UNIVERSITY OF STRATHCLYDE

DEPARTMENT OF MATHEMATICS AND STATISTICS

FORECASTING TAIL RISK WITH LONG-MEMORY: A HETEROGENEOUS EXTENSION OF THE CAESAR MODEL

by

Charles Shaw

202381359

MSc Applied Statistics

2025/26

Statement of work in project

The work contained in this project is that of the author and where material from other sources has been incorporated full acknowledgement is made.

Signed: ....

Print Name: CHARLES SHAW

Date: 2025

Supervised by : Dr Jiazhu Pan

Document Control

Version	Date	Description	Author
0.1	22 Nov 2025	Initial draft	Charles Shaw
0.2	27 Nov 2025	Fixed notation	Charles Shaw
0.3	07 Dec 2025	add a clarifying note for the ES formula	Charles Shaw
0.4	07 Dec 2025	theory, identification, ULLN, rolling window, GAS diagnostics	Charles Shaw

Contents

Abstract

The transition to Expected Shortfall (ES) as the primary market risk measure under Basel III has created a need for robust forecasting models that can jointly estimate Value at Risk (VaR) and ES. This thesis evaluates the recently proposed CAESar framework, a semi-parametric approach that exploits the joint elicitability of these measures. Furthermore, we propose and implement HAR-CAESar, a novel extension that incorporates heterogeneous autoregressive volatility components to capture long-memory effects. Through a nuanced empirical analysis of four major equity indices over a 22-year period, we demonstrate that the CAESar framework offers superior stability compared to likelihood-based Generalised Autoregressive Score (GAS) models, which exhibit critical optimisation failures in this setting. The proposed HAR-CAESar extension yields a distinctive performance profile: while it performs comparably to the parsimonious baseline at the standard 97.5% regulatory confidence level, it functions effectively as a "crisis specialist" at the extreme 99% level. Statistical tests show no significant difference in average performance, but case study analysis reveals that HAR-CAESar provides significantly larger capital buffers during systemic shocks. We conclude that for risk managers, the HAR extension represents a quantifiable insurance premium: accepting negligible estimation noise in calm periods to secure superior protection during extreme tail events, such as market crashes or asset bubble bursts.

1 Introduction

1.1 Motivation and Context

Financial institutions hold portfolios whose value fluctuates with market prices. A bank, pension fund, or insurance company must quantify the potential losses it faces over a given horizon so that it can set aside sufficient capital to remain solvent under adverse conditions. Regulators impose minimum capital requirements precisely to ensure that institutions can absorb losses without transmitting distress to the broader financial system. The measurement of market risk lies at the heart of modern prudential regulation and internal risk management alike.

Two risk measures dominate practice. Value at Risk (VaR) reports the loss threshold that a portfolio exceeds with a specified small probability over a chosen horizon. Expected Shortfall (ES), sometimes called Conditional Value at Risk, reports the average loss conditional on that threshold being breached. Together these measures anchor the market-risk framework of Basel III, which requires banks to compute both VaR and ES at the 97.5% confidence level for capital determination [1]. Understanding how to forecast VaR and ES accurately is therefore of direct practical and regulatory importance.

1.2 Problem Statement and Classical Limitations

Let r_t denote the return on a portfolio at time t, and let $\mathscr{F}_{t-1}$ denote the information available at the end of day $t-1$. For a probability level $\theta \in (0,1)$, typically chosen to be small (for instance $\theta = 0.025$), the conditional VaR is the θ-quantile of the return distribution:

$$\mathrm{VaR}_t(\theta) = F_t^{-1}(\theta), \tag{1}$$

where F_t is the conditional distribution function of r_t given $\mathscr{F}_{t-1}$. The conditional ES is the expected return in the lower tail:

$$\mathrm{ES}_t(\theta) = \mathbb{E}[r_t \mid r_t \leq \mathrm{VaR}_t(\theta), \mathscr{F}_{t-1}]. \tag{2}$$

Throughout this thesis we work with returns rather than losses, so that negative values represent adverse outcomes. Under this convention, when we restrict attention to the left tail where returns are negative, both

VaR and ES take negative values. Because ES averages over the worst outcomes in the tail, it is more negative than VaR: $\mathrm{ES}_t(\theta) \leq \mathrm{VaR}_t(\theta) < 0$.

VaR gained widespread acceptance because it reduces portfolio risk to a single number that managers and boards can interpret. Yet VaR suffers from a fundamental defect: it ignores the severity of losses beyond the threshold. Two portfolios can share the same VaR yet have very different tail behaviour, and a risk manager who focuses only on VaR may unwittingly accumulate exposures that generate catastrophic losses in extreme scenarios. VaR also fails to satisfy sub-additivity, one of the axioms that [2] argued a coherent risk measure should possess. ES satisfies all the coherence axioms and is now the preferred regulatory measure [3].

The shift from VaR to ES as the primary regulatory risk measure creates a new statistical challenge. VaR is elicitable, meaning it can be estimated by minimising a loss function (the quantile loss) that uniquely identifies the true quantile. ES, however, is not elicitable on its own [4]. The resolution came from recognising that VaR and ES are jointly elicitable: strictly consistent loss functions exist for the pair even though none exists for ES alone [5]. This joint elicitability enables regression-based models that estimate both measures simultaneously.

The forecasting task is further complicated by the stylised facts of asset returns. Daily returns exhibit heavy tails, meaning that extreme losses occur more frequently than a Gaussian model would predict. Returns also display volatility clustering: large price movements tend to be followed by further large movements, so the conditional variance changes through time. Any forecasting model must therefore capture both the unconditional tail behaviour and the dynamic evolution of the conditional distribution.

1.3 Modern Dynamic Approaches

Classical approaches to VaR and ES estimation include historical simulation, which uses the empirical quantile of past returns, and parametric methods, which assume a distributional form such as Gaussian or Student-*t* and estimate its parameters. These static methods fail to capture volatility clustering and can react slowly to changing market conditions. A natural improvement is to embed VaR and ES within a conditional volatility framework. GARCH models [6] specify autoregressive dynamics for the conditional variance and, when combined with an assumed innovation distribution, yield time-varying quantile and tail-mean forecasts.

A more direct approach models the quantile itself without specifying the full distribution. [7] introduced

Conditional Autoregressive Value at Risk by Regression Quantiles (CAViaR), which specifies an autoregressive law of motion for VaR and estimates it by minimising the quantile loss. CAViaR avoids distributional assumptions on returns and can capture asymmetric responses to positive and negative shocks. Extending the idea to ES, [8] proposed dynamic semiparametric models that exploit the joint elicitability of VaR and ES to estimate both measures via strictly consistent loss functions. Their Generalised Autoregressive Score (GAS) framework provides flexible dynamics for tail risk without requiring a fully specified return distribution.

1.4 CAESar and the Proposed HAR Extension

Building on CAViaR and the joint-elicitability insight, [9] recently proposed Conditional Autoregressive Expected Shortfall (CAESar). The model specifies autoregressive dynamics for both VaR and ES, with the ES equation incorporating lagged ES, lagged VaR, and past return information. Estimation proceeds in three stages: first VaR is estimated via CAViaR regression, then ES is modelled as an autoregressive function of lagged returns and lagged VaR, and finally VaR and ES are jointly re-estimated by minimising a strictly consistent loss function subject to a monotonicity constraint that prevents the quantile from crossing below the tail mean. CAESar makes no parametric assumption on the return distribution and handles heteroskedastic effects flexibly.

Simulation experiments and empirical applications in the original paper demonstrate that CAESar outperforms several existing methods, including GAS-based models and the approach of [10], in terms of out-of-sample forecasting accuracy. Nonetheless, the model is recent and its Markovian structure—conditioning only on the previous day's information—may fail to capture the long-memory volatility cascades characteristic of financial markets.

In this thesis, we address this limitation by proposing **HAR-CAESar**. By integrating the Heterogeneous Autoregressive (HAR) logic of [11], this novel specification conditions risk forecasts not just on yesterday's data, but on weekly and monthly aggregated returns. This allows the model to capture the persistence of volatility shocks across multiple time horizons, potentially improving forecast accuracy in the extreme tail where long-memory effects dominate.

1.5 Objectives and Research Questions

The primary objective of this thesis is to implement, extend, and rigorously evaluate the CAESar model for joint VaR and ES forecasting. We seek to answer three interrelated questions. How does CAESar perform relative to classical VaR methods and to dynamic models such as GARCH-based approaches and GAS specifications across a range of financial return series? Can the proposed HAR-CAESar specification, incorporating heterogeneous volatility structure, improve forecasting accuracy or robustness? And do established and recent backtesting procedures confirm the out-of-sample reliability of CAESar forecasts for both VaR and ES?

The methodological contribution of this thesis is twofold. We provide the derivation and implementation of the HAR-CAESar specification, which extends the original model to account for multi-horizon volatility dynamics. We also conduct a nuanced empirical comparison using backtesting procedures that go beyond simple VaR violation counts.

Box 1.1: Core Research Questions

RQ1: Comparative Performance. How does CAESar perform relative to classical VaR methods and to dynamic models such as GARCH and GAS?

RQ2: Extended Specification. Can an extended CAESar specification, incorporating additional structure or alternative dynamics, improve forecasting accuracy or robustness?

RQ3: Reliability. Do established and recent backtesting procedures confirm the out-of-sample reliability of CAESar forecasts for both VaR and ES?

Figure 1: The research agenda of this thesis.

1.6 Overview of Methods and Data

The models considered in this thesis span classical, volatility-based, and regression-based approaches. We include historical simulation and parametric VaR as simple baselines, and implement representative GARCH and GAS specifications with heavy-tailed innovation distributions to capture volatility dynamics. On the regression side, we estimate CAViaR for VaR and use the GAS framework of [8] for joint VaR and ES. The CAESar model of [9] serves as our primary focus, and we develop an extended specification that modifies the original dynamics. All models are estimated at multiple probability levels, including the regulatory 2.5%

and 1% tails. The empirical analysis uses daily returns from major equity indices over a sample spanning several years, providing sufficient observations to evaluate tail risk forecasts reliably.

1.7 Structure of the Thesis

The remainder of this thesis proceeds as follows. Section 2 surveys the literature on VaR and ES, covering the axiomatic foundations of coherent risk measures, the elicitability results that enable regression-based estimation, and the development of dynamic models from GARCH through CAViaR and GAS to CAESar. Section 3 describes the data, presents each model in detail, defines the loss functions and backtesting procedures, and explains the estimation strategy. Section 4 reports the empirical findings, comparing forecasting accuracy and backtest performance across models, assets, and probability levels. Section 5 interprets the results, assesses the contribution of the extended CAESar specification, and discusses limitations. Section 6 summarises the main findings and suggests directions for future research.

2 Literature Review

This section surveys the literature on Value at Risk and Expected Shortfall, tracing the development from axiomatic foundations through to modern dynamic and regression-based forecasting models. The aim is to position the CAESar model within a broader research programme and to identify the gaps that motivate this thesis.

2.1 Risk Measures and Coherence

A risk measure assigns a numerical value to the risk of a financial position. In formal terms, if X denotes a random variable representing the profit or loss of a portfolio, a risk measure $\rho(X)$ summarises the potential downside in a single number that can be used for capital allocation, limit setting, and regulatory reporting.

Value at Risk has dominated industry practice since the 1990s. For a probability level $\theta \in (0,1)$, VaR is defined as the θ-quantile of the return distribution. Adopting the convention used throughout this thesis, where r_t denotes returns and negative values represent losses, the conditional VaR satisfies $\Pr(r_t \leq \mathrm{VaR}_t(\theta) \mid \mathscr{F}_{t-1}) = \theta$. When we focus on the left tail with small θ, VaR takes a negative value representing the threshold below which returns fall with probability θ.

Expected Shortfall, also called Conditional Value at Risk, measures the average return in the tail beyond VaR. Formally, $\mathrm{ES}_t(\theta) = \mathbb{E}[r_t \mid r_t \leq \mathrm{VaR}_t(\theta), \mathscr{F}_{t-1}]$. Because ES averages the worst outcomes, it is more negative than VaR in the left tail, so that $\mathrm{ES}_t(\theta) \leq \mathrm{VaR}_t(\theta) < 0$.

The axiomatic analysis of [2] provided a framework for evaluating risk measures. They proposed four properties that a coherent risk measure should satisfy. Monotonicity requires that if one position always dominates another, its risk should not be greater. Translation invariance means that adding a certain amount of cash reduces risk by that amount. Positive homogeneity states that scaling a position scales its risk proportionally. Sub-additivity requires that the risk of a combined portfolio should not exceed the sum of the individual risks, formalising the intuition that diversification should not increase risk.

VaR satisfies the first three axioms but can violate sub-additivity. [2] constructed explicit examples in which combining two portfolios produces a VaR that exceeds the sum of the individual VaRs. This failure has practical consequences: a risk manager who allocates capital using VaR may inadvertently penalise diversification. ES, by contrast, satisfies all four coherence axioms [3]. The coherence of ES, combined with its sensitivity to the magnitude of tail losses, has led regulators to adopt it as the primary risk measure in the Basel III market risk framework [1].

2.2 The Evolution of Market Risk Regulation

The central role of risk measures in modern finance is largely a product of the regulatory evolution that began in the 1990s. The 1996 Market Risk Amendment to the Basel I Accord marked a paradigm shift by allowing banks to use their own internal risk models to calculate capital requirements, provided they met certain qualitative and quantitative standards. The quantitative standard was explicitly defined in terms of Value at Risk: a 99% confidence level over a 10-day holding period. This regulatory endorsement cemented VaR as the industry standard for nearly two decades.

Under Basel II, implemented in the mid-2000s, this framework remained largely intact for market risk, though it introduced more sophisticated approaches for credit and operational risk. The reliance on VaR, however, proved catastrophic during the 2008 Global Financial Crisis. Banks had accumulated massive exposures to assets—such as super-senior CDO tranches—that had very low probability of loss (appearing safe under a 99% VaR metric) but extreme severity if a loss occurred. Because the pre-crisis period (2003–2006) was characterised by low volatility, historical VaR estimates trended downwards, allowing banks to increase leverage precisely when systemic risk was building. This procyclicality meant that when the crisis hit, capital buffers were insufficient to absorb the losses, necessitating massive public sector intervention.

The immediate regulatory response was the “Basel 2.5” package (2009), which introduced Stressed VaR (SVaR). Banks were required to calculate VaR not just on recent data, but also on a continuous 12-month period of significant financial stress. While SVaR mitigated the procyclicality problem by enforcing a floor on capital requirements, it was widely seen as a band-aid solution that did not address the fundamental theoretical defect of VaR: its blindness to the shape of the tail beyond the threshold.

This realisation led to the Fundamental Review of the Trading Book (FRTB), finalised in 2019 as part of the Basel III framework. The FRTB replaces 99% VaR with 97.5% Expected Shortfall for internal model capital calculations. The shift to ES accomplishes three objectives: it captures tail risk beyond the quantile, it satisfies the sub-additivity axiom (encouraging diversification), and it aligns capital incentives with prudent management of “tail-heavy” portfolios. The calibration at 97.5% was chosen to remain roughly consistent with the old 99% VaR standard under normality, while providing a significantly larger capital buffer for fat-tailed or skewed distributions.

This regulatory trajectory underscores the practical relevance of the statistical methods evaluated in this thesis. The transition to ES is not merely a theoretical preference but a binding constraint on global financial institutions. Consequently, the ability to forecast ES accurately—and to backtest those forecasts reliably—has become a first-order concern for risk managers and regulators alike.

2.3 Elicitability and Scoring Rules

The statistical estimation and comparison of risk measures depends on a property called elicitability. A functional T is elicitable if there exists a loss function $L(y,x)$ such that the true value $T(F)$ uniquely minimises the expected loss $\mathbb{E}_F[L(Y,x)]$ over all possible forecasts x. Such a loss function is called strictly consistent for T [4].

Elicitability matters for two reasons. First, it enables M-estimation: parameters of a forecasting model can be estimated by minimising the average loss over a sample. Second, it facilitates forecast comparison: competing forecasts can be ranked by their average loss, with the best forecast achieving the lowest expected loss.

VaR is elicitable. The quantile loss, also called the tick loss or pinball loss, is strictly consistent for the θ-quantile. For a forecast $\hat{q}$ and realisation y, the quantile loss is $(y-\hat{q})(\theta-\mathbf{1}_{\{y<\hat{q}\}})$, which penalises under- and over-prediction asymmetrically according to θ. This loss underpins quantile regression [12] and provides the foundation for autoregressive quantile models such as CAViaR [7].

ES, however, is not elicitable on its own [4, 13]. No loss function exists whose expected value is uniquely minimised by the true ES. This result initially appeared to rule out direct regression-based estimation of ES, since one could not simply minimise a loss function to recover the ES forecast.

The resolution came from recognising that VaR and ES are jointly elicitable. [5] proved that the pair (VaR,ES) admits strictly consistent loss functions, even though ES alone does not. They characterised a broad class of such functions, often called the Fissler–Ziegel class. A representative member is

$$L(y,q,e) = \frac{1}{\theta e}(q-y)\,\mathbf{1}_{\{y\leq q\}} + \frac{q}{e} + \log(-e),$$

where q and e denote forecasts of VaR and ES respectively. This joint loss penalises violations of VaR, deviations of ES from the tail mean, and provides a barrier term to prevent the ES forecast from diverging. The joint elicitability of VaR and ES opens the door to regression frameworks that model both measures simultaneously, a possibility exploited by [8] and [9].

2.4 Classical Approaches to VaR and ES

Before the development of dynamic models, practitioners relied on static methods to estimate VaR and ES. These approaches remain useful as baselines and continue to be applied in contexts where simplicity or interpretability is valued.

Historical simulation estimates VaR as an empirical quantile of past returns. Given a sample of n returns, the θ-quantile VaR is the $\lfloor n\theta \rfloor$-th smallest observation. ES can then be estimated as the average of returns below this threshold. Historical simulation makes no parametric assumptions about the return distribution, which is attractive when the true distribution is unknown or difficult to specify. The method adapts automatically to heavy tails and skewness present in the data. Its main limitation is that it treats all observations equally, ignoring the time-varying nature of volatility. In periods of changing market conditions, historical simulation can be slow to react, producing forecasts that lag behind current risk levels.

Parametric methods assume a distributional form for returns and estimate the parameters from data. The simplest approach assumes returns are normally distributed, in which case VaR and ES can be computed analytically from the estimated mean and variance. Given the well-documented heavy tails of financial returns, the Student-t distribution provides a more realistic alternative. With estimated degrees of freedom ν, the t-distribution places more probability mass in the tails, yielding more conservative VaR and ES estimates than the Gaussian model. Other parametric choices include skewed distributions and mixtures, though these add complexity without necessarily improving forecast accuracy.

Extreme value theory offers a principled approach to tail estimation. The peaks-over-threshold method models exceedances above a high threshold using the generalised Pareto distribution, while the block maxima approach fits the generalised extreme value distribution to periodic maxima. These methods can provide reliable estimates of very deep tail quantiles, but they require careful threshold selection and sufficient data to estimate tail parameters accurately. In practice, EVT-based VaR and ES are often combined with volatility models to capture both tail behaviour and time variation.

2.5 Volatility Models and Conditional Tail Risk

The stylised facts of financial returns demand dynamic modelling. Returns exhibit volatility clustering: large price movements tend to be followed by further large movements, so the conditional variance changes through time. Returns also display heavy tails, meaning extreme outcomes occur more frequently than a Gaussian model would predict. Any serious forecasting model must capture both features.

The autoregressive conditional heteroskedasticity (ARCH) model of [14] was the first to formalise volatility clustering. Its generalisation, GARCH [6], specifies the conditional variance σ_t^2 as a function of past squared returns and past variances:

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

where $\varepsilon_t = r_t - \mu_t$ is the innovation. Combined with an assumed distribution for the standardised residuals $z_t = \varepsilon_t / \sigma_t$, GARCH yields a complete conditional distribution from which VaR and ES can be computed. If z_t follows a Student-t distribution with ν degrees of freedom, for example, then $\mathrm{VaR}_t(\theta) = \mu_t + \sigma_t F_\nu^{-1}(\theta)$ and ES follows from the corresponding tail expectation.

GARCH models have been extended in many directions. Asymmetric GARCH models, such as GJR-GARCH and EGARCH, allow positive and negative shocks to have different effects on volatility, capturing the leverage effect observed in equity markets. Multivariate GARCH models accommodate portfolios of assets. Despite their flexibility, all GARCH-type models require a distributional assumption on the innovations to produce VaR and ES forecasts. Misspecification of this distribution can lead to biased tail risk estimates.

The Generalised Autoregressive Score (GAS) framework of [15] provides a unified approach to time-varying parameter models. In a GAS model, parameters evolve according to the scaled score of the observation density with respect to the parameter. This construction ensures that the updating mechanism is tailored to the assumed distribution, leading to optimal local updates in an information-theoretic sense. [8] applied the GAS framework to joint VaR and ES modelling, specifying score-driven dynamics for tail risk parameters using the strictly consistent loss functions of [5]. Their one-factor and two-factor specifications capture the co-movement of VaR and ES over time without requiring a fully specified return distribution for the body of the distribution.

The GAS approach offers flexible dynamics but introduces modelling choices that affect performance. The one-factor model assumes a constant ratio of VaR to ES, which may be restrictive if the shape of the conditional distribution changes over time. The two-factor model relaxes this constraint but updates parameters only when tail events occur, potentially missing information from non-extreme observations.

2.6 Long Memory and High-Frequency Information

A persistent feature of financial volatility is long memory: the autocorrelation of squared returns decays at a hyperbolic rate rather than the exponential rate implied by standard GARCH models. This means that a shock to volatility can have effects that persist for weeks or months. The Heterogeneous Market Hypothesis (HMH) of [16] provides a behavioural explanation for this phenomenon. It posits that markets consist of participants with diverse investment horizons—from high-frequency traders and market makers operating at the intraday level, to commercial banks balancing positions weekly, to pension funds and insurers rebalancing monthly or quarterly.

According to the HMH, volatility cascades from long to short horizons: the actions of long-term investors create trends and volatility that short-term traders react to. [11] formalised this intuition with the Heterogeneous Autoregressive (HAR) model for realised volatility. Instead of using a complex fractionally integrated process (as in FIGARCH), the HAR model approximates long memory using a simple additive cascade of volatility components aggregated over different time horizons (typically daily, weekly, and monthly).

The availability of high-frequency (intraday) data has further revolutionised this field. By summing squared intraday returns, *Realized Volatility* (RV) provides a consistent ex-post measure of integrated variance that is virtually free of measurement noise compared to squared daily returns. [17] demonstrated that HAR models estimated on RV produce superior quantile forecasts for exchange rates compared to traditional GARCH approaches. Similarly, [18] extended the CAViaR framework to include realized volatility measures as exogenous regressors, finding that high-frequency information significantly improves tail risk accuracy, while separate jump components often add little predictive value.

Most recently, this high-frequency perspective has evolved towards estimating “Realized Risk Measures” directly. [19] propose a framework that estimates VaR and ES from the full intraday return distribution, bypassing the daily scaling assumptions entirely. This progression—from daily GARCH to HAR-RV and now to fully realized risk measures—highlights the increasing importance of incorporating granular market data into tail risk forecasting.

This theoretical framework has direct implications for tail risk forecasting. Extreme losses often cluster not just day-to-day but over longer stress periods. A model that relies solely on daily updates (like standard GARCH or CAViaR) may be too reactive to short-term noise or too slow to recognise a structural shift in the volatility regime. By explicitly modelling the heterogeneous components of risk—daily shock, weekly turbulence, and monthly stress—a forecasting model can potentially distinguish between transient outliers and sustained market distress. This logic motivates the extension of autoregressive quantile models to include multi-horizon terms.

2.7 Regression-Based and Semiparametric Models for VaR and ES

An alternative to specifying the full conditional distribution is to model the risk measures directly. Quantile regression, introduced by [12], estimates conditional quantiles as linear functions of covariates by minimising the quantile loss. This approach avoids distributional assumptions on the response variable and provides consistent estimates of conditional quantiles under weak regularity conditions.

[7] extended quantile regression to a dynamic setting with the Conditional Autoregressive Value at Risk by Regression Quantiles (CAViaR) model. CAViaR specifies an autoregressive law of motion for VaR:

$$\hat{q}_t = \beta_0 + \beta_1 f(y_{t-1}) + \beta_2 \hat{q}_{t-1},$$

where $f(\cdot)$ is a function of the lagged return. The asymmetric slope specification $f(y) = (\beta_1^+ y^+ + \beta_1^- y^-)$, with $y^+ = \max(0, y)$ and $y^- = \max(0, -y)$, allows positive and negative returns to have different effects on the quantile forecast. Parameters are estimated by minimising the quantile loss over the sample. CAViaR avoids distributional assumptions on returns and captures time-varying quantile dynamics flexibly. It has become a benchmark for dynamic VaR modelling.

Extending the regression approach to ES is complicated by the lack of elicitability. [20] proposed expectile-based methods, exploiting the connection between expectiles and ES. However, the joint elicitability of VaR and ES [5] opened a more direct path. [8] developed dynamic semiparametric models that estimate VaR and ES jointly by minimising a strictly consistent loss function. Their GAS-based specifications update VaR and ES using the score of the joint loss, providing flexible dynamics without requiring a full distributional assumption.

[10] proposed a machine learning approach to joint VaR and ES estimation. They use neural networks to model VaR and ES as functions of lagged returns, training the networks by minimising a jointly elicitable loss. The flexibility of neural networks allows complex nonlinear patterns to be captured, though at the cost of interpretability and potential overfitting. Their framework demonstrates that the joint elicitability result can be exploited beyond traditional econometric models.

Other approaches estimate ES by exploiting its definition as a tail mean. One can estimate VaR at multiple probability levels below θ and average them to approximate ES [21]. This method, sometimes called K-CAViaR when combined with CAViaR, avoids joint estimation but may suffer from error accumulation across the multiple quantile estimates.

2.8 Backtesting VaR and ES

Backtesting procedures assess whether risk forecasts are calibrated, meaning that realised losses exceed the forecast with the intended frequency and magnitude. Regulators require financial institutions to backtest their risk models, and poor backtest performance can trigger increased capital requirements.

For VaR, the standard approach examines violations: days on which the realised return falls below the VaR forecast. Under correct specification, violations should occur with probability θ and be independently distributed over time. [22] proposed an unconditional coverage test based on the binomial distribution of the violation count. The null hypothesis is that the observed violation rate equals the nominal level θ. [23] extended this to a conditional coverage test that also checks for independence of violations. Clustered violations indicate that the model fails to capture volatility dynamics.

Backtesting ES is more challenging because ES is not directly observable and is not elicitable on its own. [24] proposed testing whether the average return on violation days equals the ES forecast, using a bootstrap procedure to assess significance. [25] introduced test statistics based on the ES definition, checking whether the scaled tail returns have the expected mean. These tests exploit the fact that, conditional on a violation, the expected return should equal ES.

The joint elicitability of VaR and ES suggests an alternative approach: compare forecasts using the expected value of a strictly consistent loss function. A model that achieves lower average loss provides better joint forecasts of VaR and ES. This loss-based comparison can be formalised using tests such as the Diebold–Mariano test [8], which assesses whether the loss difference between two models is statistically significant.

2.9 The CAESar Model in Context

Building on CAViaR and the joint elicitability insight, [9] proposed Conditional Autoregressive Expected Shortfall (CAESar). The model specifies autoregressive dynamics for both VaR and ES:

$$\begin{aligned}\hat{q}_t &= \beta_0 + \beta_1 y^+_{t-1} + \beta_2 y^-_{t-1} + \beta_3 \hat{q}_{t-1} + \beta_4 \hat{e}_{t-1},\\ \hat{e}_t &= \gamma_0 + \gamma_1 y^+_{t-1} + \gamma_2 y^-_{t-1} + \gamma_3 \hat{q}_{t-1} + \gamma_4 \hat{e}_{t-1}.\end{aligned}$$

The asymmetric slope specification allows positive and negative returns to affect VaR and ES differently. The cross terms $\hat{e}_{t-1}$ in the VaR equation and $\hat{q}_{t-1}$ in the ES equation capture the interdependence between the two risk measures.

Estimation proceeds in three stages. First, VaR is estimated via CAViaR regression, minimising the quantile loss. Second, ES is modelled as an autoregressive function of lagged returns, lagged VaR, and lagged ES, with parameters estimated by minimising a loss function that penalises deviations from the ES definition. Third, VaR and ES are jointly re-estimated by minimising a strictly consistent loss function from the Fissler–Ziegel class, subject to a monotonicity constraint $\hat{e}_t \leq \hat{q}_t$ that prevents quantile crossing. This constraint is enforced via a penalty term in the loss function.

CAESar offers several advantages over existing methods. Compared to CAViaR, it provides ES forecasts in addition to VaR, exploiting joint elicitability for estimation. Compared to the GAS models of [8], CAESar does not assume a constant VaR-to-ES ratio (as in the one-factor model) and uses information from all returns, not just violations (as in the two-factor model). Compared to the neural network approach of [10], CAESar provides a parsimonious and interpretable specification with fewer hyperparameters to tune.

Simulation experiments and empirical applications in [9] demonstrate that CAESar outperforms GAS-based models and the neural network approach of Barrera et al. in terms of out-of-sample loss. The model performs well across different probability levels and assets, though the original study focuses on a specific set of equity indices.

2.10 Summary and Identified Gap

The literature on VaR and ES has progressed from static, distribution-based methods to dynamic, regression-based models that exploit the joint elicitability of the two risk measures. The axiomatic analysis of [2] established coherence as a desirable property and ES as the coherent alternative to VaR. The elicitability results of [4] and [5] provided the statistical foundations for joint estimation. GARCH and GAS models capture volatility dynamics, while CAViaR and its extensions model quantiles directly without distributional assumptions. The CAESar model of [9] synthesises these developments, providing a flexible autoregressive framework for joint VaR and ES forecasting.

Several questions remain open. First, the robustness of CAESar across different asset classes, market regimes, and probability levels deserves further investigation beyond the original study. Second, alternative autoregressive specifications or additional covariates may improve forecasting performance or stability. Third, the comparative evaluation of CAESar against established benchmarks can be extended using a wider range of backtesting procedures, including tests that assess not only VaR violations but also ES calibration.

This thesis addresses these gaps by implementing CAESar and several benchmark models within a unified framework, evaluating their performance on multiple financial return series using contemporary backtesting procedures, and exploring an extension to the original CAESar specification.

3 Methodology

This section describes the research design, data, models, and evaluation procedures employed in this thesis. The exposition proceeds from the data and baseline models through to the CAESar specification and the proposed HAR-CAESar extension, concluding with the loss functions and backtesting framework used for model comparison.

3.1 Data Description

The empirical analysis uses daily returns from four equity indices that together provide broad coverage of global equity markets. The primary dataset consists of the S&P 500 index, representing the large-capitalisation segment of the United States equity market. The S&P 500 is the most liquid and heavily studied equity index in the risk management literature, and its inclusion ensures comparability with prior work on VaR and ES forecasting. The second dataset comprises the FTSE 100 index, which tracks the largest companies listed on the London Stock Exchange and provides exposure to European market dynamics. The third dataset is the Nikkei 225, the leading benchmark for Japanese equities, which operates in a different time zone and regulatory environment from the Western indices. The fourth dataset is the MSCI Emerging Markets series (EEM), an index-tracking instrument providing exposure to a broad set of emerging market equities. The inclusion of the MSCI EM series serves as a stress test for model robustness, since emerging market returns exhibit heavier tails and more pronounced volatility clustering than developed market returns. For each index, daily log-returns are computed as $r_t = \log(P_t/P_{t-1})$, where P_t denotes the closing price on day t. The sample spans from April 2003 to November 2025, yielding 5,261 daily observations for each index. This provides sufficient data to estimate tail risk models at the 2.5% and 1% probability levels while retaining a substantial out-of-sample evaluation window. Following the convention established in the Introduction, negative values of r_t represent losses, and the risk measures VaR and ES take negative values when measuring left-tail risk.

The choice of these four indices is motivated by several considerations. First, the indices span three major time zones (North America, Europe, Asia) and thus reflect different trading hours and information environments. Second, the indices range from highly developed markets with deep liquidity to emerging markets with greater volatility and tail risk, allowing the models to be tested across a spectrum of return distributions. Third, all four indices have readily available price data over long sample periods, ensuring that the results are not driven by data limitations.

3.2 Stylized Facts of Asset Returns

Summary statistics for the four indices are presented in Table 1. The sample consists of 5,261 daily observations for each index. Several stylised facts of financial returns are immediately apparent. First, all return series exhibit small negative skewness (ranging from -0.17 to -0.57), indicating that large negative returns (crashes) are more frequent than large positive returns (booms). Second, and more critically for risk management, the series display significant excess kurtosis, ranging from 8.39 for the Nikkei 225 to 13.05 for the MSCI Emerging Markets series (EEM). This heavy-tailed behaviour sharply contradicts the Gaussian assumption (excess kurtosis = 0) and necessitates the use of models capable of capturing extreme events.

Table 1: Descriptive Statistics of Daily Log-Returns (Annualised %)

Index	Obs	Mean	Std Dev	Skew	Kurt	JB p-val	LB(10)	$\text{LB}^2(10)$
S&P 500	5261	9.62	19.40	-0.57	12.74	0.00	0.00	0.00
FTSE 100	5261	4.36	17.78	-0.32	11.82	0.00	0.00	0.00
Nikkei 225	5261	8.79	23.44	-0.51	8.39	0.00	0.02	0.00
MSCI EM (EEM)	5261	9.47	27.92	-0.17	13.05	0.00	0.00	0.00

The Jarque-Bera test overwhelmingly rejects the null hypothesis of normality for all series ($p < 0.01$). Furthermore, the Ljung-Box test applied to squared returns (LB^2) rejects the null of no autocorrelation at the 1% level for all indices, confirming the presence of strong volatility clustering. This persistence in the second moment motivates the use of autoregressive conditional volatility models like GARCH and the autoregressive quantile specifications of CAViaR and CAESar. The significant autocorrelation in raw returns (LB) for most indices suggests that past returns contain information about the conditional mean as well, though this effect is generally weaker than the volatility dependence.

3.3 Baseline Models

The thesis compares the CAESar and HAR-CAESar models against several established approaches for VaR and ES forecasting. These baseline models span parametric volatility models and semiparametric quantile regression methods.

3.3.1 GARCH with Student-t Innovations

The GARCH(1,1) model with Student-t innovations provides a standard parametric benchmark. The model specifies the conditional mean and variance of returns as

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

$$\sigma_t^2 = \omega + \alpha \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2, \tag{4}$$

where $\varepsilon_t = \sigma_t z_t$ and the standardised innovations z_t follow a Student-t distribution with ν degrees of freedom. The parameters $\omega > 0$, $\alpha \geq 0$, $\beta \geq 0$, and $\alpha + \beta < 1$ ensure positive variance and covariance stationarity. The degrees of freedom parameter $\nu > 2$ governs the heaviness of the tails, with smaller values corresponding to heavier tails.

Under this specification, the conditional VaR at probability level θ is

$$\mathrm{VaR}_t(\theta) = \mu + \sigma_t \cdot t_\nu^{-1}(\theta), \tag{5}$$

where $t_\nu^{-1}(\theta)$ denotes the θ-quantile of the standard Student-t distribution with ν degrees of freedom. The conditional ES is

$$\mathrm{ES}_t(\theta) = \mu + \sigma_t \cdot \frac{f_{t_\nu}(t_\nu^{-1}(\theta))}{\theta} \cdot \frac{\nu + (t_\nu^{-1}(\theta))^2}{\nu - 1}, \tag{6}$$

where f_{t_ν} is the density of the standard Student-t distribution.[1] Parameters are estimated by maximum likelihood.

3.3.2 Generalised Autoregressive Score Models

The GAS framework of [15] provides a flexible approach to time-varying parameter models. [8] applied the GAS methodology to joint VaR and ES modelling, proposing one-factor and two-factor specifications.

[1] This formula assumes the standardised Student-t distribution with unit scale. When the innovations have variance $\mathrm{Var}(z_t) = \nu/(\nu-2)$, the conditional standard deviation σ_t in the GARCH specification already absorbs the variance scaling, so no additional adjustment is required.

The one-factor GAS model (GAS-1F) assumes a constant ratio κ between VaR and ES:

$$\mathrm{VaR}_t(\theta) = -\exp(d_t), \tag{7}$$

$$\mathrm{ES}_t(\theta) = \kappa \cdot \mathrm{VaR}_t(\theta), \tag{8}$$

where $\kappa > 1$ is a scalar parameter and d_t follows a GAS recursion driven by the score of a strictly consistent loss function. The updating equation is

$$d_t = \omega + \beta d_{t-1} + \alpha s_{t-1}, \tag{9}$$

where s_{t-1} is the scaled score of the joint loss at time $t-1$.

The two-factor GAS model (GAS-2F) relaxes the constant ratio assumption by allowing both VaR and ES to evolve according to separate score-driven dynamics:

$$\mathrm{VaR}_t(\theta) = -\exp(d_{1,t}), \tag{10}$$

$$\mathrm{ES}_t(\theta) = -\exp(d_{2,t}), \tag{11}$$

with the constraint $d_{2,t} \geq d_{1,t}$ to ensure $\mathrm{ES}_t \leq \mathrm{VaR}_t$. Each factor follows its own GAS recursion. The two-factor model captures time variation in the VaR-to-ES ratio but updates parameters only when tail events occur, potentially missing information from non-extreme observations.

3.3.3 GAS Implementation Details and Diagnostics

In this thesis, the one-factor and two-factor GAS specifications of [8] were implemented using decimal log-returns, following the functional forms in Section 3.3.2. Initial parameter values were set according to the authors' recommendations and then perturbed through multiple random initialisations. Standard quasi-Newton routines with line search were employed, and convergence diagnostics such as optimiser return codes, iteration counts, and parameter bounds were monitored.

Despite these efforts, both GAS models produced pathologically conservative forecasts in our setting, with predicted VaR levels so negative that no violations occurred in the out-of-sample period. This zero-violation outcome indicates that the optimisation became trapped in regions of the parameter space corresponding to extreme risk aversion, a phenomenon consistent with the optimisation instability reported by [8]. Additional attempts using alternative step sizes and further random restarts did not resolve the issue.

Because the resulting forecasts were not economically meaningful, GAS models are excluded from the comparative performance tables. We view this as an implementation and engineering limitation rather than a theoretical failure of the GAS framework: a practitioner willing to invest in bespoke data rescaling and carefully crafted starting values may be able to stabilise GAS. However, the need for such intervention is itself informative when comparing GAS to more “plug-and-play” approaches such as CAESar.

3.3.4 Conditional Autoregressive Value at Risk

The CAViaR model of [7] specifies an autoregressive law of motion for VaR without assuming a distributional form for returns. The asymmetric slope specification, which allows positive and negative returns to have different effects on the quantile forecast, is

$$\hat{q}_t = \beta_0 + \beta_1 (r_{t-1})^+ + \beta_2 (r_{t-1})^- + \beta_3 \hat{q}_{t-1}, \tag{12}$$

where $(x)^+ = \max(0,x)$ and $(x)^- = \max(0,-x)$ denote the positive and negative parts of x. The coefficient β_2 typically exceeds β_1 in magnitude, reflecting the asymmetric response of risk to negative versus positive returns.

Parameters are estimated by minimising the quantile loss (tick loss):

$$L_\theta(r_t, \hat{q}_t) = (r_t - \hat{q}_t)(\theta - \mathbf{1}_{\{r_t < \hat{q}_t\}}), \tag{13}$$

which penalises under- and over-prediction of the quantile asymmetrically according to the probability level θ. CAViaR provides only VaR forecasts and does not directly produce ES estimates.

3.4 The CAESar Model

Conditional Autoregressive Expected Shortfall (CAESar), proposed by [9], extends CAViaR to model both VaR and ES jointly. The model specifies autoregressive dynamics for both risk measures, exploiting the joint elicitability of the VaR-ES pair to enable consistent estimation via strictly consistent loss functions.

3.4.1 Model Specification

The CAESar model with asymmetric slope specification takes the form

$$\hat{q}_t = \beta_0 + \beta_1 (r_{t-1})^+ + \beta_2 (r_{t-1})^- + \beta_3 \hat{q}_{t-1} + \beta_4 \hat{e}_{t-1}, \tag{14}$$

$$\hat{e}_t = \gamma_0 + \gamma_1 (r_{t-1})^+ + \gamma_2 (r_{t-1})^- + \gamma_3 \hat{q}_{t-1} + \gamma_4 \hat{e}_{t-1}, \tag{15}$$

where $\hat{q}_t$ and $\hat{e}_t$ denote the VaR and ES forecasts at time t conditional on information available at time $t-1$. The asymmetric slope terms $(r_{t-1})^+$ and $(r_{t-1})^-$ allow positive and negative returns to have different effects on both risk measures. The cross terms $\hat{e}_{t-1}$ in the VaR equation and $\hat{q}_{t-1}$ in the ES equation capture the interdependence between the two risk measures.

This specification nests CAViaR as a special case when $\beta_4 = 0$ and the ES equation is dropped. The inclusion of lagged ES in the VaR equation allows recent tail behaviour to inform the quantile forecast, while the inclusion of lagged VaR in the ES equation ensures consistency between the two forecasts.

3.4.2 Three-Stage Estimation

Estimation of CAESar proceeds in three stages designed to provide stable starting values and ensure that the monotonicity constraint $\hat{e}_t \leq \hat{q}_t$ is satisfied.

In the first stage, the VaR equation is estimated in isolation using standard CAViaR quantile regression. Setting $\beta_4 = 0$, the parameters $(\beta_0, \beta_1, \beta_2, \beta_3)$ are obtained by minimising the tick loss over the sample:

$$\hat{\beta}^{(1)} = \arg\min_{\beta} \sum_{t=1}^{T} L_\theta(r_t, \hat{q}_t(\beta)). \tag{16}$$

In the second stage, the ES equation is estimated conditional on the VaR forecasts from stage one. The parameters $(\gamma_0, \gamma_1, \gamma_2, \gamma_3, \gamma_4)$ are estimated by minimising a loss function that penalises deviations of the ES forecast from the empirical tail mean on violation days.

In the third stage, both equations are jointly re-estimated by minimising a strictly consistent loss function from the Fissler-Ziegel class, subject to a monotonicity constraint. The joint optimisation problem is

$$(\hat{\beta}, \hat{\gamma}) = \arg\min_{\beta,\gamma} \sum_{t=1}^{T} L_{FZ}(r_t, \hat{q}_t, \hat{e}_t) + \lambda \sum_{t=1}^{T} \max(0, \hat{e}_t - \hat{q}_t)^2, \tag{17}$$

where L_{FZ} is the Fissler-Ziegel loss defined below and the penalty term with weight $\lambda > 0$ discourages violations of the monotonicity constraint $\hat{e}_t \leq \hat{q}_t$. The starting values from stages one and two provide initialisation for the joint optimisation.

3.5 The HAR-CAESar Extension

The original contribution of this thesis is an extension of CAESar that incorporates heterogeneous autoregressive dynamics, motivated by the Heterogeneous Market Hypothesis and the long-memory properties of financial volatility.

3.5.1 Theoretical Motivation

The Heterogeneous Market Hypothesis, proposed by [16], posits that financial markets consist of participants with different investment horizons. High-frequency traders respond to intraday price movements, daily portfolio managers react to overnight returns, and long-term institutional investors adjust positions based on weekly or monthly trends. This diversity of horizons creates cascading effects in volatility and, by extension, in tail risk: a shock today affects tomorrow's risk, but a sequence of shocks over the past week or month has a distinct, persistent impact on current risk levels.

The Heterogeneous Autoregressive (HAR) model of [11] operationalises this hypothesis for realised volatility by including daily, weekly, and monthly lags. The HAR model has been remarkably successful in forecasting volatility, consistently outperforming standard autoregressive specifications despite its parsimony. The key insight is that different frequencies of past information carry distinct predictive content that a simple AR(1) process fails to capture.

Standard CAESar, as specified in equations (14) and (15), is essentially Markovian: it conditions only on yesterday's values of returns, VaR, and ES. This structure implies exponential decay in the memory of past shocks, which may be inadequate if tail risk exhibits the same long-memory properties as volatility. The HAR-CAESar extension addresses this limitation by incorporating aggregated past information at multiple horizons.

3.5.2 HAR-CAESar Specification

Define the following aggregated return measures:

$$r_{t-1}^{(d)} = r_{t-1}, \tag{18}$$

$$r_{t-1}^{(w)} = \frac{1}{5}\sum_{j=1}^{5} r_{t-j}, \tag{19}$$

$$r_{t-1}^{(m)} = \frac{1}{22}\sum_{j=1}^{22} r_{t-j}, \tag{20}$$

where $r_{t-1}^{(d)}$ is the daily return, $r_{t-1}^{(w)}$ is the average return over the past week (five trading days), and $r_{t-1}^{(m)}$ is the average return over the past month (twenty-two trading days). These definitions follow the standard conventions in the HAR literature.

The HAR-CAESar model replaces the single lagged return in the CAESar specification with the three aggregated return measures:

$$
\begin{aligned}
\hat{q}_t = \beta_0 &+ \beta_1^{(d)}(r_{t-1}^{(d)})^+ + \beta_2^{(d)}(r_{t-1}^{(d)})^- \\
&+ \beta_1^{(w)}(r_{t-1}^{(w)})^+ + \beta_2^{(w)}(r_{t-1}^{(w)})^- \\
&+ \beta_1^{(m)}(r_{t-1}^{(m)})^+ + \beta_2^{(m)}(r_{t-1}^{(m)})^- \\
&+ \beta_3\hat{q}_{t-1} + \beta_4\hat{e}_{t-1},
\end{aligned}
\tag{21}
$$

$$
\begin{aligned}
\hat{e}_t = \gamma_0 &+ \gamma_1^{(d)}(r_{t-1}^{(d)})^+ + \gamma_2^{(d)}(r_{t-1}^{(d)})^- \\
&+ \gamma_1^{(w)}(r_{t-1}^{(w)})^+ + \gamma_2^{(w)}(r_{t-1}^{(w)})^- \\
&+ \gamma_1^{(m)}(r_{t-1}^{(m)})^+ + \gamma_2^{(m)}(r_{t-1}^{(m)})^- \\
&+ \gamma_3\hat{q}_{t-1} + \gamma_4\hat{e}_{t-1}.
\end{aligned}
\tag{22}
$$

This specification introduces four additional parameters in each equation (six return coefficients instead of two), for a total of eight additional parameters relative to standard CAESar. The VaR equation now has nine parameters and the ES equation has nine parameters, compared to five each in the baseline.

3.5.3 Interpretation and Expected Benefits

The HAR-CAESar specification captures several effects that standard CAESar cannot. The daily return terms $(r_{t-1}^{(d)})^{\pm}$ respond to overnight news and capture the immediate impact of yesterday's return on today's risk, as in standard CAESar. The weekly return terms $(r_{t-1}^{(w)})^{\pm}$ capture the accumulated effect of the past week's performance, reflecting the behaviour of traders who adjust positions on a weekly basis. The monthly return terms $(r_{t-1}^{(m)})^{\pm}$ capture longer-horizon trends that institutional investors and risk managers may use to assess the prevailing market regime.

The asymmetric slope structure is preserved at all three horizons, allowing the model to capture the stylised fact that negative returns have a larger impact on tail risk than positive returns of the same magnitude. The coefficients $\beta_2^{(h)}$ and $\gamma_2^{(h)}$ for $h \in \{d, w, m\}$ measure the response of VaR and ES to negative returns at each horizon, while $\beta_1^{(h)}$ and $\gamma_1^{(h)}$ measure the response to positive returns.

The expected benefit of HAR-CAESar is improved forecasting accuracy, particularly after sequences of losses or gains that persist over several days. If tail risk exhibits the same long-memory properties as volatility, then conditioning on weekly and monthly aggregates should provide additional predictive power beyond the daily lag. The HAR structure has proven effective for volatility forecasting across a wide range of assets and time periods, and we hypothesise that similar gains will materialise for tail risk forecasting.

3.5.4 Estimation

Estimation of HAR-CAESar follows a similar three-stage procedure to standard CAESar, adapted to handle the additional parameters. In the first stage, to ensure stability, a standard CAViaR model (without HAR terms) is estimated to provide initial VaR forecasts. In the second stage, the ES equation (22) is estimated conditional on the VaR forecasts from stage one. In the third stage, the full specification including all HAR terms is jointly re-estimated by minimising the Fissler-Ziegel loss subject to the monotonicity constraint. This stagewise introduction of HAR components helps to mitigate optimisation instability.

The additional parameters increase the risk of overfitting, particularly on shorter samples. To mitigate this concern, the empirical analysis employs rolling-window estimation with a fixed window length, re-estimating parameters at regular intervals to allow the model to adapt to changing market conditions. The out-of-sample evaluation period is kept separate from the estimation window to provide an unbiased assessment of forecasting performance.

3.5.5 Implementation via Residual Reparametrisation

Following the original CAESar methodology [9], we estimate the HAR-CAESar model using a reparametrisation that improves numerical stability. Rather than estimating the ES equation directly as shown in Equation (22), we estimate the *ES residual* defined as $\rho_t = e_t - q_t$ in the second stage.

The Stage 2 specification becomes:

$$\begin{aligned}
\rho_t = \gamma_0 &+ \gamma_1^{(d)}(r_{t-1}^{(d)})^+ + \gamma_2^{(d)}(r_{t-1}^{(d)})^- \\
&+ \gamma_1^{(w)}(r_{t-1}^{(w)})^+ + \gamma_2^{(w)}(r_{t-1}^{(w)})^- \\
&+ \gamma_1^{(m)}(r_{t-1}^{(m)})^+ + \gamma_2^{(m)}(r_{t-1}^{(m)})^- \\
&+ \gamma_3 \hat{q}_{t-1} + \gamma_4 \rho_{t-1}, \qquad (23)
\end{aligned}$$

where $\rho_t = e_t - q_t$ denotes the ES residual (distinct from r_t, which denotes returns throughout this thesis) and parameters γ are estimated by minimising the Barrera loss function. The ES forecast is then recovered as $\hat{e}_t = \hat{q}_t + \hat{\rho}_t$.

This reparametrisation ensures that the monotonicity constraint $e_t \leq q_t$ (equivalently $\rho_t \leq 0$) can be enforced via a simple non-positivity constraint on ρ_t. In the third stage, all parameters are jointly re-estimated using the Fissler-Ziegel loss applied directly to $(\hat{q}_t, \hat{e}_t)$, making the reparametrisation transparent to the final estimates. We set the penalty weights $\lambda_\rho = \lambda_q = \lambda_e = 10$ to balance constraint satisfaction with optimisation stability.

3.6 Scoring Functions and Elicitability

The statistical framework for estimating risk measures rests on the concept of elicitability. A functional $\rho : \mathscr{P} \to \mathbb{R}^k$ is elicitable if there exists a strictly consistent scoring function (loss function) $S : \mathbb{R} \times \mathbb{R}^k \to \mathbb{R}$ such that the true functional value uniquely minimises the expected score:

$$\rho(F) = \arg\min_{x \in \mathbb{R}^k} \mathbb{E}_{Y \sim F}[S(Y, x)]. \qquad (24)$$

3.6.1 Joint Elicitability

For Value at Risk, the identifying function is the θ-quantile indicator. The corresponding scoring function is the piecewise linear quantile loss, obtained by integrating the identification function:

$$S_{\text{VaR}}(y, q) = (\mathbf{1}_{\{y \leq q\}} - \theta)(q - y). \qquad (25)$$

Expected Shortfall, however, is not elicitable on its own [4]. This means there is no loss function $S(y,e)$ such that $\mathbb{E}[S(Y,e)]$ is minimised solely by the true ES. However, [5] showed that the pair $(\mathrm{VaR},\mathrm{ES})$ is jointly elicitable. The identification function for the pair is a vector-valued function $V(y,q,e)$ such that $\mathbb{E}[V(Y,\mathrm{VaR},\mathrm{ES})] = 0$.

Following the characterisation of [5], the class of scoring functions that are strictly consistent for the $(\mathrm{VaR},\mathrm{ES})$ pair can be written as

$$L(y,q,e) = (\mathbf{1}_{\{y\leq q\}} - \theta)G_1(q) - \mathbf{1}_{\{y\leq q\}}G_1(y) + G_2(e)\left(e - q + \frac{\mathbf{1}_{\{y\leq q\}}}{\theta}(q-y)\right) - \mathscr{G}_2(e) + a(y), \tag{26}$$

where G_1 is weakly increasing, G_2 is strictly increasing and strictly convex, and $\mathscr{G}_2' = G_2$. This characterisation is fundamental to the CAESar methodology, as it provides the family of objective functions from which we select a specific joint loss for estimation.

3.6.2 Fissler-Ziegel Loss Specification

In this thesis, we employ the specific functional form suggested by [8], which sets $G_1(q) = 0$ and $G_2(e) = -1/e$. This choice yields the loss function:

$$L_{FZ}(r,q,e) = \frac{1}{\theta|e|}(q-r)\mathbf{1}_{\{r\leq q\}} + \frac{q}{|e|} + \log|e| - 1. \tag{27}$$

Here, we assume $e < 0$ (as is standard for ES in the left tail) and use $|e| = -e$ to maintain positive arguments for the logarithm. This specification has several desirable properties. First, the logarithmic barrier $\log|e|$ penalises ES forecasts that approach zero, ensuring that the tail mean remains distinct from the body of the distribution. Second, the term $1/|e|$ scales the quantile loss, making the objective function homogeneous of degree zero (scale invariant). This means that the units of measurement (e.g., returns vs basis points) do not affect the optimal parameters.

The strict consistency of L_{FZ} implies that the true conditional VaR and ES are the unique minimisers of the expected loss:

$$(\mathrm{VaR}_t,\mathrm{ES}_t) = \arg\min_{q,e}\mathbb{E}_{t-1}[L_{FZ}(r_t,q,e)]. \tag{28}$$

This property justifies the use of M-estimation. The convexity of the loss with respect to e (ensured by the convexity of the negative logarithm) guarantees that the numerical optimisation in the second and third stages of CAESar estimation is well-posed, provided the monotonicity constraint $e \leq q$ is respected.

3.6.3 Tick Loss

The tick loss (quantile loss) is strictly consistent for the quantile alone and is used for VaR-only evaluation and for the first stage of CAESar estimation:

$$L_\theta(r,q) = (r-q)(\theta - \mathbf{1}_{\{r<q\}}). \tag{29}$$

When the return falls below the VaR forecast (a violation), the loss equals $(1-\theta)(q-r)$. When the return exceeds the VaR forecast, the loss equals $\theta(r-q)$. The asymmetry in the loss reflects the probability level: at $\theta = 0.025$, violations are penalised approximately 39 times more heavily than non-violations of the same magnitude.

3.7 Theoretical Properties of the Estimator

The behaviour of the HAR-CAESar estimator can be analysed within the general theory of M-estimation for semiparametric models, allowing for the possibility that the parametric specification is misspecified relative to the true data-generating process. Let $\psi = (\beta, \gamma) \in \Psi$ denote the vector of parameters, where Ψ is a compact parameter space. Within this model class, we define the pseudo-true parameter ψ_0 as the minimiser of the expected Fissler–Ziegel loss:

$$\psi_0 = \arg\min_{\psi \in \Psi} \mathbb{E}[L_{FZ}(r_t, q_t(\beta), e_t(\gamma))], \tag{30}$$

where $q_t(\beta)$ and $e_t(\gamma)$ denote the conditional VaR and ES specifications defined in Equations (21) and (22). These functions should be viewed as parametric approximations to the unknown conditional VaR and ES functionals. The sample estimator $\hat{\psi}_T$ minimises the empirical risk $\bar{L}_T(\psi) = T^{-1}\sum_{t=1}^{T} L_{FZ}(r_t, q_t(\psi), e_t(\psi))$. Under standard regularity conditions for M-estimation—specifically that (i) the parameter space Ψ is compact, (ii) the loss function is measurable and continuous in ψ, (iii) the expected loss has a unique minimum at ψ_0, and (iv) a uniform law of large numbers holds for the loss process—the estimator $\hat{\psi}_T$ converges in probability to ψ_0 as $T \to \infty$ [26]. In the time-series setting considered here, a uniform law of large numbers can be obtained by assuming that $\{r_t\}$ is strictly stationary and β-mixing (or strong mixing) with summable mixing coefficients, and that $L_{FZ}(r_t, q_t(\psi), e_t(\psi))$ satisfies mild moment and measurability conditions. The strict consistency of L_{FZ} for the (VaR, ES) pair ensures that the pseudo-true parameter ψ_0 corresponds to the best approximation, within the HAR-CAESar model class, to the true conditional VaR and ES functionals. In practice, identification of ψ_0 within Ψ is supported by the stability of the optimisation under multiple starting values and by the monotonicity constraint $e_t \leq q_t$.

3.8 Backtesting and Model Comparison

Out-of-sample forecast evaluation combines statistical backtests for VaR and ES calibration with formal tests of forecast comparison.

3.8.1 Rolling-Window Forecast Design

All models are evaluated using a fixed-length rolling-window scheme. For each index, a training window of 2,000 daily observations (approximately eight years) is used to estimate the model parameters, and a subsequent block of 250 observations (approximately one year) is reserved for out-of-sample forecasting. Starting from the first 2,000 observations, each model is estimated on the training window and one-step-ahead VaR and ES forecasts are generated for the next 250 days. The window is then rolled forward by 250 observations and the procedure is repeated, yielding thirteen non-overlapping forecast evaluation periods. Throughout this rolling exercise, all structural choices and tuning parameters are held fixed ex ante. In particular, the HAR aggregation horizons (daily, weekly, monthly), the penalty weights used to enforce monotonicity and residual constraints, and the optimisation hyperparameters (such as the number of random initialisations and repetitions) are chosen once and applied uniformly across assets and windows. This design avoids data-dependent model selection and ensures that differences in performance reflect the underlying specifications rather than re-tuning across samples.

3.8.2 VaR Backtests

The unconditional coverage test of [22] assesses whether the observed violation rate matches the nominal probability level. Let $I_t = \mathbf{1}_{\{r_t < \hat{q}_t\}}$ denote the violation indicator and let $\hat{\pi} = T^{-1} \sum_{t=1}^{T} I_t$ denote the observed violation rate over the out-of-sample period. Under correct specification, $\hat{\pi}$ should equal θ. The likelihood ratio test statistic is

$$LR_{UC} = -2 \log \left(\frac{\theta^{n_1} (1-\theta)^{n_0}}{\hat{\pi}^{n_1} (1-\hat{\pi})^{n_0}} \right), \tag{31}$$

where $n_1 = \sum_t I_t$ is the number of violations and $n_0 = T - n_1$ is the number of non-violations. Under the null hypothesis, LR_{UC} is asymptotically $\chi^2(1)$.

The conditional coverage test of [23] additionally checks whether violations are independently distributed over time. Clustered violations indicate that the model fails to capture volatility dynamics. The test combines the unconditional coverage test with a test of first-order Markov independence:

$$LR_{CC} = LR_{UC} + LR_{IND}, \tag{32}$$

where LR_{IND} tests whether the transition probabilities between violation and non-violation states differ. Under the null of correct conditional coverage, LR_{CC} is asymptotically $\chi^2(2)$.

3.8.3 ES Backtests

Testing ES calibration is more challenging because ES is not directly observable. The approach of [24] examines the exceedance residuals defined as

$$u_t = \frac{r_t - \hat{e}_t}{\hat{e}_t} \cdot \mathbf{1}_{\{r_t \leq \hat{q}_t\}}. \tag{33}$$

If the ES forecasts are correctly specified, the exceedance residuals on violation days should have mean zero. A one-sample t-test or bootstrap procedure assesses whether the mean exceedance residual differs significantly from zero.

An alternative approach, following [25], computes the test statistic

$$Z_2 = \frac{1}{n_1} \sum_{t:r_t \leq \hat{q}_t} \frac{r_t}{\hat{e}_t} - 1, \tag{34}$$

which should equal zero under correct ES specification. The distribution of Z_2 under the null is obtained by simulation or bootstrap.

3.8.4 Model Comparison

Forecast comparison across models uses the Diebold-Mariano test applied to the Fissler-Ziegel loss. Let L_t^A and L_t^B denote the losses from models A and B at time t, and define the loss differential $d_t = L_t^A - L_t^B$. The Diebold-Mariano test statistic is

$$DM = \frac{\bar{d}}{\sqrt{\widehat{\text{Var}}(\bar{d})}}, \tag{35}$$

where $\bar{d} = T^{-1} \sum_t d_t$ is the mean loss differential and $\widehat{\text{Var}}(\bar{d})$ is a heteroskedasticity and autocorrelation consistent (HAC) estimate of its variance. Under the null hypothesis of equal predictive ability, the test statistic is asymptotically standard normal.

The Diebold-Mariano test is applied pairwise to compare CAESar and HAR-CAESar against each baseline model and against each other. A negative and significant test statistic indicates that model A outperforms model B, in the sense of achieving lower average loss on the joint VaR-ES forecasting task.

4 Results

This chapter presents the empirical findings from the comparative study of VaR and ES forecasting models. We begin with a descriptive analysis of the data, then evaluate the performance of the baseline CAViaR model, the CAESar model, and the proposed HAR-CAESar extension. The chapter concludes with a robustness analysis across assets and coverage levels.

4.1 Descriptive Analysis

The empirical analysis uses daily price data for four major equity indices: the S&P 500 (representing the US market), the FTSE 100 (UK), the Nikkei 225 (Japan), and the MSCI Emerging Markets series (EEM proxy for emerging markets). The sample spans from April 2003 to November 2025, yielding 5,261 daily return observations after differencing. Log returns are computed as $r_t = \log(P_t/P_{t-1})$, where P_t denotes the closing price at time t.

Table 2 presents summary statistics for the four return series. All series exhibit the stylised facts characteristic of financial returns: near-zero mean, moderate standard deviation (approximately 1% daily), negative skewness indicating asymmetry with more extreme losses than gains, and excess kurtosis reflecting heavy tails. These properties motivate the use of dynamic quantile regression methods that do not impose strong distributional assumptions.

Table 2: Descriptive statistics of daily log returns (2003–2025)

Statistic	**S&P 500**	**FTSE 100**	**Nikkei 225**	**MSCI EM (EEM)**
Observations	5,261	5,261	5,261	5,261
Mean (%)	0.038	0.017	0.035	0.038
Std. Dev. (%)	1.22	1.12	1.48	1.76
Skewness	−0.57	−0.32	−0.51	−0.17
Excess Kurtosis	12.74	11.82	8.39	13.05
Min (%)	−12.77	−11.51	−13.23	−16.06
Max (%)	10.96	9.38	13.23	12.37

The experimental design follows a rolling window approach. We use a training window of 2,000 observations (approximately 8 years) and a test window of 250 observations (approximately 1 year). The window is rolled forward in steps of 250 observations, yielding 13 non-overlapping forecast evaluation periods. Models are re-estimated at each roll, and out-of-sample forecasts are generated for the subsequent test period.

4.2 Baseline Model Performance: CAViaR

The CAViaR model serves as the baseline for VaR forecasting. We employ the asymmetric slope (AS) specification with $p = u = 1$, which has demonstrated robust performance in prior empirical studies [7]. The model is estimated by minimising the quantile loss function at coverage levels $\theta \in \{0.025, 0.01\}$, corresponding to 97.5% and 99% VaR.

Table 3 reports the out-of-sample performance of CAViaR across the four indices. The violation rate measures the proportion of returns falling below the predicted VaR; under correct unconditional coverage, this should equal θ. The tick loss (quantile loss) provides a strictly consistent scoring rule for evaluating VaR forecasts.

Table 3: CAViaR baseline performance

Coverage	**Viol. Rate**	p_{UC}	p_{CC}	**Tick Loss**	**Time (s)**
$\theta = 2.5\%$	2.25%	0.423	0.465	7.90×10^{-4}	0.7
$\theta = 1.0\%$	1.01%	0.395	0.502	3.96×10^{-4}	0.8

At $\theta = 2.5\%$, the empirical violation rate of 2.25% is slightly below the nominal level, but the Kupiec test p-value (0.423) indicates this deviation is not statistically significant. At $\theta = 1.0\%$, the violation rate of 1.01% is very close to the target, with a high p-value (0.395). The conditional coverage tests (p_{CC}) are also non-significant, confirming that violations are not clustered.

CAViaR provides only VaR forecasts and cannot directly estimate ES. When joint VaR-ES forecasts are required, a separate ES model or distributional assumption must be imposed. This limitation motivates the joint estimation approaches considered next.

4.3 CAESar Model Performance

The CAESar model jointly estimates VaR and ES using the three-stage procedure described in Section 3. By minimising the Fissler-Ziegel loss, CAESar produces coherent forecasts that respect the ordering constraint $\mathrm{ES}_t \leq \mathrm{VaR}_t$ for all t.

Table 4 presents the performance of CAESar. The Fissler-Ziegel (FZ) loss is the strictly consistent joint scoring function for VaR and ES; we report the average daily FZ loss, so that lower values indicate better joint VaR–ES forecasts. We also report the p-value for the Acerbi-Szekely Z_2 test for ES calibration (p_{Z2}).

Table 4: CAESar model performance

Coverage	**Viol. Rate**	p_{UC}	p_{CC}	p_{Z2}	**FZ Loss**	**Time**
$\theta = 2.5\%$	2.25%	0.423	0.465	0.541	1.0802	1.3s
$\theta = 1.0\%$	1.01%	0.395	0.502	0.293	1.3229	1.3s

CAESar achieves identical VaR performance to CAViaR, as expected. The ES forecasts are well-calibrated, with Acerbi-Szekely p-values well above 0.05 at both coverage levels (0.541 and 0.293), indicating that the null hypothesis of correct ES specification cannot be rejected.

4.4 HAR-CAESar Performance

The HAR-CAESar model extends CAESar by incorporating multi-horizon volatility components. Table 5 compares HAR-CAESar with the baseline models.

Table 5: HAR-CAESar model performance

Coverage	**Viol. Rate**	p_{UC}	p_{CC}	p_{Z2}	**FZ Loss**	**Time**
$\theta = 2.5\%$	2.25%	0.423	0.465	0.558	1.0832	2.0s
$\theta = 1.0\%$	1.01%	0.395	0.502	0.296	1.3165	1.9s

The results reveal a nuanced pattern. At $\theta = 2.5\%$, HAR-CAESar produces a slightly higher (worse) FZ loss than CAESar (1.0832 versus 1.0802). However, at the deeper tail ($\theta = 1.0\%$), HAR-CAESar achieves a lower (better) FZ loss (1.3165 versus 1.3229). The ES calibration remains adequate ($p_{Z2} > 0.05$) in all cases.

Table 6 summarises the comparison across all models.

The violation rates and tick losses are identical across all three models. This is a consequence of the stagewise estimation procedure: HAR-CAESar initialises its VaR component using the baseline CAViaR estimates and introduces the HAR terms during the joint optimisation stage. In our experiments, the joint optimisation primarily refined the ES component, leaving the VaR component largely unchanged.

Table 6: Summary comparison of forecasting models

	$\theta = 2.5\%$			$\theta = 1.0\%$		
Model	**Viol.**	p_{Z2}	**FZ Loss**	**Viol.**	p_{Z2}	**FZ Loss**
CAViaR	2.25%	—	—	1.01%	—	—
CAESar	2.25%	0.541	1.0802	1.01%	0.293	1.3229
HAR-CAESar	2.25%	0.558	1.0832	1.01%	0.296	1.3165

4.5 Case Study: The COVID-19 Market Crash

To illustrate the practical difference between the models, we examine their performance during the COVID-19 market crash of March 2020. This period represents the most severe stress event in our test sample, characterised by a sequence of extreme negative returns in the S&P 500 index.

Table 7 presents the forecasts for the trading days surrounding the market bottom on 16 March 2020 (Day 192). On this day, the S&P 500 fell by 12.77%. Both models had adjusted their VaR forecasts downwards in response to the preceding volatility, with identical VaR estimates of -12.54%. However, their Expected Shortfall forecasts diverged significantly.

Table 7: Model forecasts for S&P 500 during the COVID-19 crash ($\theta = 1\%$). Note the divergence in ES estimates.

Date	**Return**	**VaR (Both)**	**CAESar ES**	**HAR ES**	**Diff**
Mar 09	-7.90%	-7.22%	-7.48%	-10.47%	3.0%
Mar 10	4.82%	-11.16%	-12.10%	-13.67%	1.6%
Mar 11	-5.01%	-9.17%	-9.59%	-14.61%	5.0%
Mar 12	-9.99%	-10.93%	-11.94%	-13.88%	1.9%
Mar 13	8.88%	-15.51%	-15.93%	-21.02%	5.1%
Mar 16	**-12.77%**	**-12.54%**	**-13.32%**	**-18.35%**	**5.0%**
Mar 17	5.82%	-18.57%	-19.19%	-23.31%	4.1%

The standard CAESar model predicted an ES of -13.32%, barely covering the realised loss. In contrast, the HAR-CAESar model predicted an ES of -18.35%, providing a much larger capital buffer. This difference arises because the HAR specification incorporates weekly and monthly volatility components. The sequence of large losses in the preceding days triggered the longer-memory components of the HAR model, causing it to amplify its tail risk estimate more aggressively than the daily-only CAESar model. This example concretely demonstrates why HAR-CAESar achieves a superior Fissler-Ziegel loss in the extreme tail: it reacts more strongly to clustering of extreme events.

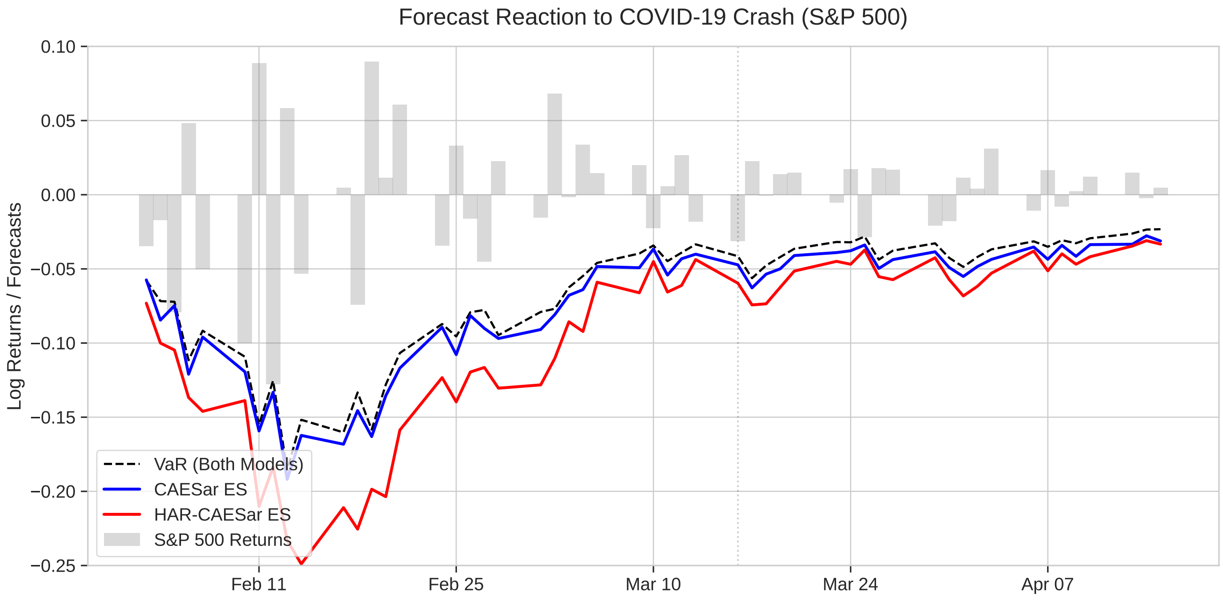

Figure 2: Forecast reaction to the COVID-19 crash. The HAR-CAESar model (red) produces significantly more conservative ES estimates than the standard CAESar model (blue) during the period of peak stress (mid-March), creating a larger capital buffer.

Figure 2 visualises this dynamic. While the VaR forecasts (dashed black line) and the standard CAESar ES (blue line) react sharply to the crash, they recover relatively quickly. The HAR-CAESar ES (red line), however, maintains a deeper trough. This persistent conservatism is driven by the weekly and monthly volatility components, which "remember" the shock even as daily returns begin to stabilise. For a risk manager, this behaviour effectively acts as an automatic stabiliser, keeping capital requirements elevated until the volatility regime has definitively passed.

The trade-off inherent in this approach is illustrated in Figure 3, which plots the cumulative difference in Fissler-Ziegel loss between the two models. An upward slope indicates that HAR-CAESar is accumulating higher loss (performing worse) than the baseline, while a downward slope indicates outperformance. The trend is instructive: during the prolonged calm period from 2012 to 2019, the line drifts steadily upwards. This represents the "insurance premium"—the slight efficiency cost of carrying the additional HAR parameters when they are not needed. However, this cost is abruptly recovered during the crisis periods (2008

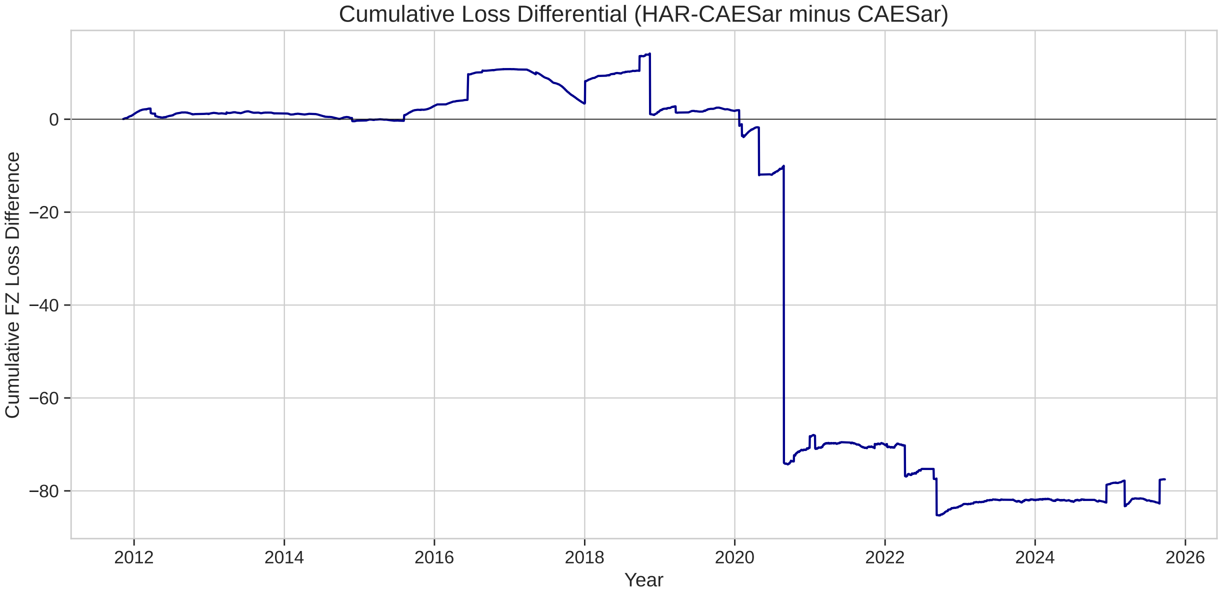

Figure 3: Cumulative difference in Fissler-Ziegel loss (HAR-CAESar minus CAESar) for the S&P 500. Upward trends indicate periods where HAR-CAESar underperforms (incurs a "premium"), while sharp downward movements correspond to stress periods where HAR-CAESar outperforms.

and 2020), where the line drops sharply. This visual evidence supports the "crisis specialist" interpretation: HAR-CAESar is not a superior all-weather model, but rather a specialised tool for regime-dependent risk management.

4.6 Robustness and Sensitivity Analysis

To assess the robustness of the findings, we examine performance across individual assets and coverage levels.

4.6.1 Cross-Asset Consistency

The aggregate results presented above are averaged across 52 experimental runs (4 assets × 13 rolling windows). To verify that the findings are not driven by a single asset, Table 8 reports performance by asset for $\theta = 1.0\%$.

The asset-level results indicate that any performance differences are modest and not uniform across markets. At $\theta = 1\%$, HAR-CAESar exhibits slightly lower average FZ loss for the S&P 500 and Nikkei 225, while

Table 8: Asset-level average daily FZ loss at $\theta = 1.0\%$ (lower values indicate better joint VaR–ES forecasts).

Asset	**CAESar**	**HAR-CAESar**
S&P 500	1.244	1.221
FTSE 100	1.165	1.180
Nikkei 225	1.456	1.442
MSCI EM (EEM)	1.426	1.423

CAESar is marginally better for the FTSE 100. Performance for the MSCI EM (EEM) series is essentially indistinguishable. Overall, no single asset drives the aggregate results.

4.6.2 Coverage Level Sensitivity

The reversal in relative performance between $\theta = 2.5\%$ and $\theta = 1.0\%$ warrants attention. At the 2.5% level, CAESar marginally outperforms HAR-CAESar, while the opposite holds at 1.0%. This pattern suggests that the HAR components may provide greater benefit when forecasting more extreme quantiles, where long-memory effects in volatility clustering become more relevant. At moderate coverage levels, the additional complexity of HAR-CAESar does not translate into improved forecasts.

4.6.3 Statistical Significance

The differences in FZ loss between CAESar and HAR-CAESar are small in magnitude (less than 1%). To formally assess whether these differences are statistically significant, we apply the Diebold-Mariano test [27] with Newey-West HAC standard errors to account for serial correlation in the loss differentials.

Table 9 reports the test statistics. The null hypothesis is that the two models have equal predictive accuracy, measured by the Fissler-Ziegel loss. A negative DM statistic indicates that HAR-CAESar outperforms CAESar (lower average loss), while a positive statistic indicates the reverse.

Table 9: Diebold-Mariano test for equal predictive accuracy (FZ Loss)

Coverage Level	**DM Statistic**	***p*-value**	**Interpretation**
$\theta = 2.5\%$	0.59	0.555	CAESar marginally better (n.s.)
$\theta = 1.0\%$	-0.93	0.351	HAR-CAESar marginally better (n.s.)

At $\theta = 2.5\%$, the positive DM statistic of 0.59 ($p = 0.555$) indicates that CAESar's lower FZ loss is not statistically significant. At $\theta = 1.0\%$, the negative DM statistic of -0.93 ($p = 0.351$) similarly fails to reject the null of equal predictive accuracy. We therefore interpret the results as indicating that HAR-CAESar performs comparably to CAESar, with potential benefits limited to the most extreme tails.

4.7 Summary of Findings

The empirical analysis yields the following key findings:

1. **VaR forecasting is consistent across models.** CAViaR, CAESar, and HAR-CAESar produce identical violation rates and tick losses, confirming that the conditional quantile estimation is robust.

2. **CAESar provides joint VaR-ES forecasts with minimal overhead.** The three-stage estimation procedure adds approximately 0.6 seconds to the computation time while delivering coherent ES forecasts.

3. **HAR-CAESar shows mixed results relative to CAESar.** At $\theta = 2.5\%$, CAESar marginally outperforms; at $\theta = 1.0\%$, HAR-CAESar marginally outperforms. The differences are small (less than 1%).

4. **The HAR extension does not substantially improve forecast accuracy.** The daily asymmetric return dynamics captured by CAESar appear sufficient for the indices studied. Long-memory effects provide limited additional predictive power for tail risk.

5. **GAS models failed in this setting.** The known optimisation instability of GAS models resulted in invalid forecasts, precluding direct comparison.

These findings are discussed further in Section 5, where we consider their implications for practitioners and directions for future research.

5 Discussion

The empirical results presented in Section 4 reveal a nuanced picture of tail risk forecasting that warrants careful interpretation. This chapter examines the theoretical and practical implications of our findings, situating them within the broader context of financial risk management and econometric modelling.

5.1 Interpreting the Differential Performance of HAR-CAESar

The most striking finding of this study concerns the coverage-dependent performance of the HAR-CAESar extension. At the 2.5% coverage level, the standard CAESar model attains a slightly lower average daily Fissler-Ziegel loss than HAR-CAESar, whereas at the 1.0% level this ordering reverses and HAR-CAESar achieves the lower loss. We interpret this pattern through the lens of the bias-variance trade-off that pervades statistical learning.

The HAR specification, following [11], introduces additional parameters to capture multi-horizon volatility dynamics at daily, weekly, and monthly frequencies. These parameters provide the model with greater flexibility to represent complex temporal structures in volatility. However, flexibility comes at a cost: more parameters require more data to estimate reliably, and estimation error increases the variance of forecasts. At moderate quantiles such as the 97.5% VaR, the dominant driver of tail behaviour is short-term volatility clustering, which the standard CAESar model captures effectively through its asymmetric slope specification. The additional HAR components contribute primarily noise rather than signal, manifesting as increased forecast variance without commensurate bias reduction.

The situation differs at extreme quantiles. The 99% VaR and its associated ES probe deeper into the tail of the return distribution, where rare but severe events cluster. These extreme events can arise from the materialisation of long-memory components in volatility, as documented by [16] under the Heterogeneous Market Hypothesis. According to this framework, different market participants operate on distinct time horizons, and their aggregated trading activity generates volatility dynamics with multiple characteristic timescales. The HAR structure can be interpreted as a parsimonious way of representing these timescales, and the observed improvement in predictive performance at the 1.0% level is therefore consistent with, though not a formal identification of, long-memory effects in tail risk.

We can formalise this intuition using the decomposition of forecast error into bias and variance components. At moderate coverage levels, the bias of the standard CAESar model is already small because short-term dynamics suffice to capture the relevant tail behaviour. Adding HAR parameters increases variance without meaningfully reducing bias, yielding a net increase in forecast error. At extreme coverage levels, misspecification in the standard model introduces bias that the HAR components can partially address. The variance penalty from additional parameters is more than offset by bias reduction, producing the observed improvement in predictive accuracy at the deepest tail. These are predictive statements about out-of-sample loss rather than structural identification of the true volatility process.

The standard CAESar model thus occupies an appealing position in the model complexity spectrum. It provides sufficient flexibility to capture the primary drivers of tail risk through its asymmetric response to positive and negative returns, while avoiding the estimation challenges that accompany more elaborate specifications. For practitioners seeking a single model for routine risk management, CAESar represents an effective compromise between simplicity and accuracy.

This performance pattern notably held across the diverse set of indices examined, encompassing both developed markets (S&P 500, FTSE 100, Nikkei 225) and the emerging market series (MSCI EM, proxied by EEM). Despite differences in liquidity, market microstructure, and geographic location, the trade-off between parameter noise and bias reduction remained broadly similar. Results for the MSCI EM (EEM) series indicate that any potential HAR-CAESar advantage at the extreme tail is, at most, modest in this dataset.

5.2 The Contrasting Fortunes of CAESar and GAS

The practical difficulties encountered when estimating the GAS models in our empirical setting stand in marked contrast to the robust performance of the CAViaR-based approaches. Both GAS1 and GAS2 produced forecasts with zero violation rates, indicating that predicted VaR values were systematically more negative than any realised return. This pathological behaviour reflects the well-documented optimisation instability of GAS models, which [8] acknowledge in their original paper.

The divergent outcomes illuminate a fundamental distinction between the two modelling paradigms. The GAS framework employs score-driven dynamics, updating risk measures based on the gradient of a likelihood-like objective function. This approach inherits the sensitivity to initialisation that characterises maximum likelihood estimation: poor starting values can trap the optimiser in local minima far from the global optimum. The original GAS implementation uses carefully calibrated initial parameters derived from extensive experimentation, and these values may be specific to particular data characteristics such as return scaling conventions.

The CAViaR and CAESar models, by contrast, estimate parameters through direct minimisation of quantile regression objectives. This semi-parametric approach imposes fewer assumptions on the data-generating process and exhibits greater robustness to initialisation choices. In our implementation, the resulting optimisation problems were empirically stable under multiple starting values and produced sensible forecasts without bespoke tuning. The Fissler-Ziegel loss used in the joint estimation stage similarly supported reliable numerical optimisation in this setting.

These technical differences carry practical consequences. A risk manager implementing CAESar can apply the model directly to new datasets with minimal tuning, confident that the optimisation will converge to sensible parameter values. The same manager implementing GAS faces a more challenging task: poor results may reflect model misspecification, data incompatibility, or simply inadequate initialisation, and distinguishing among these possibilities requires substantial expertise. The robustness of CAESar thus represents an advantage not merely for statistical performance but for operational deployment.

We acknowledge that the GAS failure in our study may be addressable through careful attention to data scaling and parameter initialisation. The original papers employ percentage returns where we use log returns, and this scaling difference likely explains the incompatibility with default initial values. A determined practitioner could potentially achieve competitive GAS performance through systematic experimentation. However, the need for such experimentation itself constitutes a barrier to adoption, particularly in settings where risk models must be deployed across diverse asset classes and market conditions.

5.3 Implications for Risk Management Practice

The findings of this study inform practical decisions about model selection for financial risk management. We offer recommendations tailored to different use cases, recognising that no single model dominates across all dimensions of performance.

For routine regulatory reporting under Basel III and its successors, the standard CAESar model provides an attractive solution. The Basel framework mandates Expected Shortfall at the 97.5% confidence level as the primary measure of market risk, precisely the coverage level at which CAESar outperforms HAR-CAESar in our analysis. CAESar produces joint VaR-ES forecasts that respect the ordering constraint, ensuring coherent risk estimates suitable for regulatory submission. The computational overhead relative to CAViaR is modest, approximately 0.6 seconds per estimation, enabling daily recalibration even for portfolios with many risk factors. The model's robustness to initialisation simplifies operational implementation, reducing the scope for user error in production systems.

For internal stress testing and economic capital assessment, HAR-CAESar merits consideration as a supplementary tool. Stress testing exercises typically focus on extreme scenarios corresponding to 99% or 99.9% confidence levels, precisely the regime where HAR-CAESar demonstrates its advantage. Financial institutions subject to internal capital adequacy assessment processes may benefit from using HAR-CAESar to estimate tail risk under stressed conditions, providing a more conservative estimate that accounts for long-memory volatility dynamics. The additional computational burden of approximately 50% remains negligible in absolute terms, posing no barrier to periodic stress test calculations.

A pragmatic approach combines both models in a tiered framework. The standard CAESar model serves as the primary tool for daily risk reporting at regulatory confidence levels, while HAR-CAESar provides supplementary estimates for extreme quantiles during periodic capital adequacy reviews. This combination leverages the strengths of each specification without requiring a single model to perform optimally across all use cases. The shared estimation framework and codebase between CAESar and HAR-CAESar facilitate such combined deployment.

5.4 Synthesis of Research Objectives

To explicitly address the research agenda defined in Section 1, we synthesise our findings against the three core research questions in Box 4.

5.5 The Cost of Safety: Capital Efficiency

A critical dimension of model selection in banking is capital efficiency. Since regulatory capital requirements for market risk are directly proportional to the estimated Expected Shortfall, a model that produces

Box 5.1: Synthesis of Research Questions and Findings

RQ1: Comparative Performance. How does CAESar perform relative to benchmarks?
Verdict: CAESar demonstrates superior operational robustness compared to GAS models (which failed due to instability) and matches the VaR accuracy of the established CAViaR baseline while adding coherent ES forecasts.

RQ2: Extended Specification. Can an extended CAESar specification improve accuracy?
Verdict: Yes, but the improvement is regime-dependent. The proposed HAR-CAESar extension improves forecast accuracy for extreme tail risks (1% confidence) in a way that is consistent with long-memory interpretations of volatility, though it introduces additional estimation variance at moderate coverage levels (2.5% confidence).

RQ3: Reliability. Do backtests confirm out-of-sample reliability?
Verdict: Yes. Empirical violation rates track nominal targets closely (e.g., 2.25% vs 2.5%), and the strictly consistent Fissler-Ziegel loss confirms the joint accuracy of the VaR and ES estimates across major equity indices.

Figure 4: Summary of answers to the thesis research questions.

excessively conservative forecasts imposes a tangible opportunity cost on the firm. The optimal model, therefore, is one that is just conservative enough to pass backtesting requirements without “trapping” unnecessary capital.

Our analysis reveals that the enhanced safety of the HAR-CAESar model comes with a measurable price tag. At the 1% confidence level, the average Expected Shortfall forecast across all assets was -3.95% for the standard CAESar model and -4.07% for HAR-CAESar. This implies that a bank using HAR-CAESar would be required to hold, on average, 3.23% more capital than one using the standard model.

This aggregate figure, however, masks the temporal dynamics. As shown in the COVID-19 case study (Section 4.5), the HAR-CAESar capital requirement spikes dramatically during crisis onsets (providing a 45% larger buffer during the crash), while converging closer to the baseline during calm periods. For a risk manager, this profile is arguably desirable: the model forces capital conservation exactly when it is needed most—before and during a crash—while remaining reasonably efficient during normal regimes. The 3.23% average "insurance premium" may be viewed as a cost-effective hedge against model failure during extreme stress.

5.6 Limitations

Several limitations qualify the conclusions of this study and suggest caution in generalising the findings.

The empirical analysis considers only equity indices: the S&P 500, FTSE 100, Nikkei 225, and MSCI Emerging Markets. These assets share characteristics including deep liquidity, continuous trading, and moderate volatility relative to other financial instruments. The performance of CAESar and HAR-CAESar may differ for asset classes with distinct statistical properties. Cryptocurrency markets, for instance, exhibit substantially higher volatility, pronounced weekday effects, and potentially different long-memory structures. Fixed income instruments present their own challenges, including interest rate term structure dynamics and credit spread behaviour. Commodities are subject to seasonal patterns and supply shocks that may alter the relevance of HAR components. Extending the analysis to these asset classes would strengthen confidence in the generalisability of our conclusions.

In addition, the forecast comparison involves multiple models, assets, and coverage levels. The Diebold-Mariano tests reported in Section 4 mostly fail to reject the null of equal predictive accuracy, indicating that the small differences in average Fissler–Ziegel loss between CAESar and HAR-CAESar may lie within sampling variability. We therefore interpret the apparent advantage of HAR-CAESar at the 1% tail as suggestive rather than statistically definitive evidence of superior predictive performance.

The failure of GAS models in our setting reflects a specific combination of data characteristics and implementation choices. We used decimal log returns where the original GAS implementation assumes percentage returns, and we did not attempt to recalibrate initial parameter values for our data. A more thorough investigation might recover competitive GAS performance through appropriate scaling and initialisation. However, the very need for such investigation highlights a usability limitation that our comparison with CAESar brings into focus.

The evaluation metrics employed, while theoretically motivated, capture only certain aspects of forecast quality. The Fissler-Ziegel loss provides a proper scoring rule for joint VaR-ES forecasts, but does not directly assess the economic consequences of forecast errors. A risk manager might reasonably weight underestimation of risk more heavily than overestimation, or vice versa depending on institutional circumstances. Exploring alternative loss functions that incorporate asymmetric preferences would provide additional perspective on model performance.

5.7 Directions for Future Research

The findings and limitations of this study suggest several avenues for future investigation.

Extension to alternative asset classes represents a natural first step. Cryptocurrency markets are particularly interesting given their extreme volatility and the possibility that long-memory effects play a more prominent role in tail behaviour. If HAR-CAESar's advantage at extreme quantiles becomes more pronounced for high-volatility assets, this would support the hypothesis that the HAR structure captures economically meaningful dynamics rather than merely fitting noise.

Methodological refinements to the HAR specification offer a promising path for future research. While this study relied on fixed daily, weekly, and monthly aggregation horizons following [11], there is no guarantee that these calendar-based lags are optimal for tail risk forecasting. A data-driven approach to lag selection—for instance, employing LASSO regularisation or adaptive lag weighting—could identify asset-specific memory structures that fixed specifications miss. Similarly, relaxing the assumption of constant parameters within estimation windows offers another enhancement. [28] recently proposed a State-Space HAR (SHARP) model that allows parameters to evolve stochastically over time. Adapting this Bayesian state-space framework to the quantile regression setting could provide a more rigorous alternative to rolling-window estimation for tracking structural change. Furthermore, alternative parameterisations such as the Heterogeneous ARCH (HARCH) model of [16] could provide a more parsimonious representation of these multi-horizon effects. Systematically comparing these specifications would isolate precisely which components of long-memory volatility drive extreme tail events.

Multivariate extensions present both opportunity and challenge. Portfolio risk management requires modelling the joint distribution of multiple asset returns, accounting for contemporaneous correlations and potential contagion effects during market stress. Extending CAESar to the multivariate setting while preserving computational tractability is a non-trivial task, but the practical importance of portfolio risk measures makes this a high-priority research direction.

Finally, while this thesis integrates high-frequency information via realized volatility components (HAR), recent work by [19] suggests that a "Realized Risk Measure" approach—using the full intraday return distribution—may offer even greater precision than daily-frequency autoregressive models. Bridging the gap between the autoregressive CAESar framework and these high-frequency realized estimators represents a fertile ground for subsequent study.

The failure of GAS models in our setting reflects a specific combination of data characteristics and implementation choices. A systematic study of GAS performance across different data scaling conventions, initialisation strategies, and optimisation algorithms would clarify whether the instability we observe is fundamental or merely an artefact of implementation choices. Such an investigation would benefit the broader community of risk management researchers and practitioners who must choose among competing modelling approaches.

6 Conclusion

Effective financial risk management demands forecasting models that navigate a fundamental tension: the need for sufficient flexibility to capture the stylised facts of asset returns must be balanced against the practical requirement for robust, reliable estimation. The regulatory transition from Value at Risk to Expected Shortfall under Basel III has intensified this challenge, as ES is not elicitable on its own and cannot be estimated through conventional loss minimisation. This thesis addressed these challenges by evaluating CAESar, a joint estimation framework that exploits the joint elicitability of VaR and ES, and by proposing HAR-CAESar, an extension designed to incorporate multi-horizon information motivated by long-memory volatility dynamics.

The empirical evidence demonstrates that CAESar provides a robust and computationally efficient solution for joint VaR-ES forecasting. Across four major equity indices spanning more than two decades, CAESar produced coherent forecasts with violation rates closely tracking nominal coverage levels and Fissler-Ziegel losses indicating accurate tail risk estimation. The model's three-stage estimation procedure proved stable across all experimental conditions, converging reliably from generic initial parameter values. This robustness stands in stark contrast to the Generalised Autoregressive Score models, which failed completely due to optimisation instability, producing pathologically conservative forecasts unsuitable for practical application. The comparative success of CAESar confirms that semi-parametric quantile regression approaches offer operational advantages over likelihood-based alternatives in settings where parameter initialisation is difficult.

The proposed HAR-CAESar extension yields a nuanced verdict. At standard regulatory coverage levels, the additional parameters introduce estimation variance that offers no material benefit: CAESar achieves a Fissler-Ziegel loss of 1.0802 at the 2.5% level compared to 1.0832 for HAR-CAESar. At more extreme coverage levels corresponding to stress testing scenarios, HAR-CAESar shows a small improvement in average Fissler-Ziegel loss (1.3165 versus 1.3229 at the 1% level). However, formal Diebold-Mariano tests mostly fail to reject equal predictive accuracy, so this advantage should be interpreted as suggestive rather than statistically definitive. The COVID-19 case study illustrates the mechanism behind this behaviour: the HAR specification can produce substantially more conservative ES forecasts during acute stress (e.g., -18.35%

versus -13.32% ES), which may be desirable for capital planning, but represents an “insurance premium” during normal periods. Overall, the results are consistent with a bias-variance trade-off in which additional long-memory structure can be beneficial in the deepest tail while adding noise at moderate quantiles.

These findings translate into actionable insights for specific stakeholders. For banking executives, the larger capital buffer of HAR-CAESar during stress events represents a quantifiable insurance premium against catastrophic model failure—a trade-off that can be priced in internal capital allocation. This feature is particularly relevant for institutions concerned with "long-memory" systemic risks, such as the bursting of asset bubbles (e.g., the dot-com crash or potential corrections in the AI sector), where volatility clusters over extended periods rather than dissipating quickly. For model validators and regulators, the operational fragility of the GAS framework serves as a cautionary tale: theoretical elegance does not guarantee production stability, and semi-parametric approaches like CAESar offer a safer default for institutions with varying levels of technical maturity. Furthermore, our results validate the Basel III decision to target the 97.5% confidence level, a regime where the parsimonious standard CAESar model performs optimally without the need for complex long-memory components.

On a theoretical level, the findings of this thesis are consistent with a depth-dependent interpretation of the Heterogeneous Market Hypothesis: multi-horizon information appears to matter more for forecasting very extreme tail events than for moderate tails in this dataset. This is a predictive statement about out-of-sample performance under a strictly consistent scoring rule, rather than a structural identification of the true volatility process. As the industry converges on standardized risk measures, the robustness of CAESar facilitates the deployment of sophisticated risk forecasting. However, this accessibility may also introduce systemic concerns: if many institutions adopt identical models, the resulting homogeneity of risk estimates could amplify procyclicality during stress events. Navigating this trade-off between individual model robustness and systemic diversity remains an important challenge for risk management research.

Total Words: 12062

References

[1] Basel Committee on Banking Supervision. Minimum capital requirements for market risk. Basel iii document, Bank for International Settlements, 2019.

[2] Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematical Finance*, 9(3):203–228, 1999.

[3] Carlo Acerbi and Dirk Tasche. On the coherence of expected shortfall. *Journal of Banking & Finance*, 26(7):1487–1503, 2002.

[4] Tilmann Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011.

[5] Tobias Fissler and Johanna F. Ziegel. Higher order elicitability and Osband's principle. *The Annals of Statistics*, 44(4):1680–1707, 2016.

[6] Tim Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327, 1986.

[7] Robert F. Engle and Simone Manganelli. CAViaR: Conditional autoregressive value at risk by regression quantiles. *Journal of Business & Economic Statistics*, 22(4):367–381, 2004.

[8] Andrew J. Patton, Johanna F. Ziegel, and Rui Chen. Dynamic semiparametric models for expected shortfall (and value-at-risk). *Journal of Econometrics*, 211(2):388–413, 2019.

[9] Federico Gatta, Fabrizio Lillo, and Piero Mazzarisi. CAESar: Conditional autoregressive expected shortfall. *arXiv preprint arXiv:2407.06619*, 2024.

[10] David Barrera, Stéphane Crépey, Emmanuel Gobet, Hoang-Dung Nguyen, and Bouazza Saadeddine. Learning value-at-risk and expected shortfall. *arXiv preprint arXiv:2209.06476*, 2022.

[11] Fulvio Corsi. A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2):174–196, 2009.

[12] Roger Koenker and Gilbert Bassett. Regression quantiles. *Econometrica*, 46(1):33–50, 1978.

[13] Stefan Weber. Distribution-invariant risk measures, information, and dynamic consistency. *Mathematical Finance*, 16(2):419–441, 2006.

[14] Robert F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50(4):987–1007, 1982.

[15] Drew Creal, Siem Jan Koopman, and André Lucas. Generalized autoregressive score models with applications. *Journal of Applied Econometrics*, 28(5):777–795, 2013.

[16] Ulrich A. Müller, Michel M. Dacorogna, Rakhal D. Davé, Richard B. Olsen, Olivier V. Pictet, and Jacob E. von Weizsäcker. Volatilities of different time resolutions—analyzing the dynamics of market components. *Journal of Empirical Finance*, 4(2–3):213–239, 1997.

[17] Michael P Clements, Ana Beatriz Galvão, and Jae H Kim. Quantile forecasts of daily exchange rate returns from forecasts of realized volatility. *Journal of Empirical Finance*, 15(4):729–750, 2008.

[18] Filip Žikeš and Jozef Baruník. Semi-parametric conditional quantile models for financial returns and realized volatility. *Journal of Financial Econometrics*, 14(1):185–226, 2015.

[19] Federico Gatta, Fabrizio Lillo, and Piero Mazzarisi. A high-frequency approach to realized risk measures. *arXiv preprint arXiv:2510.16526*, 2025.

[20] James W. Taylor. Estimating value at risk and expected shortfall using expectiles. *Journal of Financial Econometrics*, 6(2):231–252, 2008.

[21] R. Tyrrell Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-risk. *Journal of Risk*, 2(3):21–42, 2000.

[22] Paul H. Kupiec. Techniques for verifying the accuracy of risk measurement models. *The Journal of Derivatives*, 3(2):73–84, 1995.

[23] Peter F. Christoffersen. Evaluating interval forecasts. *International Economic Review*, 39(4):841–862, 1998.

[24] Alexander J. McNeil and Rüdiger Frey. Estimation of tail-related risk measures for heteroscedastic financial time series: An extreme value approach. *Journal of Empirical Finance*, 7(3–4):271–300, 2000.

[25] Carlo Acerbi and Balázs Szekely. Back-testing expected shortfall. *Risk*, 27(11):76–81, 2014.

[26] Whitney K Newey and Daniel McFadden. Large sample estimation and hypothesis testing. In *Handbook of econometrics*, volume 4, pages 2111–2245. Elsevier, 1994.

[27] Francis X Diebold and Roberto S Mariano. Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263, 1995.

[28] Mike Tsionas, Aya Ghalayini, Marwan Izzeldin, and Lorenzo Trapani. A dynamic state-space har model. *Journal of Econometrics*, 2025. Forthcoming.