option on the portfolio.

Proposition 1 (Drawdown Insurance Cost-Benefit). *For a CRRA investor with risk aversion γ , the certainty-equivalent return of a VT portfolio exceeds that of buy-and-hold when:*

$$\gamma \geq \gamma^* = \frac{\Delta\mu}{\frac{1}{2}(\sigma_{BH}^2 - \sigma_{VT}^2)}, \quad (16)$$

where $\Delta\mu = \mu_{BH} - \mu_{VT}$ is the return drag from VT, and $\sigma_{BH}^2 - \sigma_{VT}^2$ is the variance reduction.

Using our empirical estimates (Table 18), the 12/VIX strategy has a break-even risk aversion of $\gamma^* = 4.5$, and the EWMA VT strategy has $\gamma^* = 4.4$.

Table 18: VT as Drawdown Insurance: Cost-Benefit Analysis

VT Method	Drag	MDD Red.	Cost/1pp	γ^*	Welfare ($\gamma=5$)
12/VIX (5-asset mean)	3.49%/yr	−12.3 pp	0.28%	4.5	+0.01%/yr
EWMA VT (5-asset mean)	2.12%/yr	−9.6 pp	0.22%	4.4	+0.80%/yr

Notes: Return drag is the annualized difference in mean return between buy-and-hold and VT strategy. MDD reduction is the decrease in maximum drawdown. Cost per 1 pp MDD is the return drag divided by MDD improvement (in percentage points), i.e. the ratio of the two preceding columns; note that this is a ratio of cross-asset means and not the mean of the five per-asset ratios, which is the larger 0.32% and 0.31% and reverses neither the ordering nor the conclusion. Break-even γ is the minimum CRRA risk aversion at which the certainty-equivalent return of VT exceeds the buy-and-hold 50/50 benchmark. Welfare gain at $\gamma = 5$ is the annualized certainty-equivalent improvement over that same 50/50 benchmark, aggregated as the mean of the five per-asset differences. The two columns are therefore on one scale, and their ordering is a consistency check the table passes: EWMA VT crosses at the lower γ^* (4.4 versus 4.5) and accordingly shows the larger gain at $\gamma = 5$. Both welfare entries are small and rest on per-asset differences that change sign across assets, so we read the column as evidence that VT is approximately welfare-neutral for an investor at $\gamma = 5$ —just past the break-even point—and not as a point estimate of the size of the gain. Every entry in this table is a cross-asset aggregate over the same 5 assets (SPY, GLD, QQQ, EEM, 0050.TW), 2007–2026—the drag and the cost ratio are means and the break-even γ is the median—not a statistic of a two-asset SPY/GLD portfolio.

The insurance framework has a clear practical implication: VT is rational for investors who are moderately to highly risk-averse ($\gamma \geq 4.5$, which encompasses most retail investors and many institutional mandates) but not for risk-neutral or mildly risk-averse investors who should prefer buy-and-hold.

8.5 The Simplicity Premium

A natural question is whether more complex strategies justify their complexity. Our meta-analysis of 14 live-tracked strategies reveals a *simplicity premium*: the Spearman correlation between the number of estimated parameters and the Sharpe ratio is $\rho = 0.077$ ($p = 0.794$), indicating zero relationship between complexity and performance. The best-performing strategies (50/50 SPY/GLD, 12/VIX, Piecewise VT) have zero or one estimated parameter.

This echoes [DeMiguel et al.’s \(2009\)](#) finding that $1/N$ portfolios outperform optimized alternatives. In the VT context, the simplicity premium arises because:

1. Simple rules have fewer parameters to estimate, reducing overfitting.
2. Smooth weight functions (like 12/VIX) produce low turnover, saving transaction costs.
3. Zero-parameter rules are immune to estimation window and specification choices.

9 Conclusion

We set out to answer a straightforward question: can any observable signal beat VIX for equity volatility forecasting and volatility timing? After a systematic evaluation of thirteen pre-specified signal families across 33 years of data, five market eras, and a unified pipeline with strict lag enforcement, transaction-cost accounting, publication-delay correction for weekly macroeconomic series, Harvey (2016) approximate thresholds, the Clark-West test for nested comparisons, and proxy-robust evaluation per [Patton \(2011\)](#), the answer is no. That answer rests on the sign and size of the point estimates and on the Clark-West p -values, not on the Harvey threshold alone, because the threshold cannot be shown to have power against look-ahead signals and their noise-degraded versions in this window (see the power check after Table 4).

This null is not a failure but an informative finding. It reflects the efficiency of the options market, which aggregates heterogeneous information—cross-asset correlations,

macroeconomic indicators, sentiment, term structure dynamics—into a single summary statistic. Each of the thirteen signal families we tested represents one channel of information that option traders already incorporate into their pricing. VIX, as the price-weighted consensus, already contains them.

Moreover, the robustness of VIX sufficiency is underscored by two additional findings. First, volatility model rankings are criterion-dependent: GJR-GARCH dominates under proxy-robust QLIKE on r_t^2 , but AMEM dominates under $|r_t|$ targets. This instability in model rankings makes the stability of VIX sufficiency across all criteria all the more striking. Second, forecast accuracy and risk management quality are distinct dimensions—AMEM is the best risk manager (VaR/ES composite score 1.94) while GJR is the best forecaster—yet both operate entirely within the VIX information set, confirming that the relevant optimization is over model specification and application domain, not over information sources.

The practical implications are clear:

1. **For researchers:** The daily-frequency, closing-price frontier is exhausted. Gains require moving to higher frequencies (5-minute realized measures show preliminary promise with 41.8% QLIKE improvement) or genuinely exogenous information outside the option-market information set.
2. **For practitioners:** The simplest VIX-based rule (12/VIX) cannot be improved by adding signals. Complexity reduces returns through estimation error, higher turnover, and transaction costs. The optimal allocation for most investors is either buy-and-hold 50/50 SPY/GLD (if $\gamma < 4.5$) or 12/VIX VT (if $\gamma \geq 4.5$).
3. **For regulators and risk managers:** VIX sufficiency is time-invariant (CV = 0.33 across 33 years). Models based on VIX do not need frequent recalibration, and regulatory frameworks (e.g., Basel III/IV) that reference VIX for market risk are on solid empirical ground.

We conclude with a broader observation. In financial econometrics, null results are underreported relative to their information content. Our systematic horse race, with its

pre-specified signal families and stringent statistical controls, provides stronger evidence than any individual positive finding could. The 42 null results accumulated across this research program (the original 36 plus 6 from the alt-data forecasting and allocation extensions) are not 42 failures; they are 42 independent confirmations that the option market efficiently aggregates volatility-relevant information.

References

- Andersen, T. G., Bollerslev, T., Diebold, F. X., & Labys, P. (2003). Modeling and forecasting realized volatility. *Econometrica*, 71(2), 579–625.
- Baker, M., & Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. *Journal of Finance*, 61(4), 1645–1680.
- Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring economic policy uncertainty. *Quarterly Journal of Economics*, 131(4), 1593–1636.
- Barunik, J., Kocenda, E., & Vacha, L. (2016). Asymmetric connectedness on the U.S. stock market: Bad and good volatility spillovers. *Journal of Financial Markets*, 27, 55–78.
- Bekaert, G., & Hoerova, M. (2014). The VIX, the variance premium and stock market volatility. *Journal of Econometrics*, 183(2), 181–190.
- Bollerslev, T., Tauchen, G., & Zhou, H. (2009). Expected stock returns and variance risk premia. *Review of Financial Studies*, 22(11), 4463–4492.
- Bollerslev, T., Li, J., & Zhao, Y. (2020). Good volatility, bad volatility, and the cross section of stock returns. *Journal of Financial and Quantitative Analysis*, 55(3), 751–781.
- Bouri, E., Molnar, P., Azzi, G., Roubaud, D., & Hagfors, L. I. (2017). On the hedge and safe haven properties of Bitcoin: Is it really more than a diversifier? *Finance Research Letters*, 20, 192–198.
- Bouman, S., & Jacobsen, B. (2002). The Halloween indicator, “sell in May and go away”: Another puzzle. *American Economic Review*, 92(5), 1618–1635.
- Brave, S. A., & Butters, R. A. (2011). Monitoring financial stability: A financial conditions index approach. *Federal Reserve Bank of Chicago Economic Perspectives*, 35(1), 22–43.

- Bozovic, M. (2024). VIX-managed portfolios. *International Review of Financial Analysis*, 95, 103353.
- Britten-Jones, M., & Neuberger, A. (2000). Option prices, implied price processes, and stochastic volatility. *Journal of Finance*, 55(2), 839–866.
- Campbell, J. Y., & 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.
- Bucci, A. (2020). Realized volatility forecasting with neural networks. *Journal of Financial Econometrics*, 18(3), 502–531.
- Christensen, B. J., & Prabhala, N. R. (1998). The relation between implied and realized volatility. *Journal of Financial Economics*, 50(2), 125–150.
- Clark, T. E., & West, K. D. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics*, 138(1), 291–311.
- Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2), 174–196.
- Creal, D., Koopman, S. J., & Lucas, A. (2013). Generalized autoregressive score models with applications. *Journal of Applied Econometrics*, 28(5), 777–795.
- Da, Z., Engelberg, J., & Gao, P. (2011). In search of attention. *Journal of Finance*, 66(5), 1461–1499.
- DeMiguel, V., Garlappi, L., & Uppal, R. (2009). Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy? *Review of Financial Studies*, 22(5), 1915–1953.
- Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3), 253–263.

- Diebold, F. X., & Yilmaz, K. (2012). Better to give than to receive: Predictive directional measurement of volatility spillovers. *International Journal of Forecasting*, 28(1), 57–66.
- Engle, R. F., & 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.
- Estrella, A., & Mishkin, F. S. (1998). Predicting U.S. recessions: Financial variables as leading indicators. *Review of Economics and Statistics*, 80(1), 45–61.
- Fleming, J., Kirby, C., & Ostdiek, B. (2001). The economic value of volatility timing. *Journal of Finance*, 56(1), 329–352.
- Hansen, B. E. (1994). Autoregressive conditional density estimation. *International Economic Review*, 35(3), 705–730.
- Hansen, P. R., & Lunde, A. (2005). A forecast comparison of volatility models: Does anything beat a GARCH(1,1)? *Journal of Applied Econometrics*, 20(7), 873–889.
- Hansen, P. R., Lunde, A., & Nason, J. M. (2011). The model confidence set. *Econometrica*, 79(2), 453–497.
- Harvey, D., Leybourne, S., & Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, 13(2), 281–291.
- Harvey, C. R., Liu, Y., & Zhu, H. (2016). ...and the cross-section of expected returns. *Review of Financial Studies*, 29(1), 5–68.
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2), 65–70.
- Christoffersen, P. F. (1998). Evaluating interval forecasts. *International Economic Review*, 39(4), 841–862.
- Engle, R. F., & Gallo, G. M. (2006). A multiple indicators model for volatility using intra-daily data. *Journal of Econometrics*, 131(1–2), 3–27.

- Jiang, G. J., & Tian, Y. S. (2005). The model-free implied volatility and its information content. *Review of Financial Studies*, 18(4), 1305–1342.
- Kliesen, K. L., & Smith, D. C. (2010). Measuring financial market stress. *Federal Reserve Bank of St. Louis Economic Synopses*, 2010-2, 1–2.
- Kupiec, P. H. (1995). Techniques for verifying the accuracy of risk measurement models. *Journal of Derivatives*, 3(2), 73–84.
- Lou, D., Polk, C., & Skouras, S. (2019). A tug of war: Overnight versus intraday expected returns. *Journal of Financial Economics*, 134(1), 192–213.
- Huang, Z., Tong, C., & Wang, T. (2019). VIX term structure and VIX futures pricing with realized volatility. *Journal of Futures Markets*, 39(1), 72–93.
- Maillard, S., Roncalli, T., & Teiletche, J. (2010). The properties of equally weighted risk contribution portfolios. *Journal of Portfolio Management*, 36(4), 60–70.
- Moreira, A., & Muir, T. (2017). Volatility-managed portfolios. *Journal of Finance*, 72(4), 1611–1644.
- Newey, W. K., & West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3), 703–708.
- Opdyke, J. D. (2007). Comparing Sharpe ratios: So where are the p -values? *Journal of Asset Management*, 8(5), 308–336.
- Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, 160(1), 246–256.
- Patton, A. J., & Sheppard, K. (2015). Good volatility, bad volatility: Signed jumps and the persistence of volatility. *Review of Economics and Statistics*, 97(3), 683–697.
- Politis, D. N., & Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical Association*, 89(428), 1303–1313.

- Poon, S.-H., & Granger, C. W. J. (2003). Forecasting volatility in financial markets: A review. *Journal of Economic Literature*, 41(2), 478–539.
- Preis, T., Moat, H. S., & Stanley, H. E. (2013). Quantifying trading behavior in financial markets using Google Trends. *Scientific Reports*, 3, 1684.
- Schwert, G. W. (1989). Why does stock market volatility change over time? *Journal of Finance*, 44(5), 1115–1153.
- Whaley, R. E. (2000). The investor fear gauge. *Journal of Portfolio Management*, 26(3), 12–17.