Digital Finance (2026) 8:63
https://doi.org/10.1007/s42521-026-00229-9

RESEARCH

Regime-dependent bitcoin volatility forecasting under alternative information sets: evidence from econometric, deep learning, and hybrid models

Ahmed S. Wafi[1,2,3] · **Sherif El-Halaby**[4,5] · **Sameh Abd-ElMaksoud Aboul-Dahab**[6,7]

Received: 31 March 2026 / Accepted: 12 October 2026
© The Author(s) 2026

Abstract

Bitcoin volatility is challenging to forecast because accuracy depends on model choice, information sets, and market regimes. Existing studies often compare models under inconsistent volatility targets, predictors, samples, or evaluation designs, leaving unclear whether forecast gains arise from model architecture, information content, or regime shifts. This study addresses this gap by proposing a unified rolling out-of-sample framework for daily Bitcoin volatility forecasting that compares GARCH-type models, standalone LSTMs, and state-augmented GARCH–LSTM hybrids under the same volatility target, nested information sets, one-step-ahead forecast horizon, and evaluation design, using balanced daily data from 1 February 2018 to 19 July 2025. Predictors are organized into technical (T), technical plus sentiment (TS), and technical, sentiment, and blockchain (TSB) groups to measure the marginal value of added information. Forecast accuracy is assessed mainly with QLIKE, plus RMSE, MAE, and Diebold–Mariano tests across normal and crisis regimes. Predictability is strongly regime-dependent. In normal markets, LSTM-T achieves the lowest loss, with TGARCH and Hybrid_GARCH_T statistically indistinguishable from it. In crises, hybrids minimize QLIKE, but the best specification differs: Hybrid_TGARCH_TS during COVID-19, Hybrid_EGARCH_TSB during the China mining ban, and Hybrid_APARCH_TS during the crypto winter. Hence, hybrid gains are regime-specific and depend on the interaction of volatility dynamics, information sets, and regime characteristics. For daily one-step-ahead Bitcoin forecasts, the evidence supports regime-sensitive modeling and cautions against broad claims of model superiority based on pooled samples.

Keywords Bitcoin volatility · Cryptocurrency forecasting · Regime-dependent forecasting · GARCH · LSTM · Hybrid forecasting models · Alternative information sets · Diebold–Mariano test

Extended author information available on the last page of the article

Published online: 26 October 2026

JEL Classification C22 · C53 · C58 · G12 · G17

1 Introduction

Volatility forecasting is central to modern finance, underpinning risk management, derivative pricing, portfolio allocation, and capital regulation. Since Engle (1982) and Bollerslev (1986), GARCH-type models—especially simple GARCH(1,1) specifications—have been the benchmark for time-varying volatility and remain hard to beat out of sample (Hansen & Lunde, 2005). The rapid growth of cryptocurrency markets, particularly Bitcoin, poses new challenges: returns show strong volatility clustering, heavy tails, and abrupt regime shifts (Katsiampa, 2017), indicating pronounced conditional heteroskedasticity and nonlinear dynamics. Simultaneously, these markets offer rich alternative data—such as sentiment indicators and blockchain activity—suggesting that nonlinear machine learning methods may improve volatility forecasts beyond traditional econometric models.

Alongside advances in econometric volatility modeling, machine learning, and deep learning methods have gained prominence in financial forecasting. Long Short-Term Memory (LSTM) networks (Hochreiter & Schmidhuber, 1997) capture nonlinear temporal dependencies and show strong predictive performance in finance (Fischer & Krauss, 2018; Gu et al., 2020). Evidence from cryptocurrency markets underscores both their potential and limits. Some studies report that machine learning methods outperform standard econometric benchmarks for Bitcoin volatility (Huang et al., 2024). Still, performance is highly sensitive to volatility proxies, model specification, predictor sets, and evaluation criteria (Dudek et al., 2024). Related work stresses the need for systematic comparison frameworks that evaluate econometric, machine learning, and hybrid approaches under consistent conditions, while recent studies in digital finance emphasize that cryptocurrency forecasting performance depends strongly on market regimes, structural conditions, and the distinction between internal and external drivers of volatility predictability (Agakishiev et al., 2025; Kayal & Dutta, 2024; Wang et al., 2023).

Despite rapid growth in this literature, key gaps remain. First, few studies compare econometric, deep learning, and hybrid volatility models within a unified, information-consistent framework; inconsistent volatility definitions, predictors, and evaluation methods confound comparisons and mask the true sources of predictive gains (Zahid et al., 2022; Dudek et al., 2024; Feng et al., 2024). Second, although hybrid methods are increasingly popular, they usually rely on forecast combinations or preprocessing rather than embedding structured conditional volatility dynamics directly into nonlinear architectures. Third, cryptocurrency volatility is inherently regime-dependent, yet most studies pool samples and do not assess how forecasting performance varies across regimes, such as normal versus crisis periods, despite growing evidence that structural breaks, regime shifts, and time-varying dynamics materially shape cryptocurrency volatility (Feng et al., 2024; Kayal & Dutta, 2024; Agakishiev et al., 2025).

Unlike prior work centered primarily on model comparison, this study highlights how information structure and market regimes jointly shape volatility forecast perfor-

mance in cryptocurrency markets. It proposes a unified out-of-sample framework for Bitcoin volatility that compares traditional econometric models, deep learning architectures, and hybrid econometric–deep learning models under consistent volatility definitions, nested information sets, and rolling evaluations. Predictors are grouped into three tiers—technical indicators (T), technical plus sentiment variables (TS), and technical, sentiment, and blockchain variables (TSB)—to measure the incremental value of each information source. Forecasts are assessed using the QLIKE loss function, supplemented by RMSE, MAE, and Diebold–Mariano tests, in both normal and crisis regimes (Patton, 2011).

This study advances the literature in three ways. First, prior comparisons of econometric and machine learning models for forecasting cryptocurrency volatility use inconsistent volatility measures, predictors, and evaluation schemes, limiting inferences about relative performance (Zahid et al., 2022; Dudek et al., 2024; Feng et al., 2024). We address this with a unified framework that benchmarks econometric, deep learning, and hybrid models under consistent volatility definitions, nested information sets, and rolling out-of-sample evaluations. Second, although recent work recognizes that cryptocurrency volatility is regime-dependent (Kayal & Dutta, 2024; Agakishiev et al., 2025), existing evidence is fragmented and rarely separates the roles of model design, information content, and market conditions. We close this gap by comparing forecasting accuracy across normal and crisis regimes. Third, while hybrid forecasting techniques are well established (Zhang, 2003; Khashei & Bijari, 2011) and used for cryptocurrency volatility (Zahid et al., 2022), they mostly focus on forecast combinations or data preprocessing rather than embedding conditional volatility dynamics in a nonlinear learning model. We instead propose a state-augmented hybrid econometric–deep learning approach that integrates structured volatility persistence with nonlinear learning in a single framework.

The study addresses three core questions: (1) How does volatility forecast accuracy compare across econometric, deep learning, and hybrid models within a common framework? (2) What incremental predictive power do alternative information sets provide in cryptocurrency markets? (3) How does forecasting performance vary across market regimes, particularly between normal and crisis periods? These questions evaluate whether improvements in cryptocurrency volatility forecasting arise mainly from model design, information content, or market conditions, offering an integrated view of method performance across informational settings and volatility environments. The paper is organized as follows. Section 2 reviews literature on volatility forecasting, machine learning, and hybrid methods. Section 3 describes the data, models, and forecasting framework. Section 4 presents empirical results and compares performance across models, information sets, and market regimes. Section 5 interprets the findings, discusses implications, and acknowledges limitations. Section 6 concludes.

2 Related literature review

For over forty years, volatility forecasting has been central in forecasting research. Conditional heteroskedasticity models, introduced by Engle (1982) and extended by Bollerslev (1986), transformed empirical modeling of time-varying variance. Since then, volatility forecasting has followed two main paths: structured econometric models and data-driven machine learning methods. The rapid growth of cryptocurrency markets now offers a new setting where both approaches are widely applied and compared.

2.1 Volatility dynamics in cryptocurrency markets

Cryptocurrency returns exhibit standard financial econometric stylized facts—volatility clustering, heavy tails, excess kurtosis, and asymmetric shock responses (Cont, 2001). Bitcoin's volatility is particularly pronounced, with persistent variance, sharp regime shifts, and complex nonlinear dynamics (Katsiampa, 2017). Recent evidence further shows that cryptocurrency volatility exhibits rough behavior and jump dependence, reinforcing the need for flexible models that can capture nonlinear volatility adjustment (Krain et al., 2025). Extreme events such as regulatory shocks, exchange failures, and liquidity cascades highlight the importance of models that account for structural breaks and nonlinear variance dynamics. Recent evidence on digital assets also shows that shocks may dissipate at different speeds across decentralized markets, reinforcing the importance of modeling adjustment dynamics and volatility transmission in crypto-asset environments (Cheah et al., 2026). These features affect forecasting: Hansen and Lunde (2005) show that volatility models must be evaluated strictly out of sample because in-sample fit is often misleading, while Patton (2011) notes that volatility is latent and must be assessed with loss functions suited to noisy proxies. This is especially important in cryptocurrency markets, where volatility regimes shift abruptly and rigorous evaluation is essential.

2.2 Econometric conditional volatility models

The ARCH model of Engle (1982) and its GARCH extension by Bollerslev (1986) remain benchmark tools for volatility forecasting because they are parsimonious, interpretable, and empirically robust. Later variants add asymmetry and nonlinear scale transformations to capture different reactions to positive and negative shocks: GJR-GARCH (Glosten et al., 1993) introduces leverage effects; EGARCH (Nelson, 1991) models log variance and relaxes positivity constraints; and APARCH (Ding et al., 1993) allows flexible power transformations of volatility. In cryptocurrency markets, GARCH-family models often yield competitive forecasts, particularly under volatility-consistent loss functions (Katsiampa, 2017). Even with newer methods, Hansen and Lunde (2005) found little that systematically outperforms GARCH(1,1), reinforcing its status as a durable benchmark, including in highly volatile environments.

2.3 Machine learning and deep learning in financial forecasting

Parallel to advances in econometrics, machine learning has become central to forecasting. Neural networks approximate nonlinear functions without strict parametric assumptions, and combining statistical models with deep learning can improve forecasts in complex time-series settings (Teräsvirta et al., 2005), while recent evidence from cryptocurrency markets highlights that volatility dynamics evolve across distinct regimes, reinforcing the need for regime-sensitive modeling approaches (Kayal & Dutta, 2024). Long Short-Term Memory (LSTM) networks (Hochreiter & Schmidhuber, 1997) are well suited to financial sequences, as their gated memory mitigates vanishing gradients and captures long-range dependencies. Fischer and Krauss (2018) show that LSTMs capture nonlinear temporal patterns missed by linear models. Recent cryptocurrency studies reinforce this evidence: Huang et al. (2024) find that machine learning methods, including LSTMs, can outperform standard benchmarks in Bitcoin volatility forecasting, while Dudek et al. (2024) show that the relative performance of econometric and machine learning models—including HAR, GARCH, LASSO, SVR, random forests, and LSTM—depends on the forecast horizon, asset, and evaluation metric.

Recent evidence indicates that machine learning methods can outperform traditional econometric models in cryptocurrency volatility forecasting under specific settings, although accuracy remains sensitive to model design, volatility proxies, training procedures, and evaluation criteria (Feng et al., 2024). Consistent with this broader evidence, Huang et al. (2026) show that deep learning can improve cryptocurrency risk modelling and portfolio optimisation when the learning objective is aligned with downside-risk measures. Other ML-based approaches also perform well in Bitcoin markets, reflecting a broader shift toward data-driven volatility modelling (Zournatzidou et al., 2024). Taken together, these studies suggest that nonlinear ML architectures are promising for capturing complex cryptocurrency dynamics, but their performance depends critically on model specification, volatility target construction, predictor selection, and evaluation design. Rigorous benchmarking with consistent information sets and forecasting protocols is therefore essential when comparing econometric and ML approaches.

2.4 Hybrid econometric–machine learning approaches

Hybrid forecasting models combine the structural strengths of econometric models with the flexibility of neural networks. Zhang (2003) and Khashei and Bijari (2011) detail how to merge linear dynamics with nonlinear learning. In volatility settings, Kristjanpoller and Minutolo (2016) show that ANN–GARCH models improve forecasts by capturing nonlinearities beyond standard variance recursions. Digital asset studies echo this: Zahid et al. (2022) find that hybrid GARCH–machine learning models enhance Bitcoin volatility forecasts, especially for realized volatility across horizons. More broadly, Chung et al. (2025) show that the relative performance of GARCH and deep learning models shifts under structural breaks, highlighting the importance of turbulent-market evaluation over pooled samples. These studies support hybridization but do not test whether directly embedding a structured condi-

tional volatility state in an LSTM improves daily one-step-ahead Bitcoin volatility forecasts under the same target, information set, and rolling evaluation protocol—the methodological focus here.

2.5 Multi-source information in cryptocurrency forecasting

Cryptocurrency markets convey more than price information. Technical indicators summarize short-term momentum, dispersion, and trading pressure, and improve predictive performance in linear and nonlinear models (Neely et al., 2014; Fischer & Krauss, 2018). Behavioral and sentiment variables also affect volatility: investor sentiment influences asset pricing (Baker & Wurgler, 2007; Tetlock, 2007), and recent cryptocurrency research shows that news and social media sentiment enhance volatility forecasts, especially in nonlinear models.

Blockchain fundamentals add a crypto-specific information layer. Network activity, transaction intensity, and adoption metrics are linked to valuation and risk (Cong et al., 2021; Liu & Tsyvinski, 2021), with predictive power that varies across market conditions (Feng et al., 2024). Thus, the informational value of blockchain variables depends on the volatility regime. Uncertainty-related variables add another dimension: measures of economic and financial uncertainty improve volatility forecasts in a state-dependent way (Fang et al., 2019; Mokni, 2021). Evidence from cryptocurrency markets shows that external uncertainty factors also significantly affect Bitcoin dynamics (Shaikh, 2020), underscoring the need to consider broader information alongside internal market indicators.

The literature identifies many potential predictors of cryptocurrency volatility, but their effectiveness is conditional. Recent work finds that internal market variables—such as lagged volatility and trading-based indicators—often dominate predictive performance, while the value of other information sources is context-dependent (Wang et al., 2023). Predictor importance is also time-varying, with some variables becoming more informative in specific market regimes (Feng et al., 2024). Recent evidence on cross-cryptocurrency realized volatility and dynamic volatility connectedness further shows that volatility predictability depends on both asset-specific information and spillover effects across crypto markets (Korkusuz & Sahiner, 2025). Although recent studies examine regime-dependent volatility and forecasting in cryptocurrency markets (Kayal & Dutta, 2024; Agakishiev et al., 2025), they typically emphasize particular models or datasets rather than offering a unified comparison across econometric, machine learning, and hybrid frameworks. Key gaps therefore remain: studies rarely compare these models within a common, information-consistent setting, nor do they systematically assess how different information sets shape volatility forecasts across regimes, especially under stress. As a result, it is unclear whether predictive gains stem from model structure, information content, or changing market conditions.

This study develops a unified forecasting framework that compares econometric, deep learning, and hybrid models across different information sets and market regimes. It tests whether combining structured volatility dynamics with nonlinear learning improves out-of-sample Bitcoin volatility forecasts and how performance changes with information structures and market conditions.

3 Data and methodology

Rather than treating forecasting performance as a pure model-selection exercise, the empirical design is structured to identify how information sets, market regimes, and hybrid model architecture jointly shape volatility forecasting outcomes. This study proposes a conditional volatility forecasting framework to compare econometric, deep learning, and hybrid models under a unified information set. The goal is not only to estimate return variability, but to test whether model structure materially affects out-of-sample accuracy across different cryptocurrency market regimes. Bitcoin returns display well-known stylized facts—volatility clustering, heavy tails, asymmetric shock responses, and regime shifts (Cont, 2001; Bollerslev, 1986)—implying time-varying conditional variance and motivating conditional heteroskedasticity models as benchmarks. In parallel, nonlinear sequence-learning models such as Long Short-Term Memory (LSTM) networks can capture complex temporal interactions and nonlinear dependencies that may exceed the flexibility of parametric variance specifications (Hochreiter & Schmidhuber, 1997; Fischer & Krauss, 2018).

To ensure forecast comparisons reflect modeling differences rather than target or data inconsistencies, all specifications use the same conditional volatility definition and harmonized explanatory variables. Forecasts are generated strictly out of sample via rolling estimation, and performance is assessed with loss functions suited to conditional variance forecasting (Hansen & Lunde, 2005; Patton, 2011). The empirical framework proceeds as follows: first, the data construction and volatility target are defined; second, predictors are organized into nested information sets; third, econometric, deep learning, and hybrid specifications are estimated under a common forecasting design; and finally, forecast performance is evaluated across full-sample and regime-specific settings.

The empirical implementation compares econometric, LSTM, and state-augmented hybrid models under a common volatility target, nested information sets, and a consistent rolling out-of-sample design. Forecast performance is evaluated with the QLIKE loss function and Diebold–Mariano tests over the full sample and across crisis regimes. This setup allows differences in predictive performance to be assessed within a unified framework and with volatility-consistent evaluation criteria (Patton, 2011; Diebold & Mariano, 1995).

3.1 Data construction and sample design

The initial merged dataset contains 2,923 daily Bitcoin and predictor observations covering the period from 1 February 2018 to 1 February 2026. After retaining only dates for which all variables used in the analysis are jointly observed, the balanced dataset contains 2,722 complete observations spanning 1 February 2018 to 19 July 2025. All observations are historical and directly observed; no projected, forward-looking, or synthetically generated data are used. The analytical sample covers several distinct market regimes, including the 2018 post-bubble crash, the 2020 COVID-19 shock, the 2021 China mining ban, and the 2022 crypto winter associated with the LUNA–FTX collapse. These high-volatility episodes allow regime-sensitive evalu-

ation of conditional volatility forecasts rather than relying solely on pooled-sample benchmarking. Daily logarithmic returns are computed as

$$r_t = \ln\left(\frac{P_t}{P_{t-1}}\right) \tag{1}$$

Here, $_t$ denotes Bitcoin's closing price at time . Log returns are used to achieve approximate stationarity and scale invariance, in line with standard financial econometrics practice (Cont, 2001). The raw dataset combines price-based variables, technical indicators, sentiment from news and social media, and blockchain activity metrics. Because sentiment and blockchain data come from heterogeneous providers, the merged dataset is initially unbalanced, with different numbers of observations across variables.

Distributional properties are computed using all available observations for each variable and reported in Sect. 4. For model estimation and forecast comparison, however, we construct a balanced sample, so all competing models are evaluated over the same time window and information set. Formally, if $\mathcal{T}_i$ denotes the set of timestamps for which variable is observed, the balanced sample is defined as

$$\mathcal{T}^* = \bigcap {}_{i=1}^{k} \mathcal{T}_i, \tag{2}$$

Here, is the number of retained explanatory variables. The intersection ensures all models use the same observations, avoiding distortions from unequal data availability. Missing sentiment or blockchain data are neither interpolated nor forward-filled; the balanced sample includes only observed information. This conservative design avoids artificial persistence or smoothing that could bias nonlinear models. All explanatory variables are chronologically aligned to prevent look-ahead bias: at forecast origin , only information available up to is used to predict +1. This strictly recursive setup, standard in conditional volatility forecasting, guarantees genuine out-of-sample evaluation (Hansen & Lunde, 2005). The balanced dataset contains 2,722 complete daily observations. Constructing the return series removes the first observation, leaving 2,721 usable observations for model implementation. For the LSTM-based models, applying the 30-day lookback window yields 2,691 input–target sequences before model training and evaluation, while retaining all crisis subperiods examined in the analysis. This balanced-sample design is crucial for ensuring that differences in forecast performance are attributable to model structure and information content rather than unequal data availability across specifications.

3.2 Conditional volatility definition and forecast target

Volatility is not directly observable and must be inferred from the conditional second moment of returns. Let r_t denote the daily logarithmic return. The return process is modeled as

$$r_t = \mu + \epsilon_t, \tag{3}$$

where ϵ_t is a mean-zero innovation with time-varying conditional variance. Conditional heteroskedasticity models assume

$$\epsilon_t = \sigma_t z_t, \tag{4}$$

where z_t is an i.i.d. innovatio with zero mean and unit variance, and $\sigma_t^2 = \mathrm{Var}(\epsilon_t \mid \mathcal{F}_{t-1})$ denotes the conditional variance given the inormation set $\mathcal{F}_{t-1}$.

The object of interest in this study is the one-step-ahead conditional volatility,

$$\sigma_{t+1|t} = \sqrt{\mathrm{Var}(r_{t+1} \mid \mathcal{F}_t)}. \tag{5}$$

All forecasting models—econometric, LSTM-based, and hybrid—produce out-of-sample predictions of the conditional standard deviation, $\widehat{\sigma}_{t+1|t}$. Forecasts are evaluated against a conditional volatility proxy derived from the same return-generating process. This measure differs from implied volatility (e.g., VIX) and from realized volatility based on high-frequency data. It is the model-implied conditional standard deviation of daily returns, in the ARCH/GARCH tradition of Engle (1982) and Bollerslev (1986). This framework is standard when high-frequency data are unavailable or when model-based conditional variance dynamics are the main focus (Hansen & Lunde, 2005). Using a common conditional volatility target is central to the unified design of this study, as it removes an important source of asymmetry in model evaluation.

3.3 Candidate predictors and variable engineering

The forecasting framework uses predictors from three economically motivated categories: technical market indicators, sentiment measures, and blockchain activity metrics. These are chosen based on evidence that cryptocurrency volatility depends not only on past returns but also on trading dynamics, investor sentiment, and on-chain activity. Technical indicators from price dynamics capture short-horizon momentum, overbought–oversold conditions, and volatility pressure. Prior work shows that such signals add forecasting power in financial markets (Fischer & Krauss, 2018). More recent evidence also indicates that cryptocurrency volatility exhibits complex nonlinear and jump-driven dynamics, reinforcing the importance of flexible forecasting frameworks in digital asset markets (Krain et al., 2025). In particular, moving averages, the Relative Strength Index (RSI), Bollinger Band width, and Average True Range (ATR) proxy for trend persistence and volatility clustering, which are especially pronounced in Bitcoin (Katsiampa, 2017).

Sentiment variables capture narrative-driven volatility. Cryptocurrency markets are highly sensitive to information shocks and social media, and sentiment-related evidence suggests that investor attention and media tone can influence volatility and market behaviour (Baker & Wurgler, 2007; Tetlock, 2007; Wang et al., 2023). Including sentiment indicators lets the models capture expectation shifts not fully reflected in past prices. Blockchain metrics provide structural information on network usage and transactions. Variables such as transaction volume and active addresses help pre-

dict volatility and market activity (Cong et al., 2021; Liu & Tsyvinski, 2021). Unlike price-based predictors, blockchain variables reflect protocol engagement and may add forward-looking information about market stress. Because high-dimensional predictor sets can cause multicollinearity and overfitting—especially in nonlinear models—a screening procedure is applied before estimation. Pairwise correlations are computed across candidate variables, and highly collinear predictors are excluded to preserve informational diversity, following standard practice in financial machine learning (Gu et al., 2020). All competing models use the same retained predictor set, so performance differences reflect model structure rather than information-set heterogeneity. This classification provides the basis for the T, TS, and TSB information structures used throughout the empirical analysis and enables direct assessment of the incremental value of broader information environments.

3.4 Econometric conditional volatility models

To benchmark nonlinear and hybrid architectures, the study evaluates ARCH/GARCH-type conditional heteroskedasticity models. These capture key features of cryptocurrency returns—volatility clustering, persistence, heavy tails, and asymmetric shock responses—well documented in financial markets and especially pronounced in digital assets (Engle, 1982; Bollerslev, 1986; Cont, 2001). Because the goal is volatility forecasting, multiple GARCH-family variants test whether alternative treatments of persistence, asymmetry, threshold effects, and nonlinear variance transformations improve out-of-sample forecasts (Hansen & Lunde, 2005). This is crucial in cryptocurrency markets, where volatility is amplified by leverage, liquidations, and abrupt regime shifts (Katsiampa, 2017; Baur et al., 2018).

Let daily logarithmic returns be defined as

$$r_t = \ln(P_t) - \ln(P_{t-1}) \tag{6}$$

where the return process is specified as

$$r_t = \mu_t + \epsilon_t \tag{7}$$

and the innovation term satisfies

$$\epsilon_t = z_t \sigma_t, z_t \sim \ i.i.d.(0,1) \tag{8}$$

where σ_t^2 denotes the conditional variance. All econometric benchmark models are estimated using the same rolling forecasting design and generate one-step-ahead volatility forecasts $\hat{\sigma}_{t+1}$, ensuring comparability with the standalone LSTM and hybrid specifications.

The baseline specification is the standard GARCH(1,1) model (Bollerslev, 1986), defined as

$$\sigma_t^2 = \omega + \alpha \epsilon_{t-1}^2 + \beta \sigma_{t-1}^2 \tag{9}$$

with $\omega > 0$, $\alpha \geq 0$, $\beta \geq 0$, and $\alpha + \beta < 1$. This formulation captures volatility clustering and persistence and therefore serves as the conventional benchmark in volatility forecasting.

To account for highly persistent volatility, an integrated GARCH (iGARCH) specification is also considered (Engle & Bollerslev, 1986), imposing the restriction

$$\sigma_t^2 = \omega + \alpha \epsilon_{t-1}^2 + (1 - \alpha)\sigma_{t-1}^2 \tag{10}$$

so that shocks to volatility have highly persistent, potentially permanent effects. This is especially relevant in cryptocurrency markets, where prolonged stress episodes may generate near-unit-root variance dynamics.

To capture asymmetric volatility responses, the analysis further includes threshold-type GARCH specifications. In the TGARCH framework (Zakoian, 1994), conditional volatility is modeled such that negative and positive shocks may exert different effects on future volatility, reflecting the tendency of downside movements to trigger stronger volatility reactions in crash-prone markets. Similarly, the GJR-GARCH model (Glosten et al., 1993) introduces an indicator function to distinguish between negative and positive shocks:

$$\sigma_t^2 = \omega + \alpha \epsilon_{t-1}^2 + \gamma I_{t-1} \epsilon_{t-1}^2 + \beta \sigma_{t-1}^2 \tag{11}$$

where $I_{t-1} = 1\text{if } \epsilon_{t-1} < 0\text{and } 0\text{otherwise}$. A positive γ implies that negative shocks increase volatility more than positive shocks of the same magnitude. This feature is empirically important in cryptocurrency markets, where abrupt downside movements are often associated with panic selling and forced liquidations.

Additional flexibility is introduced through the APARCH model (Ding et al., 1993), which generalizes the standard GARCH structure by allowing both asymmetric responses and an estimated power transformation of the conditional scale:

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

where δ is estimated from the data. APARCH is included because it nests several standard volatility specifications and allows the data to determine the most appropriate scale transformation, which may be useful in highly nonlinear environments such as cryptocurrency markets.

Finally, the EGARCH model (Nelson, 1991) is estimated to allow for logarithmic variance dynamics and asymmetric responses to standardized shocks:

$$\ln\left(\sigma_t^2\right) = \omega + \beta \ln\left(\sigma_{t-1}^2\right) + \alpha \left| \frac{\epsilon_{t-1}}{\sigma_{t-1}} \right| + \gamma \frac{\epsilon_{t-1}}{\sigma_{t-1}} \tag{13}$$

By modeling log variance, EGARCH removes the need for non-negativity constraints and offers a flexible way to capture nonlinear volatility adjustments, especially during abrupt regime shifts with disproportionate responses to extreme shocks. Together, these econometric specifications provide structured volatility benchmarks for evaluating the standalone LSTM and hybrid econometric–deep learning models, clarifying

whether predictive gains come from richer nonlinear learning alone or from combining it with structured volatility dynamics.

3.5 LSTM architecture

To model the nonlinear temporal structure of Bitcoin's volatility, this study utilizes a stacked Long Short-Term Memory (LSTM) network. Originally proposed by Hochreiter and Schmidhuber (1997) to mitigate the vanishing gradient issue in recurrent neural networks, LSTMs have subsequently become a widely used tool for forecasting financial time series (Fischer & Krauss, 2018). In contrast to parametric conditional variance models, LSTMs do not require a specified functional form for the volatility process. Rather, they infer nonlinear relationships between historical inputs and future conditional volatility via gated memory units. Within this setup, the LSTM is configured to handle multivariate sequences comprising technical indicators, sentiment metrics, and blockchain-based variables, enabling an evaluation of whether this enriched information set enhances one-step-ahead Bitcoin volatility predictions.

3.5.1 Input representation and sequence design

At each forecast origin t, the model processes a rolling sequence of the previous 30 daily observations. The input tensor is defined as $X_t \in \mathbb{R}^{30\times 9}$ where, 30 is the sequence length and 9 is the number of explanatory variables retained after feature screening. The full dataset is thus encoded during training as a three-dimensional tensor of rolling sequences, allowing the network to capture temporal dependencies across multiple information channels.

This setup is well-suited to volatility forecasting because it preserves short-run temporal dependence and volatility clustering, key features of financial returns (Engle, 1982; Bollerslev, 1986). The 30-day lookback window, approximating one month of market activity, allows the network to capture both persistent and transitory volatility patterns while keeping sequences short enough for stable training. The same rolling-sequence procedure is applied to all information sets (T, TS, and TSB), so differences in predictive accuracy reflect informational content rather than input construction.

3.5.2 LSTM dynamics and network architecture

For each time step within the input sequence, the LSTM cell updates its internal state according to the standard gated memory mechanism (Hochreiter & Schmidhuber, 1997):

Forget gate

$$f_t = \sigma\left(W_f\left[h_{t-1}, x_t\right] + b_f\right) \tag{14}$$

Input gate

$$i_t = \sigma\left(W_i\left[h_{t-1}, x_t\right] + b_i\right) \tag{15}$$

Candidate memory

$$\widetilde{C}_t = \tanh\left(W_C[h_{t-1}, x_t] + b_C\right) \tag{16}$$

Cell state update

$$C_t = f_t \odot C_{t-1} + i_t \odot \widetilde{C}_t \tag{17}$$

Output gate

$$o_t = \sigma\left(W_o[h_{t-1}, x_t] + b_o\right) \tag{18}$$

Hidden state

$$h_t = o_t \odot \tanh\left(C_t\right) \tag{19}$$

where $\sigma(\cdot)$ denotes the logistic sigmoid function, $\odot$ denotes element-wise multiplication, W represents trainable weight matrices, and b denotes bias vectors. The hidden state h_t and cell state C_t allow the network to preserve relevant information over time while discarding less informative or transient signals.

This gated design is well-suited to cryptocurrency volatility, where persistent clustering coexists with sudden shocks and regime shifts. The architecture uses two stacked LSTM layers with 64 and 32 hidden units. The first extracts nonlinear temporal patterns from the multivariate input, and the second compresses these into a compact latent representation. Stacked LSTMs have been shown to improve financial forecasting accuracy via hierarchical temporal feature extraction (Fischer & Krauss, 2018). Here, the model is designed to capture nonlinear interactions among technical indicators, sentiment variables, and blockchain activity without a prespecified econometric framework.

3.5.3 Training, regularization, and forecast output

To reduce overfitting and improve out-of-sample performance, we apply dropout with a rate of 0.20 after each LSTM layer, randomly switching off 20% of hidden units during training (Srivastava et al., 2014). This is vital in cryptocurrency markets, where noisy inputs, abrupt volatility, and regime shifts can make nonlinear models fit short-lived artifacts rather than stable predictive signals. The network ends with a single neuron with linear activation that outputs the one-step-ahead volatility forecast:

$$\widehat{\sigma}_{t+1} = f_{\mathrm{LSTM}}\left(X_t\right) \tag{20}$$

where $f_{\mathrm{LSTM}}(\cdot)$ denotes the nonlinear mapping learned by the stacked LSTM network. The model is trained by minimizing the Mean Squared Error (MSE) loss function:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N}\sum_{t=1}^{N}(\sigma_t - \hat{\sigma}_t)^2 \tag{21}$$

where σ_t is the observed volatility proxy and $\hat{\sigma}_t$ is the model-implied volatility. We train using MSE loss for stable gradients and suitability to continuous volatility targets. However, we compare models using the QLIKE loss, which is theoretically preferable for evaluating conditional volatility forecasts with noisy volatility measures (Patton, 2011). This separates estimation from evaluation, ensuring computational stability in fitting while aligning forecast assessment with standard practice in the volatility forecasting literature.

3.5.4 Functional role in volatility forecasting

The LSTM framework models conditional volatility as a nonlinear function of past multivariate data, allowing complex interactions among predictor channels. Formally, the volatility prediction rule is

$$\hat{\sigma}_{t+1} = f(X_{t-29}, X_{t-28}, \ldots, X_t) \tag{22}$$

.

where X_t is the multivariate information set at time t, including the selected technical, sentiment, and blockchain variables. Unlike conventional parametric volatility models that impose a specific functional form for variance dynamics, the LSTM learns this relationship directly from the data.

From an econometric standpoint, this architecture is valuable because it captures nonlinear interactions and higher-order dependencies that are difficult to express in standard variance models. For example, the predictive effect of a technical volatility proxy may depend on contemporaneous sentiment or recent changes in blockchain activity. The stacked LSTM learns such conditional relationships through hierarchical temporal feature extraction. Its gating structure also separates persistent volatility from short-lived noise, which is crucial in cryptocurrency markets where volatility clustering coexists with abrupt regime shifts. Here, the LSTM serves as a standalone nonlinear benchmark for evaluating the econometric and hybrid models. Its inclusion clarifies whether nonlinear learning alone improves volatility forecasts, or whether gains require combining nonlinear flexibility with explicit conditional volatility modeling.

3.6 Hybrid volatility modeling framework

3.6.1 Motivation for hybridization

Volatility forecasting models fall into two main categories: structured parametric econometric models and flexible nonlinear machine learning methods. GARCH-type parametric models explicitly capture conditional heteroskedasticity via autoregressive variance equations (Bollerslev, 1986), offering interpretability and parsimoni-

ous descriptions of volatility persistence, leverage effects, and conditional variance dynamics. Yet their fixed functional forms can miss higher-order nonlinearities, interaction effects, and structural breaks. In contrast, neural network architectures such as LSTMs approximate volatility as a flexible nonlinear function of past multivariate information, imposing minimal structural assumptions and capturing complex cross-variable interactions and regime-dependent dynamics (Fischer & Krauss, 2018). However, purely data-driven models may not explicitly represent the well-documented autoregressive volatility clustering. Hybrid forecasting frameworks integrate these strengths by letting econometric models capture linear or structured components of volatility, while neural networks model the remaining nonlinear structure. This follows the hybrid time-series literature of Zhang (2003) and Khashei and Bijari (2011), who show that combining linear and nonlinear models can improve forecasts when the data-generating process has both components.

3.6.2 General hybrid formulation

Let σ_{t+1} denote one-step-ahead conditional volatility. Following the hybrid forecasting literature (Zhang, 2003; Khashei & Bijari, 2011), a general hybrid volatility forecasting structure can be expressed as

$$\hat{\sigma}_{t+1}^{Hybrid} = \mathcal{H}\left(\hat{\sigma}_{t+1}^{P}, \hat{\sigma}_{t+1}^{N}, \mathcal{I}_t\right), \tag{23}$$

where:

- $\hat{\sigma}_{t+1}^{P}$ represents the forecast generated by a parametric volatility model,
- $\hat{\sigma}_{t+1}^{N}$ represents the forecast generated by a nonlinear neural network model,
- $\mathcal{I}_t$ denotes the information set available at time t,
- $\mathcal{H}(\cdot)$ denotes a combining or integration function.

This formulation does not require a specific hybrid structure; it defines a general integration framework. The combining function can implement additive residual modeling, feature augmentation, or weighted forecast combination. More broadly, it aligns with the forecast combination literature beginning with Bates and Granger (1969), which shows that combining forecasts from different models can improve accuracy when models capture distinct components of the underlying process.

3.6.3 State-augmented hybrid volatility model

To combine structured econometric volatility dynamics with nonlinear multivariate learning, the study uses a state-augmented hybrid model where the conditional variance from a parametric GARCH-type model is fed into the LSTM network as an additional structured input.

Let r_t denote log returns and let the parametric volatility model be defined as:

$$r_t = \mu + \epsilon_t, \epsilon_t = \sigma_t z_t, \tag{24}$$

$$\sigma_t^2 = \omega + \sum_{i=1}^{p} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2 + (\text{asymmetry terms, if applicable}). \quad (25)$$

Denote the one-step-ahead parametric volatility forecast as $\hat{\sigma}_{t+1}^{P}$. The hybrid information set is then defined as:

$$\mathcal{I}_t^{Hybrid} = \left\{ \hat{\sigma}_t^{P}, X_t \right\}, \quad (26)$$

where $X_t$represents the multivariate tensor of explanatory variables (technical, sentiment, and blockchain indicators) observed over the preceding 30 trading days.

The hybrid forecast is obtained through the nonlinear mapping:

$$\hat{\sigma}_{t+1}^{Hybrid} = \mathcal{F}(\hat{\sigma}_t^{P}, X_t), \quad (27)$$

where $\mathcal{F}(\cdot)$denotes the stacked LSTM architecture defined in Sect. 3.5.

This specification preserves the autoregressive persistence captured by the parametric variance recursion while letting the LSTM model make nonlinear adjustments conditional on that volatility state. Unlike residual-based hybridization (Zhang, 2003), the state-augmented approach embeds the parametric conditional variance directly in the nonlinear learning mechanism. This keeps volatility clustering—well-documented in financial markets (Engle, 1982)—explicit, while higher-order nonlinear interactions among heterogeneous predictors are modeled flexibly.

The hybrid structure is well-suited to cryptocurrency markets, where persistent volatility coincides with nonlinear regime shifts and interactions across market microstructure indicators, behavioral sentiment, and blockchain activity. By conditioning nonlinear learning on a structured volatility state, the model combines econometric interpretability with machine-learning flexibility. It is empirically compared with standalone parametric and standalone nonlinear models, and its relative forecasting performance is assessed using out-of-sample loss, following standard volatility forecast evaluation methods (Patton, 2011). This hybrid specification represents the methodological core of the study, as it allows structured volatility persistence and nonlinear information processing to be evaluated jointly within a unified forecasting framework.

3.7 Forecasting design

Volatility forecasts are generated in a strictly out-of-sample framework so that model comparisons reflect predictive performance rather than in-sample fit. The sample has daily observations; the first form the initial estimation window, and forecasts are evaluated over the remaining − observations. Using a rolling window, at each forecast origin model parameters are estimated with data up to , and a one-step-ahead conditional volatility forecast $\hat{\sigma}_{t+1|t}^{2}$ is produced. The window then moves forward one observation, and the procedure repeats until the end of the sample. This ensures forecasts use only information available at prediction time and allows parameters to

adapt to changing market conditions. Such rolling evaluations are standard in volatility studies (Hansen & Lunde, 2005; Patton, 2011).

The horizon is restricted to one day ahead, appropriate for short-horizon volatility modeling in fast-moving cryptocurrency markets. To assess model stability across market environments, the out-of-sample period is split into normal and crisis regimes, where crisis windows are defined exogenously by major disruption events. Forecast performance is evaluated for the full out-of-sample period and within each regime, without altering the estimation procedure. This design allows the same forecasting architecture to be evaluated not only in aggregate but also under distinct volatility environments, which is central to the regime-dependent perspective of the study.

3.8 Forecast evaluation and statistical testing

Forecast accuracy is evaluated within a single loss-based framework applied uniformly to all econometric, deep learning, and hybrid models. Because conditional volatility is latent, forecast quality is assessed against an observable proxy rather than the true variance process. Following the volatility forecasting literature, the study uses the same volatility proxy and evaluation criteria for all models to ensure comparability and internal coherence (Hansen & Lunde, 2005; Patton, 2011). The framework combines a volatility-robust primary loss function, additional scale-based error metrics, and formal statistical tests of predictive performance.

Let $_t$ denote the benchmark volatility proxy used in forecast evaluation, and $\hat{}_t$ the one-step-ahead forecast from a competing model. Because the true latent volatility is unobservable, empirical forecasting relies on a measurable proxy for conditional variance. All models in this paper are evaluated against the same proxy so that differences in predictive performance reflect model specification, information sets, and regime dependence, rather than variation in the target. A common benchmark is essential in comparative forecasting, as heterogeneous volatility definitions can bias model rankings and obscure the true sources of forecast gains (Patton, 2011). Forecast accuracy is measured mainly with the QLIKE loss function and, for robustness, with RMSE and MAE, applied over the full out-of-sample period and within each market regime. The primary evaluation criterion is the QLIKE loss function:

$$L_t^{\text{QLIKE}} = \frac{\sigma_t^2}{\hat{\sigma}_t^2} - \log\left(\frac{\sigma_t^2}{\hat{\sigma}_t^2}\right) - 1 \tag{28}$$

QLIKE is well-suited for evaluating conditional volatility forecasts because it remains reliable with noisy or error-prone volatility proxies and provides a coherent basis for comparing forecasts (Patton, 2011). Unlike traditional scale-based error measures, it evaluates forecasts directly in terms of variance prediction rather than simple numerical closeness. Accordingly, QLIKE is used as the main model ranking criterion in the empirical study, with RMSE reported as a complement.

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(\sigma_t - \hat{\sigma}_t)^2} \tag{29}$$

where is the number of out-of-sample observations. RMSE gives more weight to large forecast errors and provides a useful, scale-dependent measure of predictive accuracy. We also assess forecast performance using the MAE:

$$MAE = \frac{1}{N}\sum_{t=1}^{N}|\sigma_t - \hat{\sigma}_t| \tag{30}$$

MAE applies a linear penalty to forecast errors and is therefore less sensitive to extreme deviations than RMSE. Here, RMSE and MAE serve as complementary diagnostics rather than primary model selection criteria. They provide additional insight into forecast performance and confirm that the QLIKE-based ranking aligns with conventional scale-dependent error measures.

To assess whether differences in forecasting performance between competing models are statistically significant, we conduct pairwise comparisons using the Diebold–Mariano (DM) test (Diebold & Mariano, 1995). This test evaluates the null hypothesis that two models have equal predictive accuracy over the out-of-sample period. Let, $L_{1,t}$. and, $L_{2,t}$. denote the QLIKE-based forecast losses of the two models at time . The loss differential is defined as:

$$d_t = L_{1,t} - L_{2,t} \tag{31}$$

The null hypothesis of equal predictive accuracy is:

$$H_0 : E(d_t) = 0 \tag{32}$$

The DM test statistic is computed as:

$$DM = \frac{\bar{d}}{\sqrt{\widehat{Var}\left(\bar{d}\right)}} \tag{33}$$

Here, d is the sample mean of the loss differential, and $\widehat{Var}\left(\bar{d}\right)$ is a heteroskedasticity- and autocorrelation-consistent (HAC) estimate of its long-run variance. Under the null hypothesis, the test statistic converges to a standard normal distribution. In this study, the DM test evaluates differences in forecast accuracy across model classes and market regimes, complementing the loss-based ranking with formal statistical inference. This is important because small loss differences may be economically or statistically irrelevant unless shown to persist over time.

Overall, the evaluation framework provides a thorough, statistically robust basis for comparing volatility forecasting models. By combining QLIKE-based rankings, scale-oriented diagnostics, and formal predictive accuracy tests, the study ensures that model comparisons are both informative and statistically sound.

4 Empirical results and forecast evaluation

The empirical analysis assesses how volatility forecasting performance varies across information sets, market regimes, and model architectures within a unified framework. Instead of a purely model-based comparison, it examines how econometric, deep learning, and hybrid models respond to different information structures under consistent forecasting protocols, ensuring that performance differences stem from model structure and information content rather than data, predictor, or evaluation inconsistencies.

4.1 Data, target variable, and forecasting design

The initial merged dataset contains 2,923 daily observations covering the period from 1 February 2018 to 1 February 2026, while the balanced dataset used for model estimation and forecast evaluation contains 2,722 complete observations spanning 1 February 2018 to 19 July 2025. Bitcoin is well suited for evaluating volatility forecasting models given its strong volatility clustering, frequent regime shifts, and sensitivity to behavioral and network-driven factors (Katsiampa, 2017). The target is one-step-ahead conditional volatility. Because true conditional variance is unobservable, all models are evaluated against a common daily volatility proxy based on realized returns. Using a single proxy across specifications ensures that performance differences reflect model characteristics rather than differences in target definition. The study therefore focuses on conditional volatility forecasts rather than realized or implied volatility (Andersen et al., 2005). The consolidated dataset includes price-based, technical, sentiment, and blockchain variables. As sentiment and blockchain data come from heterogeneous sources, the number of observations differs by variable. Descriptive statistics are therefore computed using all available historical observations for each series, while model estimation and forecast evaluation are based only on the balanced sample with jointly observed variables (see Table S1 in the Supplementary Material).

For model estimation and forecast evaluation, we use the balanced dataset containing only dates for which the volatility proxy and all selected explanatory variables are jointly available, spanning 1 February 2018 to 19 July 2025. This ensures comparable models and information sets and avoids distortions from missing or infrequently recorded variables near the sample end. Forecast evaluation uses a strictly out-of-sample rolling window: at each origin, parameters are re-estimated using information up to time , and a one-step-ahead volatility forecast is produced for +1. This setup eliminates look-ahead bias and is standard in short-horizon volatility forecasting (Hansen & Lunde, 2005).

4.2 Descriptive statistics and volatility diagnostics

4.2.1 Descriptive statistics

Descriptive statistics for all variables are reported in Supplementary Table S1. The data display standard cryptocurrency features: volatility clustering, heavy tails,

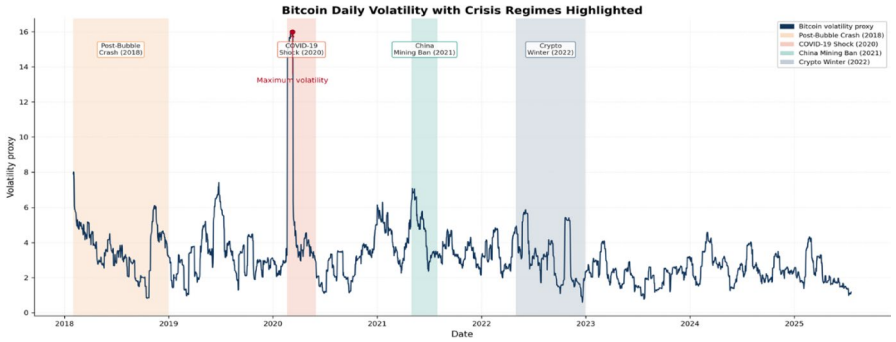


Fig. 1 Bitcoin daily volatility with crisis regimes highlighted

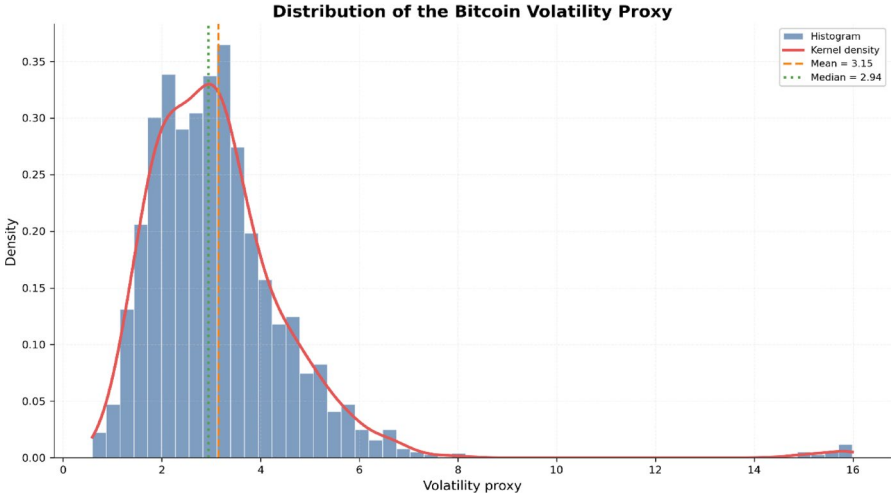


Fig. 2 Distribution of the bitcoin volatility proxy (histogram and kernel density estimate)

excess kurtosis, and non-normal returns (Cont, 2001). Jarque–Bera tests reject normality for most variables, indicating extreme observations and asymmetry. Bitcoin returns are highly volatile, negatively skewed, and strongly leptokurtic, in line with prior evidence. The volatility proxy used as the forecasting target is right-skewed with high kurtosis, reflecting frequent volatility spikes and clustering. These properties align with the volatility patterns shown in Figs. 1 and 2. Technical indicators are bounded and nonlinear, reflecting short-term trading dynamics rather than Gaussian noise. Sentiment variables are likewise non-normal, capturing behavioral effects and information-driven volatility (Tetlock, 2007). Blockchain activity is strongly right-skewed, indicating network growth and episodic transaction surges. Overall, the data exhibit heavy tails, volatility clustering, and regime-dependent variance, justifying conditional heteroskedasticity models and flexible nonlinear forecasting methods (Bollerslev, 1986; Patton, 2011).

4.2.2 Volatility dynamics and crisis identification

Figure 1 plots the daily Bitcoin volatility proxy over the full sample and highlights the main stress episodes considered in the regime-based analysis. The series displays clear volatility clustering, with large volatility shocks followed by periods of elevated volatility, which is a well-established feature of financial returns (Cont, 2001). This pattern is particularly relevant for Bitcoin, where volatility is known to be more pronounced and unstable than in many traditional asset classes (Katsiampa, 2017).

Volatility spikes cluster around major disruptions such as the 2018 post-bubble crash, the COVID-19 shock, the China mining ban, and the 2022 crypto winter, showing that Bitcoin volatility shifts sharply with market conditions. This justifies using conditional heteroskedasticity models to capture time-varying volatility and persistence (Engle, 1982; Bollerslev, 1986). Figure 1 also supports the regime-dependent forecasting design: crisis periods feature higher, more persistent volatility than normal times, so forecast accuracy should be assessed by regime rather than only on a pooled sample. This aligns with volatility forecast evaluation work stressing strict out-of-sample testing and loss functions tailored to noisy volatility proxies (Hansen & Lunde, 2005; Patton, 2011). The analysis therefore separates normal and crisis periods to test whether model performance depends on the volatility environment.

4.2.3 Volatility distribution

Figure 2 shows the unconditional distribution of the Bitcoin volatility proxy using a histogram and kernel density estimate. The distribution is strongly right-skewed with a long upper tail, indicating that extreme volatility occurs more often than under a symmetric Gaussian distribution, consistent with the heavy-tailed, non-normal behavior of financial returns (Cont, 2001).

The volatility proxy's mean exceeds its median, indicating that infrequent but economically important spikes skew the distribution to the right. This matches Fig. 1, where volatility spikes cluster in stress episodes rather than occurring uniformly over time, supporting models with time-varying volatility, such as GARCH-type and nonlinear learning models. The proxy's non-normality also matters for forecast evaluation: scale-based measures like RMSE and MAE are sensitive to extremes, especially in crises. Accordingly, the empirical analysis uses QLIKE as the primary loss function, as it is better suited to comparing volatility forecasts with noisy or imperfect proxies (Patton, 2011), while RMSE and MAE are reported only as complementary diagnostics.

4.3 Correlation analysis and feature screening

4.3.1 Correlation analysis

Figure 3 shows the Pearson correlation matrix for the candidate predictors—price-level variables, technical indicators, sentiment measures, and blockchain activity—used to detect redundant variables before model estimation. This step matters because highly collinear predictors reduce model stability and obscure whether fore-

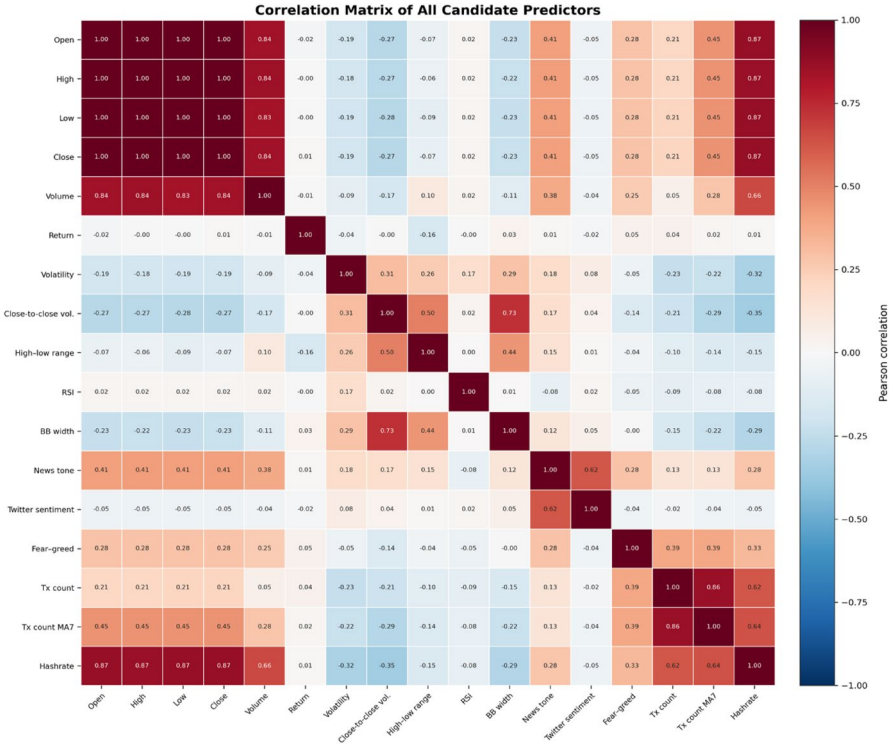


Fig. 3 Correlation matrix of candidate predictors

casting gains stem from distinct information channels or duplicated signals (Gujarati & Porter, 2009).

The strongest correlations arise among price-level variables (open, high, low, close), which all measure similar daily price levels and thus convey largely overlapping information. Including them would increase dimensionality without adding meaningful content, so they are excluded from the forecasting predictors. In contrast, return- and volatility-based transformations are retained because they better capture conditional volatility dynamics relevant for forecasting. The correlation matrix also shows moderate correlations between some volatility and blockchain variables, implying that they partly reflect common market activity or network-intensity effects. However, most technical, sentiment, and blockchain variables are not highly correlated across categories, indicating that they provide complementary information on market dynamics, investor sentiment, and on-chain activity. Keeping a parsimonious, non-redundant predictor set aligns with financial machine learning practice, where overly rich feature sets can add noise and degrade out-of-sample performance (Gu et al., 2020; James et al., 2021).

4.3.2 Selection and classification of explanatory variables

Correlation diagnostics are used to refine the predictor set, preserving informational diversity and removing redundant, highly collinear price-level variables that reflect mechanical rather than independent forecasting relationships. Among volatility indicators, retained variables are chosen to capture distinct features of market behavior—recent volatility, price-range variation, momentum, and short-term technical pressure—so the forecasting exercise focuses on economically meaningful predictors. The final predictors are arranged into three nested information sets. The technical (T) set includes only technical indicators based on price and volatility dynamics. The technical plus sentiment (TS) set adds sentiment variables to test whether behavioral and narrative information improves volatility forecasts, in line with evidence that investor sentiment influences markets (Baker & Wurgler, 2007; Tetlock, 2007). The technical plus sentiment plus blockchain (TSB) set further adds blockchain activity variables to assess whether on-chain network information provides additional predictive power in cryptocurrency markets, where network usage and adoption are economically relevant (Cong et al., 2021; Liu & Tsyvinski, 2021).

This nested information-set design lets the analysis separate gains from model architecture from gains from richer information. If forecast accuracy improves from T to TS or TSB, the added sentiment or blockchain data is informative; if performance worsens, the extra variables add noise or redundancy. This structure is central to assessing whether Bitcoin volatility predictability depends on model class, information design, or market regime.

4.4 Model classes and information sets

This section assesses normal-period out-of-sample forecasting performance for three model classes: traditional econometric volatility models, standalone LSTMs, and hybrid GARCH–LSTMs. The normal period serves as a stable first-stage benchmark, clarifying how each model behaves outside extreme stress. The goal is not to select a final regime-specific winner but to document baseline performance before turning to integrated and crisis-period comparisons. All models share the same volatility target, forecasting design, and loss functions. QLIKE is the main ranking metric, suitable when the volatility proxy is imperfect, with RMSE and MAE reported as complementary scale-based diagnostics (Patton, 2011). This setup isolates three sources of forecast performance: the econometric volatility structure, the nonlinear learning architecture, and the model's information set.

4.4.1 Traditional volatility models (econometric benchmarks)

Traditional GARCH-type models are the econometric benchmark because they explicitly capture volatility clustering, persistence, and asymmetric shock responses. The study estimates GARCH, iGARCH, TGARCH, GJR-GARCH, APARCH, and EGARCH under a common rolling out-of-sample design for direct comparison. These models serve as the baseline for evaluating whether nonlinear and hybrid

architectures improve forecasts beyond standard conditional heteroskedasticity models (Engle, 1982; Bollerslev, 1986).

Table 1 shows that TGARCH delivers the lowest QLIKE loss among traditional models in the normal period, closely followed by GARCH and GJR-GARCH. The small gap between TGARCH and GARCH indicates that threshold asymmetry offers only a modest forecasting gain, and simple GARCH dynamics remain highly relevant, consistent with evidence that parsimonious GARCH models are hard to beat in volatility forecasting (Hansen & Lunde, 2005). iGARCH underperforms TGARCH, GARCH, and GJR-GARCH, implying that highly persistent volatility dynamics are not beneficial in normal markets. APARCH and EGARCH perform worst: APARCH generates extremely large RMSE and MAE, while EGARCH yields an exceptionally large QLIKE, pointing to numerical instability and sensitivity to the loss function. Thus, added econometric flexibility does not necessarily improve out-of-sample performance. Overall, Table 1 supports TGARCH as the strongest traditional benchmark for the normal-period comparison, while simpler GARCH-type models remain competitive.

4.4.2 LSTM models

The standalone LSTM models assess whether nonlinear sequence learning improves Bitcoin volatility forecasts as information sets expand. LSTMs are suitable because their gated memory captures nonlinear temporal dependencies without a fixed parametric volatility form (Hochreiter & Schmidhuber, 1997), and prior evidence shows they learn complex financial time-series patterns (Fischer & Krauss, 2018). Three nested models are estimated: LSTM-T uses only technical indicators; LSTM-TS adds sentiment variables; and LSTM-TSB further includes blockchain activity measures. This setup tests whether richer information enhances one-step-ahead volatility forecasts or instead adds noise that degrades out-of-sample performance.

Table 2 shows that LSTM-T has the lowest QLIKE, RMSE, and MAE among standalone LSTM models, indicating that under normal market conditions, technical indicators provide the most useful short-horizon information for Bitcoin volatility forecasting. This is economically plausible because technical variables summarize recent price dynamics, volatility pressure, and market momentum, which are closely linked to short-term conditional volatility. Prior work finds that technical indicators improve financial forecasting (Neely et al., 2014), and cryptocurrency volatility stud-

Table 1 Traditional volatility models (normal period, out-of-sample)

Model	Observations	QLIKE	RMSE	MAE	Rank (QLIKE)
TGARCH	1583	0.2209	1.0038	0.8047	1
GARCH	1583	0.2218	1.0064	0.8106	2
GJR-GARCH	1583	0.2292	1.0491	0.8286	3
iGARCH	1583	0.2410	1.1596	0.9116	4
APARCH	1583	0.5465	2.2598E+30	5.6798E+28	5
EGARCH	1583	1.4943E+10	1.6868	0.8265	6

Results are for normal-period out-of-sample forecasts. Models are ranked by QLIKE, lower values indicate more accurate volatility forecasts. RMSE and MAE are reported as complementary scale-based measures. Extremely large values indicate numerical instability in the rolling forecasts

Table 2 LSTM models by information set (normal period, out-of-sample)

Model	Information set	Observations	QLIKE	RMSE	MAE	Rank (QLIKE)
LSTM-T	Technical indicators	1587	0.1934	0.8424	0.6607	1
LSTM-TS	Technical + sentiment	1587	0.2332	1.0201	0.7721	2
LSTM-TSB	Technical + sentiment + blockchain	1587	0.2468	1.2762	0.8424	3

Results are out-of-sample forecasts for standalone LSTM models in normal periods. T uses only technical indicators; TS adds sentiment variables; TSB adds blockchain variables. Models are ranked by QLIKE (lower is better). RMSE and MAE are supplementary diagnostics

ies highlight the role of internal market determinants such as lagged volatility and trading activity (Wang et al., 2023).

Adding sentiment variables in LSTM-TS worsens forecast accuracy relative to LSTM-T. This does not mean sentiment is irrelevant in cryptocurrency markets; rather, sentiment adds little incremental predictive power in normal conditions. Sentiment is more closely linked to information shocks, investor attention, and behavioral reactions, which may matter more in turbulent than in stable periods (Baker & Wurgler, 2007; Tetlock, 2007). The result is therefore consistent with the view that the predictive value of behavioral information depends on the market environment and is not constant over time.

LSTM-TSB underperforms LSTM-T, indicating that adding blockchain variables does not enhance standalone LSTM forecasts in normal periods. While blockchain activity matters for cryptocurrency valuation and market development (Cong et al., 2021; Liu & Tsyvinski, 2021), its daily short-horizon contribution to volatility forecasts appears limited in stable markets. More broadly, financial machine learning studies show that extra predictors can hurt out-of-sample performance when they add noise, redundancy, or unstable relationships instead of distinct signals (Gu et al., 2020; James et al., 2021). Thus, Table 2 indicates that nonlinear models are not automatically improved by richer information sets; additional predictors help only if they provide timely, non-redundant information for the target volatility horizon.

4.4.3 Hybrid GARCH–LSTM models

Hybrid GARCH–LSTM models feed GARCH-generated volatility states into an LSTM, combining structured conditional volatility dynamics with nonlinear sequence learning. This preserves the econometric model's ability to capture volatility persistence while enabling the LSTM to model nonlinear temporal interactions among technical, sentiment, and blockchain predictors. Such hybrids align with evidence that mixing structured time-series models with nonlinear learners improves forecasts when data exhibit both persistence and nonlinearity (Zhang, 2003; Khashei & Bijari, 2011), and with the use of LSTMs in financial forecasting, where nonlinear temporal dependencies are key (Hochreiter & Schmidhuber, 1997; Fischer & Krauss, 2018).

Table 3 shows substantial variation across hybrid models, implying that performance depends on both the econometric backbone and the information set. During the normal period, Hybrid_GARCH_T—combining a standard GARCH volatility state with technical indicators—performs best. This suggests that the most effective

Table 3 Hybrid GARCH–LSTM models (normal period, out-of-sample)

Model	Econometric backbone	Information set	Observations	QLIKE	RMSE	MAE	Rank (QLIKE)
Hybrid_GARCH_T	GARCH	T	1583	0.2062	0.8925	0.6984	1
Hybrid_APARCH_T	APARCH	T	1583	0.2078	0.9096	0.6947	2
Hybrid_iGARCH_T	iGARCH	T	1583	0.2128	0.9460	0.7230	3
Hybrid_EGARCH_T	EGARCH	T	1583	0.2133	0.9190	0.7030	4
Hybrid_EGARCH_TSB	EGARCH	TSB	1583	0.2149	0.9501	0.7095	5
Hybrid_TGARCH_TS	TGARCH	TS	1583	0.2189	1.0311	0.7562	6
Hybrid_GARCH_TS	GARCH	TS	1583	0.2207	1.0895	0.7904	7
Hybrid_GJR_GARCH_T	GJR-GARCH	T	1583	0.2222	0.9359	0.7080	8
Hybrid_iGARCH_TS	iGARCH	TS	1583	0.2238	1.1534	0.7937	9
Hybrid_APARCH_TS	APARCH	TS	1583	0.2248	1.1590	0.8145	10
Hybrid_EGARCH_TS	EGARCH	TS	1583	0.2357	1.2311	0.8257	11
Hybrid_GJR_GARCH_TSB	GJR-GARCH	TSB	1583	0.2376	1.0443	0.7578	12
Hybrid_GJR_GARCH_TS	GJR-GARCH	TS	1583	0.2589	1.1540	0.7802	13
Hybrid_APARCH_TSB	APARCH	TSB	1583	0.2615	1.1313	0.7799	14
Hybrid_TGARCH_T	TGARCH	T	1583	0.2631	1.0366	0.7746	15
Hybrid_iGARCH_TSB	iGARCH	TSB	1583	0.2677	1.2657	0.8466	16
Hybrid_GARCH_TSB	GARCH	TSB	1583	0.2761	1.3147	0.8486	17
Hybrid_TGARCH_TSB	TGARCH	TSB	1583	0.2799	1.2867	0.8524	18

Results are for normal-period out-of-sample forecasts from state-augmented GARCH–LSTM hybrid models. T, TS, and TSB are the nested information sets in Sect. 3.3. Models are ranked by QLIKE (lower is better). RMSE and MAE are additional diagnostics

normal-market hybrid is relatively parsimonious: it exploits GARCH volatility persistence while using technical indicators as the main nonlinear input. This aligns with the benchmark status of GARCH-type models in volatility forecasting (Hansen & Lunde, 2005) and with evidence that technical and internal market variables help predict financial and cryptocurrency volatility (Neely et al., 2014; Wang et al., 2023). Other technical-information hybrids (Hybrid_APARCH_T, Hybrid_iGARCH_T, Hybrid_EGARCH_T) are competitive but do not outperform Hybrid_GARCH_T in QLIKE. Thus, more flexible backbones do not necessarily improve out-of-sample performance in normal periods: added flexibility may capture extra dynamics but also raises estimation uncertainty and numerical sensitivity in rolling forecasts. Overall, under normal conditions, a standard GARCH volatility state with technical indicators offers a stable and effective hybrid structure.

The table shows that hybrids with broader information sets do not reliably beat technical-only hybrids. This matches the standalone LSTM results in Table 2: adding sentiment or blockchain variables does not inherently improve short-horizon volatility forecasts in stable periods. These variables are economically relevant, but their incremental value depends on whether they deliver timely signals about the current volatility regime (Cong et al., 2021; Liu & Tsyvinski, 2021; Wang et al., 2023). Weak or noisy additional predictors can reduce out-of-sample accuracy, a standard problem in high-dimensional financial prediction (Gu et al., 2020; James et al., 2021). Overall, Table 3 indicates that hybridization helps only when the econometric component and nonlinear information set are truly complementary. Hybrid_GARCH_T outperforms the best traditional model in Table 1, showing that adding nonlinear learning to a structured volatility state improves normal-period forecasts relative to purely econometric benchmarks. However, it does not beat LSTM-T in Table 2, so hybridization offers no unconditional advantage in calm markets. This motivates the regime-based analysis, which reassesses all model families under crisis conditions to test whether hybrids gain value when volatility is more unstable and nonlinear.

4.5 Forecast evaluation metrics and statistical inference

Forecast performance is evaluated with a loss-based framework applied consistently to all traditional, LSTM, and hybrid models. Because conditional volatility is latent, forecasts are assessed against an observable volatility proxy rather than the true variance. QLIKE is used as the primary loss function, as it is standard in volatility evaluation and well suited to noisy or imperfect proxies (Patton, 2011). It is therefore the main criterion for model ranking. RMSE and MAE are reported as complementary diagnostics: RMSE emphasizes large forecast errors and is informative in periods of extreme volatility, while MAE summarizes average absolute deviations. They are not used for ranking because scale-based errors are sensitive to the proxy's construction and magnitude, and model rankings can vary with the loss function (Hansen & Lunde, 2005). Accordingly, results are interpreted with QLIKE as the benchmark, with RMSE and MAE used to confirm the robustness of the main ranking.

Statistical inference relies on the Diebold–Mariano (DM) test, which assesses whether differences in predictive accuracy between competing forecasts are statistically significant (Diebold & Mariano, 1995). In the integrated comparison, the DM

test uses QLIKE loss differentials relative to the best benchmark in each period. A positive DM statistic indicates that the model has a higher average QLIKE loss than the benchmark. A p-value above 0.10 implies the model is not statistically worse, while smaller p-values indicate significant underperformance relative to the best model. This matters because small QLIKE differences may not be statistically meaningful, even if numerical rankings differ. Overall, the evaluation combines three elements: QLIKE-based ranking, RMSE and MAE as scale-based diagnostics, and DM tests for formal predictive-accuracy comparison. This avoids relying solely on raw loss differences and ensures that conclusions reflect both forecast accuracy and statistical significance.

4.6 Model selection under normal market conditions

Table 4 reports the normal-period cross-family comparison using the best-performing specification from each model class. Models are selected from Tables 1, 2 and 3 based on QLIKE, with RMSE and MAE as complementary diagnostics. For readability, only the family-best specifications appear in Table 4, while Supplementary Table S2 provides the full normal-period ranking of all 27 specifications and the corresponding Diebold–Mariano tests. This serves as the normal-market benchmark before examining whether rankings change in crises.

TGARCH is retained as the traditional volatility benchmark because it delivers the lowest QLIKE among the econometric models in Table 1 and, as a threshold-GARCH, captures asymmetric volatility responses to positive and negative shocks (Zakoian, 1994). Its performance, comparable to standard GARCH, underscores the

Table 4 Normal-period integrated model comparison and diebold–mariano tests

Model family	Model	Observations	QLIKE	RMSE	MAE	Overall rank	Best model in Normal	DM vs. best	*p*-value vs. best	Conclusion vs. best
Traditional	TGARCH	1583	0.2208	1.0037	0.8047	9	LSTM-T	1.0520	0.2930	Not significantly different from best
LSTM	**LSTM-T**	**1583**	**0.1934**	**0.8425**	**0.6608**	**1**	**LSTM-T**	—	—	**Best benchmark**
Hybrid	Hybrid-GARCH-T	1583	0.2062	0.8925	0.6984	2	LSTM-T	0.7941	0.4273	Not significantly different from best

Each row reports the best normal-period specification within each model family (based on QLIKE in Tables 1, 2 and 3). The DM test compares each model with the normal-period benchmark, LSTM-T, using QLIKE loss differentials. A positive DM statistic means higher average loss than the benchmark. A p-value above 0.10 indicates no statistically significant difference from the benchmark.

Boldface identifies the best-performing model for the normal period based on the lowest QLIKE value

continued relevance of parsimonious GARCH-type models as robust out-of-sample volatility benchmarks (Hansen & Lunde, 2005). LSTM-T is selected as the standalone deep learning benchmark because it achieves the lowest QLIKE, RMSE, and MAE among the LSTM models in Table 2, confirming the suitability of LSTMs for modelling nonlinear temporal dependencies in financial time series (Hochreiter & Schmidhuber, 1997; Fischer & Krauss, 2018). The dominance of the technical-information specification shows that recent market dynamics, volatility pressure, and momentum provide relevant short-horizon predictive signals, consistent with evidence on technical predictors and internal drivers of cryptocurrency volatility (Neely et al., 2014; Wang et al., 2023). Hybrid_GARCH_T is chosen as the representative hybrid specification because it attains the lowest QLIKE among the 18 hybrid models in Table 3, combining GARCH's volatility persistence with LSTM's nonlinear learning. This aligns with the hybrid forecasting literature, which finds that integrating statistical time-series structure with nonlinear learning improves predictions when data exhibit both persistence and nonlinearity (Zhang, 2003; Khashei & Bijari, 2011), and with evidence that GARCH–neural network hybrids exploit complementary econometric and nonlinear information (Kristjanpoller & Minutolo, 2016; Zahid et al., 2022).

The cross-family comparison shows that LSTM-T has the lowest numerical loss in normal periods, but Diebold–Mariano tests indicate that TGARCH and Hybrid_GARCH_T are not statistically different from LSTM-T at conventional levels. Thus, normal-period predictive differences across the leading traditional, LSTM, and hybrid models are economically small and statistically indistinguishable, motivating the broader crisis-period evaluation that follows. Because normal-market rankings are similar and crises differ in intensity, duration, and transmission, restricting stress-period analysis to normal-period winners would be too narrow. This aligns with evidence that cryptocurrency volatility is regime-dependent and sensitive to structural breaks (Kayal & Dutta, 2024; Agakishiev et al., 2025). The crisis analysis re-evaluates all 27 specifications in each regime: six traditional volatility models, three standalone LSTMs, and eighteen hybrid GARCH–LSTMs. Models are ranked by QLIKE, and the best in each family is compared with the regime-best benchmark using Diebold–Mariano tests on QLIKE loss differentials, allowing regime-specific forecast gains to be attributed to model class, information structure, and the features of each volatility episode.

4.7 Unified performance evaluation across full sample and market regimes

The previous analysis defined the normal-period benchmark and showed that traditional, LSTM, and hybrid models perform similarly in stable markets. This section evaluates crisis regimes, where volatility is unstable and model rankings may change. Rather than applying normal-period winners to all stress episodes, it re-estimates all 27 specifications for each crisis—six traditional models, three LSTMs, and eighteen GARCH–LSTM hybrids. For each regime, models are ranked by QLIKE. Table 5 reports the best specification in each family, while Supplementary Table S3 reports the full crisis-specific rankings of all 27 models and their Diebold–Mariano comparisons. These family-best models are then compared with the regime's overall best benchmark using Diebold–Mariano tests on QLIKE loss differentials. This approach

Table 5 Out of sample forecast performance and Diebold–Mariano tests across market regimes

Regime	Model family	Model	Observations	QLIKE	RMSE	MAE	Overall rank	Best model in regime	DM vs. best	*p*-value vs. best	Conclusion vs. best
COVID-19 Shock	Traditional	iGARCH	102	3.3733	6.4692	4.2240	13	Hybrid_TGARCH_TS	1.8257	0.0708	Marginally worse than best at 10%
	LSTM	LSTM-TSB	102	2.2881	4.5884	2.7851	9	Hybrid_TGARCH_TS	1.0447	0.2987	Not significantly different from best
	Hybrid	**Hybrid_TGARCH_TS**	**102**	**1.2636**	**3.9421**	**2.2698**	**1**	**Hybrid_TGARCH_TS**		—	**Best benchmark**
China Mining Ban	Traditional	GARCH	92	0.1870	1.5488	1.2220	2	Hybrid_EGARCH_TSB	0.7732	0.4414	Not significantly different from best
	LSTM	LSTM-T	92	0.4074	1.5582	1.1785	18	Hybrid_EGARCH_TSB	1.5633	0.1215	Not significantly different from best
	Hybrid	**Hybrid_EGARCH_TSB**	**92**	**0.1528**	**1.1910**	**0.9108**	**1**	**Hybrid_EGARCH_TSB**		—	**Best benchmark**
Crypto Winter	Traditional	GARCH	244	0.4686	1.6126	1.4044	22	Hybrid_APARCH_TS	3.8856	0.0001	Significantly worse than best at 1%
	LSTM	LSTM-TS	244	0.2978	1.2470	1.0744	8	Hybrid_APARCH_TS	1.0903	0.2767	Not significantly different from best
	Hybrid	**Hybrid_APARCH_TS**	**244**	**0.2638**	**1.0951**	**0.9131**	**1**	**Hybrid_APARCH_TS**		—	**Best benchmark**

For each crisis regime, all 27 model specifications are re-estimated, and the table reports the QLIKE-best model from each family. The DM test compares each family's best model with the regime's overall QLIKE-best benchmark using QLIKE loss differentials. A positive DM statistic means higher average loss than the benchmark. A p-value above 0.10 indicates no statistically significant difference from the benchmark. RMSE and MAE are reported as complementary diagnostics.

Boldface identifies the best-performing model within each crisis regime based on the lowest QLIKE value

reflects the regime-dependent nature of cryptocurrency volatility, where forecasting accuracy varies across market states and structural breaks (Kayal & Dutta, 2024; Agakishiev et al., 2025), and ensures that crisis results are based on out-of-sample performance within each episode rather than normal-market model selection. The formal crisis evaluation covers the COVID-19 shock, the China mining ban, and the 2022 crypto winter. The 2018 post-bubble crash appears in the descriptive volatility analysis but is excluded from the formal out-of-sample crisis comparison because the rolling forecast begins after the initial estimation window.

4.7.1 Regime-based statistical comparison

Table 5 reports forecast performance across the three formal out-of-sample crisis regimes.

Table 5 evaluates crisis-period forecasts after re-estimating all 27 model specifications within each stress episode. For each regime, it reports the best traditional, LSTM, and hybrid specification by QLIKE, and compares them with the regime-best benchmark using Diebold–Mariano tests. The goal is to see whether the statistically preferred specification varies with the type of volatility shock, rather than to carry over normal-period winners to crisis episodes. The results show clear regime-dependent forecast performance. In all three crisis windows, a hybrid model yields the lowest QLIKE. Yet the preferred hybrid differs by regime: Hybrid_TGARCH_TS for COVID-19, Hybrid_EGARCH_TSB for the China mining ban, and Hybrid_APARCH_TS for the crypto winter. Thus, hybrid performance is not driven by a single dominant specification; instead, predictive accuracy reflects the interaction of the volatility state, the information set, and each crisis's characteristics.

For the COVID-19 shock, the QLIKE ranking identifies Hybrid_TGARCH_TS as the best-performing specification. This is consistent with the episode's abrupt, systemic nature. The TGARCH component fits environments where volatility reacts asymmetrically to positive and negative shocks, as threshold models explicitly allow such asymmetry (Zakoian, 1994). Including sentiment is also plausible, since systemic uncertainty and rapid shifts in investor expectations increase the value of behavioural and narrative information (Baker & Wurgler, 2007; Tetlock, 2007). The traditional family-best model, iGARCH, has a much higher QLIKE and is marginally worse than the hybrid benchmark at the 10% level, suggesting persistence-based dynamics alone are inadequate when volatility adjusts abruptly. By contrast, LSTM-TSB remains statistically indistinguishable from the hybrid benchmark, indicating that richer nonlinear information sets can still perform well under extreme market uncertainty.

The China mining ban generates a distinct statistical and economic pattern. Hybrid_EGARCH_TSB attains the lowest QLIKE, but DM tests show GARCH and LSTM-T are not statistically inferior to this benchmark. Choosing Hybrid_EGARCH_TSB aligns with the crypto-specific nature of the shock, which was tied to mining activity, network conditions, and blockchain fundamentals. EGARCH is suitable because it models log variance and captures asymmetric volatility without imposing non-negativity constraints (Nelson, 1991). The TSB information set is also appropriate, as blockchain activity and network adoption are central to cryptocurrency valuation and

market dynamics (Cong et al., 2021; Liu & Tsyvinski, 2021). However, the weaker DM evidence suggests that the mining-ban shock was more localized and shorter-lived than other crises. Thus, the hybrid model's numerical QLIKE advantage reflects better performance, but not strong statistical dominance over other model families.

The crypto winter qshows that prolonged stress alters the forecasting value of model structure and information design. Hybrid_APARCH_TS achieves the lowest QLIKE, while traditional GARCH performs significantly worse at the 1% level, reflecting the regime's persistent, confidence-sensitive nature. The APARCH component fits these conditions by allowing flexible power transformations and asymmetric volatility responses, accommodating nonlinear volatility scale effects (Ding et al., 1993). The TS information set is economically relevant because extended downturns depend on both technical pressures and sentiment-driven investor confidence, consistent with evidence that sentiment shapes market behaviour, especially when information and expectations dominate (Baker & Wurgler, 2007; Tetlock, 2007). LSTM-TS remains statistically comparable to the hybrid benchmark, indicating that sentiment-enhanced nonlinear learning also performs well under sustained stress, though the APARCH-based hybrid attains the lowest loss.

Table 5 shows that the best crisis-period model depends on the crisis environment, not model class alone: abrupt systemic stress favours a threshold-based hybrid with sentiment information; a crypto-specific network disruption favours an EGARCH-based hybrid with blockchain information; and prolonged market weakness favours an APARCH-based hybrid with sentiment information. This supports the study's main claim that Bitcoin volatility forecast accuracy is jointly determined by model architecture, information structure, and market regime, and it justifies re-estimating all 27 specifications in each crisis instead of relying on the normal-period winners in Table 4. Supplementary Tables S4 and S5 further summarize the models that are statistically indistinguishable from the relevant normal-period or crisis-period benchmark, reinforcing that numerical QLIKE rankings should be interpreted alongside Diebold–Mariano test evidence.

4.7.2 Visual evidence on forecast performance across market regimes

Figures 4 and 5 show how the selected models track the timing, magnitude, and persistence of Bitcoin volatility over the out-of-sample period and the three crisis windows, complementing the statistical evidence in Tables 4 and 5. This descriptive visual analysis is interpreted alongside the QLIKE rankings and Diebold–Mariano tests, not as a substitute for formal forecast evaluation.

Figure 4 compares actual volatility with the three normal-period family representatives: TGARCH, LSTM-T, and Hybrid_GARCH_T. All three capture volatility clustering and higher volatility in stress episodes, but they differ in adjustment. TGARCH produces a smooth forecast path, reflecting GARCH persistence. LSTM-T responds more flexibly to short-run volatility changes, consistent with its nonlinear sequence-learning structure. Hybrid_GARCH_T combines GARCH persistence with nonlinear LSTM adjustment, yielding an intermediate pattern. This visual evidence matches Table 4, where LSTM-T has the lowest normal-period loss, while TGARCH

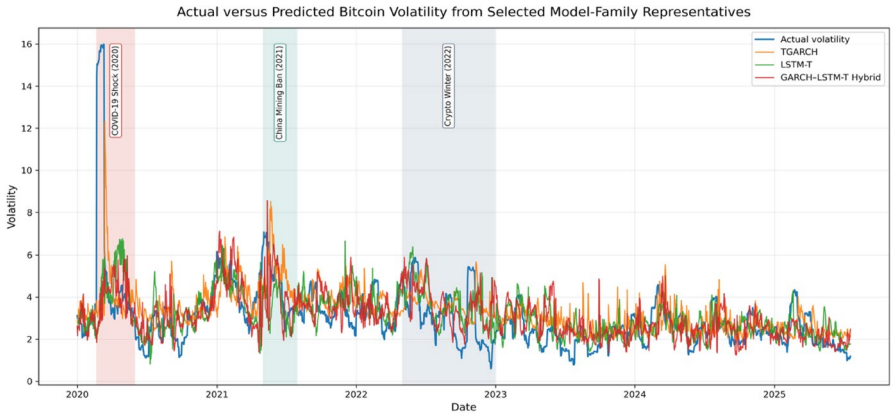


Fig. 4 Actual versus predicted volatility from the best-performing models during normal and crisis periods

and Hybrid_GARCH_T are statistically indistinguishable from the normal-period benchmark.

Figure 5A–C focus on crisis episodes and present the best-performing traditional, LSTM, and hybrid models for each crisis regime, based on the regime-specific selections in Table 5 rather than the normal-period representatives. In the COVID-19 shock, Fig. 5A shows a sudden surge in volatility to extreme levels. The hybrid model, Hybrid_TGARCH_TS, tracks this spike more closely than the traditional iGARCH benchmark, and LSTM-TSB is visually comparable. This matches Table 5, where Hybrid_TGARCH_TS has the lowest QLIKE, iGARCH is slightly worse at the 10% level, and LSTM-TSB is not statistically different from the best regime model.

Figure 5B shows a different pattern for the China mining ban: the volatility disturbance is more localized and less persistent than under COVID-19. Although Hybrid_EGARCH_TSB has the lowest QLIKE in Table 5, the visual differences among GARCH, LSTM-T, and Hybrid_EGARCH_TSB are small. This aligns with the statistical evidence that neither the traditional nor the LSTM family-best model differs significantly from the hybrid benchmark in this regime. The figure thus supports the view that the China mining ban produced weaker statistical separation across model classes.

Figure 5C most clearly illustrates regime-dependent model performance. During the crypto winter, volatility stays elevated for an extended period, posing a persistent forecasting challenge. Hybrid_APARCH_TS tracks the volatility pattern more closely than the GARCH benchmark, while LSTM-TS remains competitive. This matches Table 5, where GARCH performs significantly worse than the hybrid benchmark at the 1% level, whereas LSTM-TS is statistically indistinguishable from the regime-best model. These results indicate that prolonged stress conditions increase the value of flexible hybrid specifications, especially those combining nonlinear scale effects with sentiment-enhanced learning.

Taken together, Figs. 4 and 5 show that forecast differences vary by market state. Normal-period models perform similarly overall, while crisis-period accuracy depends on the specific volatility episode. This visual evidence supports the main



Fig. 5 **A** Actual versus predicted bitcoin volatility during COVID-19 shock.**B** Actual versus predicted bitcoin volatility during China mining ban.**C** Actual versus predicted bitcoin volatility during crypto winter

statistical result: Bitcoin volatility forecasting is regime-dependent, and the relative value of traditional, LSTM, and hybrid models changes with the intensity and persistence of market stress.

4.7.3 Time-varying loss dynamics

Figures 6 and 7 show the time variation in forecast losses using 30-day rolling QLIKE values. This complements the average loss statistics in Tables 4 and 5 by revealing whether forecast performance is stable over time or deteriorates during specific volatility episodes. As QLIKE is the primary loss function in the empirical evaluation, rolling QLIKE offers a dynamic view of model performance under changing market conditions.

Figure 6 shows rolling QLIKE losses over the full out-of-sample period for the normal-period model representatives: TGARCH, LSTM-T, and Hybrid_GARCH_T. Losses are low in calm markets but increase around stress episodes, especially during COVID-19, when abrupt volatility shifts generate large forecast errors. This indicates that cryptocurrency volatility is regime-dependent and that predictive accuracy should be evaluated across market states rather than only by pooled average losses. The models' relative performance in Fig. 6 aligns with the normal-period results in Table 4. TGARCH delivers a smoother loss path but exhibits sharper spikes when volatility changes suddenly. LSTM-T adapts more flexibly to short-run movements, yet its losses also rise in turbulent periods. Hybrid_GARCH_T shows a relatively stable loss path, reflecting persistent but nonlinearly adjusted volatility. The full-sample figure does not imply unconditional hybrid dominance, since Table 4 indicates the top normal-period models are statistically indistinguishable. Instead, Fig. 6 mainly illustrates that forecasting becomes more difficult in crisis regimes.

Figure 7A–C present crisis-specific rolling-loss results using the family-best specifications from Table 5. Figure 7A shows that the COVID-19 shock sharply increased rolling QLIKE losses, especially for the traditional iGARCH benchmark. Hybrid_TGARCH_TS delivers lower and more stable losses for most of the period, while LSTM-TSB remains competitive. This aligns with Table 5, where the hybrid model

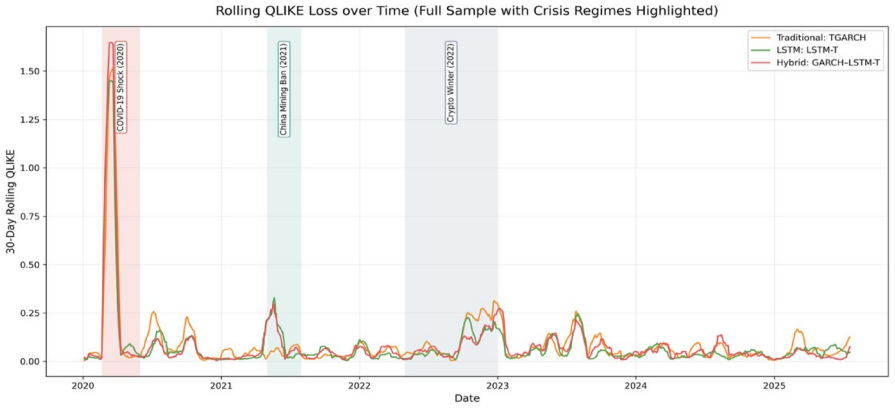
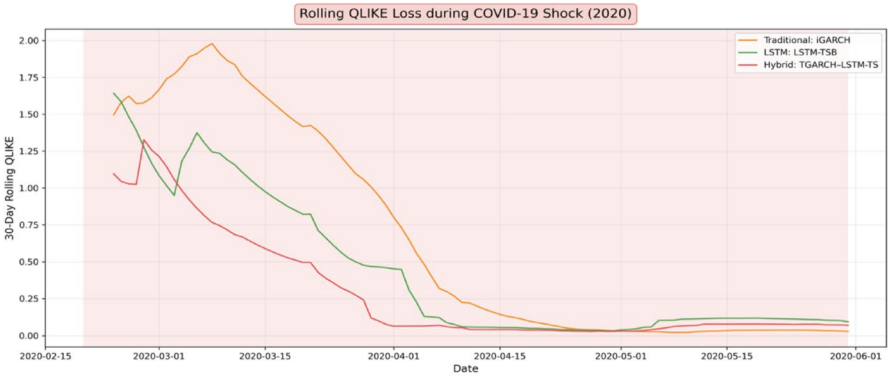
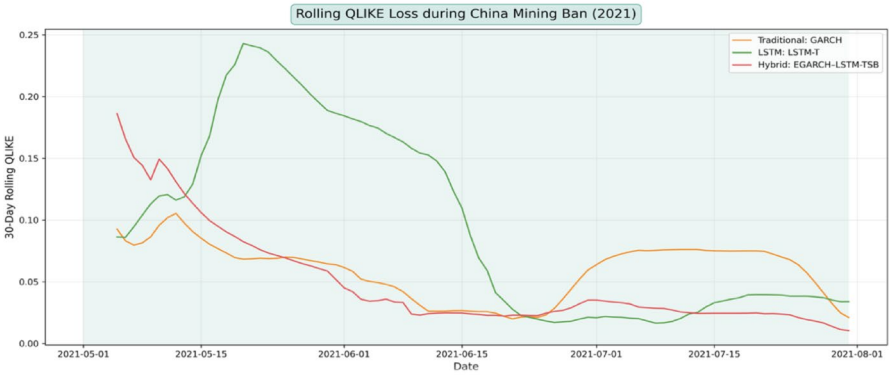


Fig. 6 Rolling QLIKE loss over time with crisis regimes highlighted



A



B



C

Fig. 7 **A** Rolling QLIKE Loss during COVID-19 shock. **B** Rolling QLIKE Loss during China mining ban.**C** Rolling QLIKE loss during crypto winter

attains the lowest QLIKE, iGARCH is slightly worse at the 10% level, and LSTM-TSB is statistically indistinguishable from the regime-best benchmark.

Figure 7B shows weaker separation across model classes during the China mining ban. Rolling losses for GARCH and Hybrid_EGARCH_TSB are similar, while LSTM-T briefly exhibits higher loss early in the episode before converging to the others. This aligns with Table 5, where Hybrid_EGARCH_TSB is numerically best but neither GARCH nor LSTM-T differs from it statistically. Thus, the China mining ban appears as a more localized volatility shock with less pronounced model differentiation.

Figure 7C shows clearer model separation during the crypto winter. Rolling losses for the traditional GARCH benchmark rise sharply late in the episode, while Hybrid_APARCH_TS maintains lower losses through most of the crisis. LSTM-TS tracks the hybrid model more closely than the benchmark, consistent with DM results indicating no statistical difference from the regime-best model. This suggests that prolonged stress increases the value of flexible volatility structures and sentiment-enhanced nonlinear learning.

The rolling-loss evidence supports a regime-dependent interpretation. Forecast accuracy varies over time with the intensity and persistence of volatility shocks. The crisis-specific figures confirm Table 5's main result: hybrid models achieve the lowest QLIKE across crisis regimes, though their advantage differs by episode—strongest in the crypto winter, moderate during COVID-19, and weakest during the China mining ban. Overall, the rolling QLIKE analysis reinforces the central claim that Bitcoin volatility forecasts depend jointly on model architecture, information structure, and market regime.

5 Discussion

The results show that Bitcoin volatility forecasting depends on the interaction between model architecture, information structure, and market regime. No single model dominates in all conditions; predictive accuracy varies with market stability or stress, the timeliness and non-redundancy of information, and the model's ability to capture persistence, asymmetry, and nonlinear adjustment. The discussion therefore: (i) compares the performance of GARCH-type models, standalone LSTMs, and state-augmented hybrids under a common rolling out-of-sample design; (ii) explains why technical, sentiment, and blockchain data have different value in normal versus crisis regimes; and (iii) examines the economic and methodological implications of regime-specific results, especially that different crises favour different hybrid models.

5.1 Main findings and interpretation

Traditional GARCH-type models remain key benchmarks for Bitcoin volatility forecasting. In normal periods, TGARCH achieves the lowest QLIKE among econometric models, with standard GARCH and GJR-GARCH close behind, showing that structured volatility persistence still explains a substantial share of short-horizon Bitcoin volatility despite more flexible machine learning methods. This aligns with

the established success of GARCH models in volatility forecasting (Engle, 1982; Bollerslev, 1986) and evidence that parsimonious GARCH specifications are hard to beat out of sample (Hansen & Lunde, 2005). For cryptocurrencies—characterized by volatility clustering, heavy tails, and regime shifts—such econometric benchmarks remain necessary rather than obsolete (Katsiampa, 2017; Cont, 2001).

The second finding concerns the standalone LSTM models. LSTM-T is the best-performing deep learning specification and yields the lowest loss in the normal-period cross-family comparison, highlighting the value of nonlinear sequence learning for short-horizon volatility forecasting when market-based technical indicators are included. These indicators summarize recent volatility pressure, price-range variation, and momentum, which are closely linked to short-run conditional volatility. This is consistent with evidence that technical indicators are predictive in financial markets (Neely et al., 2014) and with recent cryptocurrency studies emphasizing the role of internal market variables in volatility predictability (Wang et al., 2023; Dudek et al., 2024). However, the LSTM results also show that richer information sets do not automatically improve forecasts: adding sentiment (LSTM-TS) and blockchain variables (LSTM-TSB) worsens normal-period performance relative to LSTM-T. This does not imply that sentiment or blockchain information is irrelevant; rather, their incremental value depends on the volatility regime. In stable periods, extra predictors may add noise, redundancy, or unstable relationships, especially in flexible nonlinear models, consistent with machine learning evidence that out-of-sample performance can deteriorate when more predictors raise dimensionality without adding stable signals (Gu et al., 2020; James et al., 2021).

The third finding concerns hybrid models. In normal periods, Hybrid_GARCH_T outperforms the best traditional model but not LSTM-T. Diebold–Mariano tests show TGARCH and Hybrid_GARCH_T are statistically indistinguishable from LSTM-T. Thus, normal-period results do not support unconditional dominance of hybrid models. Instead, under stable market conditions, different model classes can achieve similar predictive accuracy when evaluated with a common volatility target and rolling out-of-sample design.

The crisis-period results show a more informative pattern. In all three crisis regimes, a hybrid model yields the lowest QLIKE, but the preferred hybrid changes by episode: Hybrid_TGARCH_TS for COVID-19, Hybrid_EGARCH_TSB for the China mining ban, and Hybrid_APARCH_TS for the crypto winter. This is central to the paper because it shows that the forecasting edge of hybrid models does not come from a single fixed architecture; the best specification depends on how each crisis alters volatility dynamics and which information channels matter.

The COVID-19 shock was abrupt, systemic, and heavily uncertainty-driven, making Hybrid_TGARCH_TS economically consistent with this episode. The TGARCH component captures threshold-type, asymmetric volatility responses to sharp market moves (Zakoian, 1994). Including sentiment is plausible because systemic uncertainty and rapid shifts in expectations increase the value of behavioural signals (Baker & Wurgler, 2007; Tetlock, 2007). The traditional iGARCH benchmark performs substantially worse and is only marginally inferior at the 10% level, implying persistence alone is insufficient during sudden dislocations. LSTM-TSB is statisti-

cally indistinguishable from the hybrid benchmark, showing that nonlinear models with richer information sets can remain competitive under extreme uncertainty.

The China mining ban shows weaker statistical separation across model families. Hybrid_EGARCH_TSB attains the lowest QLIKE, but GARCH and LSTM-T are not statistically different. This aligns with the episode's localized, crypto-specific nature: because the shock directly affected mining and network conditions, including blockchain information in the TSB is economically justified. Blockchain activity and network adoption are central to cryptocurrency valuation and market dynamics (Cong et al., 2021; Liu & Tsyvinski, 2021). The EGARCH component is also appropriate, as it permits logarithmic variance dynamics and asymmetric volatility without non-negativity constraints (Nelson, 1991). However, the insignificant DM tests indicate that this episode does not offer strong evidence of statistical dominance by the hybrid model, suggesting numerical superiority rather than decisive statistical outperformance.

The crypto winter offers the strongest support for crisis-specific hybrid modelling. Hybrid_APARCH_TS achieves the lowest QLIKE, while the traditional GARCH benchmark is significantly worse at the 1% level. This episode was more persistent than the COVID-19 and China mining-ban shocks, reflecting prolonged market weakness, deteriorating confidence, and sustained uncertainty after major crypto disruptions. The APARCH component is suitable here because it allows flexible power transformations and asymmetric volatility responses, capturing nonlinear scale effects in persistent stress regimes (Ding et al., 1993). The TS information set is economically relevant because prolonged downturns are shaped by both technical market pressure and sentiment-driven confidence. LSTM-TS remains statistically comparable to the hybrid benchmark, confirming the value of sentiment-enhanced nonlinear learning in extended stress, although the hybrid model delivers the lowest numerical loss.

The findings refine the study's main contribution: Bitcoin volatility forecasting is not a simple contest between econometric, deep learning, and hybrid models. Forecast accuracy depends on how volatility regimes interact with each model's information set. In normal periods, technical information and nonlinear learning yield strong forecasts, and traditional and hybrid benchmarks remain statistically competitive. In crises, hybrids gain value, but their advantage depends on the shock type: abrupt systemic stress favours a threshold-based hybrid with sentiment data; crypto-specific network disruption favours an EGARCH-based hybrid with blockchain data; and prolonged market weakness favours an APARCH-based hybrid with sentiment data. This regime-specific view better fits the evidence and offers a stronger academic contribution than claiming any single model class dominates.

5.2 Economic and methodological implications

The findings have key methodological implications for cryptocurrency volatility forecasting. First, they stress the need to compare models within a unified, information-consistent framework. Fixing the volatility target, forecast horizon, rolling window, and evaluation metrics limits the risk that performance differences arise from inconsistent data or model-specific choices—especially important in this field, where

results are highly sensitive to the volatility proxy, predictors, sample period, and loss function (Hansen & Lunde, 2005; Patton, 2011). Second, they clarify the role of model complexity. More flexible models do not automatically outperform simpler ones. In normal periods, LSTM-T has the lowest numerical loss, but TGARCH and Hybrid_GARCH_T are statistically indistinguishable. Within the hybrid class, the best normal-period model is relatively parsimonious, combining a standard GARCH volatility state with technical indicators. Model complexity should therefore match the forecasting environment: in stable markets, technical information and volatility persistence may be sufficient, and adding sentiment or blockchain variables can introduce noise rather than improve accuracy.

Crisis-period evidence shows that hybrid models are most useful when volatility is unstable, nonlinear, or regime-specific, but their benefits differ by episode. The COVID-19 shock favours a threshold-based hybrid with sentiment data, the China mining ban an EGARCH-based hybrid with blockchain data, and the crypto winter an APARCH-based hybrid with sentiment data. Thus, hybrid forecasting is not valuable merely because it combines econometric and deep learning elements; it works when the econometric backbone and information set align with the volatility episode's economic features.

The results inform how to use alternative data in cryptocurrency forecasting. Technical indicators offer the most stable short-horizon predictive power in normal times, reflecting the role of internal market dynamics in digital asset volatility. Sentiment and blockchain variables are instead state-dependent. Sentiment becomes more relevant when volatility stems from uncertainty, falling confidence, or narrative-driven reactions, as in the COVID-19 shock and the crypto winter. Blockchain data matter more when shocks are tied to network activity or crypto-specific fundamentals, as in the China mining ban. Thus, alternative data should not be added mechanically; their use should reflect the regime's economic nature and the forecast horizon.

The findings matter for risk management, derivative pricing, portfolio allocation, and volatility-based trading. Practitioners should avoid relying on a single model class in all market conditions. In stable periods, parsimonious econometric models and technical-information LSTMs can deliver forecasts that are statistically comparable to complex hybrid models. In crises, however, hybrid models can reduce volatility losses, especially under persistent stress or nonlinear adjustment. Diebold–Mariano tests also warn that not every numerical gain is statistically significant. Thus, model selection in applied risk management should combine loss-based rankings with statistical testing and economic judgment.

Bitcoin volatility forecasting should be treated as a regime-sensitive problem, not a static model-selection task. Forecasting gains depend on aligning model architecture, information set, and market conditions. This is crucial for digital assets, where volatility regimes shift abruptly and the relevance of technical, sentiment, and blockchain variables changes over time. The study thus contributes by identifying strong models and clarifying when and why specific model–information combinations are useful.

5.3 Limitations and future research

The empirical framework provides a controlled, internally consistent comparison of econometric, deep learning, and hybrid volatility models. Its main strength is the use of a common volatility target, nested information sets, and a uniform rolling out-of-sample evaluation, which enhances comparability across model classes while also defining the scope for interpreting and extending the results.

First, the analysis uses daily data and one-step-ahead volatility forecasts, which suits short-term monitoring and allows direct comparison across model specifications. Because cryptocurrency markets trade continuously and some volatility adjustment occurs intraday, future work could extend the framework to high-frequency realized volatility or multi-horizon forecasts to test whether the regime-dependent patterns found here hold at intraday, weekly, or monthly horizons. Second, all models are evaluated against a common daily volatility proxy, since conditional volatility is latent and unobservable; a single proxy ensures that forecast differences reflect model design, information, and market regime rather than changes in the target. However, model rankings may change under other volatility measures, so future studies could assess robustness using high-frequency realized volatility, range-based estimators, or proxies based on alternative return frequencies.

Third, the balanced sample ends on 19 July 2025, when technical, sentiment, and blockchain variables are all available. This conservative design avoids interpolation, forward-filling, and synthetic data, preserving the historical information set and limiting artificial persistence or look-ahead bias, at the cost of dropping later observations when any channel is missing. Future work could extend histories, use other data providers, or apply finer sentiment and blockchain measures to test the stability of model rankings as coverage improves. Fourth, the formal crisis evaluation includes only stress episodes compatible with the rolling out-of-sample framework. The 2018 post-bubble crash remains in the descriptive analysis, given its importance for Bitcoin's volatility history, but is excluded from the formal crisis comparison because the rolling scheme requires a long pre-crisis estimation window. Some crisis windows, especially for COVID-19 and the China mining ban, also contain few daily observations, so Diebold–Mariano results for these regimes should be viewed as regime-specific out-of-sample evidence, not as general evidence for all cryptocurrency market stress.

Fifth, the study focuses on Bitcoin, the most liquid and widely studied cryptocurrency and the benchmark asset in digital finance. Its long trading history and market depth make it suitable for a unified volatility-forecasting experiment. Other assets—such as smaller cryptocurrencies, stablecoins, DeFi tokens, or those with different liquidity or blockchain characteristics—may display different volatility dynamics. Applying the framework to a broader cross-section would show whether the roles of technical, sentiment, and blockchain information are asset-specific or generalizable. Sixth, the model set is intentionally restricted to preserve a clean comparison across model classes. The analysis considers six GARCH-type models, three standalone LSTMs, and eighteen hybrid GARCH–LSTM models under a common forecasting protocol. This controlled design strengthens the empirical interpretation by avoiding confounding from differences in targets, inputs, or evaluation procedures. Future work could extend the framework to other nonlinear architectures—such as GRU,

CNN-LSTM, attention-based LSTM, Temporal Convolutional Networks, XGBoost, LightGBM, random forests, and transformers—provided the strict out-of-sample design is preserved.

Seventh, the sentiment and blockchain variables are treated as predictive information channels, not causal drivers. Sentiment measures depend on source coverage, language mix, platform design, and aggregation, while blockchain indicators capture both genuine network activity and market noise. The results therefore indicate whether these variables improve forecast accuracy in specific regimes, not whether they causally determine Bitcoin volatility. Future work could combine richer alternative-data measures with causal identification to separate predictive signals from structural economic relationships. Finally, the study assesses statistical forecast accuracy rather than the direct economic value of the forecasts. QLIKE, RMSE, MAE, and Diebold–Mariano tests rigorously compare predictive performance but do not show whether gains translate into better trading, hedging, option pricing, portfolio allocation, or risk management. Future research could evaluate economic value using Value-at-Risk, Expected Shortfall, volatility-timing strategies, derivative valuation, or portfolio optimization.

These extensions define a research agenda building on this study. The framework provides a transparent benchmark to compare econometric, deep learning, and hybrid volatility models under consistent information sets and regime-specific market conditions. Future work can broaden data frequency, asset coverage, model classes, volatility targets, and the economic evaluation of forecasts.

6 Conclusion

This study develops a unified out-of-sample framework for forecasting Bitcoin volatility using econometric volatility models, standalone LSTM networks, and state-augmented hybrid GARCH–LSTM specifications. By evaluating all models under a common volatility target, nested information sets, rolling one-step-ahead forecasts, and identical loss functions, the analysis isolates the role of model architecture, information structure, and market regime in shaping predictive accuracy.

The results show that Bitcoin volatility forecasting is inherently regime-dependent. Under normal market conditions, LSTM-T records the lowest numerical loss, although TGARCH and Hybrid_GARCH_T are statistically indistinguishable from it according to the Diebold–Mariano tests. This indicates that nonlinear learning with technical information is effective in stable periods, while parsimonious GARCH-type models and state-augmented hybrids remain competitive benchmarks. The normal-period evidence therefore does not support unconditional dominance by any single model class.

Crisis-period results provide stronger evidence in favour of regime-specific hybrid modelling. Across the COVID-19 shock, the China mining ban, and the crypto winter, the lowest QLIKE is achieved by a hybrid specification. However, the selected hybrid model differs across crises: Hybrid_TGARCH_TS performs best during the abrupt and sentiment-sensitive COVID-19 shock, Hybrid_EGARCH_TSB during the crypto-specific China mining-ban episode, and Hybrid_APARCH_TS during the

prolonged crypto winter. These findings show that hybrid forecasting gains arise not simply from model complexity, but from the alignment between the econometric volatility state, the information set, and the characteristics of the market regime. The evidence also highlights the conditional value of alternative information. Technical indicators provide the most stable short-horizon predictive content in normal periods, while sentiment and blockchain variables become more useful when they match the economic nature of the shock. Sentiment information is more relevant during uncertainty-driven and confidence-sensitive regimes, whereas blockchain information is more informative when the disturbance is directly connected to network activity and crypto-specific fundamentals.

Overall, the study contributes to the cryptocurrency volatility forecasting literature by showing that predictive gains are regime-specific rather than universal. This conclusion is consistent with recent evidence that cryptocurrency volatility is shaped by regime shifts, nonlinear dynamics, and structural market conditions (Kayal & Dutta, 2024; Agakishiev et al., 2025; Krain et al., 2025). Accurate Bitcoin volatility forecasting therefore depends on the interaction between structured volatility dynamics, nonlinear learning, and economically relevant information channels. The study supports the use of unified, regime-sensitive evaluation frameworks and cautions against drawing broad model-superiority claims from pooled-sample comparisons alone.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s42521-026-00229-9.

Author contributions A.S.W. conceived the study, developed the methodology, collected and processed the data, conducted the empirical analysis, generated the results, prepared the figures and tables, and wrote the manuscript. S.E. contributed to the study through academic guidance, conceptual discussion, and critical review of the manuscript. S.A. contributed to the study through academic guidance, conceptual discussion, and critical review of the manuscript. All authors read and approved the final manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability The data used in this study are obtained from publicly available financial databases and sources cited in the manuscript. Supplementary empirical materials, including full descriptive statistics, full normal-period and crisis-period model rankings, Diebold–Mariano test comparisons, and summaries of models statistically indistinguishable from the relevant benchmark, are publicly available at: https://github.com/Wafi-Finance/Regime-Dependent-Bitcoin-Volatility-Forecasting.git. Updated data files, replication codes, further empirical outputs, and extended robustness materials will be made available through the same public repository upon publication.

Declarations

Conflict of interest The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licen

ses/by/4.0/.

References

Agakishiev, I., Härdle, W. K., Becker, D., & Zuo, X. (2025). Regime switching forecasting for cryptocurrencies. *Digital Finance*, *7*, 107–131. https://doi.org/10.1007/s42521-024-00123-2

Andersen, T. G., Bollerslev, T., & Meddahi, N. (2005). Correcting the errors: Volatility forecast evaluation using high-frequency data. *Econometrica*, *73*(1), 279–296. https://doi.org/10.1111/j.1468-0262.2005.00576.x

Baker, M., & Wurgler, J. (2007). Investor sentiment in the stock market. *Journal of Economic Perspectives*, *21*(2), 129–152. https://doi.org/10.1257/jep.21.2.129

Bates, J. M., & Granger, C. W. J. (1969). The combination of forecasts. *Operational Research Quarterly,20*(4), 451–468. https://doi.org/10.1057/jors.1969.103

Baur, D. G., Hong, K., & Lee, A. D. (2018). Bitcoin: Medium of exchange or speculative assets? *Journal of International Financial Markets, Institutions and Money,54*, 177–189. https://doi.org/10.1016/j.intfin.2017.12.004

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, *31*(3), 307–327. https://doi.org/10.1016/0304-4076(86)90063-1

Cheah, J.-T., Dao, T., Do, H., & Mishra, T. (2026). Speed of adjustment in digital assets in a decentralized financial world. *Journal of Futures Markets,46*(2), 320–333. https://doi.org/10.1002/fut.70055

Chung, V., Espinoza, J., & Quispe, R. (2025). Forecasting financial volatility under structural breaks: A comparative study of GARCH models and deep learning techniques. *Journal of Risk and Financial Management,18*(9), Article 494. https://doi.org/10.3390/jrfm18090494

Cong, L. W., Li, Y., & Wang, N. (2021). Tokenomics: Dynamic adoption and valuation. *Review of Financial Studies*, *34*(3), 1105–1155. https://doi.org/10.1093/rfs/hhaa089

Cont, R. (2001). Empirical properties of asset returns: Stylized facts and statistical issues. *Quantitative Finance*, *1*(2), 223–236. https://doi.org/10.1088/1469-7688/1/2/304

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *13*(3), 253–263. https://doi.org/10.1080/07350015.1995.10524599

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. https://doi.org/10.1016/0927-5398(93)90006-D

Dudek, G., Fiszeder, P., Kobus, P., & Orzeszko, W. (2024). Forecasting cryptocurrencies volatility using statistical and machine learning methods: A comparative study. *Applied Soft Computing,151*, Article 111132. https://doi.org/10.1016/j.asoc.2023.111132

Engle, R. F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, *50*(4), 987–1007. https://doi.org/10.2307/1912773

Engle, R. F., & Bollerslev, T. (1986). Modelling the persistence of conditional variances. *Econometric Reviews,5*(1), 1–50. https://doi.org/10.1080/07474938608800095

Fang, L., Bouri, E., Gupta, R., & Roubaud, D. (2019). Does global economic uncertainty matter for the volatility and hedging effectiveness of Bitcoin? *International Review of Financial Analysis,61*, 29–36. https://doi.org/10.1016/j.irfa.2018.12.010

Feng, L., Qi, J., & Lucey, B. (2024). Enhancing cryptocurrency market volatility forecasting with daily dynamic tuning strategy. *International Review of Financial Analysis,94*, Article 103239. https://doi.org/10.1016/j.irfa.2024.103239

Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, *270*(2), 654–669. https://doi.org/10.1016/j.ejor.2017.11.054

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. https://doi.org/10.1111/j.1540-6261.1993.tb05128.x

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *Review of Financial Studies*, *33*(5), 2223–2273. https://doi.org/10.1093/rfs/hhaa009

Gujarati, D. N., & Porter, D. C. (2009). *Basic Econometrics* (5th ed.). McGraw-Hill.

Hansen, P. R., & Lunde, A. (2005). A forecast comparison of volatility models: Does anything beat a GARCH(1,1)? *Journal of Applied Econometrics,20*(7), 873–889. https://doi.org/10.1002/jae.800

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Huang, Z.-C., Sangiorgi, I., & Urquhart, A. (2024). Forecasting Bitcoin volatility using machine learning techniques. *Journal of International Financial Markets, Institutions and Money,97*, Article 102064. https://doi.org/10.1016/j.intfin.2024.102064

Huang, X., Tan, L., Su, H., & Cheah, J. E. T. (2026). Using deep learning conditional value-at-risk based utility function in cryptocurrency portfolio optimisation. *International Journal of Finance & Economics*, *31*(2), 2845–2862. https://doi.org/10.1002/ijfe.70012

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning* (2nd ed.). Springer. https://doi.org/10.1007/978-1-0716-1418-1

Katsiampa, P. (2017). Volatility estimation for Bitcoin: A comparison of GARCH models. *Economics Letters*, *158*, 3–6. https://doi.org/10.1016/j.econlet.2017.06.023

Kayal, P., & Dutta, S. (2024). Regime switching and causal network analysis of cryptocurrency volatility: Evidence from pre-COVID and post-COVID analysis. *Digital Finance*. https://doi.org/10.1007/s42521-023-00104-x

Khashei, M., & Bijari, M. (2011). A novel hybridization of artificial neural networks and ARIMA models for time series forecasting. *Applied Soft Computing*, *11*(2), 2664–2675. https://doi.org/10.1016/j.asoc.2010.10.015

Korkusuz, B., & Sahiner, M. (2025). Coin impact on cross-crypto realized volatility and dynamic cryptocurrency volatility connectedness. *Financial Innovation*, *11*, 129. https://doi.org/10.1186/s40854-025-00881-x

Krain, L., Zuo, X., & Härdle, W. K. (2025). Cryptos have rough volatility and correlated jumps. *Digital Finance*, *7*(2), 255–273. https://doi.org/10.1007/s42521-025-00125-8

Kristjanpoller, W., & Minutolo, M. C. (2016). Forecasting volatility of oil price using an artificial neural network–GARCH model. *Expert Systems with Applications*, *65*, 233–241. https://doi.org/10.1016/j.eswa.2016.08.045

Liu, Y., & Tsyvinski, A. (2021). Risks and returns of cryptocurrency. *Review of Financial Studies*, *34*(6), 2689–2727. https://doi.org/10.1093/rfs/hhaa113

Mokni, K. (2021). When, where, and how economic policy uncertainty predicts Bitcoin returns and volatility? *Quarterly Review of Economics and Finance*, *81*, 65–83. https://doi.org/10.1016/j.qref.2021.01.017

Neely, C. J., Rapach, D. E., Tu, J., & Zhou, G. (2014). Forecasting the equity risk premium: The role of technical indicators. *Management Science*, *60*(7), 1772–1791. https://doi.org/10.1287/mnsc.2013.1838

Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica*, *59*(2), 347–370. https://doi.org/10.2307/2938260

Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, *160*(1), 246–256. https://doi.org/10.1016/j.jeconom.2010.03.034

Shaikh, I. (2020). Does economic policy uncertainty predict Bitcoin returns? *Borsa Istanbul Review*, *20*(2), 119–131. https://doi.org/10.1016/j.bir.2020.02.003

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*, 1929–1958. https://www.jmlr.org/papers/v15/srivastava14a.html

Teräsvirta, T., van Dijk, D., & Medeiros, M. C. (2005). Linear models, smooth transition autoregressions, and neural networks for forecasting macroeconomic time series. *International Journal of Forecasting*, *21*(4), 755–774. https://doi.org/10.1016/j.ijforecast.2005.04.010

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media tone in the stock market. *Journal of Finance*, *62*(3), 1139–1168. https://doi.org/10.1111/j.1540-6261.2007.01232.x

Wang, Y., Andreeva, G., & Martin-Barragan, B. (2023). Machine learning approaches to forecasting cryptocurrency volatility: Considering internal and external determinants. *International Review of Financial Analysis,90*, Article 102914. https://doi.org/10.1016/j.irfa.2023.102914

Zahid, M., Iqbal, F., & Koutmos, D. (2022). Forecasting Bitcoin volatility using hybrid GARCH models with machine learning. *Risks,10*(12), Article 237. https://doi.org/10.3390/risks10120237

Zakoian, J. M. (1994). Threshold Heteroskedastic Models. *Journal of Economic Dynamics and Control*, *18*(5), 931–955. https://doi.org/10.1016/0165-1889(94)90039-6

Zhang, G. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, *50*, 159–175. https://doi.org/10.1016/S0925-2312(01)00702-0

Zournatzidou, G., Mallidis, I., Farazakis, D., & Floros, C. (2024). Enhancing Bitcoin price volatility estimator predictions: A four-step methodological approach utilizing elastic net regression. *Mathematics,12*(9), Article 1392. https://doi.org/10.3390/math12091392

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Ahmed S. Wafi[1,2,3] · Sherif El-Halaby[4,5] · Sameh Abd-ElMaksoud Aboul-Dahab[6,7]

✉ Ahmed S. Wafi
ahmed.wafi@lmu.de; ahmed_s.wafi@euruni.edu; as.wafi@foc.cu.edu.eg

Sherif El-Halaby
s.elhalby@ajman.ac.ae

Sameh Abd-ElMaksoud Aboul-Dahab
sameh.abdelmaksod@com.kfs.edu.eg

1 Department of Statistics Seminar of Financial Econometrics, Ludwig-Maximilians-Universität München, 80539 Munich, Germany

2 Dept. of Finance, stat., & Data Analysis, EU Business School, Munich, Theresienhöhe 28, 80339 Munich, Germany

3 Business Administration Dept., Faculty of Commerce, Cairo University, 12613, Giza, Egypt

4 Dept. of Finance & Accounting College of Business Administration, Ajman University, 346, Al Jerf 1, Ajman, United Arab Emirates

5 Department of Accounting, October University of Modern Sciences and Arts, 12451 6th October City, Egypt

6 Department of Business Administration, College of Business, Jouf University, 72388 Sakakah, Saudi Arabia

7 Business Administration Department Faculty of Commerce, Kafrelsheikh University, 33516 Kafr ash Shaykh, Egypt