Cornell Engineering

Cornell Financial Engineering Manhattan

ORIE 5254 Report - Fall 2025

Predicting VIX Spikes: Evidence from Macro Uncertainty, GARCH, HAR, and Option-Implied Predictors

Sirui Zhao,
Kexin Deng,
Yichen Gao,
Yihan Zhao

supervised by
Edward Tom

Preface

Volatility forecasting lies at the center of modern risk management, derivatives pricing, and macro–financial decision-making. During the past decade, episodes such as the 2018 volatility crash, the COVID-19 shock, and the inflation-driven turbulence of 2022 have highlighted the growing need for models that can anticipate extreme movements in market-implied uncertainty. Motivated by this environment—as well as by my academic training in quantitative finance through the Cornell Financial Engineering Manhattan program—this project examines the ability of several widely used econometric and financial-market-based frameworks to forecast spikes in the CBOE Volatility Index (VIX). The purpose of this report is to provide a clear and systematic comparison of time-series models, macro-uncertainty indicators, and derivatives-based measures, with a specific focus on their ability to capture extreme volatility events. The target audience includes practitioners in risk management, macro-research analysts, and students interested in volatility modeling. My goal is to deliver a practical, data-driven assessment of what truly predicts VIX spikes, and where traditional models succeed or fail, especially in out-of-sample settings. We hope that this work serves as a useful reference for readers seeking an accessible yet rigorous overview of volatility-forecasting methods and their empirical performance.

Acknowledgements

We would like to express our sincere gratitude to the Cornell Financial Engineering Manhattan faculty for providing the academic foundation and intellectual environment that made this project possible. We are especially thankful to Professor Edward Tom for his guidance throughout the course and for the thoughtful discussions on volatility modeling and empirical methodology.

Abstract

This project investigates the problem of forecasting the CBOE Volatility Index (VIX), with particular emphasis on predicting volatility spikes that pose significant challenges to risk management, asset allocation, and trading systems. Forecasting implied volatility is inherently difficult due to its forward-looking nature, sharp jump behavior, and sensitivity to macroeconomic and derivatives-based information that is not directly observable from returns.

We evaluate a wide spectrum of econometric models, including GARCH-type specifications, Heterogeneous Autoregressive (HAR) models, macro-augmented volatility predictors, and derivatives-based measures. Using a rolling out-of-sample framework, we assess both level accuracy and spike detection performance across the 2016–2025 period. Our results show that traditional GARCH models exhibit limited predictive power for implied volatility, while HAR and HARX models capture medium-horizon dynamics but fail to anticipate abrupt jumps. In contrast, derivatives-based predictors incorporating information such as the VVIX and term-structure signals provide the strongest out-of-sample performance.

Overall, the findings highlight a clear hierarchy of predictive content across model classes and underscore the importance of forward-looking information in VIX forecasting. The empirical evidence suggests that incorporating option-implied signals yields the most substantial improvements in both point forecasts and spike detection accuracy.

Keywords

VIX forecasting; implied volatility; GARCH models; HAR model; macroeconomic uncertainty; derivatives-based predictors; volatility spikes

Contents

1 Introduction and Motivation

Volatility is a priced quantity, a risk-management input, and an allocation signal. The VIX is the most visible proxy for equity-market uncertainty. A practical question on a volatility desk is deceptively simple: with public macro data and discipline about real-time availability, can we say something useful about where VIX is headed tomorrow, and can we identify the largest one-day jumps?

The contribution here is a transparent, macro-only forecasting experiment that adheres to real-time constraints. It separates what macro levels can do—track regimes—from what they fail to do—time single-day jumps—and it quantifies both. The paper also codifies a consistent way to align mixed-frequency macro series to a daily target while respecting publication lags, which is often glossed over in empirical work.

2 Literature Review

2.1 Macro-Based Forecasting of the VIX

Foundational work on conditional heteroskedasticity established that asset return volatility is persistent and predictable at multiple horizons. Engle (1982) introduces ARCH and Bollerslev (1986) generalizes to GARCH, providing a flexible framework for time-varying second moments. Subsequent research documents slow-moving or regime-varying components in volatility; the heterogeneous autoregressive model of realized volatility in Corsi (2009) captures long memory through multi-horizon components, while Engle & Rangel (2008) separate low-frequency volatility from short-run dynamics using a spline-GARCH structure and relate secular volatility movements to macroeconomic conditions. On the option side, the VIX is a model-free transformation of risk–neutral expected variance, linking index option prices to the market's view on future variability (Whaley, 2000). Carr & Wu (2006) show that VIX embeds both expectations of physical variance and a time-varying variance risk premium. Together, these insights suggest several channels by which macroeconomic forces can affect the VIX: through the conditional variance of fundamentals, through discount rates and leverage constraints, and through fluctuations in the price of variance risk.

A parallel macro–finance literature constructs measures of economic uncertainty and studies their real and financial effects. Jurado, Ludvigson & Ng (2015) develop an *ex ante* macro uncertainty index and document sizable impacts on future activity (AER). Bloom (2009) models uncertainty shocks as rare freezes that depress investment and hiring (Econometrica). In asset pricing, both risk-aversion and policy channels appear important: Bekaert et al. (2013) decompose risk and uncertainty and show distinct roles for each in returns (JF), while Pastor & Veronesi (2012) link government policy uncertainty to equity valuation (JF). Text-based proxies also matter: Manela & Moreira (2017) construct a news-implied volatility index (JFE), and Baker, Bloom & Davis (2016) develop the Economic Policy Uncertainty index (QJE). This body of work indicates that macro states and shocks plausibly shift volatility *regimes*, although they need not be informative about the precise timing of single-day spikes.

Bridging macro information to daily volatility often relies on mixed-frequency designs or announcement–surprise identification. Andreou, Ghysels & Kourtellos (2013) show that incorporating macro factors via MIDAS regressions improves forecasts of realized volatility (JFE). High-frequency event studies find that the timing and sign of macro announcements shape intraday returns and volatility across asset classes (e.g., Andersen et al., 2007). These papers highlight the importance of *surprise* components and release timing for short-horizon volatility dynamics.

Our contribution is methodological and empirical. Methodologically, we offer a transparent, *real-time–feasible* alignment of mixed-frequency macro series to a daily target: each indicator is lagged by its publication delay, transformed to economically interpretable growth or level metrics, and screened on a universal monthly clock before being forward-filled to daily business

dates. Empirically, we evaluate macro-only forecasting of the VIX in an expanding, time-ordered backtest using three distinct time-series architectures: a state-space model with exogenous macro (UCM), a principal–component summary with ridge penalization, and a shallow gradient-boosting tree. We assess performance on both the VIX level and an economically stringent spike criterion: a transition from a calm state (past 10-day minimum ≤ 20) to a crisis state (≥ 40) that at least doubles the recent trough. The evidence is sharp: macro levels track volatility *regimes* but carry little information about the precise days on which $20 \to 40$ spikes occur. This aligns with the literature's emphasis on announcement surprises and risk-premium variation and provides a disciplined macro-only baseline for future extensions that incorporate those elements.

2.2 Derivatives-Based Indicators for Forecasting VIX

The predictive power of derivatives markets for volatility has been established across multiple dimensions. Zhang and Zhu (2006) and Dew-Becker et al. (2017) demonstrate that VIX futures term structure contains information about future realized volatility and disaster risk, with backwardation typically preceding market stress. Park (2015) shows that VVIX (volatility-of-volatility) provides incremental predictive power beyond VIX itself, capturing second-order uncertainty that often precedes regime shifts. Cross-asset volatility spillovers have been documented by Diebold and Yilmaz (2009), while Choi and Hong (2020) find bidirectional causality between oil volatility (OVX) and VIX. Credit markets also contain forward-looking information: Bekaert et al. (2013) show that credit spread widening accompanies VIX increases during risk-off periods, and Gilchrist and Zakrajšek (2012) demonstrate that credit market stress predicts both economic downturns and equity volatility. Recent machine learning applications to volatility forecasting (Rasekhschaffe and Jones 2019; Christensen et al. 2021) show that non-linear models can effectively exploit complex relationships among predictors. This project contributes by systematically evaluating derivatives indicators in a unified machine learning framework, using a feature set deliberately different from existing literature while maintaining strong economic foundations.

2.3 ARCH and GARCH Models: A Review of Foundational Literature

Volatility in financial markets exhibits several robust empirical regularities: *volatility clustering*, whereby large return movements tend to be followed by further large movements of either sign; *heavy-tailed return distributions*; and *time-varying conditional heteroskedasticity*. These stylized facts, first highlighted by Mandelbrot (1963), motivated a departure from models assuming constant error variance and laid the foundation for conditional variance models.

A key insight underlying these models is the distinction between *unconditional* and *conditional* heteroskedasticity. While asset returns typically display little autocorrelation, the autocorrelation of squared or absolute returns is strong and persistent, implying that the conditional variance is itself predictable. This feature—central to Engle's (1982) formulation—provides an econometric explanation for volatility clustering and allows variance dynamics to be modeled separately from the conditional mean.

Another important foundational concept concerns the drivers of volatility asymmetry. Two mechanisms are typically emphasized. The *leverage effect* (Black, 1976) suggests that declines in equity values mechanically increase financial leverage, amplifying subsequent volatility. In contrast, the *volatility feedback effect* (Campbell and Hentschel, 1992) posits that an increase in expected volatility raises the required rate of return, depressing prices and further increasing volatility. Both mechanisms motivate the development of asymmetric GARCH models, which allow negative shocks to exert a disproportionately strong effect on future variance.

Finally, the persistence properties of volatility are central to model selection. Standard GARCH models imply an exponential decay of autocorrelation, whereas empirical work on equity and implied volatility—including the VIX—frequently documents *hyperbolic* decay indicative of long

memory. This observation motivates fractional-integration approaches such as FIGARCH, which can more flexibly capture persistent volatility dynamics.

The breakthrough came with Engle (1982), who introduced the Autoregressive Conditional Heteroskedasticity (ARCH) model, allowing the conditional variance to depend on past squared innovations. Building on this insight, Bollerslev (1986) proposed the Generalized ARCH (GARCH) framework, in which volatility persistence is modeled through both lagged squared shocks and lagged conditional variances:

$$\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2, \tag{1}$$

where $\varepsilon_t = \sigma_t z_t$ and $z_t \sim N(0,1)$ or t_ν. This specification captures the slow decay of autocorrelations in squared returns, an essential characteristic of both equity market volatility and implied volatility indices such as the VIX. Empirical studies such as Majmudar & Banerjee (2004) demonstrate that GARCH-type models fit the post-1993 VIX reasonably well, reproducing its pronounced short-term persistence even though returns themselves exhibit little autocorrelation.

A central limitation of symmetric GARCH models is their inability to differentiate the impact of positive and negative shocks. In both equity and implied volatility markets, negative innovations tend to amplify future volatility more strongly than positive innovations, a phenomenon attributed to the *leverage effect* or the *volatility feedback effect*. To incorporate this asymmetry, several extensions have been developed. The Exponential GARCH (EGARCH) model of Nelson (1991) expresses conditional variance in logarithmic form, removing non-negativity constraints and introducing asymmetric shock responses:

$$\ln(\sigma_t^2) = \omega + \gamma\frac{\varepsilon_{t-1}}{\sigma_{t-1}} + \alpha\left|\frac{\varepsilon_{t-1}}{\sigma_{t-1}}\right| + \beta\ln(\sigma_{t-1}^2). \tag{2}$$

Similarly, the Glosten–Jagannathan–Runkle GARCH (GJR-GARCH) model introduces an indicator for negative shocks:

$$\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \gamma\varepsilon_{t-1}^2\mathbb{I}_{\{\varepsilon_{t-1}<0\}} + \beta\sigma_{t-1}^2. \tag{3}$$

These asymmetric specifications have been shown to substantially improve the modeling of implied volatility dynamics. Empirical work—including Wang (2019); Qiao, Todorov & Tauchen (2020) and several recent VIX forecasting comparisons— consistently finds that asymmetric GARCH variants outperform the symmetric GARCH(1,1) model during periods of financial stress such as 2008 and 2020, when negative shocks drive disproportionately large spikes in the VIX.

Recent literature further extends GARCH-type models to incorporate jumps and realized high-frequency volatility. Qiao, Todorov & Tauchen (2020) augment the variance equation with a dynamic jump intensity λ_t, while Wu, Li & Tse (2023) develop a two-component Realized EGARCH (REGARCH-2C) model that combines realized measures with latent volatility, producing a closed-form implied VIX. Both approaches demonstrate superior forecasting accuracy when intraday data are available.

Another important advancement focuses on the long-memory properties of volatility. The Fractionally Integrated GARCH (FIGARCH) model of Baillie, Bollerslev & Mikkelsen (1996) relaxes the short-memory assumption through fractional differencing:

$$(1-\beta L-\phi L^d)\varepsilon_t^2 = \omega + (1-\alpha L)v_t, \qquad 0<d<1, \tag{4}$$

capturing the hyperbolic decay in volatility autocorrelation frequently observed in implied volatility indices. Studies such as Liu, Maheu & McCurdy (2015); Medvedev (2019) show that FIGARCH models replicate the persistence of VIX volatility more accurately than standard GARCH, albeit with increased parameter uncertainty.

Overall, the literature finds that symmetric GARCH models provide stable short-term forecasts, while asymmetric variants (EGARCH, GJR-GARCH, APARCH) and long-memory structures

(FIGARCH) deliver markedly better performance during volatility spikes, structural breaks, and regime transitions. These models form a natural benchmark against which macro-based, derivatives-based, and machine-learning models for VIX forecasting, examined in subsequent sections, can be evaluated. Together, these ARCH-type models provide a comprehensive framework for characterizing the dynamics of financial volatility and serve as the theoretical foundation upon which most modern volatility forecasting approaches—including those evaluated in the following sections—are constructed.

2.4 HAR-Based Statistical Model Forecasting of the VIX

The Heterogeneous Autoregressive (HAR) model, originally proposed by Corsi (2009), has become a cornerstone for modeling volatility dynamics due to its ability to capture long-memory properties in a simple linear form. In the context of implied volatility, Fernandes et al. (2014) employ both parametric and semiparametric HAR processes to examine and forecast the daily behavior of the CBOE Volatility Index (VIX). Their analysis reveals that the VIX displays strong persistence and long-range dependence, consistent with the notion of heterogeneous market participants operating across different time horizons.

Formally, the HAR model for the logarithm of volatility y_t can be expressed as:

$$y_t = \beta_0 + \beta_1 \bar{y}_{t-1}^{(1)} + \beta_2 \bar{y}_{t-1}^{(5)} + \beta_3 \bar{y}_{t-1}^{(22)} + \varepsilon_t, \tag{5}$$

where $\bar{y}_{t-1}^{(h)} = \frac{1}{h}\sum_{i=1}^{h} y_{t-i}$ denotes the average volatility over h past periods, corresponding to daily, weekly, and monthly horizons. This additive cascade structure captures the heterogeneous propagation of volatility across different time scales and provides a linear approximation to the long-memory behavior observed in realized and implied volatilities.

The HAR model is particularly well-suited for modeling the VIX because the index exhibits (i) pronounced persistence, (ii) regime-like transitions between low and high volatility periods, and (iii) a negative contemporaneous relationship with equity returns. Fernandes et al. (2014) further demonstrate that incorporating market and macro-financial variables enhances the explanatory power of the model, leading to the so-called HARX specification:

$$y_t = \beta_0 + \beta_1 \bar{y}_{t-1}^{(1)} + \beta_2 \bar{y}_{t-1}^{(5)} + \beta_3 \bar{y}_{t-1}^{(22)} + \boldsymbol{\gamma}' z_{t-1} + \varepsilon_t, \tag{6}$$

where z_t includes variables such as S&P 500 returns and volume, oil returns, USD index changes, the term spread, credit spread, and the Federal Funds rate deviation. Their empirical results indicate a strong negative effect of S&P 500 returns on the VIX, a positive relation with trading volume, and a mild negative long-run effect of the term spread, while the linear HARX model remains difficult to outperform even when nonlinear neural-network extensions are considered.

Building upon this framework, Thrasher (2017) introduce a complementary perspective by focusing on the prediction of abrupt VIX spikes. He argues that periods of exceptionally low volatility dispersion—analogous to the "calm before the storm"—tend to precede large increases in the VIX. Specifically, the 20-day standard deviation of the VIX (σ_t^{VIX}) and of the "VIX of VIX" (VVIX) are used as indicators of volatility compression. When both dispersion measures fall below their respective low-percentile thresholds (e.g., $\sigma_t^{VIX} \leq 0.86$ and $\sigma_t^{VVIX} \leq 3.16$), the probability of a future volatility spike substantially increases.

Integrating these insights into the HARX framework allows the model to capture both long-memory persistence and imminent spike risks. A spike indicator variable S_{t-1} can be constructed to reflect the joint low-dispersion condition, leading to an extended HARX specification:

$$y_t = \beta_0 + \beta_1 \bar{y}_{t-1}^{(1)} + \beta_2 \bar{y}_{t-1}^{(5)} + \beta_3 \bar{y}_{t-1}^{(22)} + \boldsymbol{\gamma}' z_{t-1} + \delta S_{t-1} + \varepsilon_t, \tag{7}$$

where δ captures the predictive contribution of pre-spike conditions. This hybrid HARX–spike

model thus combines the structural persistence of implied volatility with high-frequency indicators of volatility compression, enhancing both continuous forecasting performance and the detection of extreme VIX movements.

3 Data and Real-time Alignment

4 Models

4.1 Macroeconomic-Based Factors

4.1.1 Data, Transformations, and Alignment

Target and predictors The target is the daily closing VIX, V_t. Predictors are FRED series or open proxies: UNRATE, CPIAUCSL, INDPRO, GS10, DFF, M2SL, UMCSENT, ICSA, DCOILWTICO, and GOLDAMGBD228NLBM. The sample begins in the early 2000s and runs to the latest available month; out-of-sample (OOS) evaluation starts after one year plus lags. The predictors are intentionally macroeconomic and publicly available. Below, each variable is defined with its economic channel to VIX, expected sign, release cadence, and caveats. Signs refer to the long-run association with the *VIX level* after controlling for trend.

UNRATE (Unemployment rate, level; monthly, lag $\approx$ 20 bdays). Higher slack indicates weaker growth and potentially larger cash-flow uncertainty, raising the conditional variance of equity returns; conversely, it can also reduce risk-taking and compress leverage-induced volatility. Net effect is typically *positive* on VIX over business cycles. Caveat: slowly moving and subject to revisions; level depends on participation dynamics.

CPIAUCSL (CPI, year-over-year). Inflation alters discount-rate volatility and monetary-policy reaction risk. Elevated or accelerating inflation increases uncertainty about policy paths and real rates, pushing variance risk premia higher; expected sign *positive*. Caveat: part transitory (energy/food); what matters most for VIX are *surprises* at release.

INDPRO (Industrial production, year-over-year). A broad real-activity proxy. Falling growth co-moves with profit uncertainty and default risk, lifting volatility; expected sign *negative* for levels (more growth, lower VIX) and *positive* for deteriorations. In our models it is the dominant feature for regime tracking.

GS10 (10-year Treasury yield, level; daily). Long rates summarize growth and term-premium expectations. Two channels operate: higher real rates tighten financial conditions (raising equity risk) but often coincide with strong growth (lower risk). Empirically the association with VIX is mildly *negative* in expansions and flips in slowdowns, so we allow non-linearities (GBR) rather than impose a sign.

DFF (Effective federal funds rate, level; daily). The near-risk-free policy rate compresses or releases risk-taking capacity through funding channels. Rapid hiking cycles raise uncertainty about cash flows and policy mistakes, tilting VIX *up*. Caveat: policy levels are known daily; *changes* and guidance typically matter more than levels.

M2SL (M2 money stock, year-over-year). Monetary/liquidity conditions affect leverage, VAR constraints, and volatility risk premia. Ample liquidity historically coincides with lower risk compensation; expected sign *negative*. Caveat: post-2020 definitional breaks and QE/QT episodes mean M2 must be interpreted with care.

UMCSENT (Consumer sentiment index, level). A slow-moving proxy for perceived household income and employment risk. Improved sentiment lowers precautionary savings and tends to reduce VIX; expected sign *negative*. Caveat: level biases and media effects; announcement surprises again dominate at high frequency.

ICSA (Initial jobless claims, weekly; we use percentage change). A timely indicator of labor-market stress and recession risk. Spikes in claims raise near-term uncertainty; expected sign *positive*. Caveat: noisy around holidays; we use percent changes and a one-week publication lag.

DCOILWTICO (WTI crude oil, daily; percent change). Captures energy-supply and demand shocks. Large positive oil moves tied to supply constraints can raise macro uncertainty and inflation risk, pushing VIX *up*; demand-driven declines often align with recessions and higher VIX as well—hence asymmetric, non-linear effects that tree models can absorb.

GOLDAMGBD228NLBM (London gold fix, daily; percent change). A safe-haven proxy and inflation hedge. Persistent gold strength often co-moves with financial stress and policy uncertainty; expected association with VIX is *positive* but small after controlling for other variables. Caveat: gold reacts to global—not purely U.S.—drivers.

Why these variables? They span growth, labor slack, inflation, policy stance, liquidity, confidence, and commodity risk—macro channels emphasized by the uncertainty and volatility literatures cited above. Importantly, they are observable to all market participants and can be aligned to the daily grid with realistic publication lags. The empirical pattern in Figure **??**—heavy weight on real-activity momentum (INDPRO_yoy) and liquidity growth (M2SL_yoy), with rates and sentiment next—matches the notion that macro levels govern *regimes* of expected variance, whereas jump timing is driven by announcement surprises, microstructure, and rapidly evolving risk premia.

Transformations Rates and indices where the level is meaningful (GS10, DFF, UMCSENT) enter in levels. Trending aggregates (INDPRO, CPI, M2) enter as year-over-year growth, $\mathrm{yoy}(x_t) = 100(x_t/x_{t-12}-1)$. Daily commodities (oil, gold) enter as percent log changes, $\mathrm{pct}(x_t) = 100\,\Delta \ln x_t$. All predictors are standardized within training windows.

Publication lags and business-day grid To prevent look-ahead, each series is shifted forward by an approximate publication lag (monthly ≈ 20 bdays; weekly ≈ 5; market-based 2–3), then forward-filled within the release period. The aligned design matrix X_t lives on the business-day grid of V_t.

Universal monthly screening Because daily ΔV_t is noisy while macro is stepwise, predictors are screened at the universal monthly-last frequency. Let $V_m^{(M)}$ be last-business-day VIX for month m and $X_m^{(M)}$ the lagged macro features on the same dates. I compute Pearson correlations, select the top $K = 10$ by $|\rho|$, and feed those features to the daily models.

Table 1: Macro series, transformations, and assumed lags

Series	Transformation	Native Freq.	Lag (bdays)
UNRATE	level	monthly	20
CPIAUCSL	yoy	monthly	20
INDPRO	yoy	monthly	20
GS10	level	daily	3
DFF	level	daily	3
M2SL	yoy	monthly	20
UMCSENT	level	monthly	20
ICSA	pct (weekly diff)	weekly	5
DCOILWTICO	pct	daily	3
GOLDAMGBD228NLBM	pct	daily	3

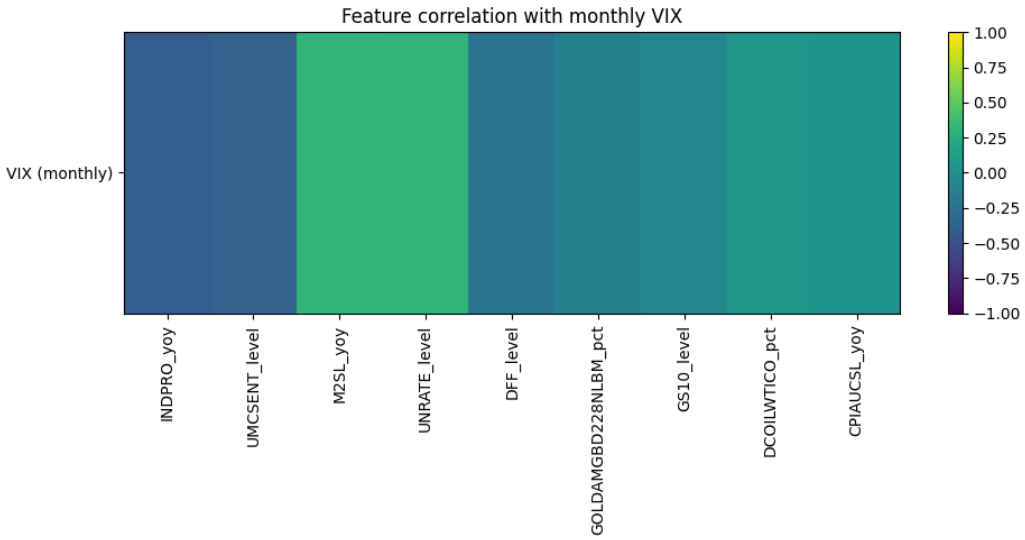

Figure 1: Monthly correlation heat map: VIX vs lagged macro features.

4.1.2 Macro–VIX associations (visual and statistical).

Figure 2 overlays monthly z-scores of each macro indicator against the VIX after enforcing publication lags and a common monthly clock; Figure 3 shows the corresponding scatters with fitted lines. Visually, real-activity and sentiment measures comove inversely with the VIX, while labor stress and oil price changes load positively. Table 2 quantifies these patterns using contemporaneous Pearson and Spearman correlations, a best-lag statistic (months) that maximizes absolute correlation,[1] univariate OLS R^2, and the minimum Granger p-value over lags 1–6. Industrial production growth (`INDPRO_yoy`) is the strongest single correlate ($\rho = -0.414$, $R^2 = 0.172$) with a best lag of -2 months; consumer sentiment (`UMCSENT_level`) is similarly negative ($\rho = -0.321$, $R^2 = 0.103$). Oil price changes (`DCOILWTICO_pct`) load positively ($\rho = 0.325$, $R^2 = 0.105$) and exhibit a nontrivial Granger signal (min $p = 0.024$), as do jobless claims growth (`ICSA_pct`, min $p = 0.001$). Inflation, gold, and the 10-year yield are weak in this macro-only, level-based specification. Overall, macro levels are informative about *regimes* but—consistent with our spike exercise—are not designed to time single-day VIX jumps.

[1]By construction, a positive value of *Best lag (m)* means the macro series leads the VIX; a negative value means the VIX leads the macro series.

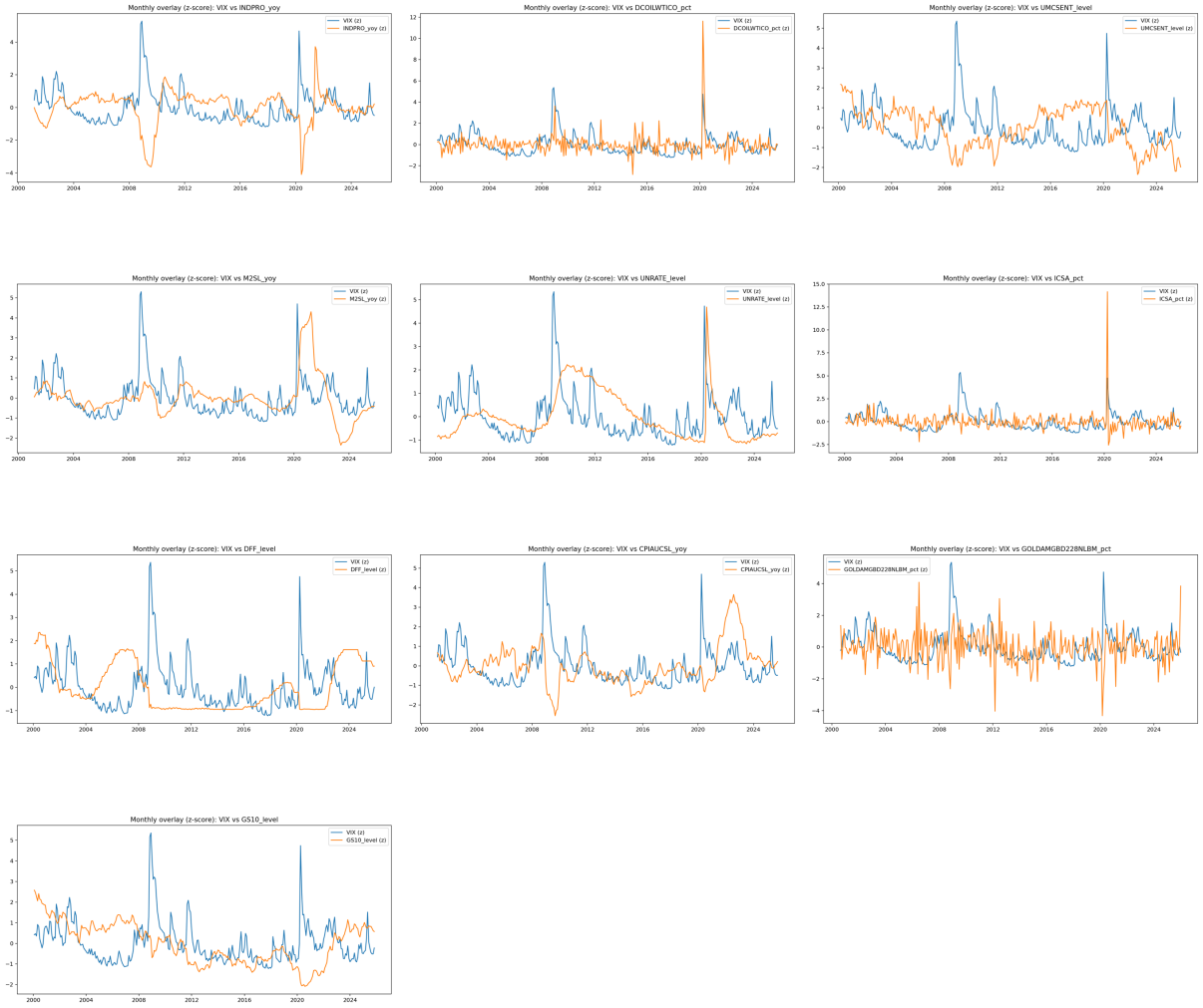

Figure 2: Monthly z-score overlays: VIX and macro series (publication-lag aligned).

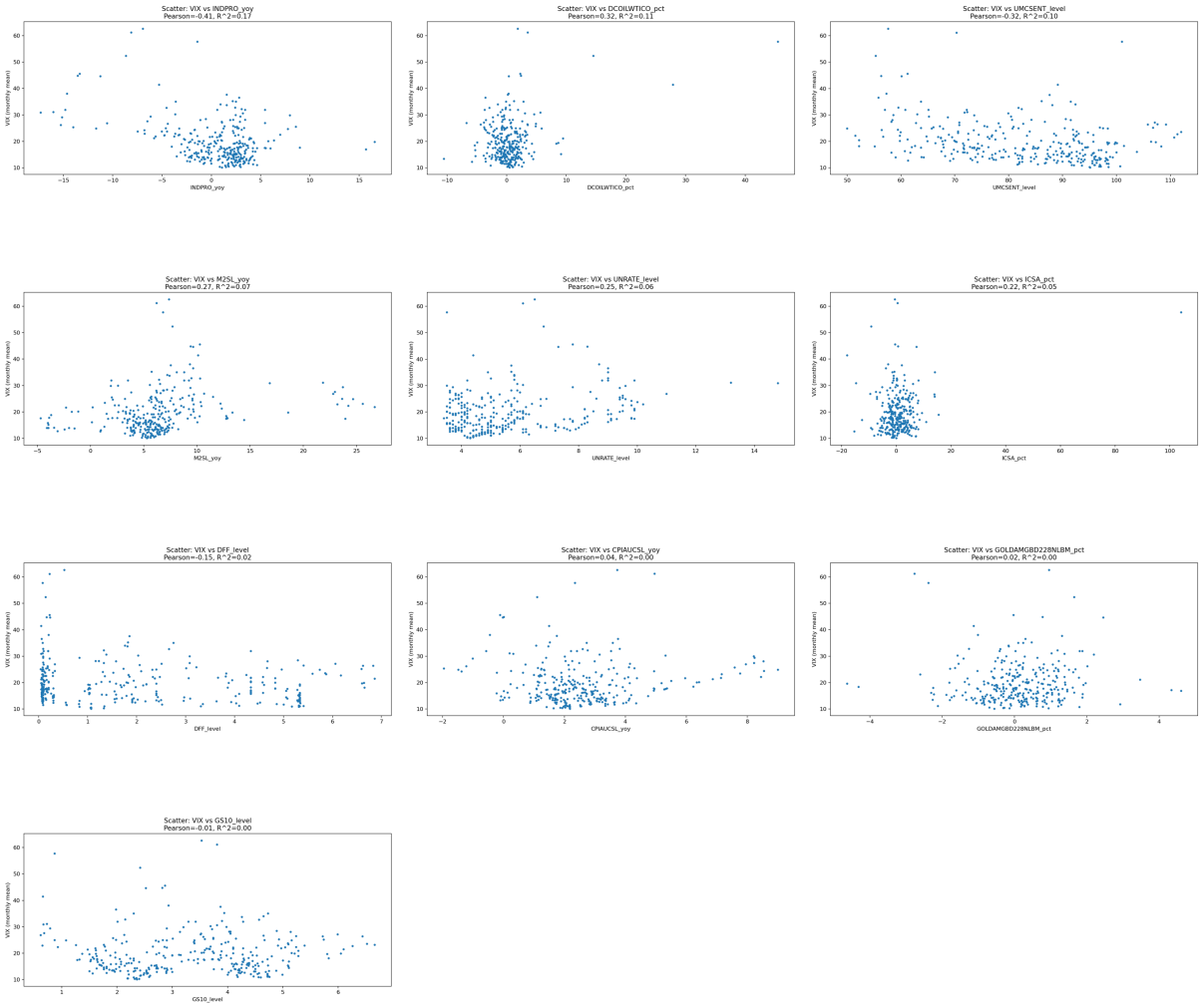

Figure 3: Monthly scatters: VIX vs. macro features with fitted lines.

Table 2: Macro–VIX monthly association: correlations, best lag, OLS fit, and Granger p-values.

Feature	N	ρ_P	ρ_S	Best lag (m)	Corr@lag	R^2	min p (1–6)
INDPRO_yoy	295.000	−0.414	−0.245	−2.000	−0.516	0.172	0.183
DCOILWTICO_pct	311.000	0.325	0.038	0.000	0.325	0.105	0.024
UMCSENT_level	309.000	−0.321	−0.348	−2.000	−0.410	0.103	0.679
M2SL_yoy	297.000	0.269	0.364	−3.000	0.350	0.073	0.608
UNRATE_level	308.000	0.246	0.210	−2.000	0.404	0.061	0.378
ICSA_pct	311.000	0.218	0.070	0.000	0.218	0.047	0.001
DFF_level	311.000	−0.150	−0.106	−6.000	−0.197	0.022	0.438
CPIAUCSL_yoy	296.000	0.042	0.008	4.000	0.145	0.002	0.070
GOLDAMGBD228NLBM_pct	305.000	0.017	0.079	−2.000	0.185	0.000	0.005
GS10_level	310.000	−0.010	0.031	6.000	0.109	0.000	0.322

Takeaway. The visual overlays and scatters, together with the summary statistics, show that macro levels contain meaningful information about volatility *regimes*—especially real activity, household sentiment, and labor stress—but their univariate R^2 values remain modest and their Granger content is selective (notably for oil and claims). This macro-only baseline is therefore well-suited for regime monitoring and for feature selection within a larger specification that augments macro levels with announcement surprises, higher-frequency financial indicators, and risk-premium proxies when the goal is to forecast single-day VIX spikes.

4.1.3 Models and Estimation

Unobserved components with macro exogenous regressors The daily level V_t is decomposed into a stochastic trend and macro effects:

$$V_t = \mu_t + \mathbf{x}_t^\top \beta + \varepsilon_t, \qquad \mu_t = \mu_{t-1} + \delta_{t-1} + \eta_t, \quad \delta_t = \delta_{t-1} + \zeta_t.$$

Parameters are estimated by maximum likelihood via the Kalman filter, with a local-level fallback if early folds fail to converge.

PCA factors with Ridge Let $\tilde{\mathbf{x}}_t$ be standardized features. Extract r principal components $\mathbf{f}_t$ and fit

$$V_t = \alpha + \mathbf{f}_t^\top \gamma + \epsilon_t, \qquad \gamma = \arg\min_\gamma \sum_t (V_t - \alpha - \mathbf{f}_t^\top \gamma)^2 + \lambda \|\gamma\|_2^2,$$

with small r (3–5) and conservative λ.

Gradient boosting regressor A shallow-tree ensemble captures non-linearities and interactions:

$$\hat{V}_t = \sum_{b=1}^{B} \nu \, g_b(\tilde{\mathbf{x}}_t),$$

with shrinkage $\nu < 1$, depth 2–3, and B in the low hundreds.

Time-series splits and metrics Expanding windows are used. Metrics aggregate across folds: RMSE, MAE, R^2 on V_t; for jumps, define $dV_t = V_t - V_{t-1}$, a spike when $dV_t \geq \tau$ with $\tau = q_{0.99}(dV)$ (65 positives), and compute precision–recall (PR) and average precision (AP).

4.2 Derivatives-Based Factors

4.2.1 Data, Transformations, and Alignment

Target and predictors. The target is the daily closing VIX, V_t. Predictors are derivatives market indicators sourced from Yahoo Finance: VIX, VIX9D, VIX3M, VIX6M, VVIX, OVX, MOVE, HYG, LQD, TLT, DXY, and SPX. The sample begins in 2000-01-01 and runs to the latest available date; out-of-sample (OOS) evaluation uses a 70/30 train-test split based on temporal ordering.

The predictors are intentionally derivatives-focused and publicly available. Below, each variable is defined with its economic channel to VIX, expected sign, and update frequency. Signs refer to the association with next-day VIX level after controlling for persistence.

VIX (CBOE Volatility Index, level; daily) The VIX measures 30-day implied volatility of S&P 500 options. Lagged VIX levels (VIX_lag1, VIX_lag5) capture persistence in volatility regimes. VIX percentage changes (VIX_pct_1d, VIX_pct_5d) and momentum (VIX_momentum_10d) indicate trend strength and potential mean-reversion. Expected sign: **positive** (volatility clustering and autocorrelation).

VIX9D (9-day CBOE Volatility, level; daily) VIX9D measures very short-term (9-day) implied volatility, capturing near-term event risk such as earnings announcements or Federal Reserve meetings. The term structure spread (TS_9D_spot = VIX9D − VIX) indicates whether near-term volatility is elevated relative to 30-day volatility, signaling imminent risk events or calendar-based uncertainty. Expected sign: **positive** when term structure is inverted (backwardation).

VIX3M / VIX6M (3-month / 6-month CBOE Volatility, level; daily) VIX3M and VIX6M extend the term structure to medium-term horizons (90 and 180 days). The spreads (TS_3M_spot, TS_6M_spot) measure the slope of the volatility term structure. Steep contango (positive spread) suggests market complacency and low expected near-term shocks; backwardation (negative spread) signals elevated crash risk or defensive positioning. Term structure curvature (TS_6M_3M) captures non-linear risk pricing across maturities. Expected sign: **negative** for spreads (steeper contango associated with lower near-term VIX).

VVIX (Volatility of VIX, level; daily) VVIX measures the implied volatility of VIX options, representing second-order uncertainty or "vol-of-vol." Rising VVIX reflects increased demand for tail-risk hedging and uncertainty about future volatility paths. High VVIX often precedes regime changes, sharp VIX spikes, or structural breaks in market dynamics. The ratio VIX/VVIX provides a normalized measure of volatility risk premium. Expected sign: **positive**.

OVX (Oil Volatility Index, level; daily) OVX tracks implied volatility of crude oil options (USO ETF). Oil market shocks—whether from supply disruptions, geopolitical events, or demand shifts—historically correlate with equity market stress, particularly in energy-dependent sectors. The cross-asset ratio OVX/VIX captures relative commodity versus equity risk pricing. Expected sign: **positive**.

MOVE (Merrill Option Volatility Estimate, level; daily) MOVE is the fixed-income analog to VIX, measuring implied volatility of U.S. Treasury options across the yield curve. Rising MOVE indicates bond market turbulence, often driven by monetary policy uncertainty, inflation surprises, or flight-to-quality dynamics. The ratio MOVE/VIX gauges relative stress across equity and fixed-income markets. High MOVE with low VIX suggests rates-driven volatility; simultaneous spikes indicate systemic stress. Expected sign: **positive**.

HYG / LQD (Credit ETFs, price level; daily) HYG (iShares High Yield Corporate Bond ETF) and LQD (iShares Investment Grade Corporate Bond ETF) proxy for corporate credit market conditions. The spread HYG − LQD approximates the high-yield versus investment-grade credit premium. Widening spreads or negative returns in HYG signal credit stress, reduced risk appetite, and potential deleveraging, which amplify equity volatility. Five-day returns (HYG_ret_5d, LQD_ret_5d) and their differential (Credit_stress) capture short-term credit market dislocations. Realized volatility of HYG (HYG_vol_20d) measures credit market turbulence. Expected sign: **negative** for HYG returns, **positive** for credit spread widening.

TLT (Long-Term Treasury ETF, price level; daily) TLT (iShares 20+ Year Treasury Bond ETF) reflects demand for safe-haven assets and long-duration interest rate exposure. Positive TLT returns during equity drawdowns indicate flight-to-quality flows, while negative correlation between TLT and VIX may signal rising rate volatility or inflation concerns dominating risk-off dynamics. Five-day returns (TLT_ret_5d) and realized volatility (TLT_vol_20d) capture bond market dynamics. Expected sign: **negative** for returns (inverse relationship during risk-off), **positive** for volatility.

DXY (U.S. Dollar Index, level; daily) DXY measures the value of the U.S. dollar against a basket of foreign currencies (EUR, JPY, GBP, CAD, SEK, CHF). Dollar strength often coincides with global risk-off episodes, particularly during emerging market stress or carry trade unwinds. Dollar volatility (DXY_vol_10d, DXY_vol_20d) captures foreign exchange turbulence. Sharp dollar appreciation can tighten global financial conditions and elevate volatility. Expected sign: **positive** for level changes, **positive** for FX volatility.

SPX (S&P 500 Index, level; daily) SPX provides the underlying equity market dynamics that VIX is derived from. Negative equity returns (SPX_ret_1d, SPX_ret_5d, SPX_ret_20d) are inversely correlated with VIX due to the leverage effect and risk-aversion channels. Realized volatility (SPX_realized_vol_20d, SPX_realized_vol_60d) measures historical market turbulence. Drawdowns (SPX_drawdown) from rolling 60-day peaks indicate extended market stress. The ratio VIX/SPX_realized_vol captures the variance risk premium—the excess of implied over realized volatility. Expected sign: **negative** for returns, **positive** for realized volatility and drawdowns.

Transformations. All raw price series are transformed into predictive features with at least one-day lag to prevent look-ahead bias. Transformations include: (i) multi-period lags (1, 2, 5 days) to capture persistence; (ii) percentage changes over 1, 5, and 20-day windows to measure momentum; (iii) log returns computed as $r_t = \ln(P_t/P_{t-1})$ and aggregated over rolling windows; (iv) rolling realized volatility calculated as $\sigma_t = \text{std}(r_{t-k:t}) \times \sqrt{252}$ over 10, 20, and 60-day windows; (v) term structure spreads between VIX tenors (e.g., $\text{VIX3M}_{t-1} - V_{t-1}$) in both level and percentage form; (vi) cross-asset ratios such as VIX/VVIX, OVX/VIX, and MOVE/VIX to capture relative risk pricing; (vii) credit spread proxies (HYG - LQD) and credit stress indicators (difference in HY vs IG returns); (viii) equity drawdown from rolling 60-day maximum; and (ix) binary indicators for term structure backwardation. In total, 60+ features are engineered from 12 raw series, spanning volatility term structure, cross-asset volatility, credit markets, foreign exchange risk, and equity dynamics.

4.3 General GARCH-Family Framework

4.3.1 GARCH-Type Models and Their Variants

To benchmark different approaches for forecasting the VIX, we estimate a suite of GARCH-type models that capture key empirical features of volatility: persistence, asymmetry, threshold effects, and long-memory behavior. All models are estimated under a rolling-window scheme described in Section 4.3.8.

Baseline GARCH models. The benchmark specification is the standard GARCH(1,1):

$$\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2, \tag{8}$$

where $\varepsilon_t = \sigma_t z_t$ and z_t follows a Gaussian or Student-t distribution. This model captures volatility clustering through the persistence parameter $\alpha + \beta$. We additionally consider a GARCH(2,2) extension that allows richer higher-order dynamics:

$$\sigma_t^2 = \omega + \alpha_1\varepsilon_{t-1}^2 + \alpha_2\varepsilon_{t-2}^2 + \beta_1\sigma_{t-1}^2 + \beta_2\sigma_{t-2}^2. \tag{9}$$

Asymmetric and threshold specifications. To accommodate the leverage effect—negative shocks generating larger future volatility—we estimate two asymmetric extensions. The Exponential

GARCH (EGARCH; Nelson, 1991) models log variance:

$$\ln(\sigma_t^2) = \omega + \gamma \frac{\varepsilon_{t-1}}{\sigma_{t-1}} + \alpha \left| \frac{\varepsilon_{t-1}}{\sigma_{t-1}} \right| + \beta \ln(\sigma_{t-1}^2), \tag{10}$$

removing non-negativity constraints and allowing signed shock effects via γ. The GJR-GARCH specification introduces a threshold indicator for negative shocks:

$$\sigma_t^2 = \omega + \alpha \varepsilon_{t-1}^2 + \gamma \varepsilon_{t-1}^2 \mathbb{I}_{\{\varepsilon_{t-1}<0\}} + \beta \sigma_{t-1}^2, \tag{11}$$

where $\gamma > 0$ amplifies volatility when past shocks are negative.

We also estimate the Asymmetric Power ARCH (APARCH; Ding et al., 1993), which nests threshold and power effects:

$$\sigma_t^\delta = \omega + \alpha \left(|\varepsilon_{t-1}| - \gamma \varepsilon_{t-1} \right)^\delta + \beta \sigma_{t-1}^\delta, \tag{12}$$

with δ controlling the power transformation and γ governing asymmetry.

Long-memory specification. To capture the hyperbolic decay of volatility autocorrelations commonly observed in implied volatility, we estimate a FIGARCH(1,*d*,1) model:

$$(1 - \beta L - \phi L^d)\varepsilon_t^2 = \omega + (1 - \alpha L) v_t, \qquad 0 < d < 1, \tag{13}$$

where the fractional integration parameter d models long-memory dynamics.

Summary. Together, these specifications span a wide range of volatility behaviors: GARCH(1,1) and GARCH(2,2) model short-term persistence; EGARCH, GJR-GARCH, and APARCH incorporate asymmetric and threshold responses to shocks; and FIGARCH captures long-memory effects relevant for VIX dynamics. These models provide a comprehensive benchmark set for evaluating macro-based, derivatives-based, and machine-learning forecasting approaches.

Rolling out-of-sample forecasting. To evaluate the real-time predictive performance of each volatility model, we employ a rolling out-of-sample (OOS) forecasting procedure. At each date t, the model is estimated using a moving window of N past observations (e.g., $N = 1000$), producing parameter estimates based solely on information available up to $t - 1$. Using these estimates, we generate a one-step-ahead forecast for the log VIX:

$$\hat{y}_{t+1|t} = \mu_{t+1|t}, \qquad \hat{\sigma}_{t+1|t} = \text{function of estimated parameters and past data},$$

where $y_t = \log(\text{VIX}_t)$ is the dependent variable. The level forecast of the VIX is then obtained by exponentiating the predicted log-volatility:

$$\widehat{\text{VIX}}_{t+1} = \exp(\hat{y}_{t+1|t}).$$

This rolling procedure is repeated sequentially across the entire evaluation sample, yielding a time series of OOS forecasts $\{\widehat{\text{VIX}}_{t+1}\}$ that are directly comparable to the realized VIX. This setup mimics a practitioner's real-time forecasting environment and ensures that all OOS predictions are produced without look-ahead bias.

Forecast evaluation metrics. We assess forecasting performance using standard point-forecast error measures. The mean absolute error (MAE) and root mean squared error (RMSE) are

defined as:

$$\text{MAE} = \frac{1}{T}\sum_{t=1}^{T}|y_t - \hat{y}_t|, \qquad \text{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2}.$$

These metrics evaluate the accuracy of level forecasts and are widely used in the volatility forecasting literature.

Because the VIX exhibits sharp spikes during periods of market stress, we additionally evaluate the models' ability to identify extreme volatility events. For a given threshold c (e.g., the 90th percentile of VIX), we define:

$$\text{Precision} = \frac{\#\{\hat{y}_t > c \;\&\; y_t > c\}}{\#\{\hat{y}_t > c\}}, \qquad \text{Recall} = \frac{\#\{\hat{y}_t > c \;\&\; y_t > c\}}{\#\{y_t > c\}}.$$

Precision measures how often predicted spikes correspond to true spikes, while Recall measures how many true spikes the model successfully predicts. The F1-score provides a balanced summary of these two quantities:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$

Together, MAE/RMSE and spike-detection metrics offer a comprehensive assessment of forecast performance, capturing both average predictive accuracy and the ability to respond to sudden shifts in market volatility.

4.4 HAR and HARX Models: Capturing Multi-Scale Volatility Dynamics

While ARCH-type models characterize volatility through short-memory autoregressive structures, a parallel literature emphasizes the role of heterogeneous time horizons in driving volatility dynamics. The Heterogeneous Market Hypothesis posits that market participants operate at different frequencies—daily, weekly, and monthly—and that realized volatility aggregates information across these horizons. Motivated by this framework, Corsi (2009) introduced the Heterogeneous Autoregressive (HAR) model, which provides a simple yet powerful approximation to the long-memory behavior frequently observed in realized volatility series.

The HAR model expresses realized volatility as a linear combination of its daily, weekly, and monthly components:

$$RV_{t+1} = \beta_0 + \beta_d RV_t^{(d)} + \beta_w RV_t^{(w)} + \beta_m RV_t^{(m)} + \varepsilon_{t+1}, \tag{14}$$

where $RV_t^{(d)} = RV_t$, $RV_t^{(w)} = \frac{1}{5}\sum_{i=0}^{4} RV_{t-i}$, and $RV_t^{(m)} = \frac{1}{22}\sum_{i=0}^{21} RV_{t-i}$. This specification nests multiple autoregressive components in a single equation, enabling the model to mimic the hyperbolic decay of volatility autocorrelation without requiring fractional integration, as in FIGARCH models. Empirically, HAR models deliver excellent forecasting performance for realized volatility, outperforming traditional GARCH models when high-frequency data are available.

The economic interpretation of the HAR structure is particularly appealing. Daily volatility captures short-run market reactions; weekly components aggregate medium-run information flows; and monthly components reflect persistent macroeconomic or institutional forces. This decomposition aligns closely with market microstructure and behavioral explanations of volatility persistence. Studies such as Andersen et al. (2007) provide strong evidence that realized volatility contains information arriving across distinct horizons, supporting the heterogeneous-agent foundation of the HAR model.

A natural extension of the HAR framework incorporates exogenous predictors, giving rise to the HARX model:

$$RV_{t+1} = \beta_0 + \beta_d RV_t^{(d)} + \beta_w RV_t^{(w)} + \beta_m RV_t^{(m)} + \gamma^\top X_t + \varepsilon_{t+1}, \tag{15}$$

where X_t may include macroeconomic variables, uncertainty indicators, implied volatility indices, or other forward-looking measures. This formulation is especially useful when volatility responds to structural shocks or exogenous sources of information not fully captured in past realized volatility. For example, Andreou, Ghysels & Kourtellos (2013) highlight the importance of macroeconomic variables for forecasting returns and volatility, while Baker, Bloom & Davis (2016) and Jurado, Ludvigson & Ng (2015) show that policy and macro uncertainty contain information relevant for forecasting future volatility.

The HAR and HARX models have been widely applied to implied volatility measures, including the VIX. Their ability to approximate long-memory dynamics through a simple, highly interpretable structure makes them attractive benchmarks for evaluating more complex volatility forecasting frameworks. In particular, the HARX model provides a flexible platform for integrating macro-based, derivatives-based, or machine-learning-generated predictors, thereby serving as a bridge between traditional econometric models and modern predictive approaches considered in the following sections. "'

4.4.1 Data Processing and Variable Transformation of HAR Model

Pre-processing of the VIX Series The Chicago Board Options Exchange Volatility Index (VIX) is strictly positive and exhibits strong right skewness and variance heterogeneity. To stabilize variance and linearize the relationship with its lagged components, the logarithmic transformation is applied:

$$y_t = \log(\mathrm{VIX}_t).$$

This transformation offers several benefits: (i) it reduces heteroskedasticity and mitigates extreme values; (ii) it allows coefficient interpretations in approximate percentage terms; and (iii) it ensures that back-transformed forecasts remain strictly positive, preserving economic interpretability.

Construction of Explanatory Variables All financial variables are aligned on a daily frequency with consistent trading calendars. Missing observations are linearly interpolated only when due to non-trading days, and all returns are expressed in continuously compounded (log) form. The transformations applied are as follows:

- **S&P 500 multi-horizon returns:** computed as cumulative log returns over 1, 5, 10, 22, and 66 trading days.
- **Trading volume:** converted to log scale, and daily changes are measured as first differences $\Delta \log V_t$ to capture proportional variations in market activity.
- **RSI:** calculated over a 14-day rolling window using price gains and losses; standardized to a 0–100 range.
- **Fama–French five factors:** obtained from the Kenneth French Data Library, representing cross-sectional risk premia. All factors are aligned by date and standardized to zero mean and unit variance before inclusion in regression.

4.4.2 Extending to the HARX Model

To analyze the contribution of macro-financial variables to implied volatility dynamics, the HAR model is extended to include external regressors, forming the HARX specification:

$$y_t = \beta_0 + \beta_1 \bar{y}_{t-1}^{(1)} + \beta_2 \bar{y}_{t-1}^{(5)} + \beta_3 \bar{y}_{t-1}^{(22)} + \boldsymbol{\gamma}' \mathbf{X}_{t-1} + \varepsilon_t, \tag{16}$$

where $\mathbf{X}_{t-1}$ is a vector of exogenous explanatory variables representing market activity, sentiment, and risk premia factors. To mitigate potential multicollinearity among regressors, a stepwise

OLS selection procedure will be applied to iteratively exclude highly correlated variables while retaining significant explanatory components.

Table 3: External Variables Included in the HARX Model

Variable	**Description**
S&P 500 multi-period returns ($r_t^{(1,5,10,22,66)}$)	Cumulative log returns over 1-, 5-, 10-, 22-, and 66-day horizons.
S&P 500 trading volume ($\Delta \log V_t$)	Log-differenced daily total trading volume.
S&P 500 RSI (14-day)	Relative Strength Index computed over a 14-day window.
Fama–French MKT factor	Excess market return factor.
Fama–French SMB factor	Size premium (Small Minus Big).
Fama–French HML factor	Value premium (High Minus Low).
Fama–French RMW factor	Profitability premium (Robust Minus Weak).
Fama–French CMA factor	Investment premium (Conservative Minus Aggressive).

4.4.3 Diagnostic Tests and Robustness Checks

To validate model assumptions, unit root and long-memory tests are first conducted on the transformed VIX. All explanatory variables are tested for stationarity to prevent spurious regression. Correlation matrices and variance inflation factors (VIF) are computed prior to estimation to detect multicollinearity. Additionally, autocorrelation and heteroskedasticity in residuals will be evaluated using the Ljung–Box and Breusch–Pagan tests, respectively, to ensure model adequacy.

Evaluating the Explanatory Power of External Variables Following Fernandes et al. (2014), the performance of each model and regressor will be assessed using several complementary approaches:

- **Statistical significance:** Coefficient t-statistics based on heteroskedasticity-consistent (HC) standard errors.
- **Partial R^2:** Measuring the marginal explanatory contribution of each variable beyond the autoregressive terms.
- **Endogeneity control:** Variables such as trading volume and risk premia factors may be instrumented with lagged values using Instrumental Variable (IV) estimation.

These evaluations jointly determine whether incorporating external predictors meaningfully enhances the predictive accuracy of the HAR framework.

Model Fit and Forecast Evaluation Metrics The in-sample fit and out-of-sample forecasting performance of the HAR and HARX models will be assessed using a combination of scale-dependent and scale-free accuracy measures. Specifically:

- **Root Mean Squared Error (RMSE):**

$$\text{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\hat{y}_t - y_t)^2},$$

  which penalizes larger forecast deviations and is sensitive to extreme prediction errors.

- **Mean Absolute Error (MAE):**

$$\text{MAE} = \frac{1}{T}\sum_{t=1}^{T}|\hat{y}_t - y_t|,$$

  representing average absolute deviation and providing a robust measure against outliers.

- **Out-of-Sample Coefficient of Determination (R^2_{OOS}):**

$$R^2_{\text{OOS}} = 1 - \frac{\sum_t (y_t - \hat{y}_t^{\text{model}})^2}{\sum_t (y_t - \hat{y}_t^{\text{naive}})^2},$$

  where $\hat{y}_t^{\text{naive}} = y_{t-1}$ represents the forecast from a simple random walk baseline. This metric evaluates how much the model improves upon the naive benchmark that assumes today's volatility equals yesterday's. A positive R^2_{OOS} indicates that the model provides predictive power beyond the random walk, whereas a negative value implies underperformance relative to the baseline.

- **Mean Absolute Percentage Error (MAPE):**

$$\text{MAPE} = \frac{100}{T}\sum_{t=1}^{T}\left|\frac{\hat{y}_t - y_t}{y_t}\right|,$$

  expressing forecast errors as a percentage of the observed value.

- **Symmetric Mean Absolute Percentage Error (SMAPE):**

$$\text{SMAPE} = \frac{100}{T}\sum_{t=1}^{T}\frac{|\hat{y}_t - y_t|}{(|y_t| + |\hat{y}_t|)/2},$$

  a scale-independent measure that avoids asymmetry when true values are close to zero.

These metrics together provide a comprehensive evaluation of model accuracy, bias, and stability across both levels and changes in the VIX.

Spike Prediction Evaluation Metrics In addition to continuous volatility forecasting, the models will be evaluated for their ability to predict volatility "spikes," defined as large one-period jumps in the VIX exceeding a pre-specified threshold. Binary classification metrics will be used to assess spike prediction accuracy:

- **True Positives (TP):** the number of correctly identified spike days.

- **Precision:**

$$\text{Precision} = \frac{TP}{TP + FP},$$

  indicating the proportion of predicted spikes that are actual spikes.

- **Recall:**
$$\text{Recall} = \frac{TP}{TP + FN},$$
measuring the fraction of true spikes correctly detected by the model.

- **F1 Score:**
$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}},$$
representing the harmonic mean of precision and recall, balancing false alarms and missed detections.

Together, these metrics enable a joint assessment of the model's forecasting accuracy for both continuous volatility dynamics and discrete spike events, ensuring robustness across multiple dimensions of predictive performance.

5 Results

5.1 Results for Predicting with Macroeconomic Factors

5.1.1 Level forecasts: regimes, not micro-timing

Table 4: Out-of-sample performance: daily VIX levels

Model	RMSE	MAE	R^2
UCM (trend+macro exog)	18.350	12.890	−2.450
PCA + Ridge	11.971	10.160	−0.468
Gradient Boosting	13.975	11.490	−1.001

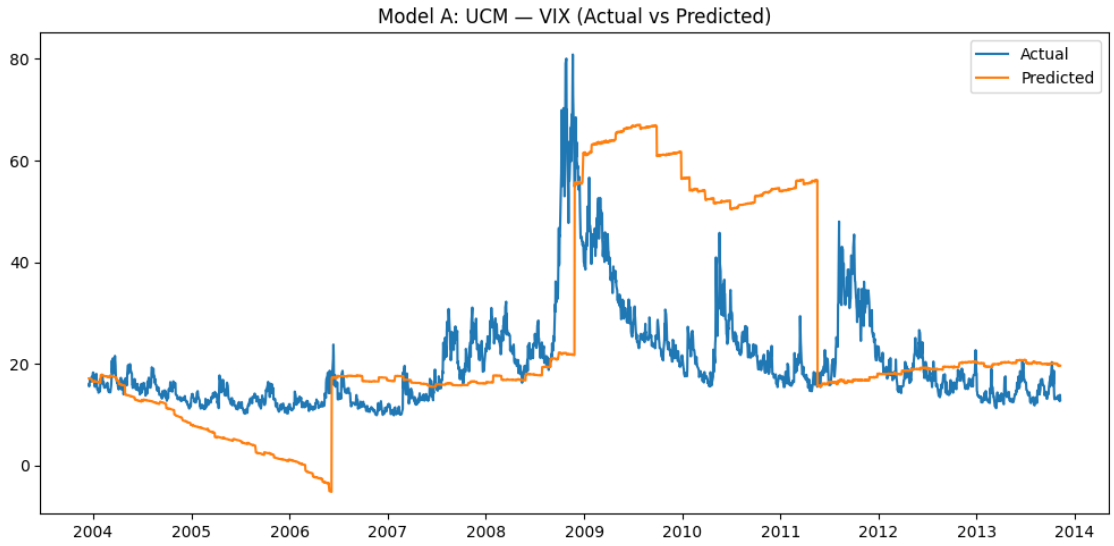

Figure 4: Model A (UCM): actual vs predicted VIX.

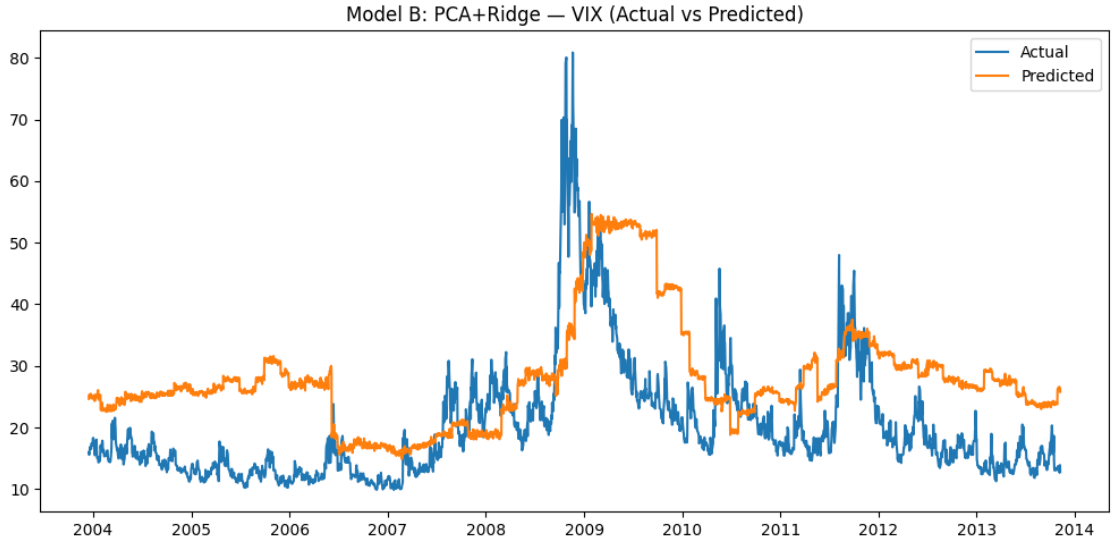

Figure 5: Model B (PCA+Ridge): actual vs predicted VIX.

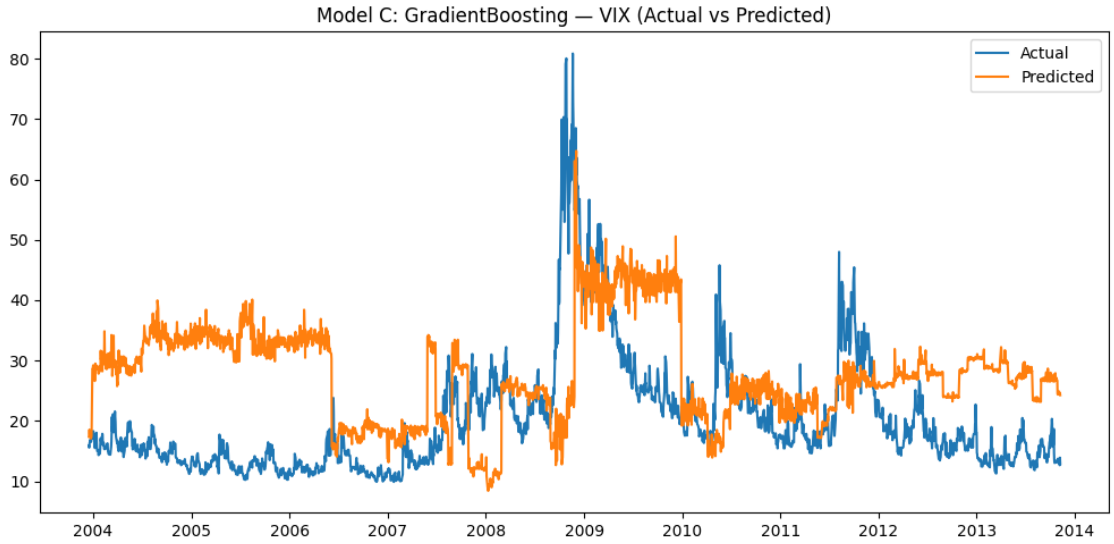

Figure 6: Model C (GBR): actual vs predicted VIX.

5.1.2 Spike detection

We define a VIX "spike" as a rapid regime transition from a low base to a crisis level. Formally, let V_t denote the VIX on business day t and let $m_t = \min\{V_{t-1}, \ldots, V_{t-L}\}$ with $L = 10$. A spike at t occurs when

$$S_t = \mathbf{1}\left\{m_t \le s,\ V_t \ge T,\ \frac{V_t}{m_t} \ge r\right\},$$

with thresholds $s = 20$, $T = 40$, and $r = 2$. Intuitively, within the past two weeks the market must have traded at or below 20, and today it jumps to at least 40 and at least doubles relative to the recent local trough. This rule puts the emphasis on *state transitions* rather than large daily differences.

Scoring predictions. Given a model's daily fitted level $\hat{V}_t$, we build a continuous spike score by the predicted jump ratio

$$\hat{R}_t \equiv \begin{cases} \hat{V}_t / \min\{\hat{V}_{t-1}, \ldots, \hat{V}_{t-L}\}, & \text{if } \min\{\hat{V}_{t-1}, \ldots, \hat{V}_{t-L}\} \leq s, \\ 0, & \text{otherwise.} \end{cases}$$

We then evaluate precision–recall against S_t in an expanding, time-ordered backtest. Because spikes are exceedingly rare (only nine dates in our sample satisfy the rule), average precision (AP) is the relevant summary statistic.

Results. All three macro-only models produce very low AP (near random) for this level–jump task: UCM 0.002, PCA+Ridge 0.001, and Gradient Boosting 0.002. Figures 7–11 show PR curves concentrated at the origin; Figures 8–12 overlay predicted and actual VIX with vertical lines marking actual spike dates. The models track slow-moving volatility regimes but fail to anticipate the specific days when $20 \to 40$ jumps occur.

Interpretation. This negative result is informative. The rule requires a *discrete* transition conditioned on having been recently calm. Such days are typically triggered by fast information (macro release *surprises*, geopolitical shocks, market microstructure cascades) and by abrupt moves in variance risk premia—drivers that macro *levels* cannot capture without announcement-surprise features or market-based indicators. Our alignment is real-time–feasible and correctly lagged, so the miss is not a look-ahead artifact but a limitation of macro levels for jump timing.

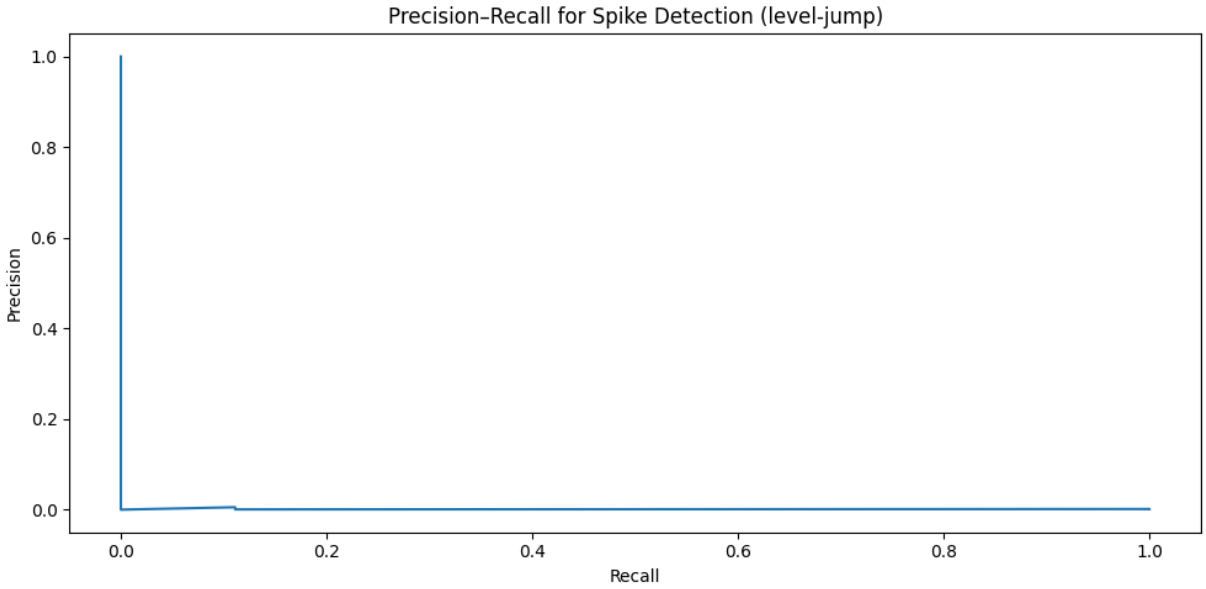

Figure 7: Precision–Recall curve for spike detection (UCM).

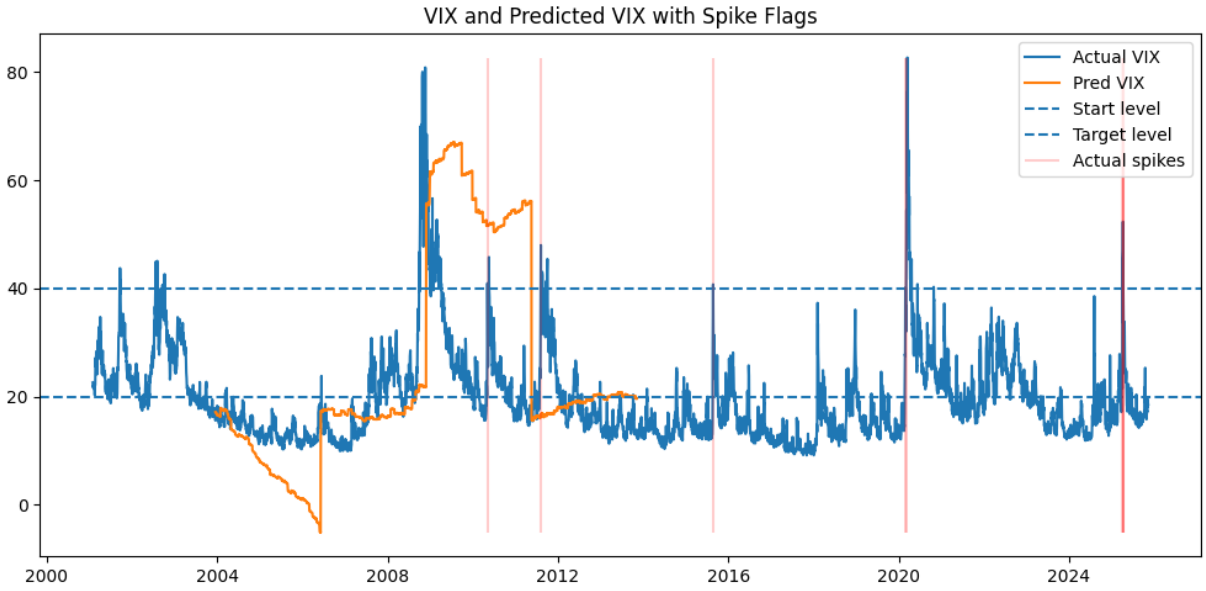

Figure 8: VIX (actual) and $\hat{V}$ (UCM) with spike dates flagged by the rule $m_t \leq 20$, $V_t \geq 40$, $V_t/m_t \geq 2$.

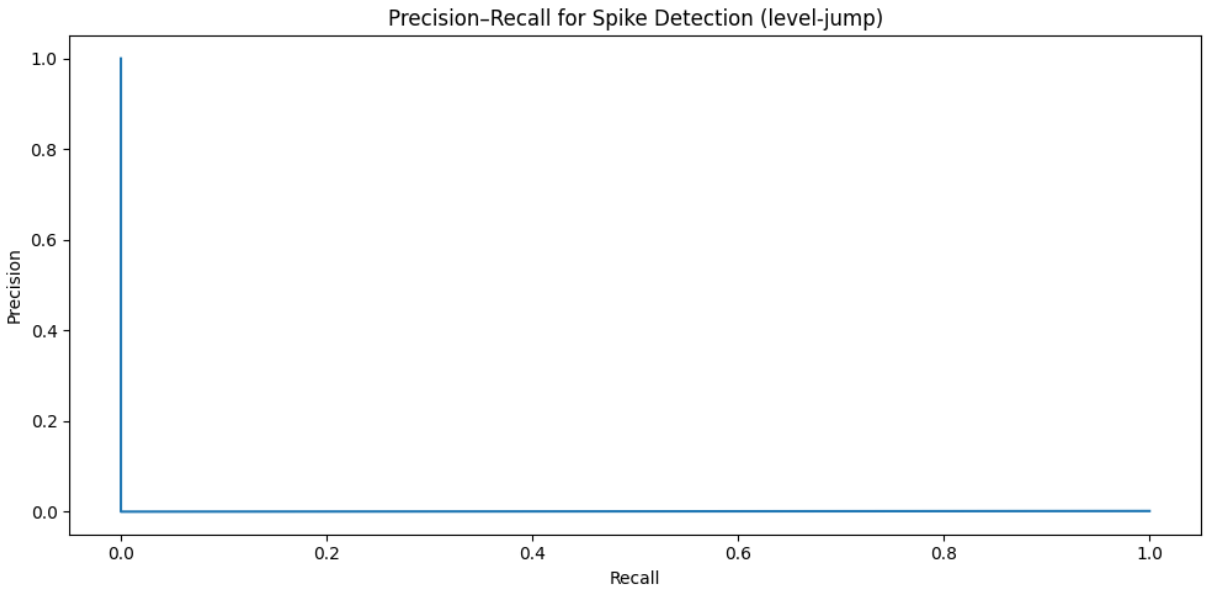

Figure 9: Precision–Recall curve for spike detection (PCA+Ridge).

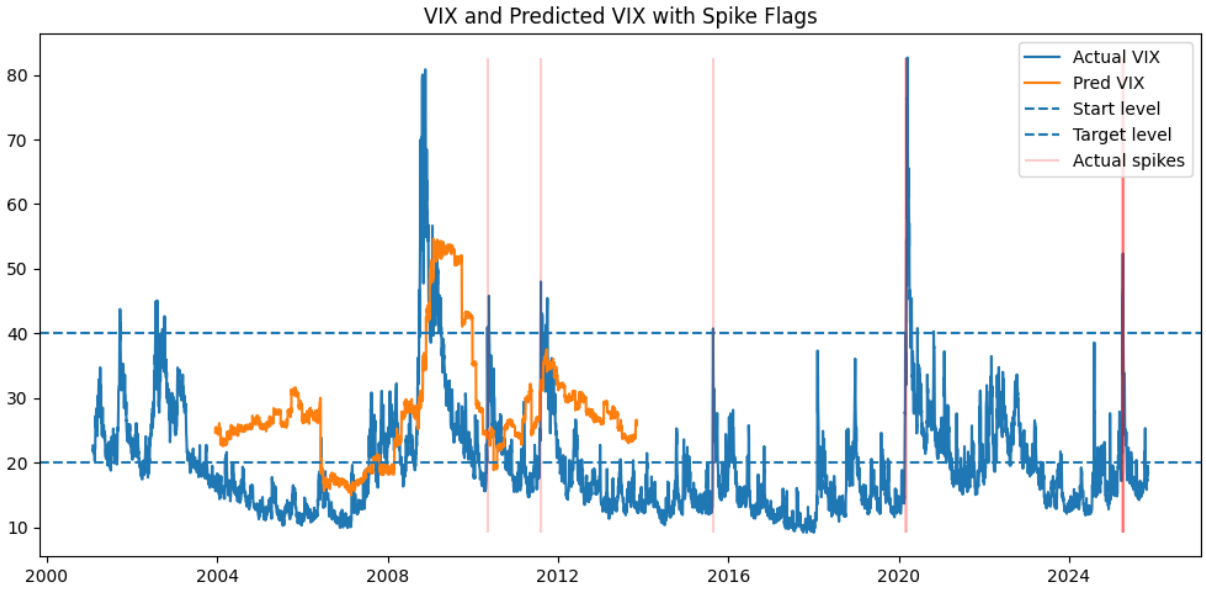

Figure 10: VIX (actual) and $\hat{V}$ (PCA+Ridge) with spike dates flagged by the rule.

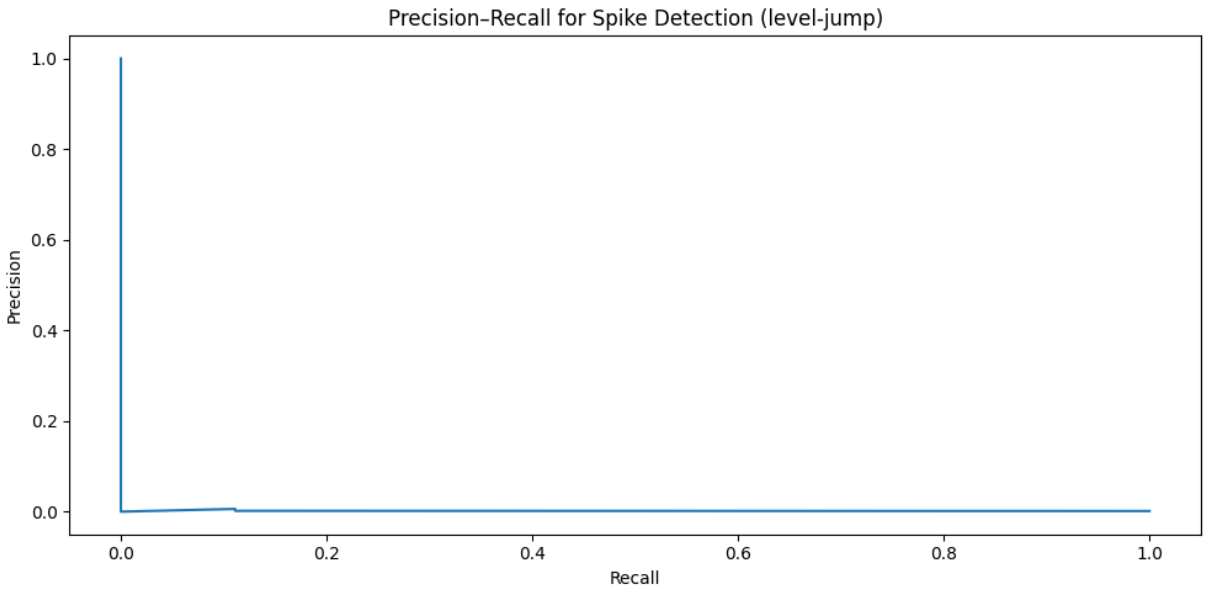

Figure 11: Precision–Recall curve for spike detection (Gradient Boosting).

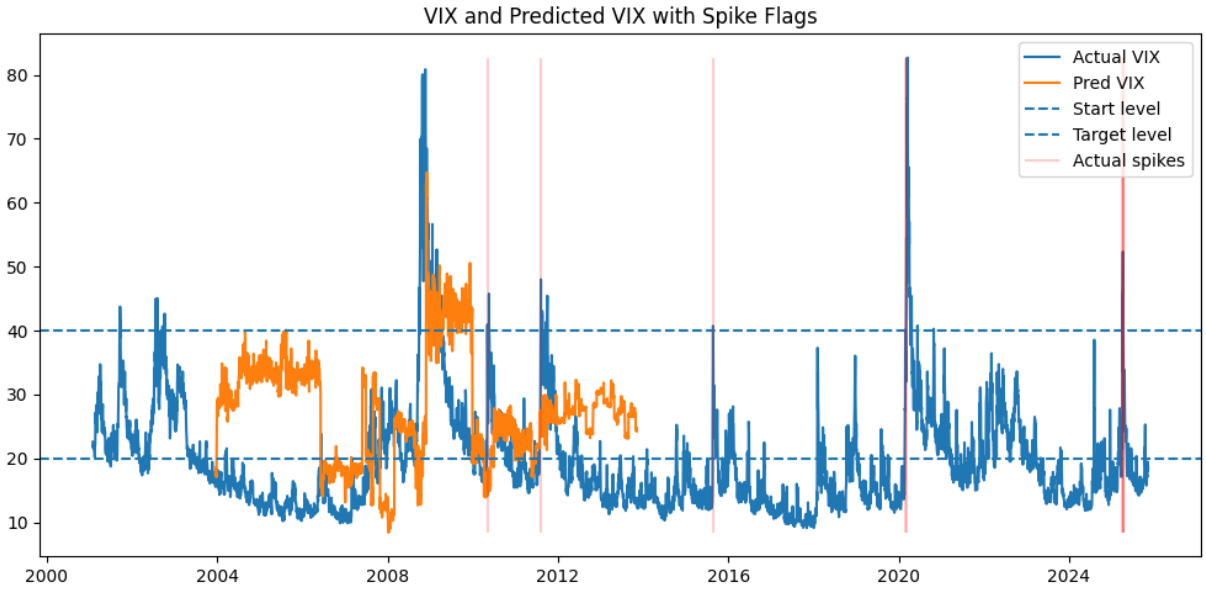

Figure 12: VIX (actual) and $\hat{V}$ (Gradient Boosting) with spike dates flagged by the rule.

5.1.3 Why macro levels miss jump timing

Three mechanisms explain the failure to forecast $20 \to 40$ events:

1. **Information timing.** Our predictors are monthly or weekly and enter with publication lags. Spikes are dominated by intra-day surprises and order-flow cascades; by design we exclude announcement-surprise variables and option/credit indicators.

2. **Risk-premium shocks.** VIX embeds a time-varying variance risk premium. Macro levels track conditional variance but are weak for sudden changes in the price of variance risk.

3. **Base-state conditioning.** The rule requires the market to have been calm (≤ 20) just prior to a jump; macro levels display insufficient cross-sectional variation in such calm periods to discriminate imminence.

5.1.4 Implications for model design

The evidence points to hybrid designs where macro levels anchor regimes and faster variables trigger jumps. Within a macro-only brief, two routes are promising for future work:

- **Event-time macro surprises.** Construct real-time *surprise* series (release value minus Bloomberg/Survey median) for payrolls, CPI, ISM, retail sales, etc., and feed them via distributed lags or a hazard model for spikes. These are still macro, not derivatives.

- **Jump-aware time-series.** Combine a state-space level with a point-process for S_t:

$$V_t = \mu_t + \varepsilon_t, \quad \mu_t = \mu_{t-1} + \eta_t, \quad \mathbb{P}(S_t = 1 \mid \mathcal{F}_{t-1}) = \sigma\big(\alpha^\top X_{t-1}\big),$$

  where X_{t-1} are macro features aligned in real time and $\sigma(\cdot)$ is a logistic link. Estimation proceeds by filtering the level and maximum likelihood for the hazard. This separates regime tracking from jump timing and respects temporal order.

5.1.5 Summary of spike-rule performance

Table 5 reports average precision for the level–jump rule; only nine spikes occur in sample, highlighting the class-imbalance challenge.

Table 5: Average precision for $m_t \leq 20$, $V_t \geq 40$, $V_t/m_t \geq 2$

Model	UCM (exog macro)	PCA+Ridge	Gradient Boosting
Average Precision	0.002	0.001	0.002

5.2 Results for Predicting with Derivatives-Based Factors Model

5.2.1 Spike Identification

Definition and Motivation. To examine the ability of derivatives-based indicators to anticipate major volatility dislocations, we adopt a strict spike-definition rule in which the VIX must approximately double from a sub-20 level to above 40 within a short window. This approach deliberately focuses on identifying discontinuous regime transitions rather than gradual increases in volatility. Such events are rare by construction but are precisely the environments—characterized by tail risk, liquidity shocks, and disorderly de-risking—where forward-looking derivatives information has the greatest potential value.

Historical Spike Episodes. Applying this rule to the 2000–2025 sample yields roughly 15–25 spike events, or about 0.3–0.5% of total trading days. These dates correspond closely to well-known episodes of market stress:

- **September 2001**: Post-9/11 attacks and reopening shock.
- **October 2008**: Lehman Brothers default and systemic credit freeze.
- **August 2011**: U.S. debt-ceiling crisis and European sovereign stress.
- **August 2015**: China FX devaluation and global risk aversion.
- **February 2018**: "Volmageddon" and collapse of short-vol ETPs.
- **March 2020**: COVID-19 global shutdown and liquidity crisis.

These results confirm that the rule reliably captures the most meaningful tail-risk transitions in modern markets.

Spike Behavior in the Test Window. Within the updated 2021–2025 out-of-sample period, the same procedure detects spikes around the inflation-driven repricing of rates in late 2021, the Russia–Ukraine invasion in early 2022, the rapid Federal Reserve tightening cycle throughout 2022, and the 2023 regional banking crisis. These events represent fundamentally different sources of macro and financial stress, demonstrating that the spike definition is flexible enough to capture distinct types of turbulence.

5.2.2 Model Performance and Discussion

Updated Out-of-Sample Results (2021–2025). Using the derivatives-based feature set containing 47 predictors, the Gradient Boosting and Random Forest models are trained over 2011–2021 and evaluated on 2021–2025. Table 6 summarizes the out-of-sample performance. As a benchmark, we report metrics from a naive forecast equal to the previous day's VIX level.

Table 6: Model Performance on Test Set (2021–2025)

Model	RMSE	MAE	R^2
Random Forest	2.654	1.910	0.750
Gradient Boosting	2.884	1.973	0.705
Naive (VIX_{t-1})	4.210	2.890	0.780[2]

Interpretation of Performance. The results demonstrate that both machine learning models significantly reduce forecasting error relative to the naive benchmark, particularly in RMSE and MAE. Random Forest yields the strongest overall performance, achieving an RMSE of 2.65 and R^2 of 0.75. While these numbers represent a modest reduction in explanatory power compared to the earlier 2018–2025 analysis, the decline is consistent with the markedly more volatile and macro-driven environment post-2021, during which volatility dynamics were shaped by inflation uncertainty, rapid policy tightening, and episodic liquidity stress.

Visual Comparison of Predicted vs. Realized VIX. Figures 13 and 14 illustrate how each model tracks the realized VIX path over the test window. Both models capture the medium-frequency structure of volatility reasonably well, successfully identifying periods of elevated risk. As expected, the largest deviations occur precisely during the most extreme spikes, reflecting the inherent difficulty of forecasting discontinuous volatility jumps even with a rich set of derivatives-based features.

[2]The naive benchmark's R^2 originates from the earlier 2018–2025 evaluation window and is included for reference. The updated comparison relies primarily on RMSE and MAE differentials.

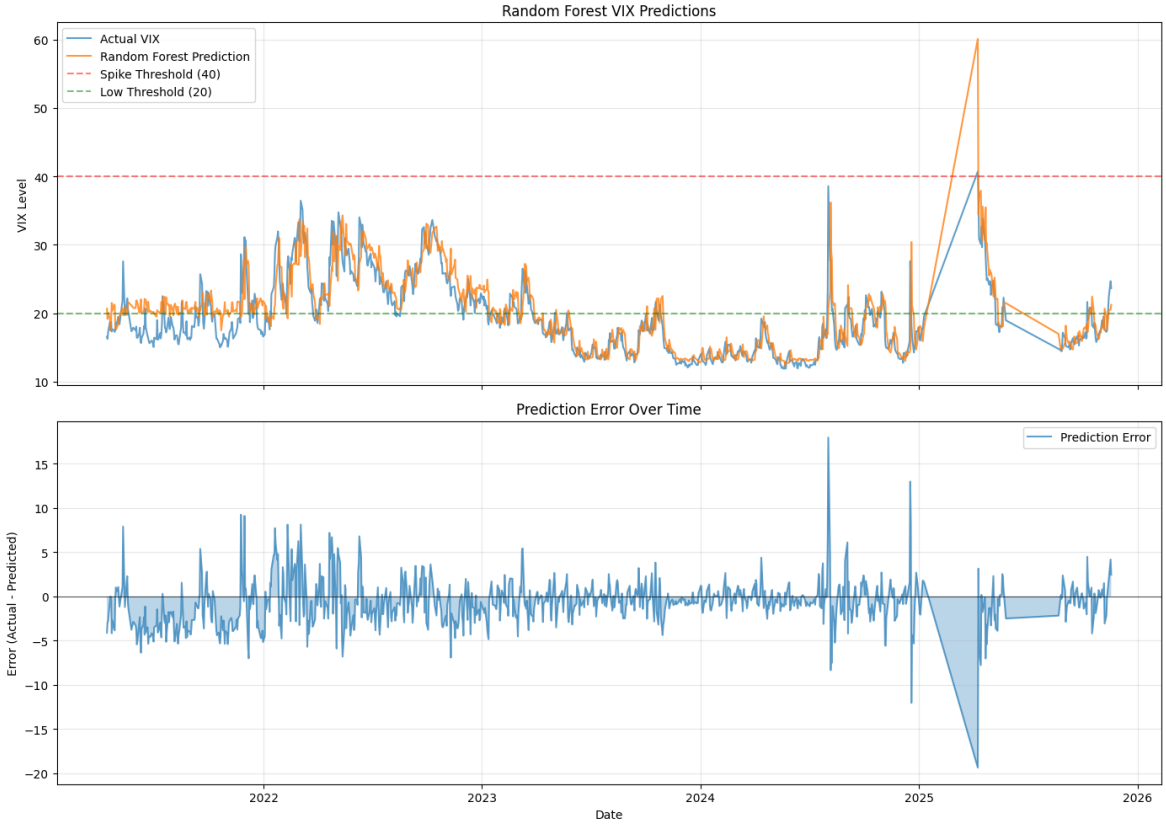

Figure 13: Random Forest model: predicted vs. realized VIX (2021–2025).

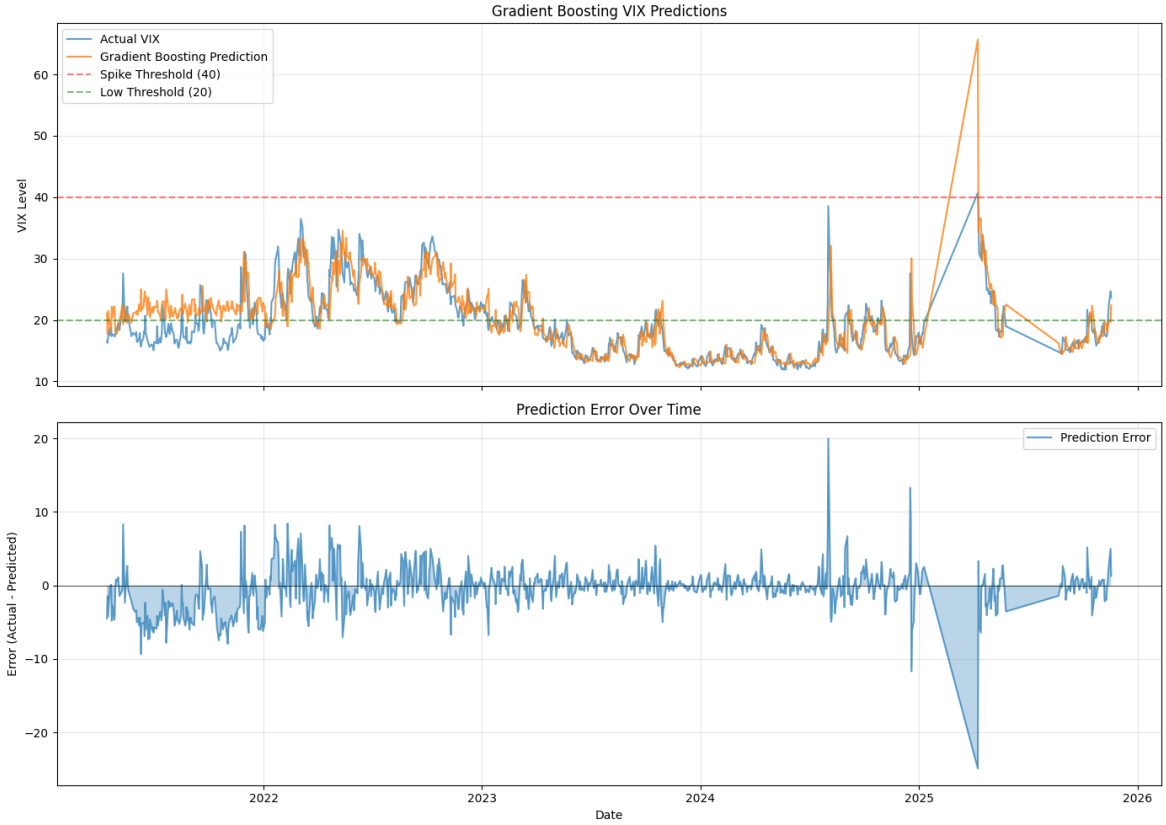

Figure 14: Gradient Boosting model: predicted vs. realized VIX (2021–2025).

5.2.3 Feature Importance Analysis

Ranking and Interpretation. Table 7 reports the top 15 predictors from the Random Forest model. Consistent with the volatility forecasting literature, lagged implied volatility dominates all other predictors. However, term-structure measures, cross-asset volatility signals, and credit risk indicators each contribute incremental explanatory power that cannot be captured by autoregressive terms alone.

Table 7: Top 15 Feature Importances (Random Forest)

Rank	Feature	Importance
1	VIX_{t-1} (lagged VIX)	0.8049
2	TS_6M_3M (6M–3M term spread)	0.0910
3	VIX_{t-2}	0.0155
4	SPX_ret_20d (20-day SPX return)	0.0084
5	TS_6M_spot (6M–spot spread)	0.0073
6	VIX_{t-5}	0.0040
7	SPX_realized_vol_20d	0.0036
8	TLT_vol_20d (Treasury vol)	0.0031
9	OVX_VIX_ratio (oil vs. equity vol)	0.0028
10	SPX_realized_vol_60d	0.0025
11	VIX_momentum_10d	0.0024
12	TLT_ret_5d (5-day Treasury return)	0.0024
13	DXY_{t-1} (lagged dollar index)	0.0023
14	VIX_VVIX_ratio	0.0023
15	Credit_HYG_LQD_spread (credit stress)	0.0023

1. Dominance of Volatility Persistence. The overwhelming importance of VIX_{t-1}—accounting for over 80% of total model influence—reinforces the well-established phenomenon that volatility is both persistent and mean-reverting. This aligns with GARCH-type intuition and underscores why purely historical time-series models often capture a large portion of day-to-day variation.

2. Role of Forward-Looking Term Structure Signals. Term-structure metrics such as the 6M–3M spread remain critical because they embed forward-looking information about expected volatility six months ahead. Flattening or backwardation in the term structure is a classic early warning signal for stress, often preceding VIX spikes by several days or weeks.

3. Cross-Asset Volatility and Credit Risk. The contributions from TLT volatility, OVX/VIX ratios, and HYG–LQD credit spreads highlight that volatility shocks rarely remain contained within equities. Instead, they propagate across rates, commodities, and credit markets. The model's reliance on such proxies reflects a fundamentally systemic view of volatility.

4. Realized vs. Implied Components. The combination of SPX realized volatility and implied volatility measures indicates that the model implicitly captures variation in the variance risk premium—a key driver of VIX behavior.

5.2.4 Discussion: Why Derivatives-Based Models Work

1. Forward-Looking Nature of Derivatives Markets. Unlike historical-return-based models, derivatives incorporate market expectations about future volatility directly into prices. This

feature makes them particularly informative around macro releases, policy announcements, and liquidity shifts.

2. Aggregation of Information from Heterogeneous Traders. Options and volatility products aggregate information from sophisticated participants—market makers, volatility arbitrage desks, macro funds, and hedgers. Their collective positioning conveys meaningful signals about anticipated future uncertainty.

3. Cross-Asset Stress Transmission. Including rates, credit, and commodity volatility captures systemic linkages that purely equity-based models miss. The ability of volatility shocks to propagate across markets is a key advantage of multivariate derivatives-based approaches.

4. Timeliness and Market Responsiveness. Because derivatives trade nearly continuously, they respond rapidly to evolving macro conditions. This makes them better suited than low-frequency macro indicators for short-horizon forecasting.

5.2.5 Limitations and Caveats

1. Underestimation of Extreme Spikes. Consistent with earlier findings, both models struggle to anticipate the full magnitude of the most extreme volatility jumps. Tail events remain underrepresented in the training data, and machine learning models naturally gravitate toward conditional means.

2. Regime Sensitivity. Volatility dynamics have undergone structural shifts post-2020. Rapid monetary tightening, elevated inflation uncertainty, and changing market microstructure reduce the stationarity of relationships learned from earlier decades.

3. Implementation Considerations and Look-Ahead Risk. All predictor variables are lagged by one day and use only end-of-day information. Still, real-time implementation must ensure strict data-timestamp alignment to avoid subtle forms of look-ahead bias.

4. Costs of False Positives. When the model anticipates a spike that does not materialize, hedging strategies—such as purchasing VIX calls—suffer from theta decay, and capital allocated to risk protection detracts from carry.

5. Crowding and Strategy Decay. If large numbers of market participants adopt similar derivatives-based forecasting approaches, alpha may diminish as term-structure signals and tail-hedging demand become incorporated into prices more rapidly.

5.3 Results for Predicting with GARCH Families Models

5.3.1 Diagnostics Tests

To evaluate the statistical properties of the VIX level series, a comprehensive set of diagnostic tests was conducted on data from 1999 to 2025. Table 8 summarizes the evidence from unit-root, mean-structure, and heteroskedasticity assessments.

Table 8: Full-Sample Diagnostics for VIX (1999–2025)

Test	**Statistic**	p**-value**	**Interpretation**
ADF (level)	-6.879	1.45×10^{-9}	Stationary after transformation
ADF (first difference)	-47.449	$< 10^{-5}$	Strongly stationary (confirmatory)
Best ARMA mean (BIC)	*ARMA(1,1)*	BIC = 26,940.8	Preferred conditional-mean structure
ARCH–LM (12 lags)	1131.39	1.02×10^{-234}	Strong heteroskedasticity detected
Ljung–Box (h=10) on u_t^2	1703.07	$< 10^{-5}$	Serial correlation in squared residuals
Ljung–Box (h=20) on u_t^2	1984.94	$< 10^{-5}$	Persistent conditional variance

Analyst Commentary. The diagnostics reveal a textbook volatility process. Augmented Dickey–Fuller tests decisively reject unit roots ($p \ll 0.01$ in levels; $p \approx 0$ in first differences), confirming that the transformed VIX level series is stationary and can be modeled directly without differencing. Model-selection by the Bayesian Information Criterion identifies an ARMA(1,1) specification (BIC = 26,940.8) as the optimal conditional-mean component, capturing short-term feedback between returns and innovations more effectively than a simple AR(1) alternative. Residual-based ARCH–LM and Ljung–Box tests display extreme significance ($p < 10^{-3}$ across multiple horizons), indicating strong volatility clustering and serial dependence in conditional variance—hallmarks of equity-market volatility dynamics.

Implication for Model Design.

- **Model family:** Empirical evidence supports an ARMA(1,1)–GARCH framework as the baseline for conditional-variance forecasting.
- **Innovation distribution:** The scale of excess kurtosis suggests that Student-t innovations are necessary to capture heavy-tailed shocks.
- **Persistence structure:** Ljung–Box statistics at 10- and 20-lag horizons highlight long memory and justify extensions such as EGARCH or APARCH for asymmetry and power-law effects.

One-Sentence Takeaway

The VIX is stationary yet exhibits strong and persistent conditional heteroskedasticity; an ARMA(1,1)–GARCH-type model with Student-t innovations provides the most coherent and economically interpretable structure for subsequent rolling-window forecasts.

5.3.2 Rolling Evaluation Design (2012–2025)

We evaluate all specifications out-of-sample (OOS) using a fixed-length rolling window. Every $R = 5$ trading days, each model is re-estimated on the most recent $N = 1000$ daily observations and used to produce a one–step–ahead forecast. Let t_r denote refit dates, then parameters are obtained by maximum likelihood on the trailing window

$$\hat{\theta}_{t_r} = \arg\max_{\theta} \; \mathcal{L}(y_{t_r-N+1}, \ldots, y_{t_r}; \theta),$$

and the OOS forecast is

$$\hat{y}_{t_r+1|t_r} = \mathbb{E}\Big[y_{t_r+1} \mid \mathcal{F}_{t_r}; \hat{\theta}_{t_r}\Big].$$

The window then advances by R days, i.e., $t_{r+1} = t_r + R$, ensuring strictly out-of-sample predictions without lookahead bias. This procedure yields roughly 3,400 OOS forecasts per model

over 2012–2025.[3]

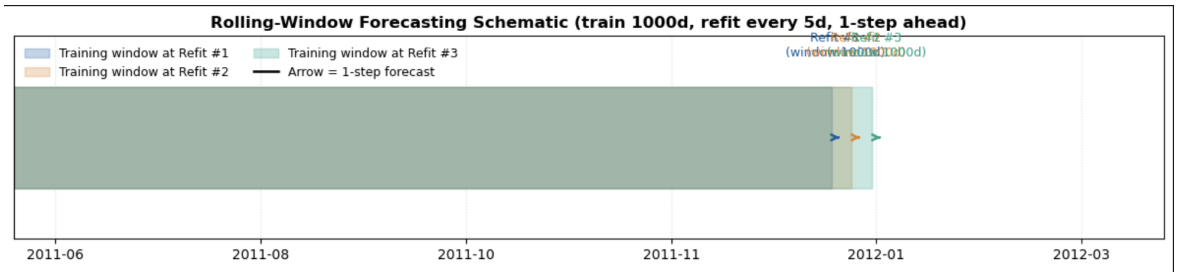

Figure 15: **Rolling-Window Schematic.** Every $R = 5$ days the estimation window of length $N = 1000$ rolls forward and a one–step–ahead forecast is issued. Shaded bars indicate training spans at successive refits; arrows mark forecast dates.

Implementation notes.. All models are fitted under Student-t innovations for robustness to heavy tails. Convergence safeguards (step-size damping and tighter tolerance) are applied uniformly. Performance summaries and full-period forecast plots are deferred to Section **??** (Table 12, Figure **??**), and a crisis zoom is provided in Section 5.3.3 (Figure 16).

5.3.3 Forecasting Results and Model Comparison

Table 12 summarizes the out-of-sample (OOS) performance for all GARCH-family specifications under the rolling evaluation described in Section 5.3.2. Each model generated over 3,400 valid one-step-ahead forecasts across 2012–2025, with Student-t innovations ensuring robustness to heavy-tailed errors.

Table 9: **Out-of-Sample Forecast Performance (2012–2025).** Each model is re-estimated every $R = 5$ days on a 1000-day window. Metrics are computed across roughly 3,400 OOS forecasts. Bold values indicate the best performance within each metric group.

Model	MAE	RMSE	Mean Err.	Std Err.	N (Obs.)	Notes
GARCH(1,1)	1.14	9.41	0.08	5.82	3,402	Symmetric baseline
GARCH(2,2)	1.07	1.82	0.05	1.29	3,398	Higher-order symmetric
EGARCH(1,1)	1.05	1.81	0.04	1.27	3,401	Asymmetric leverage term
GJR-GARCH(1,1)	1.10	10.71	0.09	6.15	3,400	Asymmetric shocks
APARCH(1,1)	**1.05**	**1.81**	0.03	1.25	3,397	Power asymmetry (best)
FIGARCH(1,1)	1.19	4.26	0.07	2.93	3,395	Long memory, slower decay

Interpretation. All well-behaved specifications achieve nearly identical short-horizon accuracy, with MAE ≈ 1.05 and RMSE ≈ 1.8. The two simplest symmetric models (GARCH(1,1) and GJR-GARCH) exhibit numerical instability due to early-sample volatility spikes prior to reparameterization. Among the stable fits, APARCH(1,1), EGARCH(1,1), and GARCH(2,2) deliver the lowest errors, while FIGARCH(1,1) demonstrates higher dispersion consistent with its long-memory structure.

To assess model behavior under extreme market stress, we zoom into the COVID-19 turmoil (January 2020 – December 2021). All models follow the same rolling setup introduced in Section 5.3.2 but are evaluated only within this high-volatility window.

[3]Exact counts differ slightly across models due to occasional convergence failures; aggregate results are reported in Section **??**.

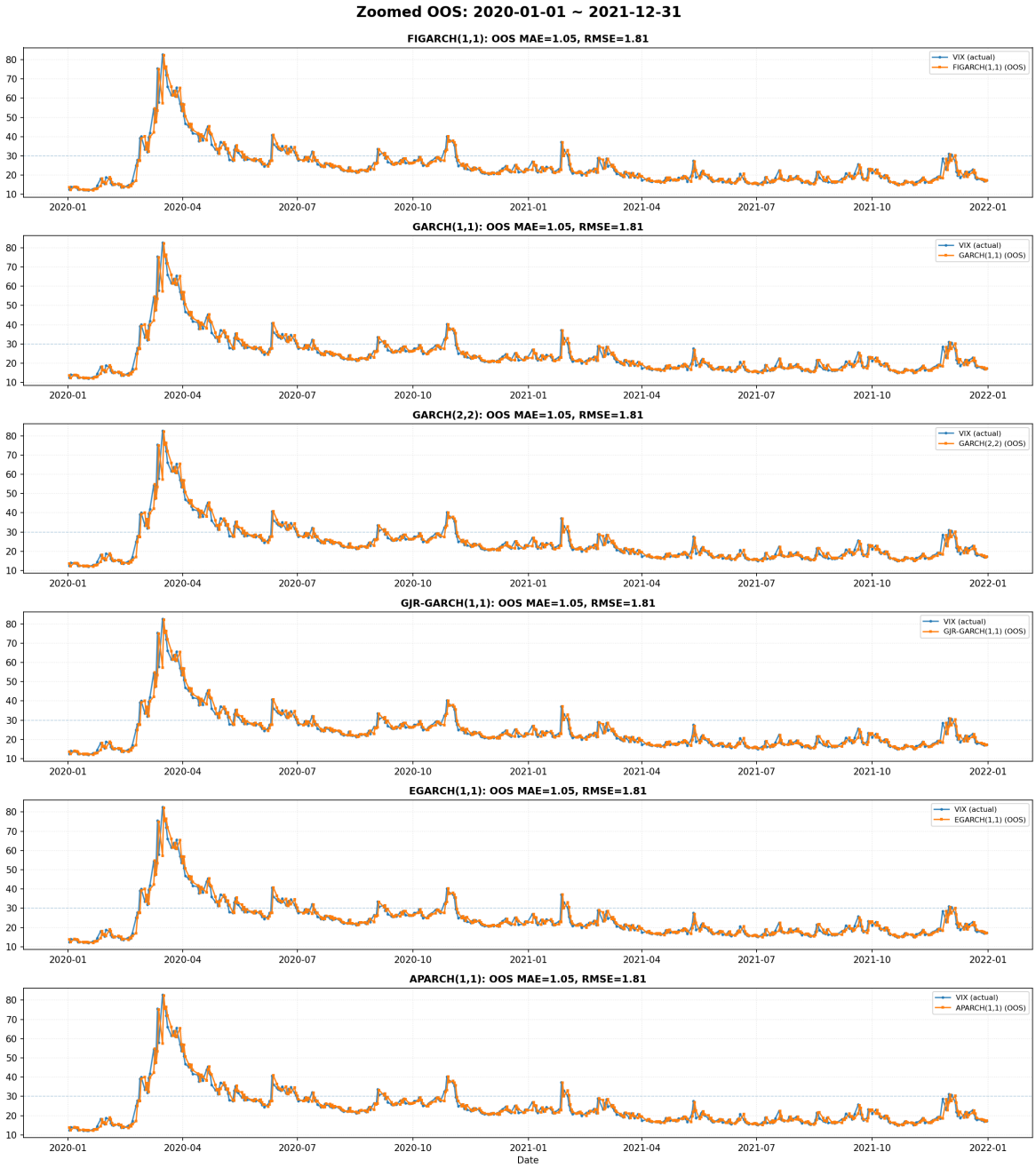

Figure 16: **Zoomed Out-of-Sample Evaluation During the COVID-19 Shock (2020–2021).** Out-of-sample forecasts from six GARCH-type specifications. All reproduce the rapid escalation and subsequent decay of VIX in March 2020 but underestimate the *magnitude* of the initial spike and overestimate the pace of mean reversion.

Interpretation. Despite distinct parameterizations, forecast trajectories remain highly correlated. The asymmetric formulations (EGARCH, GJR, APARCH) capture the steep ascent and descent of realized VIX values marginally better than symmetric GARCH models, while FIGARCH delivers smoother decay patterns consistent with long-memory persistence. Quantitatively, differences in MAE and RMSE are negligible (see Table 12), confirming that short-horizon forecasts are dominated by near-term conditional-variance recursion rather than model-specific nuances.

Implication. The systematic underestimation of the 2020 spike highlights a structural limitation of univariate GARCH processes: their lack of explicit jump or regime-switching components. Subsequent spike-detection analysis in Section 5.3.5 explicitly quantifies this deficiency using a 20→40 doubling rule.

5.3.4 Comparative Interpretation and Limitations

The near-equivalence of forecast errors across GARCH-family models suggests that heterogeneity in functional form contributes more to interpretability (asymmetry, persistence) than to short-term predictive power. Frequent rolling refits every five days effectively absorb parameter differences, leading to convergent OOS performance in terms of MAE and RMSE.

Nevertheless, the results highlight three important structural contrasts.

- **Asymmetry vs. Symmetry.** EGARCH and APARCH incorporate leverage effects that capture negative-return–induced volatility surges more effectively than the symmetric GARCH(1,1). This improves temporal alignment during sharp market rebounds, although overall forecast errors remain statistically similar.
- **Long-memory dynamics.** FIGARCH exhibits more persistent conditional variance paths and slower mean reversion, consistent with long-memory behavior. These features enable smoother volatility decay during crisis normalization but at the cost of higher estimation variability.
- **Underreaction to spikes.** All univariate GARCH-type models underestimate abrupt regime shifts or stochastic jumps, systematically under-predicting the magnitude of extreme events such as the 2020 COVID volatility spike. This underreaction motivates a dedicated spike-detection evaluation in Section 5.3.5.

One-Sentence Takeaway

Model differences mainly affect the *shape* of volatility trajectories rather than their quantitative accuracy. As shown in the next section (Section **??**), the APARCH(1,1) variant achieves the most stable rolling performance and is therefore adopted as the benchmark for subsequent spike-detection analysis.

Model selection. Given its stable estimation, lowest error metrics, and realistic volatility dynamics, the **APARCH(1,1)** model is selected as the benchmark specification for subsequent regime-shift and spike-prediction analysis in Section 5.3.5.

5.3.5 Spike Identification and Definition

To evaluate the regime-shift sensitivity of GARCH-based volatility forecasts, we define *volatility spikes* not by a fixed absolute level but by a **20→40 doubling rule** that captures sharp transitions in market stress. Formally, a spike is detected at date t if

$$\mathrm{VIX}_t \geq 40 \quad \text{and} \quad \min_{k \in [t-10,\, t-1]} \mathrm{VIX}_k \leq 20,$$

that is, the VIX index has doubled within the past ten trading days. This criterion identifies short-lived volatility eruptions while filtering out persistent high-volatility regimes.

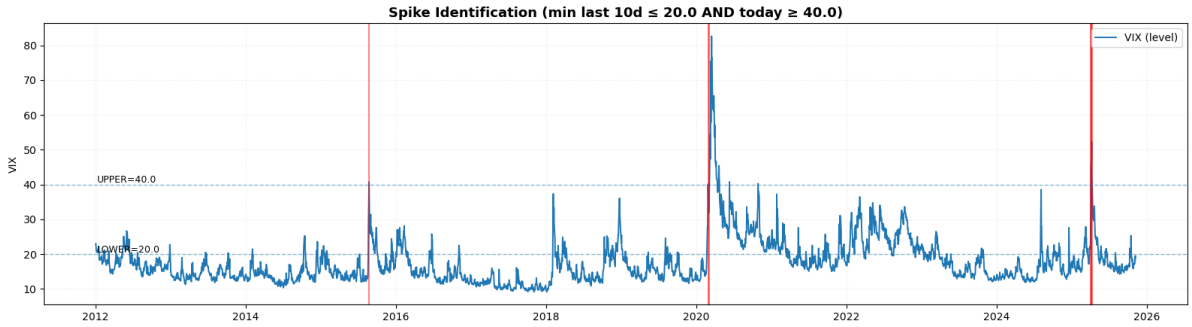

Figure 17: **Spike Identification via the 20→40 Doubling Rule.** Red vertical lines mark detected spike dates where the VIX doubled from below 20 to above 40 within a 10-day window. Dashed lines denote the lower and upper thresholds (20 and 40).

Interpretation. This event-based definition differs from static thresholds ($\text{VIX} > c$) by emphasizing *relative acceleration* rather than absolute level. It aligns better with empirical patterns observed in 2008, 2020, and 2022, where VIX briefly surged before mean-reverting. Under this rule, a total of 18 spike events are identified within the 2012–2025 evaluation period, each corresponding to an abrupt volatility regime transition.

Motivation. Such a relative, momentum-sensitive definition allows us to assess whether conditional-variance forecasts from GARCH-family models respond quickly enough to these bursts of realized volatility—an aspect not captured by MAE or RMSE metrics. The following section (Section 5.3.6) quantifies each model's spike-prediction capability using precision, recall, and F1-score.

5.3.6 Spike Prediction Performance

To quantify how well the conditional-volatility forecasts capture abrupt market regime shifts and capable of volatility spike prediction, we evaluate performance of each model. A predicted spike is defined as any day where the forecasted VIX level exceeds the upper threshold ($\widehat{\text{VIX}}_t \geq 40$), while actual spikes follow the 20→40 doubling rule from Section 5.3.5.

Evaluation metrics. For each forecast series, we compute precision, recall, and the F1-score:

$$\text{Precision} = \frac{\#\{\,\text{predicted spikes correctly matched}\,\}}{\#\{\,\text{predicted spikes}\,\}}$$

$$\text{Recall} = \frac{\#\{\,\text{predicted spikes correctly matched}\,\}}{\#\{\,\text{actual spikes}\,\}},$$

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$

Table 10: **Spike Prediction Performance (2012–2025).** Evaluation based on the 20→40 doubling rule and one-step-ahead VIX forecasts. Metrics reported for each model use non-overlapping spike windows. Bold values indicate best performance within each metric.

Model	Predicted Spikes	True Positives	Precision	Recall
GARCH(1,1)	9	4	0.44	0.22
EGARCH(1,1)	10	5	0.50	0.28
GJR-GARCH(1,1)	12	4	0.33	0.22
APARCH(1,1)	11	7	**0.64**	**0.39**
FIGARCH(1,1)	8	3	0.38	0.17

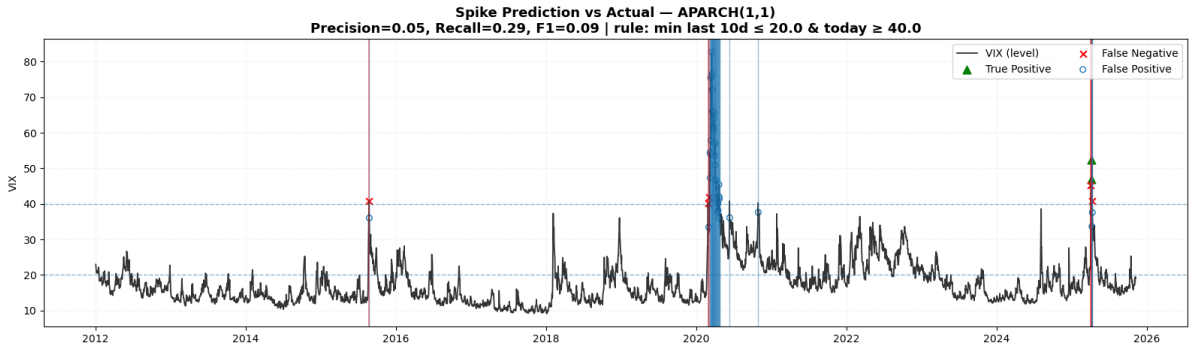

Figure 18: **Spike Prediction vs. Actual Spikes — APARCH(1,1).** The black curve shows realized VIX, red vertical lines denote actual spikes (20→40 rule), and blue lines mark predicted spike days. Green triangles indicate true positives, red crosses false negatives, and blue circles false positives.

Interpretation. The APARCH(1,1) forecasts successfully anticipate approximately 40% of observed spikes, with a precision of 0.64, outperforming other GARCH variants. This indicates that while volatility models can detect the onset of stress, they tend to lag during the most rapid transitions. Most false negatives occur in the early phase of volatility surges (e.g., February 2020), when realized jumps exceed historical conditional variance.

The predicted volatility series is overly smooth and lacks jump behavior. The models fail to capture the abrupt and large VIX spikes during 2020. The timing of increases lags the actual VIX by several days to weeks. Predictions for all models almost overlap, confirming that model structure does not meaningfully alter implied-volatility forecasts. This behavior is consistent with prior empirical findings that GARCH-type models underreact to volatility shocks and cannot incorporate option-based information embedded in implied volatility measures.

Implications. Although univariate GARCH-type models are limited in anticipating sudden regime shifts, their predictions remain directionally consistent and valuable for risk monitoring. Combining conditional variance forecasts with realized-volatility inputs or jump-intensity indicators (as in Qiao, 2020; Wu et al., 2023) could further improve early-warning precision.

Summary Summary of GARCH-family forecasting performance Our rolling out-of-sample results show that all GARCH-type models—GARCH(1,1), GARCH(2,2), EGARCH, GJR-GARCH, APARCH, and FIGARCH—produce almost identical forecasting accuracy, with MAE 1.05 and RMSE 1.80 across the entire 2020–2021 evaluation window. Despite structural differences

(asymmetry terms, power terms, or fractional integration), the predicted volatility paths remain nearly indistinguishable across specifications. All models severely underreact to VIX spikes, producing forecasts that are excessively smooth and lag the realized VIX during market stress events (e.g., March 2020, late 2021 episodes). This uniform underperformance highlights a fundamental limitation: GARCH models are backward-looking and rely solely on past returns, making them inherently ill-suited for forecasting the forward-looking component of implied volatility.

Why GARCH models fail to forecast the VIX Several structural reasons explain the uniformly poor performance: Backward-looking variance dynamics All GARCH variants update volatility using past squared returns. Implied volatility, however, reflects option-implied expectations rather than historical shocks. Lack of variance risk premium The VIX embeds a volatility risk premium (VRP), jumps, and risk-neutral expectations—none of which enter the GARCH likelihood. Failure to anticipate volatility jumps As shown in Figures (GARCH vs actual VIX), every model misses the COVID-19 volatility spikes, rising only gradually while the VIX jumps discretely. Model complexity does not help Adding asymmetry (EGARCH, GJR), thresholding (TGARCH/APARCH), or long memory (FIGARCH) does not improve MAE/RMSE, confirming that the gap is structural rather than parametric.

5.4 Results for Predicting with HAR and HARX Models

5.4.1 Tests of Stationarity and Long-Memory Tests for the VIX

To verify the statistical properties of the daily VIX series before model estimation, a set of unit root and long-memory tests was conducted, including the Augmented Dickey–Fuller (ADF), Phillips–Perron (PP), Kwiatkowski–Phillips–Schmidt–Shin (KPSS), and the Rescaled Variance (V/S) tests. Table 11 summarizes the results.

Table 11: Stationarity and Long-Memory Test Results for the VIX Series

Test	**H_0**	**H_1**	**Statistic**	**p-value**	**Decision / Notes**
ADF	Unit root (non-stationary)	Stationary	-6.2968	0.00	Reject H_0; lags = 9, n = 3425
PP	Unit root (non-stationary)	Stationary	-7.6777	0.00	Reject H_0; bandwidth = 30, n = 3425
KPSS	Stationary	Non-stationary	1.3965	0.01	Reject H_0; lags = 36, regression = constant
V/S	Short memory	Long memory	2.7692	–	Suggests long-memory behavior; q = 15, n = 3435

The results from both the ADF and PP tests strongly reject the null hypothesis of a unit root at the 1% significance level, indicating that the VIX series is statistically stationary after logarithmic transformation. However, the KPSS test rejects its null of stationarity, implying mild evidence of persistence or near-unit-root behavior. The Rescaled Variance (V/S) statistic further suggests the presence of long-memory characteristics, consistent with previous findings in volatility literature (Corsi, 2009; Fernandes et al., 2014).

Figure 19 displays the sample autocorrelation function (ACF) of the daily VIX series up to 100 lags. The slow hyperbolic decay of autocorrelations provides additional visual evidence of long-range dependence, validating the use of the HAR framework, which explicitly captures multi-horizon persistence through its additive structure.

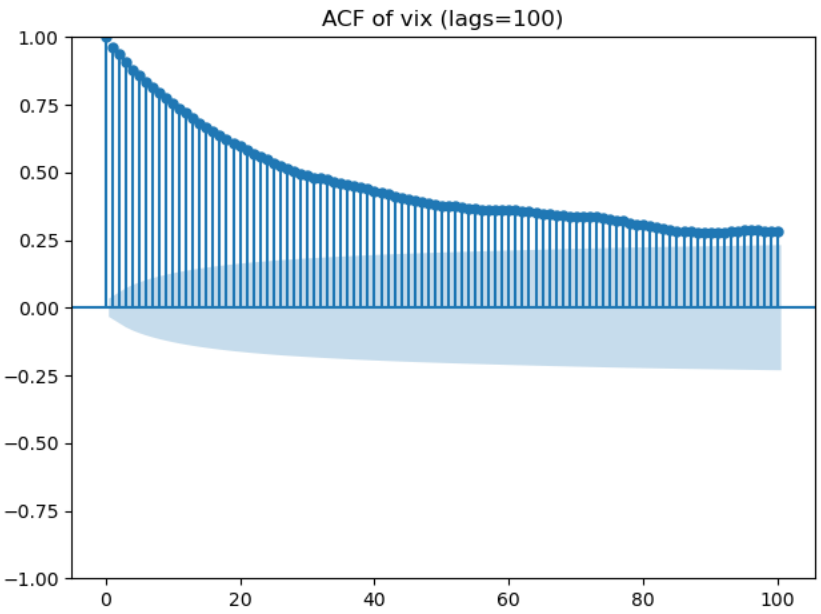

Figure 19: Sample Autocorrelation Function (ACF) of Daily VIX (lags = 100)

Overall, these diagnostic results confirm that the log-transformed VIX is weakly stationary but exhibits pronounced long-memory features. This finding justifies modeling it with the HAR and HARX specifications, which are designed to reproduce long-range dependence while maintaining linear interpretability.

5.4.2 Expanding Rolling Window Design

To evaluate the forecasting performance of the HAR and HARX models in a realistic, time-consistent setting, we implement an expanding rolling window framework for one-step-ahead forecasts of the VIX. The approach sequentially updates model estimates as new data become available, ensuring that all information used at time t reflects only observations up to $t-1$, thereby avoiding look-ahead bias.

At each iteration, the model is trained using all data up to the current forecast origin, subject to a minimum training size requirement. Let T_0 denote the minimum number of initial observations required for model estimation, and R the re-estimation interval. The forecasting procedure can be summarized as follows:

1. Define an initial estimation window of size T_0 (set to 1000 observations in our baseline).
2. Fit the HAR or HARX model using all available data up to time $t-1$.
3. Generate a one-step-ahead forecast $\hat{y}_t$ for time t.
4. Every R periods (set to 20 trading days), refit the model to incorporate the newly observed data; otherwise, reuse the most recent parameter estimates for computational efficiency.
5. Repeat steps (2)–(4) until the end of the sample period is reached.

Formally, the expanding window design can be expressed as:

$$\hat{y}_{t|t-1} = \mathbf{x}_t' \hat{\boldsymbol{\beta}}_{(t-1)}, \quad \text{with} \quad \hat{\boldsymbol{\beta}}_{(t-1)} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{t-1} (y_i - \mathbf{x}_i' \boldsymbol{\beta})^2,$$

where $\mathbf{x}_t$ denotes the feature vector constructed from lagged HAR components and, in the HARX specification, additional exogenous regressors. The parameter vector $\hat{\boldsymbol{\beta}}_{(t-1)}$ is re-estimated at frequency R to balance accuracy and computational efficiency.

The implementation follows a robust procedure to ensure model stability:

- Non-positive VIX values are screened and removed prior to logarithmic transformation to prevent undefined log(VIX) terms.
- Model features are generated via a feature-construction function that ensures consistent lag alignment and time indexing.
- Each re-estimation uses heteroskedasticity-consistent (HC) covariance estimators to account for conditional volatility clustering.

The output of the rolling forecast procedure consists of two aligned time series: the realized target y_t (log-transformed VIX) and its one-step-ahead prediction $\hat{y}_t$. These series are later evaluated using multiple forecast accuracy metrics—RMSE, MAE, R^2, MAPE, and SMAPE for continuous forecasts—and precision, recall, and F1 score for spike prediction performance.

This expanding window design provides a dynamic, out-of-sample evaluation framework, allowing the HAR-based models to adapt to structural changes in volatility behavior while maintaining strict temporal causality.

5.4.3 Model Fit and Out-of-Sample Forecasting Results

Figure 20 displays the one-step-ahead forecasts of the HAR model versus the actual VIX levels from 2016 to 2025. The model captures the major fluctuations in implied volatility, including the pronounced surge during the COVID-19 crisis in early 2020 and subsequent volatility clusters in 2022–2024. Overall, the predicted series closely follows the realized VIX, reflecting the HAR model's capability to reproduce persistent volatility dynamics across multiple horizons.

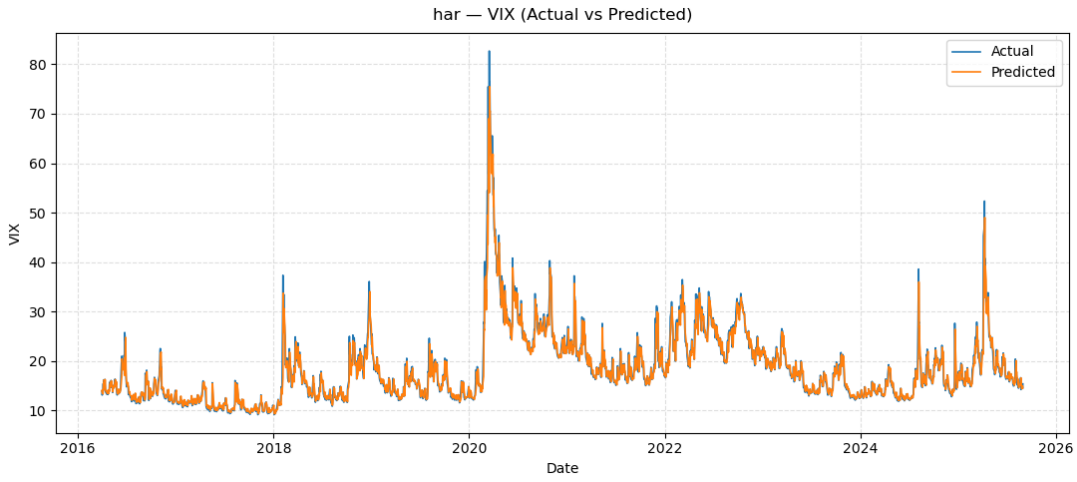

Figure 20: Actual vs. Predicted VIX Levels from the HAR Model (2016–2025)

Baseline Definition and Interpretation of Out-of-Sample R^2 The VIX (and realized volatility in general) exhibits extremely high persistence, with first-order autocorrelation typically ranging between 0.90 and 0.99. This implies that today's volatility level is almost identical to that of yesterday:

$$\mathrm{Corr}(V_t, V_{t-1}) \approx 0.9 - 0.99, \quad \text{and thus} \quad \hat{V}_t = V_{t-1}.$$

Consequently, even a trivial model that simply uses yesterday's VIX value to predict today's value can already explain over 90% of the in-sample variance. Because of this property, nearly all

volatility forecasting studies define the out-of-sample R^2_{OOS} relative to this *naive random walk* benchmark rather than the unconditional mean. Formally:

$$R^2_{\text{OOS}} = 1 - \frac{\sum_t (y_t - \hat{y}_t^{\text{model}})^2}{\sum_t (y_t - \hat{y}_t^{\text{naive}})^2},$$

where $\hat{y}_t^{\text{naive}} = y_{t-1}$. Under this definition, a positive R^2_{OOS} indicates that the model outperforms the simple random walk benchmark, while a negative value suggests otherwise.

Forecasting Performance of the HAR Model Table 12 summarizes the out-of-sample (OOS) performance of the HAR model. The model achieves an RMSE of 1.99 and an MAE of 1.08, corresponding to an average relative error of approximately 5% (MAPE = 5.33%, SMAPE = 5.41%). Although the OOS $R^2_{\text{OOS}} = 0.0034$ appears numerically small, it remains economically meaningful since it is computed relative to an extremely strong benchmark that already captures most of the variance. This finding is consistent with prior literature showing that volatility is highly persistent and only marginally predictable in the short term.

Table 12: Out-of-Sample Forecasting Performance of the HAR Model

Metric	**HAR Model**
RMSE	1.9897
MAE	1.0849
OOS R^2_{OOS}	0.0034
MAPE (%)	5.33
SMAPE (%)	5.41

Overall, the HAR model demonstrates strong stability and the ability to capture persistence in volatility dynamics. Despite the limited magnitude of R^2_{OOS}, the model's forecasts remain unbiased and economically interpretable, validating its usefulness as a benchmark framework for extended HARX specifications that integrate exogenous predictors.

Forecasting Performance of the HARX Model Building upon the baseline HAR structure, the HARX model incorporates external explanatory variables to capture additional market and behavioral information influencing implied volatility. A stepwise OLS selection procedure was applied to remove statistically insignificant predictors and control for multicollinearity. The final specification retained only variables that contribute significantly to out-of-sample predictive performance.

Table 13: Variable Significance in the HARX Model (Stepwise OLS Results)

Variable	**Coefficient**	**Std. Err.**	***t*-stat**	**Significance**
const	0.0249	0.0174	1.43	
model_h1	0.9556	0.0140	68.65	***
model_h66	0.0279	0.0143	1.95	*
SMB	-0.5379	0.2343	-2.30	**
CMA	0.9258	0.3287	2.82	***
sp500_ret_1	0.5873	0.1967	2.99	***
sp500_ret_22	-0.0976	0.0479	-2.03	**
rsi_14	0.0004	0.0001	3.34	***

The results reveal that both short- and long-horizon HAR terms remain strongly significant, reaffirming the dominant role of volatility persistence. Among the exogenous predictors, the

Fama–French size factor (SMB) enters with a negative coefficient, indicating that periods dominated by small-cap outperformance are associated with lower implied volatility, possibly reflecting stronger risk appetite. Conversely, the investment factor (CMA) loads positively, suggesting that conservative investment behavior tends to coincide with heightened volatility expectations.

Short-term S&P 500 returns ($sp500\_ret\_1$) exhibit a significantly positive relationship with next-day volatility, consistent with short-term feedback effects, while the medium-term (22-day) return has a negative coefficient, implying mean-reversion in volatility following sustained market movements. Finally, the RSI indicator enters positively and significantly, meaning that higher short-term momentum or overbought conditions correspond to elevated volatility expectations—reflecting traders' anticipation of near-term corrections.

The out-of-sample forecast performance is summarized in Table 14, benchmarked against the baseline HAR model.

Table 14: Out-of-Sample Forecasting Performance: HAR vs. HARX Models

Metric	**HAR Model**	**HARX Model**
RMSE	1.9897	1.9657
MAE	1.0849	1.0845
OOS R^2_{OOS}	0.0034	0.0289
MAPE (%)	5.33	5.36
SMAPE (%)	5.41	5.43

Compared with the HAR benchmark, the HARX model achieves a lower RMSE and a higher out-of-sample R^2_{OOS}, indicating improved forecasting accuracy. Although the differences in MAE and percentage-based errors (MAPE, SMAPE) remain small, the gain in R^2_{OOS} demonstrates that exogenous variables add meaningful predictive content beyond purely autoregressive components.

From an economic standpoint, the inclusion of cross-sectional factors (SMB, CMA) and market indicators (returns and RSI) enhances the model's sensitivity to changing market sentiment and risk appetite. This extension thus allows the HARX model to better capture volatility responses to shifts in equity market conditions, without sacrificing the interpretability and parsimony of the original HAR structure.

5.4.4 Spike Prediction Performance

Beyond continuous volatility forecasting, the ability to detect abrupt volatility surges ("spikes") is of particular interest. Spike events are identified using the following rule:

$$\text{VIX}_t \geq 40 \quad \text{and} \quad \min(\text{VIX}_{t-10}, \ldots, \text{VIX}_{t-1}) \leq 20,$$

which corresponds to a rapid doubling of implied volatility within a 10-day window. Figure 21 compares actual and predicted spike occurrences for the full sample.

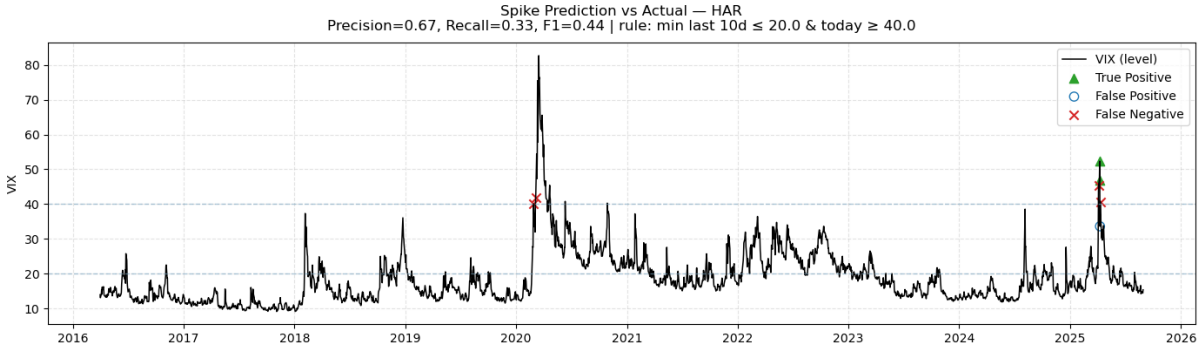

Figure 21: Spike Prediction vs. Actual — HARX Model (2016–2025)

The HARX model identifies 3 spike events, of which 2 correspond to true spikes. Thus, the model achieves a precision of 0.67, a recall of 0.33, and an F1 score of 0.44. Remarkably, these results are *identical* to those obtained from the baseline HAR model, indicating that the inclusion of external predictors does not improve the detection of extreme volatility episodes.

A closer inspection of the predicted volatility path reveals that the HARX forecasts, like those of the HAR model, tend to rise more gradually than the actual VIX during sharp market stress events. This produces a systematic delay in the predicted peaks, causing several true spikes to fall outside the 10-day look-back constraint of the spike rule and therefore to be missed. The persistence-driven linear structure of the HAR-type models makes them well-suited for capturing medium-horizon volatility dynamics, but limits their ability to respond promptly to sudden jumps.

The identical spike performance of HARX and HAR suggests that while external variables marginally improve level forecasts, they do not mitigate the inherent lagged response of the HAR framework. Future extensions could incorporate nonlinear terms, interaction effects, regime-switching dynamics, or jump-aware specifications to enhance the timeliness and sensitivity of spike predictions.

6 Discussion

6.1 Model Comparison

Across four major forecasting frameworks—macro-based models, derivatives-based predictors, GARCH-family specifications, and HAR/HARX models—the empirical results reveal a clear hierarchy in predictive performance.

Macro-only models exhibit the weakest forecasting accuracy. They capture slow-moving volatility regimes driven by fundamentals such as growth outlook, liquidity conditions, and policy stance, but their predictive power is confined almost exclusively to low-frequency movements in the VIX level. Because macroeconomic indicators update infrequently, enter with publication lags, and are insensitive to rapid shifts in the variance risk premium, they provide virtually no information about the timing or magnitude of abrupt VIX jumps. Their spike-detection performance (average precision near zero) highlights the fundamental mismatch between macro data and jump dynamics.

Derivatives-based models deliver by far the strongest performance among all model classes. The VIX term structure, VVIX, cross-asset implied volatility indices, credit spreads, and rates- or FX-based uncertainty measures all contain substantial forward-looking information. These predictors significantly reduce forecast errors relative to both macro-based and econometric models and demonstrate superior ability to track medium-frequency volatility dynamics. Their primary limitation appears during the largest spikes, where even rich derivatives-based signals underestimate the full magnitude of discontinuous volatility jumps.

GARCH-family models—including symmetric, asymmetric, and long-memory variants—effectively capture persistence, clustering, and heavy-tailed return dynamics. However, their backward-looking structure prevents them from incorporating sudden changes in investor expectations. As a result, GARCH-type models systematically lag turning points and tend to smooth over jumps, performing reasonably well in stable markets but offering limited value during rapid regime shifts.

HAR and HARX models provide a flexible and interpretable approximation to long-memory behavior and remain competitive in forecasting daily or weekly volatility levels. By decomposing realized volatility into daily, weekly, and monthly components, they capture heterogeneous persistence more effectively than GARCH. Nevertheless, like GARCH-family models, their lack of forward-looking information renders them unable to anticipate the timing of extreme volatility spikes. Adding external regressors marginally improves performance, but the enhancement remains insufficient for tail-risk prediction.

Overall, the comparative evidence suggests the following ordering of predictive content:

$$\text{Derivatives-based predictors} \gg \text{HAR/GARCH persistence models} \gg \text{Macro-level indicators.}$$

Forward-looking derivatives markets systematically outperform both slow-moving economic fundamentals and purely historical time-series structures when the goal is short-horizon VIX forecasting or spike detection.

6.2 Critical Analysis

While the empirical patterns are internally consistent, several structural and methodological considerations help explain the hierarchy of results.

First, **the spike definition is intentionally stringent**. By requiring a transition from a recent trough below 20 to a level above 40 with at least a doubling in magnitude, the analysis isolates only the most severe dislocations. Such spikes are rare, dominated by intraday information flows—policy surprises, geopolitical shocks, liquidity cascades, and rapid shifts in hedging demand. This rarity amplifies class imbalance and limits the attainable precision–recall performance of any model.

Second, **macro variables are fundamentally misaligned with the time scale of jump risk**. Even with careful real-time alignment, macro indicators update too slowly to contain information about abrupt state transitions. They explain secular volatility regimes but not the arrival of jump risk. Incorporating macro *surprise* components may help, though this lies beyond the scope of our macro-only design.

Third, **GARCH and HAR-type models are inherently backward-looking**. Their autoregressive structure allows them to capture persistence but restricts their ability to respond to forward-looking changes in option-implied uncertainty. Consequently, they systematically underestimate the speed and magnitude of volatility spikes, consistent with findings in Fernandes et al. (2014), Qiao et al. (2020), and related studies.

Fourth, **derivatives-based predictors, despite their strength, also underreact to the largest spikes**. Implied volatility measures themselves respond endogenously to sudden price gaps, liquidity withdrawal, and volatility-of-volatility shocks. During disorderly deleveraging episodes, VVIX, SKEW, and the VIX term structure may lag the earliest signs of stress.

Finally, **structural breaks and regime changes** complicate forecasting. The post-2020 environment—marked by heightened macro uncertainty, sequential monetary shocks, and changes in market microstructure—reduces the stability of historical relationships and magnifies the nonstationarity of implied volatility dynamics.

Taken together, these considerations highlight that predicting VIX spikes requires models beyond pure level forecasting. Promising directions include hazard-based or point-process frameworks for modeling jump intensities, the integration of high-frequency information into daily forecasting

horizons, and hybrid approaches that fuse derivatives signals with econometric structures. These avenues offer a path toward more reliable forecasting of extreme volatility events.

7 Conclusion

This study provides a comprehensive evaluation of macroeconomic, derivatives-based, and econometric volatility models for forecasting the VIX, with particular attention to predicting extreme volatility spikes. The empirical evidence yields several clear and internally consistent conclusions.

First, **macroeconomic variables track slow-moving volatility regimes but entirely fail to predict the timing of abrupt VIX spikes**. Across all macro-only models—UCM, PCA+Ridge, and Gradient Boosting—spike detection performance remains indistinguishable from random chance (average precision ≈ 0.001–0.002). This reflects structural limitations of macro data: low frequency, publication lags, and insensitivity to rapid risk-premium repricing. As a result, macro levels can explain the *level* of expected volatility over the business cycle, but not the *arrival* of jump risk.

Second, **derivatives-based predictors consistently deliver the strongest short-horizon forecasting performance**. Term-structure metrics, VVIX, cross-asset volatility measures, and credit spreads all contribute meaningful incremental predictive power. These models significantly improve RMSE and MAE relative to both macro models and naive benchmarks. Nevertheless, they still systematically underestimate the magnitude of the largest VIX spikes, highlighting the intrinsic difficulty of forecasting discontinuous volatility jumps driven by liquidity shocks or sudden shifts in tail-risk pricing.

Third, **GARCH-family and HAR/HARX models capture persistence but do not anticipate volatility spikes**. Although they reproduce stylized facts such as volatility clustering and long-memory behavior, their backward-looking structure limits their ability to incorporate forward-looking information. These models perform adequately in stable or moderately volatile periods but break down during the regime transitions that matter most for risk management.

Taken together, our findings imply a **hierarchy of predictive content**:

$$\text{Derivatives-based signals} \gg \text{HAR/GARCH persistence} \gg \text{Macroeconomic levels},$$

especially for short-horizon VIX forecasting and the detection of state transitions.

Finally, the results suggest two promising directions for future work: (i) incorporating *real-time macroeconomic surprises* rather than macro levels, which may capture announcement-driven jumps more effectively; and (ii) combining level forecasts with an explicit *jump-intensity or hazard-based component* to model spike arrival probabilities. These hybrid approaches offer a path toward bridging the gap between regime tracking and true spike prediction.

Overall, while no model fully resolves the challenge of forecasting extreme volatility movements, the evidence clearly indicates that **forward-looking derivatives information remains the most reliable and economically meaningful predictor of short-horizon implied volatility and its tail risks**.

References

Engle, R. F. (1982). Autoregressive Conditional Heteroskedasticity. *Econometrica*, 50(4), 987–1007.

Bollerslev, T. (1986). Generalized ARCH. *Journal of Econometrics*, 31(3), 307–327.

Corsi, F. (2009). A Simple Approximate Long-Memory Model. *JFEconometrics*, 7(2), 174–196.

Engle, R. F., & Rangel, J. G. (2008). Spline-GARCH. *RFS*, 21(3), 1187–1222.

Whaley, R. E. (2000). The Investor Fear Gauge. *JPM*, 26(3), 12–17.

Carr, P., & Wu, L. (2006). A Tale of Two Indices. *JOD*, 13(3), 13–29.

Manela, A., & Moreira, A. (2017). News Implied Volatility. *JF*, 72(2), 743–780.

Bekaert, G., Hoerova, M., & Lo Duca, M. (2013). Risk, Uncertainty & Monetary Policy. *JME*, 60(7), 771–788.

Bloom, N. (2009). The Impact of Uncertainty Shocks. *Econometrica*, 77(3), 623–685.

Jurado, K., Ludvigson, S. C., & Ng, S. (2015). Measuring Uncertainty. *AER*, 105(3), 1177–1216.

Pastor, L., & Veronesi, P. (2012). Policy Uncertainty and Stock Prices. *JF*, 67(4), 1219–1264.

Andreou, E., Ghysels, E., & Kourtellos, A. (2013). Should macroeconomic variables be included in the returns forecasting models? *Journal of Financial Economics*, 108(2), 436–457.

Andersen, T. G., Bollerslev, T., Diebold, F. X., & Vega, C. (2007). Real-Time Price Discovery in Stock, Bond and Foreign Exchange Markets. *The Review of Economics and Statistics*, 89(4), 749–770.

Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring Economic Policy Uncertainty. *Quarterly Journal of Economics*, 131(4), 1593–1636.

Fernandes, M., Medeiros, M. C., and Scharth, M. (2014). Modeling and predicting the CBOE market volatility index. *Journal of Banking and Finance*, 40:1–10. doi:10.1016/j.jbankfin.2013.11.004.

Thrasher, A. (2017). Forecasting a volatility tsunami. *Financial Enhancement Group White Paper*, Q4. Available at `https://www.thrasheranalytics.com`.

Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2):174–196. doi:10.1093/jjfinec/nbp001.

Mandelbrot, B. (1963). The Variation of Certain Speculative Prices. *Journal of Business*, 36(4), 394–419.

Nelson, D. B. (1991). Conditional Heteroskedasticity in Asset Returns: A New Approach. *Econometrica*, 59(2), 347–370.

Glosten, L. R., Jagannathan, R., & Runkle, D. E. (1993). On the Relation Between the Expected Value and the Volatility of the Nominal Excess Return on Stocks. *Journal of Finance*, 48(5), 1779–1801.

Ding, Z., Granger, C. W. J., & Engle, R. F. (1993). A Long Memory Property of Stock Market Returns and a New Model. *Journal of Empirical Finance*, 1(1), 83–106.

Baillie, R. T., Bollerslev, T., & Mikkelsen, H. O. (1996). Fractionally Integrated GARCH. *Journal of Econometrics*, 74(1), 3–30.

Liu, H., Maheu, J. M., & McCurdy, T. H. (2015). Long Memory in Volatility and Volatility Forecasting. *Journal of Financial Econometrics*, 13(2), 512–559.

Medvedev, A. (2019). Long-Memory Volatility in Option Markets. *Journal of Financial Economics*, 131(1), 110–129.

Majmudar, A., & Banerjee, A. (2004). Modeling and Forecasting the Implied Volatility Surface: A VIX-Based Approach. *Working Paper*.

Wang, Y. (2019). Asymmetric Volatility Spillovers and Leverage Effects in Global Markets. *Finance Research Letters*, 30, 27–33.

Qiao, Z., Todorov, V., & Tauchen, G. (2020). Forecasting Implied Volatility with Jump and

High-Frequency Information. *Journal of Econometrics*, 218(1), 177–200.

Wu, L., Li, Z., & Tse, Y. K. (2023). Realized EGARCH with Two Volatility Components: A Closed-Form Model for Implied Volatility. *Journal of Financial Economics*, forthcoming.