Medium-Horizon Volatility Forecasting in Corn Futures

Souhil Khiat[1]

New York University
Tandon School of Engineering

December 2025

Abstract

This paper forecasts medium-horizon volatility in corn futures by predicting 30-day realized volatility 21 trading days ahead. Using Bloomberg data on futures and options (implied volatility, term structure, skew measures), macro-financial variables, and explicit seasonality controls, we compare economically grounded linear benchmarks and the market's implied volatility to LSTM-based sequence models under a strict design. The linear economic regression delivers the best overall out-of-sample accuracy and outperforms a naïve implied-volatility benchmark. A naïve LSTM trained on the full feature set performs poorly, but a disciplined pipeline with stationarity screening, top-K feature selection, and strong regularization substantially improves LSTM performance, though it still does not beat the linear model on average. A hybrid regression–LSTM that models residual dynamics yields modest improvements over standalone nonlinear models and implied volatility, but remains slightly inferior to the strongest linear benchmark. An ablation study on the post-2008 subsample finds that USDA and NDVI features do not robustly reduce average forecast errors, suggesting that most stable predictive signal at the 30-day horizon is already captured by market-implied and economically structured variables.

[1]sk13011@nyu.edu

1 Introduction

1.1 Problem Statement and Motivation

Volatility plays a central role in agricultural commodity markets, affecting hedging costs, option pricing, inventory decisions, and risk management for producers, processors, and financial intermediaries. Among these markets, corn futures occupy a particularly important position due to corn's economic significance, strong seasonality, and sensitivity to both weather conditions and government information releases. Accurately forecasting corn price volatility is therefore of first-order importance for market participants exposed to price risk and volatility risk over horizons relevant for option maturities and hedging programs.

While short-horizon volatility dynamics in commodities have been extensively studied, forecasting volatility at medium horizons, such as one month ahead, remains challenging. At this horizon, volatility is influenced not only by recent price movements but also by expectations embedded in option markets, evolving macroeconomic conditions, supply–demand fundamentals, and time-varying uncertainty related to crop development and policy announcements. In practice, corn option implied volatility is often used as a direct forecast of future realized volatility, yet the literature documents systematic biases and slow adjustments around key information events, particularly during the growing season and major USDA report releases.

Recent advances in machine learning, and recurrent neural networks in particular, have generated optimism about their ability to capture complex nonlinear and temporal dependencies that traditional econometric models may miss. In parallel, the increasing availability of alternative data, such as remote-sensing vegetation indices and detailed government report calendars, has raised the question of whether volatility forecasts can be improved by incorporating information beyond prices and implied volatility alone. However, existing studies typically examine these channels in isolation, focus on shorter horizons, or evaluate models without stringent out-of-sample comparisons against strong market-based benchmarks.

The central problem addressed in this study is therefore whether richer information sets and more flexible modeling techniques can meaningfully improve forecasts of 30-day realized volatility for corn futures relative to simple and economically motivated baselines. In particular, we investigate whether nonlinear sequence models and hybrid approaches offer incremental predictive value beyond option-implied volatility and linear economic regressions when evaluated under a strict framework. By focusing on forecast performance rather than ex post trading profits, this study aims to provide a careful assessment of what information is already embedded in market prices and where, if anywhere, machine learning and alternative data can add value for medium-horizon volatility forecasting in agricultural commodity markets.

1.2 The Challenge of Medium-Horizon (30-Day) Volatility

Forecasting volatility over a medium-term horizon presents a distinct set of challenges that differ fundamentally from both short-term and long-term volatility modeling. In financial markets, short-horizon volatility forecasts, such as daily or intraday measures, are often driven by market microstructure effects, order flow, and rapid information assimilation. At the opposite extreme, long-horizon volatility forecasts tend to rely on slow-moving macroeconomic or structural factors. The 30-day horizon considered in this study lies between these two regimes and inherits difficulties from both.

In agricultural commodity markets, medium-horizon volatility is particularly hard to predict because it reflects the cumulative impact of heterogeneous information arriving over time. Weather conditions, crop development, planting and pollination cycles, inventory updates, and government reports all influence market expectations, but their effects unfold gradually rather than instantaneously. As a result, volatility over a 30-day window does not respond to a single dominant shock, but rather to the interaction of multiple signals with different persistence and timing. Importantly, much of this information is continuously incorporated into option prices, making implied volatility a strong and economically meaningful benchmark at this horizon.

From a modeling perspective, this horizon reduces the effectiveness of many traditional tools. Autoregressive volatility models, such as GARCH-type specifications, are well suited for capturing short-term volatility clustering but often struggle to incorporate exogenous information or regime-dependent dynamics in a flexible manner. Conversely, models that rely heavily on slowly evolving fundamentals may fail to react sufficiently to changes in market sentiment or risk conditions that materialize within a month. This creates a narrow window in which additional information may improve forecasts beyond what is already embedded in market-implied volatility.

Another key difficulty arises from the signal-to-noise ratio. While medium-horizon volatility aggregates information over several weeks, it remains highly variable and subject to abrupt changes, particularly around weather shocks or major information releases. This makes it challenging to distinguish persistent predictive signals from transient noise, especially when using high-dimensional feature sets. In such settings, more complex models risk overfitting unless carefully regularized and evaluated using strict out-of-sample procedures.

These challenges motivate the methodological choices adopted in this study. The use of a 30-day realized volatility target reflects a horizon that is economically meaningful yet empirically demanding, and one for which implied volatility constitutes a natural and difficult-to-beat benchmark. It provides a stringent test for both linear economic models and nonlinear sequence-based architectures. Any improvement over simple market-implied or economically grounded baselines at this horizon would therefore represent genuine predictive value rather than a mechanical exploitation of short-term persistence.

1.3 Research Objectives and Contribution

The primary objective of this study is to investigate whether non-linear sequence models and hybrid architectures can extract incremental predictive value for corn futures volatility at the medium-term (30-day) horizon, a window often overlooked in favor of short-term GARCH dynamics.

To the best of our knowledge, this research introduces two primary novelties to the agricultural forecasting literature, particularly concerning the medium-term horizon. First, we propose a unified forecasting framework that jointly models corn volatility using a high-dimensional feature set comprising option-implied signals, USDA information arrivals, and physical crop-stress proxies (NDVI). While existing studies typically analyze these channels in isolation, our approach allows us to capture interactions between market expectations and fundamental supply shocks. Second, we implement and evaluate a Hybrid Regression–LSTM model tailored to the corn market. This architecture is designed to bridge the gap between economically structured linear benchmarks and deep sequence learning, specifically testing whether the residual dynamics of corn volatility contain non-linear patterns that alternative data can resolve. By evaluating these models under a strict temporal validation design, this study seeks to document the limits of model complexity and the actual utility of "alt-data" for medium-term risk management.

2 Literature Review

The current literature on agricultural volatility follows a clear progression from traditional GARCH and HAR specifications toward hybrid deep learning architectures, such as GARCH-LSTM and wavelet-based models, which aim to capture complex non-linearities in grain prices [3, 6, 9]. A consistent theme across these studies is the importance of market-implied volatility and model-free variance as the strongest predictors of future realized volatility, often outperforming purely historical benchmarks at the one-month horizon [2, 11]. However, while these market-based measures are economically efficient, they frequently exhibit systematic biases and slow adjustments around critical information arrivals, particularly during the peak growing season or following major USDA announcements [8, 12].

Regarding information sources, research has identified several exogenous drivers of volatility, including seasonal weather cycles, remote-sensing indices such as NDVI, and government report surprises [4, 7, 8]. However, these data sets are frequently analyzed in isolation; for instance, remote-sensing data is typically leveraged for yield forecasting [1, 7], while volatility research often focuses strictly on price-based or implied measures. To the best of our knowledge, few studies have integrated these heterogeneous channels to forecast volatility specifically at a medium-term duration. This study seeks to bridge these separate domains by evaluating whether a unified sequence-modeling framework, incorporating market-implied signals, physical fundamentals, and structural news, provides incremental predictive value for 30-day corn volatility.

2.1 Volatility Forecasting in Agricultural Commodities

A fundamental debate in commodity markets concerns the relative predictive power of market-based expectations versus historical econometric models. At the medium-term horizon, Simon, D. [11] provides evidence that corn, soybean, and wheat option-implied volatility measured four weeks prior to expiry serves as a robust predictor of realized futures volatility, frequently outperforming seasonal GJR-GARCH specifications. This suggests that at a 30-day horizon, the forward-looking information embedded in option prices effectively anchors market expectations. Complementing this, A. Triantafyllou et al.[2] highlights that model-free implied variance and skewness for maize and other grains significantly outperform historical variance for forecasting realized volatility, while further noting that the variance risk premium (VRP) carries meaningful predictive power for future returns. Together, these findings establish that for agricultural commodities, market-implied measures are not only more accurate than simple historical lags but also encapsulate risk-neutral expectations of the "jump risk" and seasonal spikes characteristic of the growing season [5].

2.2 LSTM and Sequence Models in Financial Time Series

The shift toward deep learning in finance is driven by the need to capture non-linear temporal dependencies that traditional linear models often overlook. Recent research identifies hybrid models as an especially promising and relatively novel frontier in volatility forecasting. For instance, studies on Indian commodity markets demonstrate that hybrid architectures combining LSTMs with traditional volatility components tend to perform exceptionally well, consistently beating both pure neural networks and standalone GARCH models [9].

This trend is further supported in the context of grain markets, where hybrid GARCH-bidirectional LSTM frameworks have been shown to forecast white maize futures volatility more accurately than either standalone deep learning or pure GARCH models across several forecasting horizons [6]. Beyond volatility specifically, LSTM-based recurrent neural networks have proven superior to classical time-series models for corn futures price forecasting [10]. However, the literature notes a critical caveat: while these sequence models are highly flexible, their success in agricultural contexts often necessitates explicit de-seasonalizing and detrending of the data to account for the structural supply-and-demand cycles that characterize commodity markets [3, 10].

2.3 The Role of Alternative Data (NDVI and Government Reports)

The integration of alternative data sources addresses specific information asymmetries that market-implied measures often fail to capture instantaneously. Government reports are a primary driver of these dynamics; specifically Yang, Y. et al. [12] found that USDA crop reports, particularly those released in August, carry significant fundamental information regarding at-harvest corn volatility. Crucially, option-implied volatility has been shown to be biased and slow to adjust for several days following these releases, suggesting that the incorporation of report-day features could allow a model to exploit volatility underpricing during key information cycles.

Beyond administrative data, remote sensing offers a physical proxy for market risk. Satellite-based indicators, such as the Normalized Difference Vegetation Index (NDVI), have been shown to sharply improve forecasts of U.S. crop vegetation health compared to standard climatology [7]. Given that vegetation health is a primary determinant of yield uncertainty and, by extension, price volatility, these indices provide a valuable exogenous signal. By bridging the gap between physical crop conditions and market expectations, the inclusion of NDVI and USDA surprises allows the forecasting framework to condition volatility on the structural drivers of supply-side risk rather than relying solely on historical price action.

3 Dataset Description

All data used in this study are sourced from Bloomberg, ensuring consistency across market prices, option-implied measures, macroeconomic variables, and alternative data indicators. Bloomberg's integrated datasets allow for precise temporal alignment and standardized definitions across futures, options, government releases, and remote-sensing proxies.

3.1 Base Model Variables

The core asset analyzed in this study is the front-month (C1) and second-month (C2) corn futures contracts, retrieved from Bloomberg. For each maturity, we collect a comprehensive set of market variables including prices (open, high, low, last), trading activity (volume, open interest), historical volatility measures (10-, 30-, and 60-day), moving averages, and option-implied volatility metrics. In particular, at-the-money implied volatility and 25-delta call and put implied volatilities at the 1-month and 2-month horizons are used to characterize the implied volatility term structure and skew. All implied volatility and delta-based measures are computed directly by Bloomberg using standardized option pricing conventions and are treated as exogenous inputs to the forecasting models.

In addition, the spread between the first and second nearby contracts (C1–C2) is included as a term-structure variable. This spread serves as a proxy for market tightness and inventory expectations, capturing conditions such as backwardation or contango that are known to be associated with shifts in risk premia and volatility regimes in agricultural commodity markets.

The primary target variable in this study is the 30-day realized volatility of front-month corn futures (RV30). We utilize the Bloomberg-computed realized volatility series to minimize estimation discrepancies and ensure methodological consistency with the option-implied volatility benchmarks used elsewhere in the analysis. To align the target with a one-month forecasting horizon, we define the prediction target y_t as the realized volatility reported 21 trading days into the future. Formally, if $\widetilde{\mathrm{RV30}}_t$ denotes the realized volatility observed on date t, the target is defined as:

$$y_t \;=\; \widetilde{\mathrm{RV30}}_{t+21}.$$

This structure ensures that features available at time t are used to predict the volatility regime of the subsequent calendar month. Consistent availability for these option-implied volatility metrics begins on October 28, 2005; consequently, we define the start date of our primary dataset as late 2005 to maximize the historical training window for our deep learning models while ensuring complete feature coverage.

3.2 Additional Feature Universe (Macro)

The traditional macroeconomic feature set consists of widely used financial indicators sourced from Bloomberg. Specifically, we include the U.S. Dollar Index (DXY Curncy), short-term U.S. interest rates via the 1-month and 3-month Treasury yields (USGG1M Index and USGG3M Index), and long-term inflation expectations proxied by the 10-year U.S. breakeven inflation rate (USGGBE10 Index). These variables capture macro-financial conditions that are known to affect commodity prices and volatility through multiple channels. Movements in the U.S. dollar directly influence globally traded commodities such as corn, as prices are denominated in USD. Short-term interest rates reflect funding conditions and risk-free discounting, which impact speculative positioning and carry trades in futures markets. Finally, breakeven inflation serves as a forward-looking measure of inflation expectations, closely linked to real asset demand and commodity price dynamics. All series are obtained directly from Bloomberg and used in their standardized form.

3.3 Alternative Data Sources

3.3.1 USDA Event Releases

USDA event releases are incorporated using Bloomberg's commodity and economic event dataset. Bloomberg provides a standardized calendar of scheduled agricultural announcements, including releases issued by the U.S. Department of Agriculture, together with precise release dates and commodity relevance tags. From this event universe, we apply a commodity-level filter to retain only announcements explicitly related to corn markets. These USDA-related events are then merged with the daily futures and options data and used as exogenous indicators to capture periods of heightened information arrival in the corn market.

In addition to discrete event dates, we also include a set of USDA-related Bloomberg indices, such as HARVCORN, GRINCORN and related series, which summarize information on crop conditions, harvest progress, and supply–demand dynamics. These indices provide a continuous, market-standard representation of USDA information and expectations, complementing the event-based indicators. All such series are sourced directly from Bloomberg and aligned to the daily frequency of the futures data.

3.3.2 State-Level Normalized Difference Vegetation Index (NDVI)

While our core market and macroeconomic data is available from 2005, the satellite-based NDVI series only commences on March 3, 2008. To avoid discarding nearly three years of valuable training

data[2], we exclude these alternative datasets from the primary benchmark models and reserve them strictly for a secondary ablation study restricted to the post-2008 period.

State-level vegetation conditions are captured using NDVI series sourced from Bloomberg. These NDVI indices are constructed from satellite-based remote sensing data and provide standardized measures of vegetation greenness and crop health. For this study, we collect state-level NDVI indices corresponding to major corn-producing U.S. states, including Wisconsin, South Dakota, Ohio, Nebraska, Indiana, Missouri, Michigan, Kansas, Illinois, Iowa, and Minnesota.

Each NDVI series reflects localized crop and vegetation conditions that may influence market expectations about yields and production risk. However, because individual state-level signals are highly correlated and may introduce excessive dimensionality relative to the sample size, the NDVI series are aggregated into a single composite indicator. Specifically, we construct a weighted average NDVI index across states, with weights proportional to each state's historical share of U.S. corn production. This aggregation preserves broad regional information on crop stress while reducing noise and mitigating multicollinearity concerns.

The aggregated NDVI measure is primarily used in the ablation analysis to assess whether weather- and crop-condition information provides incremental predictive power for medium-horizon corn volatility beyond market-implied and macroeconomic variables. This design allows us to isolate the contribution of remote sensing data without materially altering the core model structure.

4 Feature Selection and Engineering

Feature engineering is used to transform raw market, options, and macroeconomic variables into representations better suited for medium-horizon volatility forecasting. These transformations encode economically meaningful structure, improve numerical stability, and facilitate learning in nonlinear sequence-based models. The goal is not causal identification, but the construction of a coherent and interpretable feature space consistent with a strictly out-of-sample forecasting framework.

4.1 Economically Motivated Feature Construction

Beyond raw price and volatility series pulled from Bloomberg, several features are constructed to encode economically meaningful transformations commonly used in commodity volatility modeling. First, price-level variables are converted to percentage changes to reduce scale effects and improve stationarity. For the asset price series P_t, percentage returns are defined as:

$$r_t = \frac{P_t - P_{t-1}}{P_{t-1}}.$$

[2] A critical loss for data-hungry sequence models like LSTMs

For short-term interest rate instruments, level changes are used instead of returns to avoid division by 0:

$$\Delta y_t = y_t - y_{t-1}.$$

Futures term-structure information is captured through level and percentage spreads between the front-month ($C1_t$) and second-month ($C2_t$) contracts. These metrics serve as proxies for storage conditions and near-term supply tightness. The calendar spread is defined as

$$\text{Spread}_t = C1_t - C2_t$$

,
and its relative version as:

$$\text{SpreadPct}_t = \frac{C1_t - C2_t}{C2_t}.$$

To capture short-term changes in term structure dynamics, the first difference of the spread is also computed:

$$\Delta\text{Spread}_t = \text{Spread}_t - \text{Spread}_{t-1}.$$

Option-implied risk measures are further summarized through skew-sensitive variables such as risk reversals and butterflies. Let σ_t^{ATM} denote the at-the-money implied volatility, and $\sigma_t^{25C}, \sigma_t^{25P}$ the 25-delta call and put implied volatilities, respectively. The risk reversal (RR25_t) and butterfly (BF25_t) are defined as:

$$\text{RR25}_t = \sigma_t^{25C} - \sigma_t^{25P}, \qquad \text{BF25}_t = \frac{1}{2}\left(\sigma_t^{25C} + \sigma_t^{25P}\right) - \sigma_t^{ATM}.$$

Relative versions normalize these quantities by the ATM level to facilitate comparison across volatility regimes:

$$\text{RR25}_t^{rel} = \frac{\text{RR25}_t}{\sigma_t^{ATM}}, \qquad \text{BF25}_t^{rel} = \frac{\text{BF25}_t}{\sigma_t^{ATM}}.$$

We also incorporate the variance risk premium (VRP), defined as the difference between option-implied and realized volatility:

$$\text{VRP}_t = \sigma_{t,30}^{IV} - \sigma_{t,30}^{RV}.$$

To capture nonlinear regime effects, the VRP is further discretized using expanding empirical quantiles. Letting $Q_{0.25,t}$ and $Q_{0.75,t}$ denote the expanding 25th and 75th percentiles, the regime state is defined as:

$$\text{Regime}_t = \begin{cases} -1, & \text{if } \text{VRP}_t \le Q_{0.25,t}, \\ 0, & \text{if } Q_{0.25,t} < \text{VRP}_t < Q_{0.75,t}, \\ 1, & \text{if } \text{VRP}_t \ge Q_{0.75,t}. \end{cases}$$

We apply an identical expanding-window discretization to the realized volatility series itself ($\sigma_{t,30}^{RV}$) to construct a Realized Volatility Regime indicator (Regime_t^{RV}). This classifies the market into low, neutral, and high volatility states $\{-1, 0, 1\}$ based on the expanding 25th and 75th percentiles

of the historical realized volatility distribution up to time t.[3]

Seasonality is explicitly encoded using both discrete and continuous representations. Calendar effects are captured via month-based cyclical transformations:

$$\text{MonthSin}_t = \sin\left(\frac{2\pi \cdot \text{Month}_t}{12}\right), \quad \text{MonthCos}_t = \cos\left(\frac{2\pi \cdot \text{Month}_t}{12}\right).$$

In addition, agronomic regimes specific to U.S. corn production are introduced through binary indicators for planting, pollination, and harvest periods:

$$\text{Planting}_t = 1_{Month_t \in \{4,5\}}, \text{Pollination}_t = 1_{Month_t \in \{6,7\}}, \text{Harvest}_t = 1_{Month_t \in \{9,10\}}.$$

These regime variables allow the models to condition volatility dynamics on biologically and economically meaningful phases of the crop cycle, rather than relying solely on generic calendar effects. All engineered features are computed deterministically from information available at time t, ensuring consistency with the forecasting framework and preventing any look-ahead bias.

4.2 Data Partitioning and Temporal Validation

To ensure robust model evaluation and prevent look-ahead bias, the dataset is partitioned chronologically into training, validation, and test sets. Given the sequential nature of LSTM networks and the requirement for stable temporal dependencies, a standard 70/15/15 split is applied. The training set (0% to 70%) is used for weight optimization; the validation set (70% to 85%) serves for hyperparameter tuning and early stopping; and the final 15% of the data is reserved strictly for out-of-sample performance evaluation. After removing initial missing values from alternative data sources, the split results in the following date ranges:

- Training Set: 2005 – 2019
- Validation Set: 2019 – 2022
- Test Set: 2022 – 2025

This chronological split is critical for financial time series to avoid data leakage, as it ensures the model is never trained on information that would be chronologically "future" relative to the prediction point.

4.2.1 Addressing Outliers and Generalization

An analysis of the historical volatility distribution reveals that the selected test set contains several of the most challenging periods in the sample. Specifically, years such as 2021, 2022, and 2023 are identified as having the most significant variance risk premium (VRP) deviations and extreme tail-risk events (p_{99} volatility spikes). These periods represent structural shifts and high-uncertainty regimes that are notoriously difficult to forecast.

[3] Crucially, the use of an expanding window means that the quantile thresholds for day t are computed using only data available up to that day. This ensures the regime classification is strictly out-of-sample and prevents look-ahead bias.

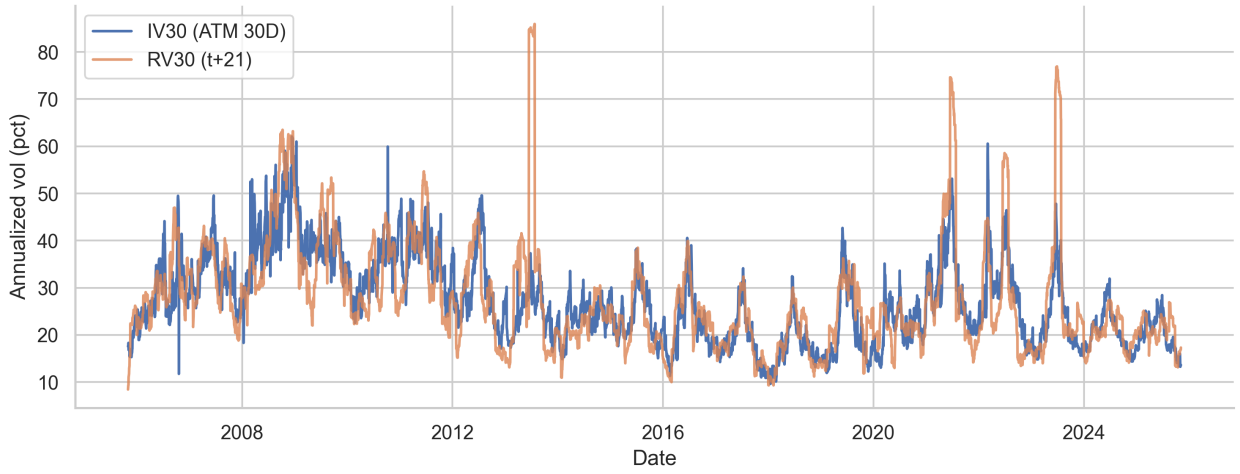

Figure 1: C1 Corn Futures: 30-Day Implied vs. 21-Day-Ahead Realized Volatility. The plot highlights the presence of extreme volatility spikes in the latter portion of the sample, particularly within the test period.

Despite the concentration of these outliers in the test set, we maintain this partition for two primary reasons. First, it maximizes the amount of historical data available to the "data-hungry" LSTM architecture, providing the model with sufficient examples of past regimes to learn complex patterns. Second, evaluating the model on a "high-difficulty" test set serves as a stringent test of generalization. A model that performs well during these tail-risk years is significantly more valuable for production-grade risk management than one that only succeeds during calm, "non-tail" market regimes. This design ensures that the reported metrics reflect the model's ability to handle genuine market stress and regime shifts.

5 Methodology and Model Architectures

The empirical methodology of this study follows a hierarchical modeling approach. We first establish strong benchmark models for forecasting 30-day realized volatility, using both market-based and econometric specifications that are standard in the volatility forecasting literature. These benchmarks serve as reference points to assess whether more complex machine-learning models deliver genuine incremental predictive value. Once robust baselines are identified, we evaluate nonlinear sequence-based models, including LSTM architectures and hybrid specifications, with the explicit goal of outperforming the best-performing benchmark models rather than weaker or ad hoc comparators.

5.1 Baseline Models

Prior to implementing deep learning models, we estimate a broad set of baseline volatility forecasting models, including historical autoregressive specifications (e.g., HAR-type models), regularized linear regressions (Ridge and Lasso), implied-volatility-based forecasts, and economically motivated regressions combining market-implied volatility with macro and term-structure variables. Model selection among these candidates is conducted using out-of-sample validation performance. Only two

strong-performing baseline models are retained for comparison with LSTM and hybrid approaches, ensuring that any reported performance gains are economically and statistically meaningful.[4]

5.1.1 Naive Market-Implied Volatility

As a primary benchmark, we consider a naive market-implied volatility forecast. Let IV_t denote the at-the-money option-implied volatility with approximately 30 days to maturity observed at time t. The naive forecast assumes that implied volatility is an unbiased predictor of future realized volatility over the same horizon, such that:

$$\widehat{\mathrm{RV}}^{\mathrm{IV}}_{t,t+30} = \mathrm{IV}_t$$

This benchmark is well motivated theoretically. Under standard no-arbitrage arguments and risk-neutral pricing, option-implied volatility reflects the market's expectation of future variance. Empirically, implied volatility has been shown to be a strong predictor of future realized volatility in commodity markets and often outperforms purely historical volatility models.

In our empirical analysis, this naive implied-volatility forecast emerges as one of the strongest baseline models, providing a demanding benchmark for more complex econometric and machine-learning approaches. Any improvement beyond this baseline therefore represents economically meaningful predictive content rather than mechanical gains from model flexibility.

5.1.2 Linear Economic Regression

Our main econometric benchmark is a linear regression that maps a small, economically motivated feature set to next-month realized volatility. Let y_t denote the 30-day realized volatility target. We estimate:

$$y_t = \alpha + \beta^\top x_t^{\mathrm{econ}} + \varepsilon_t$$

where x_t^{econ} includes (i) market-implied and historical volatility signals and (ii) simple seasonality or regime indicators.

Concretely, the economic feature block contains:

- Options and realized volatility anchors: IV_t and RV_t
- Short and medium historical volatility measures on corn futures: VOL_t^{10D}, VOL_t^{60D}
- A variance risk premium proxy: $\mathrm{VRP30}_t$
- Term-structure information from futures: $\mathrm{spread}_t = C1_t - C2_t$ and $\mathrm{spread\_pct}_t$

[4] Across the full benchmark set (Appendix 8), Lasso achieves the lowest test MAE, but the economic regression is retained as the primary linear benchmark because it is more interpretable and uses a small, economically motivated feature set, while delivering very similar out-of-sample performance.

- Calendar or regime dummies: month/season/regime indicators.

We standardize regressors using a StandardScaler and fit an OLS Linear Regression. This regression serves as a strong baseline because it captures the core economic drivers of medium-horizon volatility, especially information already embedded in implied volatility, volatility persistence, and predictable seasonal patterns in agricultural markets.

5.1.3 Baseline LSTM

As a nonlinear benchmark, we estimate a baseline LSTM model trained on the full set of available numeric features, without prior feature selection or systematic hyperparameter optimization. All input variables are scaled using Min–Max normalization, and the model is trained using a fixed 21-day lookback window. The network consists of a single LSTM layer with 64 hidden units, followed by a dropout layer with rate 0.2 and a linear output layer. Training is performed using the Adam optimizer with early stopping based on validation loss.

This benchmark LSTM is intentionally specified using standard rules of thumb commonly adopted in applied deep-learning settings, rather than through data-driven tuning. Its purpose is not to represent a best-in-class neural architecture, but to provide a reference point for assessing whether structured feature engineering, dimensionality reduction, and targeted hyperparameter selection materially improve LSTM-based volatility forecasts relative to a naïve high-capacity sequence model.

5.2 LSTM Framework

To model the dynamics of medium-horizon corn volatility, we employ Long Short-Term Memory (LSTM) neural networks, a class of recurrent neural networks designed to capture nonlinear temporal dependencies in sequential data. LSTMs are particularly well suited for financial time series where predictive signals may persist over multiple periods and interact in a nonlinear manner.

In this study, the LSTM is used to map a 21 days rolling window of past observations into a forecast of future 30-day realized volatility. The objective is not to replace economically interpretable benchmarks, but to assess whether a flexible sequence-based model can extract incremental predictive information from a rich, high-dimensional feature set.

Prior to sequence construction, all LSTM input features are scaled using a Min–Max normalization fitted on the training sample and applied consistently to the validation and test sets. This transformation maps each feature to the $[0, 1]$ interval, ensuring comparable magnitudes across heterogeneous inputs such as returns, volatility measures, spreads, and macroeconomic variables.[5]

[5]The target variable is scaled separately using the same approach and inverse-transformed after prediction to recover forecasts in the original volatility units.

5.2.1 LSTM Architecture and Gating Logic

Unlike standard feedforward networks, LSTMs process input sequences $(x_1, ..., x_T)$ by maintaining an internal memory state C_t, which serves as an information accumulator over time. This design is specifically chosen to address the vanishing gradient problem, allowing the model to capture both short-term shocks (e.g., daily price moves) and long-term seasonal regimes (e.g., harvest cycles) that characterize agricultural volatility.

The core logic of the LSTM relies on three "gates" that regulate the flow of information into and out of the cell state C_t.[6] Formally, at each time step t, the network computes:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad \text{(Forget Gate)} \tag{1}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad \text{(Input Gate)} \tag{2}$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad \text{(Candidate Update)} \tag{3}$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad \text{(Cell State Update)} \tag{4}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad \text{(Output Gate)} \tag{5}$$

$$h_t = o_t * \tanh(C_t) \quad \text{(Hidden State/Output)} \tag{6}$$

Economic Interpretation of the Gates:

- The Forget Gate (f_t): This gate determines how much of the past market history should be discarded. In the context of volatility forecasting, this allows the model to "forget" transient shocks (e.g., a one-day weather panic) once they become irrelevant, preventing noise accumulation.
- The Input Gate (i_t): This controls the extent to which new information x_t (e.g., today's USDA announcement or VRP change) alters the long-term memory. A high value indicates that current news is structurally significant for future volatility.
- The Output Gate (o_t): This gates the final prediction h_t, ensuring that the forecasted realized volatility reflects both the long-term regime (C_t) and immediate market conditions.

5.2.2 Feature Selection Framework

The initial feature universe is intentionally broad, combining market, options, macroeconomic, seasonal, and alternative data variables. However, with a sample starting in 2005, this richness creates a trade-off between expressive power and noise accumulation, particularly for sequence-based models. The benchmark LSTM trained directly on the full feature set without prior filtering exhibited very poor out-of-sample performance, consistent with overfitting and the sensitivity of recurrent architectures to irrelevant or redundant inputs. This motivated a disciplined feature selection framework

[6] LSTMs augment a standard RNN with a persistent cell state and gating mechanisms (forget, input, and output gates) that control information flow and help mitigate vanishing gradients, enabling the model to learn longer-range dependencies.

aimed at retaining economically meaningful signals while limiting algorithmic complexity and noise.

As a first step, all candidate features were subjected to Augmented Dickey–Fuller (ADF) tests on the training sample. This diagnostic revealed that many price-level variables and moving-average features were non-stationary at conventional significance levels. Since their predictive content is largely captured by the returns and volatility transformations, these non-stationary series were excluded from the LSTM feature set. This step ensured that the sequence model was trained primarily on stationary or weakly stationary inputs, improving numerical stability and learning efficiency.

To rank the remaining features by predictive relevance, a Random Forest regressor was trained on the cleaned training set. Feature importance was computed using permutation importance, which measures the marginal deterioration in forecast accuracy when a feature is randomly permuted. This approach captures nonlinear dependencies and interaction effects while remaining model-agnostic. The resulting ranking provides an ordered list of features according to their contribution to predicting future realized volatility.

The optimal dimensionality of the LSTM input was then determined through a Top-K selection procedure. For values of $K \in \{10, 20, 30, 40, 50, 60, 70\}$, the top-K ranked features were used to train a simple, fixed LSTM architecture consisting of a single LSTM layer with 32 units, followed by a dropout layer (rate 0.3) and a linear output layer. This architecture was deliberately kept simple to ensure that differences in validation performance across K reflected the information content of the feature set rather than architectural or optimization choices. Each configuration was evaluated over multiple runs with controlled random seeds, and validation mean absolute error (MAE) was averaged across runs to assess robustness.

This procedure identified $K = 30$ as the best-performing feature subset, followed really closely by $K = 10$. Fixing this dimensionality, a final hyperparameter search was conducted over a small grid of LSTM sizes $\{16, 32, 64\}$ and dropout rates $\{0.2, 0.5\}$. The selected "Optimized" configuration uses 16 LSTM units and a dropout rate of 0.5, striking a balance between model capacity and regularization. This final architecture and feature set are used consistently in all subsequent LSTM and hybrid-model experiments.

Throughout this procedure, a rolling lookback window of 21 trading days is used to construct the LSTM input sequences. This choice follows common practice in volatility forecasting, where one trading month is treated as a natural unit of information aggregation. While this heuristic is standard and economically intuitive, future work could treat the lookback length as a tunable hyperparameter and determine it using more formal, data-driven selection criteria.

5.2.3 Regularization and Early Stopping

To mitigate overfitting and ensure robust out-of-sample performance, several regularization mechanisms are incorporated into the LSTM training procedure. First, dropout is applied to the LSTM hidden state, randomly deactivating a fraction of units during training. This prevents the network from relying excessively on any single feature or temporal pattern and encourages the learning of more stable representations. Dropout rates are treated as hyperparameters and selected through validation-based tuning.

Second, early stopping is employed during training, monitoring validation loss and halting optimization when no further improvement is observed. This prevents the model from over-optimizing on the training sample once generalization performance begins to deteriorate. When early stopping is triggered, the model weights corresponding to the lowest validation loss are restored.

Finally, all model selection and hyperparameter tuning decisions are based exclusively on validation-set performance, with the test set held out until the final evaluation. Together, these design choices ensure that improvements in forecasting accuracy reflect genuine predictive structure rather than overfitting or training noise.

5.2.4 Hybrid Regression-LSTM Methodology

The hybrid model decomposes the 30-day realized volatility target into two components:
(i) an economically interpretable linear forecast, and
(ii) a residual component capturing the remaining variation unexplained by the linear model.

The central hypothesis is that the linear regression absorbs the dominant structured signal (options-implied volatility, volatility term structure, variance risk premium, calendar spreads, and seasonality controls), while a sequence model can learn nonlinear, regime-dependent dynamics in the unexplained residual.
Step 1: The usual economic regression model we used as a benchmark is fitted again. This produces the values $\hat{y}_t^{econ}$ on the training and validation sets, and out-of-sample predictions on the test set using the previously computed baseline forecasts.
Step 2: Residuals are constructed in daily time as: $r_t = y_t - \hat{y}_t^{econ}$, where y_t denotes the forecasting target (the forward-shifted 30-day realized volatility). Residuals are computed separately for the training, validation, and test samples.
Step 3: An LSTM is trained to forecast the residual series r_t using a 21-day rolling input window. The model inputs consist of the top-K features (with $K = 30$) selected from the Random-Forest permutation-importance ranking and scaled using the same feature scaler as the main LSTM pipeline. The residual target is scaled separately using a Min–Max scaler fitted on training residuals only. The LSTM architecture is kept identical to the base specification fine-tuned precedently to ensure comparability across models.
Step 4: Hybrid forecast construction At test time, the residual-LSTM produces predicted residuals $\hat{r}_t$, which are inverse-transformed back to residual units. The final hybrid forecast is obtained by

recombining the linear and nonlinear components: $\hat{y}_t^{hyb} = \hat{y}_t^{econ} + \hat{r}_t$.

Because the LSTM relies on a 21-day lookback window, all hybrid forecasts are aligned after trimming the first 21 observations of the test set. Model performance is evaluated out-of-sample using MAE and RMSE.

5.3 Hands-On Example: Linear Regression (OLS) with Standardization

This section provides a practical, fully worked example of a linear regression model similar to our empirical baseline. The goal is to show, on a tiny pseudo-dataset, how the pipeline works end-to-end:

- Fit the scaler on the training set only;
- Standardize train and test with training statistics;
- Fit OLS using the closed-form solution $\hat{\beta} = (X'X)^{-1}X'y$;
- Predict on the test point;
- Interpret how changing inputs affects the forecast.

5.3.1 Dataset Description

Features

- **IV30**: 30-day implied volatility (%, annualized)
- **RV30**: 30-day realized volatility (%, annualized)

Target

- y: realized volatility in 21 days (%, annualized)

5.3.2 Sample Data

We use four fictional observations. We keep the dataset very small for hand computation.

Table 1: Pseudo dataset for the hands-on OLS example

Row	Split	IV30	RV30	Target y
1	Train	22	20	21
2	Train	28	24	26
3	Train	26	21	23
4	Test	18	16	17

5.3.3 Standardization (Fit on Train Only)

We fit a StandardScaler on the **training rows only**.

Training means

$$\mu_{IV} = \frac{22 + 28 + 26}{3} = 25.333, \qquad \mu_{RV} = \frac{20 + 24 + 21}{3} = 21.667.$$

Training standard deviations StandardScaler uses population variance (divide by n):

$$\sigma_{IV} = \sqrt{\frac{(22 - 25.333)^2 + (28 - 25.333)^2 + (26 - 25.333)^2}{3}} = 2.494,$$

$$\sigma_{RV} = \sqrt{\frac{(20 - 21.667)^2 + (24 - 21.667)^2 + (21 - 21.667)^2}{3}} = 1.700.$$

Standardization rule The same training statistics are applied to train and test:

$$z_{IV} = \frac{IV30 - \mu_{IV}}{\sigma_{IV}}, \qquad z_{RV} = \frac{RV30 - \mu_{RV}}{\sigma_{RV}}.$$

5.3.4 Standardize the Data

Training rows (standardized)

Row 1 $(22, 20)$:

$$z_{IV} = \frac{22 - 25.333}{2.494} = -1.336, \qquad z_{RV} = \frac{20 - 21.667}{1.700} = -0.981.$$

Row 2 $(28, 24)$:

$$z_{IV} = \frac{28 - 25.333}{2.494} = 1.069, \qquad z_{RV} = \frac{24 - 21.667}{1.700} = 1.373.$$

Row 3 $(26, 21)$:

$$z_{IV} = \frac{26 - 25.333}{2.494} = 0.267, \qquad z_{RV} = \frac{21 - 21.667}{1.700} = -0.392.$$

Test row (standardized using training statistics)

Row 4 $(18, 16)$:

$$z_{IV} = \frac{18 - 25.333}{2.494} = -2.940, \qquad z_{RV} = \frac{16 - 21.667}{1.700} = -3.334.$$

5.3.5 OLS by Hand on Standardized Features

We fit the regression on standardized features with an intercept:

$$\hat{y} = \beta_0 + \beta_1 z_{IV} + \beta_2 z_{RV}.$$

Design matrix and target vector (train only)

Each row is $[1, z_{IV}, z_{RV}]$:

$$X = \begin{bmatrix} 1 & -1.336 & -0.981 \\ 1 & 1.069 & 1.373 \\ 1 & 0.267 & -0.392 \end{bmatrix}, \qquad y = \begin{bmatrix} 21 \\ 26 \\ 23 \end{bmatrix}.$$

We compute $\hat{\beta} = (X'X)^{-1}X'y$.

Step 1: Compute $X'X$ By definition, each entry is a sum over training rows:

$$\begin{aligned}
(1,1) &= \sum 1^2 = 3, \\
(1,2) &= \sum z_{IV} = (-1.336) + (1.069) + (0.267) = 0, \\
(1,3) &= \sum z_{RV} = (-0.981) + (1.373) + (-0.392) = 0, \\
(2,2) &= \sum z_{IV}^2 = (-1.336)^2 + (1.069)^2 + (0.267)^2 = 2.998946, \\
(3,3) &= \sum z_{RV}^2 = (-0.981)^2 + (1.373)^2 + (-0.392)^2 = 3.001154, \\
(2,3) &= \sum z_{IV} z_{RV} = (-1.336)(-0.981) + (1.069)(1.373) + (0.267)(-0.392) = 2.673689.
\end{aligned}$$

So:

$$X'X = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 2.998946 & 2.673689 \\ 0 & 2.673689 & 3.001154 \end{bmatrix}.$$

Step 2: Compute $X'y$

$$\begin{aligned}
\sum y &= 21 + 26 + 23 = 70, \\
\sum z_{IV} y &= (-1.336)(21) + (1.069)(26) + (0.267)(23) = 5.879, \\
\sum z_{RV} y &= (-0.981)(21) + (1.373)(26) + (-0.392)(23) = 6.081.
\end{aligned}$$

So:

$$X'y = \begin{bmatrix} 70 \\ 5.879 \\ 6.081 \end{bmatrix}.$$

Step 3: Compute $(X'X)^{-1}$ Because the intercept column is orthogonal to standardized features in this toy dataset, $X'X$ splits into an intercept block [3] and a 2×2 block

$$A = \begin{bmatrix} 2.998946 & 2.673689 \\ 2.673689 & 3.001154 \end{bmatrix}.$$

The inverse of the intercept block is $1/3 = 0.333333$.
Determinant:

$$\det(A) = 2.998946 \cdot 3.001154 - (2.673689)^2 = 9.000000 - 7.148314 = 1.851686.$$

Thus:

$$A^{-1} = \frac{1}{1.851686} \begin{bmatrix} 3.001154 & -2.673689 \\ -2.673689 & 2.998946 \end{bmatrix} = \begin{bmatrix} 1.620768 & -1.443921 \\ -1.443921 & 1.619576 \end{bmatrix}.$$

Therefore:

$$(X'X)^{-1} = \begin{bmatrix} 0.333333 & 0 & 0 \\ 0 & 1.620768 & -1.443921 \\ 0 & -1.443921 & 1.619576 \end{bmatrix}.$$

Step 4: Compute $\hat{\beta} = (X'X)^{-1}X'y$ Intercept:

$$\hat{\beta}_0 = 0.333333 \cdot 70 = 23.333333.$$

Slopes:

$$\hat{\beta}_1 = 1.620768 \cdot 5.879 - 1.443921 \cdot 6.081 = 0.748011,$$
$$\hat{\beta}_2 = -1.443921 \cdot 5.879 + 1.619576 \cdot 6.081 = 1.359827.$$

So:

$$\hat{\beta} = \begin{bmatrix} 23.333333 \\ 0.748011 \\ 1.359827 \end{bmatrix}, \qquad \hat{y} = 23.333333 + 0.748011\, z_{IV} + 1.359827\, z_{RV}.$$

5.3.6 Prediction on the Test Observation

The standardized test features are $z_{IV} = -2.940$ and $z_{RV} = -3.334$, so:

$$X_{\text{test}} = \begin{bmatrix} 1 & -2.940 & -3.334 \end{bmatrix}.$$

Prediction:

$$\hat{y}_{\text{test}} = 23.333333 + 0.748011(-2.940) + 1.359827(-3.334) \approx 16.60.$$

The true value is $y = 17$.

5.3.7 Sensitivity: Change Inputs and Observe the Effect

Because the model is linear in standardized inputs:

- Increasing IV30 by $+1\sigma_{IV}$ increases z_{IV} by $+1$, so $\hat{y}$ increases by $+0.748011$.
- Increasing RV30 by $+1\sigma_{RV}$ increases z_{RV} by $+1$, so $\hat{y}$ increases by $+1.359827$.

5.4 Hands-On Example: Attempting a Hybrid LSTM on Residuals (Why It Cannot Work Here)

In a hybrid architecture, we first fit a linear model and then train an LSTM to predict the residuals of that linear model:

$$e_t = y_t - \hat{y}_t^{OLS}.$$

The idea is that OLS captures the linear component, and the LSTM learns any remaining non-linear, time-dependent structure in e_t.

5.4.1 Step 1: Residual Targets for the Hybrid Model

In a hybrid Regression + LSTM approach, the LSTM targets are the residuals:

$$y_t^{(LSTM)} = e_t = y_t - \hat{y}_t^{OLS}.$$

The LSTM inputs would normally be sequences of the past T days of features (for example IV30 and RV30), so each sample has shape:

$$X_i^{(LSTM)} \in \mathbb{R}^{T\times 2}.$$

5.4.2 Step 2: The Windowing Constraint (Sequence Sufficiency)

With a lookback window of $T = 21$, the number of training sequences that can be formed from N observations is:

$$N_{\text{seq}} = N - T.$$

In our toy dataset, the training set contains only $N = 3$ observations, so:

$$N_{\text{seq}} = 3 - 21 < 0.$$

This makes it mathematically impossible to construct even a single LSTM training sample.

5.4.3 Step 3: Even If We Force a Tiny Window, the Model Is Underdetermined

If we artificially reduce the window to $T = 2$, then with $N = 3$ we obtain only:

$$N_{\text{seq}} = 3 - 2 = 1$$

training sequence, which is not enough to train or validate any neural network model.

Moreover, an LSTM with u units and input dimension d has: $4\times((d+u)u+u)$ trainable parameters. With $d = 2$ (two features per time step) and $u = 16$: $4 \times ((2 + 16) \cdot 16 + 16) = 1216$ parameters. Attempting to fit 1216 parameters with only 1 sequence is a severely underdetermined learning problem, the model will either overfit instantly or fail to learn anything meaningful.

5.4.4 Step 4: Additional Toy-Data Issue (Residuals Are Trivially Zero)

In this toy example we used 3 training points and an OLS model with 3 parameters (intercept + 2 slopes). In that case, OLS fits the training data exactly (up to rounding), so the training residuals satisfy $e_t \approx 0$ for all training points. Therefore, even ignoring windowing, there is no residual signal available to learn.
Conclusion: This is why deep learning methods such as LSTMs require large time series datasets, while OLS can be demonstrated by hand on a tiny dataset.

---

6 Empirical Results

6.1 Baseline Performance Benchmarks

We begin by comparing the performance of the baseline models on the held-out test set using root mean squared error (RMSE) and mean absolute error (MAE) as evaluation metrics.

Table 2: Baseline Model Performance Metrics (Test Set)

Model Name	**Test MAE**	**Test RMSE**
Linear Economic Regression	3.91	5.89
Naive Market-Implied Volatility	4.42	8.11
Benchmark LSTM (Naive)	11.97	13.60

Note: The Benchmark LSTM was trained on the full feature set without optimization, illustrating the noise sensitivity of high-capacity models at this horizon.

The Linear Economic regression emerges as the strongest baseline, achieving the lowest forecast error among all benchmark models (RMSE = 5.89, MAE = 3.91). This result highlights the substantial predictive content embedded in economically motivated variables such as implied volatility, recent realized volatility, term-structure information, and seasonal indicators. It also underscores the relevance of linear economic structure for medium-horizon volatility forecasting in corn futures. The Naive Market-Implied Volatility benchmark performs reasonably well, with an MAE of 4.42, but exhibits a notably higher RMSE (8.11). This pattern suggests that implied volatility provides a useful average forecast of future realized volatility, consistent with its theoretical role as a market expectation, but tends to overreact during periods of elevated uncertainty, leading to large forecast errors in extreme episodes.

The Benchmark LSTM, trained on the full feature set without prior feature selection or targeted hyperparameter tuning, performs substantially worse than all other baselines (RMSE = 13.60, MAE = 11.97). This result illustrates the sensitivity of sequence-based models to noisy and high-dimensional inputs and confirms that naive application of deep learning architectures does not automatically yield superior performance in medium-horizon volatility forecasting.
Overall, the results demonstrate that linear models based on simple, economically motivated features and seasonality already outperform at-the-money implied volatility on a full-year basis. Beating the implied volatility benchmark with any model suggests a potential window for a profitable trading strategy, and this opportunity has already been identified by our linear baseline. Any nonlinear or hybrid approach must therefore aim to further improve upon the strong performance of the Linear Economic regression to demonstrate genuine incremental predictive value beyond these foundational signals.

6.2 LSTM vs. Linear Economic Models

Building on the baseline benchmarks, we now compare the performance of the optimized LSTM model against the Linear Economic regression. The LSTM configuration used in this comparison corresponds to the Optimized specification selected through the feature-selection and hyperparameter-tuning procedure, with $K = 30$ input features, 16 LSTM units, and a dropout rate of 0.5.

Under this configuration, the Optimized LSTM achieves a test-set MAE of 4.54 and an RMSE of 6.91. While this represents a substantial improvement relative to the naive Benchmark LSTM trained on the full feature set, the optimized LSTM does not outperform the Linear Economic regression, which remains the best-performing model in terms of both MAE and RMSE.

This result highlights an important empirical finding: once economically meaningful variables and seasonality are appropriately incorporated, a parsimonious linear model remains highly competitive for medium-horizon volatility forecasting. Although the LSTM is capable of capturing nonlinear temporal dependencies, these effects do not translate into superior out-of-sample accuracy relative to the linear benchmark at the 21-day forecasting horizon considered in this study.

We note that alternative LSTM capacities can perform better for smaller feature subsets, highlighting interactions between model complexity and feature dimensionality. A full joint optimization over network architecture and feature selection is left for future work. In the present analysis, architectural choices are deliberately fixed conditional on the selected feature dimension to ensure a transparent and disciplined model comparison.

The next figure provides a visual comparison of the Optimized LSTM forecasts against realized 30-day volatility and our 3 benchmarks on the test set. The LSTM tracks broad volatility regimes and medium-term movements reasonably well but tends to lag during abrupt volatility spikes and turning points. In contrast, the Linear Economic model exhibits more stable performance across regimes, contributing to its superior average forecast accuracy.

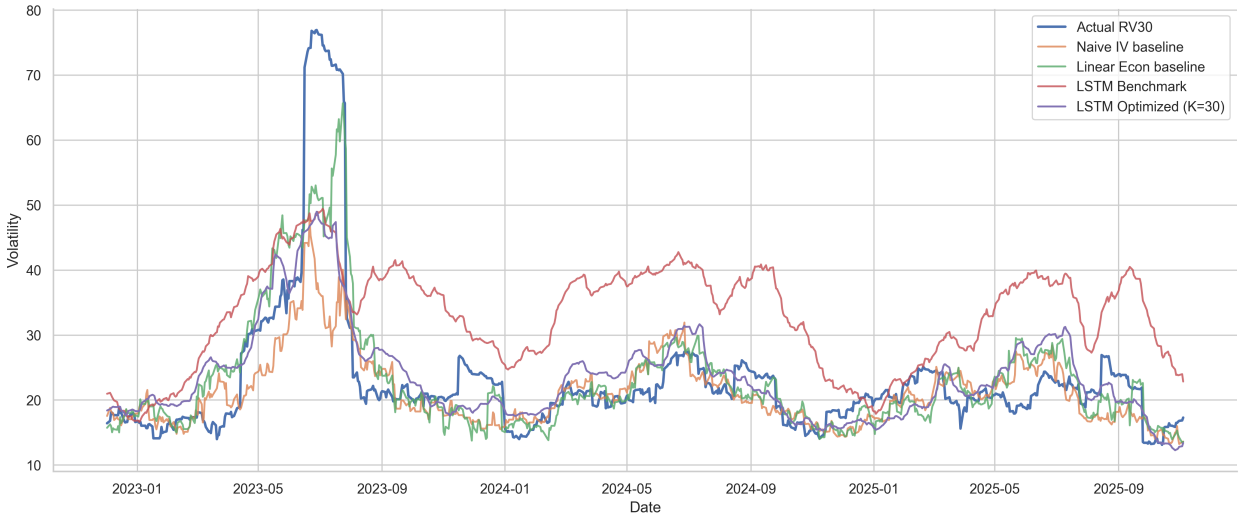

Figure 2: Out-of-Sample Volatility Forecasts: Baselines vs. Optimized LSTM ($K = 30$)

Overall, these results suggest that while carefully engineered LSTM models can meaningfully improve upon naive deep-learning baselines, nonlinear sequence models do not automatically dominate economically structured linear approaches at medium forecasting horizons. Any gains from increased model flexibility appear to be modest relative to the strong signal already captured by the Linear Economic regression.

6.3 Hybrid Model Performance

We now evaluate the performance of the hybrid modeling approach that combines the Linear Economic regression with an LSTM trained on its residuals. The hybrid forecast is constructed as the sum of the linear regression prediction and the LSTM-based residual forecast, using the same feature set ($K = 30$) and architecture as the Optimized LSTM.

Table 3 reports the out-of-sample performance of the hybrid model relative to the main benchmarks. The hybrid approach achieves a test-set MAE of 3.98 and an RMSE of 6.21. This represents a clear improvement over both the Naive Market-Implied Volatility benchmark and the standalone LSTM models, including the Optimized configuration.

Table 3: Hybrid Model Performance Comparison (Test Set)

Model Name	**Test MAE**	**Test RMSE**
Linear Economic Regression	3.91	5.89
Hybrid (Economic + Residual LSTM)	3.98	6.21
Naive Market-Implied Volatility	4.42	8.11
Final LSTM (K=30)	4.54	6.91

Despite these gains, the hybrid model does not outperform the Linear Economic regression, which remains the best-performing specification across both evaluation metrics. This suggests that while the residual LSTM is able to extract some additional nonlinear structure beyond the linear baseline,

its incremental contribution is not sufficient to consistently improve overall forecast accuracy at the 21-day horizon.

Figure 3 illustrates the time-series behavior of the hybrid forecast relative to realized volatility and the Optimized LSTM. The hybrid model tracks medium-term volatility regimes more closely than the standalone LSTM and exhibits reduced forecast errors during certain periods. However, similar to the LSTM-only models, it remains challenged by sharp volatility spikes and rapid regime transitions.

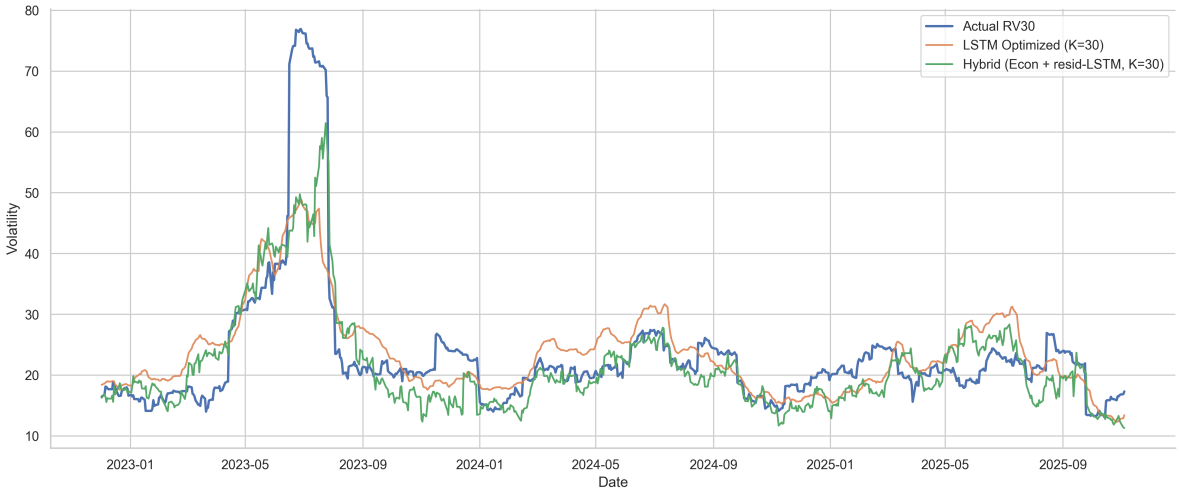

Figure 3: Test Set Volatility Forecasts: Actual RV30, Optimized LSTM, and Hybrid Model. Forecasts are aligned after the 21-day lookback window.

Overall, the hybrid results indicate that augmenting a strong linear economic model with a nonlinear residual component can yield modest improvements over purely nonlinear approaches and naive benchmarks. However, the simple Linear Economic regression remains difficult to outperform at the medium-horizon volatility horizon considered in this study. These findings motivate a cautious interpretation of hybrid deep-learning gains and highlight the robustness of economically structured linear models in agricultural volatility forecasting.

6.4 Ablation Study: Incremental Value of Alternative Data

This supplementary experiment evaluates whether USDA- and NDVI-related features add incremental predictive value beyond the baseline feature universe. A key practical constraint is that the alternative-data panel is shorter, starting on March 5, 2008; consequently, all benchmarks and LSTM-family models are re-estimated on this reduced sample to ensure a fair comparison.

We replicate the modeling pipeline used in the main experiment: identical data alignment, feature scaling, sequence construction (21-trading-day lookback), and feature selection based on Random Forest permutation importance. The alternative-data block contributes four features in the top 40:
- GCFPNPKI Index VOLATILITY 10D
- GCFPNPKI Index VOLATILITY 30D

- GCFPNPKI Index VOLATILITY 60D
- NDVI CornBelt PCT CHG

Here, *GCFPNPKI* is the Green Markets North American fertilizer price index, and `NDVI CornBelt PCT CHG` is the percentage change in the average NDVI over the U.S. Corn Belt region.

To isolate the effect of alternative data while holding model capacity and input dimensionality constant, we estimate four LSTM variants (with $K \in \{10, 40\}$ inputs):

- AUG_K10 and AUG_K40: use the top-K predictors from a Random-Forest permutation-importance ranking computed on the training set (alternative-data variables are allowed to enter if they rank in the top K).
- BASE_K10 and BASE_K40: use the top-K predictors from the same ranking after excluding any variables belonging to the alternative-data block (identified by keyword filters such as `NDVI`, `USDA`, `GCFPNPKI`, etc.).

All four models use the same "Optimized" configuration: a single-layer LSTM with 16 units, dropout of 0.5, Adam optimizer, MSE loss, and early stopping on validation loss (patience of 10, restoring best weights). This ensures the comparison reflects differences in information content rather than architecture or training procedure.

Table 4: Ablation Study Results (Reduced Sample)

Model	**Test MAE**	**Test RMSE**
LSTM_Optimized_BASE_K10	4.0034	6.8567
LSTM_Optimized_BASE_K40	4.2882	6.2235
LSTM_Optimized_AUG_K10	4.3857	6.8274
LSTM_Optimized_AUG_K40	5.2026	8.0814
Linear_Econ (Benchmark)	3.9745	6.0354
Naive_IV (Benchmark)	4.7233	8.5735

Overall, adding the USDA/NDVI block does not improve average forecast accuracy in this setup. The augmented models underperform their BASE counterparts at both $K = 10$ and $K = 40$, and the degradation is substantial for AUG_K40. However, the time-series plot in Figure 4 suggests a nuance: during the large volatility episode in mid-2023, AUG_K10 appears to track the spike more closely than BASE_K10. One plausible explanation is that fertilizer-market volatility, captured by GCFPNPKI_Index_VOLATILITY_30D (a top-ranked augmented feature), contains information about stress conditions that are not fully reflected in the baseline set.

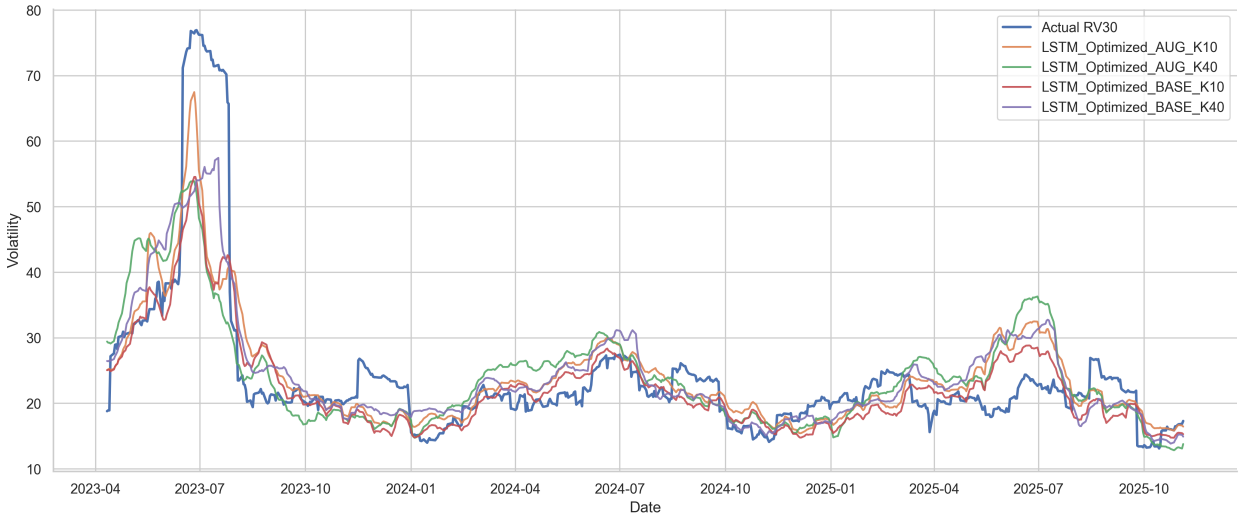

Figure 4: Ablation Study: Comparing Baseline and Augmented LSTMs

The results of this ablation support the following conclusions:

1. The strongest overall benchmark remains the simple Linear_Econ model.
2. USDA/NDVI features, as included here, do not systematically improve the LSTM's out-of-sample error metrics.
3. Some event-period dynamics may be better captured by the augmented specification (notably around the 2023 spike), but this is localized and does not translate into improved average performance.

7 Discussion

Overall, the empirical results show that simple, economically motivated linear models remain highly competitive for medium-horizon corn volatility forecasting, consistently outperforming both naive and optimized nonlinear benchmarks on average. While LSTM-based and hybrid approaches capture certain nonlinear and state-dependent dynamics, their gains are episodic rather than systematic, highlighting important limitations in extracting stable predictive value beyond established economic signals at this horizon.

7.1 Interpretation of Economic vs. Nonlinear Signals

The empirical results reveal a systematic performance gap between linear economic models and LSTM-based approaches at the 30-day forecasting horizon. Linear models achieve lower average forecast errors, indicating that a substantial share of the predictive signal is captured by stable, economically motivated variables. While LSTM models are capable of modeling nonlinear and sequential dependencies, their inferior average performance suggests that such nonlinear dynamics do not dominate volatility formation at this horizon in the available data. Nevertheless, episodic improvements observed during specific periods are consistent with the possibility that nonlinear

effects may become relevant under particular market conditions. In this regard, the hybrid modeling approach explored in this study may hold genuine predictive potential, but its implementation remains deliberately simple. Further refinement, including richer residual structures, alternative decompositions, or joint optimization strategies, was beyond the scope of this short academic paper and is left for future research.

7.2 Risks and Limitations

7.2.1 Forecast Horizon Dependence

All empirical results in this study are conditional on a fixed medium-term forecast horizon, defined as 30-day realized volatility shifted 21 trading days ahead. This choice reflects common practice in volatility forecasting and aligns with the typical maturity of short-dated options, but it also imposes a specific structure on the prediction problem. Volatility dynamics are known to vary substantially across horizons, with short-horizon forecasts dominated by persistence and clustering effects, and longer horizons increasingly influenced by seasonality, macroeconomic conditions, and regime shifts.

As a consequence, the relative performance of linear, econometric, and sequence-based models observed in this paper should not be interpreted as horizon-invariant. In particular, GARCH-type models are widely documented to perform best at short horizons, while their comparative disadvantage at medium horizons may reflect structural limitations rather than misspecification.

Future work could explicitly treat the forecast horizon as a tunable dimension of the modeling problem, re-estimating the full pipeline at alternative horizons (e.g., t+1, t+5, or t+60) to assess how model performance and the incremental value of alternative data sources vary across time scales.

7.2.2 Model Capacity and Sample Size Constraints

A central limitation of this study relates to the trade-off between model capacity and effective sample size. Although the dataset spans a relatively long calendar period (> 20 years), the usable sample is substantially reduced once medium-horizon targets, rolling lookback windows, and train–validation–test splits are applied. This constraint is particularly relevant for sequence-based models, which require learning high-dimensional temporal mappings from a limited number of independent sequences.

To mitigate overfitting risk, the LSTM architectures and hyperparameter search space were intentionally kept lightweight. While this design choice improves stability and interpretability, it may also limit the model's ability to capture more complex nonlinear interactions, especially those potentially embedded in alternative data sources. More expressive architectures, such as deeper recurrent networks, attention-based mechanisms, or Transformer-style models, may extract additional signal but would require larger datasets or stronger regularization strategies to be reliably estimated.

As a result, the empirical findings should be interpreted as conditional on a conservative modeling choice that prioritizes robustness over maximal flexibility. Future research could explore richer architectures in combination with longer samples, higher-frequency data, or cross-asset pooling to better assess the scalability of nonlinear models in agricultural volatility forecasting.

7.2.3 Feature Timing, Lags, and Information Alignment

To avoid any form of look-ahead bias, all event-based and alternative data features are incorporated with conservative timing assumptions. In particular, USDA-related variables and other announcement-based indicators are shifted such that the model only observes them on the trading day following their release. While this ensures strict information alignment, it also implies that the model does not react contemporaneously to surprise announcements, potentially understating their short-term impact on volatility.

Future research could relax this conservative treatment by explicitly modeling intraday release timestamps, announcement-time jumps, or high-frequency price responses. Access to intraday data would allow for a more precise alignment between information arrival and market reaction, and could materially improve the ability of nonlinear models to capture announcement-driven volatility dynamics.

7.2.4 Limits of the Ablation Study Design

The ablation study evaluates the incremental contribution of NDVI and USDA-related variables within the specific modeling pipeline adopted in this paper. Feature selection is performed using Random Forest permutation importance, which ranks variables based on their marginal predictive power in a non-sequential setting. While this approach is well suited for identifying strong predictors, it may fail to surface features whose value emerges primarily through nonlinear temporal interactions, which LSTM architectures are designed to exploit.

As a result, the absence of strong average gains from alternative data in the ablation study should not be interpreted as evidence that NDVI or USDA information lacks predictive content for corn volatility. Rather, it reflects the limits of a pipeline in which feature relevance is assessed outside the sequence model itself. Future work could explore joint feature selection and sequence learning, or end-to-end architectures in which alternative data are allowed to influence hidden-state dynamics without prior filtering.

7.2.5 Other limitations

Potential Extensions and Alternative Specifications:

1. While seasonality is explicitly encoded through calendar and agronomic indicators, future work could explore alternative preprocessing strategies such as detrending or seasonal adjustment prior to LSTM training. Such approaches may help isolate residual nonlinear dynamics, particularly in settings with strong periodic structure.

2. In addition, the economic value of alternative data sources such as NDVI indices and USDA releases could be assessed using simpler benchmark models, including linear or low-capacity nonlinear specifications. This would allow a clearer distinction between genuine economic signal and model-specific extraction capabilities, and help validate whether these data sources add value independently of deep sequence architectures.

8 Conclusion

This study set out to assess whether modern sequence-based machine learning models can improve the forecasting of medium-horizon (30-day) realized volatility in corn futures relative to economically grounded benchmarks. Using a comprehensive dataset combining market-implied volatility, term-structure information, macro variables, seasonal indicators, and selected alternative data sources, we implemented a disciplined modeling pipeline that emphasized out-of-sample evaluation, feature hygiene, and comparability across models.

The empirical results deliver a clear message. Simple, economically motivated linear models remain extremely competitive at the 30-day horizon, outperforming both naive deep-learning baselines and more carefully optimized LSTM architectures. While optimized LSTM models substantially improve upon naive neural-network implementations and outperform implied volatility in some configurations, they do not consistently dominate the linear economic regression. Hybrid residual-based architectures show encouraging behavior and can outperform implied volatility, but still fall short of the strongest linear benchmark in this setting.

The ablation study on NDVI and USDA-related variables highlights the difficulty of extracting incremental predictive value from alternative data at this horizon using conservative architectures and limited samples. While specific episodes suggest that such data may help anticipate extreme volatility events, these gains are not robust across configurations and sample splits. As a result, the findings should be interpreted as conditional on the modeling choices and constraints of a short academic project, rather than as definitive evidence against the usefulness of alternative data.

Beyond the quantitative results, this project provided valuable methodological insight. As a first applied machine learning project focused on financial time series, it required moving beyond mechanical model implementation and critically questioning each modeling choice, from data transformations and feature selection to evaluation design and horizon selection. While the initial ambition was to develop a trading-oriented model, the empirical evidence reinforced an important lesson: in many financial applications, simple models built on well-understood economic structure already capture a large share of the available signal, and more complex methods must be carefully justified rather than adopted by default.

Overall, this work demonstrates both the promise and the limits of nonlinear sequence models for medium-horizon volatility forecasting in agricultural markets. It suggests that future progress

may come less from increasing architectural complexity in isolation, and more from improved data alignment, horizon-specific modeling strategies, and hybrid frameworks evaluated under rigorous economic criteria.

Acknowledgement

My sincere thanks go to Prof. Aboussalah for his guidance and invaluable support throughout this semester and project. I would also like to thank Abdessalam Ed-Dib for his advice and availability throughout the project.

References

1. Adegoke, J. O., & Carleton, A. (2021). *Relations between soil moisture and satellite vegetation indices in the u.s. corn belt*. Retrieved December 18, 2025, from https://pubs.usgs.gov/publication/70024386.
2. Athanasios Triantafyllou, A. S., George Dotsis. (2015). *Volatility forecasting and time-varying variance risk premiums in grains commodity markets*. Retrieved December 18, 2025, from https://repository.essex.ac.uk/30100/1/Volatility%20Forecasting%20and%20Time-Varying%20Variance%20Risk%20Premiums%20in%20Agricultural%20Commodity%20Markets.pdf?utm_source=chatgpt.com.
3. Chengjin Yang, Z. L., Yanzhong Zhai. (2025). *Enhancing corn industry sustainability through deep learning hybrid models for price volatility forecasting*. Retrieved December 18, 2025, from https://pubmed.ncbi.nlm.nih.gov/40489457/.
4. CMEGroup. (2025). *Vol is high by the fourth of july*. Retrieved December 18, 2025, from https://www.cmegroup.com/articles/whitepapers/vol-is-high-by-the-fourth-of-july.html.
5. He, X. (2023). *The pricing of variance risks in agricultural futures markets: Do jumps matter?* Retrieved December 19, 2025, from https://academic.oup.com/erae/article/50/4/1428/7246304?login=true.
6. Jordaan, H. (2010). *Factors affecting the price volatility of july futures contracts for white maize in south africa*. Retrieved December 18, 2025, from https://www.researchgate.net/publication/233140774_Factors_affecting_the_price_volatility_of_July_futures_contracts_for_white_maize_in_South_Africa.
7. Le, M. (2024). *On the use of smap soil moisture for forecasting ndvi over conus cropland regions*. Retrieved December 18, 2025, from https://ntrs.nasa.gov/citations/20240012812.
8. Olga Isengildina Massa, S. H. I., Berna Karali. (2020). *What do we know about the accuracy and impact of usda reports in the corn market?* Retrieved December 18, 2025, from https://aaec.vt.edu/content/dam/aaec_vt_edu/faculty-research/NCGA%20Report_Final.pdf.
9. R, M. (2025). *A novel hybrid neural network-based volatility forecasting of agricultural commodity prices: Empirical evidence from india*. Retrieved December 18, 2025, from https://www.researchgate.net/publication/390704912_A_novel_hybrid_neural_network-based_volatility_forecasting_of_agricultural_commodity_prices_empirical_evidence_from_India.
10. Sckokai, P. (2024). *Machine learning to predict grains futures prices*. Retrieved December 19, 2025, from https://ideas.repec.org/a/bla/agecon/v55y2024i3p479-497.html.
11. Simon, D. (2002). *Implied volatility forecasts in the grains complex*. Retrieved December 18, 2025, from https://www.researchgate.net/publication/229480771_Implied_Volatility_Forecasts_in_the_Grains_Complex.

12. Yang, Y., & McKenzie, A. (2024). *The impacts of usda reports on pre-harvest volatility expectations: The case of corn new crop futures*. Retrieved December 18, 2025, from https://farmdoc.illinois.edu/assets/meetings/nccc134/conf_2024/pdf/Yang_McKenzie_NCCC-134_2024.pdf.

Appendices

Appendix A - Benchmark Models Performance Metrics

Table 5: Model Performance Metrics

Model	Test MAE	Test RMSE	Test R2	Test Corr	Val MAE	Val R2
LinearReg (Econ)	3.8675	5.8313	0.7153	0.8468	5.1888	0.7293
Lasso (All Feats)	3.7776	6.0575	0.6928	0.8620	5.0405	0.7548
Ridge (All Feats)	5.6003	7.1741	0.5691	0.8216	5.1823	0.7192
Naive IV Baseline	4.3717	8.0177	0.4619	0.7447	6.1030	0.5861
LinearReg (IV Only)	4.4717	8.0948	0.4515	0.7447	6.2582	0.5624
HistGBR	5.8378	8.3397	0.4178	0.6849	5.7312	0.6423
HAR-RV (Vol Only)	5.0500	8.5116	0.3935	0.6447	6.8845	0.4161
RandomForest	5.5946	8.5325	0.3905	0.6618	7.3046	0.2139
Naive RV Baseline	5.5482	10.8735	0.0102	0.5059	8.5021	0.0215

Appendix: Questions to Test Your Understanding

Q1. What caused the Hands-On LSTM (hybrid residual) example to fail?

(a) Because OLS cannot be used before an LSTM in a hybrid model

(b) Because the residuals are always non-stationary, so an LSTM cannot be trained on them

(c) Because with a 21-day lookback window and only 3 training observations, you cannot form any training sequences ($N_{\text{seq}} = N - T < 0$)

(d) Because Min–Max scaling requires at least 30 observations to work properly

Q2. How does an LSTM differ from a "classical" (vanilla) RNN?

(a) LSTMs remove recurrence entirely and behave like feedforward networks

(b) LSTMs use an explicit cell state and gating mechanisms (forget/input/output gates) to control information flow and mitigate vanishing gradients

(c) LSTMs guarantee better performance than vanilla RNNs on any dataset

(d) LSTMs do not require sequence data, they only need cross-sectional features

Q3. What is the main reason the paper uses an expanding-window quantile rule to build regime indicators (for VRP and RV)?

(a) It forces regimes to be equally frequent in train, validation, and test

(b) It makes the regime indicators stationary by construction

(c) It prevents look-ahead bias by computing thresholds at time t using only information available up to t

(d) It guarantees lower MAE for all models

Q4. Why is option-implied volatility a strong forecast of future realized volatility at a 30-day horizon?

(a) Because it is computed from past realized volatility, so it mechanically matches the future

(b) Because under no-arbitrage option pricing, implied volatility reflects the market's forward-looking expectation of future variance (risk-neutral), making it a natural benchmark

(c) Because implied volatility is always unbiased and never exhibits systematic errors

(d) Because implied volatility is independent of risk premia and market conditions

Q5. Which statement best explains why the "naive" LSTM trained on the full feature set performed so poorly out of sample?

(a) LSTMs cannot forecast volatility, they only work for price direction

(b) A large, noisy, partly non-stationary and redundant feature set increases overfitting risk without disciplined filtering and regularization

(c) Min–Max scaling always destroys macroeconomic signal

(d) LSTM performance is always worse than OLS when the horizon is 21 days

Q6. In the hybrid Regression–LSTM approach, what is the final test-set forecast $\hat{y}_t^{hyb}$ and how is it constructed? Let y_t be the target (the 30-day realized volatility observed at $t+21$), $\hat{y}_t^{econ}$ the linear economic regression forecast made using information available at time t, $r_t = y_t - \hat{y}_t^{econ}$ the residual, and $\hat{r}_t$ the LSTM forecast of that residual (trained on past sequences).

(a) $\hat{y}_t^{hyb} = \hat{r}_t$ (use only the residual forecast)

(b) $\hat{y}_t^{hyb} = \hat{y}_t^{econ} - \hat{r}_t$ (subtract the predicted residual)

(c) $\hat{y}_t^{hyb} = \hat{y}_t^{econ} + \hat{r}_t$ (add the predicted residual to the linear forecast)

(d) $\hat{y}_t^{hyb} = IV_t + \hat{r}_t$ (add residuals to implied volatility instead of the regression)

Appendix: Answers to the Questions

Q1. Correct Answer: (c)
Explanation: With a lookback window $T = 21$ and only $N = 3$ training observations, the number of sequences is $N_{\text{seq}} = N - T < 0$, so it is impossible to construct even a single LSTM training sample.

Q2. Correct Answer: (b)
Explanation: LSTMs introduce a cell state and gating mechanisms (forget/input/output gates) that regulate information flow and mitigate the vanishing-gradient problem, helping them learn longer-range dependencies than vanilla RNNs.

Q3. Correct Answer: (c)
Explanation: Expanding-window quantiles compute thresholds at time t using only data available up to t, which prevents look-ahead bias and preserves a strict out-of-sample design.

Q4. Correct Answer: (b)
Explanation: Option-implied volatility is backed out from option prices, which embed market expectations about future variance (under risk-neutral pricing). This makes it a strong, forward-looking benchmark at the 30-day horizon, even if biases can occur.

Q5. Correct Answer: (b)
Explanation: Training a high-capacity LSTM on a large, noisy feature set (including redundant and potentially unstable series) increases overfitting risk. Without disciplined filtering, feature selection, and strong regularization, out-of-sample performance degrades sharply.

Q6. Correct Answer: (c)
Explanation: The hybrid model adds the residual forecast back to the linear prediction:

$$\hat{y}_t^{hyb} = \hat{y}_t^{econ} + \hat{r}_t.$$

Here $\hat{r}_t$ is the LSTM's forecast of the regression residual $r_t = y_t - \hat{y}_t^{econ}$.