FINANCIAL DYNAMICS UNDER SHOCKS: A COMPARATIVE ANALYSIS OF DEEP LEARNING, GARCH MODELS, AND BLACK-SCHOLES ASSUMPTION FAILURES

KONRAD BARANEK AND MAKSYMILIAN DĘBOWSKI

Abstract. Classical Black-Scholes assumptions, like constant volatility, often fail during market shocks. This study investigates these limitations using a case study of a company experiencing a specific external shock. We employ GARCH(1,1) and asymmetric GJR-GARCH models to analyze pre-shock volatility dynamics. We generate and compare post-shock price path simulations and volatility forecasts using: (1) fixed pre-shock GARCH/GJR parameters (static forecast), (2) adaptive expanding-window GARCH/GJR parameters (dynamic forecast), (3) a fixed naive approach based on pre-shock averages (static forecast), (4) a dynamic naive approach using lagged historical data (dynamic forecast), (5) Deep Learning (DL) models trained pre-shock using the AutoGluon-TimeSeries framework (static forecast), (6) adaptive standard Machine Learning (ML) models (LGBM, XGBoost, RandomForest, Ridge) with feature engineering and hyperparameter optimization (dynamic volatility forecast), and (7) the TimesFM foundation model in an adaptive rolling forecast setting (dynamic price forecast). Performance is assessed against the actual price path and realized volatility. Results highlight the failure of static forecasts (Fixed GARCH/Naive, AutoGluon). Adaptive models excelled post-shock: adaptive LightGBM and a dynamic naive approach provided the most accurate volatility forecasts (depending on the metric), while the adaptive TimesFM rolling forecast yielded the best median price path simulation. Findings underscore the critical role of model adaptability and appropriate feature engineering post-shock.

1. Introduction

Accurate modeling and forecasting of financial asset volatility are crucial for effective risk management, derivative pricing, and portfolio allocation [5, 3]. The Black-Scholes model [2], while foundational, describes the asset price dynamics S_t via a GBM (geometric Brownian motion).

$$dS_t = \mu S_t dt + \sigma S_t dW_t \tag{1}$$

where μ is the "percentage drift" (or expected return), W_t is a standard Wiener process, and crucially, σ is the volatility, which is assumed to be constant.

This assumption is recognized as empirically unrealistic, particularly during periods of market stress or some idiosyncratic shocks affecting specific assets [4, 1]. Such shocks often induce significant, abrupt changes in volatility dynamics, making standard B-S pricing and hedging formulas (σ = constant) unreliable just when accurate risk assessment is most critical. This project addresses the challenge of modeling and forecasting volatility under market shocks, focusing on addressing the limitations of the B-S framework and comparing the performance of traditional econometric models with modern machine learning techniques. Specifically, we examine the impact of a documented external shock on the volatility dynamics of a selected company's stock price.

Standard approaches to capture time-varying volatility include the family of GARCH models [3]. The canonical GARCH(1,1) model, for instance, describes the conditional variance $\sigma_t^2 = \mathrm{Var}(r_t|\mathcal{F}_{t-1})$ of returns r_t (given the past information $\mathcal{F}_{t-1}$) as:

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

where $\epsilon_t = r_t - \mathbb{E}[r_t|\mathcal{F}_{t-1}]$ are the residuals, and $\omega > 0$, $\alpha \geq 0$, $\beta \geq 0$ are parameters usually constrained to ensure stationarity (e.g., $\alpha + \beta < 1$).

While GARCH models effectively capture "groups" of different volatilities, asymmetric variants such as GJR-GARCH models [7] also take into account leverage (negative shocks increase volatility more than positive ones):

$$\sigma_t^2 = \omega + (\alpha + \gamma \mathbb{1}_{\epsilon_{t-1}<0})\epsilon_{t-1}^2 + \beta\sigma_{t-1}^2, \tag{3}$$

where $\mathbb{1}_{\epsilon_{t-1}<0}$ is equal to 1 if the previous shock was negative, and $\gamma \geq 0$ captures the asymmetry.

Despite their broad utility, the responsiveness and predictive accuracy of GARCH models during sudden, high-severity shocks require careful investigation, particularly concerning parameter stability and the need for adaptive estimation versus relying on pre-shock parameters.

In recent years, Deep Learning (DL) architectures have become powerful tools for modeling complex time series data [10, 8]. Automated Machine Learning (AutoML) frameworks like AutoGluon [6, 12] simplify their application. However, their effectiveness in forecasting through major structural breaks when trained only on pre-break data remains an open question compared to models that adapt using post-break information. Furthermore, standard Machine Learning (ML) models like Gradient Boosting Machines (e.g., LightGBM, XGBoost) or ensemble methods (Random Forest), when used adaptively with appropriate feature engineering, offer another powerful alternative for time series forecasting [?]. Large pre-trained foundation models for time series, such as TimesFM [?], represent another recent advancement, potentially offering strong performance, especially when used adaptively.

1.1. **Main Contributions.** The main contribution of this paper is divided into three parts:

(1) We demonstrate the failure of Black-Scholes assumptions during a specific shock event through empirical analysis. We quantify the shock's impact by comparing actual post-shock dynamics to counterfactual simulations based on fixed pre-shock GARCH/Naive parameters.
(2) We conduct a comparative analysis of post-shock simulation/forecasting performance between models using fixed pre-shock parameters (Fixed GARCH, Fixed Naive, AutoGluon trained pre-shock) and models incorporating post-shock adaptation (Adaptive GARCH, Dynamic Naive, Adaptive standard ML models, Adaptive TimesFM).

(3) We assess the impact of model adaptability and feature engineering on simulation/forecasting accuracy in the post-shock period, highlighting the benefits of adaptation (Adaptive GARCH/Naive/ML/TimesFM) versus static forecasts (Fixed GARCH/Naive, AutoGluon) for both volatility tracking and price path prediction in this specific shock scenario.

Our methodology uses data from Rheinmetall AG. We simulate/forecast post-shock price paths and volatility using the different modeling approaches. The results aim to shed light on the relative strengths and weaknesses, particularly regarding adaptability and the role of feature engineering, in volatile, shock-prone conditions.

The remainder of this article is organized as follows. Section 2 describes the data, models, simulation/forecasting strategies, and evaluation metrics. Section 3 presents empirical analysis and comparative performance. Section 4 discusses the findings. Section 5 provides information about the code repository.

2. Methodology

For the analysis, we chose the German joint-stock company Rheinmetall AG, which is a leading international manufacturer of systems in the defense industry.

2.1. Data and Preparation.

Data was acquired using the `yfinance` Python library for the period starting from January 1, 2019 up to April 4, 2025 (though analysis focuses on the period around the shock). Data was retrieved at a daily frequency. We utilize the daily Adjusted Close price (P_t), which accounts for dividends and stock splits.

Logarithmic returns (r_t) were calculated using the standard formula:

$$r_t = \ln\left(\frac{P_t}{P_{t-1}}\right). \tag{4}$$

The first observation in the returns series was subsequently removed. We also calculate rolling mean log returns over a window $W = 21$:

$$\bar{r}_{roll,t} = \frac{1}{W}\sum_{i=0}^{W-1} r_{t-i}. \tag{5}$$

2.2. Realized Volatility Calculation.

The target variable for volatility forecasting is the realized annualized volatility ($\sigma_{RV,t}$). We estimate it using a rolling window ($W = 21$ trading days) of daily logarithmic returns (r_t). We also define the daily realized volatility $\sigma_{RV,daily,t}$. The calculation, based on [9], is:

$$\sigma_{RV,t} = \sqrt{N} \times \sigma_{RV,daily,t} = \sqrt{N} \times \sqrt{\frac{1}{W-1}\sum_{i=0}^{W-1}(r_{t-i} - \bar{r}_{roll,t})^2}, \tag{6}$$

where $N = 260$ is the annualization factor and $\bar{r}_{roll,t}$ is the rolling average return (Eq. 5). This $\sigma_{RV,t}$ serves as the ground truth proxy for actual annualized volatility. $\sigma_{RV,daily,t}$ (without the $\sqrt{N}$ factor but often scaled by 100 for percentage representation) is used as the target for

AutoGluon and standard ML volatility prediction. For consistency in evaluation (Table 8), all model volatility forecasts are compared against the annualized $\sigma_{RV,t}$.

2.3. **Shock Event.** The chosen shock date is February 24, 2022, marking the start of the full-scale Russian invasion of Ukraine. This event significantly impacted the defense sector, including Rheinmetall AG. Figure 1 visually confirms a structural change around this date, with sharp increases in both price and realized volatility.

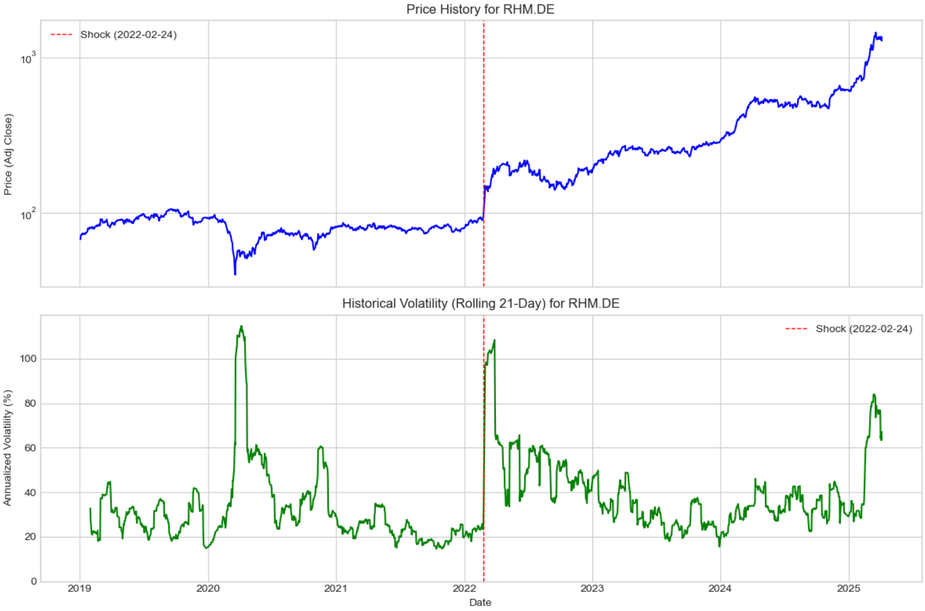

FIGURE 1. Adjusted Closing Price and Annualized 21-Day Realized Volatility for RHM.DE. The vertical dashed red line indicates the identified shock date (2022-02-24).

2.4. **GARCH Model Estimation.** We utilize GARCH(1,1) (Eq. 2) and GJR-GARCH(1,1) (Eq. 3) models with Student's t-distributed innovations. Parameters are estimated via MLE using the `arch` library. Model orders (1,1) were selected based on BIC using pre-shock data (Tables 4, 5).

2.5. **Naive Baseline Approaches.** Two simple baselines are used for price path simulation and volatility forecasting:

2.5.1. *Dynamic Naive Approach (Rolling History T-1 Year / T-1 Day).* For price path simulation, uses rolling historical mean log-return $\hat{\mu}_{log,hist}$ and historical daily realized volatility $\hat{\sigma}_{daily,hist}$ calculated over the period $[t-\Delta, t-1]$ (where Δ is 1 year or 1 day) to simulate $r_t^{(j)} \sim N(\hat{\mu}_{log,hist}, \hat{\sigma}^2_{daily,hist})$ for each post-shock day t. For volatility forecasting, it uses the lagged historical daily realized volatility $\hat{\sigma}_{daily,hist}$ (annualized for comparison). This approach adapts based on recent data.

2.5.2. *Fixed Naive Approach (Average Pre-Shock Parameters).* Uses fixed average mean log-return $\bar{\mu}_{log,fixed}$ and average daily realized volatility $\hat{\sigma}_{daily,fixed}$ calculated over the year `before` the shock ($[t_{shock} - \Delta_{1yr}, t_{shock} - 1]$) to simulate $r_t^{(j)} \sim N(\bar{\mu}_{log,fixed}, \hat{\sigma}^2_{daily,fixed})$ for all post-shock days t. The volatility forecast is the constant $\hat{\sigma}_{daily,fixed}$ (annualized). This is a static approach.

2.6. **AutoGluon-TimeSeries Setup.** We use AutoGluon-TimeSeries [12] version 1.2 to generate DL-based forecasts. Crucially, AutoGluon was trained `only` on data available `before` the shock date (up to Feb 23, 2022). Two separate predictors were trained:

(1) **Volatility Predictor:** Trained to forecast `Realized Volatility Daily` (non-annualized, scaled by 100) over the entire post-shock horizon.
(2) **Price Predictor:** Trained to forecast `Adj Close` over the same horizon.

Both used the `best_quality` preset, RMSE evaluation, a 3600s time limit, GPU acceleration, and available past covariates (e.g., price, volume, returns for volatility predictor). AutoGluon selected a `WeightedEnsemble` as the best model based on validation performance (using validation folds within the pre-shock data). This pre-trained ensemble was then used to generate a **single, static forecast** for the entire post-shock period (median quantile "0.5" for price, "mean" for volatility) without any updates using post-shock data. The mean daily volatility forecast was annualized for evaluation.

2.7. **Adaptive Standard Machine Learning Models.** Beyond Deep Learning, we also evaluate standard ML models in an adaptive forecasting framework for volatility. We tested LightGBM, XGBoost, RandomForest, and Ridge regression.

- **Features:** We experimented with different feature sets, including basic lags, rolling window statistics (mean, std) of key variables (price, returns, volatility, volume) over multiple horizons (e.g., 5, 10, 21, 63 days), calendar features, and technical indicators. Rolling window features proved most effective.
- **Adaptive Training:** Similar to Adaptive GARCH, these models were retrained daily (or at each step) using all data up to $t - 1$ to predict the volatility at time t.
- **Target:** The target variable was the daily realized volatility (non-annualized, scaled by 100).
- **Optimization:** Hyperparameters for the best performing model (LightGBM with rolling features) were optimized using Optuna based on time series cross-validation on the pre-shock data.
- **Evaluation:** The resulting daily volatility forecasts were annualized for comparison in Table 8. These models were used only for volatility forecasting in this study, not price path simulation.

2.8. **Adaptive TimesFM Rolling Forecast.** We utilize the TimesFM foundation model [**?**] (specifically, `google/timesfm-2.0-500m-pytorch`) for price forecasting in an adaptive, rolling framework.

- **Strategy:** A rolling forecast simulation is performed starting from the shock date. At predefined intervals (`step size`), the model uses the most recent `contextlen` days of historical price data to forecast the next `horizon len` days. Only the first `step size` days of this forecast are kept and evaluated. The historical window then rolls forward by `step size` days.
- **Parameters:** We used `context len = 128` days and `horizon len = 30` days. We tested `step size` values of 1, 5, 10, 15, and 20 days.
- **Target:** The model directly forecasts the price (**Adj Close**).
- **Evaluation:** The median forecast (`timesfm` column from the model output) for the relevant `step size` days is compared against the actual price (`y`) to calculate MAE and RMSE. Results for the best performing `step size=1` are reported in Table 8. No volatility forecast was generated or evaluated in this setup.

2.9. **Simulation and Forecasting Strategies.** We compare the following strategies post-shock, starting from $S_{shock-1}$:

(1) **Fixed GARCH/GJR:** Models estimated once on pre-shock data. Used for static price simulation ($n_{sims} = 20000$) and static volatility forecast over the entire post-shock period.
(2) **Adaptive GARCH/GJR:** Parameters re-estimated daily using an expanding window including post-shock data. Used for dynamic one-step-ahead price simulation ($n_{sims} = 20000$) and dynamic volatility forecast ($\hat{\sigma}_t$).
(3) **Fixed Naive:** Uses fixed pre-shock averages for static price simulation ($n_{sims} = 20000$) and static volatility forecast.
(4) **Dynamic Naive:** Uses rolling history (T-1 year or T-1 day) for dynamic price simulation ($n_{sims} = 20000$) and dynamic volatility forecast (lagged rolling volatility).
(5) **AutoGluon Forecast:** Uses the pre-trained AutoGluon ensembles (Sec 2.6) to generate a single, static forecast path for price and volatility over the entire post-shock horizon.
(6) **Adaptive ML Forecast (Volatility Only):** Uses standard ML models (best: Optimized LightGBM with rolling features) retrained daily (Sec 2.7) to generate a dynamic volatility forecast.
(7) **Adaptive TimesFM Forecast (Price Only):** Uses the TimesFM model in a rolling forecast simulation (Sec 2.8) to generate an adaptive price forecast.

Strategies (1), (3), and (5) represent forecasts made *at the time of the shock* using only past information (static). Strategies (2), (4), (6), and (7) represent forecasts that **adapt** using information revealed after the shock (dynamic). Note that strategy (6) only provides volatility forecasts, and strategy (7) only provides price forecasts in this study.

2.10. **Evaluation Metrics.** Performance is evaluated over the post-shock period T.

2.10.1. *Price Path Accuracy Metrics.* Median simulated price ($\hat{S}_{t,median}$ for GARCH/Naive), AutoGluon median price forecast ($\hat{S}_{t,AG,0.5}$), or TimesFM median price forecast ($\hat{S}_{t,TFM}$) compared against actual price (S_t). Lower is better.

(1) **RMSE**: $RMSE_{price} = \sqrt{\frac{1}{T}\sum_{t=t_0}^{t_0+T-1}(S_t - \hat{S}_t)^2}$
(2) **MAE**: $MAE_{price} = \frac{1}{T}\sum_{t=t_0}^{t_0+T-1}|S_t - \hat{S}_t|$

where $\hat{S}_t$ is the respective model's price forecast/simulation median.

2.10.2. *Volatility Forecast Accuracy Metrics.* Model's annualized forecast volatility ($\hat{\sigma}_{model,t}$) compared against actual annualized realized volatility ($\sigma_{RV,t}$) over T_{vol} days (from $t_0 + W$ to $t_0 + T - 1$). AutoGluon's daily mean forecast $\hat{\sigma}_{AG,daily,mean}$ is annualized: $\hat{\sigma}_{AG,t} = \sqrt{N} \times \hat{\sigma}_{AG,daily,mean}$. GARCH/Naive daily σ_t are similarly annualized. Adaptive ML daily forecasts are also annualized. Lower is better.

(1) **RMSE**: $RMSE_{vol} = \sqrt{\frac{1}{T_{vol}}\sum(\sigma_{RV,t} - \hat{\sigma}_{model,t})^2}$
(2) **MAE**: $MAE_{vol} = \frac{1}{T_{vol}}\sum|\sigma_{RV,t} - \hat{\sigma}_{model,t}|$
(3) **QLIKE** [11]: $QLIKE_{vol} = \frac{1}{T_{vol}}\sum\left(\frac{\sigma^2_{RV,t}}{\hat{\sigma}^2_{model,t}} - \ln\left(\frac{\sigma^2_{RV,t}}{\hat{\sigma}^2_{model,t}}\right) - 1\right)$

3. Results and Comparisons

3.1. **Empirical Analysis of Black-Scholes Assumptions.** Before proceeding with volatility forecasting, we begin by empirically examining the key mathematical assumptions of the Black-Scholes model, namely constant volatility and normality of logarithmic returns. This is crucial to demonstrate the context of our study and the need to go beyond the B-S framework, especially in a market environment highly susceptible to shocks.

3.1.1. *Volatility Dynamics Pre- and Post-Shock.* Descriptive statistics in Table 1 and Figure 2 clearly demonstrate a significant change in volatility dynamics after the shock. The mean annualized realized volatility increased significantly, rising from 31.37% pre-shock to 40.11% post-shock. The standard deviation also rose substantially, indicating wider volatility swings.

Table 1. Descriptive Statistics of Annualized Realized Volatility Before and After Shock (2022-02-24)

Statistic	Before Shock	After Shock
Mean (%)	31.37	40.11
Std Dev (%)	26.15	35.42
Median (%)	16.60	16.63
Min (%)	14.53	15.53
Max (%)	114.81	108.48

Figure 2 visually confirms this shift, showing the post-shock volatility distribution shifted towards higher levels and exhibiting greater dispersion compared to the pre-shock period.

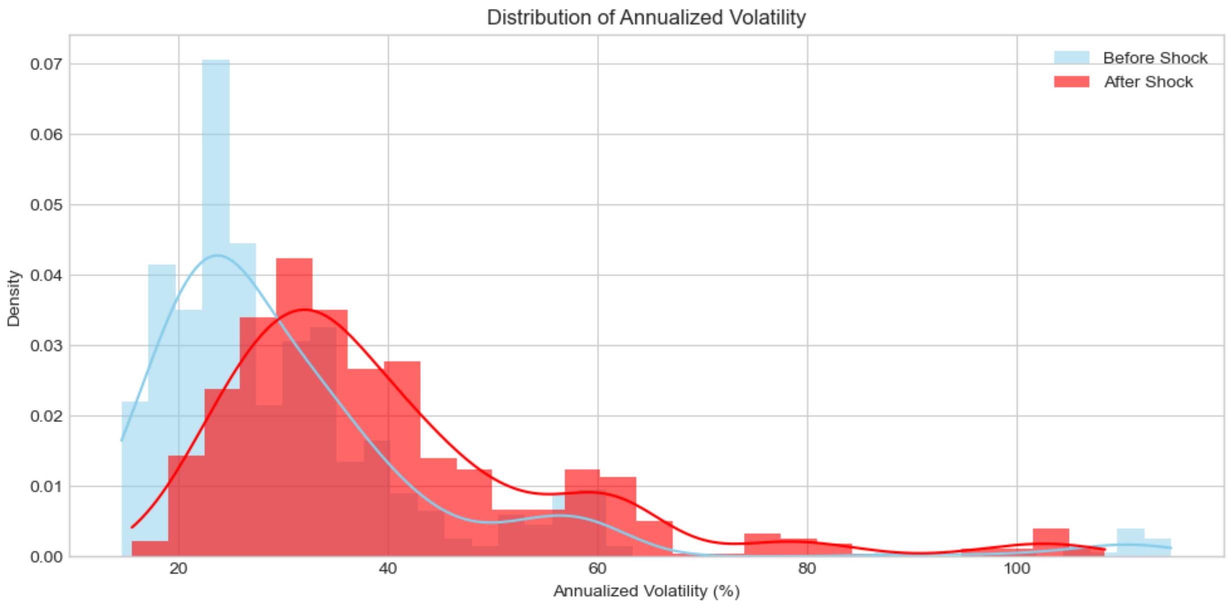

FIGURE 2. Distribution of Annualized Realized Volatility Before and After Shock (2022-02-24)

This evidence points to a clear structural break in volatility behavior coinciding with the shock, directly challenging the Black-Scholes assumption of constant volatility.

3.1.2. *Normality of Log-Returns.* Analysis shows logarithmic returns fail to meet the Black-Scholes normality assumption. Statistics in Table 2 indicate positive skewness and high kurtosis (leptokurtosis) in both periods (pre and post-shock), demonstrating a right-skewed, heavy-tailed distribution.

TABLE 2. Descriptive Statistics of Daily Log-Returns Before and After Shock (2022-02-24)

Statistic	Before Shock	After Shock
Mean	0.000338	0.003321
Std Dev	0.022061	0.027382
Skewness	0.375239	0.738859
Kurtosis	9.976945	9.294488

Histograms (Figure 3) and Q-Q plots (Figure 4) confirm significant deviations from normality, especially in the tails. The Shapiro-Wilk test (Table 3) firmly rejects normality ($p \approx 0.0000$) for both periods. Increased volatility after the shock and persistent non-normality weaken the Black-Scholes model's applicability, particularly during shocks when extreme price movements are more likely.

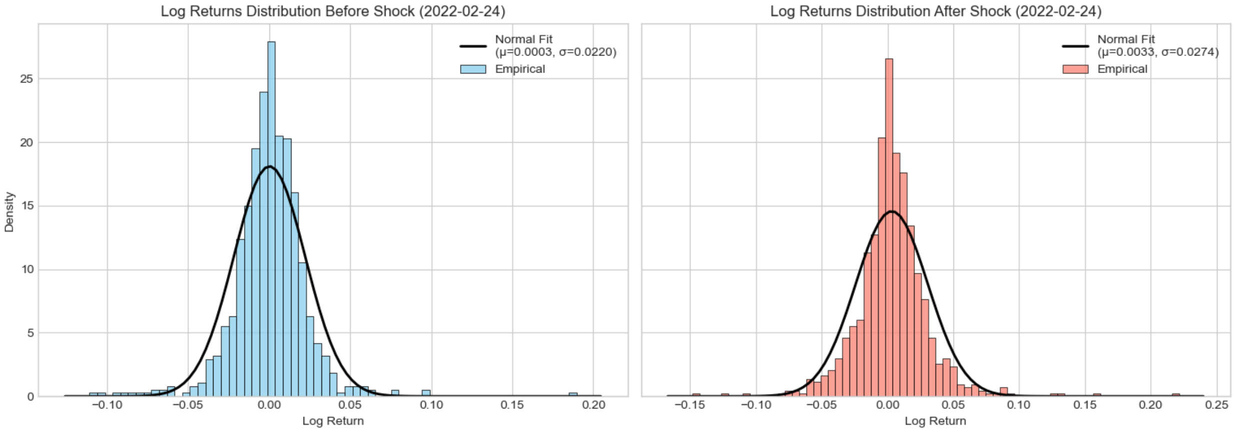

Figure 3. Histograms of Daily Log Returns Before and After Shock (2022-02-24) with Fitted Normal Distributions

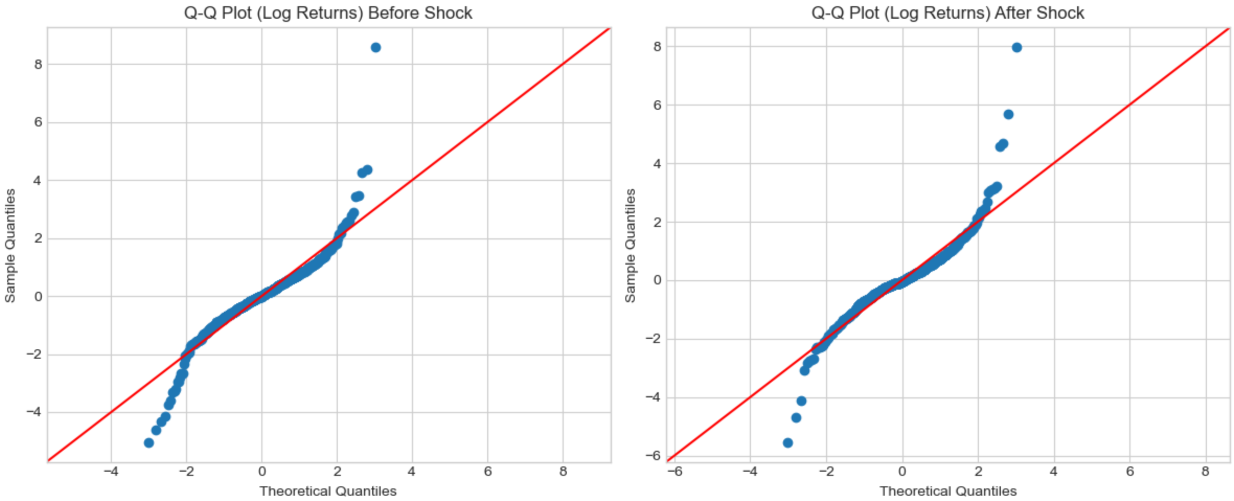

Figure 4. Q-Q Plots of Daily Log Returns Against Normal Distribution Before and After Shock (2022-02-24)

Table 3. Shapiro-Wilk Normality Test for Daily Log Returns

Test Period	Statistic	p-value
Before Shock	0.9062	< 0.0001
After Shock	0.9055	< 0.0001

3.1.3. *Implications for Modeling.* Empirical evidence confirms Black-Scholes assumptions fail for the analyzed asset, especially during the market shock. Volatility shows a structural shift post-shock, while log-returns display non-normality and heavy tails throughout. These findings highlight the standard model's limitations for pricing and risk management during market stress, necessitating more sophisticated approaches like GARCH, standard ML, and potentially Deep Learning (like AutoGluon or TimesFM) to capture time-varying volatility and non-normality.

3.2. **Post-Shock Simulation with Fixed Pre-Shock Parameters.**

We first examine simulations based on models and parameters derived solely from pre-shock data to represent a counterfactual "no-shock adaptation" scenario. This includes GARCH(1,1) and GJR-GARCH(1,1,1) models estimated on data up to Feb 23, 2022 (BIC selection in Tables 4, 5), and the Fixed Naive approach using average parameters from the year before the shock. These models generate static forecasts for the entire post-shock period.

Table 4. GARCH(p,q) Model Comparison (Pre-Shock Data, Sorted by BIC)

p	q	AIC	BIC	LogLikelihood	$\alpha + \beta$
1	1	3252.34	3275.75	−1621.17	0.9676
1	2	3254.26	3282.35	−1621.13	0.9509
2	1	3254.28	3282.36	−1621.14	0.9662
2	2	3255.79	3288.55	−1620.89	0.9450
1	3	3256.26	3289.03	−1621.13	0.9509
3	1	3256.28	3289.04	−1621.14	0.9662
3	2	3257.26	3294.71	−1620.63	0.9415
2	3	3257.64	3295.09	−1620.82	0.9425
3	3	3258.71	3300.84	−1620.35	0.9186

Models fitted using pre-shock data ($N = 797$) for RHM.DE returns, constant mean, Student's t-distribution.

Visual inspection of the simulation paths (not shown here) and quantitative metrics (Table 6) confirm the failure of these static approaches. Actual prices diverge sharply upwards from the simulated paths, and actual volatility remains significantly elevated compared to the low, decaying, or constant volatility forecasts generated by these models using only pre-shock parameters. This demonstrates the inadequacy of relying solely on pre-shock information after a major structural break has occurred.

3.3. **Post-Shock Simulation/Forecast with Adaptive/Dynamic Parameters.**

Next, we examine models that adapt using post-shock data. This includes Adaptive GARCH/GJR (re-estimated daily), Dynamic Naive (using rolling T-1 year or T-1 day data), and Adaptive standard ML models (Sec 2.7, best: Optimized LGBM with rolling features, for volatility

TABLE 5. GJR-GARCH(p,1,q) Model Comparison (Pre-Shock Data, Sorted by BIC)

p	o	q	AIC	BIC	γ
1	1	1	3244.49	3272.57	0.0793
1	1	2	3246.20	3278.97	0.1213
2	1	1	3246.25	3279.02	0.0801
2	1	2	3247.72	3285.17	0.1253
3	1	1	3248.11	3285.56	0.0829
1	1	3	3248.20	3285.65	0.1213
3	1	2	3249.67	3291.79	0.1238
2	1	3	3249.72	3291.85	0.1252
3	1	3	3251.67	3298.47	0.1238

Models fitted using pre-shock data ($N = 797$) for RHM.DE returns, constant mean, GJR-GARCH($o = 1$), Student's t-distribution.

TABLE 6. Evaluation Metrics for Fixed Pre-Shock Simulations/Forecasts

Model (Fixed Pre-Shock Params)	Price Path		Volatility		
	RMSE	MAE	RMSE	MAE	QLIKE
GARCH(1,1) Fixed	320.8753	235.1205	20.3088	13.3936	0.7943
GJR-GARCH(1,1,1) Fixed	336.6622	248.7592	21.4038	14.4542	0.9942
Fixed Naive (Avg Pre-Shock)	343.7507	255.4850	24.3017	17.9503	1.7356

Lower values indicate better performance. Price metrics compare median simulated price to actual price. Volatility metrics compare model's implied/used volatility to actual realized volatility (proxy). QLIKE is for variance/volatility forecast quality. Volatility metrics are for annualized volatility (%).

only). We analyze the simulation paths (where applicable) and volatility forecasts generated by these adaptive approaches against the actual data.

Analysis of the underlying data (plots not shown) reveals improved volatility tracking by adaptive GARCH/GJR compared to fixed models. Their conditional volatility follows the actual realized volatility more closely after incorporating post-shock information. The Dynamic Naive volatility also adapts but might appear lagged compared to the GARCH models. The adaptive ML models, particularly the optimized LightGBM, demonstrate strong volatility tracking capabilities, benefiting from feature engineering (rolling features). Price path simulations generated by adaptive GARCH/Naive models, while still underestimating the strong upward trend observed in the actual data, tend to be closer to reality than those from the fixed models. Quantitative performance metrics are provided in Table 7.

3.4. **Deep Learning Approach using AutoGluon (Static Forecast).** We employed AutoGluon-TimeSeries (Sec 2.6), training models solely on pre-shock data to generate static

Table 7. Evaluation Metrics for Adaptive/Dynamic Simulations/Forecasts (Excluding TimesFM)

Model (Adaptive/Dynamic Params)	Price Path		Volatility		
	RMSE	MAE	RMSE	MAE	QLIKE
GARCH(1,1) Adaptive	306.4631	220.1101	7.1723	5.0485	0.0518
GJR-GARCH(1,1,1) Adaptive	301.9481	197.8874	8.0955	5.3340	0.0611
Dynamic Naive (T-1 Year)	221.7467	160.5942	27.7763	21.9810	1.5855
Dynamic Naive (T-1 Day)	14.8315	7.3015	3.6707	1.5967	0.0167
Optimized LGBM (Rolling Feats)	—	—	3.77	1.80	0.0039

Lower values indicate better performance. Price metrics compare median simulated price to actual price. Volatility metrics compare model's conditional/used volatility to actual realized volatility (proxy). QLIKE is for variance/volatility forecast quality. Volatility metrics are for annualized volatility (%). LGBM provides volatility forecast only (—). This table excludes TimesFM results shown in Table 8.

forecasts for the entire post-shock period. The best validation model (WeightedEnsemble) was used for prediction.

3.4.1. *AutoGluon Volatility Forecasting.* The AutoGluon mean volatility forecast, evaluated against actual realized volatility, yielded the metrics presented later in Table 8. Textually, the forecast captured the generally elevated post-shock level but was smoother and less reactive than adaptive GARCH and adaptive ML models. Its performance was better than the naive models based on QLIKE but worse than adaptive GARCH and significantly worse than the optimized adaptive LGBM (on QLIKE) and the Dynamic Naive (T-1 day) (on RMSE/MAE) across all metrics. Notably, the AutoGluon leaderboard indicated **ChronosFineTuned** had a better test RMSE (14.42) than the chosen **WeightedEnsemble** (test RMSE 17.33 based on calculation, 58.03 on leaderboard), suggesting pre-shock validation was not optimal for selecting the best model for the post-shock regime.

3.4.2. *AutoGluon Price Forecasting.* The AutoGluon median price forecast significantly underestimated the post-shock trend. Performance metrics (Table 8) show higher errors than adaptive GARCH and substantially higher than Dynamic Naive (T-1 day). Similar to volatility, the test leaderboard suggested **ChronosFineTuned** (test RMSE 213.61) performed better on the post-shock data than the validation-selected **WeightedEnsemble** (test RMSE 328.92 calculated, 335.45 on leaderboard), again highlighting the challenge of forecasting through the regime shift using only pre-shock data for model selection and training.

3.5. **Deep Learning Approach using TimesFM (Adaptive Forecast).** We employed the TimesFM foundation model in an adaptive rolling forecast setting (Sec 2.8). Unlike the static AutoGluon approach, TimesFM used updated historical context windows to generate forecasts periodically throughout the post-shock period. We evaluated its performance in forecasting the price path. The quantitative results for the best configuration (step size=1 day) are presented in Table 8. This adaptive approach allowed the model to incorporate post-shock information, potentially leading to better performance compared to static forecasts.

3.6. **Comparative Analysis of Simulation/Forecast Performance.** Table 8 consolidates the evaluation metrics for all strategies, grouping them by whether they used adaptation post-shock.

Table 8. Consolidated Evaluation Metrics Post-Shock

Model / Strategy	Price Path Accuracy		Volatility Forecast Accuracy		
	RMSE$_{price}$	MAE$_{price}$	RMSE$_{vol}$	MAE$_{vol}$	QLIKE$_{vol}$
Static Forecasts (Trained/Calibrated Pre-Shock Only)					
GARCH(1,1) Fixed	320.88	235.12	20.31	13.39	0.794
GJR-GARCH(1,1,1) Fixed	336.66	248.76	21.40	14.45	0.994
Fixed Naive (Avg Pre-Shock)	343.75	255.49	24.30	17.95	1.736
AutoGluon (WeightedEnsemble)	328.92	262.49	17.33	11.26	0.084
Adaptive / Dynamic Forecasts (Using Post-Shock Data)					
GARCH(1,1) Adaptive	306.46	220.11	7.17	5.05	0.052
GJR-GARCH(1,1,1) Adaptive	301.95	197.89	8.10	5.33	0.061
Dynamic Naive (T-1 Year)	221.75	160.59	27.78	21.98	1.586
Dynamic Naive (T-1 Day)	**14.83**	**7.30**	**3.67**	**1.60**	0.017
Optimized LGBM (Rolling Feats)	—	—	3.77	1.80	**0.004**
TimesFM (Rolling, step=1d)	14.90	7.77	—	—	—

Lower values better. Price/Volatility: Forecast vs Actual (best in **bold**). AutoGluon: static pre-shock `WeightedEnsemble`. Optimized LGBM: adaptive forecast using rolling features, volatility only (—). TimesFM: adaptive rolling forecast (1-day step), price only (—). Volatility metrics are for annualized volatility (%).

Key observations from Table 8:

(1) **Failure of Static Forecasts:** All models relying solely on pre-shock information (Fixed GARCH/GJR, Fixed Naive, AutoGluon) performed poorly compared to adaptive alternatives, especially in price path simulation and volatility forecasting accuracy (except perhaps AutoGluon's QLIKE relative to Fixed GARCH/Naive). This underscores the danger of using models calibrated on pre-shock regimes after a structural break.

(2) **Benefit of Adaptability:** Introducing adaptability via daily updates (Adaptive GARCH/GJR, Adaptive ML), rolling history (Dynamic Naive), or rolling context windows (Adaptive TimesFM) yielded significant improvements.

- **Volatility Forecasting:** Adaptive models dramatically outperformed static forecasts. Among the adaptive methods, the **Dynamic Naive (T-1 day)** model achieved the lowest RMSE (3.67) and MAE (1.60), indicating strong performance in terms of average magnitude of errors. However, the **Optimized Adaptive LightGBM** with rolling features yielded the best QLIKE score (0.004). Since QLIKE is often preferred for evaluating volatility forecasts due to its asymmetric loss function [11], the LGBM result is particularly noteworthy. Both the Dynamic Naive (T-1 day) and the Optimized LGBM substantially outperformed the adaptive GARCH models (best GARCH(1,1) RMSE 7.17, QLIKE 0.052). (TimesFM volatility was not evaluated here).

- **Price Path Simulation/Forecast:** The adaptive models significantly outperformed static ones. The **Dynamic Naive model (T-1 day variant)** achieved the best price path accuracy (lowest RMSE 14.83, MAE 7.30), slightly outperforming the **Adaptive TimesFM (Rolling, step=1d)** (RMSE 14.90, MAE 7.77). Both were substantially better than adaptive GARCH models and all static models. This highlights the effectiveness of both a sophisticated foundation model (TimesFM) used adaptively and a simple adaptive naive approach for tracking the price path in this specific post-shock scenario.

(3) **Deep Learning Performance (Static vs Adaptive):** The static AutoGluon baseline, trained only pre-shock, did not match the performance of adaptive models post-shock. In contrast, the adaptive TimesFM rolling forecast delivered state-of-the-art price forecasting accuracy in this comparison (very close to the best naive model). This strongly suggests that for DL models to be effective through structural breaks, adaptive strategies (like rolling forecasts or retraining) are likely necessary. The discrepancy between validation-best and test-best models within AutoGluon further highlights the difficulty of forecasting through the shock using only pre-shock data for model selection and training.

(4) **Best Performing Model:** The best approach depends on the objective and whether adaptation is possible.

- For **volatility forecasting post-shock**, the **Dynamic Naive (T-1 day)** (best RMSE/MAE) and **adaptive Optimized LightGBM** (best QLIKE) were superior.
- For simulating/forecasting the **median price path post-shock**, the **Dynamic Naive model (T-1 day)** was most accurate, closely followed by the **Adaptive TimesFM (Rolling, step=1d)**.
- If only a **static** forecast from the shock date onwards is possible, none of the tested methods performed well for price, although AutoGluon offered the best volatility QLIKE among the static options.

(5) **Relative Performance (Naive, GARCH, ML, TimesFM):** Dynamic Naive strongly outperformed Fixed Naive. Adaptive GARCH/GJR strongly outperformed Fixed GARCH/GJR. Adaptive Optimized LGBM (best QLIKE) and Dynamic Naive T-1 day (best RMSE/MAE) strongly outperformed Adaptive GARCH for volatility

forecasting. Dynamic Naive T-1 day and Adaptive TimesFM strongly outperformed Adaptive GARCH for price forecasting. The effectiveness of rolling features for the adaptive ML models was crucial for their strong volatility performance (especially QLIKE). The adaptive nature of the TimesFM rolling forecast was key to its strong price performance.

4. Conclusions

This study empirically demonstrated the failure of Black-Scholes assumptions, particularly constant volatility, during a significant market shock affecting Rheinmetall AG. Our comparative analysis revealed the inadequacy of static forecasting models (Fixed GARCH/Naive, pre-trained AutoGluon) that rely solely on pre-shock information. Conversely, models incorporating post-shock adaptation showed marked improvements. For volatility forecasting, an adaptive Optimized LightGBM with rolling features achieved the best QLIKE score, while a simple Dynamic Naive (T-1 day) model excelled in RMSE/MAE terms, both significantly outperforming adaptive GARCH models. For price path forecasting, a Dynamic Naive (T-1 day) approach and an adaptive rolling forecast using the TimesFM foundation model yielded the most accurate results, substantially better than adaptive GARCH and all static methods. These findings highlight the critical importance of model adaptability – whether through parameter re-estimation, rolling historical windows, rolling context updates for foundation models, or adaptive ML retraining with relevant features – when forecasting through structural breaks in financial markets. Simple adaptive methods can be surprisingly effective, while sophisticated adaptive approaches like TimesFM or feature-engineered ML can offer state-of-the-art performance depending on the specific task (price vs. volatility). Relying on models trained only on pre-shock data appears highly unreliable in such scenarios.

5. Code and Implementation

The analysis was implemented in Python (versions 3.12), using `pandas`, `numpy`, `arch`, `yfinance`, `autogluon.timeseries` (v1.2), `timesfm` (v2.0), `statsmodels`, `lightgbm`, `xgboost`, `scikit-learn`, `optuna`, and `matplotlib/seaborn`. The code, including GARCH modeling, naive approaches, AutoGluon setup, adaptive ML implementations (LGBM, XGBoost, RandomForest, Ridge), TimesFM rolling forecast simulation, feature engineering experiments, and Optuna hyperparameter optimization, potentially covering analyses beyond the scope of this paper, is available at:

https://github.com/maksdebowski/mathematical_foundations_of_ML_DL.git

in the **project** directory. Details might be found in the codebase and associated README files.

Exclusive footage of one of the authors deep into a late-night coding session.

References

1. Jean-Noël Barrot and Julien Sauvagnat, *Input specificity and the propagation of idiosyncratic shocks in production networks*, The Quarterly Journal of Economics **131** (2016), no. 3, 1543–1592.
2. Fischer Black and Myron Scholes, *The pricing of options and corporate liabilities*, The Journal of Political Economy **81** (1973), no. 3, 637–654.
3. Tim Bollerslev, *Generalized autoregressive conditional heteroskedasticity*, Journal of Econometrics **31** (1986), no. 3, 307–327.
4. Rama Cont, *Empirical properties of asset returns: stylized facts and statistical issues*, Quantitative Finance **1** (2001), no. 2, 223–236.
5. Robert F. Engle, *Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation*, Econometrica **50** (1982), no. 4, 987–1007.
6. Nick Erickson, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, and Alexander Smola, *AutoGluon-Tabular: Robust and accurate automl for structured data*, Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 11700–11710.

7. Lawrence R. Glosten, Ravi Jagannathan, and David E. Runkle, *On the relation between the expected value and the volatility of the nominal excess return on stocks*, The Journal of Finance **48** (1993), no. 5, 1779–1801.
8. Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep learning*, MIT Press, 2016.
9. John C. Hull, *Options, futures, and other derivatives*, 10th ed., Pearson, 2018.
10. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, *Deep learning*, Nature **521** (2015), no. 7553, 436–444.
11. Andrew J. Patton, *Volatility forecast comparison using imperfect volatility proxies*, Journal of Econometrics **161** (2011), no. 2, 244–256.
12. Oleksandr Shchur, Caner S. Aussignac, Jan Gasthaus, and Syama Sundar Rangapuram, *AutoGluon-TimeSeries: Automl for probabilistic time series forecasting*, 2023.