Universitat Pompeu Fabra Barcelona

Faculty of Economics and Business

Academic year: 2024-2025

Final Year Project

Extraction and Analysis of Implied Probabilities in Commodity Options

Gorka - Bravo Díaz (Degree in Economics)

Tutor

Sergi – Barberán Benedicto

CODE: ESC01

(This page is intentionally left blank)

Declaration of Authorship

I, Gorka Bravo Díaz, hereby declare that I am the sole author of this Final Year Project, Extraction and Analysis of Implied Probabilities in Commodity Options, submitted to Universitat Pompeu Fabra in fulfillment of the requirements for the Degree in Economics.

I affirm that all work presented herein is my own, except where explicitly stated otherwise through proper citation. Any sources, ideas, or data derived from external references have been duly acknowledged per academic standards.

Furthermore, I confirm that the use of AI-based tools was for non-substantive purposes, such as grammar checking, formatting assistance, and readability improvements. These tools have not contributed to the development of the core arguments, analysis, or original insights of this thesis.

Gorka Bravo Díaz **03/03/2025**

Acknowledgements

First of all, I want to thank my supervisor sincerely, **Professor Sergi Barberán Benedicto**, for his guidance and support throughout this project. His suggestion to study implied probabilities and his advice on methodology and writing were invaluable.

I am also grateful to **Dr. Montserrat Díaz, M.D. Ph.D.**, for her help with proofreading and ensuring proper citation formatting.

Additionally, I appreciate the assistance of **Professor Elisa Alòs**, who provided helpful insights during office hours regarding implied volatility surfaces.

Finally, I thank my family and friends for their support.

Abstract

This Final Year Project examines the extraction of implied probabilities from WTI crude oil options by leveraging the Breeden-Litzenberger approach and the SABR (Stochastic Alpha, Beta, Rho) model. By calibrating a smooth implied volatility surface, we derive risk-neutral probability density functions (PDFs) that reveal market expectations regarding future crude oil prices and volatility. Our analysis links these PDFs to the WTI futures curve, focusing on how changes in market structure relate to option-implied variance, skewness, and kurtosis.

Furthermore, this project analyzed changes in market dynamics in early 2025 after increased tensions between the US and China. These increased tensions revealed significant changes in skewness and volatility dynamics. Our findings detail the importance of integrating both options-implied probabilities and futures market indicators as a crucial part of a more comprehensive framework for effective risk management and informed trading strategies.

Keywords: Implied Probabilities, Commodity Options, SABR Model, Risk-Neutral Density, WTI Futures, Market Structure, Volatility Dynamics.

Contents

1 Introduction **1**

2 Theoretical Framework **2**

2.1 Butterfly Spreads as proxies for implicit probabilities . . . 2

2.2 The Breeden–Litzenberger Formula . . . 3

2.2.1 Key Assumptions . . . 4

2.3 The need for interpolation . . . 4

2.3.1 Interpolation in volatility space . . . 5

2.4 Black-76 vs Black-Scholes . . . 5

2.5 The SABR model . . . 6

2.6 Characteristics of the Crude Oil Market . . . 7

2.6.1 Oil Futures Curve . . . 7

2.6.2 Contango and Backwardation: Implications for Volatility . . . 7

3 Data **9**

3.1 Collection of the Data . . . 9

3.2 Description of the Data . . . 9

3.2.1 Options Data . . . 9

3.2.2 Futures Data . . . 10

3.3 Cleaning the Data . . . 11

3.3.1 Options Data Processing . . . 11

3.3.2 Implied Volatility Computation . . . 11

3.3.3 Data Validation and Enhancement . . . 11

3.3.4 Quality Control Visualization . . . 11

4 Methodology **12**

4.1 Overview of the Pipeline . . . 12

4.1.1 Options Branch . . . 12

4.1.2 Futures Branch . . . 12

4.2 SABR Calibration . . . 13

4.2.1 Multi-β Comparison . . . 13

4.2.2 Numerical Implementation . . . 14

4.3 Implied PDF Generation and Moment Computation . . . 14

4.3.1 Constructing a Dense Grid of Strikes . . . 14

4.3.2 Computing Call Prices Using SABR-Implied Volatility . . . 15

4.3.3 Approximating the Risk-Neutral PDF . . . 15

4.3.4 Numerical Integration and Moment Calculation . . . 16

4.4 Futures Curve Analysis . . . 17

4.4.1 Data Preprocessing and Term Structure Construction . . . 17

4.4.2 Spreads and Normalization . . . 18
4.4.3 Directional Persistence Measure . . . 18
4.4.4 Composite Index Construction . . . 18
4.5 Statistical Analysis . . . 19
4.5.1 Data Merging and Exploratory Checks . . . 19
4.5.2 Regression Models and Component Analyses . . . 19
4.5.3 Time-Series Extensions and Transformations . . . 20

5 Results **21**
5.1 SABR Calibration and PDF Generation . . . 21
5.2 Statistical Analysis of PDF Moments . . . 22
5.2.1 Regression Results . . . 22
5.3 Market Regime Shift in Early February 2025 . . . 23
5.4 Time Series Analyses . . . 24

6 Discussion **25**
6.1 SABR Calibration and Market Dynamics . . . 25
6.2 PDF Moments and Term Structure . . . 25
6.3 Time Series Properties . . . 25
6.4 Limitations and Future Research . . . 26

7 Conclusion **26**

References **27**

A Supplementary Figures and Tables **A1**
A.1 SABR Multi-β Comparison Tables . . . A1
A.2 SABR Fits of IV . . . A2
A.3 SABR Fits of IV Errors . . . A3
A.4 SABR PDFs . . . A4
A.5 Scatter Plots of OLS Regressions . . . A5

B Extended Derivations and Results **B1**
B.1 2.1 Extended Derivation . . . B1
B.2 Breeden-Litzenberger Derivation . . . B2
B.3 Time Series tests . . . B3

C Code Snippets **C1**
C.1 Full Code . . . C1
C.2 Code Snippets . . . C1
C.2.1 Multi-β Comparison Code . . . C1
C.2.2 SABR & PDF Generation . . . C3

D Pipeline Overview **D1**
D.1 Overview not including statistical analysis D1

E Additional Regressions **E1**
E.1 GARCH Results E17

List of Symbols

Symbol	Definition
$C(K,T)$	Price of a European call option with strike K and maturity T
$P(K,T)$	Price of a European put option with strike K and maturity T
S_t	Spot price of the underlying asset at time t
F_t	Futures price of the underlying at time t
r	Risk-free rate
σ	Volatility of the underlying asset
σ_{IV}	Implied volatility computed from market option prices
σ_{SABR}	Implied volatility under the SABR model
α, β, ρ, ν	SABR model parameters
$f_Q(K)$	Risk-neutral probability density function
$\Phi(x)$	Cumulative distribution function of the standard normal
$\mathbb{E}_Q[X]$	Expectation under the risk-neutral measure Q

List of Abbreviations

Abbreviation	Definition
WTI	West Texas Intermediate
IV	Implied Volatility
ADF	Augmented Dickey-Fuller Test
DTE	Days to Expiration
RND	Risk-Neutral Density
SABR	Stochastic Alpha, Beta, Rho Model
MSE	Mean Squared Error
RMSE	Root Mean Squared Error
ADF	Augmented Dickey-Fuller Test

1 Introduction

The concept of financial options dates back to ancient times, with one of the earliest documented examples found in Aristotle's "Politics." He recounts how Thales of Miletus used his knowledge of astronomy to predict a great olive harvest come the following year and secured options on olive presses, effectively cornering the market when his prediction proved correct. Although not formalized, this Aristotelian anecdote presents the first occurrence of an effective call option.

The first systematic attempt to mathematically model option prices came from Louis Bachelier's doctoral thesis "Théorie de la Spéculation" in 1900. Bachelier's work laid the foundation for modern financial mathematics by introducing the concept of stochastic processes to financial modeling. Before this work, prices of financial options were primarily determined through bargaining. However, his insights remained largely overlooked for over half a century.

The most crucial moment in option pricing theory came in 1973 with the publication of the Black and Scholes model, a year later the Chicago Board Options Exchange (CBOE) opened. Since then, financial mathematics has evolved, producing increasingly sophisticated models to capture market dynamics. The Stochastic Alpha Beta Rho (SABR) model, introduced by Hagan et al. (2002), represents one of those advancements, offering a more flexible framework for modeling volatility behavior.

Today, derivatives markets play a crucial role in price discovery and risk management, particularly in commodity markets where supply and demand factors can create complex term structures. The crude oil market, specifically, has become one of the most actively traded commodity derivatives markets globally, with West Texas Intermediate (WTI) futures and options serving as key benchmarks for global oil pricing (Fattouh, 2011).

This study is motivated by the question: How does the shape of the WTI futures curve relate to the market's assessment of future price uncertainty, as reflected in the implicit probabilities derived from option prices? To answer this question, we aim to study whether and how the term structure of the futures curve affects the moments of the implied probability distribution of option prices.

This research contributes to the existing literature by developing a framework for analyzing the information content of futures and options markets. By examining how the term structure of futures prices relates to the moments of option-implied probability distributions, we aim to better understand how market participants incorporate information across different derivative instruments.

The remainder of this thesis is organized as follows: Chapter 2 provides a review of relevant literature and theoretical background. Chapter 3 describes our data and preliminary processing steps. Chapter 4 details our methodology, including the SABR model calibration and construction of our futures curve index. Chapter 5 presents our empirical results, and Chapter 6 discusses their implications and concludes.

2 Theoretical Framework

In the following chapter, the theoretical underpinnings of this project are discussed. Moving from the first approximations of how implicit probabilities may be calculated, via butterfly spreads, to more exact and mathematically complete models. We also discuss the need for a continuous function of call prices as a function of the strike price and, thus, the need to interpolate the market data. Furthermore, we also provide a review of the concepts of the necessary oil futures market structure in this chapter.

2.1 Butterfly Spreads as proxies for implicit probabilities

Butterfly spreads are a specific option strategy designed to, with limited risk, only yield a positive gross payoff if the underlying security trades in a narrow range. The construction of a long butterfly spread requires three strikes, K_1, K_2, K_3. Where ΔK represents the difference between K_3 and K_1, and K_2 is the midpoint of this difference. Then, a long butterfly spread is constructed by buying one long call option with strike K_1, two short calls with strike K_2, and one long call with strike K_3. (Natenberg, 2014).

To understand why this type of spread may be useful to approximate probability distributions, we consider an arbitrary spread. The maximum payoff possible is achieved at K_2 and is equal to ΔK. Furthermore, consider this payoff is binary. That is, with probability p we receive ΔK and with probability $1-p$ we receive 0. Then the fair price (X) of such a position should be:

$$X = p\Delta K \tag{2.1}$$

Rearranging terms, we get:

$$p = \frac{X}{\Delta K} \tag{2.2}$$

This relation shows that the price X of a butterfly spread implies the probability p of obtaining a positive payoff within the strike interval ΔK. If market prices are correct, constructing such spreads along the option chain should, in theory, allow for the approximation of a risk-neutral density (RND) function (see Martin (2020), part 1).

The intuition derived above can be expressed more formally as described by Cox and Ross (1976). Start with the definition of a probability density function for a small interval ΔK around a strike K. Then the probability that the underlying falls within that range can be approximated by:

$$P(K \leq S \leq K + \Delta K) \approx f(K)\Delta K. \tag{2.3}$$

As ΔK becomes infinitesimally small this probability, which can be denoted as p_K, satisfies:

$$p_K = \lim_{\Delta K \to 0} \{f(K)\Delta K\} \quad \Longrightarrow \quad f_Q(K) = \lim_{\Delta K \to 0} \frac{p_K}{\Delta K}. \tag{2.4}$$

By recognizing that a butterfly spread over ΔK approximates a finite-difference scheme for the second

derivative of the call price with respect to K, one sees how, in the limit as $\Delta K \to 0$, this discrete payoff maps onto $\frac{\partial^2 C(K)}{\partial K^2}$. Under the risk-neutral measure, it follows that $f_Q(K) \propto \frac{\partial^2 C(K)}{\partial K^2}$. A full step-by-step derivation of this result is provided in Appendix B.1, but the essential intuition remains: the price of a narrow butterfly spread pinpoints the local probability mass implied by the market's option prices.

2.2 The Breeden–Litzenberger Formula

In 1978, Breeden and Litzenberger laid the foundation for extracting the risk-neutral probability distribution of future asset prices directly from observed option prices. This result, widely referred to as the *Breeden–Litzenberger formula*, represents a crucial insight in asset pricing theory.

To illustrate the essence of their approach, consider a European call option on an underlying asset S with maturity T and strike price K. Under the risk-neutral measure Q, its fair market value is given by the standard risk-neutral valuation formula:

$$C(K,T) = e^{-rT}\,\mathbb{E}^{Q}\big[\max(S_T - K, 0)\big], \tag{2.5}$$

where r is the constant risk-free rate, and $\mathbb{E}^{Q}$ denotes the expectation with respect to the measure Q. From this starting point, Breeden and Litzenberger's main insight arises from manipulating the payoff $(S_T - K)^+$ as an integral above the strike price and then systematically differentiating with respect to K. This procedure takes advantage of the Fundamental Theorem of Calculus to transform integrals of the risk-neutral density into expressions involving the first and second derivatives of the call price.

More specifically, the integral form of the option payoff can be decomposed into two parts: one capturing the contribution of S_T above K, and another capturing the "discounted strike" component. By differentiating these terms with respect to K, one isolates the probability $Q(S_T > K)$ in the first derivative and then obtains the value of the risk-neutral probability density function $f_Q(\cdot)$ in the second derivative. Thus, while the initial expression in (2.5) is an integral transform of f_Q, differentiation effectively inverts that transform, revealing the density at a specific point.

The full step-by-step derivation, including all intermediate steps, can be found in Appendix B.2. Nevertheless, the fundamental outcome can be stated in the following relationship:

$$f_Q(K) \;=\; e^{rT}\,\frac{\partial^2 C(K,T)}{\partial K^2}. \tag{2.6}$$

This formula captures the risk-neutral PDF of the underlying asset price at maturity T by taking the second partial derivative of the call option price with respect to the strike. In practice, 2.6 allows practitioners and researchers to recover an implied distribution from market option data, thereby making it possible to infer market expectations about volatility, skewness, and tail behavior in a model-free manner. Given its broad utility, the Breeden–Litzenberger formula remains an indispensable tool in modern option pricing theory and market microstructure analysis.

2.2.1 Key Assumptions

The first most relevant assumption is that call prices are continuous in K. This ensures that derivatives of the option price with respect to the strike exist and are well-defined. Without continuity, the process of differentiation could produce erratic or undefined results, undermining the applicability of the Breeden-Litzenberger formula.

The second assumption, tied to no-arbitrage, is crucial for ensuring logical consistency in pricing. Arbitrage-free pricing enforces smooth and coherent relationships between option prices across different strikes (Hull, 2021).

Concerning assumption two, implicit in it, is that $C(K,T)$ must be convex $\forall K$ (Bahra, 1997). Convexity ensures that the second derivative, which corresponds to the implied risk-neutral density, is non-negative. If $C(K,T)$ were not convex, this would imply arbitrage opportunities in the market, as butterfly spreads could be constructed to generate risk-free profits.

2.3 The need for interpolation

As stated in section 2.2.1, under Breeden and Litzenberger (1978) call prices must be continuous in K. But in the real world, option chains are discrete. That is, only some strikes are traded, and the distance between strikes is normally related to moneyness. Strikes closer to being ATM are traded in smaller intervals than deep ITM or OTM strikes. Thus, to use equation 2.6 with market data, we need to construct a continuous $C(K,T)$ function. There are several ways of achieving this goal, and the exact methodology used in this project is explained in Chapter 4. But, there are some practical considerations worth pointing out in this chapter.

First, a straightforward approach is polynomial interpolation, but higher-order polynomials are prone to oscillations, sometimes referred to as "Runge's phenomenon" (Epperson, 1987), especially if the strikes are unevenly spaced. As a result, practitioners often favor lower-order piecewise polynomials or spline-based methods that achieve continuity while reducing unwanted oscillatory behavior.

Spline interpolation first introduced by Schoenberg (1946), is particularly appealing because it stitches together multiple low-degree polynomials in a way that maintains smoothness at the boundaries between intervals. Despite this benefit, spline fitting remains sensitive to how strikes are distributed and to the presence of illiquid or outlier points, which can distort the curve in local regions.

Regardless of the method used to generate $C(K,T)$, there is one essential issue when interpolating in this space. Splines or polynomials don't inherently preserve convexity as noted by Le Floc'h (2021). In the no-arbitrage framework, a European call price must be a non-increasing and convex function in the strike dimension. A non-convex price function is in direct contradiction with the assumptions of Breeden and Litzenberger (1978), creating not only a price function that would yield negative values for $f_Q(K)$, but also arbitrage opportunities as discussed in 2.2.1.

Various specialized methods have been proposed to mitigate these issues. Some practitioners incorpo-

rate additional constraints into spline-fitting routines, ensuring monotonicity or convexity in segments of the strike domain (Aràndiga et al., 2021). Others adopt parametric curves calibrated to market data, but such curves may lack the flexibility needed to capture local imperfections of market data. In many cases, the computational complexity increases as one tries to enforce these no-arbitrage constraints directly on discrete call prices. Consequently, an alternative, and often more stable, approach is to shift the interpolation problem into implied volatility space, where the function to be fitted tends to behave more regularly (see Martin (2020), part 2).

2.3.1 Interpolation in volatility space

Instead of creating a continuous function directly for $C(K,T)$, one can focus on the implied volatility (IV) for a given strike price K and time to maturity T. For each option quote, a unique IV is obtained by inverting the pricing formula to match the observed market price. Specifically, let $\sigma_{\text{IV}}(K,T)$ denote the IV corresponding to strike K and maturity T, it is defined implicitly as

$$C(K,T;\sigma_{\text{IV}}(K,T)) = C_{\text{market}}(K,T), \tag{2.7}$$

where $C(K,T;\sigma)$ is the theoretical option price provided by a chosen pricing model. By doing so, the interpolation challenge shifts to constructing a smooth volatility function $\sigma(K)$ across strikes K for a given maturity T rather than a direct price curve.

Working in volatility space offers several advantages. First, market participants commonly quote implied volatilities rather than direct option prices since IV is often viewed as a more stable or "model-neutral" parameter (Ackerer et al., 2019). Second, the empirical shape of implied volatility curves, for instance, the skew or smile observed in WTI markets, often varies more smoothly than prices themselves, reducing the likelihood of abrupt changes between adjacent strikes or maturities (Romo, 2012). This smoothing effect can help ensure that, once mapped back into prices, the resultant $C(K,T)$ inherits the no-arbitrage properties that a coherent volatility smile (or skew) typically embodies.

Nevertheless, enforcing no-arbitrage when interpolating implied volatilities can still be challenging. A poorly chosen interpolation method can inadvertently create local violations of convexity in the resulting prices. To address this, practitioners frequently resort to parametric forms such as the Stochastic Alpha, Beta, Rho (SABR) model, introduced by Hagan et al. (2002) and discussed further in 2.5, or the Stochastic Volatility Inspired (SVI) parametrization. These frameworks impose conditions that make it easier to preserve vertical arbitrage constraints across the strike axis. For commodities in particular, SABR often captures the pronounced skew and term-structure effects seen in crude oil options (Soini and Lorentzen, 2019), offering a more robust fit than simpler polynomial or spline methods.

2.4 Black-76 vs Black-Scholes

A crucial part of this project is a way to compute implied volatility (IV), as discussed in 2.3.1. To this effect, in 1973 Black and Scholes derived the first option pricing formula. Under assumptions of

frictionless markets, risk neutrality, and arbitrage-free pricing, a European call option with strike K and time to maturity T should be priced as follows:

$$C = S_0 \cdot \Phi(d_1) - K \cdot e^{-rT} \cdot \Phi(d_2) \tag{2.8}$$

where d_1 and d_2:

$$d_1 = \frac{\ln(S_0/K) + (r + \sigma^2/2)T}{\sigma\sqrt{T}} \tag{2.9}$$

$$d_2 = d_1 - \sigma\sqrt{T} \tag{2.10}$$

There is no closed-form solution for obtaining σ_{IV} in this formula, but with computational methods, we can, from observed market prices, derive what level of volatility would yield those same prices. This formula and the others derived by Black and Scholes (1973) were the basis for most option pricing models since. But in the case of equation 2.8, the underlying security trades in a spot market.

In cases where the underlying security is a futures contract. There are significant changes to both the price of the option and the implied volatility. To address forward contracts as underlying securities Black 1976 presented a revised model, under this model the price of a European call option with strike K, time to maturity T and underlying F is:

$$C = e^{-rT}[F \cdot \Phi(d_1) - K \cdot \Phi(d_2)] \tag{2.11}$$

where d_1 and d_2:

$$d_1 = \frac{\ln(F/K) + (\sigma^2/2)T}{\sigma\sqrt{T}} \tag{2.12}$$

$$d_2 = d_1 - \sigma\sqrt{T} \tag{2.13}$$

2.5 The SABR model

As noted in 2.3.1, the SABR (Stochastic Alpha Beta Rho) is a parametric model of IV. First introduced by Hagan et al. (2002), this model offers a flexible and well-established method to capture both the skew and the smile commonly observed in implied volatilities.

The SABR model centers on a joint stochastic process for the underlying forward price $F(t)$ and its instantaneous volatility $\sigma(t)$, which both evolve under the risk-neutral measure. In its most common formulation, the forward price follows:

$$dF_t = \alpha_t F_t^{\beta} \, dW_t^1 \tag{2.14}$$

$$d\alpha_t = \nu \, \alpha_t \, dW_t^2 \tag{2.15}$$

Where W_t^1 and W_t^2 are two correlated Weiner processes with correlation coefficient ρ, thus:

$$\mathbb{E}\left[dW_t^1 \, dW_t^2\right] = \rho \, dt \tag{2.16}$$

The parameters α, β, ρ, and ν encapsulate the model's flexibility. Here, α_t often denotes the initial level of volatility, β in $[0,1]$ governs the elasticity of the underlying price move with respect to its volatility, ρ captures the correlation between price and volatility fluctuations, and ν sets the volatility of volatility.

Hagan et al. (2002) derived an asymptotic formula expressing implied volatility in closed form as a function of $F(t)$, the strike K, and the model parameters. This formula will be discussed further in Chapter 4. Nevertheless, calibrating SABR to a dense set of strikes yields an implied volatility smile (or skew) that behaves smoothly and can reproduce observed market conditions without resorting to high-degree spline interpolations in volatility space.

2.6 Characteristics of the Crude Oil Market

Crude oil occupies a singular position among global commodities, underpinning energy production, transportation, and manufacturing. Its pricing is molded by a broad spectrum of influences, including geopolitical events (Kilian et al., 2024), macroeconomic trends (Le et al., 2024), technological innovations (such as shale extraction), and OPEC production strategies. These factors affect both spot prices and the structure of derivatives markets. In particular, oil futures and options, standardized by the CME, reflect real-time supply and demand conditions as well as market expectations for future scenarios.

2.6.1 Oil Futures Curve

The oil futures curve illustrates the prices of crude oil futures contracts across sequential delivery months, typically plotted with price on the vertical axis and contract expiration on the horizontal axis. This curve can shift significantly in response to changes in production levels, demand outlooks, storage capacity, and broader macroeconomic trends, serving as a barometer of market sentiment on short- to medium-term price movements.

From a theoretical standpoint, the slope of the curve captures more than just supply and demand expectations. In a *contango* market, longer-dated futures trade at higher prices than near-dated contracts, often indicating anticipation of ample supply or highlighting the costs (e.g., storage, insurance, financing) of holding physical oil. Conversely, in a *backwardation* scenario, near-term contracts are priced above later maturities. Such pricing patterns often suggest tighter short-term supply or robust immediate demand, prompting buyers to pay a premium for prompt delivery (Gabillon, 1991).

2.6.2 Contango and Backwardation: Implications for Volatility

While contango and backwardation primarily describe the slope of the oil futures curve, they also hold theoretical implications for volatility and option pricing (see Hogenhout (2024), Kearney and Shang (2019) and Schwartz (1997)). In a *contango* environment, the market tends to view short-term supply as sufficient, potentially leading to relatively stable spot prices and, thus, lower volatility. However,

higher storage costs and financing charges factored into longer-dated contracts can still spur volatility further along the curve.

In *backwardation*, elevated near-term prices typically reflect immediate scarcity or strong short-term demand. This situation can heighten the market's sensitivity to unforeseen disruptions, ranging from extreme weather events to geopolitical tensions, potentially increasing spot price volatility.

Shifts between contango and backwardation can be abrupt, driven by changing fundamentals or geopolitical shocks, with effects on futures prices and implied volatilities. Recognizing when and how the market transitions between these two regimes is, therefore, vital for interpreting the option-implied probability distributions central to this study. In the chapters that follow, we build on these theoretical insights by detailing our data sources and methodological framework, specifically examining whether variations in the shape of the WTI futures curve influence how the market perceives and prices future price uncertainty, including volatility and tail risk. This approach is designed to test our core hypothesis that the term structure of crude oil futures exerts a measurable impact on the implied distributions extracted from option prices.

3 Data

The purpose of this chapter is to present an overview of the data underlying this study and to describe how the raw information was transformed into a format suitable for analysis. We begin by detailing the process through which the data was collected, focusing on the rationale behind choosing Barchart as the primary source. Subsequently, we provide a description of both the options and futures data, emphasizing their structural composition and relevance to the research objectives. Finally, we discuss the procedures undertaken to clean the data, highlighting the methods used to address any gaps, inconsistencies, or other issues that arose during the acquisition phase.

3.1 Collection of the Data

As stated in 3, data for this project was obtained primarily from Barchart, a well-known platform for financial instruments, including equities, options, and futures, offering real-time and historical market data. Although Barchart provides some automation for retrieval, for this project, data was mainly collected manually.

The manual retrieval method typically involved navigating to each instrument's dedicated page on Barchart and downloading real-time market data multiple times a day (approximately 5–10 times) to capture intraday moments. Specifically, these downloads were performed simultaneously for monthly WTI futures contracts, covering expiration dates from 2025 through 2036 and for the 60 DTE WTI option chain. Collecting both datasets in parallel ensured consistent snapshots across instruments, preserving temporal alignment.

Despite the added effort of manual downloads, this method provided opportunities to verify consistency. Each CSV file was reviewed to ensure alignment in column naming conventions, date formatting, and instrument identifiers. This oversight helped prevent typographical errors and inconsistencies that could have propagated to subsequent steps in the analysis. Once all raw files were gathered, a preliminary screening was performed to identify and document any missing values, date mismatches, or anomalous entries. In several cases, the manual process itself facilitated the early detection of irregularities, which could then be addressed before data cleaning.

3.2 Description of the Data

In the following subsection, we aim to dive into the details of the data for both branches of the project, as stated in 3.1, and to lay the foundation of some of the decisions of the data cleaning logic. First, we describe the options data and some of the challenges faced in this dataset, and then we explain the future dataset.

3.2.1 Options Data

The options dataset for this project consists of various fields relevant to the pricing and trading of WTI options. Table 1 illustrates a reduced sample of the raw data, highlighting some of the primary

columns. These columns typically include *Strike*, *Open*, *High*, *Low*, *Last*, *Change*, *Bid*, *Ask*, *Volume*, *Open Int*, *Premium*, *Time*, and *Type*. In the dataset, calls are denoted by an appended "C" (e.g., "50.00C"), making the *Strike* column a string rather than a numeric value, and redundant as there was a existing *Type* column. This formatting detail is addressed and standardized during the data cleaning process (see Section 3.3).

Additionally, it is noteworthy that the raw data files do not provide a direct measure of market-implied volatility (IV). Instead, implied volatility is computed during the cleaning phase. The absence of a dedicated IV column, combined with the need to convert textual strikes to numeric values, were the main challenges that the cleaning process tackled, as well as N/A volumes.

Strike	Open	High	Low	Last	Bid	Ask	Volume	Premium	Time	Type
52.00C	0	20.21	20.21	20.21	18.09	21.84	N/A	20210	01/30/25	Call
53.00C	0	19.22	19.22	19.22	17.24	20.72	50	19220	01/30/25	Call
60.00C	0	12.41	12.41	12.41	12.11	12.21	108	12410	01/30/25	Call
75.00C	1.75	1.83	1.51	1.69	1.66	1.69	416	1690	01/30/25	Call
80.00C	0.64	0.64	0.64	0.64	0.64	0.66	1	640	04:12 ET	Call
85.00C	0.34	0.34	0.34	0.34	0.27	0.29	3	340	20:17 ET	Call

Table 1: *Sample of WTI options data from (reduced rows for illustration). Source: Barchart*

3.2.2 Futures Data

The futures dataset used in this research comprises intraday quotations of monthly WTI contracts with expiration dates ranging from the near future to several years out. Each entry typically includes information on the *Contract* symbol, along with pricing fields such as *Last*, *Change*, *Open*, *High*, *Low*, as well as market activity measures like *Volume* and *Open Interest*. The sample in Table 2 presents a condensed view of the columns and a subset of the records obtained.

Entries for "CLY00 (Cash)" with a *Volume* or *Open Int* marked as N/A reflect reference cash prices that are not tradable contracts. In addition, not all contract months are actively traded, resulting in lower *Volume* and *Open Int* for more distant expirations. These factors are handled during the cleaning process, where inactive or anomalous records can be flagged or removed as necessary (see Section 3.3).

Contract	Last	Change	Open	High	Low	Volume	Open Int	Time
CLY00 (Cash)	72.75	0.14	72.75	72.75	72.75	N/A	N/A	2025-01-30
CLH25 (Mar '25)	72.49	-0.24	73.20	73.49	72.42	51,213	312,178	05:06 CT
CLJ25 (Apr '25)	71.93	-0.25	72.62	72.85	71.86	21,668	159,095	05:06 CT
CLK25 (May '25)	71.34	-0.27	71.95	72.19	71.28	12,521	139,176	05:05 CT
CLZ25 (Dec '25)	67.82	-0.15	68.20	68.41	67.73	3,988	164,651	05:05 CT
CLZ26 (Dec '26)	65.19	-0.06	65.48	65.60	65.12	1,276	77,470	05:01 CT

Table 2: *Sample of WTI futures (reduced rows for illustration). "Cash" indicates a reference price. Source: Barchart*

3.3 Cleaning the Data

The data cleaning process involves several critical steps to ensure the quality and reliability of our options dataset. The primary cleaning script handles both data validation and enrichment, particularly focusing on the computation of implied volatilities using the Black-76 model since the raw data from Barchart lacks market volatility information. In the case of the futures contract information,

3.3.1 Options Data Processing

The cleaning process for options data starts with validation and standardization. The script loads raw CSV files, standardizes column names for consistency, removes trailing option type indicators (C/P) from strike prices, and standardizes option types to lowercase.

3.3.2 Implied Volatility Computation

A key part of the cleaning process is computing implied volatilities. Barchart's raw data lacks market volatility, so we use the Black-76 model, which is ideal for future options. It calculates option prices using the forward price, fitting our WTI futures options analysis. This implementation features both call and put pricing functions and an implied volatility solver that employs the Brent method for effective root finding, ensuring convergence in a reasonable time.

3.3.3 Data Validation and Enhancement

The data validation process includes several filtering steps. Records with near-zero implied volatilities (below 0.0001) are removed as they typically represent pricing anomalies. Similarly, entries with missing or negative bid/ask prices are filtered out, though zero values are permitted to maintain market-depth information. Volume data is treated specially: missing values are filled with zeros instead of being removed since zero-volume options still offer valuable insights into the strike price grid and potential market sentiment.

Mid-prices are computed to enhance the quality of our price data. Additionally, we implement a Gaussian smoothing filter on these to reduce noise while preserving the underlying price structure. This smoothed price series proves particularly useful for subsequent volatility surface construction and analysis.

3.3.4 Quality Control Visualization

The cleaning process concludes with comprehensive data visualization and quality checks. For each cleaned dataset, we generate implied volatility plots that separate calls and puts, allowing visual inspection of the volatility smile and potential put-call parity violations. These visualizations serve as both a quality control mechanism and a tool for identifying market anomalies that might require further investigation.

4 Methodology

This chapter describes the methodology and the entire Python-based workflow used in this research. First, we detail how the SABR model is calibrated and how the implied probability density functions (PDFs) are generated. Second, we explain the construction of the WTI futures curve and the numerical approach used to quantify its shape. Finally, we outline the statistical analysis employed to test our main hypothesis.

A flowchart of the full pipeline can be found in Appendix D.1, while key portions of the code and a link to the corresponding GitHub repository are provided in Appendix C.1. Data cleaning procedures, such as filtering out stale quotes and adjusting for strikes, are covered in detail in Chapter 3.

4.1 Overview of the Pipeline

Although the entire workflow can be executed in a sequential manner, it can also be divided into two main branches, referred to in the code as "OptionsF" (or simply "Options") and "Futures." These names correspond to the types of financial instruments each branch handles.

4.1.1 Options Branch

The raw options data, once downloaded, first undergoes a data cleaning phase. This step removes stale quotes, standardizes the strikes, and generates implied volatilities (IVs). Next, a module constructs butterfly spreads for each option chain, normalizes these spreads, and produces the corresponding implied PDFs. These PDFs can sometimes be unreliable due to the limited range of valid strikes, but they are useful for preliminary comparisons.

The next phase involves examining different values of β in the SABR model (see Section 4.2.1 for details). For each β, we calibrate the remaining SABR parameters (α, ρ, and ν)and compute error metrics (MSE and RMSE). The results are stored in a table (an example is shown in Appendix A). Finally, once the preferred β is determined, the model is calibrated to produce a final set of parameters and an implicit PDF is generated for each date in the sample.

4.1.2 Futures Branch

The Futures branch is relatively straightforward. It uses snapshots of currently traded WTI futures contracts, which may range in expiration from, for example, 2025 up to 2036, depending on availability at the time of download. Each contract is then assigned a numerical month code (1 to 12) based on its expiration month, this process is further explained in 4.4. After labeling, various components of the futures-based index are computed, such as front-month vs. back-month spreads, to capture the slope or shape of the term structure. Finally, the resulting index values and their components are saved in a CSV file. These data are later combined with the outputs from the Options branch for the final statistical analysis.

4.2 SABR Calibration

As discussed in section 2.5, the SABR model proposed by Hagan et al. (2002) characterizes the forward price F_t and its volatility α_t as correlated stochastic processes under the risk-neutral measure. Under these dynamics, the Hagan approximation provides a closed-form expression for the implied volatility, which we refer to as σ_{SABR}. In its extended form, it can be written as

$$\sigma_{\text{SABR}} = \frac{\alpha}{\left(FK\right)^{\frac{1-\beta}{2}}} \frac{z}{x(z)} \left[1 + \left(\frac{(1-\beta)^2}{24} \frac{\alpha^2}{\left(FK\right)^{1-\beta}} + \frac{1}{4}\rho\,\beta\,\nu\,\alpha\left(FK\right)^{\frac{\beta-1}{2}} + \frac{2-3\rho^2}{24}\nu^2\right)T\right], \tag{4.1}$$

where

$$z = \frac{\nu}{\alpha}\left(F^{1-\beta} - K^{1-\beta}\right), \quad x(z) = \ln\left(\frac{\sqrt{1-2\rho z+z^2}+z-\rho}{1-\rho}\right). \tag{4.2}$$

In the Python code implementation, these correction terms

$$\frac{(1-\beta)^2}{24}\frac{\alpha^2}{\left(FK\right)^{1-\beta}}, \quad \frac{1}{4}\rho\,\beta\,\nu\,\alpha\left(FK\right)^{\frac{\beta-1}{2}}, \quad \frac{2-3\rho^2}{24}\nu^2$$

are separately labeled as `A1`, `A2`, and `A3` for convenience. These are the common correction factors in the extended Hagan approximation.

4.2.1 Multi-β Comparison

The parameter β in the SABR model determines the elasticity of volatility with respect to the underlying forward price. In industry practice, it is sometimes set to a fixed value (often 0.5) or calibrated jointly with the other parameters if sufficient market data are available. Here, we explore a range of possible β values in order to determine which one yields the most reasonable calibration fit.

Our approach involves defining a set of candidate β values, each lying between 0.05 and 0.90 (see C.2.1 line 49 for the exact values), and calibrating α, ρ, and ν for each fixed β. The calibration procedure minimizes the mean squared error (MSE) between the observed market implied volatilities σ_i^{mkt} and the SABR-implied volatilities $\sigma_i^{\text{SABR}}(\alpha,\beta,\rho,\nu)$ across a set of strikes. For each β, we therefore solve

$$\min_{\alpha,\rho,\nu} \frac{1}{N}\sum_{i=1}^{N}\left(\sigma_i^{\text{mkt}} - \sigma_i^{\text{SABR}}(\alpha,\beta,\rho,\nu)\right)^2, \tag{4.3}$$

subject to $\alpha \geq 0$, $\nu \geq 0$, and $\rho \in [-1,1]$. Once the calibration converges, we record the resulting parameters (α,ρ,ν) and the calibration errors (MSE and RMSE) as can be seen in Table 4 in Appendix A.1.

A simple criterion might be to select the β that produces the lowest MSE. However, we also examine the parameter estimates to ensure that they lie within a plausible range. In particular, if a particular β leads to unreasonably large α or extremely small ν (or vice-versa), we disregard that choice even if it has a slightly lower MSE. This two-step criterion (balancing fit quality and parameter plausibility)

provides overall good fits while maintaining sensible parameter values that still maintain interpretability.

4.2.2 Numerical Implementation

Calibrating the SABR model requires solving a nonlinear least-squares problem for each candidate β. We implement this using `scipy.optimize.least_squares`. The solver takes an initial guess for α, ρ, and ν, as well as upper and lower bounds to keep each parameter in a financially feasible domain. The choice of initial conditions is $\alpha_0 = 0.5$, $\rho_0 = 0.0$, and $\nu_0 = 0.5$, although we occasionally vary these starting points for testing purposes. Once the solver is complete, we retrieve the final parameter estimates and compute the corresponding MSE and RMSE.

The calibration routine is repeated for every β in the candidate set, which produces a mapping $\beta \mapsto (\alpha_\beta, \rho_\beta, \nu_\beta, \text{MSE}_\beta)$. This mapping is then examined to identify which β leads to the most consistent fit and sensible parameter values. Although our procedure is heuristic, it aligns with common industry practice, where practitioners often prefer a stable, interpretable parameter set over a marginally smaller MSE that might accompany highly volatile or non-intuitive parameter estimates (West (2005) and Kennedy et al. (2012)).

Final parameter values are logged alongside the calibration errors. Depending on the richness of the option data, the solver generally converges for each β. If a non-convergent run is encountered, we discard that solution and continue, although this has been a rare occurrence. Once a final β is identified, we move to a separate document where the calibration is repeated once more to confirm the parameters $(\alpha^*, \rho^*, \nu^*)$, and for further steps.

4.3 Implied PDF Generation and Moment Computation

This subsection details the procedure for extracting the risk-neutral probability density function (PDF) implied by the calibrated SABR model. The overarching idea is to construct a dense set of call prices across a range of strikes, using the SABR-implied volatility within the Black–76 framework, and then approximate the PDF via the second derivative of the discounted call price. Once this PDF is obtained, numerical integration allows for the calculation of the first four moments of the PDF (mean, variance, skewness, and kurtosis). The following pages describe each step in detail, starting with how the strike grid is chosen and proceeding to the derivation of model-based call prices.

4.3.1 Constructing a Dense Grid of Strikes

The first step in generating an implied PDF is to define a sufficiently dense grid of strikes. In practice, the grid should include values below and above the forward price F so that any relevant tail behavior is captured. A common heuristic is to set $K_{\min} = \max(F - \Delta, 1)$ to ensure that the lower bound remains positive, where Δ might be a fraction of F (e.g., 80%) or a fixed number of points. The upper bound $K_{\max}$ can then be set to $F + \Delta$, ensuring that both ITM and OTM regions are included.

Within this interval $[K_{\min}, K_{\max}]$, an evenly spaced set of N strike points $\{K_1, K_2, \ldots, K_N\}$ is chosen. The exact value of N depends on the desired trade-off between computational speed and numerical accuracy, but it often ranges from a few hundred up to a few thousand points, in the case of this project we find that a good number is 4000. A larger N tends to smooth out noise in the finite-difference calculations that will later be used to estimate the second derivative of the call price. Once the grid is established, the methodology proceeds by mapping each strike K_i to a model-implied call price.

4.3.2 Computing Call Prices Using SABR-Implied Volatility

After defining the strike grid, the code computes the corresponding call prices by combining the SABR-implied volatility with the Black–76 pricing formula. For each strike K_i, the extended Hagan approximation is used to evaluate

$$\sigma_{\text{SABR}}(K_i) \;=\; \sigma_{\text{SABR}}\big(F, K_i; \hat{\alpha}, \beta, \hat{\rho}, \hat{\nu}\big),$$

where $\hat{\alpha}$, $\hat{\rho}$, and $\hat{\nu}$ are the calibrated parameters. This volatility is then fed into the Black–76 model:

$$C(K_i) \;=\; e^{-rT}\big[F\,\Phi(d_1) \;-\; K_i\,\Phi(d_2)\big],$$

with d_1 and d_2 derived in the usual way, as discussed in 2.4, but using $\sigma_{\text{SABR}}(K_i)$ instead of a constant volatility. Repeating this step across the entire strike grid yields a smooth array of call prices $\{C(K_1), \ldots, C(K_N)\}$. These model-generated prices form the basis for the subsequent PDF estimation, as the risk-neutral density $f(K)$ can be approximated by applying a central-difference scheme to the discounted prices $e^{-rT}C(K)$.

4.3.3 Approximating the Risk-Neutral PDF

Once a discrete set of model-based call prices $\big\{C(K_1), C(K_2), \ldots, C(K_N)\big\}$ has been computed across a strike grid $\{K_1, K_2, \ldots, K_N\}$, the risk-neutral PDF can be estimated by invoking equation 2.6 as discussed in 2.2 the discounted call price and the density in the Black–76 framework.

$$f(K_i) \;= e^{rT}\,\frac{\partial^2}{\partial K^2}\,C(K_i), \tag{4.4}$$

where the factor e^{rT} compensates for the discounting inherent in $C(K)$. This formula reflects the fact that under the risk-neutral measure and Black–76 assumptions, the second derivative of the discounted call price recovers the implied PDF. In practice, we implement this second derivative numerically using finite differences, as described below.

To evaluate $\frac{\partial^2 C(K)}{\partial K^2}$ in a discrete setting, we rely on a central-difference scheme. Suppose we have three consecutive strike points $\big(K_{i-1}, K_i, K_{i+1}\big)$ and their corresponding call prices $\big(C_{i-1}, C_i, C_{i+1}\big)$. We first

compute the first-order differences on each subinterval:

$$\Delta_+ \;=\; \frac{C_{i+1}-C_i}{K_{i+1}-K_i}, \qquad \Delta_- \;=\; \frac{C_i-C_{i-1}}{K_i-K_{i-1}}. \tag{4.5}$$

We then subtract these slopes and scale by the distance between K_{i-1} and K_{i+1}, which yields an approximation of the second derivative at K_i:

$$\left.\frac{\partial^2 C}{\partial K^2}\right|_{K_i} \;\approx\; \frac{2\left(\Delta_+ - \Delta_-\right)}{K_{i+1}-K_{i-1}}. \tag{4.6}$$

Finally, we multiply by e^{rT} to obtain the PDF value $f(K_i)$. More explicitly,

$$f(K_i) \;=\; e^{rT}\,\frac{\partial^2 C}{\partial K^2}\Big(K_i\Big).$$

These PDF estimates $\big\{f(K_2), f(K_3), \ldots, f(K_{N-1})\big\}$ are computed at each interior node of the grid, since a central-difference scheme requires valid neighboring points on both sides. The outermost grid points K_1 and K_N generally require alternative finite-difference approximations (Floc'h and Kennedy, 2016) or are omitted to avoid boundary artifacts. Once the PDF is approximated at these interior strikes, it can be plotted and numerically integrated to verify normalization and compute statistical moments, as will be discussed in subsequent sections.

4.3.4 Numerical Integration and Moment Calculation

An essential step after approximating the PDF is to verify how closely it integrates to one and to extract important summary statistics such as mean, variance, skewness, and kurtosis. In practice, this is achieved by a discrete trapezoidal-rule approximation. Suppose the strike grid is given by $\{K_1, K_2, \ldots, K_N\}$ and the corresponding PDF values are $\{f(K_1), f(K_2), \ldots, f(K_N)\}$. The total probability is estimated by summing up trapezoids from K_1 to K_N:

$$\sum_{i=1}^{N-1} \frac{1}{2}\Big[f(K_i)+f(K_{i+1})\Big]\,\big(K_{i+1}-K_i\big). \tag{4.7}$$

This expression should be close to 1 if the grid and the second-derivative approximations have adequately captured the PDF's mass.

Once the total area under the PDF is confirmed, the first moment (mean) is obtained by integrating $K\,f(K)$ over the same interval and dividing by the total area:

$$\mu \;=\; \frac{\displaystyle\int K\,f(K)\,dK}{\displaystyle\int f(K)\,dK} \;\approx\; \frac{\sum_{i=1}^{N-1}\frac{1}{2}\big[K_i f(K_i)+K_{i+1}f(K_{i+1})\big]\,(K_{i+1}-K_i)}{\sum_{i=1}^{N-1}\frac{1}{2}\big[f(K_i)+f(K_{i+1})\big]\,(K_{i+1}-K_i)}. \tag{4.8}$$

The variance σ^2 follows from the second moment minus the square of the mean:

$$\sigma^2 = \frac{\int K^2 f(K)\,dK}{\int f(K)\,dK} - \mu^2, \tag{4.9}$$

again approximated numerically by applying the trapezoidal rule to $K^2 f(K)$.

Higher-order statistics such as skewness and kurtosis can then be computed in a similar fashion. The skewness measures asymmetry in the distribution, with

$$\text{Skewness} = \frac{\int (K-\mu)^3 f(K)\,dK}{\sigma^3 \int f(K)\,dK}, \tag{4.10}$$

while the kurtosis gauges the degree of tail heaviness or peakedness:

$$\text{Kurtosis} = \frac{\int (K-\mu)^4 f(K)\,dK}{\sigma^4 \int f(K)\,dK}. \tag{4.11}$$

In the code, these calculations are performed by looping through the strike grid, summing partial trapezoids for expressions like $K^n f(K)$, and then dividing by the total probability. This process provides insights into the implied distribution's central tendency, dispersion, and tail behavior. By logging these moment estimates alongside the calibrated SABR parameters, one can observe how shifts in market conditions may influence the underlying risk-neutral distribution.

4.4 Futures Curve Analysis

Understanding whether the shape of the WTI futures curve influences the option-implied perception of volatility and tail risk first requires a systematic way to quantify market structure. A comprehensive analysis must consider both the temporal spread across multiple contract maturities (1 month, 3 months, and 1 year) and whether the market is in contango or backwardation. To capture these features in a single measure, we develop a composite index that also includes a directional persistence factor, reflecting consecutive price changes along the curve. This index becomes an essential input to our broader hypothesis: namely, that fluctuations in the WTI futures curve may correlate with, and potentially help explain, how the market prices future uncertainty as reflected in option-implied distributions.

4.4.1 Data Preprocessing and Term Structure Construction

The raw dataset, as explained in 3.2.2 consists of WTI futures contracts identified by standardized CME symbols (e.g., `CLH25`, `CLM25`). Each symbol encodes its expiration month and year using a

letter code for the month (e.g., H for March, M for June) followed by two digits for the year (e.g., 25). A dedicated function, `parse_contract`, interprets these components and converts them into a proper `YYYY-MM-DD` format. The dataset is then sorted chronologically, ensuring that each contract's price aligns with its correct time to expiration.

4.4.2 Spreads and Normalization

After constructing the ordered series of contracts, three primary spreads are computed relative to the front-month (nearest) contract. Let the front-month price be denoted by P_t, and suppose P_{t+i} represents the price of a contract expiring approximately i months (or a specified time horizon) after P_t. The raw percentage spread for that horizon is

$$\frac{P_{t+i} - P_t}{P_t}. \tag{4.12}$$

In this work, we focus on spreads at 1 month, 3 months, and 1 year from the front contract. These spreads are further normalized against expected benchmarks (e.g., 1% for the 1-month spread, 3% for the 3-month spread, and 5% for the 1-year spread) to form *spread severities*. A hyperbolic tangent function is then applied:

$$\text{NormSpread} = \tanh\Big(\text{SpreadSeverity}\Big), \tag{4.13}$$

which bounds each normalized spread, mitigating the influence of extreme market outliers.

4.4.3 Directional Persistence Measure

In addition to these time-based spreads, a *directional persistence* term is computed to capture the extent to which consecutive contracts on the curve move in the same direction (all increasing or all decreasing). Mathematically, the code identifies consecutive downward moves via $\Delta_j = (P_{j+1} - P_j) < 0$ and then computes the fraction of Δ_j that evaluate to `True`. This fraction is shifted and scaled to lie in the range $[-1, 1]$, with negative values implying predominantly upward moves and positive values indicating primarily downward moves. Moreover, the persistence measure is further modulated by the absolute magnitude of the 1-year spread, so that strong contango or backwardation amplifies the effect of persistence, whereas mild curve shapes dampen it.

4.4.4 Composite Index Construction

The final composite index I combines the three normalized spreads (for 1 month, 3 months, and 1 year) and the directional persistence component. Specifically, let $n_{1\text{m}}, n_{3\text{m}}, n_{1\text{y}}$ be the three normalized spreads, and let p denote the persistence impact. A linear combination of these terms yields.

$$I = 0.25\, n_{1\text{m}} + 0.35\, n_{3\text{m}} + 0.25\, n_{1\text{y}} + 0.15\, p. \tag{4.14}$$

Finally, I is clipped to lie in $[-1, 1]$ to avoid runaway values. Negative values indicate backwardation (nearby contracts priced higher than those farther out), whereas positive values signal contango

(longer-dated contracts priced higher). The absolute magnitude of I gauges the intensity of the given regime.

Because each spread is normalized and further passed through $\tanh(\cdot)$, the index remains robust to outliers. At the same time, incorporating directional persistence provides insight into whether the market's shape has been stable or is undergoing frequent directional shifts. The resulting single measure thus unifies disparate aspects of the futures term structure into a concise numerical indicator, convenient for tracking changes over time or comparing different market snapshots.

4.5 Statistical Analysis

4.5.1 Data Merging and Exploratory Checks

In order to study the relationship between the WTI futures curve and the implied probability density functions (PDFs) of WTI options, two datasets were combined. The first dataset, `Stats.csv`, contains all pertinent information about the structure of the futures curve, while the second dataset, `Stats2.csv`, includes the main PDF-based statistics (mean, variance, skewness, and kurtosis) extracted from the options data. Each file provides a time-stamped "snapshot" of market conditions, ensuring that rows from both files could be merged seamlessly once an appropriate ordering key was established.

After merging, basic data-cleaning checks were performed to remove any missing or anomalous values and verify that each row accurately represented a single time point. Descriptive statistics (e.g., mean, standard deviation, minimum, and maximum) were then computed for all variables of interest. These summary measures provided an initial sense of the data's range and distribution. Additionally, a correlation matrix of the key variables was generated to gauge the strength and direction of their linear relationships. To supplement this numerical check, scatterplots and pairplots were reviewed, confirming whether any obvious nonlinearities or outliers might affect subsequent modeling steps.

4.5.2 Regression Models and Component Analyses

Once exploratory analysis indicated plausible connections between the futures-based predictor(s) and the PDF moments, a series of Ordinary Least Squares (OLS) regressions were run. In each model, the dependent variable was one of the PDF moments (mean, variance, skewness, or kurtosis), while the explanatory variable(s) originated from the merged dataset's futures-curve information. Standard errors were corrected for heteroskedasticity using HC3 adjustments to ensure robust inference.
For a univariate OLS regression of the form

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t, \tag{4.15}$$

where y_t might be, for instance, the skewness of the implied PDF at time t and x_t the futures-based predictor, the coefficient β_1 was estimated via least squares. In practice, we iterated this specification for each of the PDF moments to identify the most salient relationships. Graphical checks, including

scatter plots with best-fit lines, were used to illustrate each model's fit.

Because the key predictor could be decomposed into multiple components (e.g., short-term vs. long-term spreads), additional regressions were conducted on each component individually. This step helped pinpoint which part of the curve exerted the strongest relationship on each PDF moment. The outcome of these component-wise analyses clarified whether shorter-dated or longer-dated futures carried the most significant explanatory power for market-implied probabilities, as reflected in skewness, variance, or other moments.

4.5.3 Time-Series Extensions and Transformations

To capture potential time-dependent effects, we introduced lagged versions of the futures-based predictor. Specifically, a one-period lag (x_{t-1}) was added to some models to account for any delayed influence the term structure may exert on the PDF moments. We also explored transformations of key PDF variables to address potential nonlinearities. Given that variance can grow at a nonlinear rate, we examined the standard deviation, the logarithm of variance, and the logarithm of the standard deviation, each of which was regressed on both the current and lagged values of the primary predictor. Comparing these regression outputs helped reveal nonlinear relationships and reduced the impact of large outliers.

Additionally, because volatility is often recognized as non-stationary, we tested GARCH and GARCH-X models to allow for time-varying volatility. While these results are presented in Appendix E.1, the intricacies of these models largely fall outside the scope of this thesis. Nevertheless, this multi-phase procedure provides a robust framework for understanding how the term structure of WTI futures might inform or correlate with the risk-neutral PDF moments implied by options. In the following chapter, we present the findings for each phase of the pipeline, as well as the outcomes of the accompanying statistical analyses.

5 Results

This chapter presents the key findings of our study, focusing on whether and how the WTI futures curve influences the market's perception of price uncertainty. First, we describe the SABR calibration process that generates the implied volatility surface and the resulting risk-neutral probability density functions (PDFs). Next, we detail the statistical analysis examining how our composite index of the futures curve relates to key PDF moments, namely variance and skewness. Finally, we investigate evidence of a possible market regime shift and discuss the time–series diagnostics performed, ensuring that our conclusions about the futures curve's impact on implied volatility and tail risk remain robust across different market conditions.

5.1 SABR Calibration and PDF Generation

For each observation, the SABR model was calibrated to the observed market volatilities. Representative SABR calibration plots (see Figure 1) show the fitted IV curve alongside the market data, while a corresponding set of SABR error plots (Figure 2) illustrates that the model residuals remain within acceptable bounds.

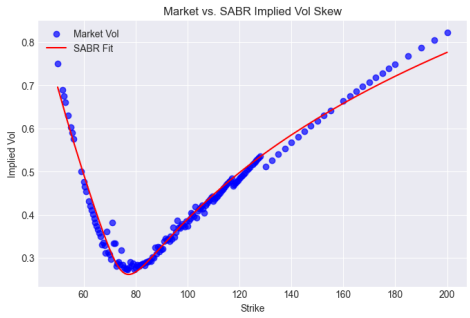

Figure 1: *Representative SABR Fits for IV. Additional SABR fit figures are provided in A.2. Source: own.*

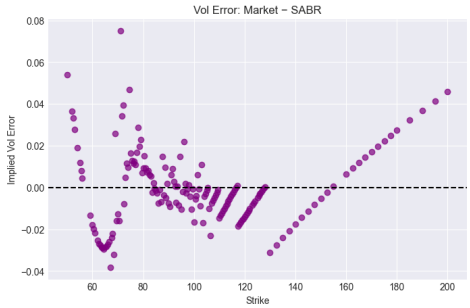

Figure 2: *SABR Calibration Error Plots. Full error plots for all observations are provided in A.3. Source: own.*

Using the calibrated IV smiles, we constructed risk–neutral densities (RNDs) via the Breeden–Litzenberger

approach as discussed in 4.3.3. Figure 3 displays a representative set of PDFs derived from the SABR–implied volatility surfaces. These PDFs capture important characteristics of the market's implied view on future volatility, downside risk, and the underlying future price.

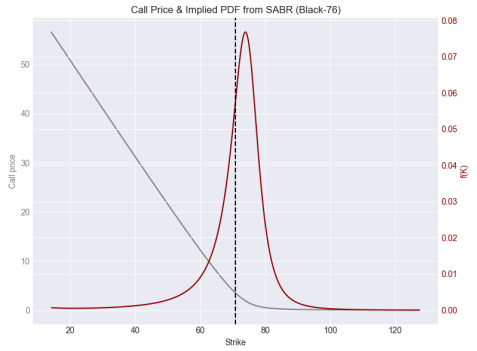

Figure 3: *Representative Risk–Neutral Densities (RND) Derived from SABR. Full PDFs for all observations are provided in A.4. Source: own*

5.2 Statistical Analysis of PDF Moments

As discussed in the methodology chapter, as a preliminary investigation of the impact of the futures curve on the option–implied PDF, we computed key statistical moments (mean, variance, skewness, and kurtosis) from the derived RNDs. We then estimated a series of robust regressions relating these moments to our composite index.

5.2.1 Regression Results

Our regression analysis reveals that:

- **Variance and Kurtosis:** The regressions for variance and kurtosis show very high explanatory power (with R^2 values exceeding 0.80) and strong statistical significance. These results indicate that the dispersion and tail behavior of the PDF are closely associated with the shape of the futures curve.
- **Mean and Skewness:** The relationships for mean and skewness are statistically significant, though the models explain a more moderate share of the variation (with R^2 values around 0.27–0.37). Nevertheless, when inspecting the scatter plot of both of these variables (see A.5), we can observe some behavior that may signal a deeper relationship (see 5.3).

Presented here is the regression output table for variance as the dep. variable, and index as a simple explanatory variable (all regression outputs can be found in E).

Table 3: *OLS Regression Results for Variance: Regression of* `variance` *on* `index`*. Source: own.*

Dep. Variable:	variance	**R-squared:**	0.825
Model:	OLS	**Adj. R-squared:**	0.820
Method:	Least Squares	**F-statistic:**	246.8
Date:	Wed, 12 Feb 2025	**Prob (F-statistic):**	3.29e-18
Time:	15:36:12	**Log-Likelihood:**	-153.37
No. Observations:	40	**AIC:**	310.7
Df Residuals:	38	**BIC:**	314.1
Df Model:	1		
Covariance Type:	HC3		

	coef	**std err**	**z**	**P$>\|$z$\|$**	**[0.025**	**0.975]**
const	-145.4364	13.922	-10.447	0.000	-172.723	-118.150
composite_calculated	-398.7826	25.382	-15.711	0.000	-448.531	-349.034

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)

5.3 Market Regime Shift in Early February 2025

Our time series analysis of the option-implied PDF skewness reveals that the overall low R^2 in the skewness regression is driven by the presence of two distinct regimes. In particular, a clear structural break is observed on February 5, 2025. Prior to this date, the relationship between the composite index—which measures the shape of the WTI futures curve—and the PDF skewness exhibits a strong negative association. A regression based solely on the pre–break period yielded an R^2 of 0.905 and a coefficient of -6.7129 (with $p < 0.001$). During this period, the environment was characterized by relatively high variance levels, around 100. In contrast, after February 5, 2025, the relationship, although still significant, becomes weaker, with an R^2 of 0.824 and a coefficient of -2.9047 (with $p < 0.001$); here, the variance level drops substantially to approximately 40. When the two regimes are combined in a single regression, these contrasting effects tend to cancel each other out partially, resulting in a lower overall R^2 and a regression coefficient that does not fully capture the distinct dynamics present in each period.

Figure 4 illustrates the time series of PDF skewness with a marked jump on February 5, 2025, clearly signaling the structural break. In addition, Figure 5 presents the evolution of the implied volatility surface around the breakpoint. Prior to February 5, the volatility surface exhibited a pronounced smile and skew, with the SABR β parameter in the range of approximately 0.2 to 0.4, indicating that market participants priced in the likelihood of large price jumps. Following the break, the volatility surface became noticeably flatter, and the SABR β parameter increased to a range of approximately 0.7 to 0.9, reflecting a transition toward more log-normal dynamics. This shift in the SABR β parameter suggests a fundamental change in market behavior, wherein the risk pricing moved away from anticipating extreme movements toward a more standard distribution of returns.

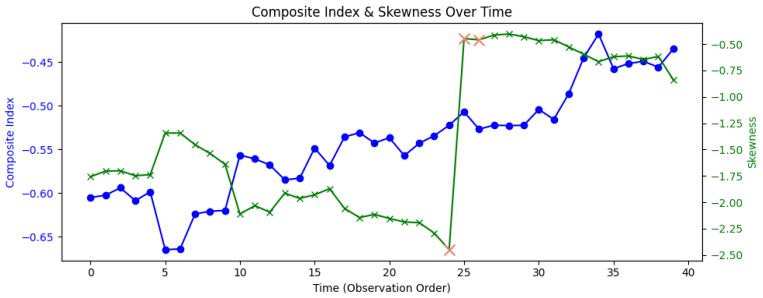

Figure 4: *Time series of PDF skewness showing a structural break on February 5, 2025. ($t = 25$). Source: own.*

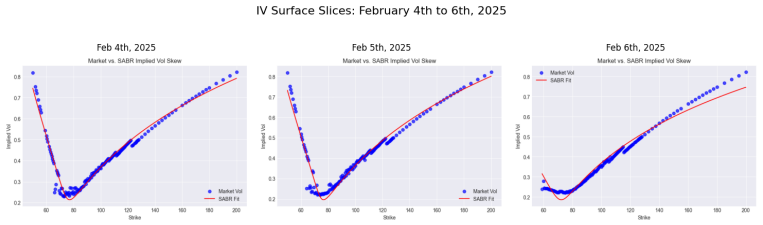

Figure 5: *Evolution of the IV smile around February 5, 2025. Source: own.*

The timing of this shift coincided with WTI hitting its lowest settlement price of the year on February 5, 2025, driven by rising U.S. crude inventories and escalating U.S.–China trade tensions. These developments triggered both a significant price movement and a shift in market sentiment. As traders grew more risk-averse, the futures curve and options-implied probabilities adjusted, leading to changes in both PDF skewness and volatility surface dynamics.

5.4 Time Series Analyses

In addition to the abovementioned analysis, further time series analyses were conducted to understand the dynamics between the futures curve and the option-implied PDF moments. Stationarity tests using the ADF procedure confirmed that the composite index and PDF moment series are stationary or can be so, while cointegration tests validated that the observed relationships are not spurious (a summary of these results can be found in B.3, with detailed output tables in Appendix E.1).

Diagnostic tests on the residuals of the variance regression showed significant autocorrelation, indicating the model could use additional lags or dynamic specifications. An ARCH test revealed borderline evidence of volatility clustering, consistent with persistent volatility shocks. Granger causality tests indicated that while the composite index does not Granger-cause variance, variance Granger-causes the composite index at lag 1, suggesting feedback effects may exist.

These findings underscore the complex, dynamic relationships in our data. While the appendices provide additional detail, the key insight is that the structural break and time series diagnostics reveal both regime shifts and persistent volatility shocks, enhancing our understanding of how the futures curve influences option-implied probabilities.

6 Discussion

The findings presented in this study offer several insights into the relationship between the WTI futures curve and option-implied probability distributions. And contribute to the broader literature on derivatives markets by underscoring how these instruments facilitate price discovery and risk management. The results indicate that the shape of the WTI futures curve is closely tied to shifts in the distribution of expected returns. Nevertheless, this relationship may be more complex than anticipated.

6.1 SABR Calibration and Market Dynamics

The calibration of the SABR model across varying market regimes highlights its versatility in capturing the volatility dynamics of WTI options. Around February 5, 2025, the observed shift in the β parameter from values around 0.2–0.4 to higher levels of approximately 0.7–0.9 aligns with the observations of Yuan and Kwon (2023), who noted that commodity options often exhibit regime-switching behavior. This notable increase in β suggests a fundamental change in how market participants were assessing tail risks, moving toward more lognormal-like dynamics. By extending the work of Moni (2014), who focused primarily on the volatility surface, this research demonstrates that changes in the futures curve are not only reflected in higher moments, such as variance, but also in the broader distribution of returns implied by options prices.

6.2 PDF Moments and Term Structure

A strong relationship, exceeding an R^2 of 0.80, between the shape of the futures curve and the variance/kurtosis of option-implied probability distributions provides further evidence that the term structure contains significant information about market uncertainty. These results build upon the work of Luo et al. (2024), who examined how futures spreads could predict realized volatility in crude oil markets. The structural break observed in February 2025 offers an especially revealing example of how rapidly market expectations can shift. Before and after this break, the skewness coefficients change dramatically (from approximately -6.7129 to -2.9047), indicating that market participants significantly revised their tail risk assessments. This shift also underscores that the informational content of the futures curve may vary across different market regimes and raises the possibility that the relationship between physical probabilities and their risk-neutral counterparts could be time-varying. These conclusions align with Bianchi and Piana (2017), who reported similarly dynamic behavior in risk premia across commodity futures markets.

6.3 Time Series Properties

Patterns in the time series data support the idea that WTI options reflect complex market forces. Significant autocorrelation in variance regression residuals and borderline ARCH effects confirm volatility clustering in commodity markets. Granger causality tests suggest variance leads to changes in the futures curve, indicating implied volatility may signal shifting market expectations. This adds to the debate over whether futures or options markets better gauge near-term supply and demand shifts.

6.4 Limitations and Future Research

Although these insights deepen our understanding of WTI options and futures markets, several limitations warrant consideration. The relatively short sample period of 40 observations may constrain the generalizability of the results. Furthermore, during the data collection phase, the futures curve was backwardated (although the index could capture the slight shift in the curve). This limits any inference on the results as they may not extend to periods where the curve is in contango. Additionally, the composite index relies on somewhat heuristic weightings that could be refined. Future work could involve extending the analysis to longer time series and additional commodities, refining index weighting schemes to capture more nuanced market signals, and assessing how well option-implied probability distributions predict actual price movements.

Overall, this discussion highlights the complex interplay between the futures curve and the option-implied PDF, emphasizing the importance of integrating multiple sources of market information. By recognizing the structural shifts and time-varying relationships revealed in this study, researchers and practitioners can develop more tools for analyzing and forecasting oil price behavior, ultimately contributing to more efficient and effective risk management strategies.

7 Conclusion

This thesis provides evidence of a strong yet complex relationship between the WTI futures curve and option-implied probability distributions. Our findings suggest that the futures curve contains significant information about market expectations and risk assessments, though this relationship can shift dramatically during periods of market stress.

The identification of a market-perceived change in tail risk and volatility in February 2025 highlights the dynamic nature of these relationships and underscores the importance of considering regime changes when analyzing derivatives markets. The strong connection between curve shape and higher moments of the implied PDF (particularly variance and kurtosis) suggests that market participants incorporate term structure information when pricing options.

This project's methodological contribution lies in developing a comprehensive pipeline for extracting and analyzing option-implied PDFs and constructing a composite index that captures multiple aspects of the futures curve.

The results have important implications for risk management and trading strategies, suggesting that changes in the futures curve might presage shifts in market uncertainty and tail risk assessments. However, the observed regime-dependent nature of these relationships implies that practitioners should remain vigilant to potential structural breaks and adjust their models accordingly.

Finally, this work opens several avenues for future research, particularly in understanding the dynamics of derivatives markets and the construction of term structure indicators. Future studies could explore how these relationships behave in other commodity derivatives markets.

References

Damien Ackerer, Natasa Tagasovska, and Thibault Vatter. Deep smoothing of the implied volatility surface. *SSRN Electronic Journal*, 1 2019. doi: 10.2139/ssrn.3402942. URL https://doi.org/10.2139/ssrn.3402942.

Aristotle. *Aristotle's politics*. University of Chicago Press, 3 2013.

Francesc Aràndiga, Antonio Baeza, and Dionisio F. Yáñez. Monotone cubic spline interpolation for functions with a strong gradient. *Applied Numerical Mathematics*, 172:591–607, 11 2021. doi: 10.1016/j.apnum.2021.11.007. URL https://doi.org/10.1016/j.apnum.2021.11.007.

L. Bachelier. Théorie de la spéculation. *Annales Scientifiques de l École Normale Supérieure*, 17: 21–86, 1 1900. doi: 10.24033/asens.476. URL https://doi.org/10.24033/asens.476.

B. Bahra. Implied risk-neutral probability density functions from option prices. Working Paper 66, Bank of England, March 1997.

Daniele Bianchi and Jacopo Piana. Expected spot prices and the dynamics of commodity risk premia. *SSRN Electronic Journal*, 6 2017. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2785563.

Fischer Black. The pricing of commodity contracts. *Journal of Financial Economics*, 3(1-2): 167–179, 1 1976. doi: 10.1016/0304-405x(76)90024-6. URL https://doi.org/10.1016/0304-405x(76)90024-6.

Fischer Black and Myron Scholes. The pricing of options and corporate liabilities. *Journal of Political Economy*, 81(3):637–654, 5 1973. doi: 10.1086/260062. URL https://doi.org/10.1086/260062.

Douglas T. Breeden and Robert H. Litzenberger. Prices of State-Contingent claims implicit in option prices. *The Journal of Business*, 51(4):621, 1 1978. doi: 10.1086/296025. URL https://doi.org/10.1086/296025.

John C. Cox and Stephen A. Ross. The valuation of options for alternative stochastic processes. *Journal of Financial Economics*, 3(1):145–166, 1976. ISSN 0304-405X. doi: https://doi.org/10.1016/0304-405X(76)90023-4. URL https://www.sciencedirect.com/science/article/pii/0304405X76900234.

James F. Epperson. On the Runge Example. *American Mathematical Monthly*, 94(4):329, 4 1987. doi: 10.2307/2323093. URL https://doi.org/10.2307/2323093.

Bassam Fattouh. An anatomy of the crude oil pricing system. *The Oxford Institute for Energy Studies*, 1 2011. URL https://ora.ox.ac.uk/objects/uuid:8b957970-239c-4a4f-9cbe-21830381de16/download_file?safe_filename=WPM40.pdf&file_format=application%2Fpdf&type_of_work=Working+paper.

Fabien Le Floc'h and Gary Kennedy. Finite difference techniques for arbitrage-free SABR. *The Journal of Computational Finance*, 8 2016. doi: 10.21314/jcf.2016.320. URL https://doi.org/10.21314/jcf.2016.320.

Jacques Gabillon. The Term Structures of Oil Futures Prices. *Oxford Institute for Energy Studies*, 1 1991. URL https://www.oxfordenergy.org/wpcms/wp-content/uploads/2010/11/WPM17-TheTermStructureofOilFuturesPrices-JGabillon-1991.pdf.

Patrick Hagan, Deep Kumar, Andrew S. Lesniewski, and Diana E. Woodward. Managing Smile Risk. *Wilmott magazine*, 1:84–108, 2002.

Hidde Hogenhout. Predicting volatility using the term structure curves of commodity futures, 2024.

John. Hull. *Options, Futures, and other Derivatives, Global Edition*. Pearson, 6 2021.

Fearghal Kearney and Han Lin Shang. Uncovering predictability in the evolution of the WTI oil futures curve. *European Financial Management*, 26(1):238–257, 1 2019. doi: 10.1111/eufm.12212. URL https://doi.org/10.1111/eufm.12212.

Joanne E. Kennedy, Subhankar Mitra, and Duy Pham. On the Approximation of the SABR Model: A Probabilistic Approach. *Applied Mathematical Finance*, 19(6):553–586, 2 2012. doi: 10.1080/1350486x.2011.646523. URL https://doi.org/10.1080/1350486x.2011.646523.

Lutz Kilian, Michael D. Plante, and Alexander W. Richter. Geopolitical oil price risk and economic fluctuations. *Federal Reserve Bank of Dallas, Working Papers*, 2024(2403), 11 2024. doi: 10.24149/wp2403r1. URL https://doi.org/10.24149/wp2403r1.

Yiyi Le, Jing Wen, Yuchen Wu, Jia Liu, and Yuchen Zhu. Investigating factors influencing oil volatility: a GARCH-MIDAS model analysis. *Frontiers in Energy Research*, 12, 6 2024. doi: 10.3389/fenrg.2024.1392905. URL https://doi.org/10.3389/fenrg.2024.1392905.

Fabien Le Floc'h. An arbitrage-free interpolation of Class C 2 for option prices. *Journal of Derivatives*, 28(4), 2021.

Jiawen Luo, Tony Klein, Thomas Walther, and Qiang Ji. Forecasting realized volatility of crude oil futures prices based on machine learning. *Journal of Forecasting*, 43(5):1422–1446, 2 2024. doi: 10.1002/for.3077. URL https://doi.org/10.1002/for.3077.

Robert Andrew Martin. Option-implied probability distributions, part 2 · Reasonable Deviations, 10 2020. URL https://reasonabledeviations.com/2020/10/10/option-implied-pdfs-2/.

Claudio Moni. Risk Managing Smile Risk with SABR Model. *SSRN Electronic Journal*, 1 2014. doi: 10.2139/ssrn.2527480. URL https://doi.org/10.2139/ssrn.2527480.

Sheldon Natenberg. *Option Volatility and Pricing: Advanced Trading Strategies and Techniques, 2nd edition*. McGraw Hill Professional, 11 2014.

Jacinto Marabel Romo. Dynamics of the implied volatility surface. Theory and empirical evidence. *Quantitative Finance*, 14(10):1829–1837, 6 2012. doi: 10.1080/14697688.2012.686668. URL https://doi.org/10.1080/14697688.2012.686668.

I. J. Schoenberg. Contributions to the problem of approximation of equidistant data by analytic functions. Part B. On the problem of osculatory interpolation. A second class of analytic approximation formulae. *Quarterly of Applied Mathematics*, 4(2):112–141, 7 1946. doi: 10.1090/qam/16705. URL https://doi.org/10.1090/qam/16705.

Eduardo S. Schwartz. The Stochastic Behavior of Commodity Prices: Implications for valuation and hedging. *The Journal of Finance*, 52(3):923–973, 7 1997. doi: 10.1111/j.1540-6261.1997.tb02721.x. URL https://doi.org/10.1111/j.1540-6261.1997.tb02721.x.

Vesa Soini and Sindre Lorentzen. Option prices and implied volatility in the crude oil market. *Energy Economics*, 83:515–539, 7 2019. doi: 10.1016/j.eneco.2019.07.011. URL https://doi.org/10.1016/j.eneco.2019.07.011.

Graeme West. Calibration of the SABR model in illiquid markets. *Applied Mathematical Finance*, 12(4):371–385, 12 2005. doi: 10.1080/13504860500148672. URL https://doi.org/10.1080/13504860500148672.

Zhushun Yuan and Roy H. Kwon. A lognormal/normal regime-switching commodity pricing model. *International Journal of Financial Engineering*, 6 2023. doi: 10.1142/s2424786323500147. URL https://doi.org/10.1142/s2424786323500147.

A Supplementary Figures and Tables

A.1 SABR Multi-β Comparison Tables

Table 4: *SABR Multi-Beta Comparison example, picked* $\beta = 0.3$*, source: own.*

Beta	Alpha	Rho	Nu	MSE	RMSE	Converged
0.05	16.5981	-0.6665	3.4265	3.289905e-04	1.813810e-02	True
0.10	13.4124	-0.6719	3.6098	3.354918e-04	1.831643e-02	True
0.20	8.7494	-0.6824	4.0394	3.486214e-04	1.867141e-02	True
0.25	7.0618	-0.6876	4.2932	3.552964e-04	1.884931e-02	True
0.30	**5.6961**	**-0.6927**	**4.5796**	**3.620782e-04**	**1.902835e-02**	**True**
0.35	4.5906	-0.6977	4.9054	3.689932e-04	1.920919e-02	True
0.40	3.6956	-0.7026	5.2790	3.760672e-04	1.939245e-02	True
0.50	2.3832	-0.7123	6.2178	3.907869e-04	1.976833e-02	True
0.55	1.9065	-0.7169	6.8176	3.984680e-04	1.996166e-02	True
0.60	1.5191	-0.7215	7.5382	4.063713e-04	2.015865e-02	True
0.65	1.2036	-0.7260	8.4188	4.144825e-04	2.035884e-02	True
0.70	0.9454	-0.7303	9.5170	4.227618e-04	2.056117e-02	True
0.80	0.5563	-0.7387	12.7876	4.394960e-04	2.096416e-02	True
0.90	0.2776	-0.7467	19.4591	4.557798e-04	2.134900e-02	True

Table 5: *SABR Multi-Beta Comparison example, picked* $\beta = 0.8$*, source: own*

Beta	Alpha	Rho	Nu	MSE	RMSE	Converged
0.05	9.8072	-0.0920	2.8599	6.984572e-04	2.642834e-02	True
0.10	7.9030	-0.0968	2.9942	6.944160e-04	2.635177e-02	True
0.20	5.1227	-0.1061	3.3046	6.839805e-04	2.615302e-02	True
0.25	4.1193	-0.1106	3.4854	6.775974e-04	2.603070e-02	True
0.30	3.3088	-0.1151	3.6872	6.704484e-04	2.589302e-02	True
0.35	2.6541	-0.1194	3.9141	6.625449e-04	2.573995e-02	True
0.40	2.1252	-0.1236	4.1710	6.539003e-04	2.557147e-02	True
0.50	1.3524	-0.1318	4.8023	6.344552e-04	2.518839e-02	True
0.55	1.0731	-0.1358	5.1962	6.236948e-04	2.497388e-02	True
0.60	0.8471	-0.1396	5.6615	6.122736e-04	2.474416e-02	True
0.65	0.6639	-0.1434	6.2199	6.002186e-04	2.449936e-02	True
0.70	0.5152	-0.1470	6.9040	5.875594e-04	2.423963e-02	True
0.80	**0.2945**	**-0.1541**	**8.8934**	**5.605567e-04**	**2.367608e-02**	**True**
0.90	0.1423	-0.1607	12.8933	5.315169e-04	2.305465e-02	True

A.2 SABR Fits of IV

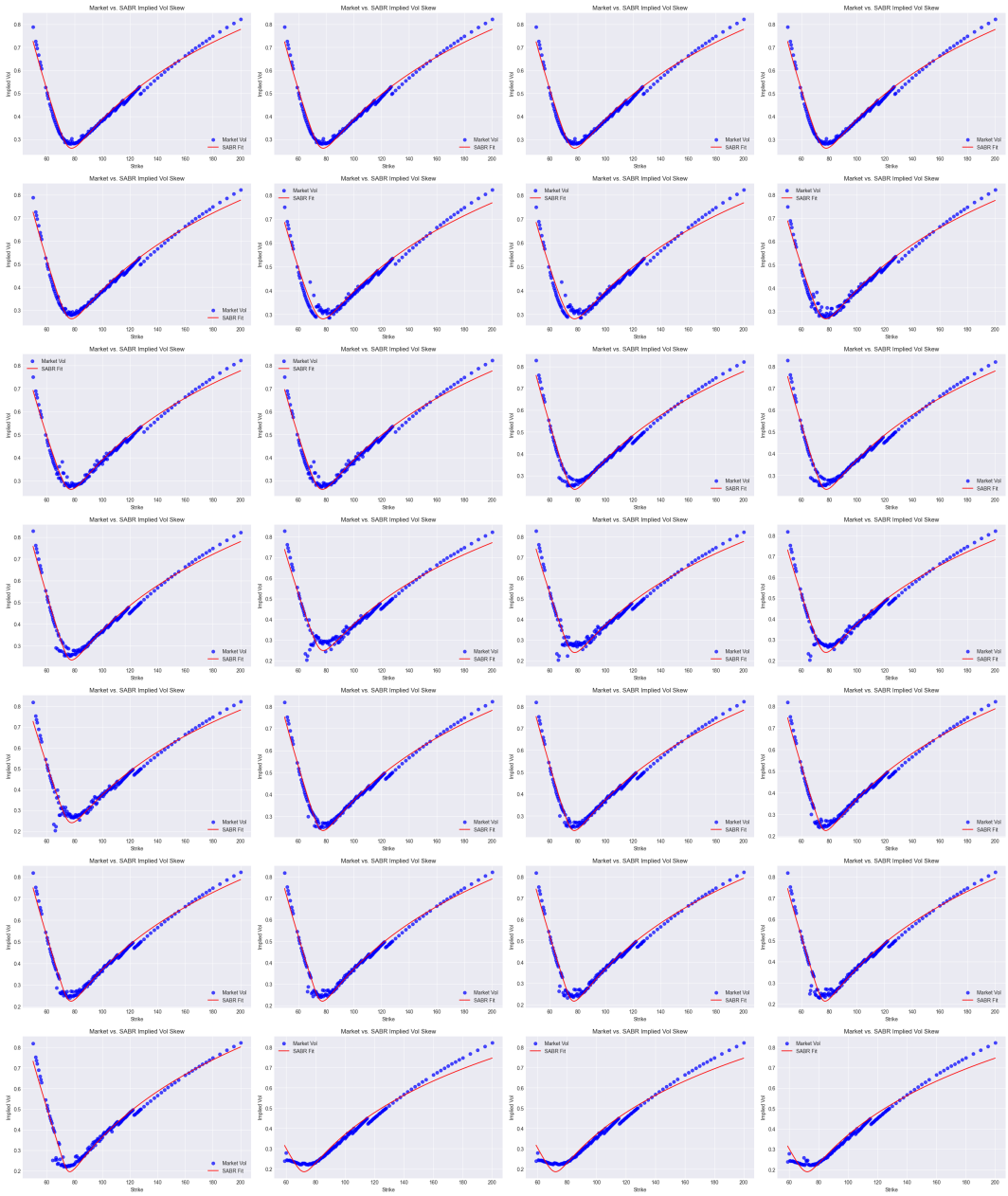

Figure 6: *Mosaic of SABR fits to Market Volatilities. Source: Own*

A.3 SABR Fits of IV Errors

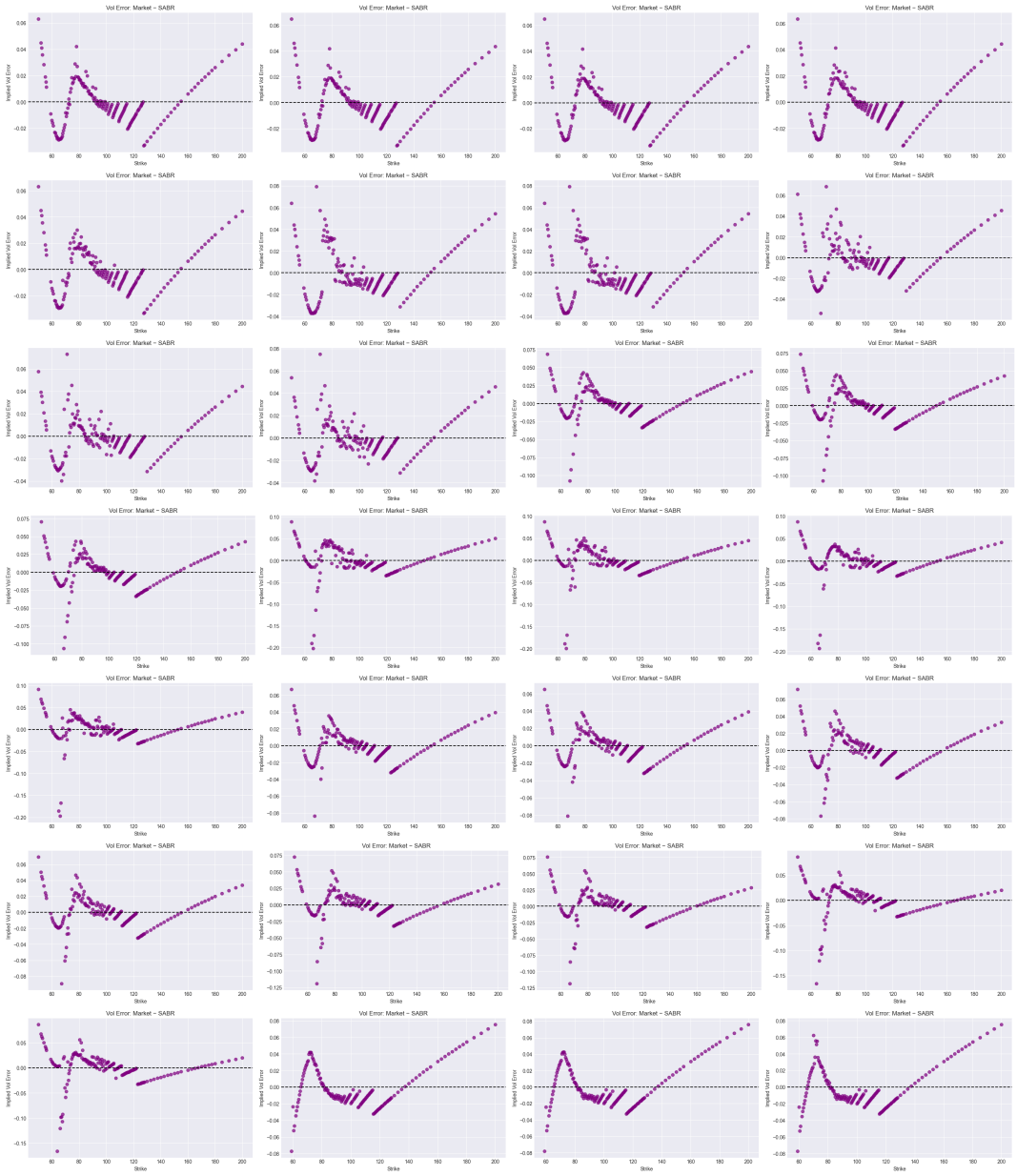

Figure 7: *Mosaic of SABR fit errors. Source: Own*

A.4 SABR PDFs

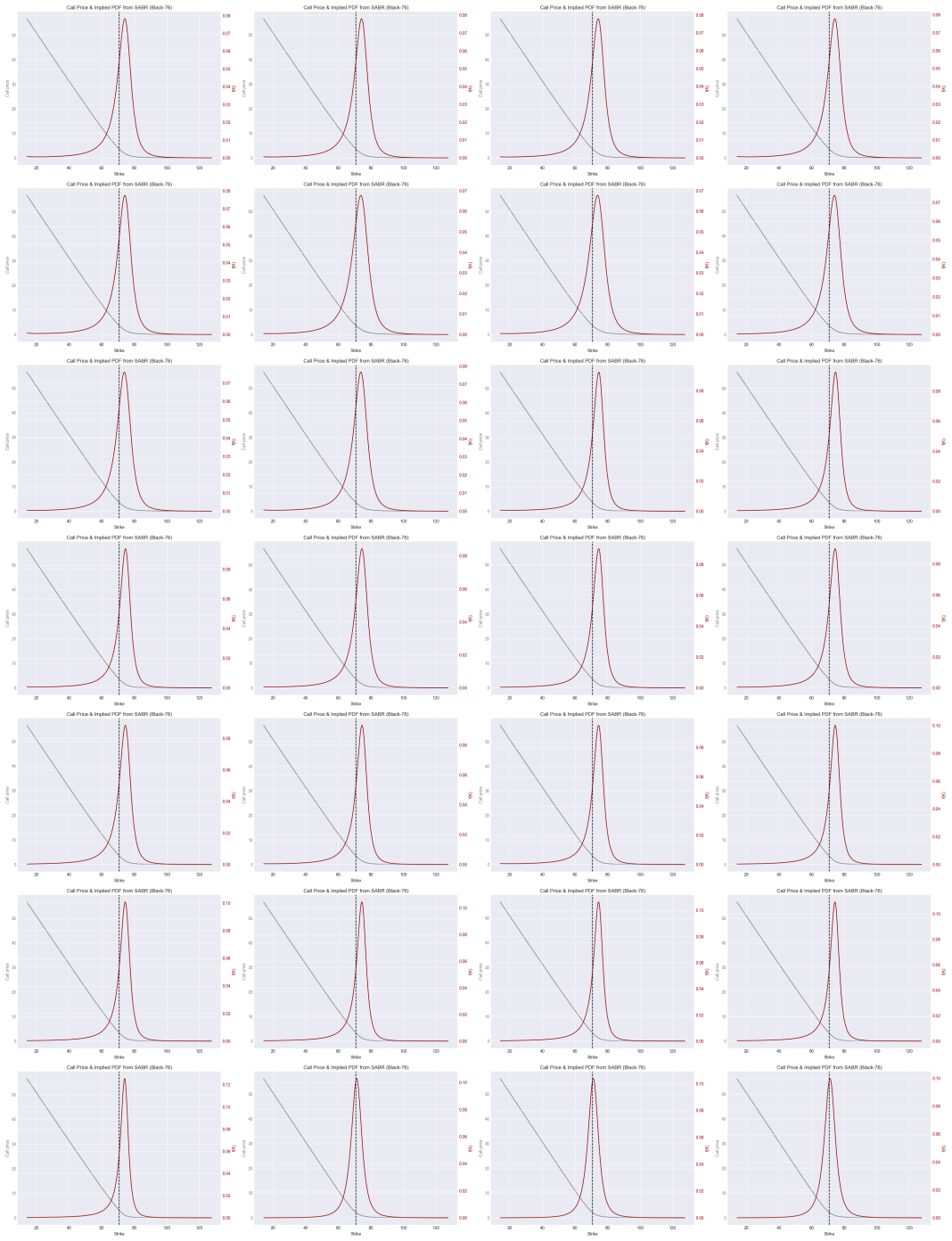

Figure 8: *Mosaic of SABR generated Risk–Neutral Densities. Source: Own*

A.5 Scatter Plots of OLS Regressions

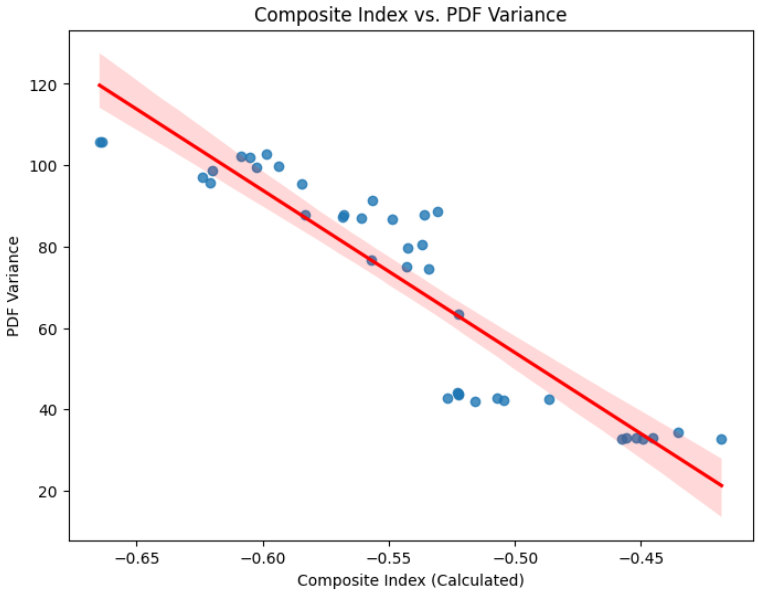

Figure 9: *OLS Regression of Variance as dep. variable. Source: Own*

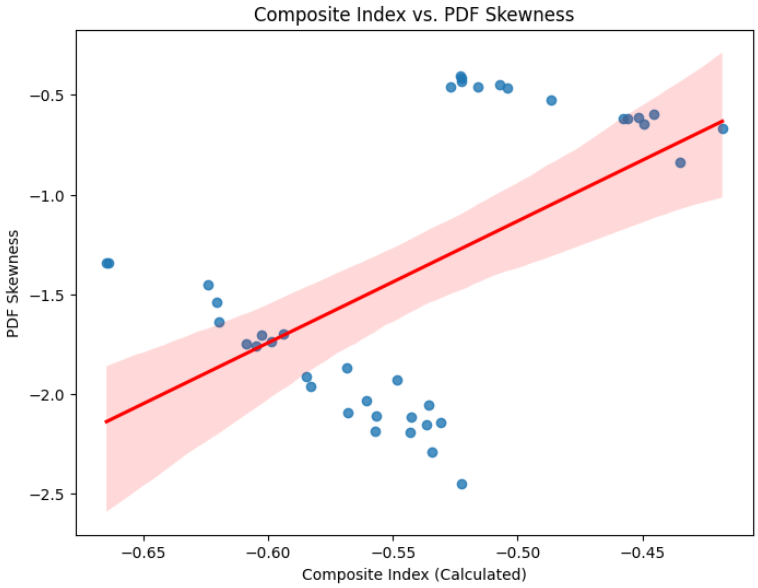

Figure 10: *OLS Regression of Skewness as dep. variable. Source: Own*

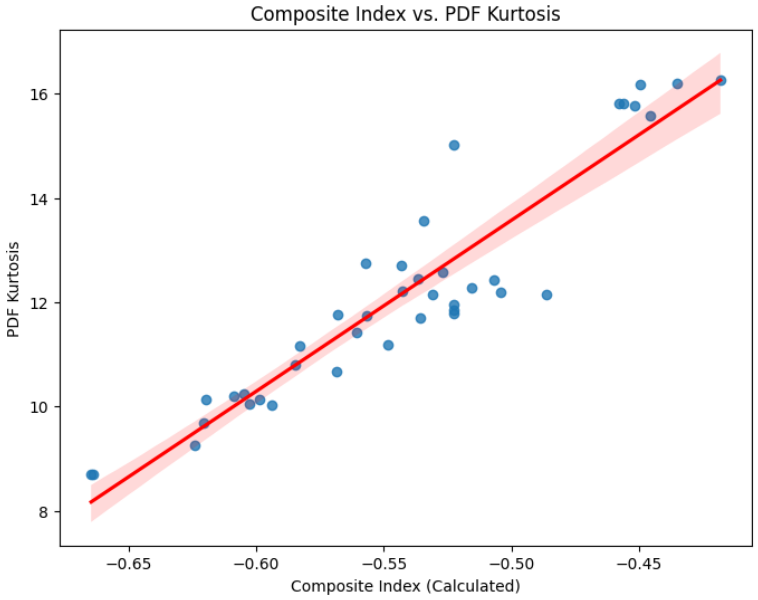

Figure 11: *OLS Regression of Kurtosis as dep. variable. Source: Own*

B Extended Derivations and Results

B.1 2.1 Extended Derivation

We start from the definition of PDF:

$$\mathrm{P}(A \leq x \leq B) = \int_A^B f(x)\,dx \tag{B.1}$$

We can approximate the probability of that the underlying price ends up inside our range ΔK can be written as

$$P(K_0 \leq K \leq K_0 + \Delta K) \approx f(K_0)\Delta K \tag{B.2}$$

In the limit as ΔK tends to 0, the approximation above must become an equality, and in the limit $P(K_0 \leq K \leq K_0 + \Delta K)$ becomes p_K. This can be expressed as follows

$$p_K = \lim_{\Delta K \to 0} \{f(K)\Delta K\} \implies f(K) = \lim_{\Delta K \to 0} \frac{p_K}{\Delta K} \tag{B.3}$$

Recall equation 2.2 from Chapter 2. This gives us a relation between p_K and the price of the butterfly spread. Substituting 2.2 in B.3,

$$f(K) = \lim_{\Delta K \to 0} \left\{ \frac{X}{\Delta K \Delta K} \right\} = \lim_{\Delta K \to 0} \left\{ \frac{X}{(\Delta K)^2} \right\} \tag{B.4}$$

Lastly recall that by the law of one price two equivalent products must have the same price thus, the price of our butterfly X must be equal to the sum of the prices of the calls involved in constructing the spread. Thus,

$$X = C(K + \Delta K) - 2C(K) + C(K - \Delta K) \tag{B.5}$$

Substituting this relationship into B.4 we obtain the following expression,

$$f(K) = \lim_{\Delta K \to 0} \frac{C(K + \Delta K) - 2C(K) + C(K - \Delta K)}{(\Delta K)^2} \tag{B.6}$$

Equation B.6 yields and interesting result as by definition this is the second partial derivative with respect to K of C. That is:

$$f(K) = \frac{\partial^2 C(K)}{\partial K^2} \tag{B.7}$$

B.2 Breeden-Litzenberger Derivation

The price of a European call option with maturity T and strike price K is given by the risk-neutral valuation formula:

$$C(K,T) = e^{-rT}\mathbb{E}^{Q}\left[\max(S_T - K, 0)\right], \tag{B.8}$$

Expanding the expectation:

$$C(K,T) = e^{-rT}\int_{K}^{\infty}(S_T - K) f_Q(S_T)\, dS_T, \tag{B.9}$$

where $f_Q(S_T)$ is the risk-neutral PDF of S_T.
The integrand can be split as follows:

$$C(K,T) = e^{-rT}\left[\int_{K}^{\infty} S_T f_Q(S_T)\, dS_T - \int_{K}^{\infty} K f_Q(S_T)\, dS_T\right]. \tag{B.10}$$

The second term simplifies:

$$\int_{K}^{\infty} K f_Q(S_T)\, dS_T = K\int_{K}^{\infty} f_Q(S_T)\, dS_T = KQ(S_T > K), \tag{B.11}$$

where $Q(S_T > K)$ is the risk-neutral probability that $S_T > K$. Thus,

$$C(K,T) = e^{-rT}\left[\int_{K}^{\infty} S_T f_Q(S_T)\, dS_T - KQ(S_T > K)\right]. \tag{B.12}$$

Differentiating $C(K,T)$ with respect to K:

$$\frac{\partial C(K,T)}{\partial K} = e^{-rT}\left[-Q(S_T > K) + K\frac{\partial}{\partial K}\int_{K}^{\infty} f_Q(S_T)\, dS_T\right]. \tag{B.13}$$

The derivative of the integral $\int_{K}^{\infty} f_Q(S_T)\, dS_T$ is $-f_Q(K)$ by the Fundamental Theorem of Calculus, so:

$$\frac{\partial C(K,T)}{\partial K} = -e^{-rT}Q(S_T > K). \tag{B.14}$$

Differentiating again with respect to K:

$$\frac{\partial^2 C(K,T)}{\partial K^2} = -\frac{\partial}{\partial K}\left(e^{-rT}Q(S_T > K)\right). \tag{B.15}$$

Since $Q(S_T > K) = \int_{K}^{\infty} f_Q(S_T)\, dS_T$, its derivative is $-f_Q(K)$, giving:

$$\frac{\partial^2 C(K,T)}{\partial K^2} = e^{-rT} f_Q(K). \tag{B.16}$$

Rearranging, we obtain the risk-neutral PDF $f_Q(S_T)$:

$$f_Q(K) = e^{rT} \frac{\partial^2 C(K,T)}{\partial K^2}. \tag{B.17}$$

B.3 Time Series tests

As commented in 5.4 several tests where conducted to better understand this time series. Here we a relevant summary of the results of the tests:

- **Stationarity and Cointegration:** Augmented Dickey–Fuller tests indicate that the composite index and PDF moment series are (or can be made) stationary. Cointegration tests confirm that the observed relationships are not spurious.
- **Residual Autocorrelation:** The Ljung–Box test on the variance regression residuals yielded a significant statistic ($p < 0.00001$), suggesting that the model's residuals are autocorrelated. This indicates that additional dynamic components might be incorporated in future models.
- **ARCH Effects:** An ARCH test on the variance regression residuals produced an LM statistic of 14.29 ($p \approx 0.075$), which is borderline at the 5% significance level. This hints at potential volatility clustering.
- **Granger Causality:** Granger causality tests showed that the composite index does not Granger–cause variance (with p-values consistently above 0.74), whereas variance appears to Granger–cause the composite index at lag 1 ($p \approx 0.045$). These results suggest possible feedback effects, with variance influencing the observed term structure.

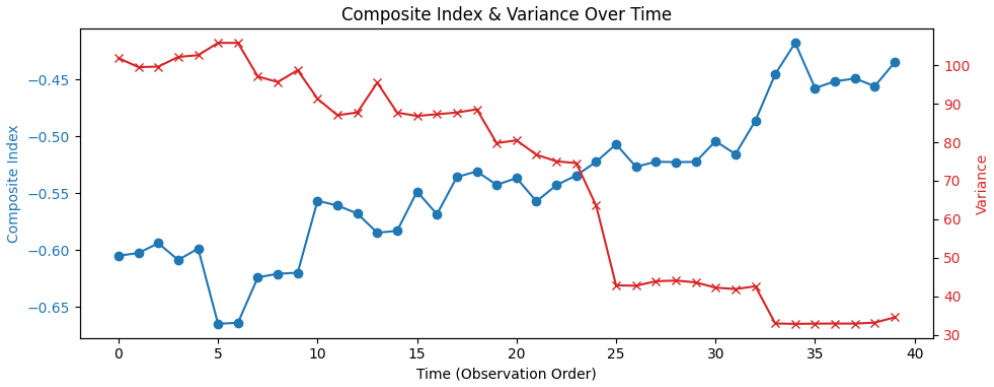

Figure 12: *Time series of PDF variance. Source: own.*

C Code Snippets

C.1 Full Code

All the code used in this thesis—including the options and futures branches, data cleaning procedures, butterfly spread PDFs, SABR multi-beta comparisons, SABR PDFs, and the relevant statistical testing—can be found in the following GitHub repository.

Link to the GitHub repository: `https://github.com/gorkbravo/Undergraduate-Thesis`

C.2 Code Snippets

C.2.1 Multi-β Comparison Code

```python
import numpy as np
import pandas as pd
import math
from scipy.optimize import least_squares


# 1) Extended Hagan SABR

def sabr_hagan_black_vol_extended(alpha, beta, rho, nu, F, K, T):
    if np.isclose(F, K, atol=1e-14):
        leading = alpha / (F**(1.0 - beta))
        A1 = ((1-beta)**2 / 24.0) * (alpha**2 / (F**(2*(1-beta))))
        A2 = 0.25 * rho * beta * nu * alpha / (F**(1.0-beta))
        A3 = ((2.0 - 3.0*rho**2)/24.0) * (nu**2)
        correction = 1.0 + (A1 + A2 + A3)*T
        return max(leading*correction, 1e-8)

    z = (nu/alpha)*(F**(1.0 - beta) - K**(1.0 - beta))
    eps = 1e-14
    numerator = np.sqrt(1.0 - 2.0*rho*z + z*z) + z - rho
    denominator = (1.0 - rho)
    x_z = np.log((numerator + eps)/(denominator + eps))

    denom = (F*K)**((1.0 - beta)/2.0)
    fk_1minusbeta = (F*K)**(1.0 - beta)
    A1 = ((1.0-beta)**2 / 24.0) * (alpha**2 / fk_1minusbeta)
    A2 = 0.25 * rho * beta * nu * alpha / ((F*K)**((1.0-beta)/2.0))
    A3 = ((2.0 - 3.0*rho**2)/24.0) * (nu**2)
    correction = 1.0 + (A1 + A2 + A3)*T

    sabr_vol = (alpha/denom)*(z/x_z)*correction
    return max(sabr_vol, 1e-8)


```

```python
# 2) SABR Objective

def sabr_objective_extended(params, beta, F, T, strikes, market_vols):
    alpha, rho, nu = params
    model_vols = [
        sabr_hagan_black_vol_extended(alpha, beta, rho, nu, F, K, T)
        for K in strikes
    ]
    return np.array(market_vols) - np.array(model_vols)


# 3) Main Loop Over Beta

def main(csv_path, F, T):
    beta_values = [0.05, 0.10, 0.15 ,0.20, 0.25, 0.30, 0.35, 0.40, 0.50, 0.55, 0.60, 0.65,
    ↪  0.70, 0.80, 0.90]

    # Bounds & initial guess
    lower_bounds = [0.0, -1.0, 0.0]
    upper_bounds = [np.inf, 1.0, 1000.0]
    x0 = [0.5, 0.0, 0.5]

    # Load Data
    df = pd.read_csv(csv_path)
    df = df[df["type"] == "call"].copy()  # Only calls
    strikes = df["strike"].values
    market_vols = df["impliedVolatility"].values

    # Store results
    results = []

    # Loop over Beta values
    for beta in beta_values:
        fun = lambda p: sabr_objective_extended(
            p, beta, F, T, strikes, market_vols
        )

        # Perform calibration
        res = least_squares(
            fun,
            x0,
            bounds=(lower_bounds, upper_bounds),
            method='trf',
            ftol=1e-12,
            xtol=1e-12,
            gtol=1e-12
        )

        alpha_calib, rho_calib, nu_calib = res.x
```

```
        # Compute MSE & RMSE
        sabr_vols = [
            sabr_hagan_black_vol_extended(alpha_calib, beta, rho_calib, nu_calib, F, K, T)
            for K in strikes
        ]
        errors = np.array(market_vols) - np.array(sabr_vols)
        mse = np.mean(errors**2)
        rmse = math.sqrt(mse)

        # Save results
        results.append({
            "Beta": beta,
            "Alpha": alpha_calib,
            "Rho": rho_calib,
            "Nu": nu_calib,
            "MSE": mse,
            "RMSE": rmse,
            "Converged": res.success
        })

    # Print Summary
    print("============= SABR Multi-Beta Comparison (Unweighted) =============")
    print(f"{'Beta':>5}  {'Alpha':>9}  {'Rho':>9}  {'Nu':>9}  {'MSE':>12}  {'RMSE':>12}
    ↪  Converged")
    for r in results:
        print(f"{r['Beta']:5.2f}  {r['Alpha']:9.4f}  {r['Rho']:9.4f}  {r['Nu']:9.4f}"
              f"  {r['MSE']:12.6e}  {r['RMSE']:12.6e}  {r['Converged']}")

if __name__ == "__main__":
    # Example for standalone testing
    main(
        csv_path="C:/Users/User/Desktop/UPF/TGF/Data/SPY_opt_1mo_cleaned.csv",
        F=607,
        T=4.0 / 52.0
    )
```

C.2.2 SABR & PDF Generation

```
import os
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from scipy.optimize import least_squares
from math import sqrt
from scipy.stats import norm
from datetime import datetime
```

```python

# Extended SABR volatility using Hagan's formula (abbreviated)
def sabr_hagan_black_vol_extended(alpha, beta, rho, nu, F, K, T):
    if np.isclose(F, K, atol=1e-14):
        # ATM case: return leading term with correction (details omitted)
        return max(alpha / (F**(1-beta)) * (1 + ... * T), 1e-8)
    # Off-ATM: compute z and x(z) for corrections
    z = (nu / alpha) * (F**(1-beta) - K**(1-beta))
    x_z = np.log((np.sqrt(1 - 2*rho*z + z*z) + z - rho) / (1 - rho))
    return max((alpha / ((F*K)**((1-beta)/2))) * (z/x_z) * (1 + ... * T), 1e-8)

# Calibration objective function
def sabr_objective_extended(x, beta, F, T, strikes, market_vols):
    alpha, rho, nu = x
    model_vols = [sabr_hagan_black_vol_extended(alpha, beta, rho, nu, F, K, T)
                  for K in strikes]
    return np.array(market_vols) - np.array(model_vols)

# Black-76 call price formula
def black_76_call_price(F, K, r, T, sigma):
    if T <= 0:
        return max(F - K, 0.0)
    d1 = (np.log(F/K) + 0.5 * sigma**2 * T) / (sigma * sqrt(T))
    d2 = d1 - sigma * sqrt(T)
    return np.exp(-r*T) * (F * norm.cdf(d1) - K * norm.cdf(d2))

# Main function (abbreviated details for plotting and PDF stats)
def main(csv_path, F, T, beta):
    # Load and filter call option data
    df = pd.read_csv(csv_path)
    df = df[df["type"]=="call"]
    strikes = df["strike"].values
    market_vols = df["impliedVolatility"].values
    r = 0.04163  % Risk-free rate

    # Calibrate SABR parameters
    x0 = [0.5, 0.0, 0.5]
    res = least_squares(sabr_objective_extended, x0, args=(beta, F, T, strikes, market_vols))
    alpha_calib, rho_calib, nu_calib = res.x
    print("Calibration complete:", res.success)

    # Compute errors and display basic calibration output
    sabr_vols = [sabr_hagan_black_vol_extended(alpha_calib, beta, rho_calib, nu_calib, F, K,
    ↪  T)
                 for K in strikes]
    errors = np.array(market_vols) - np.array(sabr_vols)
    print(f"MSE: {np.mean(errors**2):.6f}, RMSE: {np.sqrt(np.mean(errors**2)):.6f}")

    # Additional plotting and PDF calculations are abbreviated...
```

```
    # e.g., plt.plot(...), compute PDF, statistics, etc.

if __name__ == "__main__":
    main("data.csv", 71.98, 4.0/52.0, 0.25)
```

D Pipeline Overview

D.1 Overview not including statistical analysis

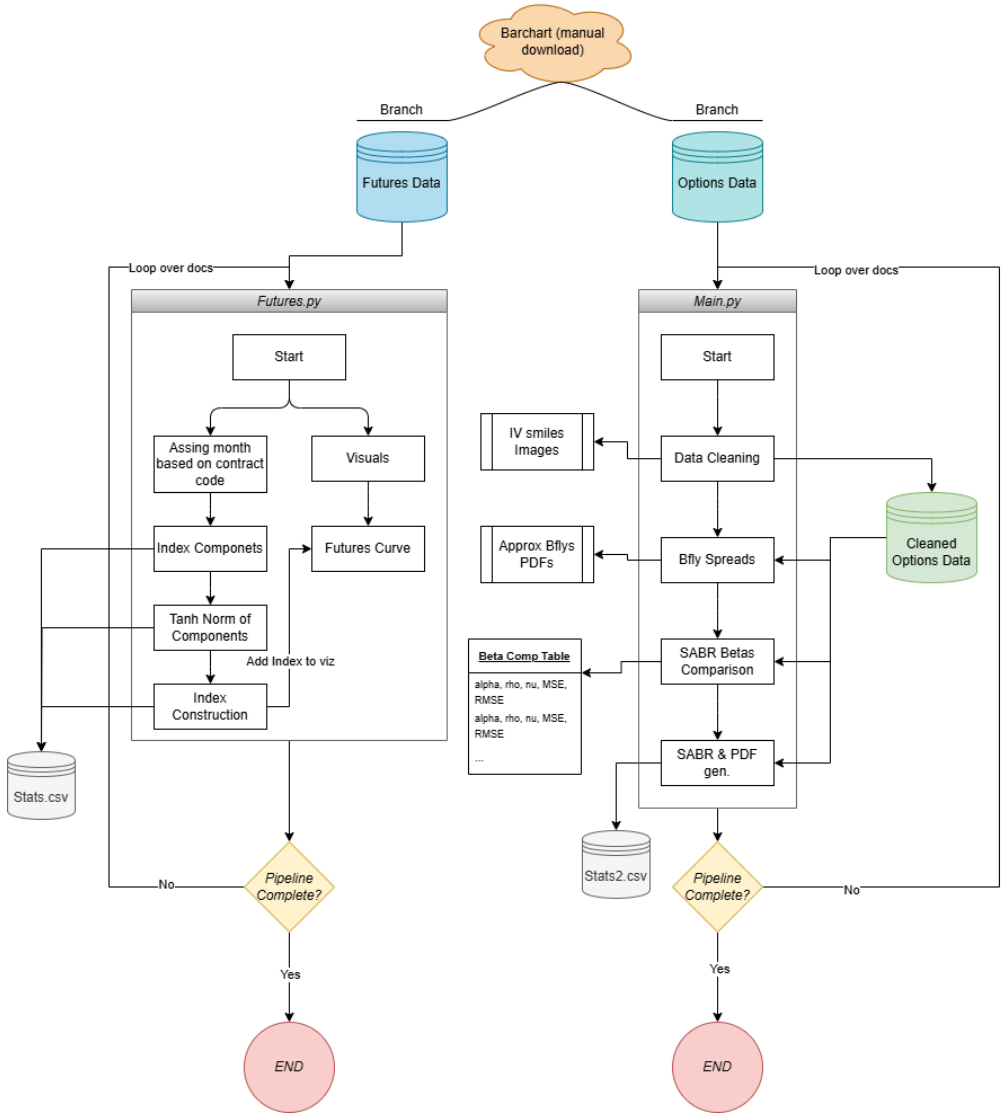

Figure 13: *Overview of the two main branches of the pipeline, source: own*

E Additional Regressions

Table 6: *OLS Regression Results for Mean (Current): Regression of* `mean` *on* `composite_calculated`.

Dep. Variable:	mean	**R-squared:**	0.270
Model:	OLS	**Adj. R-squared:**	0.251
Method:	Least Squares	**F-statistic:**	19.67
Date:	Wed, 12 Feb 2025	**Prob (F-statistic):**	7.62e-05
Time:	15:36:12	**Log-Likelihood:**	-11.424
No. Observations:	40	**AIC:**	26.85
Df Residuals:	38	**BIC:**	30.23
Df Model:	1		
Covariance Type:	HC3		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	69.5616	0.388	179.349	0.000	68.801	70.322
composite_calculated	-3.2174	0.725	-4.435	0.000	-4.639	-1.795

Omnibus:	14.834	**Durbin-Watson:**	0.383
Prob(Omnibus):	0.001	**Jarque-Bera (JB):**	3.713
Skew:	0.336	**Prob(JB):**	0.156
Kurtosis:	1.667	**Cond. No.:**	21.3

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)

Table 7: *OLS Regression Results for Variance (Current): Regression of* `variance` *on* `composite_calculated`.

Dep. Variable:	variance	**R-squared:**	0.825
Model:	OLS	**Adj. R-squared:**	0.820
Method:	Least Squares	**F-statistic:**	246.8
Date:	Wed, 12 Feb 2025	**Prob (F-statistic):**	3.29e-18
Time:	15:36:12	**Log-Likelihood:**	-153.37
No. Observations:	40	**AIC:**	310.7
Df Residuals:	38	**BIC:**	314.1
Df Model:	1		
Covariance Type:	HC3		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	-145.4364	13.922	-10.447	0.000	-172.723	-118.150
composite_calculated	-398.7826	25.382	-15.711	0.000	-448.531	-349.034

Omnibus:	1.706	**Durbin-Watson:**	0.574
Prob(Omnibus):	0.426	**Jarque-Bera (JB):**	1.418
Skew:	-0.294	**Prob(JB):**	0.492
Kurtosis:	2.289	**Cond. No.:**	21.3

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)

Table 8: *OLS Regression Results for Skewness (Current): Regression of* `skewness` *on* `composite_calculated`.

Dep. Variable:	skewness	**R-squared:**	0.285
Model:	OLS	**Adj. R-squared:**	0.266
Method:	Least Squares	**F-statistic:**	23.34
Date:	Wed, 12 Feb 2025	**Prob (F-statistic):**	2.25e-05
Time:	15:36:12	**Log-Likelihood:**	-35.611
No. Observations:	40	**AIC:**	75.22
Df Residuals:	38	**BIC:**	78.60
Df Model:	1		
Covariance Type:	HC3		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	1.9214	0.685	2.807	0.005	0.580	3.263
composite_calculated	6.1087	1.265	4.831	0.000	3.630	8.587

Omnibus:	6.289	**Durbin-Watson:**	0.437
Prob(Omnibus):	0.043	**Jarque-Bera (JB):**	2.185
Skew:	-0.116	**Prob(JB):**	0.335
Kurtosis:	1.879	**Cond. No.:**	21.3

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)

Table 9: *OLS Regression Results for Kurtosis (Current): Regression of* `kurtosis` *on* `composite_calculated`.

Dep. Variable:	kurtosis	**R-squared:**	0.874
Model:	OLS	**Adj. R-squared:**	0.871
Method:	Least Squares	**F-statistic:**	378.4
Date:	Wed, 12 Feb 2025	**Prob (F-statistic):**	2.37e-21
Time:	15:36:12	**Log-Likelihood:**	-45.695
No. Observations:	40	**AIC:**	95.39
Df Residuals:	38	**BIC:**	98.77
Df Model:	1		
Covariance Type:	HC3		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	29.9798	0.985	30.444	0.000	28.050	31.910
composite_calculated	32.8132	1.687	19.453	0.000	29.507	36.119

Omnibus:	1.639	**Durbin-Watson:**	1.099
Prob(Omnibus):	0.441	**Jarque-Bera (JB):**	0.726
Skew:	0.137	**Prob(JB):**	0.695
Kurtosis:	3.601	**Cond. No.:**	21.3

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)

Table 10: *OLS Regression Results for Mean vs.* `norm_month1`*: Regression of* `mean` *on* `norm_month1`.

Dep. Variable:	mean	**R-squared:**	0.354
Model:	OLS	**Adj. R-squared:**	0.337
Method:	Least Squares	**F-statistic:**	35.27
Date:	Wed, 12 Feb 2025	**Prob (F-statistic):**	6.90e-07
Time:	15:36:12	**Log-Likelihood:**	-8.9741
No. Observations:	40	**AIC:**	21.95
Df Residuals:	38	**BIC:**	25.33
Df Model:	1		
Covariance Type:	HC3		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	69.5091	0.297	234.115	0.000	68.927	70.091
norm_month1	-2.2206	0.374	-5.939	0.000	-2.953	-1.488

Omnibus:	9.848	**Durbin-Watson:**	0.400
Prob(Omnibus):	0.007	**Jarque-Bera (JB):**	3.446
Skew:	0.391	**Prob(JB):**	0.179
Kurtosis:	1.793	**Cond. No.:**	16.4

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)

Table 11: *OLS Regression Results for Variance vs. norm_month1: Regression of variance on norm_month1.*

Dep. Variable:	variance	**R-squared:**	0.876
Model:	OLS	**Adj. R-squared:**	0.873
Method:	Least Squares	**F-statistic:**	244.5
Date:	Wed, 12 Feb 2025	**Prob (F-statistic):**	3.86e-18
Time:	15:36:12	**Log-Likelihood:**	-146.41
No. Observations:	40	**AIC:**	296.8
Df Residuals:	38	**BIC:**	300.2
Df Model:	1		
Covariance Type:	HC3		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	-129.6880	13.586	-9.546	0.000	-156.316	-103.060
norm_month1	-247.7448	15.845	-15.635	0.000	-278.801	-216.688

Omnibus:	0.453	**Durbin-Watson:**	0.760
Prob(Omnibus):	0.797	**Jarque-Bera (JB):**	0.588
Skew:	-0.078	**Prob(JB):**	0.745
Kurtosis:	2.426	**Cond. No.:**	16.4

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)

Table 12: *OLS Regression Results for Skewness vs. norm_month1: Regression of skewness on norm_month1.*

Dep. Variable:	skewness	**R-squared:**	0.367
Model:	OLS	**Adj. R-squared:**	0.351
Method:	Least Squares	**F-statistic:**	39.90
Date:	Wed, 12 Feb 2025	**Prob (F-statistic):**	2.10e-07
Time:	15:36:12	**Log-Likelihood:**	-33.157
No. Observations:	40	**AIC:**	70.31
Df Residuals:	38	**BIC:**	73.69
Df Model:	1		
Covariance Type:	HC3		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	1.4447	0.589	2.454	0.014	0.291	2.598
norm_month1	4.1817	0.662	6.317	0.000	2.884	5.479

Omnibus:	3.855	**Durbin-Watson:**	0.472
Prob(Omnibus):	0.146	**Jarque-Bera (JB):**	1.883
Skew:	-0.208	**Prob(JB):**	0.390
Kurtosis:	2.022	**Cond. No.:**	16.4

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)

Table 13: *OLS Regression Results for Kurtosis vs. norm_month1: Regression of kurtosis on norm_month1.*

Dep. Variable:	kurtosis	**R-squared:**	0.846
Model:	OLS	**Adj. R-squared:**	0.842
Method:	Least Squares	**F-statistic:**	283.1
Date:	Wed, 12 Feb 2025	**Prob (F-statistic):**	3.35e-19
Time:	15:36:12	**Log-Likelihood:**	-49.661
No. Observations:	40	**AIC:**	103.3
Df Residuals:	38	**BIC:**	100.3
Df Model:	1		
Covariance Type:	HC3		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	27.9360	1.023	27.309	0.000	25.931	29.941
norm_month1	19.4615	1.157	16.826	0.000	17.194	21.728

Omnibus:	0.647	**Durbin-Watson:**	0.917
Prob(Omnibus):	0.724	**Jarque-Bera (JB):**	0.103
Skew:	-0.021	**Prob(JB):**	0.950
Kurtosis:	3.245	**Cond. No.:**	16.4

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)

Table 14: *OLS Regression Results for Mean vs. norm_month3: Regression of mean on norm_month3.*

Dep. Variable:	mean	**R-squared:**	0.289
Model:	OLS	**Adj. R-squared:**	0.270
Method:	Least Squares	**F-statistic:**	19.55
Date:	Wed, 12 Feb 2025	**Prob (F-statistic):**	7.95e-05
Time:	15:36:12	**Log-Likelihood:**	-10.913
No. Observations:	40	**AIC:**	25.83
Df Residuals:	38	**BIC:**	29.20
Df Model:	1		
Covariance Type:	HC3		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	69.8122	0.331	210.859	0.000	69.163	70.461
norm_month3	-2.1730	0.492	-4.421	0.000	-3.136	-1.210

Omnibus:	12.144	**Durbin-Watson:**	0.369
Prob(Omnibus):	0.002	**Jarque-Bera (JB):**	3.498
Skew:	0.343	**Prob(JB):**	0.174
Kurtosis:	1.724	**Cond. No.:**	15.8

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)

Table 15: *OLS Regression Results for Variance vs. norm_month3: Regression of variance on norm_month3.*

Dep. Variable:	variance	**R-squared:**	0.846
Model:	OLS	**Adj. R-squared:**	0.842
Method:	Least Squares	**F-statistic:**	264.0
Date:	Wed, 12 Feb 2025	**Prob (F-statistic):**	1.08e-18
Time:	15:36:12	**Log-Likelihood:**	-150.82
No. Observations:	40	**AIC:**	305.6
Df Residuals:	38	**BIC:**	309.0
Df Model:	1		
Covariance Type:	HC3		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	-110.6299	11.061	-10.002	0.000	-132.308	-88.952
norm_month3	-263.8958	16.242	-16.248	0.000	-295.729	-232.062

Omnibus:	2.051	**Durbin-Watson:**	0.545
Prob(Omnibus):	0.359	**Jarque-Bera (JB):**	1.794
Skew:	-0.399	**Prob(JB):**	0.408
Kurtosis:	2.336	**Cond. No.:**	15.8

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)

Table 16: *OLS Regression Results for Skewness vs. norm_month3: Regression of skewness on norm_month3.*

Dep. Variable:	skewness	**R-squared:**	0.304
Model:	OLS	**Adj. R-squared:**	0.286
Method:	Least Squares	**F-statistic:**	22.92
Date:	Wed, 12 Feb 2025	**Prob (F-statistic):**	2.57e-05
Time:	15:36:12	**Log-Likelihood:**	-35.066
No. Observations:	40	**AIC:**	74.13
Df Residuals:	38	**BIC:**	77.51
Df Model:	1		
Covariance Type:	HC3		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	1.4447	0.589	2.454	0.014	0.291	2.598
norm_month3	4.1247	0.862	4.787	0.000	2.436	5.813

Omnibus:	5.269	**Durbin-Watson:**	0.422
Prob(Omnibus):	0.072	**Jarque-Bera (JB):**	2.033
Skew:	-0.130	**Prob(JB):**	0.362
Kurtosis:	1.927	**Cond. No.:**	15.8

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)

Table 17: *OLS Regression Results for Kurtosis vs. norm_month3: Regression of kurtosis on norm_month3.*

Dep. Variable:	kurtosis	**R-squared:**	0.869
Model:	OLS	**Adj. R-squared:**	0.866
Method:	Least Squares	**F-statistic:**	339.3
Date:	Wed, 12 Feb 2025	**Prob (F-statistic):**	1.54e-20
Time:	15:36:12	**Log-Likelihood:**	-46.438
No. Observations:	40	**AIC:**	96.88
Df Residuals:	38	**BIC:**	100.3
Df Model:	1		
Covariance Type:	HC3		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	26.8889	0.874	30.778	0.000	25.177	28.601
norm_month3	21.3845	1.161	18.421	0.000	19.109	23.660

Omnibus:	2.169	**Durbin-Watson:**	1.131
Prob(Omnibus):	0.338	**Jarque-Bera (JB):**	1.235
Skew:	-0.009	**Prob(JB):**	0.539
Kurtosis:	3.861	**Cond. No.:**	15.8

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)

Table 18: *OLS Regression Results for Mean vs. norm_year1: Regression of mean on norm_year1.*

Dep. Variable:	mean	**R-squared:**	0.229
Model:	OLS	**Adj. R-squared:**	0.209
Method:	Least Squares	**F-statistic:**	13.15
Date:	Wed, 12 Feb 2025	**Prob (F-statistic):**	0.000840
Time:	15:36:12	**Log-Likelihood:**	-12.521
No. Observations:	40	**AIC:**	29.04
Df Residuals:	38	**BIC:**	32.42
Df Model:	1		
Covariance Type:	HC3		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	64.7736	1.800	35.981	0.000	61.245	68.302
norm_year1	-7.1827	1.980	-3.627	0.000	-11.064	-3.301

Omnibus:	7.627	**Durbin-Watson:**	0.493
Prob(Omnibus):	0.022	**Jarque-Bera (JB):**	2.954
Skew:	0.344	**Prob(JB):**	0.228
Kurtosis:	1.860	**Cond. No.:**	72.8

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)

Table 19: *OLS Regression Results for Variance vs. `norm_year1`: Regression of `variance` on `norm_year1`.*

Dep. Variable:	variance	**R-squared:**	0.659
Model:	OLS	**Adj. R-squared:**	0.650
Method:	Least Squares	**F-statistic:**	96.69
Date:	Wed, 12 Feb 2025	**Prob (F-statistic):**	5.44e-12
Time:	15:36:12	**Log-Likelihood:**	-166.68
No. Observations:	40	**AIC:**	337.4
Df Residuals:	38	**BIC:**	340.7
Df Model:	1		
Covariance Type:	HC3		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	-715.3522	79.691	-8.977	0.000	-871.544	-559.161
norm_year1	-864.3903	87.904	-9.833	0.000	-1036.679	-692.101

Omnibus:	6.187	**Durbin-Watson:**	0.739
Prob(Omnibus):	0.045	**Jarque-Bera (JB):**	2.203
Skew:	-0.141	**Prob(JB):**	0.332
Kurtosis:	1.886	**Cond. No.:**	72.8

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)

Table 20: *OLS Regression Results for Skewness vs. `norm_year1`: Regression of `skewness` on `norm_year1`.*

Dep. Variable:	skewness	**R-squared:**	0.222
Model:	OLS	**Adj. R-squared:**	0.201
Method:	Least Squares	**F-statistic:**	13.90
Date:	Wed, 12 Feb 2025	**Prob (F-statistic):**	0.000627
Time:	15:36:12	**Log-Likelihood:**	-37.296
No. Observations:	40	**AIC:**	78.59
Df Residuals:	38	**BIC:**	81.97
Df Model:	1		
Covariance Type:	HC3		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	10.5007	3.194	3.287	0.001	4.240	16.762
norm_year1	13.0753	3.507	3.728	0.000	6.201	19.949

Omnibus:	4.359	**Durbin-Watson:**	0.523
Prob(Omnibus):	0.113	**Jarque-Bera (JB):**	1.834
Skew:	-0.106	**Prob(JB):**	0.400
Kurtosis:	1.973	**Cond. No.:**	72.8

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)

Table 21: *OLS Regression Results for Kurtosis vs. norm_year1: Regression of kurtosis on norm_year1.*

Dep. Variable:	kurtosis	**R-squared:**	0.732
Model:	OLS	**Adj. R-squared:**	0.725
Method:	Least Squares	**F-statistic:**	86.41
Date:	Wed, 12 Feb 2025	**Prob (F-statistic):**	2.50e-11
Time:	15:36:12	**Log-Likelihood:**	-60.764
No. Observations:	40	**AIC:**	125.5
Df Residuals:	38	**BIC:**	128.9
Df Model:	1		
Covariance Type:	HC3		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	78.4281	7.170	10.938	0.000	64.375	92.482
norm_year1	72.8333	7.835	9.296	0.000	57.476	88.190

Omnibus:	4.945	**Durbin-Watson:**	0.983
Prob(Omnibus):	0.084	**Jarque-Bera (JB):**	1.898
Skew:	0.043	**Prob(JB):**	0.387
Kurtosis:	1.936	**Cond. No.:**	72.8

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)

Table 22: *OLS Regression Results for Variance (Current) [Repeated]: Regression of variance on composite_calculated.*

Dep. Variable:	variance	**R-squared:**	0.825
Model:	OLS	**Adj. R-squared:**	0.820
Method:	Least Squares	**F-statistic:**	246.8
Date:	Wed, 12 Feb 2025	**Prob (F-statistic):**	3.29e-18
Time:	15:36:12	**Log-Likelihood:**	-153.37
No. Observations:	40	**AIC:**	310.7
Df Residuals:	38	**BIC:**	314.1
Df Model:	1		
Covariance Type:	HC3		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	-145.4364	13.922	-10.447	0.000	-172.723	-118.150
composite_calculated	-398.7826	25.382	-15.711	0.000	-448.531	-349.034

Omnibus:	1.706	**Durbin-Watson:**	0.574
Prob(Omnibus):	0.426	**Jarque-Bera (JB):**	1.418
Skew:	-0.294	**Prob(JB):**	0.492
Kurtosis:	2.289	**Cond. No.:**	21.3

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)

Table 23: *OLS Regression Results for Variance (With Lag): Regression of* `variance` *on* `composite_calculated` *and its lag (*`composite_calculated_lag1`*).*

Dep. Variable:	variance	**R-squared:**	0.840
Model:	OLS	**Adj. R-squared:**	0.831
Method:	Least Squares	**F-statistic:**	134.1
Date:	Wed, 12 Feb 2025	**Prob (F-statistic):**	2.07e-17
Time:	15:36:12	**Log-Likelihood:**	-147.61
No. Observations:	39	**AIC:**	301.2
Df Residuals:	36	**BIC:**	306.2
Df Model:	2		
Covariance Type:	HC3		

	coef	**std err**	**z**	**P$>$ |z|**	**[0.025**	**0.975]**
const	-153.8069	14.439	-10.652	0.000	-182.107	-125.506
composite_calculated	-242.2048	62.935	-3.849	0.000	-365.554	-118.855
composite_calculated_lag1	-170.3960	69.774	-2.442	0.015	-307.151	-33.641

Omnibus:	2.855	**Durbin-Watson:**	0.331
Prob(Omnibus):	0.240	**Jarque-Bera (JB):**	1.883
Skew:	-0.324	**Prob(JB):**	0.390
Kurtosis:	2.141	**Cond. No.:**	78.7

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)

Table 24: *OLS Regression Results for Standard Deviation (Current): Regression of* `std_dev` *on* `composite_calculated`.

Dep. Variable:	std_dev	**R-squared:**	0.820
Model:	OLS	**Adj. R-squared:**	0.815
Method:	Least Squares	**F-statistic:**	217.3
Date:	Wed, 12 Feb 2025	**Prob (F-statistic):**	2.66e-17
Time:	15:36:12	**Log-Likelihood:**	-43.317
No. Observations:	40	**AIC:**	90.63
Df Residuals:	38	**BIC:**	94.01
Df Model:	1		
Covariance Type:	HC3		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	-5.3227	0.919	-5.793	0.000	-7.124	-3.522
composite_calculated	-25.0200	1.697	-14.741	0.000	-28.347	-21.693

Omnibus:	2.034	**Durbin-Watson:**	0.572
Prob(Omnibus):	0.362	**Jarque-Bera (JB):**	1.326
Skew:	-0.165	**Prob(JB):**	0.515
Kurtosis:	2.171	**Cond. No.:**	21.3

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)

Table 30: *OLS Regression Results for Skewness vs. Composite (First 24): Regression of* `skewness` *on* `composite_calculated` *for the first 24 observations.*

Dep. Variable:	skewness	**R-squared:**	0.905
Model:	OLS	**Adj. R-squared:**	0.901
Method:	Least Squares	**F-statistic:**	259.9
Date:	Wed, 12 Feb 2025	**Prob (F-statistic):**	1.14e-13
Time:	15:36:12	**Log-Likelihood:**	25.518
No. Observations:	24	**AIC:**	-47.04
Df Residuals:	22	**BIC:**	-44.68
Df Model:	1		
Covariance Type:	HC3		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	-5.7786	0.249	-23.165	0.000	-6.268	-5.290
composite_calculated	-6.7129	0.416	-16.122	0.000	-7.529	-5.897

Omnibus:	0.958	**Durbin-Watson:**	1.314
Prob(Omnibus):	0.619	**Jarque-Bera (JB):**	0.853
Skew:	0.218	**Prob(JB):**	0.653
Kurtosis:	2.186	**Cond. No.:**	34.7

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)

Table 25: *OLS Regression Results for Standard Deviation (With Lag): Regression of* `std_dev` *on* `composite_calculated` *and its lag (*`composite_calculated_lag1`*).*

Dep. Variable:	std_dev	**R-squared:**	0.836
Model:	OLS	**Adj. R-squared:**	0.827
Method:	Least Squares	**F-statistic:**	118.4
Date:	Wed, 12 Feb 2025	**Prob (F-statistic):**	1.48e-16
Time:	15:36:12	**Log-Likelihood:**	-40.214
No. Observations:	39	**AIC:**	86.43
Df Residuals:	36	**BIC:**	91.42
Df Model:	2		
Covariance Type:	HC3		

	coef	**std err**	**z**	**P$>\|$z$\|$**	**[0.025**	**0.975]**
const	-5.9019	0.940	-6.277	0.000	-7.745	-4.059
composite_calculated	-14.7242	4.251	-3.464	0.001	-23.055	-6.393
composite_calculated_lag1	-11.2629	4.631	-2.432	0.015	-20.339	-2.187

Omnibus:	3.131	**Durbin-Watson:**	0.328
Prob(Omnibus):	0.209	**Jarque-Bera (JB):**	1.640
Skew:	-0.175	**Prob(JB):**	0.440
Kurtosis:	2.059	**Cond. No.:**	78.7

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)

Table 26: *OLS Regression Results for Log Variance (Current): Regression of* `log_var` *on* `composite_calculated`.

Dep. Variable:	log_var	**R-squared:**	0.813
Model:	OLS	**Adj. R-squared:**	0.808
Method:	Least Squares	**F-statistic:**	183.9
Date:	Wed, 12 Feb 2025	**Prob (F-statistic):**	3.86e-16
Time:	15:36:12	**Log-Likelihood:**	10.189
No. Observations:	40	**AIC:**	-16.38
Df Residuals:	38	**BIC:**	-13.00
Df Model:	1		
Covariance Type:	HC3		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	0.6880	0.254	2.706	0.007	0.190	1.186
composite_calculated	-6.4283	0.474	-13.561	0.000	-7.357	-5.499

Omnibus:	2.896	**Durbin-Watson:**	0.568
Prob(Omnibus):	0.235	**Jarque-Bera (JB):**	1.459
Skew:	-0.031	**Prob(JB):**	0.482
Kurtosis:	2.066	**Cond. No.:**	21.3

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)

Table 31: *OLS Regression Results for Skewness vs. Composite (Remaining): Regression of* `skewness` *on* `composite_calculated` *for the remaining observations.*

Dep. Variable:	skewness	**R-squared:**	0.824
Model:	OLS	**Adj. R-squared:**	0.809
Method:	Least Squares	**F-statistic:**	27.86
Date:	Wed, 12 Feb 2025	**Prob (F-statistic):**	0.000195
Time:	15:36:12	**Log-Likelihood:**	22.065
No. Observations:	14	**AIC:**	-40.13
Df Residuals:	12	**BIC:**	-38.85
Df Model:	1		
Covariance Type:	HC3		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	-1.9474	0.278	-7.015	0.000	-2.491	-1.403
composite_calculated	-2.9047	0.550	-5.278	0.000	-3.983	-1.826

Omnibus:	18.467	**Durbin-Watson:**	1.117
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	16.541
Skew:	-1.847	**Prob(JB):**	0.000256
Kurtosis:	6.835	**Cond. No.:**	33.1

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)

Table 27: *OLS Regression Results for Log Variance (With Lag): Regression of* `log_var` *on* `composite_calculated` *and its lag (*`composite_calculated_lag1`*).*

Dep. Variable:	log_var	**R-squared:**	0.832
Model:	OLS	**Adj. R-squared:**	0.822
Method:	Least Squares	**F-statistic:**	99.51
Date:	Wed, 12 Feb 2025	**Prob (F-statistic):**	2.15e-15
Time:	15:36:12	**Log-Likelihood:**	12.018
No. Observations:	39	**AIC:**	-18.04
Df Residuals:	36	**BIC:**	-13.05
Df Model:	2		
Covariance Type:	HC3		

	coef	**std err**	**z**	**P$>\|$z$\|$**	**[0.025**	**0.975]**
const	0.5265	0.258	2.042	0.041	0.021	1.032
composite_calculated	-3.6782	1.196	-3.075	0.002	-6.022	-1.334
composite_calculated_lag1	-3.0222	1.275	-2.370	0.018	-5.521	-0.523

Omnibus:	3.639	**Durbin-Watson:**	0.329
Prob(Omnibus):	0.162	**Jarque-Bera (JB):**	1.619
Skew:	-0.016	**Prob(JB):**	0.445
Kurtosis:	2.002	**Cond. No.:**	78.7

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)

Table 28: *OLS Regression Results for Log Standard Deviation (Current): Regression of* `log_std_dev` *on* `composite_calculated`.

Dep. Variable:	log_std_dev	**R-squared:**	0.813
Model:	OLS	**Adj. R-squared:**	0.808
Method:	Least Squares	**F-statistic:**	183.9
Date:	Wed, 12 Feb 2025	**Prob (F-statistic):**	3.86e-16
Time:	15:36:12	**Log-Likelihood:**	37.915
No. Observations:	40	**AIC:**	-71.83
Df Residuals:	38	**BIC:**	-68.45
Df Model:	1		
Covariance Type:	HC3		

	coef	**std err**	**z**	**P$>\|$z$\|$**	**[0.025**	**0.975]**
const	0.3440	0.127	2.706	0.007	0.095	0.593
composite_calculated	-3.2142	0.237	-13.561	0.000	-3.679	-2.750

Omnibus:	2.896	**Durbin-Watson:**	0.568
Prob(Omnibus):	0.235	**Jarque-Bera (JB):**	1.459
Skew:	-0.031	**Prob(JB):**	0.482
Kurtosis:	2.066	**Cond. No.:**	21.3

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)

Table 29: *OLS Regression Results for Log Standard Deviation (With Lag): Regression of* `log_std_dev` *on* `composite_calculated` *and its lag (*`composite_calculated_lag1`*).*

Dep. Variable:	log_std_dev	**R-squared:**	0.832
Model:	OLS	**Adj. R-squared:**	0.822
Method:	Least Squares	**F-statistic:**	99.51
Date:	Wed, 12 Feb 2025	**Prob (F-statistic):**	2.15e-15
Time:	15:36:12	**Log-Likelihood:**	39.051
No. Observations:	39	**AIC:**	-72.10
Df Residuals:	36	**BIC:**	-67.11
Df Model:	2		
Covariance Type:	HC3		

	coef	**std err**	**z**	**P$>\|$z$\|$**	**[0.025**	**0.975]**
const	0.2633	0.129	2.042	0.041	0.011	0.516
composite_calculated	-1.8391	0.598	-3.075	0.002	-3.011	-0.667
composite_calculated_lag1	-1.5111	0.638	-2.370	0.018	-2.761	-0.262

Omnibus:	3.639	**Durbin-Watson:**	0.329
Prob(Omnibus):	0.162	**Jarque-Bera (JB):**	1.619
Skew:	-0.016	**Prob(JB):**	0.445
Kurtosis:	2.002	**Cond. No.:**	78.7

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)

E.1 GARCH Results

Table 32: *GARCH(1,1) Model Results on Variance: Constant Mean - GARCH Model Results*

Dep. Variable:	variance	R-squared:	0.000
Mean Model:	Constant Mean	Adj. R-squared:	0.000
Vol Model:	GARCH	Log-Likelihood:	-170.275
Distribution:	Normal	AIC:	348.551
Method:	Maximum Likelihood	BIC:	355.306
		No. Observations:	40
Date:	Sun, Feb 16 2025	Df Residuals:	39
Time:	16:16:41	Df Model:	1

Table 33: *Mean Model*

	coef	std err	t	P>—t—	95.0% Conf. Int.
mu	87.9659	1.617	54.393	0.000	[84.796, 91.136]

Table 34: *Volatility Model*

	coef	std err	t	P>—t—	95.0% Conf. Int.
omega	16.6124	10.556	1.574	0.116	[-4.077, 37.301]
alpha[1]	1.0000	0.378	2.648	8.091e-03	[0.260, 1.740]
beta[1]	3.1268e-17	0.367	8.529e-17	1.000	[-0.719, 0.719]

Notes: [1] Covariance estimator: robust

Table 35: *GARCH-X Model Results (Variance with Composite Index as exogenous variable): Constant Mean - GARCH Model Results*

Dep. Variable:	variance	R-squared:	0.000
Mean Model:	Constant Mean	Adj. R-squared:	0.000
Vol Model:	GARCH	Log-Likelihood:	-170.275
Distribution:	Normal	AIC:	348.551
Method:	Maximum Likelihood	BIC:	355.306
		No. Observations:	40
Date:	Sun, Feb 16 2025	Df Residuals:	39
Time:	16:16:42	Df Model:	1

Table 36: *Mean Model*

	coef	std err	t	P>—t—	95.0% Conf. Int.
mu	87.9659	1.617	54.393	0.000	[84.796, 91.136]

Table 37: *Volatility Model*

	coef	std err	t	P>—t—	95.0% Conf. Int.
omega	16.6124	10.556	1.574	0.116	[-4.077, 37.301]
alpha[1]	1.0000	0.378	2.648	8.091e-03	[0.260, 1.740]
beta[1]	3.1268e-17	0.367	8.529e-17	1.000	[-0.719, 0.719]

Notes: [1] Covariance estimator: robust