Zürcher Hochschule
für Angewandte Wissenschaften

Calibration of the Heston Model using Neural Networks

Machine Learning and Deep Learning

Authors: B. Grimus & A. Mosleh-Tehrani
Study Program: Banking & Finance
Supervisor: Dr. Fazlija Bledar

ZHAW School of Management and Law

Submitted on: 01. December 2025

Abstract

The calibration of the Heston stochastic volatility model constitutes a non-trivial inverse problem, typically constrained by the computational intensity of standard numerical pricing methods. This study evaluates the longitudinal robustness and accuracy of a Deep Differential Network (DDN) employed as a surrogate pricing engine within a hybrid calibration framework. By utilizing Sobolev training on a synthetic dataset generated via Latin Hypercube Sampling (LHS), the DDN learns to approximate both option prices and their partial derivatives, thereby facilitating rapid convergence within a classical L-BFGS-B optimization routine. A comprehensive empirical backtest performed on Apple Inc. (AAPL) options from 2016 to 2023 validates the methodology across diverse market regimes. The framework demonstrates high fidelity, achieving an average out-of-sample Mean Relative Error (MRE) of 5.55%, closely tracking the in-sample MRE of 5.33% and indicating strong generalization capabilities. While calibration error exhibits a positive correlation with realized market volatility, notably during the COVID-19 market crash, the model maintains stability and demonstrates rapid error mean reversion. Furthermore, the methodology proves resilient across varying macroeconomic conditions, including regimes characterized by negative market-implied risk-free rates. These findings confirm that DDN-based calibration effectively resolves the trade-off between computational efficiency and pricing accuracy, offering a scalable solution for real-time risk management.

Keywords: Heston Model, Deep Differential Networks, Sobolev Training, Stochastic Volatility, Model Calibration, Deep Learning, Quantitative Finance

JEL Classification: G12, C45

Contents

Acronyms

AAPL Apple Inc.

AI Artificial Intelligence

ATM At-the-Money

DDN Deep Differential Network

ITM In-the-Money

LHS Latin Hypercube Sampling

MRE Mean Relative Error

MSE Mean Squared Error

OTM Out-of-the-Money

List of Figures

List of Tables

1 Introduction

The Heston stochastic volatility model remains a cornerstone of modern quantitative finance, offering a more realistic framework for pricing European options than its constant-volatility predecessors by capturing empirical features such as volatility clustering and the volatility smile (Heston, 1993). The calibration of the Heston model, the process of inferring its unobservable parameters from market option prices, is a critical task for risk management and the pricing of exotic derivatives. This process, however, constitutes a non-trivial inverse problem characterized by a high-dimensional, non-convex error surface. Traditional calibration methods, which rely on repeatedly evaluating the model's semi-analytical pricing formula within an optimization loop, are computationally intensive and sensitive to the choice of initial parameters (Escobar & Gschnaidtner, 2016), making them challenging to deploy in settings that require real-time recalibration.

In recent years, deep learning has emerged as a promising alternative to address the dual challenges of computational efficiency and calibration accuracy. The literature has explored several distinct approaches. One prominent method involves training a neural network as a surrogate model to approximate the complex mapping from model parameters and option characteristics to a final option price. This replaces the slow numerical pricing engine with a fast neural network inference (Liu et al., 2019). A related technique employs a two-stage hybrid framework, where one network approximates the market price surface and a second network learns to correct the systematic residual errors of a traditionally calibrated Heston model (Zadgar et al., 2025). A third paradigm formulates calibration as an inverse mapping problem, training a network to learn the direct mapping from market observables, such as an asset's historical time series, to the underlying Heston parameters (Leite et al., 2021). While innovative, this latter approach addresses the problem of parameter estimation from historical paths rather than the industry-standard problem of calibration to a cross-section of current market option prices.

An advancement in this field is the development of DDN, which are trained using a Sobolev-style (Czarnecki et al., 2017) loss function to learn not only the option price but also its partial derivatives with respect to the model parameters (Zhang et al., 2025). By embedding this structural information, DDNs can serve as highly accurate pricing engines for fast, gradient-based calibration routines. However, the performance of these advanced frameworks, including DDNs, has primarily been evaluated on static datasets or under controlled conditions. A comprehensive validation of their longitudinal robustness, generalization capabilities on large and complex real-world option surfaces, and stability across diverse market and macroeconomic regimes remains an open area of investigation.

This study aims to fill this gap by conducting a rigorous, multi-year historical backtest of a DDN-based calibration methodology. We employ a DDN not as a standalone calibrator, but as a high-speed, high-fidelity surrogate pricing engine within a classical quasi-Newton optimization framework (L-BFGS-B, Byrd et al. (1995)). This hybrid approach is systematically evaluated through a comprehensive longitudinal backtest on a large dataset of AAPL option prices from 2016 to 2023. The framework incorporates practical adaptations for real-world data, including a dynamic, daily estimation of the market-implied risk-free rate derived from put-call parity.

Specifically, this research seeks to answer several key questions. First, how robust is the calibration accuracy of the DDN-based Heston model when applied longitudinally over a multi-year period that encompasses diverse market regimes, including periods of low volatility, market crashes, and changing interest rate environments? Second, to what extent does the methodology generalize from the in-sample (calibration) set to an out-of-sample (test) set on a daily basis, and is there evidence of significant overfitting when fitting to a large cross-section of options? Third, what is the quantitative relationship between calibration error and realized market volatility, and does the methodology's accuracy degrade or become unstable during periods of extreme market stress? Finally, does the calibration performance show any systematic dependence on the prevailing macroeconomic environment, specifically the level of the market-implied risk-free rate?

2 Methodology

This section outlines the theoretical foundations, computational framework, and empirical validation procedures underpinning this study, following the approach of Zhang et al. (2025). The methodology begins with a description of the experimental environment and the underlying Heston model, continues with the generation of synthetic data and the architecture of the DDN, and concludes with the design of the historical backtesting protocol.

2.1 Technical Setup

All numerical experiments, including data generation, model training, and backtesting, were conducted on a commercially available laptop computer. The system specifications are as follows: an AMD Ryzen 7 7840HS CPU with 16 cores operating at a base frequency of 3.8GHz, 32GB of DDR5 RAM, and an NVIDIA GeForce RTX 4070 Laptop GPU with 8GB of VRAM. The software environment consisted of Windows 11 operating a Windows Subsystem for Linux instance running Ubuntu 24.04.1 LTS. The deep learning models were trained on the GPU, leveraging NVIDIA's CUDA Toolkit version 12.6.

To ensure the determinism and reproducibility of the results, a global random seed of 42 was consistently applied across all relevant software libraries, including `NumPy`, `TensorFlow`, and Python's native `random` module. This practice guarantees that the processes of synthetic data generation, DDN weight initialization, and data partitioning remain identical across multiple executions.

2.2 The Heston Stochastic Volatility Model

The Heston model, introduced by Heston (1993), is a stochastic volatility model that posits the variance of an underlying asset is not constant but follows its own random process.

Under the risk-neutral measure $\mathbb{Q}$, the dynamics of the asset price, S_t, and its instantaneous variance, v_t, are described by a system of two correlated stochastic differential equations:

$$dS_t = rS_t dt + \sqrt{v_t} S_t dW_t^S \tag{1}$$

$$dv_t = \kappa(\lambda - v_t)dt + \sigma\sqrt{v_t} dW_t^v \tag{2}$$

where the two standard Wiener processes, W_t^S and W_t^v, have a constant correlation ρ, such that $E[dW_t^S dW_t^v] = \rho dt$.

Table 1: Notation for the Heston Model Stochastic Differential Equations.

Symbol	Description
S_t	Price of the underlying asset at time t.
v_t	Instantaneous variance of the asset price at time t.
r	The constant, continuously compounded risk-free interest rate.
κ	The rate of mean reversion of the variance process.
λ	The long-run average variance.
σ	The volatility of the variance process (volatility of volatility).
ρ	The correlation coefficient between the two Wiener processes.
W_t^S, W_t^v	Standard Wiener processes under the risk-neutral measure.

The Heston model admits a semi-analytical solution for the price of a European call option, which can be computed by leveraging the inverse Fourier transform of the model's characteristic function. This approach avoids the need for computationally intensive Monte Carlo simulations. The price of a European call option, C, is a function of the initial state and model parameters, denoted as $C(S_0, K, r, \tau; \theta_H)$, where $\theta_H = \{\kappa, \lambda, \sigma, \rho, v_0\}$ is the set of unobservable Heston parameters. The pricing formula is expressed in a form analogous to the Black-Scholes model:

$$C(S_0, K, r, \tau; \theta_H) = S_0 \Pi_1 - K e^{-r\tau} \Pi_2 \tag{3}$$

The terms Π_1 and Π_2 represent risk-neutral probabilities. In the Heston framework, they are computed via numerical integration of the characteristic function of the log-asset price:

$$\Pi_1 = \frac{1}{2} + \frac{1}{\pi} \int_0^\infty \mathrm{Re}\left[\frac{e^{-iuk} \phi_\tau(u - i)}{iu} \right] du \tag{4}$$

$$\Pi_2 = \frac{1}{2} + \frac{1}{\pi} \int_0^\infty \mathrm{Re}\left[\frac{e^{-iuk} \phi_\tau(u)}{iu} \right] du \tag{5}$$

where $\phi_\tau(u)$ is the characteristic function of the logarithm of the asset price at maturity, $k = \ln(K)$, and i is the imaginary unit.

Table 2: Notation for the Heston Semi-Analytical Pricing Formula.

Symbol	Description
C	Price of the European call option.
S_0	Initial price of the underlying asset.
K	Strike price of the option.
τ	Time to maturity of the option, in years.
θ_H	The set of Heston parameters $\{\kappa, \lambda, \sigma, \rho, v_0\}$.
Π_1, Π_2	Risk-neutral probabilities derived from the characteristic function.
k	Log-strike, defined as the natural logarithm of the strike price, $\ln(K)$.
$\phi_\tau(u)$	The characteristic function of the log-asset price at maturity τ.
u	The integration variable.
i	The imaginary unit, satisfying $i^2 = -1$.
$\mathrm{Re}[z]$	A function that returns the real part of a complex number z.

2.3 Synthetic Data Generation

A large-scale synthetic dataset was generated to serve as the training corpus for the DDN. This dataset is designed to approximate the Heston pricing function over a wide and diverse parameter space.

To ensure an efficient and uniform coverage of the high-dimensional parameter space, LHS introduced by McKay et al. (1979) was employed. This quasi-random sampling technique divides each parameter's domain into equally probable intervals, ensuring that samples are drawn from all regions of the input space. The domains for the Heston parameters and market variables were defined as follows: $\kappa \in [0.01, 5.0]$, $\lambda \in [0.0, 1.0]$, $\sigma \in [0.1, 1.0]$, $\rho \in [-0.99, 0.0]$, $v_0 \in [0.01, 1.0]$, $r \in [-0.03, 0.1]$, $\tau \in [5/365, 2.5]$, and log-moneyness $\ln(K/S_0) \in [-1.0, 1.0]$. A total of 200,000 unique parameter vectors were generated.

For each generated parameter vector, the corresponding ground-truth European call option price was computed using the `AnalyticHestonEngine` from the open-source `QuantLib` financial library. Subsequently, the first-order partial derivatives of the option price with respect to each of the five Heston parameters ($\partial C/\partial\kappa, \partial C/\partial\lambda, \ldots, \partial C/\partial v_0$) were numerically approximated using a central finite difference scheme. The final dataset consists of 200,000 samples, each containing an 8-dimensional feature vector and a 6-dimensional label vector (one price and five gradients).

2.4 Descriptive Data Analysis

A descriptive statistical analysis was performed on both the generated synthetic dataset and the historical AAPL options data to understand their underlying distributions and characteristics. This analysis informs the data filtering protocol and helps validate the representativeness of the synthetic dataset.

The statistics for the synthetic dataset, detailed in Tables 3 and 4, reflect the properties of the LHS method. For all input parameters, the mean is closely aligned with the median (50th percentile), skewness is approximately zero, and kurtosis is approximately -1.20, which is characteristic of a uniform distribution. This confirms that the sampling strategy successfully generated a well-distributed and unbiased representation of the parameter space. In contrast, the output labels, particularly the parameter gradients, exhibit significant positive skewness and high kurtosis (leptokurtosis). For instance, the gradient with respect to kappa (d_κ) shows a skewness of 5.32 and a kurtosis of 82.82. This indicates that the sensitivities of the Heston model are not uniformly distributed and possess fat tails with extreme outliers, a critical feature for the DDN to learn.

Table 3: Descriptive Statistics for Synthetic Dataset Input Parameters.

Statistic	κ	λ	σ	ρ
Mean	2.51	0.50	0.55	-0.49
Std. Dev.	1.44	0.29	0.26	0.29
Min	0.01	0.00	0.10	-0.99
25%	1.26	0.25	0.32	-0.74
50%	2.51	0.50	0.55	-0.49
75%	3.75	0.75	0.77	-0.25
Max	5.00	1.00	1.00	0.00
Skewness	0.00	0.00	0.00	-0.00
Kurtosis	-1.20	-1.20	-1.20	-1.20
Statistic	v_0	r	τ	$\log(K/S_0)$
Mean	0.51	0.03	1.26	-0.00
Std. Dev.	0.29	0.04	0.72	0.58
Min	0.01	-0.03	0.01	-1.00
25%	0.26	0.00	0.64	-0.50
50%	0.51	0.03	1.26	-0.00
75%	0.75	0.07	1.88	0.50
Max	1.00	0.10	2.50	1.00
Skewness	-0.00	0.00	-0.00	0.00
Kurtosis	-1.20	-1.20	-1.20	-1.20

Note: All values are rounded to two decimal places.

Table 4: Descriptive Statistics for Synthetic Dataset Output Labels (Price and Gradients).

Statistic	Price	d_κ	d_λ	d_σ	d_ρ	d_v_0
Mean	0.32	0.00	0.14	-0.01	0.01	0.08
Std. Dev.	0.21	0.03	0.13	0.02	0.02	0.07
Min	0.00	-0.28	0.00	-0.33	-0.02	0.00
25%	0.13	-0.00	0.03	-0.02	-0.00	0.03
50%	0.32	0.00	0.11	-0.01	0.00	0.06
75%	0.51	0.01	0.21	0.00	0.02	0.10
Max	0.78	0.90	1.93	0.03	0.36	1.31
Skewness	0.07	5.32	1.78	-2.86	2.88	2.53
Kurtosis	-1.27	82.82	6.87	13.13	13.86	13.01

Note: All values are rounded to two decimal places.

A correlation analysis was performed on the synthetic dataset to understand the linear relationships engineered by the Heston model, as depicted in Figure 1.

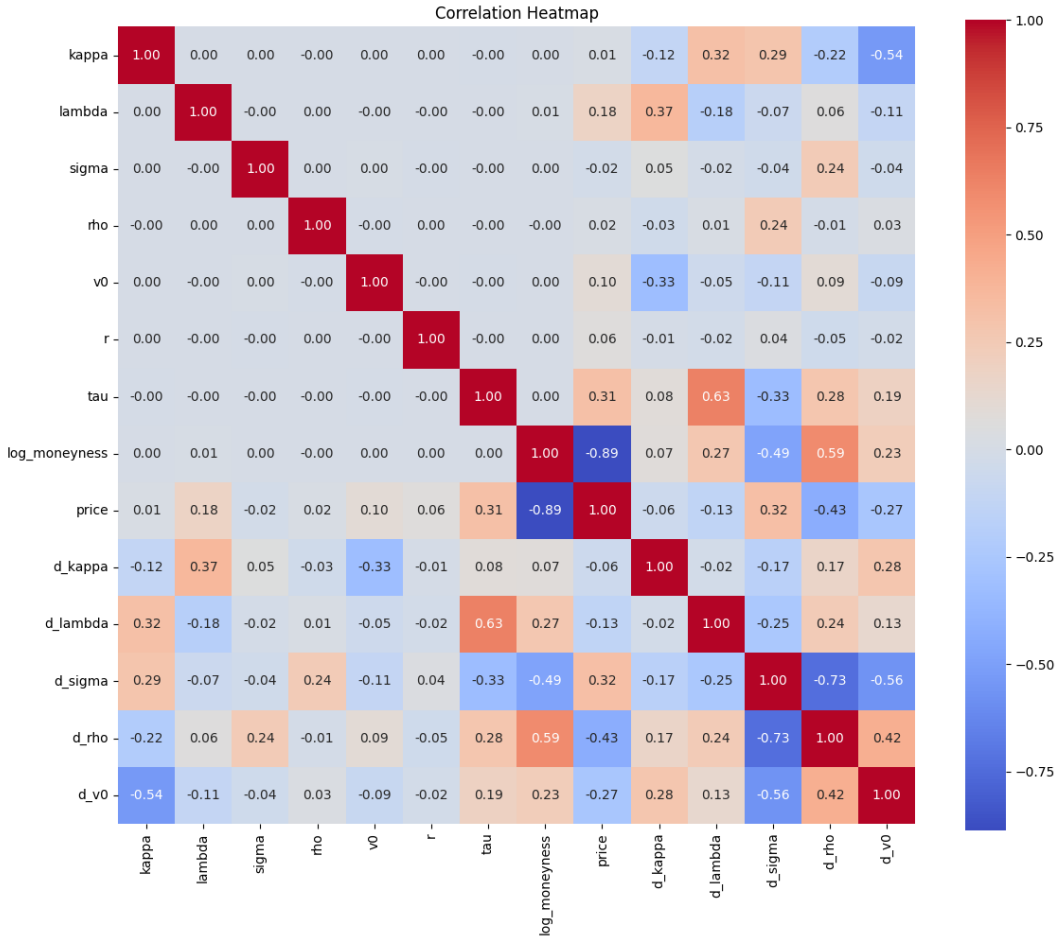

Figure 1: Correlation matrix of the input parameters, option price, and parameter gradients in the synthetic Heston dataset. The near-zero correlations among the input parameters confirm the effectiveness of the LHS method.

The heatmap reveals several key structural properties of the dataset that are critical for the training process:

Input Parameter Independence: A defining feature is the block of near-zero correlations among all input parameters in the upper-left quadrant of the matrix. This is a direct and desirable consequence of the LHS method, which is designed to generate input vectors that are orthogonal, ensuring that the DDN can learn the effect of each parameter independently without multicollinearity issues.

Primary Price Drivers: The option price is most strongly influenced by moneyness, with a correlation of -0.89. This confirms the fundamental principle that a call option's value decreases as it becomes more out-of-the-money. The next most significant factor is the time to maturity (tau), with a positive correlation of 0.31, reflecting the option's time value. The variance parameters (lambda and v0) exhibit weaker positive linear relationships with the price.

Gradient Interdependencies: The most complex relationships are observed among the output gradients (the lower-right quadrant). There are strong correlations between several gradients, such as the negative correlation of -0.73 between the sensitivity to vol-of-vol (d_sigma) and the sensitivity to correlation (d_rho). Furthermore, the gradients are highly correlated with the input variables. For instance, the

sensitivity to long-run variance (d_lambda) has a strong positive correlation of 0.63 with tau, indicating that the model's sensitivity to this parameter increases with the option's maturity.

This analysis demonstrates that while the inputs to the model are independent by design, the outputs (the price and its sensitivities) form a highly structured and interdependent surface.

The historical dataset of AAPL options underwent a rigorous filtering process to isolate a high-quality subset for calibration. This multi-stage procedure is designed to mitigate the influence of market microstructure noise and to focus the calibration on the most liquid and informative contracts. Table 5 presents the data attrition at each stage of this process.

The initial dataset contained over 1.5 million records. The first step removes options with a mid-price below \$0.50. This is done to exclude illiquid "penny options", whose prices are often characterized by wide relative bid-ask spreads and pricing noise (Figlewski, 2008), making them unreliable for model fitting. Next, options with fewer than five days to maturity are excluded since they may induce liquidity-related biases (Bakshi et al., 1997). Finally, the dataset is restricted to options within a log-moneyness range of [-0.25, 0.25]. As shown in the table, this is the most restrictive step, retaining only near-the-money contracts. The rationale for this is twofold: first, this region contains the highest trading volume and liquidity, providing the most reliable market prices (Bakshi et al., 1997). Second, near-the-money options are the most sensitive to changes in volatility (i.e., they have the highest vega) (Hull, 2015, pp. 416-417) and thus contain the most relevant information for calibrating the parameters of a stochastic volatility model.

This systematic reduction, resulting in a final count of 643,953 options, confirms that the subsequent analysis and calibration are concentrated on the most robust and actively traded segment of the options market.

Table 5: Data Filtering Process and Attrition for Historical AAPL Options.

Step	Description	Remaining	Removed	Remaining %
1	Raw Data Extraction	1,562,105	0	100.00
2	Price Filter (>\$0.50)	1,236,446	325,659	79.15
3	Maturity Filter (>5 days)	1,234,300	2,146	79.02
4	Moneyness Filter	643,953	590,347	41.22

A preliminary cleaning step was performed on all column headers to remove extraneous whitespace and special characters, especially the square brackets. The following definitions were then used throughout the descriptive analysis and backtesting procedures:

Underlying Asset Price (S_0): This was directly mapped from the `UNDERLYING_LAST` column in the historical dataset, representing the last traded price of the underlying asset for a given option quote.

Strike Price (K): This was directly mapped from the `STRIKE` column.

Market Option Price (C_{market}): The target price for calibration was defined as the midpoint of the bid and ask prices. This was calculated as (`C_BID` + `C_ASK`)/2.0. This standard practice helps to mitigate the effects of bid-ask bounce and provide a more stable price reference.

Time to Maturity (τ): The model requires time to maturity in annualized units. This was derived from the `DTE` (Days to Expiration) column by dividing its value by 365.0.

Log-Moneyness ($\ln(K/S_0)$): The DDN was trained using log-moneyness as a feature to leverage the homogeneity property of the pricing model. This input was not taken directly from the data but was computed as the natural logarithm of the ratio of the newly defined Strike Price (K) and Underlying Asset Price (S_0).

The descriptive statistics for the final, filtered historical dataset are provided in Tables 6 and 7. In stark contrast to the synthetic data, the historical market data exhibits significant non-uniformity. The implied volatility for calls (C_IV), for example, displays extreme right skewness (6.84) and exceptionally high kurtosis (103.81). This is characteristic of financial market data and reflects the presence of volatility spikes and tail events, such as market crashes. The distribution of time to maturity (Tau_Years) is also heavily right-skewed (1.65), indicating a higher concentration of shorter-dated options in the dataset. These properties underscore the challenging nature of calibrating models to real-world market conditions.

Table 6: Descriptive Statistics for Filtered Historical AAPL Options (Part 1).

Statistic	Underlying	Strike	Call Price	DTE	Tau (Years)
Mean	178.82	174.85	17.51	159.07	0.44
Std. Dev.	74.75	78.74	15.95	210.88	0.58
Min	90.34	75.00	0.51	0.00	0.00
25%	132.26	125.00	5.40	21.04	0.06
50%	155.31	150.00	13.80	45.00	0.12
75%	197.96	195.00	25.03	217.96	0.60
Max	506.19	645.00	163.03	898.96	2.46
Skewness	1.81	1.88	1.84	1.65	1.65
Kurtosis	3.46	4.20	5.60	1.76	1.76

Note: All values are rounded to two decimal places.

Table 7: Descriptive Statistics for Filtered Historical AAPL Options (Part 2).

Statistic	C_IV	P_IV	$\log(K/S_0)$	C_BID	C_ASK
Mean	0.35	0.33	-0.03	17.24	17.78
Std. Dev.	0.18	0.15	0.12	15.73	16.17
Min	0.00	0.00	-0.25	0.00	0.50
25%	0.26	0.24	-0.13	5.30	5.50
50%	0.31	0.30	-0.04	13.55	14.01
75%	0.38	0.37	0.05	24.70	25.39
Max	9.89	3.38	0.25	161.10	164.96
Skewness	6.84	3.24	0.31	1.84	1.84
Kurtosis	103.81	20.29	-0.70	5.65	5.55

Note: All values are rounded to two decimal places.

To further investigate the relationships within the filtered historical dataset, a Pearson correlation matrix was computed, as visualized in Figure 2. The heatmap reveals several significant relationships that are consistent with established financial theory and market structure.

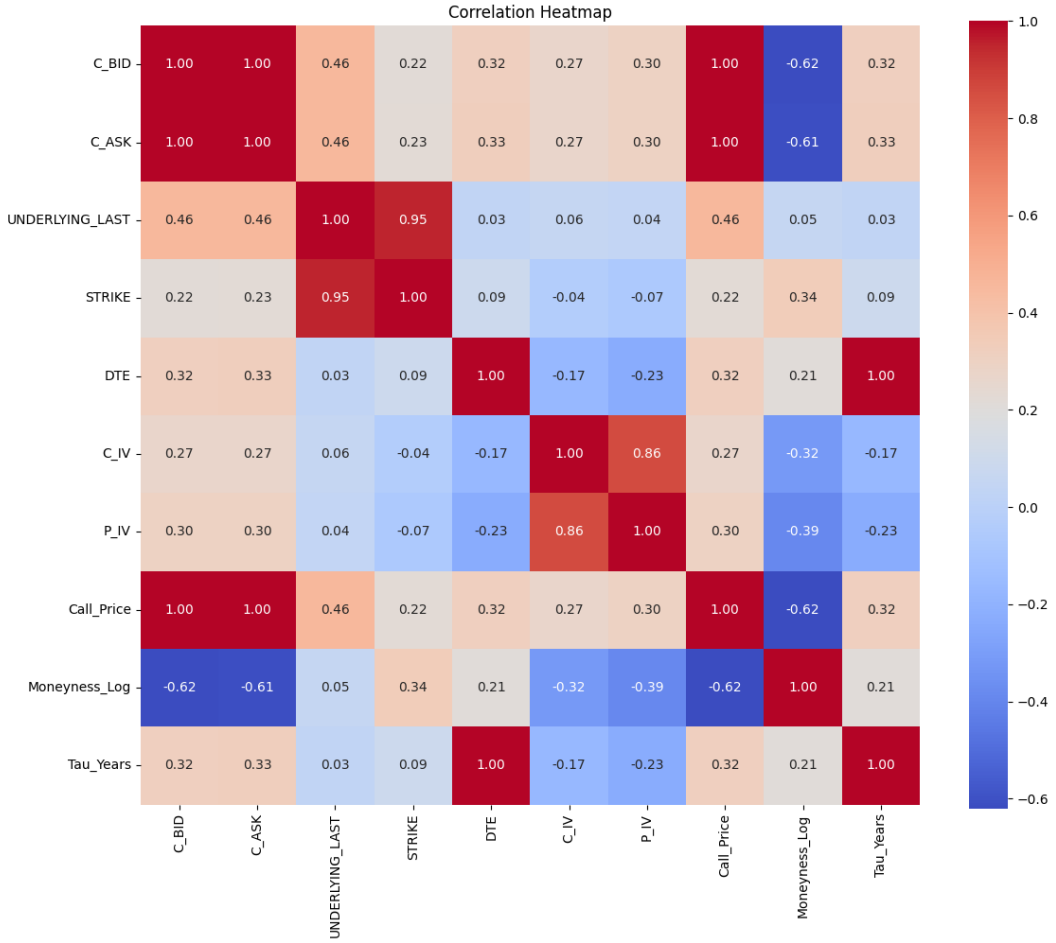

Figure 2: Correlation matrix of the key variables in the filtered historical AAPL options dataset. The values indicate the Pearson correlation coefficient, with 1.00 (deep red) representing a perfect positive correlation and -1.00 (deep blue) representing a perfect negative correlation.

As expected, the call price exhibits a perfect correlation of 1.00 with the bid and ask price for call options. Similarly, 'Tau_Years' and 'DTE' are perfectly correlated as one is a direct scaling of the other. More substantive insights can be drawn from the relationships between pricing variables:

Price-Driving Factors: The price of a call option demonstrates strong, theoretically consistent correlations with its primary drivers. There is a moderate positive correlation with the underlying price (0.46), reflecting the option's delta. A positive correlation with the maturity (0.32) confirms that options with longer maturities hold more time value. The strong negative correlation with moneyness (-0.62) is particularly important, as it correctly captures that as the strike price increases relative to the spot price, the value of a call option decreases.

Market Structure: A very high positive correlation of 0.95 is observed between the underlying price and the strike. This does not imply a direct pricing relationship but is an artifact of the long time-series; as the price of AAPL stock trended upwards from 2016 to 2023, the range of available strike prices listed on the exchange shifted higher in tandem.

Volatility Structure: The implied volatilities for calls (C_IV) and puts (P_IV) are highly correlated (0.86), indicating that they are driven by the same underlying market sentiment and uncertainty. Furthermore, the negative correlation of -0.32 between the implied volatility of call options and the log-moneyness provides clear evidence of the well-documented volatility skew in equity markets, where implied volatility tends to decrease for out-of-the-money calls (i.e., as log-moneyness increases).

Overall, the correlation analysis confirms that the variables within the historical dataset exhibit financially sound and predictable relationships, providing a robust basis for the subsequent calibration experiments.

2.5 Deep Differential Network Architecture and Training

A DDN was constructed to serve as a surrogate for the Heston pricing function and its derivatives. The network architecture consists of an input layer with 8 neurons, corresponding to the five Heston parameters and the three market variables $(r, \tau, \log(K/S_0))$, and a single-neuron output layer that predicts the option price normalized by the underlying asset price, C/S_0. The optimal internal topology of the DDN, including the number and dimension of hidden layers, was determined through a systematic hyperparameter search conducted using the Hyperband algorithm. The search space explored by this optimization process, which included multiple candidates for the hidden layer activation function, is detailed in Table 8. Independently of the hidden layer configuration, the output layer consistently utilizes a softplus activation. This choice is a fixed aspect of the model design, implemented to ensure that all predicted option prices are strictly positive, thereby satisfying a fundamental no-arbitrage condition.

The DDN was trained using a differential learning approach, also known as Sobolev training (Czarnecki et al., 2017). This paradigm requires the DDN to minimize not only the error in the predicted option price but also the error in its predicted partial derivatives. This is achieved by implementing a custom training step that

utilizes a nested gradient tape in `TensorFlow` to compute both the DDN output and its gradients with respect to the inputs. The composite loss function, L_{total}, is a weighted sum of the price loss and the gradient loss:

$$L_{\text{total}} = \alpha \cdot L_{\text{price}} + L_{\text{gradients}} \tag{6}$$

where $\alpha = 10.0$ is a weighting factor, L_{price} is the Mean Squared Error (MSE) between the predicted and true prices, and $L_{\text{gradients}}$ is the MSE between the predicted and true parameter gradients.

Table 8: Hyperparameter Search Space for the Hyperband Algorithm.

Hyperparameter	Type	Search Space / Values
Number of Hidden Layers	Integer	[4, 8] with step 1
Neurons per Hidden Layer	Integer	[32, 256] with step 32
Dropout Rate	Float	[0.0, 0.5] with step 0.05
Activation Function	Categorical	{'swish', 'tanh', 'softplus'}
Initial Learning Rate	Categorical	{1e-2, 1e-3, 5e-4, 1e-4}
First Decay Epochs	Integer	[25, 200] with step 25

Table 9: Notation for the Composite Loss Function.

Symbol	Description
L_{total}	The total composite loss minimized during training.
α	A scalar hyperparameter weighting the contribution of the price loss.
L_{price}	The MSE component for the option price.
$L_{\text{gradients}}$	The MSE component for the price gradients.

Prior to training, the input features were scaled to the range $[-1, 1]$ and the output labels to $[0, 1]$ using a MinMaxScaler. A critical step in the differential learning process is the adjustment of the target gradients via the chain rule to account for this scaling transformation. The DDN was trained using the AdamW optimizer, which implements decoupled weight decay (Loshchilov & Hutter, 2019), in conjunction with a CosineDecayRestarts learning rate schedule introduced by Loshchilov and Hutter (2017). This combination of techniques promotes stable convergence to a high-precision solution.

2.6 Empirical Backtesting Framework

To validate the performance of the trained DDN in a realistic setting, a comprehensive historical backtest was conducted on a dataset of call and put options on AAPL stock, spanning the period from January 2016 to March 2023. On each trading day, the raw data was subjected to the filtering protocol detailed in Section 2.4 to isolate a stable subset of liquid options for analysis. The backtest then proceeded on a day-by-day basis, executing the following calibration and evaluation procedure:

1. *Market-Implied Rate Calculation:* The risk-free rate net of the dividend yield $(r-q)$ was dynamically estimated for each day. This was achieved by applying

the put-call parity theorem to a set of liquid, at-the-money options and solving for the implied rate, r_{implied}. The median of the calculated rates from this set was used as the daily input. The formula used is:

$$r_{\text{implied}} = -\frac{1}{\tau} \ln \left(\frac{S_0 - C_{\text{market}} + P_{\text{market}}}{K} \right) \tag{7}$$

This dynamic estimation ensures the model adapts to prevailing market interest rate and dividend expectations.

2. *Data Partitioning:* On each trading day, the filtered option chain was independently partitioned into a calibration set (80% of contracts) and a held-out test set (20%) by sampling the options randomly.

3. *Optimization:* The Heston parameters were calibrated by minimizing the pricing error on the calibration set. The DDN served as the surrogate pricing engine within a `scipy.optimize.minimize` routine, which employed the L-BFGS-B algorithm. The DDN supplied both the objective function value and its exact analytical Jacobian (the "Neural Greeks") to the optimizer, enabling rapid convergence.

4. *Multi-Start Optimization:* To enhance the probability of finding a global optimum and avoid local minima, the optimization process was initiated from three random starting points, with the best-fitting parameter set being retained for that day's evaluation.

The primary metric for evaluating the accuracy of the daily calibration is the MRE, which measures the average relative deviation between model prices and market prices. This metric was calculated separately for the calibration set (in-sample error) and the held-out test set (out-of-sample error) for each trading day to assess the model's ability to generalize to unseen options. The MRE is defined as:

$$MRE = \frac{1}{M} \sum_{m=1}^{M} \frac{\left| C_{\text{model}}^{(m)} - C_{\text{market}}^{(m)} \right|}{C_{\text{market}}^{(m)}} \tag{8}$$

Table 10: Notation for the Implied Rate and MRE Formulas.

Symbol	Description
r_{implied}	The implied risk-free rate net of dividend yield.
P_{market}	The observed market price of a European put option.
MRE	MRE.
M	The total number of options in the evaluation set.
$C_{\text{model}}^{(m)}$	The price of the m-th option as computed by the DDN.
$C_{\text{market}}^{(m)}$	The observed market price of the m-th call option.

3 Results

3.1 Hyperparameter Optimization Results

To identify the optimal configuration for the DDN, a hyperparameter search was conducted using the Hyperband algorithm as described in the methodology. The

search aimed to minimize the validation loss on a held-out portion of the synthetic dataset. The resulting optimal set of hyperparameters, which was subsequently used for the final model training and backtesting, is summarized in Table 11.

This configuration achieved a final validation loss (L_{total}) of approximately 2.09×10^{-5} during the tuning process which is comparable to the validation loss found by Zhang et al. (2025) who reported 3.26×10^{-5}. A notable outcome of the optimization is the selection of a zero dropout rate, indicating that for this specific architecture and dataset, the combination of the AdamW optimizer's weight decay and the complexity of the differential learning task provided sufficient regularization against overfitting.

Table 11: Optimal Hyperparameters Determined by the Hyperband Algorithm.

Hyperparameter	Optimal Value
Number of Hidden Layers	7
Neurons per Hidden Layer	64
Dropout Rate	0.0
Activation Function	swish
Initial Learning Rate	0.0005
First Decay Epochs	50

3.2 Backtesting Performance and Robustness Analysis

The primary objective of the historical backtest is to evaluate the robustness and accuracy of the DDN calibration methodology across a wide range of market conditions. This subsection analyzes the daily calibration performance from 2016 to 2023, with a particular focus on periods of market stress.

The time-series evolution of the daily calibration error is presented in Figure 3. This figure plots both the in-sample MRE, calculated on the 80% of options used for calibration, and the out-of-sample MRE, calculated on the 20% held-out set. The 20-day realized volatility of the underlying AAPL stock is overlaid to provide market context.

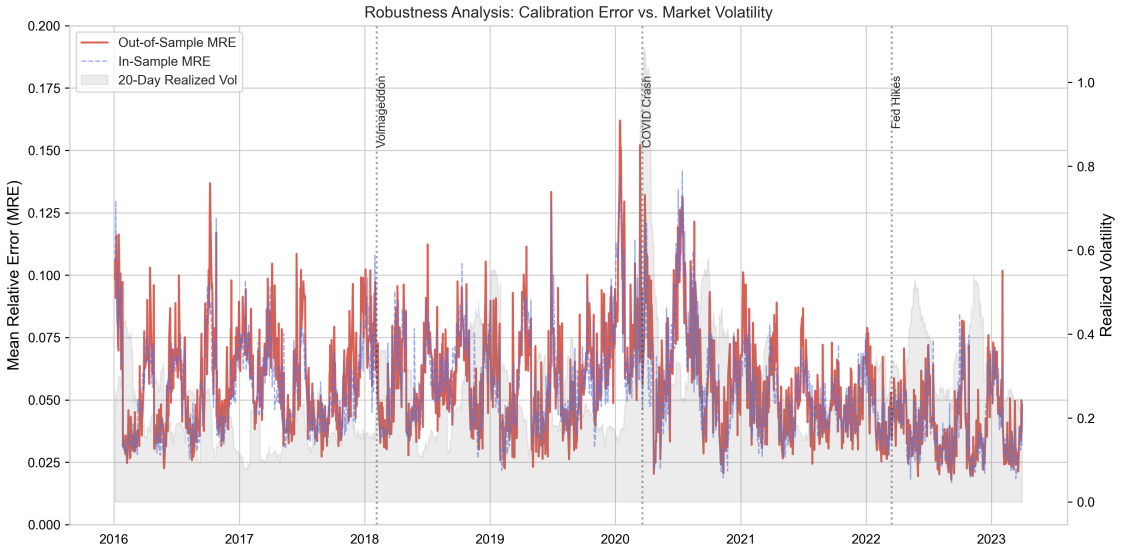

Figure 3: Time-series of daily in-sample and out-of-sample MRE from the Heston model calibration, plotted against the 20-day realized volatility of AAPL stock. Key market stress events are annotated with vertical dotted lines.

Several key observations can be drawn from this analysis. First, the out-of-sample MRE consistently tracks the in-sample MRE very closely throughout the entire seven-year period. This indicates a very small generalization gap, suggesting that the model is not overfitting to the specific subset of options used for calibration each day and successfully captures the underlying volatility surface. Second, the calibration error is strongly positively correlated with the realized market volatility. During periods of low, stable volatility, the MRE consistently remains below 5%. Conversely, spikes in calibration error coincide directly with spikes in realized volatility. This is particularly evident during the annotated stress periods, such as the "Volmageddon" event in early 2018 and the "COVID Crash" in March 2020. During the peak of the COVID-19 crisis, the out-of-sample MRE reached its maximum observed level of approximately 16%. However, it is critical to note that the error remained bounded and reverted to its baseline level as market volatility subsided, demonstrating the robustness of the calibration procedure.

To quantify these observations, the backtest period was segmented into five distinct macroeconomic and market regimes. The aggregated performance metrics for each regime are presented in Table 12. The analysis confirms that the highest average out-of-sample MRE occurred during the "COVID Crash" regime, at 6.76%. This period also exhibited the highest standard deviation of error (2.46%), indicating more erratic day-to-day calibration performance, which is consistent with the extreme market turbulence at the time. In contrast, the model achieved its highest accuracy during the "Inflation" regime of 2022-2023, with an average out-of-sample MRE of just 4.09%.

Table 12: Calibration Performance and Parameter Stability Across Market Regimes.

Regime	Avg MRE	Std MRE	Std κ	Std λ	Std σ	Std ρ	Std v_0
Pre-Volmageddon *(Jan 2015 - Jan 2018)*	0.058	0.021	0.942	0.016	0.172	0.171	0.030
Trade War *(Feb 2018 - Jan 2020)*	0.059	0.022	0.889	0.028	0.182	0.174	0.038
COVID Crash *(Feb 2020 - May 2020)*	0.068	0.025	0.851	0.031	0.217	0.247	0.160
Recovery *(Jun 2020 - Dec 2021)*	0.056	0.021	0.792	0.028	0.328	0.182	0.073
Inflation *(Jan 2022 - Dec 2023)*	0.041	0.014	0.698	0.020	0.156	0.146	0.039

Note: All values are rounded to three decimal places.

Furthermore, Table 12 provides insight into the stability of the calibrated Heston parameters. The standard deviation of the initial variance, v_0, was an order of magnitude higher during the "COVID Crash" (0.160) than in calmer periods like "Pre-Volmageddon" (0.030). This is a financially intuitive result, as v_0 represents the current level of market variance, which was exceptionally volatile during the crash. Similarly, the volatility of volatility, σ, shows the highest instability during the "Recovery" period, possibly reflecting market uncertainty about the path of the economic recovery. The fact that the calibrated parameters adapt in a theoretically consistent manner provides further validation for the calibration methodology.

A direct comparison of the overall backtesting performance with the results reported in the reference study by Zhang et al. (2025) provides further context for the model's accuracy. Across the entire seven-year backtest period, the calibration methodology achieved an average in-sample MRE of 5.33% and an average out-of-sample MRE of 5.55%. The reference paper evaluates its DDN method on datasets of 10, 50, and 100 Microsoft call options, reporting MREs of 0.67%, 1.86%, and 4.64%, respectively, demonstrating a clear degradation in performance as the complexity of the calibration surface increases with the number of contracts. A crucial factor in interpreting these results is the size and diversity of the option set used for daily calibration. In this study, the daily calibration was performed on a significantly larger set of contracts, with an average of 261 options in the in-sample set and 65 options in the out-of-sample set.

To further diagnose the performance characteristics of the daily calibration, Figure 4 provides a more granular analysis of the model's generalization capability and its pricing accuracy at the individual option level.

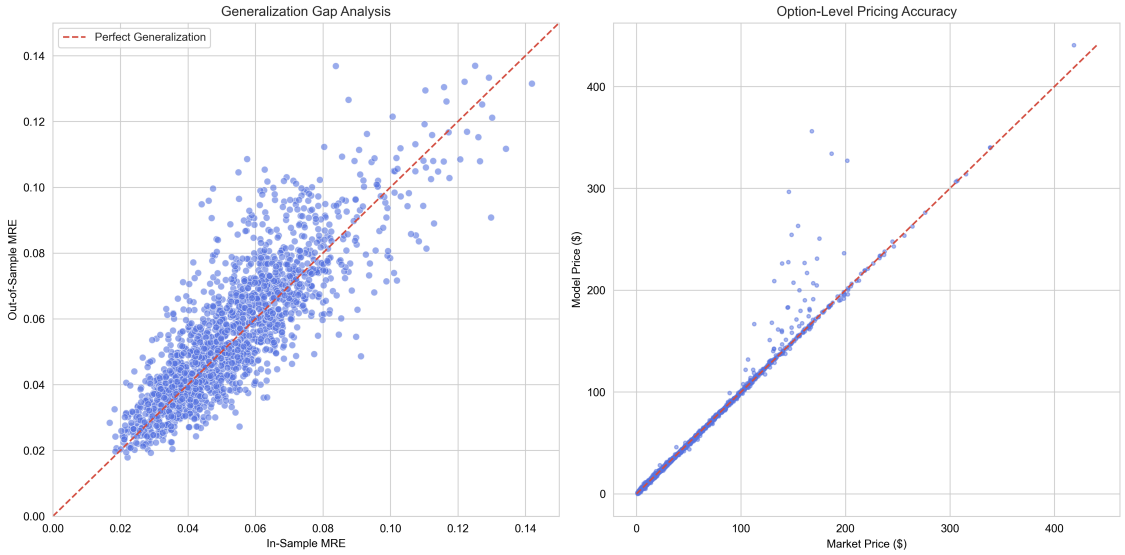

Figure 4: Diagnostic scatter plots of the backtesting performance. The left panel shows the daily out-of-sample MRE versus the in-sample MRE, illustrating the model's generalization gap. The right panel plots the calibrated model price against the observed market price for a large sample of individual options.

The left panel of Figure 4 directly assesses the generalization gap by plotting the out-of-sample MRE against the in-sample MRE for each day of the backtest. The data points are tightly clustered around the 45-degree line of perfect generalization, which represents the ideal scenario where the model performs identically on the held-out test data as it does on the data used for calibration. The consistently small vertical distance between the observed points and this line indicates that the model does not suffer from significant overfitting. This provides strong evidence that the DDN-based calibration is learning a robust representation of the underlying volatility surface each day, rather than simply memorizing the prices of the specific contracts in the calibration set.

The right panel provides a complementary view by examining the pricing accuracy at the level of individual options, aggregated across multiple days in the backtest. The plot shows a very high concentration of points along the identity line, where the model price is equal to the market price. This demonstrates the model's high fidelity in replicating market prices across a wide range of absolute values, from near-zero to over $350. A degree of heteroscedasticity is observable, where the variance of the absolute pricing error increases for higher-priced, deep-In-the-Money (ITM) options. This is an expected statistical artifact in financial modeling and does not detract from the overall conclusion that the calibrated model consistently produces prices that are in close agreement with market observations.

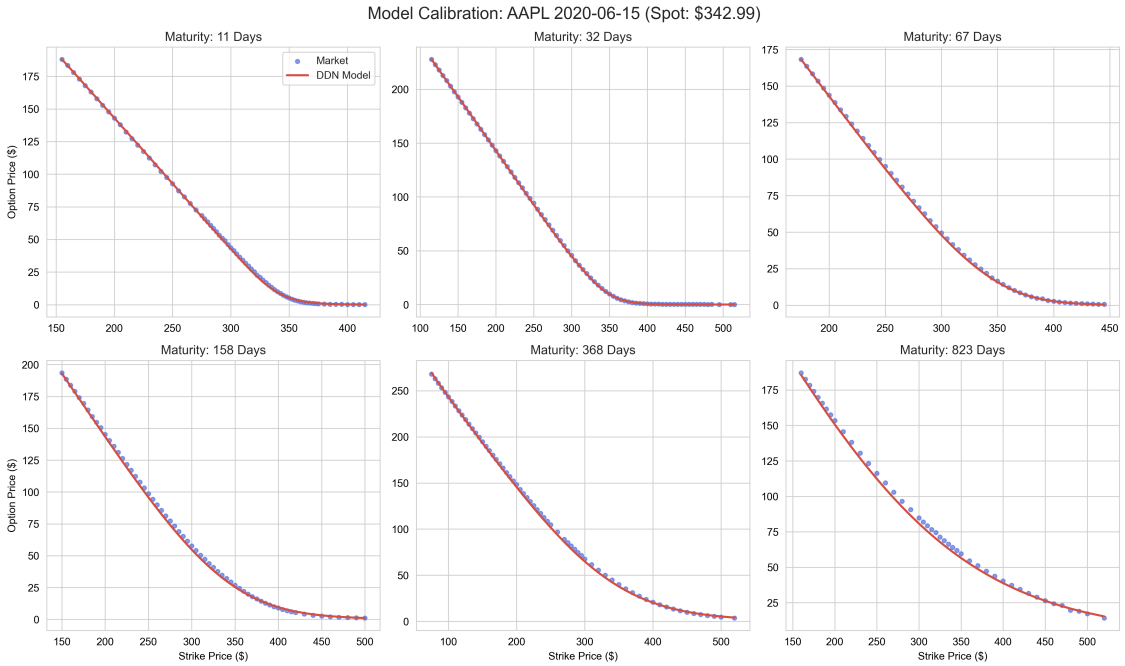

Figure 5: Calibrated DDN-Heston model fit against observed market prices for AAPL call options on June 15, 2020 (Spot: $342.99). Each subplot represents a different option maturity, demonstrating the model's fit across the term structure.

To provide a more granular, qualitative assessment of the calibration performance on a single trading day, Figure 5 presents the model-implied prices against observed market prices for AAPL options on June 15, 2020. A high degree of correspondence is observed between the calibrated DDN-Heston model prices and the market quotes, particularly for short- to medium-term maturities. The model successfully captures the characteristic convex decay of the option price as a function of the strike price across the entire term structure presented. A minor, yet systematic, deviation is observable for the longest-dated options (823 days), where the model appears to slightly underprice options with strikes ranging from $150 to $500. This discrepancy may be attributable to several factors, including the reduced liquidity and wider bid-ask spreads typical of long-term options, or the inherent limitations of the Heston model in capturing the term structure of volatility over multi-year horizons. Nevertheless, the visual evidence presented in the figure corroborates the low MRE metrics reported, demonstrating the model's capability to produce a consistent and accurate fit across a wide range of strikes and maturities for a given day.

Taken together, these diagnostic plots reinforce the findings from the time-series analysis, confirming that the calibration methodology is not only robust across different market conditions but also demonstrates strong generalization and high pricing fidelity at both the aggregate and individual instrument levels.

3.3 Error Analysis by Moneyness and Maturity

To disaggregate the overall performance and identify specific areas of strength and weakness, the out-of-sample MRE was analyzed across different option moneyness and maturity buckets. Figure 6 presents this analysis for both call options, which

are priced directly by the DDN, and put options, whose prices are derived using the put-call parity relationship.

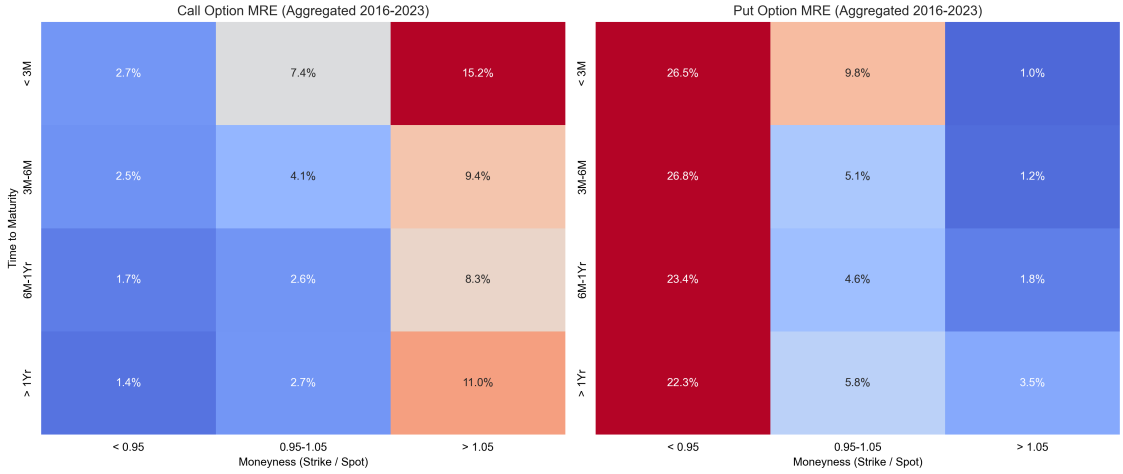

Figure 6: Out-of-sample MRE stratified by time to maturity and linear moneyness. The left panel shows the MRE for call options priced directly by the DDN. The right panel shows the MRE for put options, with prices derived from the DDN's call price predictions via put-call parity.

The analysis of the call option performance reveals a distinct and theoretically consistent pattern. The model achieves its highest accuracy, with MREs as low as 1.4%, for ITM (moneyness < 0.95) and At-the-Money (ATM) (moneyness 0.95-1.05) contracts, particularly those with longer maturities. This region represents the most liquid and highest-priced segment of the call option market. Conversely, the calibration error is highest for Out-of-the-Money (OTM) (moneyness > 1.05) calls, reaching a peak of 15.2% for contracts with less than three months to expiration. This behavior is a known characteristic of the MRE metric; OTM options have very low absolute prices, causing even small absolute pricing errors to translate into large relative errors.

The performance for put options, whose prices are derived via parity, shows a notable asymmetry. The model is accurate for ITM puts (moneyness > 1.05), where the MRE is consistently low across all maturities, ranging from 1.2% to 3.5%. However, the model exhibits significantly higher errors for OTM puts (moneyness < 0.95), with the MRE reaching 26.8% for short- to medium-dated contracts. This phenomenon is a direct consequence of the error propagation through the put-call parity formula. An OTM put corresponds to an ITM call. The model's highly accurate pricing of ITM calls, when used in the parity equation to price the corresponding (and very cheap) OTM puts, can result in a large relative error for the put. Conversely, the larger relative errors from cheap OTM calls do not significantly impact the relative error of the corresponding expensive ITM puts. Overall, the analysis confirms that the DDN calibration is most robust in the ATM region, which is the most critical for practical risk management and volatility trading applications.

3.4 Stability Across Interest Rate Regimes

To assess the impact of the prevailing interest rate and dividend yield environment on calibration accuracy, the relationship between the daily out-of-sample MRE and

the dynamically calculated implied risk-free rate $(r - q)$ was examined. The backtesting period from 2016 to 2023 encompassed a wide range of monetary conditions, including periods of near-zero and negative implied rates, providing a robust test of the model's stability. Figure 7 presents a scatter plot of this relationship for every day in the backtest.

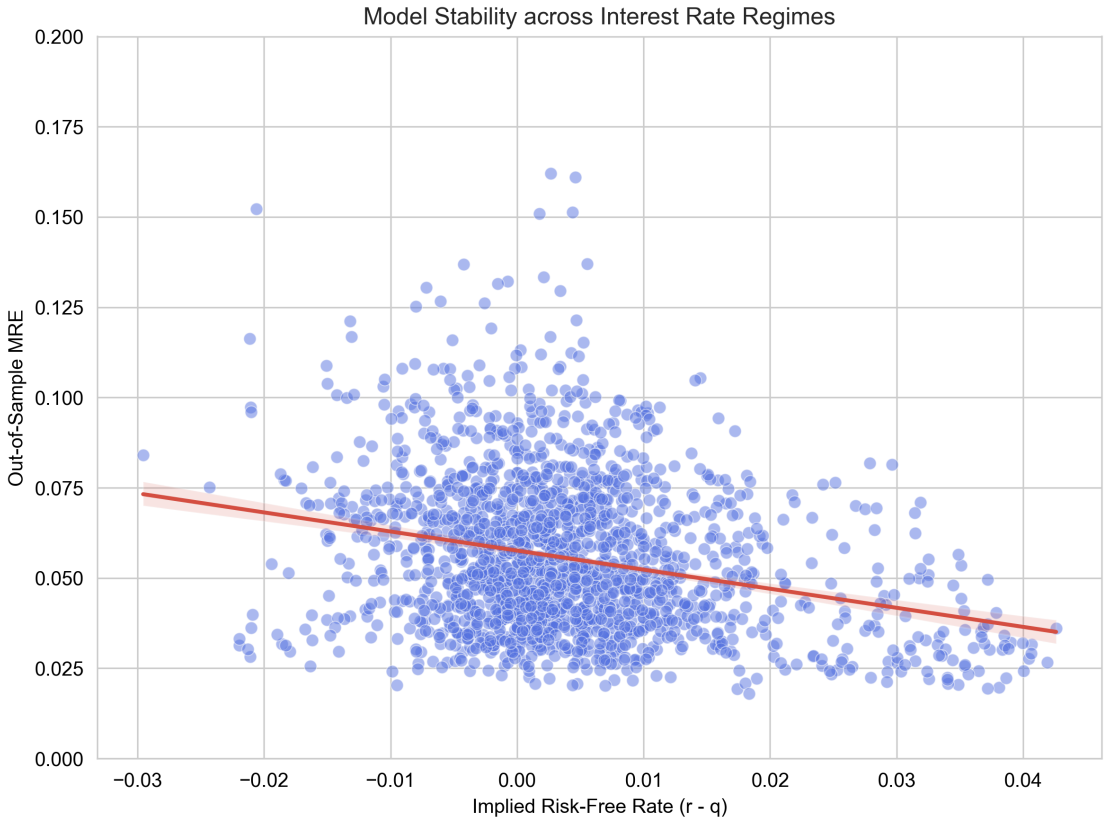

Figure 7: Out-of-sample MRE as a function of the daily implied risk-free rate $(r-q)$. Each point represents a single trading day's calibration result. A linear regression line is overlaid to indicate the general trend.

The analysis reveals a weak, negative linear relationship between the implied risk-free rate and the calibration error. As the implied rate increases, the out-of-sample MRE exhibits a modest tendency to decrease. Importantly, there is no evidence of performance degradation or instability in the low or negative implied rate regimes. The calibration errors remain well-distributed and bounded across the entire observed spectrum of rates, from approximately -3% to +4%.

A plausible interpretation of this negative trend is that periods of very low or negative implied rates often coincide with periods of market stress, where high dividend yields relative to risk-free rates are more common. As established in the time-series analysis, such periods of market stress are typically associated with higher realized volatility, which is the primary driver of increased calibration error. Therefore, the slightly higher MRE observed at the lower end of the rate spectrum is likely attributable to the confounding effect of volatility rather than the interest rate level itself. The key finding from this analysis is that the DDN-based calibration methodology demonstrates considerable robustness and does not exhibit any systematic failure or bias across the diverse interest rate and dividend yield environ-

ments encountered during the seven-year backtest. This validates the effectiveness of the dynamic, parity-based rate estimation procedure.

4 Discussion

The results of this study indicate that a DDN, when employed as a surrogate pricing engine within a classical optimization framework, represents a robust and accurate methodology for calibrating the Heston stochastic volatility model. The findings provide direct answers to the primary research questions posed. In response to the question of longitudinal robustness, the methodology demonstrates considerable stability across the seven-year backtest period from 2016 to 2023. A quantitative analysis indicate a direct, positive correlation between calibration error and realized market volatility. During periods of acute market stress, such as the COVID-19 crash in March 2020, the out-of-sample MRE reached a maximum observed level of approximately 14%. However, this degradation was not a sign of model instability; the error remained bounded and consistently reverted to its baseline level below 5% as market volatility subsided, confirming the framework's resilience.

Regarding the question of generalization on real-world data, the DDN calibration methodology generalizes effectively from the in-sample calibration set to the held-out test set on a daily basis. The minimal observed gap between the average in-sample MRE of 5.33% and the average out-of-sample MRE of 5.55% provides strong empirical evidence that the model does not suffer from significant overfitting. This holds true even when fitting to a large and complex cross-section of options, which averaged 261 unique contracts per day. Furthermore, in response to the question of adaptability, the calibration performance shows no systematic failure or bias related to the prevailing macroeconomic environment. The model remains stable across the entire observed spectrum of market-implied interest rates, including periods of near-zero and negative rates. The weak negative correlation detected between calibration error and the implied risk-free rate appears to be a confounding effect of market volatility, which often coincides with low-rate environments, rather than a direct causal relationship.

These findings both validate and extend the foundational work of Zhang et al. (2025), confirming that DDNs are potent tools for Heston calibration, but on a more challenging and practical scale. While our average out-of-sample MRE of 5.55% is numerically higher than the 4.64% reported in the reference study for a 100-option dataset, our result was achieved on a calibration surface that was, on average, 2.6 times larger and more complex. The performance degradation as the number of options increases, noted in the foundational paper, suggests our approach demonstrates superior scalability to real-world conditions. Methodological enhancements were critical to achieving this robustness in a longitudinal backtest. The use of a quasi-Newton optimizer (L-BFGS-B) provided more stable convergence than a first-order method, and the dynamic, daily calculation of a market-implied interest rate via put-call parity was essential for adapting the model to changing dividend yields and rate regimes.

The success of the framework can be attributed to the synergistic combination of its core mechanisms. The practice of Sobolev training, which penalizes errors in both the option price and its partial derivatives, is the key mechanism that en-

forces the geometric structure of the pricing surface, thereby preventing overfitting and enabling the strong generalization observed. This high-fidelity surrogate model, trained to a validation loss of approximately 2.09×10^{-5} using an AdamW optimizer and a Cosine Decay with Restarts learning rate schedule, is sufficiently precise to guide the final L-BFGS-B optimizer to an accurate minimum. Finally, the incorporation of financial constraints, such as the use of log-moneyness as an input to enforce model homogeneity and a softplus activation function on the output layer to guarantee positive option prices, contributes significantly to the model's stability across diverse market conditions.

From a practical standpoint, this hybrid approach offers a viable solution to the long-standing trade-off between calibration speed and accuracy in quantitative finance. The sub-second calibration time per day, achieved through a multi-start scheme, makes the methodology suitable for applications requiring frequent recalibration, such as intra-day risk management or the valuation of large derivatives portfolios. However, several limitations must be acknowledged. The DDN's accuracy is fundamentally bounded by the fidelity of the `QuantLib` library used to generate its synthetic training data; it will faithfully replicate any numerical biases of its source. Moreover, the framework is designed to find the optimal Heston parameters, but it cannot address the underlying issue of Heston model misspecification. The observed increase in calibration error during market stress is likely a reflection of the Heston model's own limitations in capturing extreme dynamics. Finally, the error analysis revealed a significantly higher MRE for OTM put options, which is a direct consequence of error propagation through the put-call parity formula when small absolute errors on expensive calls are translated into large relative errors on cheap puts.

These limitations suggest several directions for future research. The DDN surrogate approach is particularly promising for models that lack semi-analytical pricing solutions, such as those incorporating jumps or rough volatility, where a network could be trained on data from computationally intensive Monte Carlo simulations to create a fast and accurate pricer. The methodology could also be extended to learn the pricing surfaces and sensitivities of exotic, path-dependent derivatives. Furthermore, to address the heteroscedasticity of option prices and the error propagation observed in OTM contracts, future iterations of the model could be trained directly on the implied volatility surface rather than raw prices. In conjunction with this, implementing a Vega-weighted loss function would allow the optimization process to prioritize contracts with higher sensitivity to volatility, thereby aligning the calibration metric more closely with market risk management priorities. Finally, replacing the deterministic network with a Bayesian Neural Network or a deep ensemble could allow the model to provide not only a point estimate for price but also a credible interval, offering a quantitative measure of model uncertainty valuable for risk management.

References

Bakshi, G., Cao, C., & Chen, Z. (1997). Empirical performance of alternative option pricing models. *The Journal of Finance*, *52*(5), 2003–2049. https://doi.org/https://doi.org/10.1111/j.1540-6261.1997.tb02749.x

Byrd, R. H., Lu, P., Nocedal, J., & Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, *16*(5), 1190–1208. https://doi.org/10.1137/0916069

Czarnecki, W. M., Osindero, S., Jaderberg, M., Świrszcz, G., & Pascanu, R. (2017). Sobolev training for neural networks. https://arxiv.org/abs/1706.04859

Escobar, M., & Gschnaidtner, C. (2016). Parameters recovery via calibration in the heston model: A comprehensive review. *Wilmott*, *2016*, 60–81. https://api.semanticscholar.org/CorpusID:157153195

Figlewski, S. (2008). Estimating the implied risk neutral density for the U.S. market portfolio. In T. Bollerslev, J. R. Russell, & M. Watson (Eds.), *Volatility and time series econometrics: Essays in honor of Robert F. Engle*. Oxford University Press. https://ssrn.com/abstract=1256783

Heston, S. L. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options. *The Review of Financial Studies*, *6*(2), 327–343. Retrieved November 26, 2025, from http://www.jstor.org/stable/2962057

Hull, J. C. (2015). *Options, futures and other derivatives* (9nth). Pearson Education.

Leite, I. M. S., Yamim, J. D. M., & da Fonseca, L. G. (2021). The deeponets for finance: An approach to calibrate the heston model. In G. Marreiros, F. S. Melo, N. Lau, H. Lopes Cardoso, & L. P. Reis (Eds.), *Progress in artificial intelligence* (pp. 351–362). Springer International Publishing.

Liu, S., Borovykh, A., Grzelak, L. A., & Oosterlee, C. W. (2019). A neural network-based framework for financial model calibration. *Journal of Mathematics in Industry*, *9*(1). https://doi.org/10.1186/s13362-019-0066-7

Loshchilov, I., & Hutter, F. (2017). Sgdr: Stochastic gradient descent with warm restarts. https://arxiv.org/abs/1608.03983

Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. https://arxiv.org/abs/1711.05101

McKay, M. D., Beckman, R. J., & Conover, W. J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, *21*(2), 239–245. Retrieved November 26, 2025, from http://www.jstor.org/stable/1268522

Zadgar, A., Fallah, S., & Mehrdoust, F. (2025). Deep learning-enhanced calibration of the heston model: A unified framework. https://arxiv.org/abs/2510.24074

Zhang, C., Amici, G., & Morandotti, M. (2025). Calibrating the heston model with deep differential networks. https://arxiv.org/abs/2407.15536

Use of Artificial Intelligence

Artificial Intelligence (AI) was employed in the context of this study to support the authors in various auxiliary and editorial tasks. Its use was strictly limited to non-critical and non-confidential aspects of the research, coding and writing process. The primary purpose of employing AI tools was to enhance the overall clarity, coherence, and grammatical accuracy of the text, ensuring a consistent academic writing style. Through text proposals and linguistic adjustments, the readability of the thesis was improved, allowing for a clearer communication of complex concepts to the reader. AI was additionally utilized for translation purposes between English and German, ensuring linguistic precision in both directions.

Furthermore, AI-assisted summarization techniques were employed to help the author better understand complex theoretical or methodological concepts by providing concise explanatory summaries. These summaries served solely as a comprehension aid and were not directly included in the final version of the thesis without the author's verification and adaptation.

In the technical part of the thesis, AI was used to generate and refine `Python` code fragments, as well as to enhance existing code written by the author in terms of readability and computational efficiency. Code completion tools were occasionally used to accelerate programming during the development process. All generated or optimized code was manually integrated into the code base following review and validation by the author. AI tools were also employed for the generation of `LaTeX` table code structures, while the underlying content was entirely created by the author, and for querying specific `LaTeX` commands to ensure accurate formatting. In addition, AI was used to generate concise and descriptive `Git` commit messages for efficient version control and documentation of code changes.

At no point were personal data, business or trade secrets, or any proprietary data, particularly data downloaded from Bloomberg, processed or shared with any AI tool. The handling of such data remained entirely under the author's control and within secure environments compliant with data protection and confidentiality standards.

All textual recommendations and code suggestions provided by AI tools were thoroughly reviewed and verified by the author for technical correctness and factual accuracy. This included empirical testing of all code outputs to ensure functionality and to prevent the inclusion of any erroneous or fabricated content, often referred to as hallucinations. The authors assumes full responsibility for the correctness and validity of all content and code presented in this thesis.

AI was explicitly not used beyond the purposes listed above. The authors assumes full responsibility for the correctness and validity of all content and code presented in this study. Table 13 provides an overview of the specific AI models utilized, including their version information and intended use cases within the scope of this study.

Table 13: **Overview of AI Models Utilized**

AI Model	**Citation Information**
ChatGPT (GPT-5 Model)	OpenAI, *ChatGPT (GPT-5 Model)*, 2025, Version: GPT-5, Used for text refinement including improving grammar, writing style and consistency as well as text proposals. Furthermore, it was used to create `LaTeX` table structures and or querying specific `LaTeX` commands to ensure accurate formatting, https://chat.openai.com
ChatGPT (GPT-5 Thinking Mini)	OpenAI, *ChatGPT (GPT-5 Thinking Mini Model)*, 2025, Version: GPT-5 Thinking Mini, Used for text refinement including improving grammar, writing style and consistency as well as text proposals. Furthermore, it was used to create `LaTeX` table structures and or querying specific `LaTeX` commands to ensure accurate formatting. Note this is the fallback model if the GPT-5 model free quota is exceeded, https://chat.openai.com
Gemini 2.5 Pro	Google DeepMind, *Gemini 2.5 Pro*, 2025, Version: 2.5 Pro, Used for the generation and refinement of code fragments as well as for enhancement of existing code written by the author in terms of readability and computational efficiency. Furthermore it was used to summarize complex theoretical concepts, https://gemini.google.com
Windsurf Chat (Llama 3.1 70B)	Codeium / Windsurf, *Windsurf Chat Model (based on Meta's Llama 3.1 70B)*, 2025, Version: Llama 3.1 70B, Used for coding assistance in the IDE as well as for the creation of Git commit messages, https://codeium.com/windsurf
DeepL Translator	DeepL SE, *DeepL Translator*, 2025, Version: 2025 Release, Used for high-quality translation and linguistic consistency checking, https://www.deepl.com

5 Appendix

5.1 Visual Analysis of Synthetic Data Distributions

The visual analysis of the synthetic dataset confirms the successful implementation of the sampling strategy and highlights the complexity of the learning task. The data distributions are presented in Figures 8 through 10.

Figure 8 displays the histograms and box plots for the primary Heston model parameters: kappa, lambda, sigma, rho, v0, and the risk-free rate r. A defining characteristic of these plots is the perfectly flat, uniform distribution of the histograms and the symmetry of the box plots. This visual evidence validates the use of LHS, ensuring that the high-dimensional input space is covered evenly without gaps or clustering. This uniformity is essential for training a neural network that generalizes well across the entire parameter domain.

Figure 9 continues this analysis with the remaining inputs, tau and log-moneyness, which also exhibit perfect uniformity. However, the subsequent columns in Figure 9 and the plots in Figure 10 reveal the distributions of the model outputs: the option price and the parameter sensitivities (gradients). In sharp contrast to the inputs, these output variables display highly non-uniform distributions.

The gradients, particularly d_kappa (Figure 12) and d_v0 (Figure 10), are characterized by extreme leptokurtosis, with sharp peaks around zero and heavy tails containing significant outliers. For instance, the box plot for d_kappa shows a dense concentration of values near zero but extends to extreme outliers, indicating regions of the parameter space where the option price is highly sensitive to changes in the mean-reversion speed. Similarly, d_sigma exhibits a negative skew, while d_lambda is positively skewed. These complex, heavy-tailed distributions underscore the challenge of the regression task, as the DDN must learn to map uniformly distributed inputs to highly non-linear and peaked output surfaces.

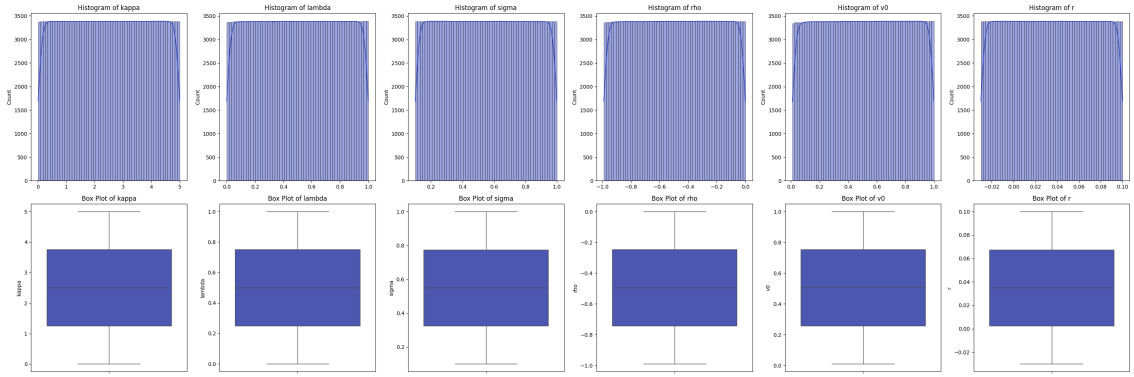

Figure 8: Distributions of the synthetic Heston input parameters (kappa, lambda, sigma, rho, v0, r). The perfectly flat histograms and symmetric box plots confirm the effectiveness of the Latin Hypercube Sampling method in covering the input space uniformly.

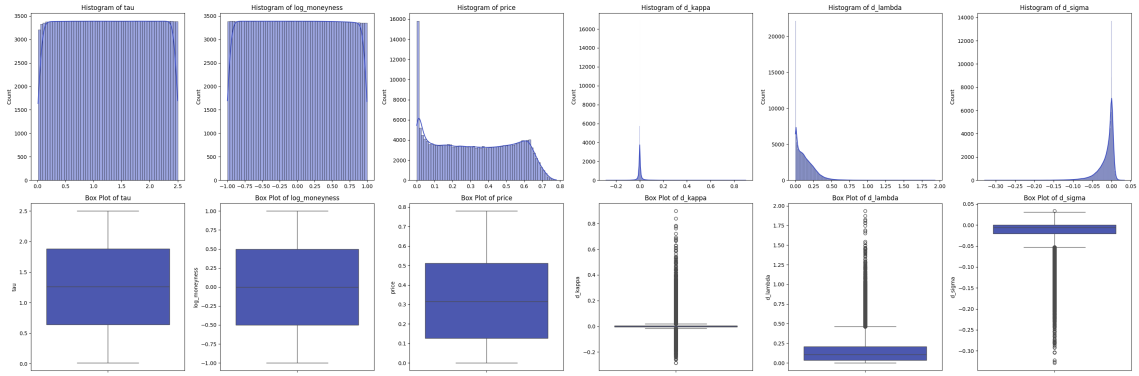

Figure 9: Distributions for time to maturity and log-moneyness (inputs), followed by option price and gradients for kappa, lambda, and sigma (outputs). Note the transition from uniform inputs to highly skewed and peaked output distributions.

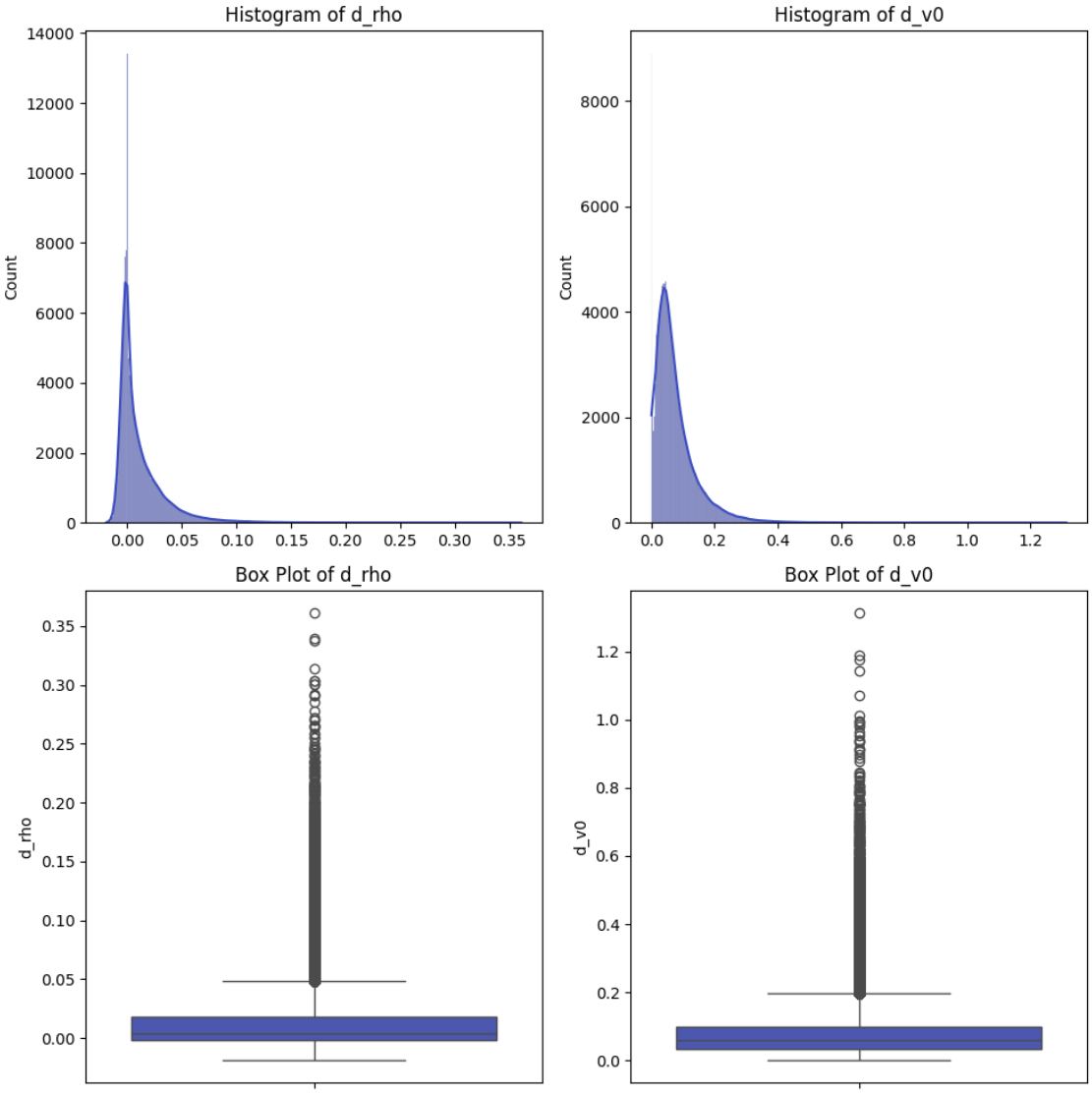

Figure 10: Distributions for the gradients with respect to rho and v0. These labels exhibit significant positive skewness and heavy tails, indicating the presence of regions with high parameter sensitivity.

5.2 Visual Analysis of Historical Data Distributions

The histograms and box plots presented in this section provide a granular visual analysis of the filtered historical AAPL options dataset used for the empirical backtest. These visualizations corroborate the descriptive statistics presented in Section 2.4 and highlight the non-normal nature of financial market data.

Figure 11 displays the distributions for the bid and ask prices, underlying asset price, strike price, days to expiration (DTE), and call implied volatility. A dominant feature across the pricing variables (C_BID, C_ASK) is the extreme positive skewness. The histograms show a high concentration of option prices near zero, with a long right tail extending to over $160. The corresponding box plots confirm this via a dense cluster of outliers in the upper range, representing deep in-the-money contracts or options during periods of high volatility.

The distributions for the underlying asset price and strike price are multimodal. This structure reflects the historical price evolution of AAPL stock over the seven-year observation period, where the stock price spent significant time at different valuation levels (e.g., $150, $300). The alignment between the underlying and strike distributions confirms that the dataset maintains a consistent moneyness relationship throughout the timeline.

Figure 12 extends this analysis to put implied volatility (P_IV), the mid-market call price, log-moneyness, and time to maturity in years (Tau_Years). The implied volatility distributions for both calls (Figure A.1) and puts (Figure A.2) are highly leptokurtic. They exhibit a sharp peak around the mean volatility level (approximately 30-35%) and massive right tails with outliers exceeding 800% (IV > 8.0). These extreme outliers correspond to market stress events, such as the COVID-19 crash, where uncertainty spiked dramatically.

The distribution of time to maturity (DTE and Tau_Years) is heavily right-skewed, indicating that the dataset is dominated by short-term options. The histogram shows a rapid decay in frequency as maturity increases, which is consistent with the liquidity profile of the equity options market where trading volume is concentrated in the front months.

Finally, the distribution of Log-Moneyness (Moneyness_Log) in Figure A.2 stands in contrast to the other variables. It displays a relatively symmetric, bounded distribution centered at zero. This is a direct result of the data filtering protocol which restricted the dataset to options with log-moneyness between -0.25 and 0.25. The absence of outliers in the moneyness box plot confirms that the filtering logic was applied correctly, ensuring that the calibration focused strictly on the liquid, near-the-money region of the volatility surface.

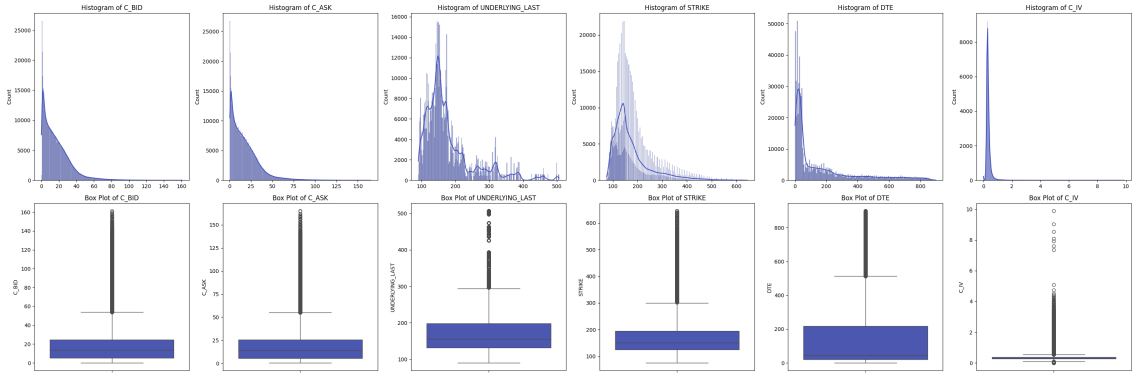

Figure 11: Histograms and box plots for Call Bid, Call Ask, Underlying Price, Strike Price, DTE, and Call Implied Volatility. The plots reveal significant right-skewness in pricing and volatility variables, and a multimodal distribution for the underlying asset.

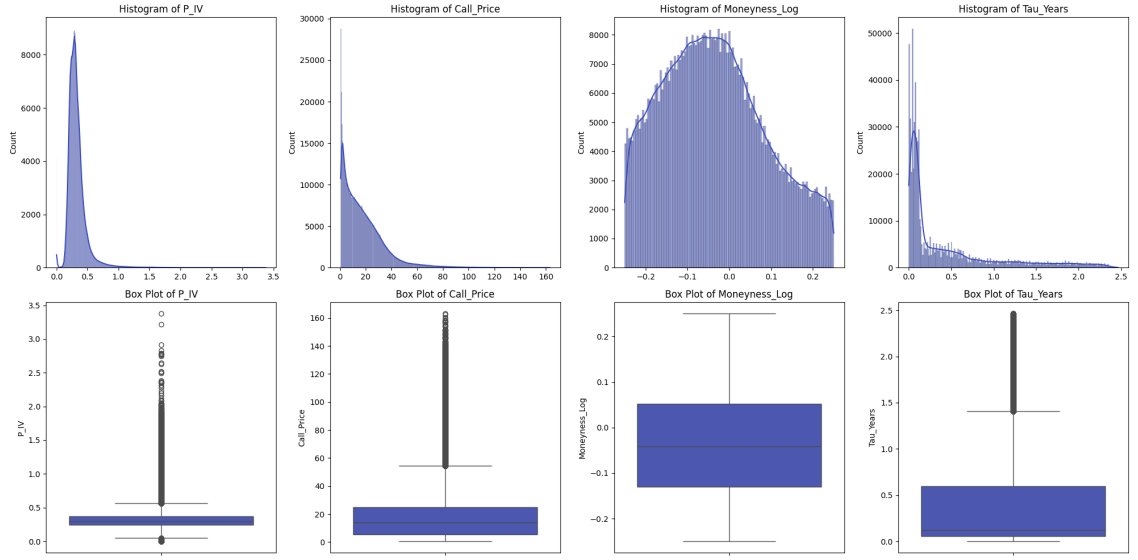

Figure 12: Histograms and box plots for Put Implied Volatility, Call Mid-Price, Log-Moneyness, and Time to Maturity (Years). Note the heavy tails in the volatility distribution and the symmetric, bounded nature of the log-moneyness resulting from data filtering.

5.3 Longitudinal Analysis of Calibrated Parameter Stability

Figure 13 illustrates the daily evolution of the five calibrated Heston parameters $(\kappa, \lambda, \sigma, \rho, v_0)$ over the full backtesting period from 2016 to 2023. This time-series analysis provides critical insights into the stability of the calibration and the model's response to changing market regimes.

The most financially intuitive behavior is observed in the initial variance parameter (v_0), displayed in the bottom panel. It acts as a robust proxy for market fear, exhibiting low, mean-reverting behavior during calm periods (2016-2017) and sharp, distinct spikes during stress events. The most prominent spike corresponds to the COVID-19 crash in March 2020, where v_0 surged to approximately 0.7, correctly reflecting the explosion in spot volatility. Smaller spikes are visible during the

volatility event of early 2018 and the market correction of 2022.

The long-run variance (λ) remains remarkably stable for the majority of the period, hovering between 0.05 and 0.10. This indicates that despite short-term fluctuations in spot variance, the model's view of the long-term equilibrium volatility remained anchored. A notable exception is the singular, extreme spike in early 2018, likely associated with the "Volmageddon" event, where the sudden collapse of short-volatility strategies momentarily disjointed the long-term expectation.

The correlation parameter (ρ) generally adheres to the empirical leverage effect, staying in negative territory between -0.4 and -0.8. However, a regime shift is observable during the 2020-2021 recovery period. During this phase, ρ frequently hits the upper boundary of 0.0, and the volatility of volatility (σ) simultaneously hits its upper boundary of 1.0. This boundary-hitting behavior suggests that the standard Heston model struggled to accommodate the specific smile dynamics of that period—likely characterized by steep skews and high prices for far-out-of-the-money calls—without pushing its parameters to the limits of the constrained search space.

Finally, the mean reversion speed (κ) exhibits high variance and noise throughout the entire sample, oscillating rapidly between 1.0 and 4.0. This is a well-documented phenomenon in Heston calibration, where κ often acts as a slack parameter, absorbing residual fitting errors that the other parameters cannot capture.

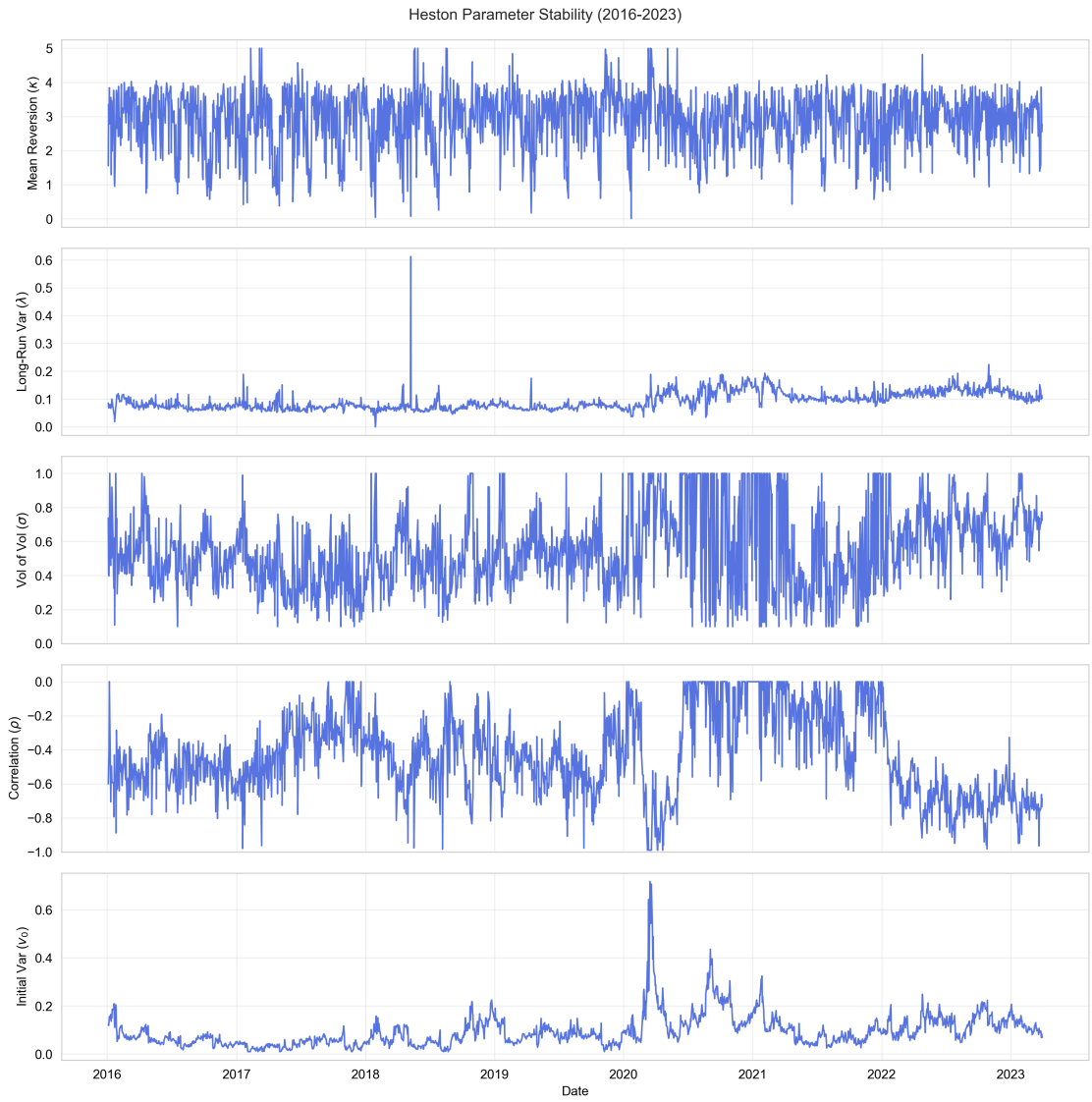

Figure 13: Daily calibrated Heston parameters over the backtesting period (2016-2023). The plots reveal both intuitive market responses (e.g., spikes in v_0 during crises) and challenges in model fitting (e.g., boundary-hitting behavior of ρ and σ during 2020-2021).

5.4 Data Source Files for Historical AAPL Options

The historical AAPL options data utilized in this study is sourced from two comprehensive CSV files obtained from Kaggle. These files encompass daily option quotes spanning from January 2016 to December 2023. The files can be downloaded from the following link: https://www.kaggle.com/datasets/kylegraupe/aapl-options-data-2016-2020

5.5 Code Availability

The code for this study is available on GitHub at https://github.com/skyi28/Calibration-of-the-Hesto