Aalto University
School of Science

Master’s Programme in Systems and operations research

Assessing Volatility

Transparent, Reproducible VIX Forecasts from Public Data

Mika Viirret

Master’s Thesis
2025

Aalto University
School of Science

Author Mika Viirret

Title Assessing Volatility — Transparent, Reproducible VIX Forecasts from Public Data

Degree programme Systems and operations research

Major Systems and operations research

Supervisor Prof. Pauliina Ilmonen

Advisor Prof. Pauliina Ilmonen

Date 18 December 2025 **Number of pages** 54 **Language** English

Abstract

This thesis examines one-day-ahead forecasting of the CBOE Volatility Index (VIX) using public daily data from 2020–2024. The forecast target is the next-day VIX level. Predictors include daily/weekly/monthly VIX components (HAR-style), realized volatility computed from S&P 500 returns, the one-day change in VIX, and a high-volatility indicator.

We compare classical econometric models ARIMA, GARCH and the HAR with modern machine-learning methods (Random Forest, XGBoost, and a shallow neural network). Models are trained on 2020–2023 and evaluated out-of-sample in 2024 using a rolling-origin design that avoids look-ahead bias. Forecast accuracy is summarized by RMSE, MAE, MAPE, R^2, and directional accuracy. We apply DM tests to assess pairwise differences in predictive accuracy. To evaluate economic value, forecasts are translated into stylized volatility-timing rules using both discrete sign and continuous sizing variants.

Across rolling windows, HAR and tree-based methods deliver the strongest level accuracy, while a returns-only GARCH proxy tends to underpredict VIX, consistent with a positive volatility risk premium. DM tests indicate statistically significant improvements over weaker baselines. The best statistical models also yield positive Sharpe ratios in the timing exercise. The contribution is a transparent, reproducible VIX-forecasting pipeline based solely on public data, together with an integrated statistical-and-economic evaluation that clarifies the distinction between risk-neutral and realized volatility measures.

Keywords VIX, implied volatility, volatility forecasting, HAR, ARIMA, GARCH, random forest, XGBoost, neural network, DM test, Sharpe ratio

Aalto-yliopisto
Perustieteiden
korkeakoulu

Tekijä Mika Viirret

Työn nimi Volatiliteetin mittaaminen — Läpinäkyvät ja toistettavat VIX-ennusteet julkisesta datasta

Koulutusohjelma Systeemi- ja Operaatiotutkimus

Pääaine Mathematics and Operations research

Työn valvoja Prof. Pauliina Ilmonen

Työn ohjaaja Prof. Pauliina Ilmonen

Päivämäärä 18 joulukuuta 2025 **Sivumäärä** 54 **Kieli** englanti

Tiivistelmä

Tässä työssä tutkitaan CBOE:n volatiliteetti-indeksin (VIX) yhden päivän eteenpäin tapahtuvaa ennustamista käyttäen vain julkista päivädataa vuosilta 2020–2024. Ennustekohteena on seuraavan pörssipäivän VIX-taso. Selittäjinä käytetään VIXin päivittäisiä, viikoittaisia ja kuukausittaisia komponentteja (HAR-tyyli), S&P 500 -indeksin tuotoista laskettua realisoitunutta volatiliteettia, VIXin yhden päivän muutosta sekä yksinkertaista korkean volatiliteetin indikaattoria.

Menetelmävertailussa ovat klassiset ekonometriset mallit ARIMA, GARCH ja HAR sekä modernit koneoppimismenetelmät (satunnaismetsä, XGBoost ja ohut neuroverkko). Mallit opetetaan vuosilla 2020–2023 ja testataan vuodelta 2024 käyttäen rullaavaa testi-ikkunaa, jolla pyritään estämään ennakointiharhaa. Ennustetarkkuutta mitataan RMSE-, MAE-, MAPE- ja R^2-mittareilla sekä suunnan osumatarkkuudella. Mallien välisiä eroja testataan Diebold–Mariano -testeillä. Taloudellista arvoa arvioidaan muuntamalla ennusteet tyylitellyiksi volatiliteetin ajoitussäännöiksi, joissa hyödynnetään sekä diskreettiä suuntamerkkiä että jatkuvaa positiokokoa.

Tulosten mukaan HAR ja puupohjaiset menetelmät tuottavat parhaan ennusteen, kun taas tuottoihin perustuva GARCH-vertailu aliarvioi VIX-tasoa, mikä on yhdenmukaista positiivisen volatiliteettipreemion kanssa. Diebold–Mariano -testit osoittavat tilastollisesti merkitseviä parannuksia heikompiin malleihin verrattuna. Parhaat tilastolliset mallit tuottavat myös positiivisen Sharpe-suhteen ajoituskokeessa. Työn panos on läpinäkyvä ja toistettava VIX-ennustamisen tulos pelkällä julkisella datalla yhdistettynä tilastolliseen ja taloudelliseen arviointiin, mikä selkeyttää riskineutraalin ja realisoituneen volatiliteetin eroa käytännössä.

Avainsanat VIX, implisiittinen volatiliteetti, volatiliteetin ennustaminen, HAR, ARIMA, GARCH, satunnaismetsä, XGBoost, neuroverkko, Diebold–Mariano, Sharpe-suhde

Contents

Symbols and abbreviations

Symbols

P_t	S&P 500 close level at day t
VIX_t	VIX close at day t
$r_t = \ln(P_t) - \ln(P_{t-1})$	Daily log-return of the S&P 500
$RV_t^{(n)}$	n-day realised volatility (annualised), from returns
$\mathrm{VIX}_t^{(d)},\ \mathrm{VIX}_t^{(w)},\ \mathrm{VIX}_t^{(m)}$	Daily, 5-day (weekly) and 22-day (monthly) VIX components
$\Delta\mathrm{VIX}_t = \mathrm{VIX}_t - \mathrm{VIX}_{t-1}$	One-day change in VIX
$\widehat{\mathrm{VIX}}_{t+1}$	One-step-ahead forecast for VIX (made at t)
$y_t,\ \widehat{y}_t,\ e_t$	Actual, forecast, and error ($e_t = y_t - \widehat{y}_t$)
$C(S, K, \tau, \sigma),\ P(S, K, \tau, \sigma)$	Black–Scholes call/put price
S_t	Underlying index level (generic)
K	Option strike price
τ	Time to maturity in years
$r,\ q$	Continuously compounded risk-free rate and dividend yield
σ	Volatility parameter
$d_1,\ d_2$	Black–Scholes terms used in C and P
$\Phi(\cdot),\ \phi(\cdot)$	Standard normal CDF and PDF
$F,\ K_0,\ \Delta K_i,\ T$	VIX formula terms: forward level, pivot strike, strike interval, horizon
θ	Trading threshold for sign-based timing rule
SR	Annualised Sharpe ratio of strategy P&L

Operators

$\mathbb{E}[\cdot]$	Expectation
$\mathrm{Var}(\cdot),\ \mathrm{Cov}(\cdot,\cdot)$	Variance and covariance
$\nVdash\{\cdot\}$	Indicator function (1 if condition holds, else 0)
$\mathrm{sign}(\cdot)$	Sign function $(-1, 0, +1)$
$\Delta x_t = x_t - x_{t-1}$	First difference (daily change)
$Lx_t = x_{t-1}$	Lag operator
$\bar{x}_t^{(n)} = \frac{1}{n}\sum_{i=0}^{n-1} x_{t-i}$	n-day rolling mean at t
$\arg\min_\theta\ \mathcal{L}(\theta)$	Argument that minimizes a loss $\mathcal{L}$

Abbreviations

VIX	CBOE Volatility Index (model-free 30-day implied volatility of S&P 500)
SPX	S&P 500 index
RV	Realised volatility (from historical returns)
IV	Implied volatility
ARIMA	Autoregressive Integrated Moving Average
GARCH	Generalized Autoregressive Conditional Heteroskedasticity
HAR	Heterogeneous Autoregression (daily/weekly/monthly components)
RF	Random Forest
XGB	Extreme Gradient Boosting (XGBoost)
NN	Neural Network (MLP in this thesis)
OOS	Out-of-sample
OLS	Ordinary Least Squares
DM	Diebold–Mariano (test of equal predictive accuracy)
HAC	Heteroskedasticity- and Autocorrelation-Consistent (e.g. Newey–West)
FFT	Fast Fourier Transform
PSD	Power Spectral Density (Welch method)
P&L	Profit and Loss
CV	Cross-Validation
RMSE	Root Mean Squared Error
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
R^2	Coefficient of determination

1 Introduction

Volatility, the magnitude of fluctuations in asset prices, is central to option pricing, risk management, and portfolio construction. In the Black–Scholes framework the value of a European option depends sensitively on the volatility input (Black and Scholes, 1973; Hull, 2014). Under the model's assumptions one obtains closed-form prices, yet empirical evidence is at odds with constant volatility: option markets display smiles and skews across strikes and asset returns exhibit fat tails and volatility clustering (Mandelbrot, 1963; Fama, 1965; Cont, 2001). These facts have motivated volatility models that vary over time or state, including local and stochastic volatility and jump diffusions (Dupire, 1994; Heston, 1993; Cont and Tankov, 2004), as well as data-driven forecasting approaches.

A practical challenge that unites these strands is that future volatility is not observed. Machine-learning methods provide flexible, nonparametric mappings from predictors to a volatility target and can capture nonlinear interactions that classical linear models may miss (Gu, Kelly and Xiu, 2020). Neural networks are universal function approximators (Hornik, Stinchcombe and White, 1989) and large-scale evidence suggests that they can extract persistent structure from financial time series (Sirignano and Cont, 2019). The objective in this thesis is not to replace pricing models, but to evaluate whether machine learning can deliver more accurate and practically useful volatility forecasts relative to interpretable econometric baselines.

A central design decision is the choice of forecast target. One possibility is realized volatility computed from future S&P 500 returns, for example a forward 21 to 30 day measure (Andersen et al., 2001). In this thesis the primary target is the VIX, which aggregates option implied expectations of 30 day variance for the S&P 500 in a model-free way (Whaley, 2009). This choice is motivated by implementability, reproducibility, and conceptual alignment with option pricing. VIX is directly tradable via futures and options, which allows a clean assessment of economic value from forecasts. VIX is publicly available at daily frequency, so results are replicable without proprietary option chain data or high frequency microstructure choices. VIX is also the market's risk-neutral volatility proxy, which is the quantity that matters for Black–Scholes style valuation. VIX and subsequent realized volatility are not identical. VIX typically exceeds realized volatility because option prices embed a positive volatility risk premium (Blair, Poon and Taylor, 2001; Bondarenko, 2014). For reproducibility, all software versions and Python libraries are documented in Appendix A.

The empirical work uses daily data from 1 January 2020 to 31 December 2024. The dependent variable is the next day VIX level, denoted VIX_{t+1}. Predictors include lagged and averaged VIX components inspired by the HAR framework, realized volatility proxies from S&P 500 returns, and simple dynamics such as the daily change in VIX (Corsi, 2009). All forecasts are strictly one step ahead. At time t the models form $\widehat{\mathrm{VIX}}_{t+1}$ using only information available at time t.

The modelling suite combines classical econometric baselines and modern machine-learning methods. The baselines include ARIMA, GARCH, and HAR. The machine learning models include RF, XGBoost, and a shallow NN. Models are trained on 2020 to 2023 and evaluated out-of-sample in 2024 using root mean squared error, mean absolute error, the coefficient of determination, directional accuracy, and pairwise DM tests for predictive accuracy (Engle, 1982; Bollerslev, 1986; Diebold and Mariano, 1995). To assess economic value we implement simple timing rules that take positions based on the predicted change in VIX and we report annualized Sharpe ratios under transparent assumptions that ignore trading frictions.

This thesis makes three contributions. First, it provides a transparent and reproducible comparison of econometric and machine learning models for one day ahead VIX forecasting over 2020 to 2024 using public data and out-of-sample validation. Second, it quantifies both statistical and economic value through standard accuracy metrics, DM tests, and simple trading rules, and it highlights when tree based machine learning can materially outperform linear baselines. Third, it discusses how the relationship between VIX and forward realized volatility, via the volatility risk premium, affects the interpretation of the forecasting results.

The research questions are as follows. Do machine-learning methods improve one step ahead VIX forecasts relative to ARIMA, GARCH, and HAR baselines? Do such improvements translate into economically meaningful gains in VIX timing strategies? How does the relationship between VIX and forward realized volatility, through the volatility risk premium, affect the interpretation of the results? The remainder of the thesis is structured as follows. Section 2 reviews the option pricing background and volatility concepts. Section 3 describes data, features, models, and validation design. Section 4 presents out-of-sample results and the economic evaluation. The software environment and libraries used are listed in Appendix A.

Large language models were used in a limited manner to support literature discovery and language polishing. Specifically, OpenAI's GPT models assisted in searching for references asked by the author, translating paragraphs originally drafted in Finnish and helped with LaTeX–syntax. All modeling decisions, code, data handling, and empirical results are the author's own. AI outputs were verified and edited for accuracy by author (OpenAI, 2025).

2 Theoretical framework

2.1 Overview of options and derivatives

Derivatives are financial contracts whose value depends on the price of an underlying asset or index, such as equities, commodities, interest rates, or currencies. Common derivative types include forwards, futures, and swaps, which have linear payoffs, and options, which have nonlinear payoffs. A European call option gives the holder the right but not the obligation to buy the underlying at a specified strike price K at expiration T, while a European put option gives the right to sell at the strike. The payoff of a call is $\max(S_T - K, 0)$, and for a put $\max(K - S_T, 0)$, so option payoffs depend nonlinearly on the underlying price S_T. Because of this nonlinearity and leverage, pricing options requires careful modeling of the underlying's uncertainty (Hull, 2014).

The price of an option at time t depends on several factors: the current underlying level S_t, the strike K, the time to maturity $\tau = T - t$, the volatility of the underlying, the risk-free interest rate r, and the continuous dividend yield q (if any). These inputs are central to option valuation. In particular, understanding how volatility enters option pricing is crucial. That is why we derive the classical option pricing formula in the following subsection.

2.2 Symbols and conventions

We work under continuous time and continuous compounding. The risk-free rate r and dividend yield q are assumed to be continuously compounded rates (possibly time-dependent). An overdot (e.g. $\dot{S}$) or time derivative denotes differentiation with respect to t. Other symbols used include:

- S_t: underlying index (e.g. S&P 500) level at time t.
- K: strike price of an option.
- $\tau = T - t$: time remaining to option maturity (in years).
- r_t: risk-free interest rate at time t.
- q_t: continuous dividend yield at time t.
- $\Phi(x)$, $\phi(x)$: the standard normal CDF and PDF.
- σ: instantaneous volatility parameter.

- ΔK_i: strike interval used in the VIX calculation.

2.3 Dividend-adjusted Black–Scholes and the PDE

Under the Black–Scholes assumptions, under the risk-neutral measure the underlying price satisfies the stochastic differential equation

$$dS_t = (r - q)S_t\,dt + \sigma S_t\,dW_t,$$

where r is the continuously compounded risk-free rate, q is the continuous dividend yield, and σ is the (constant) volatility. By applying Itô's lemma and no-arbitrage (delta-hedging) arguments, one shows that the price $V(S, t)$ of a derivative must satisfy the Black–Scholes partial differential equation:

$$\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + (r - q)S\frac{\partial V}{\partial S} - rV = 0.$$

This PDE is solved backward from the terminal payoff at expiry T. For example, for a European call option with payoff $V(S, T) = \max(S_T - K, 0)$, the closed-form solution is given by the Black–Scholes formula (Black and Scholes, 1973; Hull, 2014):

$$C(S, K, \tau, \sigma) = Se^{-q\tau}\,\Phi(d_1) - Ke^{-r\tau}\,\Phi(d_2), \tag{1}$$

$$P(S, K, \tau, \sigma) = Ke^{-r\tau}\,\Phi(-d_2) - Se^{-q\tau}\,\Phi(-d_1), \tag{2}$$

where

$$d_1 = \frac{\ln(S/K) + (r - q + \frac{1}{2}\sigma^2)\tau}{\sigma\sqrt{\tau}}, \qquad d_2 = d_1 - \sigma\sqrt{\tau},$$

and $\tau = T - t$. Here $\Phi(\cdot)$ is the standard normal cumulative distribution. Put–call parity holds: $C - P = Se^{-q\tau} - Ke^{-r\tau}$. From these formulas one can derive the option *Greeks* (sensitivities), which we omit for brevity.

2.4 Limitations and model extensions

The Black–Scholes model's assumptions of constant volatility, continuous trading, and log-normal returns imply a flat implied-volatility surface across strikes and maturities. In contrast, empirical market data show a pronounced skew/smile in the implied-volatility surface (Cont, 2001; Hull, 2014), and return distributions exhibit fat tails and volatility clustering (Mandelbrot, 1963; Fama, 1965). To address these discrepancies, various extensions have been proposed:

- Local volatility models: Allow volatility to be a deterministic function $\sigma_{\text{loc}}(S, t)$ calibrated to match today's implied-volatility surface exactly. Dupire (1994) showed how to compute $\sigma_{\text{loc}}(K, T)$ from market prices of calls $C(K, T)$. Local-vol models fit current prices perfectly but assume future volatility is fully determined by the current state.

- Stochastic volatility models: Volatility itself follows a random process. For example, the Heston model (Heston, 1993) assumes the variance has its own mean-reverting dynamics. Stochastic-volatility models capture volatility clustering and produce more realistic dynamics of the implied surface. They require estimation of additional parameters and often lack closed-form solutions.

- Jump-diffusion models: Incorporate occasional jumps in the underlying price. Merton's jump-diffusion or more general models (Cont and Tankov, 2004) add a jump component to capture fat tails and extreme events. These models better fit option prices across strikes but add complexity and computational challenges.

These extensions can improve the fit to option market data, but at the cost of additional calibration complexity and potential overfitting.

2.5 Implied volatility, realized volatility and the surface

Market option prices are often expressed in terms of implied volatility. For a given observed option price C^{mkt}, the implied volatility σ^{IV} is defined by solving $C(S, K, \tau, \sigma^{IV}) = C^{\text{mkt}}$ in the Black–Scholes formula. The mapping $(K, \tau) \mapsto \sigma^{IV}(K, \tau)$ defines the implied-volatility surface. Equity index options typically exhibit a downward-sloping skew: out-of-the-money puts (low strikes) have higher implied volatilities than at-the-money options. In Dupire's framework, a local volatility function can be extracted from a smooth call price surface $C(K, T)$. Dupire's formula is

$$\sigma^2_{\text{loc}}(K, T) = \frac{\frac{\partial C}{\partial T} + rK\,\frac{\partial C}{\partial K}}{\frac{1}{2}K^2\,\frac{\partial^2 C}{\partial K^2}},$$

which shows how the curvature of option prices in strike and time determines the local variance.

Realized volatility is a backward-looking measure computed from historical returns. Given daily log-returns r_i of the S&P 500, an n-day realized volatility is often estimated as

$$\widehat{\sigma}_{\mathrm{RV}} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(r_i - \bar{r})^2},$$

which is essentially the sample standard deviation (annualized by multiplying by $\sqrt{252/n}$ if desired) (Andersen et al., 2001). Empirically, realized volatility clusters over time: large returns tend to be followed by large returns (of either sign), and similarly for small returns. This persistence is why econometric models like GARCH (Bollerslev, 1986) are used to capture volatility clustering in return series.

2.6 The VIX as model-free implied variance

The VIX is a widely-followed measure of the S&P 500's expected 30-day volatility, derived from option prices. It is constructed to represent the risk-neutral expectation of variance over the next month (Whaley, 2009). In practice, VIX is computed as 100 times the square root of a weighted sum of out-of-the-money option prices across strikes. The CBOE formula is:

$$\mathrm{VIX}_t = 100 \times \sqrt{\frac{2}{T}\sum_i \frac{\Delta K_i}{K_i^2} e^{rT} Q(K_i) \; - \; \frac{1}{T}\left(\frac{F}{K_0} - 1\right)^2},$$

Where the sum is over strike prices K_i, ΔK_i is the spacing between strikes, $Q(K_i)$ are the mid-quote prices of out-of-the-money options (calls for $K_i > K_0$ and puts for $K_i < K_0$), F is the forward S&P 500 level, K_0 is the first strike below F, and T is the time to the 30-day horizon via interpolation (CBOE, 2019). Because VIX is derived under the risk-neutral measure, it typically exceeds the actual realized volatility that follows as figure 4 shows. This difference is known as the volatility risk premium (Blair, Poon and Taylor, 2001; Bondarenko, 2014) embedded in option prices.

2.7 Volatility forecasting approaches

Forecasts of volatility (or VIX) can be obtained from several sources:

- Econometric time-series models: These include ARIMA models on volatility or returns to capture autoregressive patterns, and GARCH models to capture volatility clustering (Engle, 1982; Bollerslev, 1986). For example, an ARIMA(p, d, q) can model VIX dynamics directly, while GARCH(1,1) models the conditional variance of returns.

- Implied-volatility measures: One may use market-based forecasts directly. For instance, today's VIX is often used as a predictor for tomorrow's volatility, since it summarizes current option market expectations (Hull, 2014; Poon and Granger, 2003).

- Hybrid/long-memory models: The HAR model (Corsi, 2009) regresses future volatility on its recent daily, weekly, and monthly averages, capturing persistent long-range dependence.

- machine-learning methods: Techniques like RF, XGB, and NN can flexibly model nonlinear interactions among many predictors (past volatilities, returns, etc.). Such methods have been applied to asset pricing and may uncover complex patterns (Gu, Kelly and Xiu, 2020).

In this thesis, we implement representatives of each category (as described in 3.4) and compare their one-step-ahead VIX forecasts on the same data. The short-term dynamics and potential predictability of implied volatility have also been studied by Konstantinidi et al. (2008), who report limited but statistically significant forecasting power in the evolution of implied volatility surfaces.

3 Data and methodology

3.1 Scope and sample selection

We do not have access to detailed option-chain data, so we forecast a market-wide VIX instead. Our primary target is the daily close of the VIX. To complement this, we use the S&P 500 index to derive historical return-based volatility measures. We select the five-year period January 1, 2020 to December 31, 2024 for analysis. This window contains both low-volatility periods and high-volatility events (for example, the market crash of March 2020), allowing a robust evaluation of forecasting methods. All data are aligned to trading days: if the S&P market is closed, that day is omitted. If VIX is missing while S&P is open (rare), we forward-fill the last VIX value. In practice, data gaps are minimal.

3.2 Data sources and basic transformations

- VIX index (VIX_t): Daily closing values from January 2020 through December 2024. Obtained from the CBOE and cross-checked with Yahoo Finance (`^VIX` historical data) (Yahoo Finance, 2025). This is our forecast target series.

- S&P 500 index (P_t): Daily closing values GSPC from Yahoo Finance (Yahoo Finance, 2025). We compute log-returns $r_t = \ln(P_t) - \ln(P_{t-1})$ from this series.

- Realized volatility (RV): From the daily returns r_t, we construct backward-looking volatility measures. For example, the n-day realized volatility is defined as

$$RV_t^{(n)} = \sqrt{\frac{252}{n}\sum_{i=0}^{n-1} r_{t-i}^2},$$

  where 252 is the trading-day annualization factor (Andersen et al., 2001). In this study we use $RV_t^{(5)}$ and $RV_t^{(30)}$ to capture short-term and medium-term historical volatility.

- Data alignment: All series are aligned by date. If a trading day is missing in one series, it is removed from all series (or VIX is forward-filled if missing). This ensures that the feature and target vectors for forecasting line up in time.

3.3 Feature engineering

Volatility exhibits persistence over multiple horizons (Corsi, 2009). We therefore construct lagged volatility features analogous to the HAR model components:

$$\begin{aligned}
\mathrm{VIX}_t^{(d)} &= \mathrm{VIX}_t, \\
\mathrm{VIX}_t^{(w)} &= \frac{1}{5}\sum_{i=0}^{4} \mathrm{VIX}_{t-i}, \\
\mathrm{VIX}_t^{(m)} &= \frac{1}{22}\sum_{i=0}^{21} \mathrm{VIX}_{t-i}.
\end{aligned}$$

These represent the current VIX level, the 5-day (weekly) average, and the 22-day (monthly) average. We also include the one-day change $\Delta\mathrm{VIX}_t = \mathrm{VIX}_t - \mathrm{VIX}_{t-1}$ to capture short-term momentum. From realized volatility, we add $RV_t^{(5)}$ and $RV_t^{(30)}$ as features, capturing recent market volatility measured from returns. Finally, we include simple regime indicators such as $\nVdash\{\mathrm{VIX}_t > 30\}$ to flag high-volatility states. All numeric features are standardized (zero mean, unit variance) when used in machine learning models. This ensures stable training and comparability of coefficients (but does not affect the interpretation of linear models).

3.4 Models

At date t, let $\mathcal{X}_t$ denote all features available (Section 3.3) and let the target be $y_{t+1} \equiv \mathrm{VIX}_{t+1}$. All models produce a one–step–ahead forecast $\widehat{y}_{t+1} = \widehat{\mathrm{VIX}}_{t+1}$ using only information up to t. Econometric baselines are specified mathematically, while ML models include a short pseudocode box.

Statistical baselines

ARIMA on VIX: ARIMA provides a transparent linear benchmark for short-run dependence in VIX (Box–Jenkins tradition). Let z_t be either VIX_t or $\log(\mathrm{VIX}_t)$ (optional variance stabilization). After d differences, $w_t = \Delta^d z_t$, the ARMA form is

$$\phi(L)\, w_t \;=\; \theta(L)\, \varepsilon_t, \qquad \phi(L) = 1 - \phi_1 L - \cdots - \phi_p L^p, \quad \theta(L) = 1 + \theta_1 L + \cdots + \theta_q L^q,$$

with white-noise innovations ε_t and lag operator $Lx_t = x_{t-1}$. Orders (p, d, q) are chosen by BIC on the training window, subject to stationarity. One-step forecasts use the standard ARMA recursion and, if modeling $\log(\mathrm{VIX})$, are exponential back to levels with optional bias correction (Box et al., 2015).

```
Algorithm ARIMA$(p, d, q)$ (estimate & one-step forecast)
Input: univariate series $\{y_t\}_{t=1}^{T}$ (here $y_t = \mathrm{VIX}_t$); candidate grids $\mathcal{P}, \mathcal{D}, \mathcal{Q}$.
Model selection (training window)
1. for $(p, d, q) \in \mathcal{P} \times \mathcal{D} \times \mathcal{Q}$ do
    1.1 Difference $d$ times to obtain $y_t^{(d)}$.
    1.2 Fit ARMA$(p, q)$ on $y_t^{(d)}$ by MLE; record BIC.
end for
2. Choose $(\hat{p}, \hat{d}, \hat{q})$ with the lowest BIC and refit by MLE on the full training window.
One-step forecast at time t
3. Update the ARIMA state with $y_t$ (innovation filter).
4. Compute $\widehat{y}_{t+1|t}$ from the fitted ARMA$(\hat{p}, \hat{q})$ representation of $y^{(\hat{d})}$.
Rolling origin (if used): re-estimate at scheduled refit dates; otherwise only update the state each day.
```

GARCH(1,1) on returns (variance proxy):To capture volatility clustering, we estimate a GARCH model on SPX log-returns and use the one-step conditional variance as a proxy for next-day risk under the physical measure (Engle, 1982; Bollerslev, 1986; Nelson, 1991):

$$
\begin{aligned}
r_t &= \mu + \varepsilon_t, \\
\varepsilon_t &= \sigma_t z_t, \qquad z_t \sim t_{\nu}^{(0,1)}, \ \ \nu > 2, \\
\sigma_t^2 &= \omega + \alpha\, \varepsilon_{t-1}^2 + \beta\, \sigma_{t-1}^2, \qquad \omega > 0, \ \alpha, \beta \geq 0, \ \alpha + \beta < 1.
\end{aligned}
$$

We annualize the one-day-ahead variance forecast via 252 trading days and convert to an annualized standard deviation in percent as

$$x_{t+1} \ \equiv \ 100 \sqrt{252\, \widehat{\sigma}_{t+1|t}^2}.$$

Because VIX is a risk-neutral 30-day expectation, while $\widehat{\sigma}_{t+1|t}^2$ is a one-day conditional variance under the physical measure, this proxy can systematically underpredict VIX in the presence of a volatility risk premium (Blair, Poon and Taylor, 2001; Bondarenko, 2014; Whaley, 2009). To map the proxy into a VIX level, we fit an affine calibration on the training set by OLS:

$$\mathrm{VIX}_t \ = \ a + b\, x_t + u_t,$$

and then predict

$$\widehat{\text{VIX}}_{t+1} \;=\; a + b\, x_{t+1}.$$

If a through-the-origin mapping is desired, set a=0 and estimate a single scale b by OLS.

```
**Algorithm GARCH(1,1) variance proxy (estimate & one-step VIX level)**
*Input:* SPX log-returns $r_t$, $t = 1, \ldots, T$; training set indices $\mathcal{T}_{\text{train}}$.
**Specification**
$r_t = \mu + \varepsilon_t,\;\; \varepsilon_t = \sigma_t z_t,\;\; z_t \overset{\text{iid}}{\sim} t_\nu^{(0,1)};\;\; \sigma_t^2 = \omega + \alpha\, \varepsilon_{t-1}^2 + \beta\, \sigma_{t-1}^2.$
**Estimation (MLE with constraints)**
1. Initialize $\sigma_0^2$ (e.g., sample variance) and choose starting $\theta = (\mu, \omega, \alpha, \beta, \nu)$.
2. Maximize the Student-$t$ log-likelihood over $\mathcal{T}_{\text{train}}$ subject to $\omega > 0$, $\alpha, \beta \geq 0$, $\alpha + \beta < 1$, $\nu > 2$.
3. With $\hat{\theta}$, filter $\widehat{\sigma}_t^2$ recursively over the training window.
**Map to a VIX-level proxy**
4. For each training date $t$: $x_t \leftarrow 100\sqrt{252\, \widehat{\sigma}_{t|t-1}^2}$.
5. Fit an affine calibration by OLS on the training set:  $\text{VIX}_t = a + b\, x_t + u_t$; store $(\hat{a}, \hat{b})$.
*One-step forecast at time t*
6. Compute $\widehat{\sigma}_{t+1|t}^2 = \hat{\omega} + \hat{\alpha}\, \hat{\varepsilon}_t^2 + \hat{\beta}\, \widehat{\sigma}_t^2$ with $\hat{\varepsilon}_t = r_t - \hat{\mu}$.
7. Set $x_{t+1} \leftarrow 100\sqrt{252\, \widehat{\sigma}_{t+1|t}^2}$ and output  $\widehat{\text{VIX}}_{t+1} = \hat{a} + \hat{b}\, x_{t+1}$.
*Rolling origin (if used):* re-estimate $\hat{\theta}$ and $(\hat{a}, \hat{b})$ at refit points. Otherwise update recursively.
```

HAR on VIX: To approximate long memory with interpretable components (Corsi, 2009), we regress next-day VIX on daily/weekly/monthly VIX aggregates:

$$\text{VIX}_t^{(d)} = \text{VIX}_t, \quad \text{VIX}_t^{(w)} = \frac{1}{5}\sum_{i=0}^{4} \text{VIX}_{t-i}, \quad \text{VIX}_t^{(m)} = \frac{1}{22}\sum_{i=0}^{21} \text{VIX}_{t-i},$$

and estimate by OLS

$$\text{VIX}_{t+1} = a_0 + a_1\, \text{VIX}_t^{(d)} + a_2\, \text{VIX}_t^{(w)} + a_3\, \text{VIX}_t^{(m)} + \varepsilon_{t+1}.$$

HAC standard errors are optional. The forecasting formula is the fitted linear combination.

```
Algorithm HAR (estimate & one-step forecast)
Input: $y_t = \mathrm{VIX}_t$, $t = 1, \dots, T$.
Feature construction (requires $t \geq 22$)
1. For each $t$: compute $\mathrm{VIX}_t^{(d)} = \mathrm{VIX}_t$, $\mathrm{VIX}_t^{(w)} = \frac{1}{5}\sum_{i=0}^{4} \mathrm{VIX}_{t-i}$, $\mathrm{VIX}_t^{(m)} = \frac{1}{22}\sum_{i=0}^{21} \mathrm{VIX}_{t-i}$.
Estimation (OLS)
2. Regress $y_{t+1}$ on $(1,\ \mathrm{VIX}_t^{(d)},\ \mathrm{VIX}_t^{(w)},\ \mathrm{VIX}_t^{(m)})$ over the training window. Store $\hat{a}_0, \hat{a}_1, \hat{a}_2, \hat{a}_3$.
One-step forecast at time t
3. Form predictors at $t$ and compute $\widehat{y}_{t+1|t} = \hat{a}_0 + \hat{a}_1\,\mathrm{VIX}_t^{(d)} + \hat{a}_2\,\mathrm{VIX}_t^{(w)} + \hat{a}_3\,\mathrm{VIX}_t^{(m)}$.
Rolling origin (if used): recompute $(\hat{a}_j)$ on each expanding/rolling window.
```

Machine-learning models

Preprocessing (common to all ML models). Features are standardized on the training window. Time-series (rolling-origin) validation selects hyperparameters and guards against look-ahead. Inputs may include $\{\mathrm{VIX}_t^{(d)}, \mathrm{VIX}_t^{(w)}, \mathrm{VIX}_t^{(m)}, \Delta\mathrm{VIX}_t, RV_t^{(5)}, RV_t^{(30)}\}$.

RF: An ensemble of decorrelated CART trees reduces variance and captures threshold interactions with minimal tuning (Breiman, 2001). With T trees $\{h_b\}_{b=1}^{T}$, prediction is the mean $\widehat{y}_{t+1} = \frac{1}{T}\sum_{b=1}^{T} h_b(\mathcal{X}_t)$. Leaf limits control complexity.

Algorithm: RF

```
Inputs: training data $\{(\mathcal{X}_i, y_i)\}_{i=1}^{N}$, number of trees $T$, maximum depth $D$, max_features $m$.
for $t = 1, \ldots, T$ do
    1. Draw a bootstrap sample of size $N$ with replacement.
    2. Train a CART regression tree $h_t$ on the bootstrap:
    At each node, sample $m$ features uniformly at random;
    choose the split that minimizes mean squared error (MSE).
    Stop if depth = $D$ or leaf size below threshold.
end for
Prediction: For a new $\mathcal{X}$, output $\widehat{y} = \frac{1}{T} \sum_{t=1}^{T} h_t(\mathcal{X})$.
Notes: No shuffling across time in CV (expanding windows). Set random seed for replicability.
```

Gradient Boosting / XGBoost: Stage-wise additive trees fit pseudo-residuals with shrinkage, depth constraints, and regularization, well-suited to weak nonlinear signals (Friedman, 2001; Chen and Guestrin, 2016). Learning rate $\eta \in (0, 1)$ and early stopping control overfitting.

```
**Algorithm: XGB**
*Inputs:* ${(\mathcal{X}_i, y_i)}_{i=1}^{N}$, rounds $M$, learning rate $\eta$, max depth $d$, regularization $(\lambda, \gamma)$.
Initialize $F_0(\mathcal{X}) = \bar{y}$.
**for** $m = 1, \ldots, M$ **do**
    For each $i$: gradient $g_i = \partial\ell/\partial\hat{y}_i = \hat{y}_i - y_i$, hessian $h_i = \partial^2\ell/\partial\hat{y}_i^2 = 1$.
    Fit a depth-$d$ tree by maximizing split gain using node sums $G = \sum g_i$, $H = \sum h_i$:
    *Leaf weight:* $w^* = -\dfrac{G}{H + \lambda}$.
    *Split gain:* $\text{Gain} = \tfrac{1}{2}\left(\dfrac{G_L^2}{H_L + \lambda} + \dfrac{G_R^2}{H_R + \lambda} - \dfrac{G^2}{H + \lambda}\right) - \gamma$.
    Add the new tree $h_m$ to the ensemble: $F_m(\mathcal{X}) = F_{m-1}(\mathcal{X}) + \eta\, h_m(\mathcal{X})$.
    Validate; apply early stopping (patience $P$).
**end for**
*Prediction:* $\widehat{y} = F_M(\mathcal{X})$.
*Notes:* Column/row subsampling can be used to regularize. With squared error the residuals equal gradients.
```

NN: A single hidden layer MLP with ReLU approximates smooth nonlinear maps. L_2 regularization and early stopping aid generalization in noisy, low-signal data (Hornik, Stinchcombe and White, 1989; Goodfellow, Bengio and Courville, 2016). With H hidden units:

$$h = \sigma(W_1\mathcal{X}_t + b_1), \quad \sigma(u) = \max\{0, u\}, \qquad \widehat{y}_{t+1} = w_2^{\top} h + b_2,$$

trained to minimize $\text{MSE} + \lambda\|W_1\|_2^2 + \lambda\|w_2\|_2^2$.

```
Algorithm: NN (regression)
Inputs: $\{(\mathcal{X}_i, y_i)\}_{i=1}^{N}$, hidden size $H$, optimizer (Adam), batch size $B$, learning rate $\eta$, $L_2$ weight decay $\lambda$.
Preprocess: standardize features (zero mean, unit variance) on training set.
Initialize $W^{(1)}, b^{(1)}, W^{(2)}, b^{(2)}$.
for epochs = $1, 2, \dots$ do
    Split data into mini-batches of size $B$. For each batch:
    Forward: $h = \mathrm{ReLU}(W^{(1)}\mathcal{X} + b^{(1)})$, $\widehat{y} = W^{(2)}h + b^{(2)}$.
    Loss: $\mathcal{L} = \frac{1}{B}\sum(\widehat{y} - y)^2 + \lambda\|W\|_2^2$.
    Backward: compute gradients and update parameters with Adam$(\eta)$.
    Track validation MSE. Stop early if no improvement for $P$ epochs.
end for
Prediction: apply the forward pass to standardized inputs.
Notes: Linear output, MSE loss. Fixed random seed improves replicability.
```

3.5 Estimation design and validation

We split the data into a training sample (2020–2023) and an (out-of-sample) OOS test period (2024). Econometric models (ARIMA, GARCH, HAR) are estimated once on the full training set, and their fitted parameters remain fixed when forecasting the test period. Machine learning models (RF, XGBoost, NN) are tuned using time-series CV (rolling-origin): we repeatedly fit the model on expanding windows of the training data and validate on the next holdout block. This simulates real-time forecasting and prevents look-ahead bias. At each forecast date, only information available up to that date is used (no future data is accessed), ensuring a strict OOS evaluation.

3.6 Evaluation metrics

Forecast accuracy is summarized by standard error metrics. Let y_t be the actual observed value (VIX) and $\widehat{y}_t$ the forecast. We compute where $\bar{y}$ is the mean of y_t in the test period. RMSE penalizes larger errors more strongly, while MAE measures the average absolute error. R^2 indicates the fraction of variance explained (1 is perfect fit, 0 is as good as the sample mean).

$$RMSE = \sqrt{\frac{1}{N} \sum_{t \in \mathcal{T}_{\text{test}}} \left(\widehat{y}_t - y_t\right)^2}, \qquad MAE = \frac{1}{N} \sum_{t \in \mathcal{T}_{\text{test}}} \left|\widehat{y}_t - y_t\right|,$$

$$MAPE = \frac{100}{N} \sum_{t \in \mathcal{T}_{\text{test}}} \left|\frac{\widehat{y}_t - y_t}{y_t}\right|, \qquad R^2 = 1 - \frac{\sum_{t \in \mathcal{T}_{\text{test}}} \left(\widehat{y}_t - y_t\right)^2}{\sum_{t \in \mathcal{T}_{\text{test}}} \left(y_t - \bar{y}\right)^2}.$$

We also report MAPE for interpretability. To assess directional accuracy, we compute

$$\Pr\{\text{sign}(\widehat{\Delta\text{VIX}_t}) = \text{sign}(\Delta\text{VIX}_t)\},$$

the proportion of days where the forecasted change has the same sign as the actual change. For statistical comparison between models, we use the DM test (Diebold and Mariano, 1995) on the forecast errors.

For economic evaluation, we simulate a simple volatility-timing strategy. Each day t, if the model predicts a sufficiently large increase in VIX, we take a long volatility position (e.g. buy VIX futures). If it predicts a decrease, we go short volatility. We then compute the strategy's daily P&L and annualized Sharpe ratio. This idealized strategy (ignoring transaction costs and slippage) provides a proxy for the potential alpha that could be harvested from the forecast signals.

3.7 Data summary

Table 1 summarizes the variables and their sources. Figure 1 plots the daily VIX index over 2020–2024, highlighting periods of extreme volatility. Figure 2 shows an example segment of OOS actual vs. predicted VIX for one model.

Table 1: Variables and sources (01 Jan 2020–31 Dec 2024).

Variable	**Symbol**	**Freq.**	**Source**
VIX close	VIX_t	Daily	CBOE; Yahoo Finance
S&P 500 close	P_t	Daily	Yahoo Finance
Log-return	$r_t = \ln(P_t) - \ln(P_{t-1})$	Daily	Computed
RV (5-day)	$RV_t^{(5)}$	Daily	Computed
RV (30-day)	$RV_t^{(30)}$	Daily	Computed
VIX daily comp.	$\text{VIX}_t^{(d)}$	Daily	Computed
VIX weekly avg	$\text{VIX}_t^{(w)}$	Daily	Computed
VIX monthly avg	$\text{VIX}_t^{(m)}$	Daily	Computed

Table 2: Model settings (concise).

Model	**Key choices**
ARIMA	Orders by BIC on training window
GARCH(1,1)	Student–t errors; annualize $\widehat{\sigma}_{t+1}$
HAR (VIX)	OLS on $(\mathrm{VIX}_t^{(d)}, \mathrm{VIX}_t^{(w)}, \mathrm{VIX}_t^{(m)})$
RF	∼200 trees; max depth ≤ 5
XGBoost	Depth 3; shrinkage $\eta \in [0.03, 0.1]$; early stopping
Neural Net	1 hidden layer (ReLU, ≈ 10 units), L_2, early stop

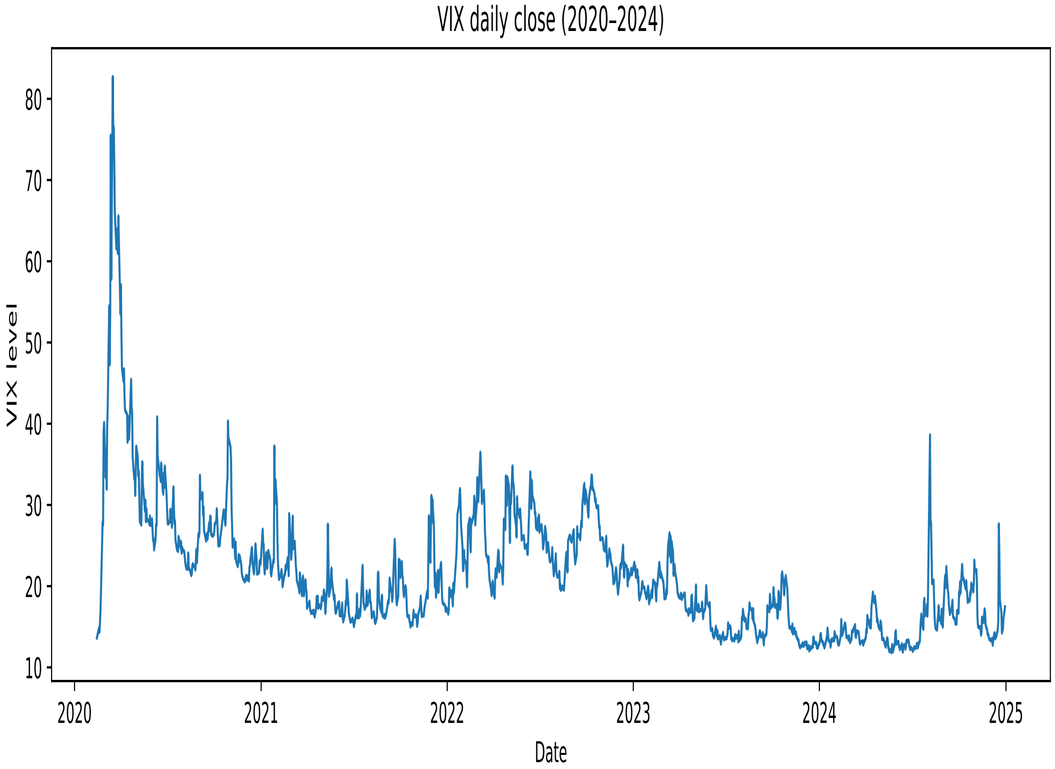

Figure 1: VIX daily close (2020–2024). The sample spans calm and stressed regimes, providing variation for model evaluation.

3.8 Notes on interpretation

It is important to recognize that VIX is a forward-looking, risk-neutral volatility measure. In practice, VIX tends to exceed the realized volatility that follows, reflecting a positive volatility risk premium (Whaley, 2009; Blair, Poon and Taylor, 2001; Bondarenko, 2014). Therefore, a model based purely on historical returns (such as GARCH) will often underpredict VIX. Models that directly include implied volatility information such as using past VIX levels in a HAR model or training on the VIX series can partially account for this effect (Corsi, 2009; Chen and Guestrin, 2016). As a result, we expect models leveraging VIX data directly to produce higher average forecasts than purely returns based models. For package versions, see Appendix A.

3.9 Visual and spectral diagnostics

To complement the descriptive statistics, we examine whether the VIX series exhibits patterns that forecasting models can exploit. The first set of panels contrasts OOS predictions against the realised VIX level under two rolling estimation windows. The 300-day window reacts more quickly to regime changes (Figure 2), whereas the 500-day window stabilises parameter estimates in calm periods but adapts more slowly during spikes (Figure 3).

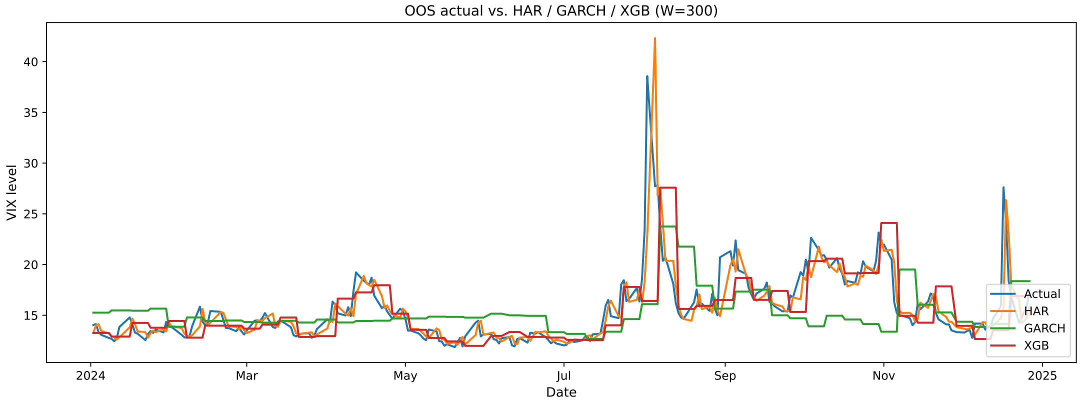

Figure 2: OOS VIX forecasts under a rolling 300-day training window for representative models (e.g., HAR, GARCH, XGBoost). Forecast errors expand around sharp spikes. Otherwise the models track level dynamics well.

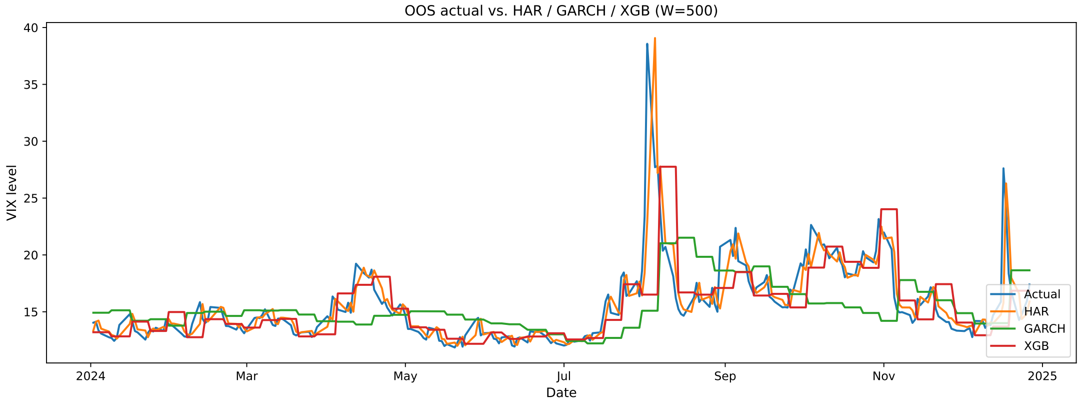

Figure 3: Same comparison with a 500-day rolling window. The longer window dampens noise in tranquil markets but adapts more slowly to abrupt transitions.

Next, we compare the VIX level with the forward 21-day realized volatility (annualized) computed from S&P 500 returns (Figure 4). VIX typically exceeds future realized volatility, consistent with a positive volatility risk premium, and motivates forecasting VIX directly when the aim is to anticipate option-implied risk.

Finally, we assess the frequency content of the core series. Fourier magnitude and Welch power spectral density indicate that most energy lies at low frequencies, reflecting persistent regimes rather than short-horizon periodicity (Figures 5–6).

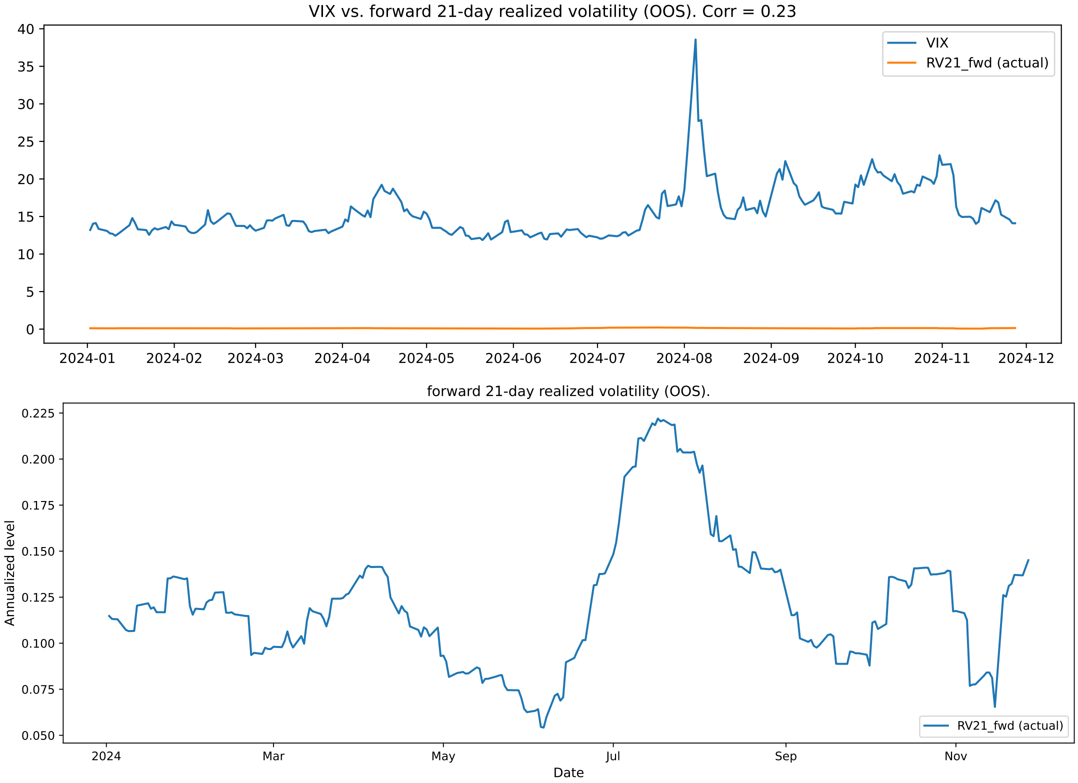

Figure 4: VIX versus forward 21-day realized volatility over the OOS period. The level gap reflects the volatility risk premium embedded in option prices.

Distinct narrow band cycles are weak, supporting models that capture long-memory and regime changes instead of strict seasonality.

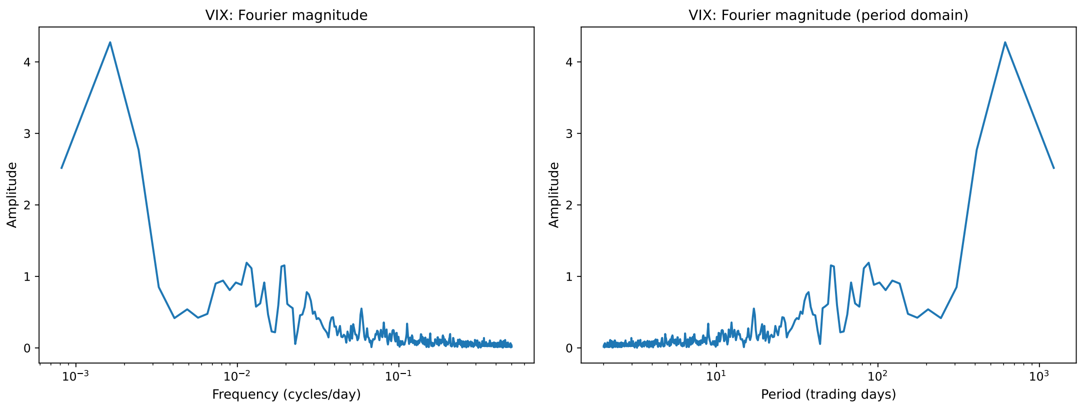

Figure 5: Fourier magnitude of the VIX series. Power concentrates at low frequencies (regime variation). Strong short-period cycles are not evident.

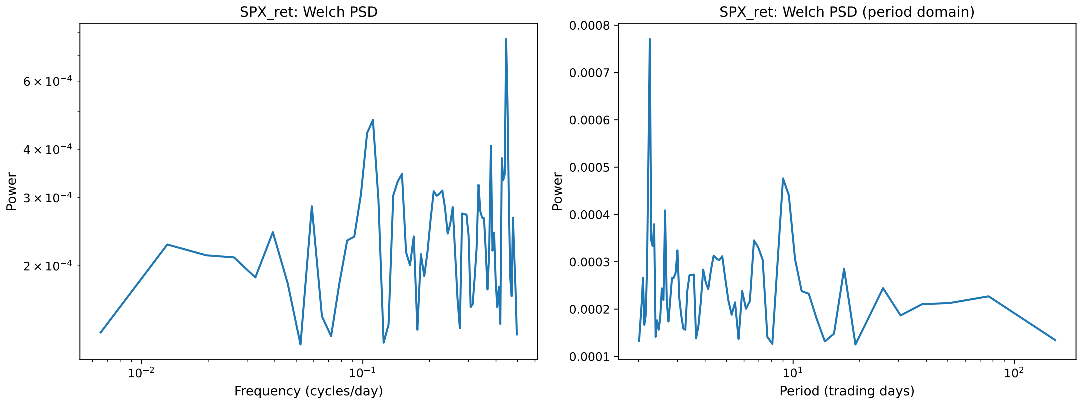

Figure 6: Welch power spectral density of S&P 500 daily log-returns. The spectrum is relatively flat at high frequencies with modest low-frequency power, consistent with weak daily serial dependence.

4 Results and evaluation

This section reports OOS forecasting results for the 2024 test period using the models and rolling-estimation design described in Section 3. We first compare statistical accuracy across econometric and machine-learning (ML) models under two rolling windows (300 and 500 trading days) in tables 3–4. We then assess the significance of performance differences via DM tests (Tables 5–6), examine calibration and error profiles through visual diagnostics (Figures 7–8), translate forecasts into a stylised volatility-timing rule and finally discuss computational efficiency.

4.1 Forecast accuracy

Tables 3–4 summarise RMSE, MAE, MAPE, R^2, and directional accuracy for all models. Consistent with heterogeneous volatility persistence, the HAR specification performs strongly in levels, and tree-based ML methods (RF, XGBoost) deliver competitive or superior RMSE by capturing non-linear interactions among lagged VIX components and realised-volatility covariates. ARIMA and the returns-based GARCH proxy underperform in level prediction, reflecting the imperfect and state-dependent link between conditional return variance and the option-implied VIX level.

Table 3: OOS forecast performance in 2024 (rolling window W=300). Lower is better for RMSE/MAE/MAPE and higher is better for R^2.

Model	RMSE	MAE	MAPE (%)	R^2	Dir. Acc. (%)
Mean	3.733	2.841	17.567	-0.231	52.0
ARIMA	3.075	1.850	10.676	0.165	50.4
GARCH(1,1)	3.595	2.444	14.380	-0.142	54.8
HAR	1.941	0.938	5.241	0.667	53.6
RF	3.040	1.649	9.354	0.184	50.0
XGB	2.923	1.590	9.049	0.245	48.0
NN	7.264	5.055	31.140	-3.662	53.6

Table 3 shows that HAR attains the strongest statistical accuracy (lowest RMSE/MAE/MAPE and highest R^2), followed by XGB and ARIMA. The returns-based GARCH proxy underperforms in level prediction despite the highest directional accuracy, reflecting an imperfect mapping from conditional return variance to the VIX level. RF is weaker than XGB in this setting, and the simple MLP (NN) is clearly overfit/mis-specified. Directional accuracy is clustered near 50–55%, with limited separation across models during the 2024 OOS window.

Table 4 HAR remains the strongest in level accuracy under W=500, while tree

Table 4: OOS forecast performance in 2024 (rolling window W=500). Lower is better for RMSE/MAE/MAPE and higher is better for R^2.

Model	RMSE	MAE	MAPE (%)	R^2	Dir. Acc. (%)
Mean	5.130	4.375	30.046	-1.325	50.0
ARIMA	3.147	1.903	11.031	0.125	48.4
GARCH(1,1)	3.414	2.294	13.441	-0.030	53.6
HAR	1.857	0.946	5.347	0.695	50.0
RF	2.904	1.603	9.102	0.255	50.0
XGB	2.913	1.607	9.134	0.250	48.0
NN	5.355	4.352	27.365	-1.533	54.8

ensembles (RF, XGB) are competitive, ARIMA is middling, the GARCH-based level proxy remains weak and the simple MLP (NN) underperforms.

Figure 7 overlays actual and predicted VIX for core models under W=300, illustrating that errors concentrate around abrupt volatility spikes. Per-model calibration scatters (Figures 9–18) align closely with the 45-degree line for the best-performing specifications, indicating good level calibration.

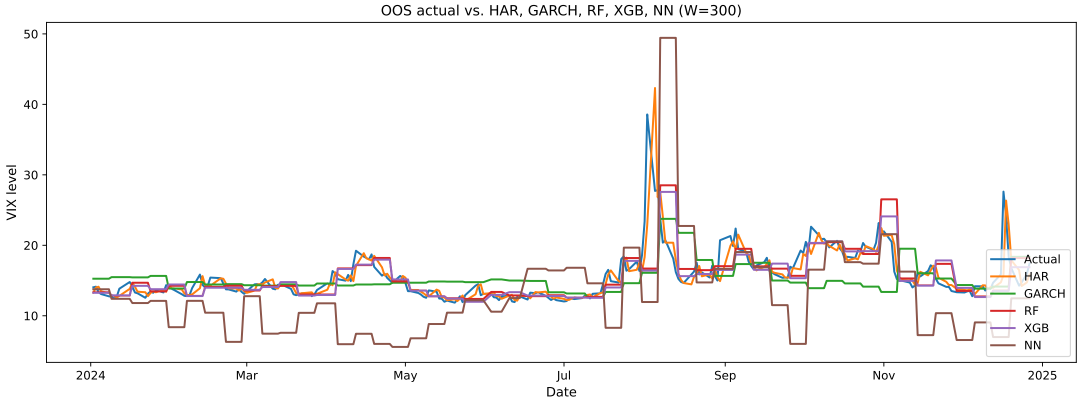

Figure 7: OOS actual VIX vs. predictions for HAR, GARCH, RF, XGB, and NN under a rolling window W=300.

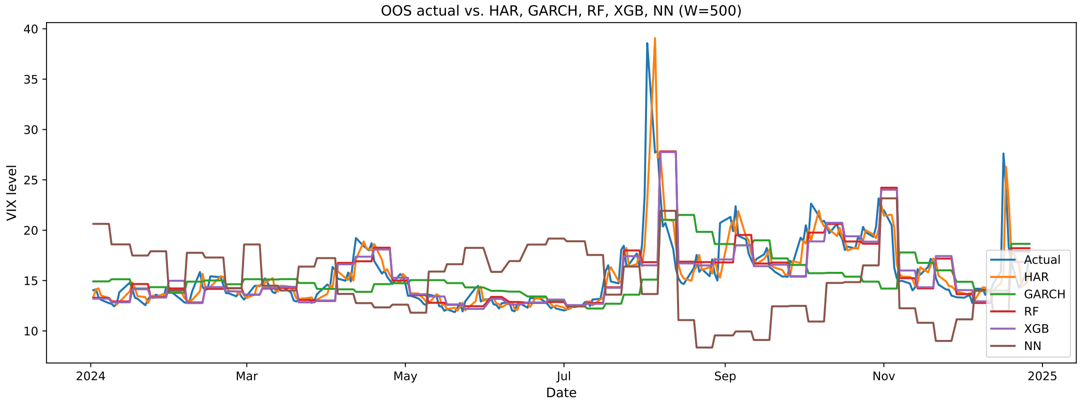

Figure 8: OOS actual VIX vs. predictions for HAR, GARCH, RF, XGB, and NN under a rolling window W=500.

4.2 DM tests

Pairwise DM tests use squared-error loss with a Newey–West HAC variance estimator and lag $L = 1$ (one-step-ahead, non-overlapping errors). We report two-sided p-values.

To assess whether differences in predictive accuracy are statistically significant, we apply the DM test to pairwise loss differentials under squared-error loss. The matrices

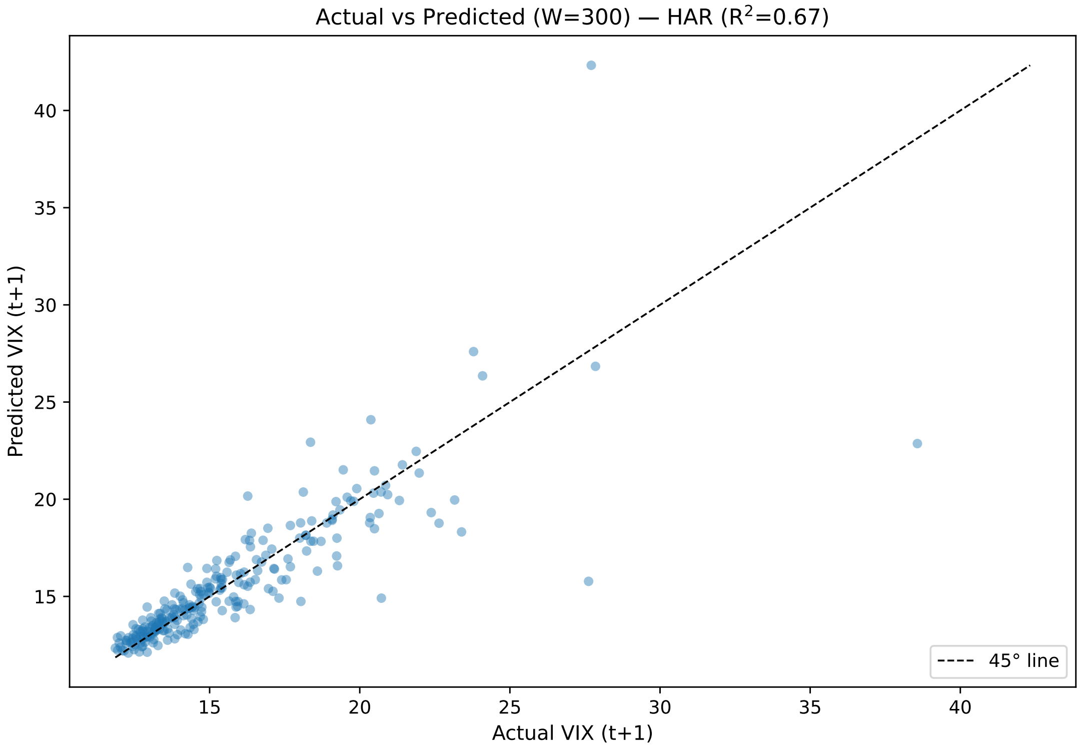

Figure 9: Calibration scatter (actual vs. predicted VIX) for HAR, rolling W=300. The 45° line is shown for reference.

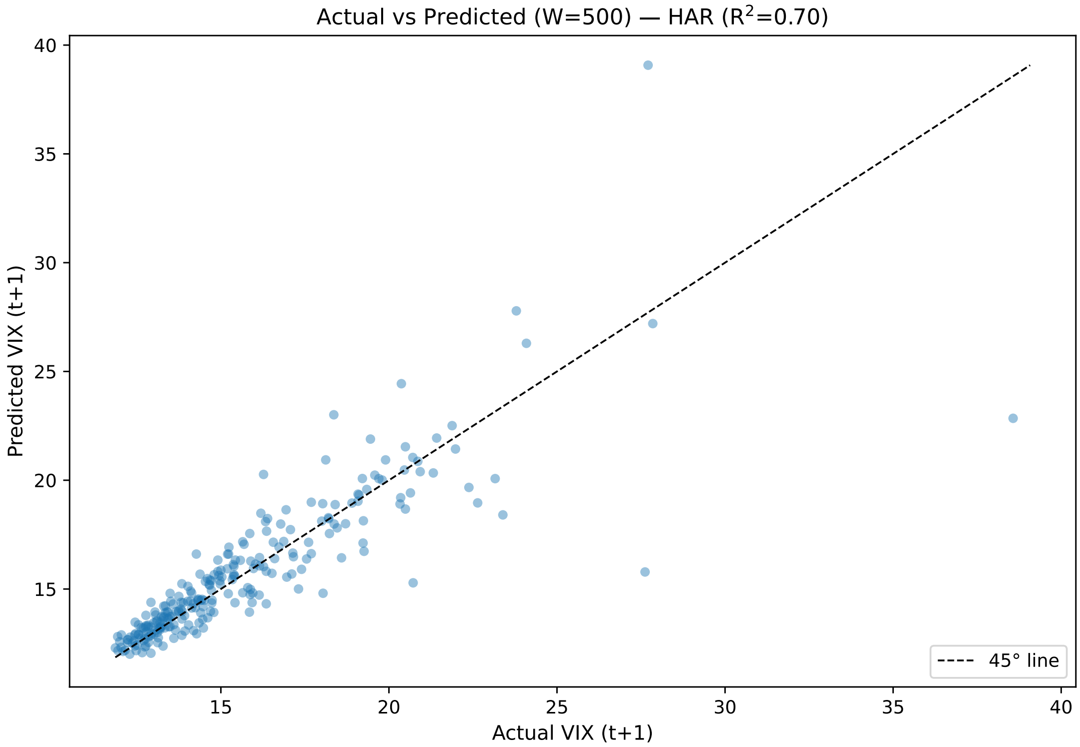

Figure 10: Calibration scatter (actual vs. predicted VIX) for HAR, rolling W=500. The 45° line is shown for reference.

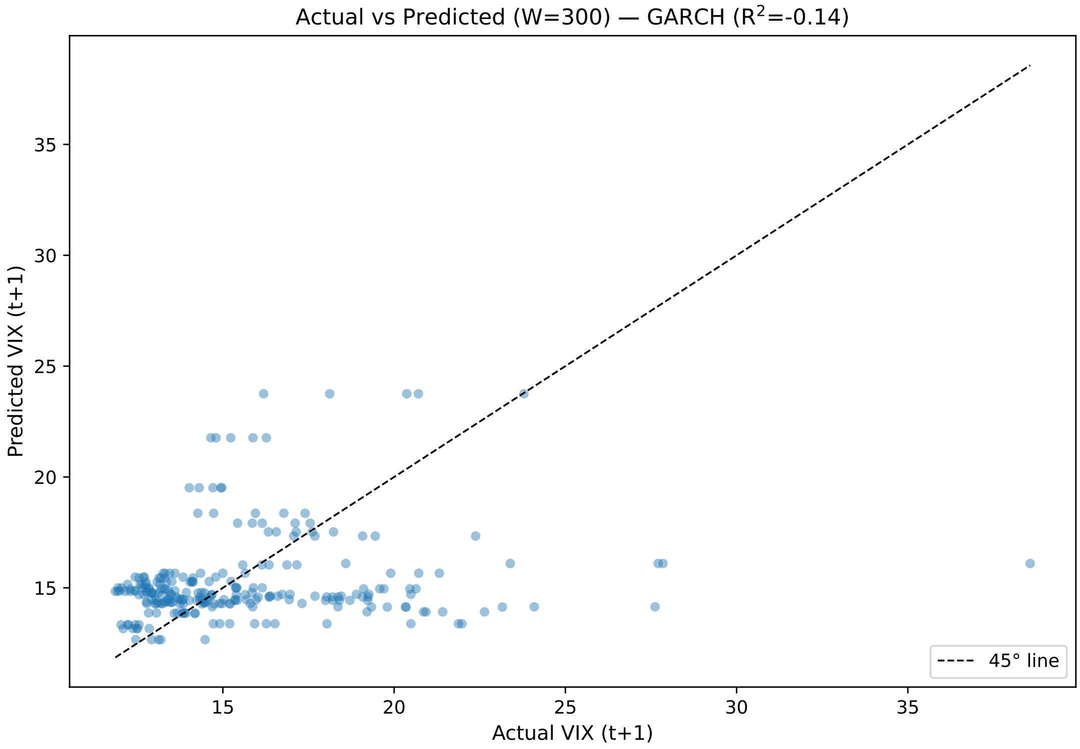

Figure 11: Calibration scatter (actual vs. predicted VIX) for GARCH, rolling W=300. The 45° line is shown for reference.

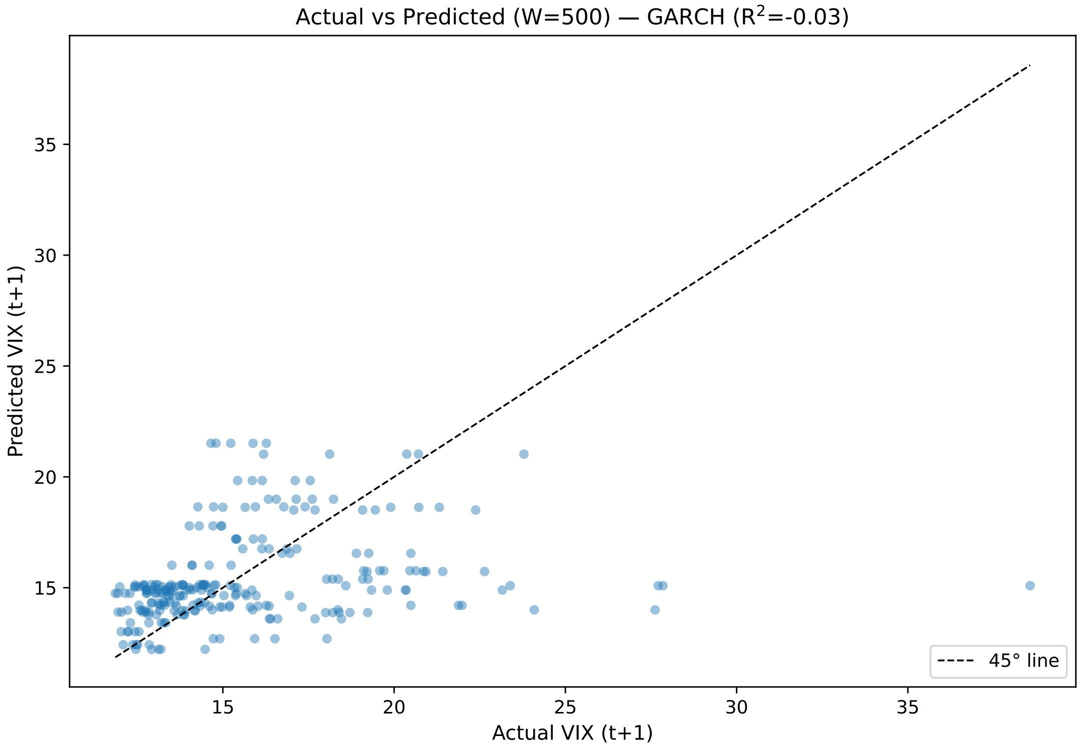

Figure 12: Calibration scatter (actual vs. predicted VIX) for GARCH, rolling W=500. The 45° line is shown for reference.

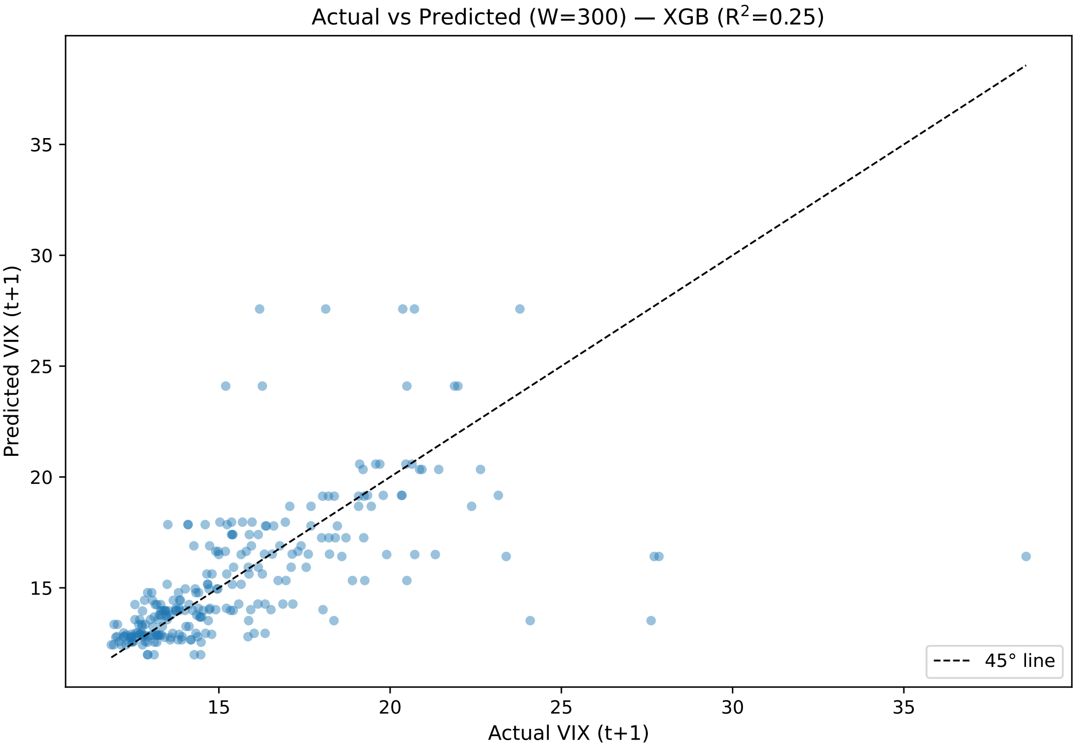

Figure 13: Calibration scatter (actual vs. predicted VIX) for XGB, rolling W=300. The 45° line is shown for reference.

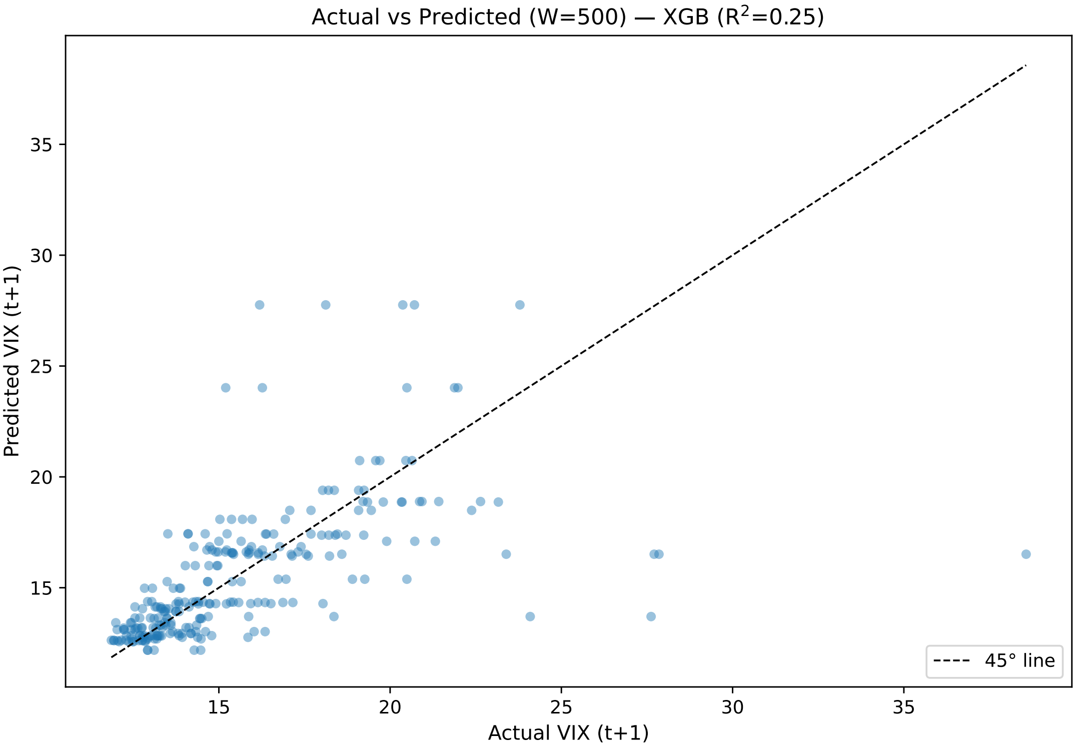

Figure 14: Calibration scatter (actual vs. predicted VIX) for XGB, rolling W=500. The 45° line is shown for reference.

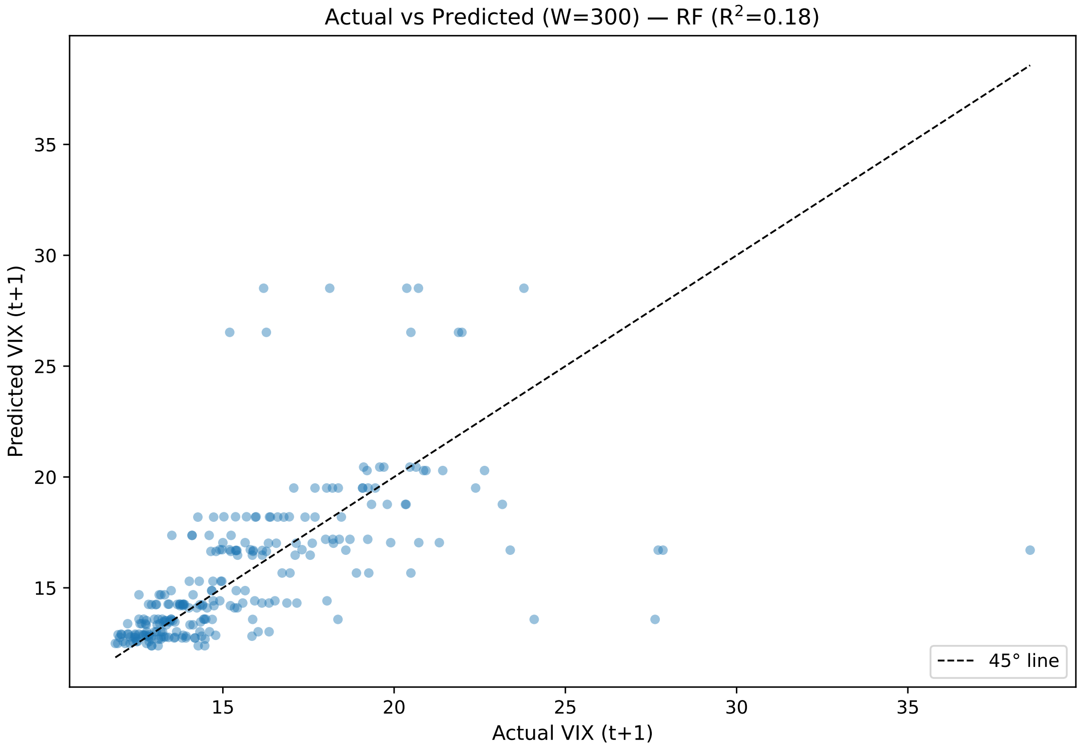

Figure 15: Calibration scatter (actual vs. predicted VIX) for RF, rolling W=300. The 45° line is shown for reference.

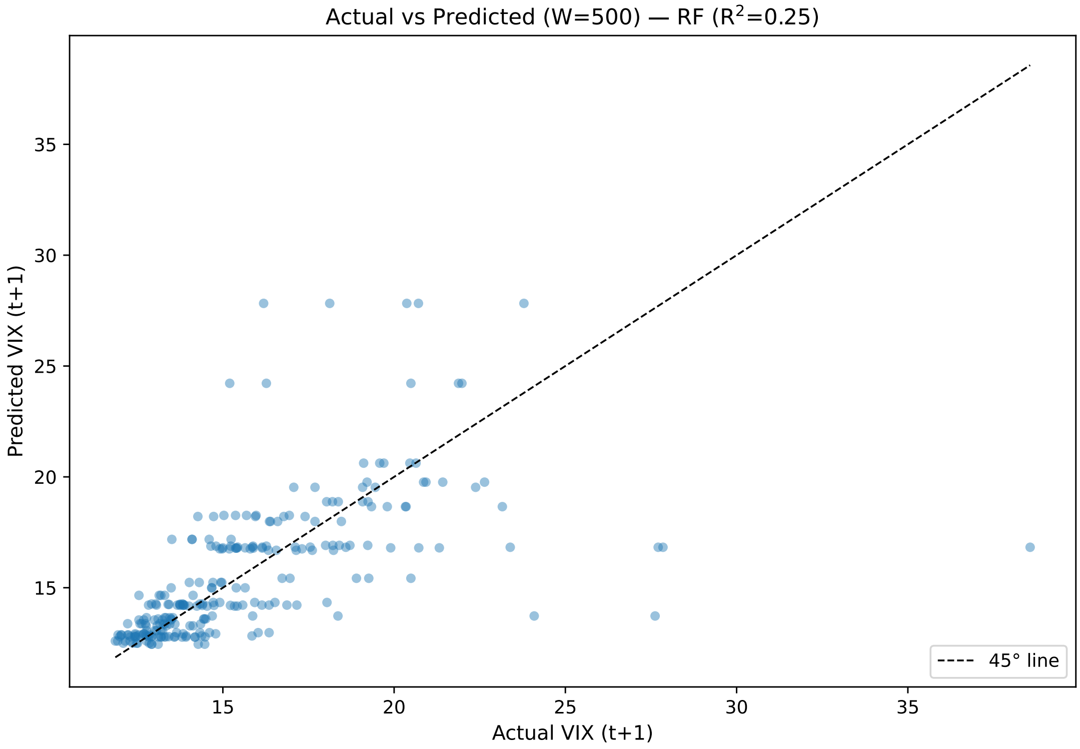

Figure 16: Calibration scatter (actual vs. predicted VIX) for RF, rolling W=500. The 45° line is shown for reference.

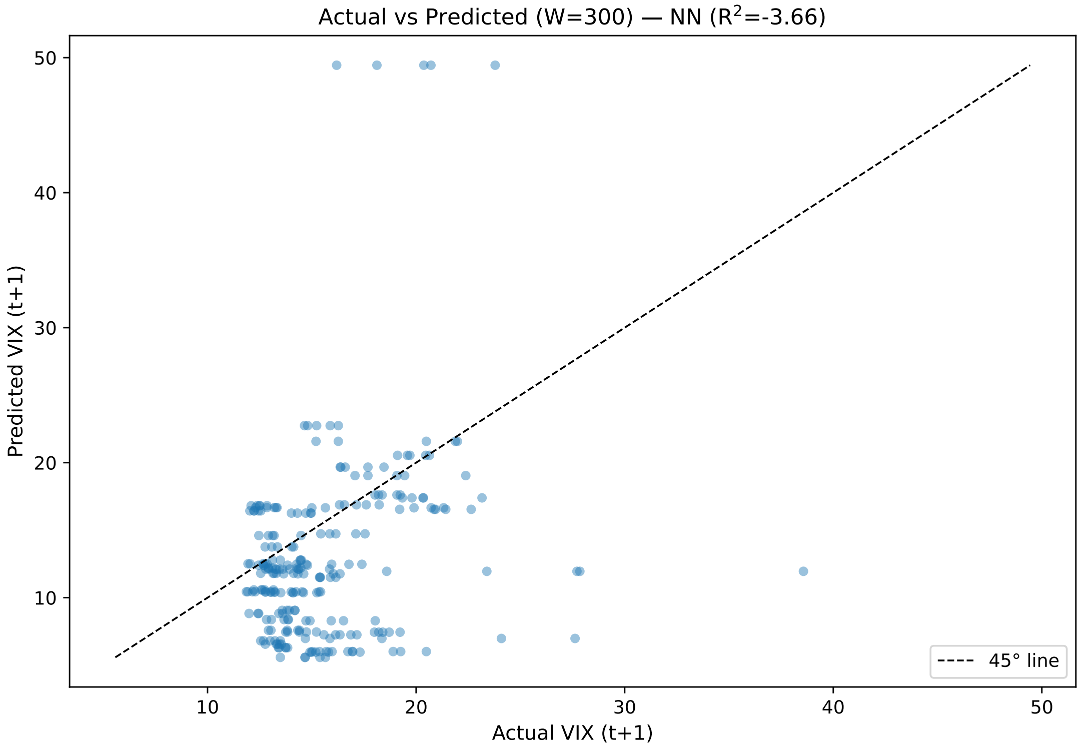

Figure 17: Calibration scatter (actual vs. predicted VIX) for NN, rolling W=300. The 45° line is shown for reference.

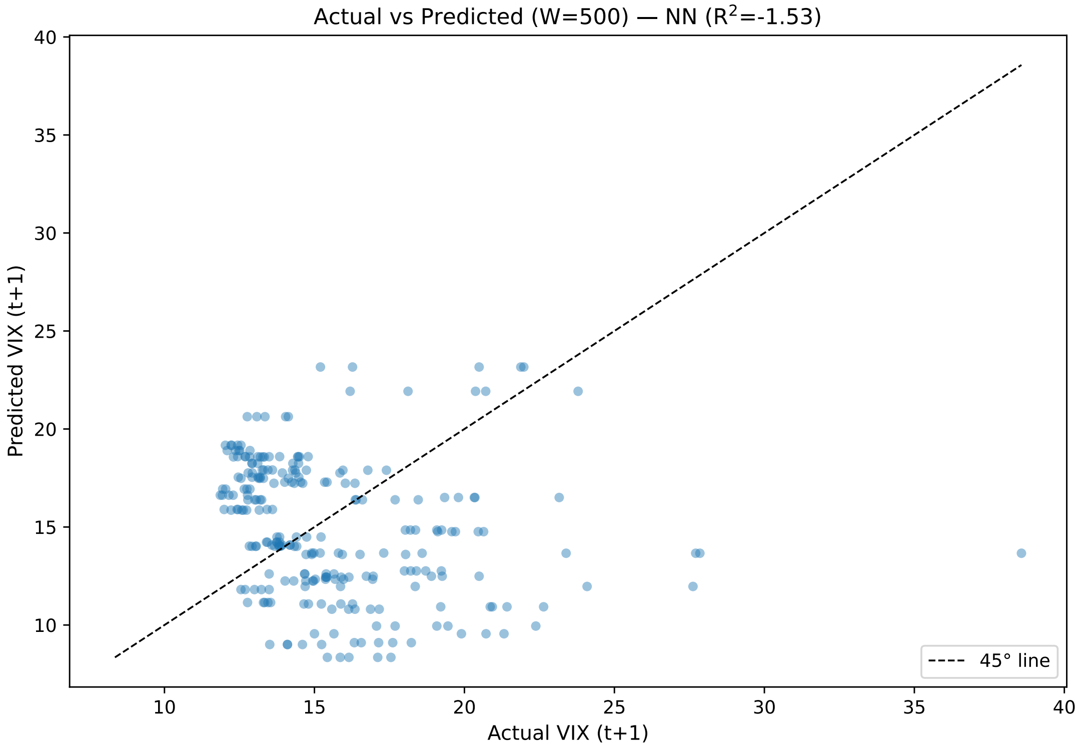

Figure 18: Calibration scatter (actual vs. predicted VIX) for NN, rolling W=500. The 45° line is shown for reference.

below report the test statistic and p-value (Tables 5–6). Positive entries indicate the row model has higher loss than the column model. Under both rolling windows, HAR and the leading ML specifications significantly outperform weaker baselines. Differences among the top ML models are smaller and frequently not significant.

Table 5: DM tests (squared-error loss), OOS 2024, rolling window W=300. Positive entries mean the row model has higher loss than the column model.

	HAR	GARCH	RF	XGB	NN
HAR	—	-4.21, p=0.000	-2.53, p=0.011	-2.59, p=0.010	-2.90, p=0.004
GARCH	4.21, p=0.000	—	2.04, p=0.042	2.46, p=0.014	-2.48, p=0.013
RF	2.53, p=0.011	-2.04, p=0.042	—	1.14, p=0.254	-2.85, p=0.004
XGB	2.59, p=0.010	-2.46, p=0.014	-1.14, p=0.254	—	-2.87, p=0.004
NN	2.90, p=0.004	2.48, p=0.013	2.85, p=0.004	2.87, p=0.004	—

Table 6: DM tests (squared-error loss), OOS 2024, rolling window W=500.

	HAR	GARCH	RF	XGB	NN
HAR	—	-3.59, p=0.000	-2.49, p=0.013	-2.45, p=0.014	-6.30, p=0.000
GARCH	3.59, p=0.000	—	2.13, p=0.033	2.14, p=0.033	-4.75, p=0.000
RF	2.49, p=0.013	-2.13, p=0.033	—	-0.30, p=0.762	-5.33, p=0.000
XGB	2.45, p=0.014	-2.14, p=0.033	0.30, p=0.762	—	-5.38, p=0.000
NN	6.30, p=0.000	4.75, p=0.000	5.33, p=0.000	5.38, p=0.000	—

4.3 Error profiles and robustness

We analyse bias and dispersion through rolling means of signed errors and rolling MAE. Figure 19 shows no persistent bias for the strongest models Short-lived drifts coincide with regime shifts. Rolling MAE (Figure 21–22) increases around volatility spikes, as expected. Binning errors by the actual VIX level (Figure 23–24) reveals larger absolute errors at high levels, although relative errors remain controlled. Error distributions (Figure 25) exhibit heavier tails during stress episodes.

4.4 Economic value

We map forecasts into a stylised volatility-timing rule using the predicted change in VIX and either a discrete sign position with a data-driven threshold or a continuous, standardised sizing rule. Tables 7–10 report annualised Sharpe ratios, active-day hit

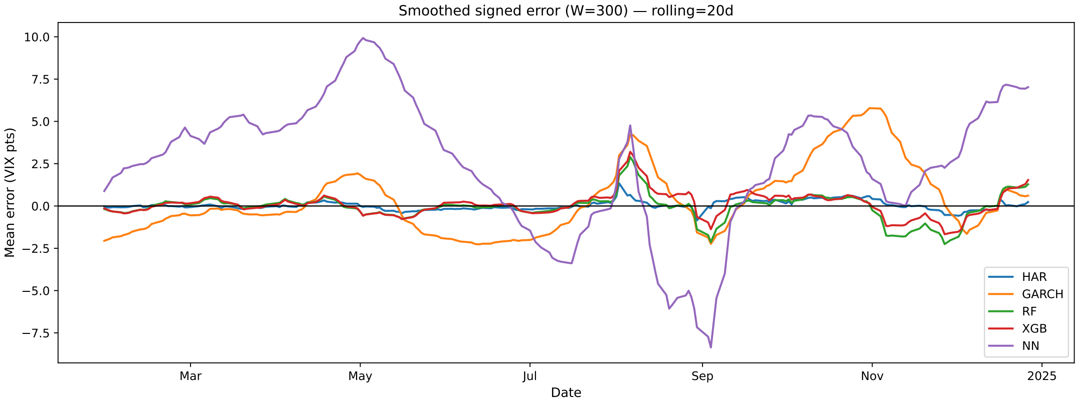

Figure 19: Smoothed (20-day) signed errors by model, rolling W=300. No persistent bias is evident.

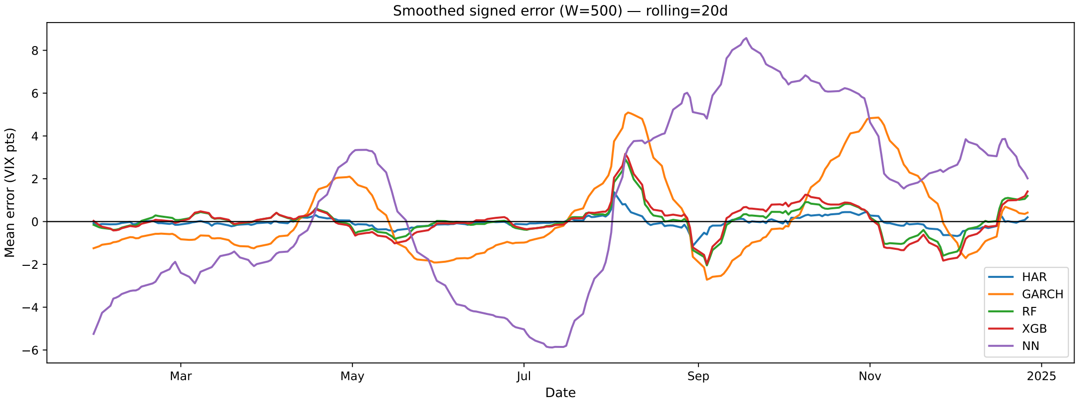

Figure 20: Smoothed (20-day) signed errors by model, rolling W=500. No persistent bias is evident.

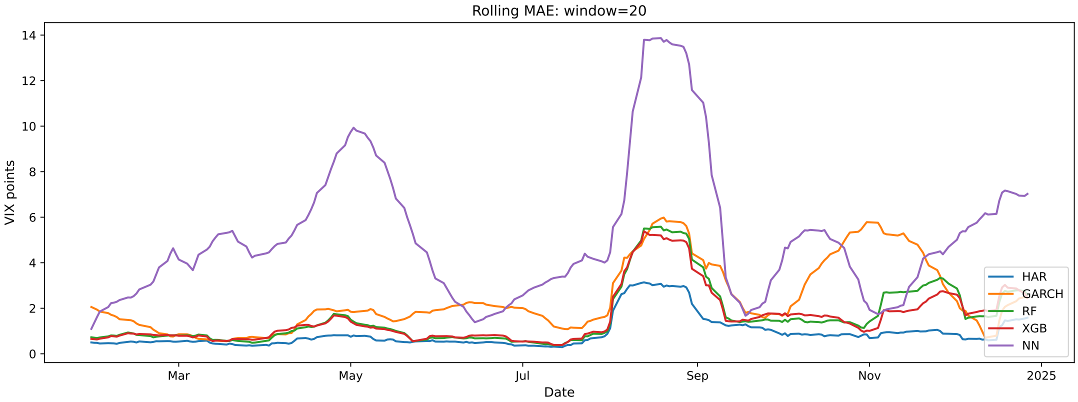

Figure 21: Rolling MAE (20-day window) by model, rolling W=300. Errors peak around volatility spikes.

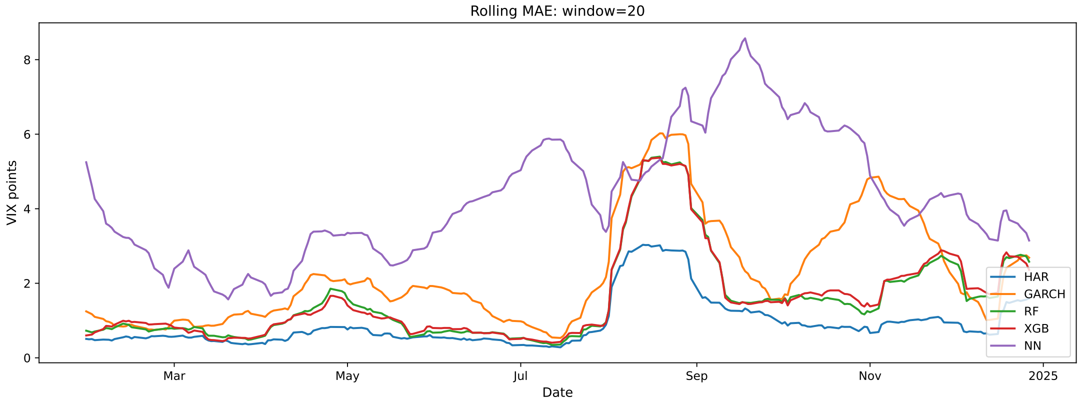

Figure 22: Rolling MAE (20-day window) by model, rolling W=500. Errors peak around volatility spikes.

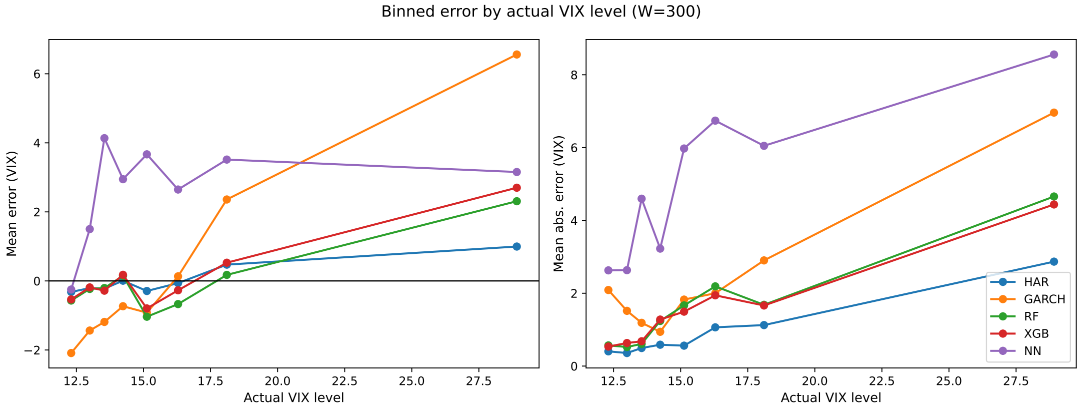

Figure 23: Binned errors by actual VIX level, rolling W=300. Absolute errors increase with level and relative errors remain contained.

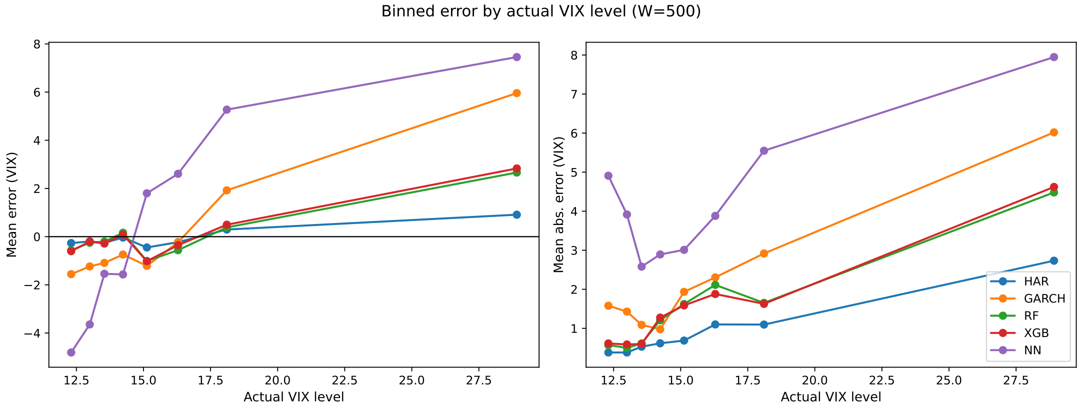

Figure 24: Binned errors by actual VIX level, rolling W=500. Absolute errors increase with level and relative errors remain contained.

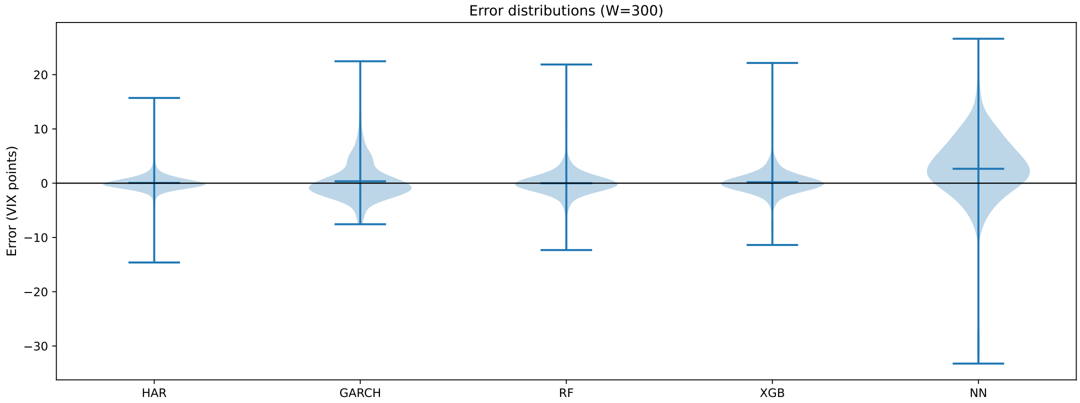

Figure 25: Error distributions (violin plots), rolling W=300. Tails widen during stress episodes.

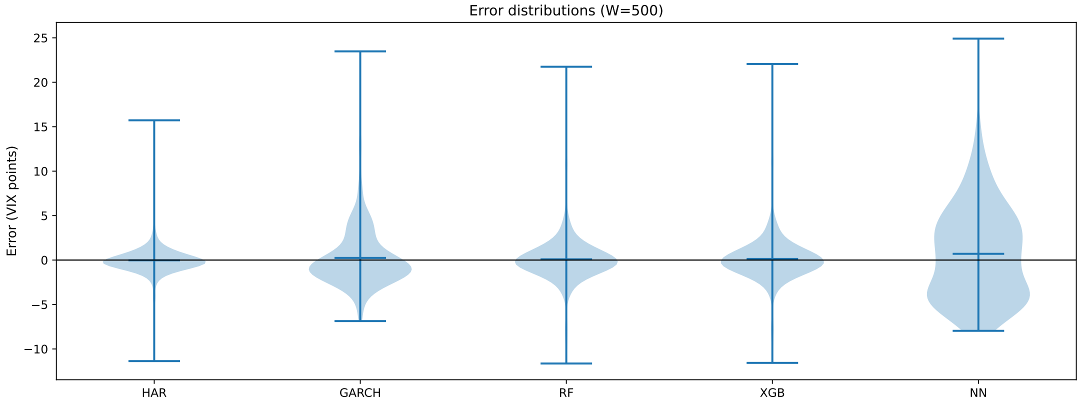

Figure 26: Error distributions (violin plots), rolling W=300. Tails widen during stress episodes.

rates, and activity shares. The strongest statistical models yield the highest economic performance, with sensitivity to the threshold and clustering of spikes.

Table 7: Economic evaluation (discrete sign rule), OOS 2024, rolling W=300.

	Sharpe	HitActive(%)	ActiveDays(%)
Mean	1.61	52.86	90.8
ARIMA	-0.97	46.78	68.4
GARCH	0.42	54.9	81.6
HAR	0.73	71.43	2.8
RF	-0.62	50.74	54.4
XGB	-0.50	52.67	60.0
NN	-1.43	53.78	90.0

Table 8: Economic evaluation (continuous sizing), OOS 2024, rolling W=300.

	Sharpe	HitActive(%)	ActiveDays(%)
Mean	0.94	51.63	98.4
ARIMA	-0.25	49.79	97.2
GARCH	0.47	55.42	96.0
HAR	0.03	53.09	97.2
RF	-0.73	49.17	96.0
XGB	-0.62	47.76	98.0
NN	-1.24	53.66	98.4

Table 9: Economic evaluation (discrete sign rule), OOS 2024, rolling W=500.

	Sharpe	HitActive(%)	ActiveDays(%)
Mean	1.87	50.22	92.4
ARIMA	-0.49	47.73	70.4
GARCH	0.28	51.66	84.4
HAR	1.36	100.0	0.8
RF	-0.79	49.65	56.4
XGB	-0.65	49.04	62.8
NN	-0.51	55.84	92.4

4.5 Computational efficiency

We compare training time and accuracy to assess efficiency. Figure 27 contrasts total training time by model across windows, while Figure 28 plots RMSE against training time for W=300. HAR and RF offer favorable accuracy–cost trade–offs deeper

Table 10: Economic evaluation (continuous sizing), OOS 2024, rolling W=500.

	Sharpe	**HitActive(%)**	**ActiveDays(%)**
Mean	2.19	50.2	99.6
ARIMA	-0.37	49.19	98.4
GARCH	0.35	53.63	99.2
HAR	0.34	49.8	98.8
RF	-0.67	49.79	96.4
XGB	-0.71	47.48	95.2
NN	-0.45	55.06	98.8

boosting and NN can be competitive at higher computational expense. In addition to RMSE–time frontiers (Figures 28–29), Sharpe–time frontiers (Figures 30–31) summarize the economic efficiency trade-off.

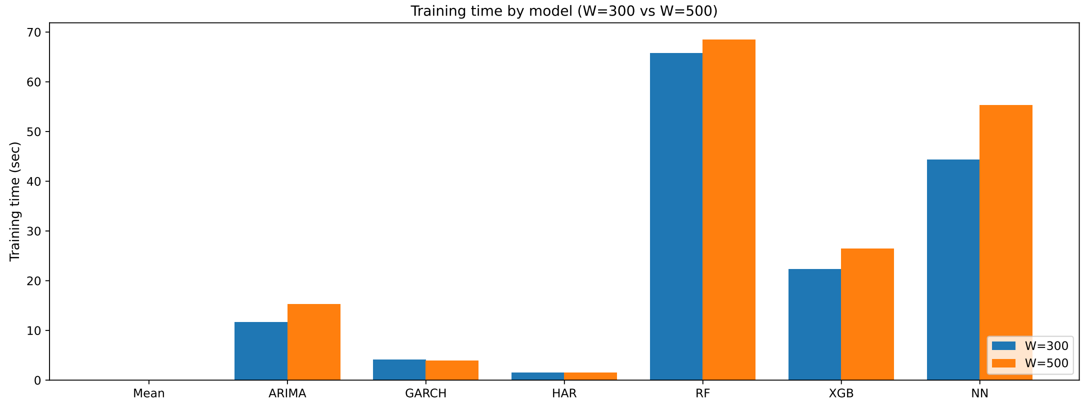

Figure 27: Total training time by model under rolling windows W=300 and W=500.

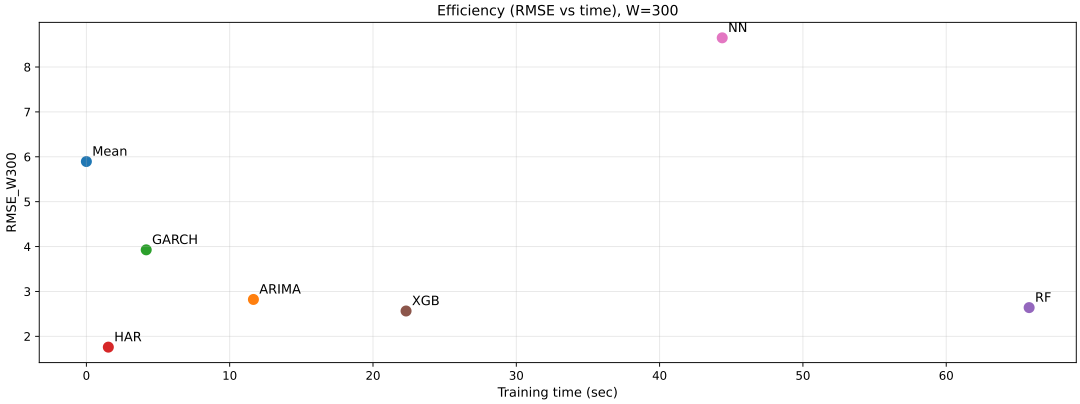

Figure 28: Efficiency frontier: RMSE versus training time, rolling W=300.

4.6 Summary

We constructed a daily one–step–ahead volatility forecasting framework centered on VIX over 2020–2024 using public inputs (variables and sources in Table 1 and model settings in Table 2). VIX is a model-free, risk-neutral proxy for 30-day variance (Whaley, 2009; CBOE, 2023), while S&P 500 returns provide backward-looking realized-volatility covariates (Andersen et al., 2001). The sample spans calm and stressed regimes (Figure 1). Because risk-neutral expectations embed a positive volatility risk premium, returns-based conditional variance models often under predict VIX levels (Blair, Poon and Taylor, 2001; Bondarenko, 2014; Whaley, 2009).

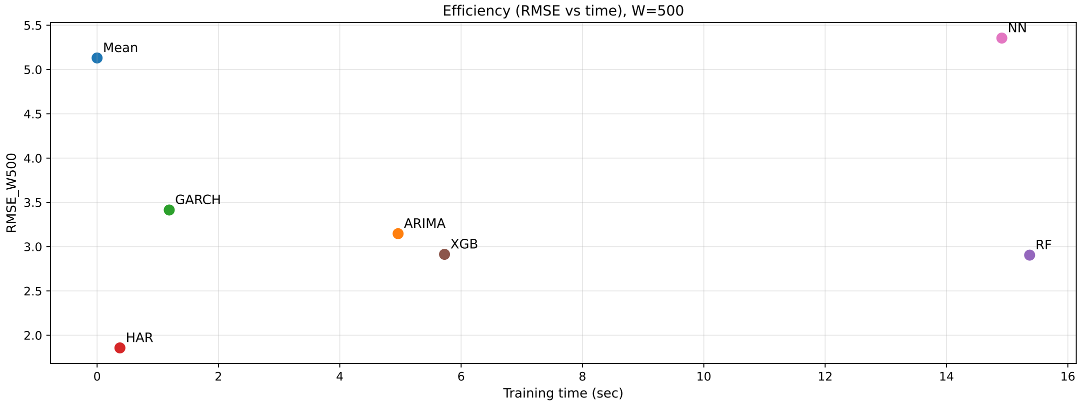

Figure 29: Efficiency frontier: RMSE versus training time, rolling W=500.

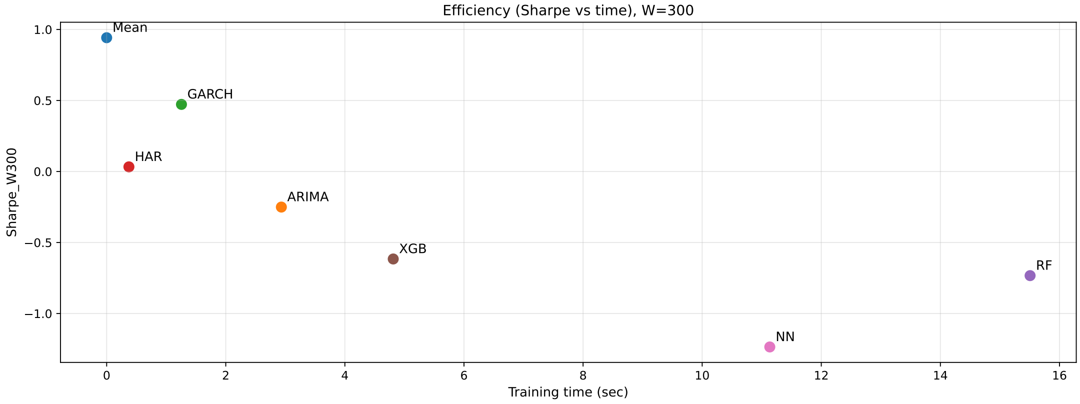

Figure 30: Efficiency frontier: Sharpe versus training time, rolling W=300.

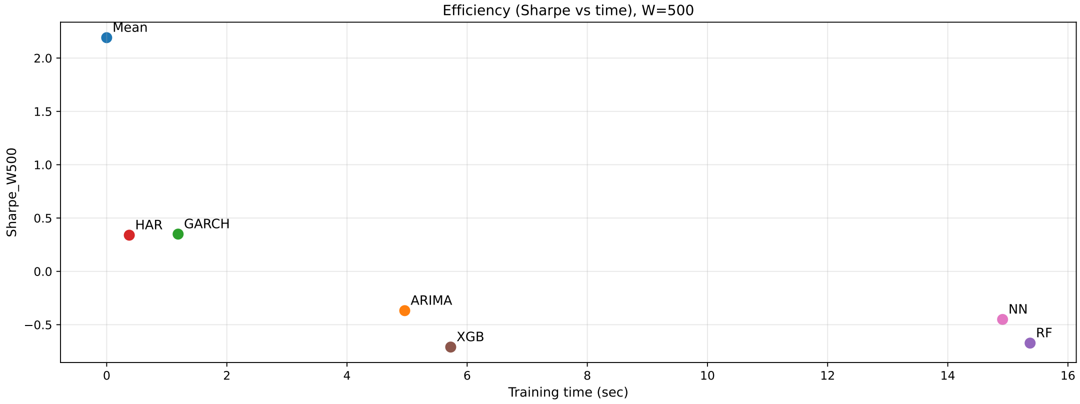

Figure 31: Efficiency frontier: Sharpe versus training time, rolling W=500.

Our model set balances interpretability and flexibility: ARIMA, GARCH and HAR capture linear autocorrelation, volatility clustering and multi-horizon persistence, respectively (Engle, 1982; Bollerslev, 1986; Corsi, 2009), while RF, XGB and a shallow NN allow non-linear interactions across daily/weekly/monthly VIX components and realised volatility inputs (Breiman, 2001; Chen and Guestrin, 2016; Goodfellow, Bengio and Courville, 2016). Oos overlays illustrate tracking and spike misspecification patterns (Figures 7–8). Calibration scatters show level fit (Figures 9–18). Spectral/volatility-gap diagnostics motivate targeting VIX directly (Figures 4–6).

Accuracy comparisons (Tables 3–4) and DM tests (Tables 5–6) show that specifications leveraging lagged VIX (e.g., HAR) attain the strongest level RMSE/MAE/R^2 across windows, while tree-based ML is competitive by capturing nonlinearities. Error profiles concentrate around volatility spikes (Figures 19–26). Mapping signals into simple timing rules translates statistical gains into economic terms (Tables 7–10). Training-time frontiers summarize accuracy–cost trade-offs (Figures 27–29). The framework is transparent and extensible to alternative horizons/targets without changing the validation logic (Andersen et al., 2001; Corsi, 2009; Chen and Guestrin, 2016).

5 Discussion

Across both rolling windows (W=300, 500), models that explicitly exploit multi-horizon persistence in implied volatility deliver the most accurate level forecasts. The HAR specification, which aggregates daily, weekly and monthly VIX components, attains the best RMSE/MAE and maintains good level calibration (Tables 3–4; Figures 7–8, 9–10). This is consistent with the long-memory view of volatility and the idea that implied volatility behaves as a persistent state variable (Corsi, 2009). A returns-based GARCH proxy underpredicts the VIX level throughout the out-of-sample period (Tables 3–4; Figures 11–12). The gap lines up with prior evidence on the volatility risk premium embedded in option prices and with the well-known difference between risk-neutral and physical-measure objects (Blair, Poon and Taylor, 2001; Bondarenko, 2014; Whaley, 2009). In practical terms, when the aim is to anticipate option-implied risk, forecasting the VIX directly is preferable to mapping conditional return variance into a VIX level. Our findings reinforce this and are consistent with Ahoniemi (2008), who found that GARCH did not outperform simpler ARIMA models in VIX prediction, and with Liu, Guo and Qiao (2015), who documented that GARCH forecasts systematically underpredict VIX levels. Wang (2019) further supports HAR's superiority by showing that lagged VIX components improve forecast accuracy over naïve AR(1) alternatives.

Flexible learners help insofar as they capture nonlinear interactions among lagged VIX components and realized-volatility covariates. Tree ensembles (RF, XGBoost) are competitive on RMSE and level calibration (Figures 13–16), which is consistent with recent results on machine learning in asset pricing and scalable boosting (Gu, Kelly and Xiu, 2020; Chen and Guestrin, 2016; Breiman, 2001). Our evidence is in line with Grefhorst (2024), who used XGBoost to predict VIX and achieved a MAPE around 4.8%, and with Degiannakis, Filis and Hassani (2018), who showed that nonlinear and non-parametric methods can improve global volatility index forecasts. Wu, He and Xie (2023) report that enhanced GARCH-MIDAS models outperform standard GARCHs by accounting for time-varying risk aversion, consistent with our finding that more flexible structures capture shifts better. Forecasting volatility spikes remains challenging therefore errors widen in stress, and signed errors drift near regime shifts (Figures 25–26; 19–20). This pattern is acknowledged in recent work such as Bai and Cai (2024), who note the limits of predictive capacity under turbulence. Regime-adaptive strategies and covariate expansion help, but perfect spike capture remains out of reach.

Statistical gains do not automatically translate into tradable gains. The stylised timing rules show that Sharpe improvements depend on sizing choices, participation thresholds and regime clustering (Tables 7–10). Selective participation can produce high hit rates with few active days (HAR under discrete rules), whereas continuous sizing can add value even without markedly better direction. The efficiency views make the trade-offs explicit RMSE–time frontiers capture statistical efficiency (Figures 28–29) and Sharpe–time frontiers summarize economic efficiency (Figures 30–31), while Figure 27 shows absolute training-time profiles. Prior studies also caution against assuming statistical advantage implies economic value. Ahoniemi (2008) reported strategy profits that diminished after costs. Poon and Granger (2003) highlight that even highly accurate models may not lead to arbitrage opportunities. Net Sharpe must account for slippage, turnover and capacity. In our results, HAR's smoother signals likely hold up better under costs, whereas tree models could suffer unless post-processed. This mirrors conclusions in S. Wang *et al.* (2024), who emphasise signal sparsity and cost-aware tuning in ML-based VIX trading systems. Our findings support the view that models should be evaluated not only by predictive power, but also implementation robustness an approach aligned with best practices in recent volatility-timing literature.

5.1 Limitations and threats to validity

Our conclusions are conditioned on the 2024 out-of-sample window and daily close data. Intraday information and official settlement mechanics are not used. The mapping from statistical to economic value is sensitive to transaction costs, fees, slippage and execution timing, and model rankings can change once frictions are included. The GARCH-to-level mapping is approximate because VIX is risk-neutral while conditional return variance is a physical-measure object (Blair, Poon and Taylor, 2001; Bondarenko, 2014; Whaley, 2009). These caveats frame the contribution as a transparent, reproducible benchmark rather than a production trading system.

5.2 Trading costs and sensitivity analysis

When frictions are accounted for, signal stability becomes central. Linear and HAR-type models typically require fewer adjustments and generate lower turnover, which can make them look relatively better after costs. Tree ensembles may introduce jitter near split thresholds unless signals are post-processed. In practice, models should be tuned on cost-aware objectives for example a turnover-penalized loss or net Sharpe, rebalancing should be disciplined and execution assumptions should be explicit. Psaradellis et al. (2016) further demonstrate that high predictive accuracy in implied volatility models does not necessarily translate into economic gains once transaction costs and execution frictions are accounted for. Sensitivity checks should vary costs, rebalancing frequency and thresholds, key hyperparameters, regimes and execution timing. Model rankings that persist across these scenarios are more credible for live use. The Sharpe–time frontiers (Figures 30–31) together with turnover statistics from the timing results provide a concise view of how much friction each approach can absorb before underperforming (Gu, Kelly and Xiu, 2020; Chen and Guestrin, 2016; Breiman, 2001).

5.3 Conclusions

In conclusion, this study demonstrates that machine-learning methods can improve one-day-ahead VIX forecasts relative to ARIMA and GARCH. However, the statistical gains translate into economic value only under realistic frictions and when signals are sufficiently stable. HAR remains a strong benchmark because of volatility persistence (Figures 9–10), while tree-based models (RF (Figures 15–16) and XGBoost (Figures 13–14)) capture nonlinearities that lift forecast accuracy.

6 References

Ahoniemi, K. (2008). *Modeling and Forecasting the VIX Index*. SSRN Working Paper. [online] Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1033812 [Accessed 1 Nov. 2025].

Andersen, T.G., Bollerslev, T., Diebold, F.X. and Labys, P. (2001). *The distribution of realized exchange rate volatility*. Journal of the American Statistical Association, 96(453), pp. 42–55.

Aroussi, R. (2015). *yfinance:* Yahoo! Finance market data downloader. [online] Available at: https://github.com/ranaroussi/yfinance [Accessed 1 Nov. 2025].

Bai, C. and Cai, Y. (2024). *Predicting the VIX using adaptive machine learning: dynamic hyperparameter search and online feature selection*. Quantitative Finance (forthcoming). [online] Available at: https://livrepository.liverpool.ac.uk/3161274/1/Predicting%20the%20VIX%20using%20adaptive%20machine%20learning.pdf [Accessed 1 Nov. 2025].

Black, F. and Scholes, M. (1973). *The pricing of options and corporate liabilities*. Journal of Political Economy, 81(3), pp. 637–654.

Blair, B.J., Poon, S.-H. and Taylor, S.J. (2001). *Forecasting S&P 100 volatility: the incremental information content of implied volatilities and high-frequency index returns*. Journal of Econometrics, 105(1), pp. 5–26.

Bollerslev, T. (1986). *Generalized autoregressive conditional heteroskedasticity*. Journal of Econometrics, 31(3), pp. 307–327.

Bondarenko, O. (2014). *Why are put options so expensive?* Quarterly Journal of Finance, 4(3), 1450015.

Box, G.E.P., Jenkins, G.M., Reinsel, G.C. and Ljung, G.M. (2015). *Time Series Analysis: Forecasting and Control*. 5th ed. Hoboken, NJ: Wiley.

Breiman, L. (2001). *Random forests*. Machine Learning, 45(1), pp. 5–32.

Cboe Global Markets (2019). *The Cboe Volatility Index — VIX®: White Paper*. [online] Available at: https://cdn.cboe.com/resources/vix/vixwhitepaperjan2019.pdf [Accessed 1 Nov. 2025].

Cboe Global Markets (2023). *Historical data for Cboe Volatility Index (VIX)*. [online] Available at: https://www.cboe.com/us/products/vix-index-volatility/vix-options-and-futures/vix-index/historical_data/ [Accessed 1 Nov. 2025].

Chen, T. and Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794.

Cont, R. (2001). *Empirical properties of asset returns: stylized facts and statistical issues*. Quantitative Finance, 1(2), pp. 223–236.

Cont, R. and Tankov, P. (2004). *Financial Modelling with Jump Processes*. Boca Raton: Chapman & Hall/CRC.

Corsi, F. (2009). *A simple approximate long-memory model of realized volatility*. Journal of Financial Econometrics, 7(2), pp. 174–196.

Degiannakis, S., Filis, G. and Hassani, H. (2018). *Forecasting global stock market implied volatility indices*. Journal of Empirical Finance, 46, pp. 111–129.

Diebold, F.X. and Mariano, R.S. (1995). *Comparing predictive accuracy*. Journal of Business & Economic Statistics, 13(3), pp. 253–263.

Dupire, B. (1994). *Pricing with a Smile*. Risk Magazine, 7(1), pp. 18–20.

Engle, R.F. (1982). *Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation*. Econometrica, 50(4), pp. 987–1007.

Fama, E.F. (1965). *The behavior of stock-market prices*. Journal of Business, 38(1), pp. 34–105.

Friedman, J.H. (2001). *Greedy function approximation: A gradient boosting machine*. Annals of Statistics, 29(5), pp. 1189–1232.

Goodfellow, I., Bengio, Y. and Courville, A. (2016). *Deep Learning*. MIT Press.

Grefhorst, J. (2024). *Predicting the VIX index using time series modelling and machine learning*. Bachelor's Thesis, Erasmus University Rotterdam. [online] Available at: https://thesis.eur.nl/pub/70166/Grefhorst_527447.pdf [Accessed 1 Nov. 2025].

Gu, S., Kelly, B. and Xiu, D. (2020). *Empirical asset pricing via machine learning*. Review of Financial Studies, 33(5), pp. 2223–2273.

Harris, C.R., Millman, K.J., van der Walt, S.J., et al. (2020). *Array programming with NumPy*. Nature, 585, pp. 357–362.

Heston, S.L. (1993). *A closed-form solution for options with stochastic volatility*. Review of Financial Studies, 6(2), pp. 327–343.

Hornik, K., Stinchcombe, M. and White, H. (1989). *Multilayer feedforward networks* *are universal approximators*. Neural Networks, 2(5), pp. 359–366.

Hull, J.C. (2014). *Options, Futures and Other Derivatives*. 9th ed. Harlow: Pearson Education.

Hunter, J.D. (2007). *Matplotlib: A 2D graphics environment*. Computing in Science & Engineering, 9(3), pp. 90–95.

Konstantinidi, E., Skiadopoulos, G. and Tzagkaraki, E. (2008). *Can the evolution of implied volatility be forecasted? Evidence from European and US implied volatility indices*. Journal of Banking & Finance, 32(11), pp. 2401–2411.

Liu, J.Y., Guo, P. and Qiao, Z. (2015). *VIX forecasting and variance risk premium: A new GARCH approach*. North American Journal of Economics and Finance, 34, pp. 314–322.

Mandelbrot, B. (1963). *The variation of certain speculative prices*. Journal of Business, 36(4), pp. 394–419.

McKinney, W. (2010). *Data structures for statistical computing in Python*. Proceedings of the 9th Python in Science Conference, pp. 51–56.

Nelson, D.B. (1991). *Conditional heteroskedasticity in asset returns: A new approach*. Econometrica, 59(2), pp. 347–370.

OpenAI (2025). *ChatGPT (GPT-5 Thinking): large language model used as a translation aid*. [online] Available at: https://chat.openai.com [Accessed 1 Nov. 2025].

Pedregosa, F. et al. (2011). *Scikit-learn: Machine learning in Python*. Journal of Machine Learning Research, 12, pp. 2825–2830.

Poon, S.-H. and Granger, C.W.J. (2003). *Forecasting volatility in financial markets: A review*. Journal of Economic Literature, 41(2), pp. 478–539.

Psaradellis, I., Sermpinis, G., Theofilatos, K., Stasinakis, C. and Dunis, C. (2016). *Modelling and trading the US implied volatility indices*. Working paper, University of Glasgow. [online] Available at: https://eprints.gla.ac.uk/120240/7/120240.pdf [Accessed 1 Nov. 2025].

Python Software Foundation (2025). *Python Language Reference*. [online] Available at: https://docs.python.org/3/ [Accessed 1 Nov. 2025].

Seabold, S. and Perktold, J. (2010). *Statsmodels: Econometric and statistical modeling with Python*. Proceedings of the 9th Python in Science Conference, pp. 57–61.

Sheppard, K. (2014). *arch: Econometric analysis of volatility and correlation*. [online] Available at: https://github.com/bashtage/arch [Accessed 1 Nov. 2025].

Sirignano, J. and Cont, R. (2019). *Universal features of price formation in financial markets: perspectives from deep learning*. Quantitative Finance, 19(9), pp. 1449–1459.

Virtanen, P., Gommers, R., Oliphant, T.E., et al. (2020). *SciPy 1.0: Fundamental algorithms for scientific computing in Python*. Nature Methods, 17, pp. 261–272.

Wang, H. (2019). *VIX and volatility forecasting: A new insight*. Physica A: Statistical Mechanics and its Applications, 533, 121951.

Wang, S., Li, K., Liu, Y., Chen, Y. and Tang, X. (2024). *VIX constant maturity futures trading strategy: A walk-forward machine learning study*. PLOS ONE, 19(4), e0302289.

Whaley, R.E. (2009). *Understanding the VIX*. Journal of Portfolio Management, 35(3), pp. 98–105.

Wu, X., He, Q. and Xie, H. (2023). *Forecasting VIX with time-varying risk aversion*. International Review of Economics & Finance, 88, pp. 458–475.

Yahoo Finance (2025). *CBOE Volatility Index (ˆVIX) Historical Data*. [online] Available at: https://finance.yahoo.com/quote/%5EVIX/history [Accessed 1 Nov. 2025].

Appendix A

All analyses were performed in Python 3.x (Python Software Foundation, 2025). The following third-party libraries were used.

- **NumPy**: array programming, linear algebra utilities (Harris *et al.*, 2020).
- **pandas**: tabular data structures, I/O, time-series handling (McKinney, 2010).
- **SciPy**: scientific routines; specifically `scipy.stats` and `scipy.signal` (Virtanen *et al.*, 2020).
- **Matplotlib**: plotting (`pyplot`) and date formatting (`matplotlib.dates`) (Hunter, 2007).
- **scikit-learn**: preprocessing (`StandardScaler`), linear models (`LinearRegression`), ensembles (`RandomForestRegressor`), neural nets (`MLPRegressor`), and metrics (`mean_squared_error`, `mean_absolute_error`, `r2_score`) (Pedregosa *et al.*, 2011).
- **XGBoost**: gradient boosting regressor (`xgboost.XGBRegressor`) (Chen and Guestrin, 2016).
- **statsmodels**: time-series modelling (`ARIMA`) and related utilities (Seabold and Perktold, 2010).
- **arch**: volatility models (`arch.univariate.arch_model`) for GARCH-type specifications (Sheppard, 2014).
- **yfinance**: historical market data downloader used for VIX and related series (Aroussi, 2015).
- **OpenAI**: Machine-translation assistance for the English text was provided by ChatGPT. All outputs were reviewed and edited by the author (2025).