

7543758-do-regime-conditional-diffusion-models-improve-expected-shortfall-for-ecasts-evid

SUMMARY: This study examines whether diffusion-based generative scenario models, particularly when conditioned on market regimes, can improve Expected Shortfall (ES) forecasting in Asia-Pacific equity markets. The analysis uses daily data from six major markets—India, Japan, Hong Kong, South Korea, Australia, and Singapore—from 2007 to August 2026. Five approaches are compared: Historical Simulation (HS), regime-conditioned HS (Regime-HS), GARCH-filtered HS (GARCH-FHS), a diffusion-based scenario model, and a regime-conditioned diffusion model. The results show that incorporating Hidden Markov Model regime probabilities does not significantly improve diffusion-based ES forecasts. GARCH-FHS delivers the strongest numerical forecasting performance and tail-risk calibration, although formal pairwise tests do not establish its superiority over competing models at conventional significance levels. The findings suggest that greater generative flexibility and explicit regime conditioning do not necessarily translate into superior tail-risk forecasts, highlighting the importance of accurately modelling extreme distributional characteristics when applying generative models to financial risk management.

METHOD: The study uses daily data from six major Asian-Pacific equity markets (India, Japan, Hong Kong, South Korea, Australia, and Singapore) from 2007 to August 2026. Five approaches are compared: Historical Simulation (HS), regime-conditioned HS (Regime-HS), GARCH-filtered HS (GARCH-FHS), a diffusion-based scenario model, and a regime-conditioned diffusion model. The analysis evaluates 1,151 genuine out-of-sample forecasts between January 2020 and August 2026.

KEY FINDINGS:

- Incorporating Hidden Markov Model regime probabilities does not provide a statistically significant improvement in diffusion-based ES forecasts.
- GARCH-FHS delivers the strongest numerical forecasting performance and tail-risk calibration.
- Scenario-fidelity diagnostics indicate that the diffusion models underrepresent the negative skewness, excess kurtosis, and joint downside dependence observed in realized returns.
- Greater generative flexibility and explicit regime conditioning do not necessarily translate into superior tail-risk forecasts.

IMPLICATION FOR TRADING: The findings suggest that diffusion-based generative models, especially when conditioned on market regimes, can improve ES forecasting in Asia-Pacific equity markets. However, the superiority of these models over competing methods is not established at conventional significance levels. The study highlights the importance of accurately modelling extreme distributional characteristics when applying generative models to financial risk management, particularly in the context of market regimes and tail risks. This research can inform the development of more robust and accurate ES forecasting models for financial institutions, asset managers, and portfolio managers.

Do Regime-Conditional Diffusion Models Improve Expected Shortfall Forecasts? Evidence from Asia-Pacific Equity Markets

ARTICLE INFO

Keywords

Expected Shortfall; Diffusion Models; Market Regimes; Historical Simulation; Filtered Historical Simulation; Tail Risk; Asia-Pacific Equity Markets

JEL Classification

G17 — Financial Forecasting and Simulation

C53 — Forecasting and Prediction Methods

C58 — Financial Econometrics

G15 — International Financial Markets

Abstract

Historical Simulation (HS) remains widely used for market-risk estimation because of its transparency and limited distributional assumptions, but its reliance on past observations may restrict its ability to adapt to changing volatility, tail behaviour, and cross-market dependence. This study examines whether diffusion-based generative scenario models, particularly when conditioned on market regimes, improve Expected Shortfall (ES) forecasting in Asia-Pacific equity markets. The analysis uses daily data from six major markets—India, Japan, Hong Kong, South Korea, Australia, and Singapore—from 2007 to August 2026, with 1,151 genuine out-of-sample forecasts evaluated between January 2020 and August 2026. Five approaches are compared: Historical Simulation, regime-conditioned Historical Simulation (Regime-HS), GARCH-filtered Historical Simulation (GARCH-FHS), a diffusion-based scenario model, and a regime-conditioned diffusion model. The results show that incorporating Hidden Markov Model regime probabilities does not provide a statistically significant improvement in diffusion-based ES forecasts. GARCH-FHS delivers the strongest numerical forecasting performance and tail-risk calibration, although formal pairwise tests do not establish its superiority over competing models at conventional significance levels. Scenario-fidelity diagnostics further indicate that the diffusion models underrepresent the negative skewness, excess kurtosis, and joint downside dependence observed in realized returns. Overall, the findings suggest that greater generative flexibility and explicit regime conditioning do not necessarily translate into superior tail-risk forecasts, highlighting the importance of accurately modelling extreme distributional characteristics when applying generative models to financial risk management.

1. Introduction

1.1 Market tail risk and Expected Shortfall

Extreme market movements remain one of the central challenges in financial risk management. Periods of severe market stress can generate losses far beyond those implied by normal market conditions, while simultaneously altering volatility, liquidity, and dependence across assets and markets. For financial institutions, asset managers, and portfolio managers, the ability to estimate such losses is therefore not only a question of forecasting average market behaviour, but of accurately characterizing the conditional left tail of the return distribution.

Value-at-Risk (VaR) has historically been one of the most widely used measures of market risk. At a specified confidence level, VaR identifies a loss threshold that is expected to be exceeded only with a given probability. Its main limitation, however, is that it does not describe the magnitude of losses once that threshold has been crossed. This limitation contributed to the increasing use of Expected Shortfall (ES), which measures the expected loss conditional on losses occurring beyond the corresponding VaR threshold. ES therefore provides information about the severity of extreme losses and possesses desirable theoretical properties as a coherent measure of risk (Artzner et al., 1999; Acerbi and Tasche, 2002). Its practical importance has also increased through the Basel market-risk framework, which places 97.5% Expected Shortfall at the centre of the internal-model approach to market-risk measurement (Basel Committee on Banking Supervision, 2026).

Accurate ES estimation, however, is inherently difficult. Unlike measures focused on the centre of the distribution, ES depends on observations in a region where data are naturally scarce. Forecast performance consequently depends on how effectively a model captures time-varying volatility, asymmetric losses, heavy tails, and the possibility that several markets may experience large negative movements at the same time.

1.2 The Historical Simulation problem

Historical Simulation (HS) remains an attractive approach to market-risk measurement because it is intuitive, transparent, and largely non-parametric. Rather than imposing a particular distribution on returns, HS constructs the loss distribution from observed historical market movements. This allows the method to retain empirical features of past data and, in a multivariate setting, preserve the cross-market relationships embedded in historical observations.

These advantages are accompanied by important limitations. First, the number of extreme observations contained in any finite historical window is necessarily small, making tail estimates sensitive to a relatively limited set of past events. Second, the method depends entirely on the historical sample and may therefore adjust slowly when current volatility differs materially from the volatility prevailing over the estimation window. Pritsker (2006), among others, highlights the potential sluggishness of conventional Historical Simulation when conditional market risk changes rapidly.

A further problem arises because observations from fundamentally different market environments are pooled into the same empirical distribution. Calm-market observations may receive the same treatment as observations generated during periods of financial stress even though the underlying return process may differ substantially across those states. Finally, conventional HS can only reuse scenarios that have already occurred. It cannot naturally generate new combinations of shocks that are statistically plausible but absent from the observed historical record.

Conceptually, the distinction can be expressed as follows. Historical Simulation treats a future risk scenario as approximately a replay of a previously observed market scenario. A generative model, by contrast, seeks to draw future scenarios from an estimated conditional distribution learned from historical data. This distinction motivates the use of modern generative methods for financial scenario generation.

1.3 Generative scenario models and diffusion

Recent developments in generative modelling offer a potentially different approach to scenario-based risk measurement. Earlier work has shown that neural generative models can reproduce several stylized properties of financial time series and produce synthetic market observations beyond those contained in the original sample (Wiese et al., 2020). More recently, diffusion models have emerged as a powerful class of generative models and have been adapted to multivariate probabilistic time-series forecasting and conditional time-series generation (Ho et al., 2020; Rasul et al., 2021; Yuan and Qiao, 2024).

For market-risk applications, diffusion models are potentially attractive for several reasons. They can generate large numbers of synthetic scenarios, model multivariate distributions jointly, condition generation on current market information, and potentially capture nonlinear relationships that are difficult to represent through simple empirical resampling. In principle, these properties could be particularly valuable for Expected Shortfall, where the quality of the estimated extreme-loss distribution is critical.

However, the ability to generate flexible and visually realistic synthetic data does not necessarily imply superior tail-risk forecasting. A model may reproduce broad features of a return distribution while failing to reproduce the negative asymmetry, excess kurtosis, or simultaneous downside movements that determine portfolio losses in the extreme tail. **Better generative flexibility does not necessarily imply better tail-risk forecasts.** Whether diffusion-generated scenarios provide meaningful improvements over established market-risk methods is therefore an empirical question rather than a consequence of model complexity alone.

1.4 Why market regimes may matter

One potential way of improving conditional scenario generation is to explicitly account for market regimes. Financial returns are not characterized by constant behaviour over time. Periods of market stress are typically associated with higher volatility, larger negative returns, changing dependence structures, and more frequent simultaneous market declines. Regime-switching approaches provide a natural framework for representing such changes in market conditions (Hamilton, 1989). In international financial markets, regime-dependent dynamics have also been shown to affect correlations and asset-allocation decisions (Ang and Bekaert, 2002; Guidolin and Timmermann, 2007), while regime-switching methods have been applied directly to VaR and Expected Shortfall estimation (Taamouti, 2009).

These observations provide an economic rationale for conditioning a diffusion model on the prevailing market state. A scenario generator that recognizes whether the market is currently calm, normal, or stressed may generate a more appropriate conditional distribution than a model that ignores such state information.

There is, however, an important competing possibility. If a diffusion model already conditions on recent market returns and measures of realized volatility, estimated regime probabilities may summarize information that is already embedded in those continuous state variables. In that case, adding Hidden Markov Model (HMM) regime probabilities may provide little incremental forecasting information. The value of explicit regime conditioning therefore also requires direct out-of-sample testing.

1.5 Asia-Pacific empirical setting

This study examines these questions using six major Asia-Pacific equity markets: India, Japan, Hong Kong, South Korea, Australia, and Singapore. The region provides a useful empirical environment because it combines large and economically significant equity markets with considerable heterogeneity in market structure, economic development, investor composition, and exposure to global and regional shocks.

At the same time, these markets are sufficiently interconnected for joint downside behaviour to matter for diversified regional portfolios. International equity research has long shown that dependence can become stronger in adverse market conditions and that correlations observed during severe downturns may differ materially from those estimated under ordinary conditions (Longin and Solnik, 2001; Ang and Chen, 2002). This makes a multivariate Asia-Pacific portfolio a relevant setting in which to examine whether synthetic scenario models can reproduce the cross-market tail dependence required for accurate Expected Shortfall estimation.

The purpose is not to claim that the selected markets represent the entire Asia-Pacific region, but rather to evaluate the competing risk models across **six major Asia-Pacific equity markets** with long daily return histories and meaningful regional diversification.

1.6 Research gap

Existing research provides substantial evidence on Historical Simulation and volatility-filtered risk models, regime-dependent financial behaviour, generative financial time series, and diffusion-based probabilistic forecasting. These strands of literature, however, have largely developed separately. Comparatively limited evidence exists on whether regime-conditioned diffusion-generated multivariate scenarios translate into superior **out-of-sample Expected**

Shortfall forecasts relative to conventional historical and volatility-adaptive approaches, particularly in an Asia-Pacific setting.

This distinction is important. Demonstrating that a generative model can produce plausible synthetic financial data is not equivalent to demonstrating that it improves the forecasting of a specific tail-risk functional such as ES. Similarly, showing that markets exhibit regimes does not establish that explicit regime probabilities improve a model that already incorporates current returns and volatility. The present study connects these questions within a common forecasting framework.

1.7 Research questions

The empirical analysis is organized around three research questions.

RQ1: Do diffusion-generated scenarios improve 97.5% Expected Shortfall forecasts relative to conventional Historical Simulation for an Asia-Pacific equity portfolio?

RQ2: Does conditioning diffusion models on estimated market-regime probabilities provide incremental forecasting value?

RQ3: Can differences in Expected Shortfall performance be explained by the models' ability to reproduce negative skewness, fat tails, and cross-market downside dependence?

1.8 Contributions

The study makes three principal contributions.

First, it provides a controlled comparison of five distinct approaches to scenario-based market-risk forecasting: conventional Historical Simulation, regime-conditioned Historical Simulation, GARCH-filtered Historical Simulation, a diffusion-based scenario model, and a regime-conditioned diffusion model. This design allows the effects of historical resampling, volatility adaptation, regime information, and generative modelling to be evaluated within the same out-of-sample experiment.

Second, the comparison between the Diffusion and Regime-Diffusion specifications directly isolates the incremental value of explicit regime information. Because the baseline diffusion model already observes recent returns and volatility, the experiment tests whether HMM regime probabilities contribute forecasting information beyond those continuous market-state variables.

Third, the study connects forecast accuracy to **scenario fidelity** rather than simply ranking models according to a single performance statistic. In addition to Expected Shortfall performance, the analysis examines whether competing methods reproduce negative skewness, excess kurtosis, cross-market correlation, and joint lower-tail exceedances. This allows differences in risk forecasts to be interpreted in terms of the underlying characteristics of the generated scenario distributions.

1.9 Preview of findings

The empirical analysis uses daily observations from 2007 through August 2026 and evaluates 1,151 genuine out-of-sample forecasts between January 2020 and August 2026. The results provide limited support for the proposition that additional generative or regime complexity necessarily improves tail-risk forecasting. GARCH-filtered Historical Simulation produces the strongest numerical Expected Shortfall performance and the closest VaR calibration among the competing models, although formal pairwise inference does not establish its superiority over all alternatives at conventional significance levels.

More importantly, adding HMM regime probabilities to the diffusion model does not produce a statistically significant improvement relative to the corresponding non-regime diffusion specification. Scenario diagnostics help explain this result: the diffusion-generated return distributions exhibit substantially less negative skewness and excess kurtosis than the realized data and reproduce joint downside dependence less accurately than the strongest historical benchmark. These findings suggest that the main challenge for generative market-risk models is not simply generating more scenarios, but generating scenarios with sufficiently accurate behaviour in the extreme joint left tail.

1.10 Organization of the paper

The remainder of the paper is organized as follows. Section 2 reviews the literature on Expected Shortfall, Historical and Filtered Historical Simulation, regime-switching risk models, generative financial time series, diffusion methods, and cross-market tail dependence. Section 3 describes the Asia-Pacific dataset, return construction, portfolio definition, and preliminary empirical characteristics. Section 4 presents the regime model, the five competing risk-forecasting approaches, the out-of-sample design, and the forecast-evaluation methodology. Section 5 reports the main empirical results and examines the distributional and dependence properties of the generated scenarios. Section 6 presents robustness analyses and the multi-day stress-scenario extension. Section 7 discusses the economic and risk-management implications of the findings and the limitations of the study. Section 8 concludes.

2. Literature Review and Research Positioning

2.1 Value-at-Risk and Expected Shortfall

Market-risk measurement has traditionally relied heavily on Value-at-Risk (VaR), which summarizes the loss threshold that should not be exceeded over a specified horizon at a given confidence level. Its practical appeal lies in reducing portfolio risk to a single interpretable number. However, VaR provides no information about the magnitude of losses once the threshold has been exceeded. Two portfolios can therefore have the same VaR while exhibiting substantially different loss distributions beyond that quantile. (Engle and Manganelli, 2004; Berkowitz and O'Brien, 2002).

This limitation became particularly important in the development of coherent risk-measure theory. Artzner et al. (1999) formalized coherence through properties including monotonicity, translation invariance, positive homogeneity, and subadditivity. Quantile-based measures such as VaR need not satisfy subadditivity in general, implying that measured portfolio risk can, under some distributions, behave inconsistently with diversification.

Expected Shortfall (ES) addresses a different aspect of the loss distribution. Instead of reporting only a threshold, it measures the average loss conditional on being in the tail beyond the corresponding VaR level. Acerbi and Tasche (2002) show that, when appropriately defined, ES satisfies the coherence requirements and provides a theoretically stronger measure of extreme portfolio risk. This makes ES especially relevant when the research question concerns not merely the frequency of large losses but also their severity. (Rockafellar and Uryasev, 2000, 2002; McNeil and Frey, 2000).

The regulatory importance of ES has also increased substantially. Under the Basel market-risk framework, banks using the internal models approach calculate ES daily at a one-tailed 97.5% confidence level, with a 10-day base liquidity horizon forming part of the regulatory construction. This shift reflects the view that tail severity is central to market-risk capital assessment rather than the quantile threshold alone.

The move from VaR to ES, however, does not eliminate the empirical difficulty of tail-risk measurement. ES is estimated from observations lying beyond an already extreme quantile, which means that available information in the relevant region of the distribution is limited. Its evaluation is also more complicated than simple VaR exception counting. Later work on forecast evaluation establishes that although ES is not elicitable as a stand-alone functional, the pair consisting of VaR and ES is jointly elicitable, allowing competing forecasts to be compared using consistent joint scoring functions (Fissler and Ziegel, 2016; Fissler et al., 2016). (Gneiting, 2011; Acerbi and Szekely, 2014; Bayer and Dimitriadis, 2022; Patton et al., 2019).

These developments have an important implication for the present study. Accurate ES forecasts require more than an accurate estimate of unconditional volatility. They depend on the quality of the estimated **conditional extreme-loss distribution**, including its asymmetry, fat tails, and dependence across assets. ES accuracy therefore depends critically on modelling the conditional distribution of extreme losses.

2.2 Historical Simulation and Filtered Historical Simulation

Historical Simulation (HS) is one of the most intuitive approaches to market-risk estimation. Instead of specifying a parametric distribution for returns, it uses the empirical distribution of historical shocks. This gives HS several practical advantages. It is transparent, relatively simple to implement, and does not require the researcher to assume that returns follow a Gaussian or other pre-specified distribution. In a multivariate application, historical observations also preserve the contemporaneous relationships that actually occurred across assets.

These advantages explain why HS has remained important in institutional risk management. Yet the same empirical structure that makes the method transparent also produces important limitations. Each historical observation is effectively drawn from a finite window, so estimates of very high confidence-level tail risk depend on relatively few extreme observations. In addition, conventional HS usually gives similar status to shocks generated under very different volatility conditions. A return observed in a calm market may enter the same empirical distribution as a return observed during a crisis.

Pritsker (2006) emphasizes that apparently straightforward historical-simulation methods can have undesirable risk-measurement properties, particularly when market risk changes over time. A fixed historical window may respond only gradually to rapidly changing current conditions, while the sample itself can contain observations drawn from different volatility environments.

Filtered Historical Simulation (FHS) was developed to address this limitation while retaining the empirical appeal of historical resampling. Rather than resampling raw returns directly, FHS first filters them by an estimate of conditional volatility. Historical shocks are converted into standardized residuals and then rescaled using the volatility forecast relevant for the current period. In effect, the method separates the **shape of historical innovations** from their historical volatility level. (Boudoukh et al., 1998).

Hull and White (1998) proposed volatility updating within Historical Simulation using GARCH or exponentially weighted volatility estimates, adjusting historical market movements to reflect the difference between past and current volatility. Barone-Adesi et al. (1999) similarly proposed filtering historical simulation through GARCH processes, specifically noting that ordinary HS assigns equal probability to past returns without accounting sufficiently for current market conditions.

The GARCH component is important because financial returns display conditional heteroskedasticity: periods of large movements tend to cluster together. By standardizing shocks with historical conditional volatility and applying current forecast volatility, GARCH-FHS allows a historical shock originating in a low-volatility period to be translated into a larger current shock when prevailing volatility is high, and vice versa. (Engle, 1982; Bollerslev, 1986; Nelson, 1991; Glosten et al., 1993).

FHS is therefore a particularly relevant benchmark for this study. It retains an empirical innovation distribution while explicitly adapting to changing volatility. This provides a stronger comparison than conventional HS alone. If a complex generative model cannot outperform a method that combines historical residuals with dynamic volatility scaling, the result is informative about whether generative flexibility genuinely adds value for tail-risk forecasting rather than merely improving upon a relatively static benchmark.

2.3 Regime-Dependent Financial Risk

Volatility filtering assumes that risk changes continuously, but financial markets may also display more discrete changes in behaviour. Regime-switching approaches model this possibility by allowing returns to arise from different latent states. Hamilton's regime-switching framework provided a general basis for modelling economic processes in which the underlying state is not directly observed but evolves probabilistically over time. (Gray, 1996; Ang and Timmermann, 2012).

In financial applications, these states are often interpreted economically as calm, intermediate, and stressed environments. The relevance of such regimes extends beyond differences in variance.

International equity markets may exhibit different expected returns, volatilities, and correlations across states. Ang and Bekaert (2002), for example, show that international asset allocation can be materially affected by regime-dependent investment opportunities, including higher correlations and volatilities in adverse states.

Regime-switching methods have also been applied directly to tail-risk measurement. Taamouti (2009) develops VaR and ES measures under regime switching and emphasizes the ability of such models to accommodate features such as volatility clustering and heavy-tailed risk that Gaussian specifications may underestimate.

Hidden Markov Models (HMMs) provide a particularly useful formulation because the regime itself is latent. Instead of assuming that an observed indicator perfectly labels the market as calm or stressed, an HMM estimates the probability that each regime is active given currently available information. This permits uncertainty about the prevailing state to be retained rather than forcing a deterministic classification.

The potential benefit for scenario generation is straightforward:

Regime information $\rightarrow$
better identification of the current
market state $\rightarrow$
better conditional tail forecasts.

If extreme losses arise from a different market state than ordinary observations, explicitly conditioning simulations on state probabilities may improve the relevance of generated scenarios.

There is, however, a competing mechanism. Recent returns and realized volatility are themselves strong indicators of market state. A flexible model that already receives these continuous variables may infer much of the same information captured by the discrete HMM probabilities. In such a setting,

returns + volatility

may already encode much of what the regime variable summarizes. Explicit regime conditioning could therefore be informative in a simpler empirical-resampling model but largely redundant in a flexible conditional generative model. Distinguishing between these mechanisms requires a controlled comparison rather than assuming *ex ante* that regime conditioning must improve forecasts.

2.4 Generative Models in Financial Time Series

The rise of generative machine learning created a different approach to scenario simulation. Instead of resampling existing observations, generative models attempt to learn an underlying data distribution from which new synthetic observations can be drawn. Variational autoencoders introduced latent-variable learning based on probabilistic encoding and decoding, while generative adversarial networks (GANs) formulated generation as an adversarial learning problem between a generator and discriminator (Kingma and Welling, 2014; Goodfellow et al., 2014).

In time-series applications, the challenge is not only to match marginal distributions but also to preserve temporal relationships. TimeGAN, for example, combines adversarial and supervised components specifically to improve the reproduction of temporal dynamics in generated sequences (Yoon et al., 2019).

Financial applications have similarly explored whether deep generative models can reproduce stylized properties of market data. Wiese et al. (2020) develop Quant GANs for deep generation of financial time series, illustrating the potential for synthetic return generation that goes beyond replaying historical observations. (Liao et al., 2024).

For risk management, the attraction is clear: generative methods may produce a large number of previously unseen but data-consistent scenarios. The central question, however, is whether synthetic realism measured over the overall distribution also translates into sufficient realism in the small portion of the distribution that determines ES.

2.5 Diffusion Models for Time-Series Forecasting and Simulation

Diffusion models represent a more recent class of generative methods. In the denoising diffusion probabilistic framework of Ho et al. (2020), the forward process gradually corrupts data by adding

noise, while a learned reverse process attempts to reconstruct samples by sequentially removing that noise. Generation therefore begins from noise and proceeds through a sequence of denoising steps toward observations drawn from the learned distribution.

The approach has several features that make it attractive for financial scenario generation. It provides a flexible probabilistic mechanism, avoids some of the adversarial-training difficulties associated with GANs, and can naturally be adapted to conditional generation. More importantly for portfolio risk, diffusion methods can be designed to generate multivariate outcomes jointly rather than modelling each market independently. (Nichol and Dhariwal, 2021; Song et al., 2021).

TimeGrad provides an early example of applying denoising diffusion methods to multivariate probabilistic time-series forecasting. Rasul et al. (2021) construct an autoregressive diffusion approach that generates multivariate forecasts by sampling from the conditional distribution at successive time steps.

More recently, Diffusion-TS extends diffusion-based generation to general multivariate time-series tasks and demonstrates that diffusion models can support conditional generation, forecasting, and imputation within a unified framework (Yuan and Qiao, 2024). (Tashiro et al., 2021).

These developments suggest several theoretical advantages for market-risk analysis. A diffusion model can generate large scenario sets, accommodate multiple correlated risk factors, condition on observable market information, and represent distributions that need not belong to a simple parametric family. Such flexibility could, in principle, allow the model to capture nonlinear and state-dependent behaviour missed by conventional approaches.

Yet this flexibility creates an important evaluation problem. Generative models are frequently assessed through measures of overall distributional resemblance, sample realism, average forecast quality, or predictive usefulness. Market-risk forecasting places a much more specific demand on the model. For ES, a model must reproduce the **extreme joint left tail** accurately. A generator may match the centre of the distribution and even unconditional volatility while producing too little negative skewness, insufficient kurtosis, or too few simultaneous downside events. Thus, realistic synthetic data in a broad statistical sense need not imply reliable tail-risk forecasts.

2.6 International Equity Dependence and Tail Dependence

The multivariate dimension is particularly important for international portfolios. Portfolio ES is determined not only by extreme losses in individual markets but also by whether those losses occur simultaneously. Standard Pearson correlation summarizes average linear dependence and may therefore conceal substantial differences between normal and stressed conditions. (Poon et al., 2004; Patton, 2006).

Longin and Solnik (2001) show that dependence in international equity markets is asymmetric across extreme market conditions: correlations rise particularly in bear markets rather than uniformly with volatility. Ang and Chen (2002) similarly document stronger correlations for downside equity movements than for comparable upside movements.

This distinction is central to tail-risk measurement. For two markets i and j , the probability

$$P(r_i < q_i, r_j < q_j)$$

captures the likelihood that both returns simultaneously fall below their respective downside thresholds q_i and q_j . For portfolio ES, this quantity may be more informative than unconditional correlation because simultaneous extreme declines directly magnify aggregate portfolio losses.

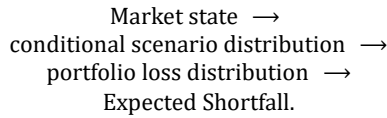
This issue is particularly relevant in internationally connected markets, where common global and regional shocks can generate dependence patterns that differ sharply between ordinary and stressed conditions. Recent research in *Asia-Pacific Financial Markets* continues to emphasize tail dependence and state-dependent interconnectedness as distinct empirical phenomena rather than relying solely on average correlations. For example, Mar'i and Seraj (2025) analyze tail dependence and lead-lag relationships across

financial markets in bullish and bearish states, illustrating the continuing relevance of asymmetric dependence for diversification and risk assessment. (Forbes and Rigobon, 2002; Bae et al., 2003; Bekaert et al., 2005; Chiang et al., 2007; Fong and Koh, 2002; Su et al., 2025).

For the present study, this literature implies that scenario quality must be evaluated at two levels: the model must reproduce the marginal tails of individual markets and the joint occurrence of downside events across markets.

2.7 Synthesis and Research Positioning

The preceding literature suggests a common chain linking market conditions to portfolio tail risk:



Historical Simulation relies primarily on the empirical distribution of past scenarios. GARCH-FHS modifies those scenarios to reflect current conditional volatility. Regime-based approaches introduce discrete information about latent market states. Generative methods instead attempt to learn a flexible conditional distribution from which new multivariate observations can be sampled. Diffusion models combine this generative flexibility with conditional scenario generation.

The unresolved issue is whether these additional layers of complexity improve the particular part of the return distribution that matters for ES. In particular, this study asks whether explicit HMM regime information

adds predictive value beyond the information already contained in continuous returns and volatility state variables.

The empirical design therefore compares conventional Historical Simulation, regime-conditioned Historical Simulation, GARCH-filtered Historical Simulation, a diffusion model conditioned on continuous market information, and an otherwise comparable diffusion model augmented with HMM regime probabilities.

The contribution is not to assume that the most flexible model should dominate. Instead, the study evaluates whether additional state information and generative flexibility translate into superior out-of-sample ES forecasts and then examines whether differences in forecast performance can be traced to the quality of the generated marginal tails and cross-market downside dependence. This positioning connects the literatures on coherent risk measurement, volatility adaptation, regime switching, generative time-series modelling, and international tail dependence within a single market-risk forecasting framework. (Brauneis and Sahiner, 2026).

3. Data and Preliminary Analysis

3.1 Market Universe

The empirical analysis is based on six major Asia-Pacific equity markets: India, Japan, Hong Kong, South Korea, Australia, and Singapore. These markets were selected to provide a diversified regional sample comprising both developed and emerging equity markets, while also offering sufficiently long daily histories for out-of-sample risk forecasting. Broad market indices are used rather than individual securities so that the analysis captures market-level risk and cross-country dependence rather than firm-specific effects.

Table 1A. Market Universe

Country/Region	Equity Index
India	Nifty 50
Japan	Nikkei 225
Hong Kong	Hang Seng Index
South Korea	KOSPI
Australia	S&P/ASX 200
Singapore	Straits Times Index

These six indices form the basis of both the multivariate scenario-generation models and the equal-weight Asia-Pacific portfolio used for Expected Shortfall estimation. The objective is not to construct an optimal or directly investable portfolio, but to create a transparent and consistent setting in which differences in risk forecasts can primarily be attributed to the competing modelling approaches.

3.2 Data Source and Sample Period

Daily closing index levels were obtained from Yahoo Finance through the yfinance interface. The requested sample period extends from 1 January 2007 to 31 August 2026.

Actual data availability differs slightly across the six indices. Japan, Hong Kong, South Korea, Australia, and Singapore have observations available from the beginning of 2007, whereas the Nifty 50 series begins on 17 September 2007. After constructing returns and aligning the six markets, the first usable common one-session target observation occurs on 17 October 2007.

The resulting matched multivariate dataset contains 3,258 common target observations through 31 August 2026.

The empirical design separates model development from final forecast evaluation. The final out-of-sample period runs from **7 January 2020 to 31 August 2026** and contains **1,151 genuine out-of-sample forecasts**. These forecasts form the basis of the principal Value-at-Risk and Expected Shortfall comparisons reported later in the paper. Earlier observations are used for model estimation, regime identification, and pre-test model development according to the chronological framework described in Section 4.

3.3 Trading-Calendar Alignment

A practical issue in a multivariate Asia-Pacific study is that the six exchanges do not operate on identical trading calendars. National holidays, exchange-specific closures, and differences in local trading schedules imply that one market may be closed on a date when other markets remain open.

Treating such non-trading days as zero-return observations could artificially reduce measured volatility and distort estimated cross-market dependence. To avoid this problem, daily returns are first calculated independently on each market's native trading calendar.

No artificial zero return is inserted merely because an exchange is closed on a particular calendar date. Once native-session returns have been calculated, the individual return series are aligned to construct the common multivariate dataset required for joint modelling.

This procedure preserves actual consecutive trading-session movements while avoiding distortions that could arise from mechanically forward-filling prices across exchange holidays.

Because trading calendars differ across countries, the number of native-session observations varies across the six indices. The final native-session samples range from 4,649 observations for the Nifty 50 to 4,965 observations for the S&P/ASX 200. The common multivariate sample is smaller because an observation is retained only when the information required for the six-market modelling framework is jointly available.

3.4 Return Construction and Preprocessing

For equity market i , the continuously compounded daily return at time t is defined as

$$r_{i,t} = \ln \left(\frac{P_{i,t}}{P_{i,t-1}} \right),$$

where $r_{i,t}$ denotes the log return of market i at time t , $P_{i,t}$ denotes the closing index level at time t , and $P_{i,t-1}$ denotes the closing index level observed during the preceding trading session.

The raw price series were examined for missing observations, duplicate dates, invalid price values, and unusually large one-session movements before model estimation.

Extreme observed returns were not winsorized or removed solely because of their magnitude. This treatment is important because the objective of the study is specifically to evaluate tail-risk forecasting. Removing genuine extreme returns would mechanically reduce the empirical tails of the return distributions and weaken the very behaviour that the competing models are expected to capture.

For model components requiring standardized inputs, the transformation is performed using parameters estimated only from the relevant training sample. For a variable x_t , the standardized value is therefore defined as

$$z_t = \frac{x_t - \mu_{\text{train}}}{\sigma_{\text{train}}},$$

where μ_{train} and σ_{train} denote the mean and standard deviation estimated using only observations available within the training period.

This procedure avoids look-ahead bias because observations from the held-out evaluation period are not used to scale earlier model inputs.

3.5 Portfolio Construction

The principal risk portfolio is constructed as an equally weighted combination of the six equity-index returns.

Table 1. Descriptive Statistics of Daily Index Returns

Market	Observations	Mean (%)	SD (%)	Skewness	Excess Kurtosis	Minimum (%)	Maximum (%)	1% Quantile (%)	2.5% Quantile (%)	5% Quantile (%)
Nifty 50	4,649	0.036	1.295	-0.28	15.16	-13.90	16.33	-3.62	-2.54	-1.85
Nikkei 225	4,805	0.028	1.504	-0.45	8.25	-13.23	13.23	-4.13	-3.03	-2.30
Hang Seng	4,839	0.005	1.535	-0.11	8.22	-14.18	13.41	-4.27	-3.02	-2.32
KOSPI	4,841	0.032	1.428	-0.43	13.94	-12.85	16.48	-4.44	-2.94	-1.97
S&P/ASX 200	4,965	0.009	1.063	-0.67	7.74	-10.20	6.77	-3.06	-2.21	-1.65
Straits Times	4,913	0.013	1.028	-0.36	8.62	-8.70	7.53	-3.15	-2.05	-1.50

Several features of the data are particularly relevant for the subsequent tail-risk analysis.

First, all six markets exhibit negative skewness. This indicates that their return distributions are asymmetric and that large negative movements tend to be more pronounced than would be expected under a symmetric distribution.

Second, excess kurtosis is high across all markets. It ranges from 7.74 for the S&P/ASX 200 to 15.16 for the Nifty 50. Since a Gaussian distribution has excess kurtosis equal to zero, these values indicate substantial fat-tailed behaviour.

Third, unconditional volatility differs materially across markets. The Hang Seng Index and Nikkei 225 exhibit the highest daily standard deviations, whereas the Straits Times Index and S&P/ASX 200 display comparatively lower volatility.

Extreme one-session movements are also economically significant. The minimum daily return ranges from approximately

Let the portfolio weight assigned to each market be

$$w_i = \frac{1}{6}, i = 1, \dots, 6.$$

The portfolio return at time t is therefore

$$R_{p,t} = \sum_{i=1}^6 w_i r_{i,t}.$$

Under equal weighting, this becomes

$$R_{p,t} = \frac{1}{6} \sum_{i=1}^6 r_{i,t}.$$

Portfolio loss is defined as the negative of portfolio return:

$$L_t = -R_{p,t}.$$

Accordingly, a negative portfolio return corresponds to a positive portfolio loss.

An equal-weight portfolio is deliberately employed instead of an optimized portfolio. This choice is intended to ensure that measured performance primarily reflects the quality of the risk models rather than portfolio-construction assumptions. Introducing optimized portfolio weights would add another layer of estimation and could make it more difficult to determine whether differences in Expected Shortfall forecasts originate from the risk model or from the allocation methodology.

A more concentrated portfolio specification is considered separately as a robustness exercise.

3.6 Descriptive Statistics

Table 1 reports descriptive statistics for the native-session daily log returns of the six equity indices. Return-based statistics are expressed in percentage terms, while skewness and excess kurtosis are unit-free.

−8.70% for Singapore to −14.18% for Hong Kong. Similarly, the empirical 1% downside quantiles range from approximately −3.06% for Australia to −4.44% for South Korea.

These results demonstrate clear departures from Gaussian return behaviour. Negative skewness, high excess kurtosis, and large downside quantiles are particularly relevant for Expected Shortfall because the quality of an ES forecast depends on the model's ability to represent the extreme left tail of the loss distribution.

3.7 Preliminary Market Behaviour and Dependence

Figure 1 presents the normalized evolution of the six equity indices over the sample period. The series display substantial differences in long-term market performance, together with several episodes in which multiple markets experience simultaneous draw-downs.

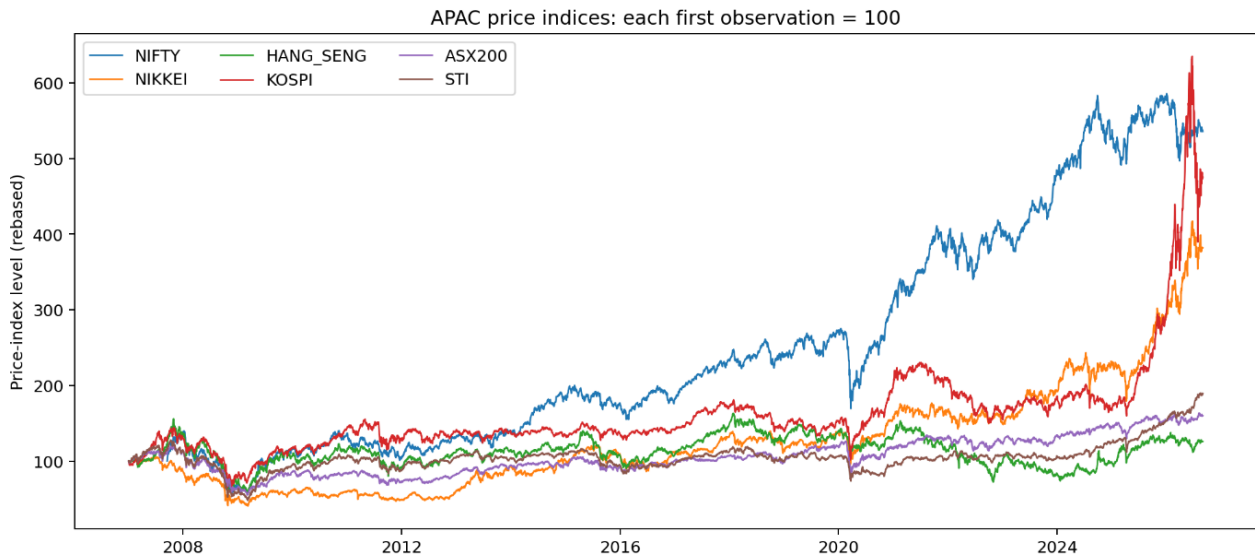


Figure 1. Normalized Performance of the Six Asia-Pacific Equity Indices

Figure 2 presents trailing 20-session volatility across the six underlying equity markets.

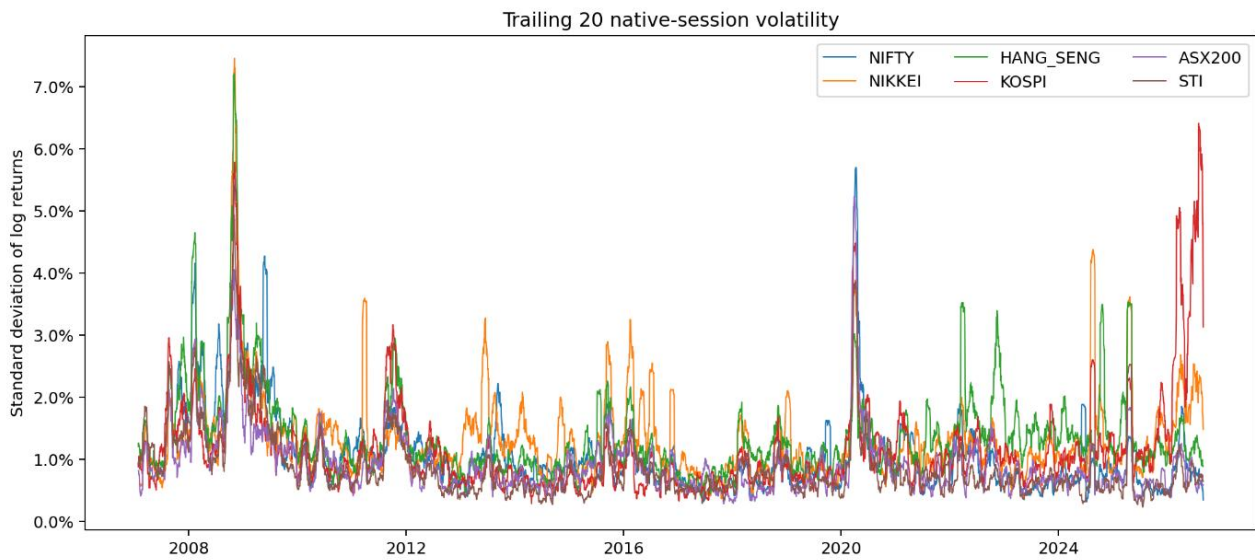


Figure 2. Twenty-Session Rolling Volatility across Asia-Pacific Equity Markets

The figure shows pronounced volatility clustering. Periods of elevated volatility tend to occur in persistent episodes rather than as isolated observations. Several episodes of high volatility also occur contemporaneously across multiple markets, although their magnitude and persistence differ across countries. This behaviour provides

direct empirical motivation for volatility-adaptive approaches such as GARCH-filtered Historical Simulation and for the latent market-regime framework introduced in Section 4.

Figure 3 presents the correlation matrix for the matched one-session returns.

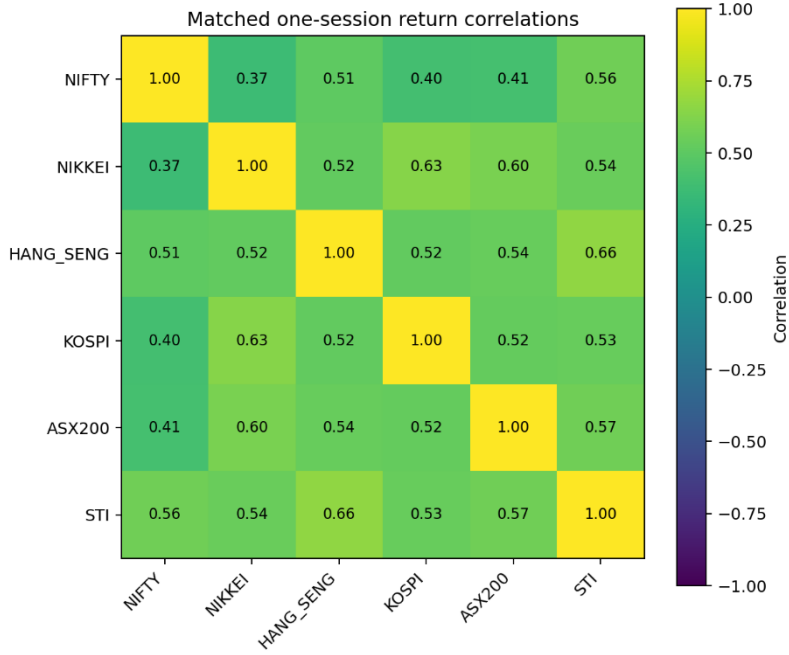


Figure 3. Cross-Market Return Correlation Matrix

Pairwise correlations are positive across all six markets but vary materially in magnitude. The correlation between the Nifty 50 and Nikkei 225 is approximately 0.37, while the correlation between the Hang Seng Index and the Straits Times Index is approximately 0.66. The Nikkei 225 and KOSPI also exhibit relatively strong dependence, with a correlation of approximately 0.63. Most remaining market pairs exhibit correlations in the range of approximately 0.50 to 0.60.

The evidence therefore indicates meaningful regional integration without suggesting that the six markets follow a homogeneous return process.

Taken together, the preliminary analysis identifies four features of direct relevance for the subsequent empirical investigation. First, volatility changes substantially through time. Second, the return distributions are negatively skewed. Third, all six markets exhibit pronounced excess kurtosis and fat tails. Fourth, cross-market dependence is positive but heterogeneous.

These characteristics create a demanding environment for market-risk estimation and provide the empirical motivation for comparing Historical Simulation, volatility-filtered methods, explicit regime conditioning, and diffusion-based scenario generation in forecasting the conditional tail of the Asia-Pacific portfolio loss distribution.

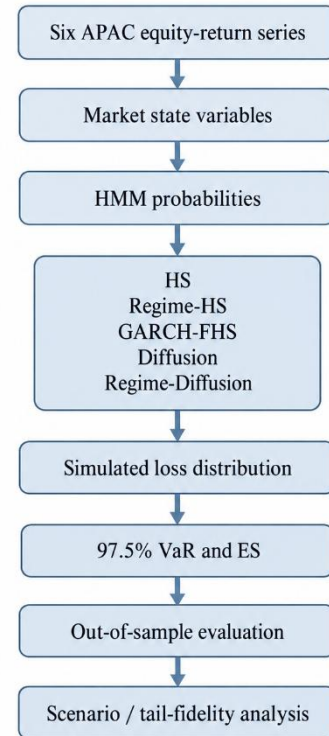
4. Empirical Methodology

4.1 Research Design

The empirical design is structured to isolate three distinct sources of potential improvement in market-risk forecasting: the use of historical versus generated scenarios, the incorporation of time-varying volatility, and the addition of explicit market-regime information. Five models are therefore evaluated within a common out-of-sample framework: Historical Simulation (HS), Regime-conditioned Historical Simulation (Regime-HS), GARCH-Filtered Historical Simulation (GARCH-FHS), a non-regime conditional diffusion model (Diffusion), and a regime-conditioned diffusion model (Regime-Diffusion).

All models ultimately produce a conditional distribution of one-session portfolio losses. From this distribution, 97.5% Value-at-Risk and Expected Shortfall forecasts are calculated and compared with the realized portfolio loss. The primary experiment therefore evaluates whether increasingly flexible approaches to scenario generation translate into more accurate tail-risk forecasts.

Figure 4. Empirical Research Framework



Note: HMM denotes Hidden Markov Model; VaR denotes Value-at-Risk; ES denotes Expected Shortfall.

4.2 Hidden Markov Regime Model

Market regimes are estimated using a three-state Gaussian Hidden Markov Model. The purpose of the HMM is not to forecast returns directly, but to summarize the prevailing market environment in a probabilistic form that can subsequently be used by Regime-HS and Regime-Diffusion.

The state model uses two observable variables derived from the six equity markets. The first is the contemporaneous equal-weight basket return. Because the individual series are modelled as log returns, the corresponding simple return for market i is

$$g_{i,t} = e^{r_{i,t}} - 1.$$

The synchronized equal-weight basket return used in the HMM is therefore

$$R_{p,t} = \frac{1}{6} \sum_{i=1}^6 g_{i,t}.$$

The second state variable is a volatility proxy based on the individual markets' trailing 20-session realized volatilities. Let $\sigma_{i,t}^{(20)}$ denote the 20-session rolling standard deviation of the log returns of market i . The cross-market volatility proxy is constructed as

$$\sigma_{20,t} = \sqrt{\frac{1}{6} \sum_{i=1}^6 (\sigma_{i,t}^{(20)})^2}.$$

The HMM is estimated using the standardized vector

$$x_t = \begin{bmatrix} 100R_{p,t} \\ \ln(\sigma_{20,t}) \end{bmatrix}.$$

A Gaussian emission distribution with diagonal covariance is assumed within each latent state. The three latent states are ordered ex post according to their estimated volatility coordinate and labelled **Calm**, **Normal**, and **Stress**, with **Calm** corresponding to the lowest conditional volatility and **Stress** to the highest.

Let S_t denote the unobserved regime at time t . Rather than assigning only a deterministic state label, the model retains the complete filtered probability vector

$$\pi_t = \begin{bmatrix} P(S_t = C | \mathcal{F}_t) \\ P(S_t = N | \mathcal{F}_t) \\ P(S_t = S | \mathcal{F}_t) \end{bmatrix},$$

where C , N , and S represent **Calm**, **Normal**, and **Stress** states, respectively, and $\mathcal{F}_t$ denotes information available through the forecast origin.

A critical design choice is the use of **filtered rather than smoothed regime probabilities**. Smoothed probabilities incorporate information observed after date t and would therefore introduce look-ahead bias if used as forecasting inputs. Filtered probabilities are obtained through the forward recursion only and depend exclusively on information available at the time the risk forecast is produced.

The HMM is re-estimated annually using observations available before the beginning of each forecast year. Three independent initialization attempts are used for estimation, and the highest-likelihood converged solution is retained.

4.3 Choice of Three Regimes

The three-state specification was selected **ex ante** because it provides an economically interpretable distinction between **Calm**, **Normal**, and **Stress** market conditions and avoids unnecessarily fragmenting an already limited number of extreme observations.

The baseline three-state model should not, however, be interpreted as the specification mechanically preferred by an information criterion. As part of the subsequent robustness analysis, two-, three-, and four-state Gaussian HMMs are compared. The four-state specification produces lower Bayesian Information Criterion values across the annual estimation windows. Accordingly, the study does not claim that three states provide the statistically optimal in-sample representation of the regime process.

The central reason for retaining the three-state specification in the primary experiment is that the regime structure was fixed before evaluation of the held-out test results and provides economically transparent labels for the forecasting exercise. The robustness analysis subsequently examines whether increasing the number of latent states improves risk forecasting. Importantly, the statistical fit of the regime model and the quality of the resulting Expected Shortfall forecasts are treated as separate questions.

4.4 Historical Simulation

Historical Simulation serves as the primary non-parametric benchmark. At each forecast origin t , the method uses the most recent **1,000 eligible historical observations** available before the

forecast date. No future observation is allowed to enter the scenario window.

Let L denote portfolio loss and let $F_{L,t}$ denote the empirical conditional distribution constructed at forecast origin t . For confidence level α , Value-at-Risk is defined as

$$VaR_{\alpha,t} = F_{L,t}^{-1}(\alpha).$$

Expected Shortfall is the average loss in the upper tail beyond the corresponding VaR threshold:

$$ES_{\alpha,t} = E[L | L \geq VaR_{\alpha,t}].$$

The main analysis uses

$$\alpha = 0.975.$$

Thus, the principal risk measure is 97.5% Expected Shortfall.

In the empirical implementation, ES is calculated from the empirical quantile function rather than by simply averaging observations strictly greater than VaR. Fractional probability mass at the VaR threshold is included where necessary, ensuring that the estimate remains well-defined in finite and weighted scenario distributions.

The 1,000-observation window represents the primary specification. Alternative 500- and 1,500-observation windows are examined as robustness checks.

4.5 Regime-Conditioned Historical Simulation

Regime-HS modifies standard Historical Simulation by assigning greater weight to historical observations whose estimated regime composition resembles the regime probabilities prevailing at the current forecast origin.

Let

$$\pi_s = (\pi_{s,1}, \pi_{s,2}, \pi_{s,3})^\top$$

denote the filtered HMM probability vector associated with historical observation s , and let

$$\pi_t = (\pi_{t,1}, \pi_{t,2}, \pi_{t,3})^\top$$

denote the regime probabilities at the current forecast origin.

Conceptually, the regime-conditioned distribution can be written as

$$F_t^{RHS} = \sum_{k=1}^3 \pi_{t,k} F_k,$$

where F_k denotes the historical loss distribution associated with regime k .

Operationally, historical observations receive probability weights according to their probabilistic membership in each regime and the current probability of that regime. A small 5% shrinkage toward uniform weights is applied to reduce instability when one regime contains relatively little historical probability mass.

This model serves an important identification purpose. A comparison of Regime-HS with HS measures whether explicit market-state information improves an otherwise historical resampling framework. Similarly, the later comparison of Regime-Diffusion with Diffusion isolates the incremental contribution of the same regime information within a generative setting.

4.6 GARCH-Filtered Historical Simulation

GARCH-FHS is included as a main benchmark because it provides a well-established mechanism for adapting empirical historical shocks to current volatility conditions without requiring a neural generative model. (Hull and White, 1998; Barone-Adesi et al., 1999).

The portfolio return process is modelled using a GARCH(1,1) volatility specification,

$$R_{p,t} = \mu + \varepsilon_t,$$

with

$$\varepsilon_t = \sigma_t z_t,$$

and conditional variance

$$\sigma_t^2 = \omega + \alpha_G \varepsilon_{t-1}^2 + \beta_G \sigma_{t-1}^2.$$

A Student- t likelihood is used when estimating the GARCH parameters to accommodate heavy-tailed innovations.

After estimating the conditional volatility process, historical standardized residuals are calculated as

$$z_t = \frac{R_{p,t} - \mu}{\sigma_t}.$$

The residual pool is standardized to zero mean and unit variance. The one-step-ahead conditional volatility forecast is then

$$\sigma_{t+1} = \sqrt{\omega + \alpha_G \varepsilon_t^2 + \beta_G \sigma_t^2}.$$

Historical residuals from the most recent 1,000 observations are rescaled by this forecast volatility to construct the conditional return scenarios:

$$R_{p,t+1}^{(s)} = \mu + \sigma_{t+1} z^{(s)}.$$

Although a Student- t likelihood is used to estimate the GARCH parameters, forecast innovations are drawn from the **empirical standardized residual pool**, rather than from newly simulated Student- t innovations. This preserves the historical innovation distribution while adapting its scale to current volatility.

GARCH-FHS therefore provides a strong conventional comparison: it explicitly addresses volatility clustering and current market conditions while remaining substantially more parsimonious than the diffusion models.

4.7 Non-Regime Conditional Diffusion Model

The baseline diffusion specification is referred to as **Diffusion** or **Non-Regime Conditional Diffusion**. It should not be interpreted as an unconditional generative model, because scenario generation is explicitly conditioned on observable market information.

At each forecast origin, the conditioning vector contains the most recently observed log return for each of the six markets together with the logarithm of each market's trailing 20-session volatility. The resulting baseline conditioning vector therefore contains 12 observable state variables.

Let

$$r_{t+1} = (r_{1,t+1}, \dots, r_{6,t+1})^\top$$

denote the six-dimensional next-session log-return vector. The model seeks to approximate the conditional distribution

$$p_\theta(r_{t+1} | c_t),$$

where c_t contains recent returns and volatility measures.

The forward diffusion process gradually corrupts the standardized target return vector with Gaussian noise. At diffusion step m , (Ho et al., 2020; Song et al., 2021).

$$x_m = \sqrt{\bar{\alpha}_m} x_0 + \sqrt{1 - \bar{\alpha}_m} \epsilon,$$

where x_0 is the standardized return vector, $\epsilon \sim N(0, I)$, and $\bar{\alpha}_m$ follows a cosine noise schedule.

A residual multilayer perceptron learns the corresponding reverse denoising process. The implemented model uses velocity parameterization and minimizes mean-squared prediction error during training. The main diffusion process contains 100 noise steps, while scenario generation uses 20 deterministic DDIM sampling steps.

The neural architecture contains two residual blocks. Hidden-layer width is selected from candidate values of 64 and 128 using only the 2017–2019 validation period. The selected width is 128. Optimization uses AdamW with a learning rate of 0.001, weight decay of 0.0001, batch size of 128, and a maximum of 150 epochs. Early stopping is applied with a patience of 15 epochs. Complete architecture and training details are provided in Appendix B.

4.8 Regime-Conditional Diffusion Model

Regime-Diffusion uses the same architecture and continuous conditioning variables as the baseline Diffusion model but augments the conditioning vector with the three filtered HMM probabilities.

Its conditional distribution is therefore

$$p_\theta(r_{t+1}, | c_t \pi_t),$$

where c_t contains recent returns and volatility measures and π_t contains the Calm, Normal, and Stress probabilities.

All other major architectural and training choices are held common across the two diffusion specifications. Hyperparameters are selected jointly using pre-test validation data rather than separately optimizing the regime model after observing its test-period performance.

The comparison between Diffusion and Regime-Diffusion provides a direct test of whether explicit regime probabilities add forecasting information beyond the continuous return and volatility state variables.

4.9 Model Training, Validation, and Out-of-Sample Design

The empirical design preserves chronological separation between model development and final evaluation.

The historical sample prior to 2017 provides the initial estimation base. The period from **2017 through 2019** is reserved for model-development and hyperparameter validation. Candidate neural widths are evaluated during this period using mean 97.5% FZ0 performance across both diffusion specifications. The same selected architecture is then frozen for the final experiment.

The final held-out evaluation begins in January 2020 and extends through 31 August 2026. Because the six markets operate on different trading calendars and require aligned forecast origins, the first eligible final forecast occurs on **7 January 2020**.

The final test contains

$$N = 1,151$$

one-session out-of-sample forecasts.

During the held-out period, the HMM and diffusion models are re-estimated annually. For a forecast year y , only observations dated before 1 January of year y enter model estimation. Within each diffusion training sample, the observations are kept in chronological order; the first 85% are used for optimization and the final 15% for selecting the appropriate number of training epochs through early stopping. The network is then reinitialized and fitted to all data available at the annual cutoff for the selected number of epochs.

This design ensures that neither model architecture nor training length is selected using the held-out forecast outcomes.

4.10 Multiple Diffusion Seeds

Neural-network estimation is inherently stochastic because initialization, minibatch ordering, and noise draws can influence the fitted parameters. A result based on a single initialization could therefore reflect an unusually favorable or unfavorable training run.

To address this concern, five independently initialized diffusion models are trained using the pre-specified seeds

$$11, 29, 47, 71, 101.$$

Each seed generates a separate set of out-of-sample risk forecasts.

The seeds are treated as repeated model estimations, **not as additional financial observations**. Forecast scores are first evaluated across the actual 1,151 market dates and subsequently summarized across seeds. Robustness results report both mean performance and cross-seed dispersion.

4.11 Scenario Generation and Portfolio Losses

For every out-of-sample date and each diffusion seed, the fitted model generates

$$5,000$$

six-dimensional next-session return scenarios.

For scenario s ,

$$r_{t+1}^{(s)} = (r_{1,t+1}^{(s)}, \dots, r_{6,t+1}^{(s)})^\top.$$

Because the diffusion model generates log returns, each simulated market return is converted to a simple return:

$$g_{i,t+1}^{(s)} = \exp(r_{i,t+1}^{(s)}) - 1.$$

For the equal-weight portfolio,

$$R_{p,t+1}^{(s)} = \frac{1}{6} \sum_{i=1}^6 g_{i,t+1}^{(s)}.$$

Portfolio loss is then defined as

$$L_{t+1}^{(s)} = -R_{p,t+1}^{(s)}.$$

The resulting set of 5,000 portfolio losses forms the conditional loss distribution from which VaR and ES are estimated.

A Monte Carlo sensitivity check using 2,000, 5,000, and 10,000 scenarios is also conducted to assess whether the selected scenario count materially affects the principal risk estimates.

4.12 VaR Backtesting

VaR forecasts are evaluated using two standard tests. (Kupiec, 1995; Christoffersen, 1998).

The **Kupiec unconditional coverage test** examines whether the observed proportion of VaR violations is consistent with the theoretical violation probability

$$1 - \alpha.$$

At the 97.5% confidence level, a correctly calibrated model should therefore produce violations on approximately 2.5% of forecast dates.

The **Christoffersen independence test** examines whether violations occur independently through time. A model may achieve the correct average violation frequency while still producing clusters of exceedances during stressed periods, indicating that the risk forecast is not adapting adequately to changing market conditions.

Detailed likelihood expressions for the two tests are provided in Appendix D.

4.13 Expected Shortfall Forecast Evaluation

Because VaR and ES describe different aspects of the same conditional tail distribution, the principal model comparison uses a joint VaR-ES scoring function from the Fissler-Ziegel class. (Fissler and Ziegel, 2016; Patton et al., 2019; Bayer and Dimitriadis, 2022).

Under the positive-loss convention used in this study, the FZ0 score for realized loss L_t , VaR forecast q_t , ES forecast e_t , and confidence level α is

$$S_\alpha(L_t, q_t, e_t) = \frac{q_t}{e_t} + \ln(e_t) - 1 + \frac{(L_t - q_t)_+}{(1 - \alpha)e_t},$$

where

$$(x)_+ = \max(x, 0).$$

The score requires positive ES and

$$e_t \geq q_t.$$

The central interpretation is:

Lower FZ0 scores indicate better joint VaR-ES forecasting performance.

This criterion is particularly useful because ES cannot be meaningfully evaluated through VaR exception counts alone. The score rewards forecasts that jointly place the VaR threshold and expected tail severity appropriately relative to realized losses.

4.14 Statistical Forecast Comparison

Differences in average FZ0 scores are supplemented with formal statistical inference. (Diebold et al., 1998).

For two competing models A and B , the paired score differential is

$$d_t = S_{A,t} - S_{B,t}.$$

A negative mean differential implies that model A performs better because lower FZ0 scores are preferred.

Inference is conducted using heteroskedasticity- and autocorrelation-consistent covariance estimation so that serial dependence in forecast-score differences is not treated as independent sampling variation. In addition, a stationary bootstrap is employed as a complementary non-parametric procedure. The bootstrap uses 2,000 replications with an expected block length of 20 observations, pre-

serving local time dependence in the paired score sequence. (Newey and West, 1987). (Politis and Romano, 1994).

The two pre-specified primary comparisons are Regime-Diffusion versus Diffusion and Regime-Diffusion versus Regime-HS. Holm adjustment is used to control multiplicity within the primary comparison family. Additional comparisons involving GARCH-FHS are reported to evaluate the strongest conventional benchmark. (Holm, 1979).

The bootstrap and HAC analyses condition on the fitted forecasts. They therefore quantify uncertainty in forecast-score differences but do not repeatedly retrain all neural and econometric models within each bootstrap sample.

4.15 Scenario-Fidelity Diagnostics

Forecast rankings alone do not identify why one risk model performs better than another. The final component of the methodology therefore evaluates whether the generated return scenarios reproduce distributional characteristics that are economically relevant for portfolio tail risk.

First, **skewness** is used to assess whether models reproduce the asymmetry of realized returns. A scenario generator that produces an overly symmetric distribution may underrepresent severe negative outcomes.

Second, **excess kurtosis** measures whether the generated distributions contain sufficiently heavy tails.

Third, the empirical **2.5% return quantile** is compared with the corresponding generated quantile because it lies directly in the region relevant for 97.5% risk estimation.

Fourth, the **Wasserstein distance** measures the overall discrepancy between realized and generated marginal return distributions.

Fifth, multivariate dependence is evaluated using the distance between generated and realized correlation matrices. The Frobenius norm of the correlation-matrix difference is used, with smaller values indicating closer reproduction of the observed dependence structure.

Finally, lower-tail dependence is assessed using simultaneous 5% downside exceedances. For markets i and j , the event of interest is

$$I_{i,j,t} = 1(r_{i,t} < q_{i,0.05} \text{ and } r_{j,t} < q_{j,0.05}).$$

The generated probability of each joint downside event is compared with its realized occurrence. A Brier score summarizes the accuracy of these probabilistic co-exceedance forecasts across market pairs and forecast dates. (Brier, 1950).

These diagnostics allow the analysis to move beyond the question of **which model has the lowest ES score** to the more economically informative question of **which distributional features explain the observed differences in tail-risk forecasting performance**. In particular, they provide the empirical basis for evaluating whether deficiencies in negative skewness, fat-tail behaviour, or joint downside dependence account for the performance of the diffusion-based scenario models.

5. Empirical Results

5.1 Market-Regime Identification

The three-state Hidden Markov Model produces economically distinct market environments that can be interpreted as **Calm**, **Normal**, and **Stress** regimes. The classification used in the out-of-sample analysis is based on the regime with the highest **filtered probability at the forecast origin**. Consequently, the regime labels use only information available when the forecast is formed and are not ex-post classifications based on the subsequently realized return.

Figure 5 plots the filtered probabilities assigned to the three regimes over the out-of-sample period.

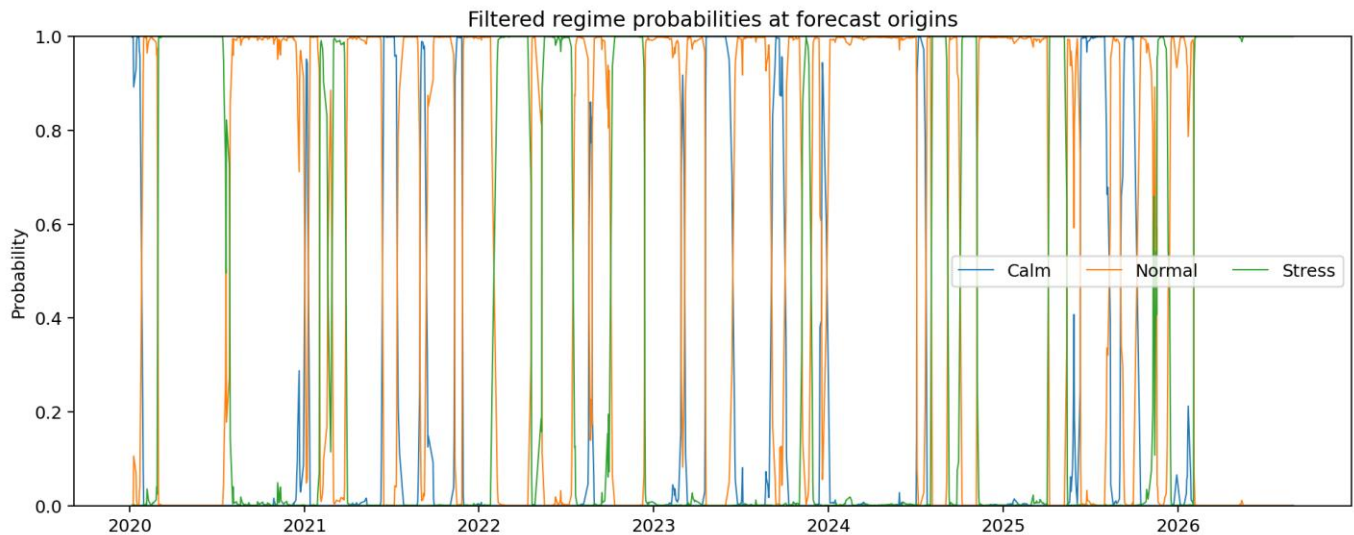


Figure 5. Filtered Probabilities of Calm, Normal, and Stress Market Regimes

The regime probabilities vary substantially through time rather than remaining concentrated permanently in one state. Periods identified as Stress exhibit visibly higher persistence during episodes of elevated market volatility, while Calm probabilities dominate during lower-volatility intervals. The Normal state accounts for the largest

share of the out-of-sample observations and functions as an intermediate state between these two environments.

Table 2 summarizes the characteristics of the observations associated with each regime. The 2.5% downside quantile is reported because it corresponds naturally to the left-tail region relevant for the study's 97.5% VaR and Expected Shortfall analysis.

Table 2. Characteristics of Inferred Market Regimes

Regime	OOS Observations	Mean Portfolio Return (%)	Volatility (%)	2.5% Return Quantile (%)	Average Pairwise Correlation	Worst Realized Loss (%)
Calm	159	0.071	0.593	-1.130	0.360	2.170
Normal	627	0.012	0.736	-1.595	0.440	3.265
Stress	365	0.039	1.276	-2.425	0.460	5.512

The most pronounced difference across regimes is in volatility. Daily portfolio volatility increases from approximately **0.59% in the Calm state to 1.28% in the Stress state**, more than doubling between the two environments. The downside distribution also becomes considerably more severe. The empirical 2.5% return quantile deteriorates from approximately -1.13% in Calm periods to -2.42% in Stress periods, while the worst realized portfolio loss increases from approximately 2.17% to 5.51%.

Cross-market dependence also rises with market stress. Average pairwise correlation increases from approximately 0.36 in the Calm regime to 0.44 in the Normal regime and 0.46 in the Stress regime. Although the increase is less dramatic than the rise in volatility, its direction is consistent with the idea that diversification benefits weaken when regional markets experience adverse conditions simultaneously.

The average return in the Stress regime remains slightly positive. This should not be interpreted as evidence that the regime is economically benign. Regimes are assigned using information available at the **forecast origin**, whereas the reported return is the subsequent

one-session outcome. A market classified as stressed at time t can therefore experience either a positive or negative return at $t + 1$. Volatility, downside quantiles, dependence, and worst realized losses provide more direct evidence of the economic differences between the states.

Overall, the HMM identifies market environments with meaningfully different risk characteristics. This supports the use of regime probabilities as a potential conditioning variable, while leaving open the empirical question of whether those probabilities improve tail-risk forecasts once other market-state information is already incorporated.

5.2 Main Expected Shortfall Forecasting Performance

Table 3 presents the principal out-of-sample results for the five competing models at the 97.5% confidence level. The primary comparative criterion is the Fissler-Ziegel FZ0 joint VaR-ES score. **Lower FZ0 values indicate better joint forecasting performance.** The table also reports the VaR violation rate and the Kupiec and Christoffersen backtesting results.

Table 3. Main 97.5% VaR and Expected Shortfall Forecasting Results

Model	Mean FZ0	Mean ES	VaR Violation Rate	Kupiec p -value	Christoffersen Independence p -value
HS	-3.4896	2.77%	3.04%	0.255	0.022
Regime-HS	-3.6153	2.81%	2.52%	0.966	0.005
GARCH-FHS	-3.7746	2.54%	2.52%	0.966	0.205
Diffusion	-3.6060	2.02%	3.86%	0.019	0.512
Regime-Diffusion	-3.5913	2.07%	3.82%	0.012	0.344

GARCH-FHS achieves the lowest mean FZ0 score, at -3.7746 , and therefore provides the strongest **numerical** forecasting performance in the primary specification. Regime-HS ranks second at -3.6153 , followed closely by Diffusion at -3.6060 and Regime-Diffusion at -3.5913 . Conventional HS produces the weakest mean FZ0 score, at -3.4896 .

The VaR calibration results reinforce the strong numerical performance of GARCH-FHS. At a 97.5% confidence level, the expected violation frequency is 2.5%. GARCH-FHS records a violation rate of approximately **2.52%**, almost exactly equal to the nominal rate. The Kupiec test produces a p -value of 0.966, providing no evidence against correct unconditional coverage. Its Christoffersen independence p -value of 0.205 also indicates no statistically significant clustering of violations at conventional levels.

Regime-HS records the same 2.52% overall violation rate and therefore also performs well in terms of unconditional coverage. However, its Christoffersen independence p -value is only 0.005, indicating that the timing of violations is more clustered than would be desirable under a correctly specified dynamic risk model.

The two diffusion specifications exhibit a different pattern. Diffusion records a violation rate of approximately **3.86%**, while Regime-Diffusion records approximately **3.82%**. Both exceed the nominal 2.5% tail probability. Their Kupiec p -values, 0.019 and 0.012 respectively, reject correct unconditional coverage at the 5% level. Thus, although the diffusion models achieve better FZ0 scores than conventional HS, they generate VaR thresholds that are breached too frequently. This result indicates a tendency to understate the severity of the empirical tail.

Conventional HS generates a 3.04% violation rate. Its Kupiec test does not reject unconditional coverage, but the Christoffersen test indicates statistically significant violation dependence. This is consistent with the concern that a fixed historical distribution may adjust too slowly when market risk changes.

Taken together, Table 3 provides an important initial result. GARCH-FHS has the strongest combination of numerical FZ0 performance and VaR calibration. However, the ranking of average scores alone does not establish statistical superiority. Formal pairwise inference is therefore considered in Section 5.7.

5.3 Does Generative Modelling Improve upon Basic Historical Simulation?

The first research question asks whether diffusion-generated scenarios improve 97.5% Expected Shortfall forecasts relative to conventional Historical Simulation.

On the principal FZ0 criterion, the answer is partially favorable to generative modelling. The mean score improves from

$$-3.4896$$

for HS to

$$-3.6060$$

for Diffusion. Because lower FZ0 scores are preferred, the diffusion model delivers a meaningful numerical improvement relative to the simplest historical benchmark.

However, this result changes when forecast calibration is considered. HS records a 3.04% VaR violation rate, whereas Diffusion records 3.86%. The generative model therefore experiences substantially more losses beyond its forecast VaR threshold. Its lower mean Expected Shortfall of approximately 2.02%, compared with 2.77% for HS, is also consistent with a synthetic loss distribution that is less severe in the extreme tail.

The distinction is important. Diffusion performs better than basic HS according to the joint FZ0 scoring criterion, but it does not provide better VaR coverage. Accordingly, the evidence does not support a simple conclusion that generative modelling dominates Historical Simulation across all dimensions of tail-risk performance.

The comparison with GARCH-FHS is even more informative. GARCH-FHS attains a mean FZ0 of -3.7746 , considerably lower than the Diffusion score of -3.6060 , while simultaneously producing an almost exactly nominal violation frequency. Thus, improving upon naïve Historical Simulation is not sufficient to establish the practical superiority of a generative risk model. A relatively conventional volatility-adaptive framework remains highly competitive once current conditional volatility is explicitly incorporated.

This finding addresses an important distinction in evaluating advanced risk models: the relevant benchmark should not be only the simplest historical method, but also strong econometric approaches designed specifically to capture the conditional heteroskedasticity that characterizes financial returns.

5.4 Does Regime Conditioning Improve Diffusion Forecasts?

The second research question concerns the incremental forecasting value of explicit HMM regime information.

The mean FZ0 score of the baseline Diffusion model is

$$-3.6060,$$

compared with

$$-3.5913$$

for Regime-Diffusion. The difference, expressed as Regime-Diffusion minus Diffusion, is therefore approximately

$$\Delta FZ0 = +0.0147.$$

Since lower scores indicate superior performance, the positive difference implies that Regime-Diffusion performs **slightly worse**, rather than better, in the main out-of-sample experiment.

The magnitude of the difference is small, and subsequent statistical testing shows that it is not significantly different from zero. The result therefore does not imply that regime conditioning materially damages the diffusion model. More importantly, however, it provides no evidence that the addition of HMM regime probabilities improves Expected Shortfall forecasting.

VaR calibration points to the same conclusion. The violation rate declines only marginally from 3.86% for Diffusion to 3.82% for Regime-Diffusion and remains well above the nominal 2.5% rate. The Kupiec test rejects correct unconditional coverage for both specifications.

This result is particularly relevant because the baseline Diffusion model is not unconditional. It already observes recent six-market returns and volatility-related state variables. HMM probabilities may therefore summarize information that is substantially contained in the continuous conditioning set. Explicitly representing the market as Calm, Normal, or Stress does not appear to contribute enough additional information to materially alter the model's extreme-loss forecasts.

5.5 The Effect of Regime Conditioning within Historical Simulation

The comparison between HS and Regime-HS provides a complementary perspective on the value of regime information.

Conventional HS produces a mean FZ0 score of -3.4896 , while Regime-HS improves this to -3.6153 . Its violation frequency also falls from 3.04% to 2.52%, bringing unconditional VaR coverage almost exactly in line with the nominal 2.5% level.

The numerical improvement is considerably more visible than the difference between Diffusion and Regime-Diffusion. This suggests that regime information may have greater value when applied to a relatively simple empirical resampling framework. Conventional HS treats observations from different market environments within a common historical distribution. Introducing regime probabilities changes the relevance assigned to historical scenarios and can therefore materially alter the estimated tail.

By contrast, the diffusion model already conditions scenario generation on continuous state variables. Regime information may consequently be more redundant in that environment.

This interpretation should nevertheless be qualified. Although Regime-HS achieves strong unconditional coverage, the low Christoffersen independence p -value indicates that violations remain temporally clustered. Regime conditioning therefore improves several aspects of Historical Simulation without fully resolving its dynamic calibration limitations.

5.6 Regime-Specific Model Performance

Table 4 decomposes model performance according to the Calm, Normal, and Stress regime classifications assigned at each forecast origin.

Table 4. FZ0 Performance by Market Regime

Model	Overall	Calm $N = 159$	Normal $N = 627$	Stress $N = 365$
HS	-3.4896	-3.9223	-3.8281	-2.7197
Regime-HS	-3.6153	-4.1627	-3.9071	-2.8756
GARCH-FHS	-3.7746	-4.1401	-3.8886	-3.4196
Diffusion	-3.6060	-4.1965	-3.8330	-2.9589
Regime-Diffusion	-3.5913	-4.1843	-3.8297	-2.9236

The ranking varies across regimes. During Calm periods, Diffusion produces the lowest mean FZ0 score, narrowly outperforming

Regime-Diffusion, Regime-HS, and GARCH-FHS. This suggests that the generative model is capable of producing competitive conditional risk forecasts in relatively stable market environments.

The picture changes in the Normal regime. Regime-HS achieves the best score of -3.9071 , followed closely by GARCH-FHS at -3.8886 . The two diffusion models perform similarly to each other but do not improve on the strongest historical approaches.

The largest difference appears during Stress periods. GARCH-FHS achieves a mean FZ0 of

$$-3.4196,$$

compared with -2.9589 for Diffusion and -2.9236 for Regime-Diffusion. Conventional HS performs worst at -2.7197 .

The corresponding violation rates make the stress-period result particularly important. GARCH-FHS records a Stress-regime viola-

tion rate of approximately **2.19%**, whereas Diffusion and Regime-Diffusion record approximately **4.77%** and **5.15%**, respectively. Conventional HS reaches approximately 6.58%.

Thus, diffusion does **not** become relatively more effective precisely when tail-risk modelling is most important. Its performance is strongest in Calm conditions and deteriorates materially in the Stress state. Moreover, regime conditioning does not correct this deterioration; Regime-Diffusion actually exhibits a slightly worse stress-period FZ0 score and violation frequency than baseline Diffusion.

These findings suggest that the principal limitation of the diffusion framework lies not in representing ordinary conditional variation, but in generating sufficiently severe and appropriately dependent observations during stressed market conditions.

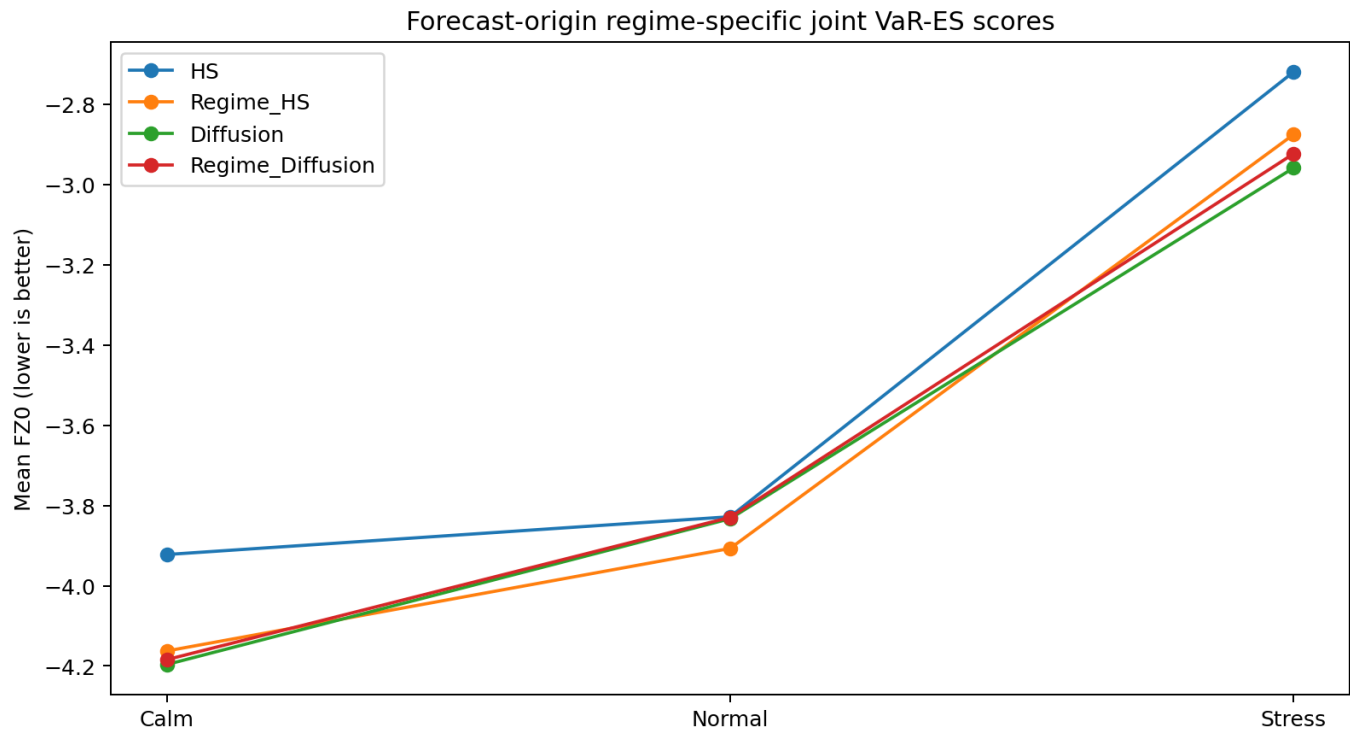


Figure 6. FZ0 Forecast Performance across Market Regimes

5.7 Formal Pairwise Forecast Comparisons

Average rankings alone cannot establish whether observed differences reflect persistent forecasting improvements rather than sampling variability. Table 5 therefore reports pairwise comparisons

Table 5. Pairwise Predictive-Accuracy Comparisons

Model A	Model B	Mean FZ0 Difference	HAC SE	95% CI	Raw p -value	Adjusted p -value
Regime-Diffusion	Diffusion	+0.0147	0.0156	[-0.0159, 0.0453]	0.346	0.692
Regime-Diffusion	Regime-HS	+0.0240	0.1107	[-0.1929, 0.2408]	0.829	0.829
GARCH-FHS	Diffusion	-0.1685	0.1030	[-0.3705, 0.0334]	0.102	0.407
GARCH-FHS	Regime-Diffusion	-0.1833	0.1047	[-0.3885, 0.0220]	0.080	0.400
GARCH-FHS	Regime-HS	-0.1593	0.1114	[-0.3776, 0.0590]	0.153	0.458

The comparison between Regime-Diffusion and Diffusion confirms the descriptive result. The mean score difference is positive, indicating slightly worse performance after regime conditioning, but the confidence interval contains zero and the raw p -value is 0.346. There is therefore no statistically significant evidence that the two diffusion specifications differ in forecasting performance.

Similarly, Regime-Diffusion does not significantly differ from Regime-HS.

The comparisons involving GARCH-FHS show consistently negative mean differences, indicating numerical superiority over Diffusion, Regime-Diffusion, and Regime-HS. The largest difference is against Regime-Diffusion at approximately -0.1833 . However, the corresponding HAC p -value is 0.080, and the adjusted

based on FZ0 score differentials using HAC inference. The difference is calculated as the score of Model A minus the score of Model B; because lower FZ0 is preferable, a negative value favors Model A.

value is 0.400. None of the priority pairwise comparisons is statistically significant at the conventional 5% threshold after adjustment for multiple comparisons.

The appropriate interpretation is therefore more nuanced than a simple ranking. **GARCH-FHS provides the strongest numerical and calibration profile, but the evidence does not establish blanket statistical superiority over all competing methods.** Likewise, there is no statistical support for the proposition that explicit HMM conditioning improves the diffusion forecasts.

5.8 Distributional Fidelity of Generated Scenarios

The third research question considers whether differences in ES performance can be related to the statistical properties of the under-

lying scenario distributions. Table 6 compares the realized out-of-sample distribution with the scenarios produced by HS, Regime-HS,

Diffusion, and Regime-Diffusion.

Table 6. Scenario Distributional Fidelity

Model	Mean Skewness	Mean Excess Kurtosis	Wasserstein Distance	2.5% Quantile Absolute Error	Correlation Error	Tail Brier Score
Realized Data	-1.340	9.798	—	—	—	—
HS	-0.582	9.940	0.00198	0.00354	0.659	0.01596
Regime-HS	-0.616	9.501	0.00183	0.00287	0.637	0.01316
Diffusion	-0.090	3.124	0.00210	0.00450	0.870	0.02276
Regime-Diffusion	-0.091	3.369	0.00206	0.00425	0.870	0.02290

The differences in skewness are particularly striking. The realized return distributions exhibit average skewness of approximately -1.34 ,

whereas the five-seed diffusion averages are only approximately -0.09 .

The diffusion-generated distributions are therefore substantially more symmetric than the realized market returns. Regime conditioning produces almost no improvement in this dimension.

The difference in kurtosis is similarly substantial. Realized average excess kurtosis is approximately 9.80 ,

while Diffusion produces approximately 3.12

and Regime-Diffusion approximately 3.37 .

The generative distributions therefore contain materially thinner tails than those observed in the actual Asia-Pacific return data.

Historical approaches perform much better on this dimension. HS produces excess kurtosis of approximately 9.94 and Regime-HS approximately 9.50 , both close to the realized value. This advantage follows naturally from retaining historical extreme observations in the empirical scenario distribution.

The 2.5% quantile error leads to a similar conclusion. Regime-HS records the smallest mean absolute error, followed by HS, while both diffusion specifications show larger discrepancies from realized downside quantiles.

These findings provide a direct mechanism for understanding the VaR and ES results. If synthetic scenarios are too symmetric and insufficiently fat-tailed, the model will generate too few sufficiently severe losses. The resulting VaR threshold will be breached more

frequently than intended, which is exactly what is observed for the two diffusion models.

5.9 Cross-Market Dependence

Marginal distributions are only one component of multivariate portfolio risk. Accurate Expected Shortfall forecasting also requires the model to reproduce relationships among the six equity markets.

Figure 7 compares the realized and generated dependence structures.

The correlation-matrix error in Table 6 is measured using the Frobenius distance between generated and realized correlation matrices, with lower values indicating closer replication of observed dependence.

Regime-HS achieves the lowest correlation error, at approximately 0.637 , followed by HS at 0.659 . By comparison, both diffusion models record errors close to 0.87 . Regime conditioning generates virtually no improvement in the correlation structure: the error declines only marginally from 0.8704 for Diffusion to 0.8697 for Regime-Diffusion.

This suggests that the diffusion framework has difficulty simultaneously matching the multivariate dependence structure and marginal return tails. That issue is particularly consequential for a regional portfolio because even an accurate marginal forecast for each market can understate portfolio risk if the model fails to reproduce the extent to which markets decline together.

5.10 Joint Downside Dependence

To examine dependence specifically in the region relevant for tail risk, the analysis considers simultaneous lower-tail exceedances. For a pair of markets i and j , the relevant event is

$$P(r_{i,t} < q_{i,0.05}, r_{j,t} < q_{j,0.05}),$$

where $q_{i,0.05}$ and $q_{j,0.05}$ denote the corresponding 5% downside return quantiles.

Table 7. Lower-Tail Dependence Accuracy

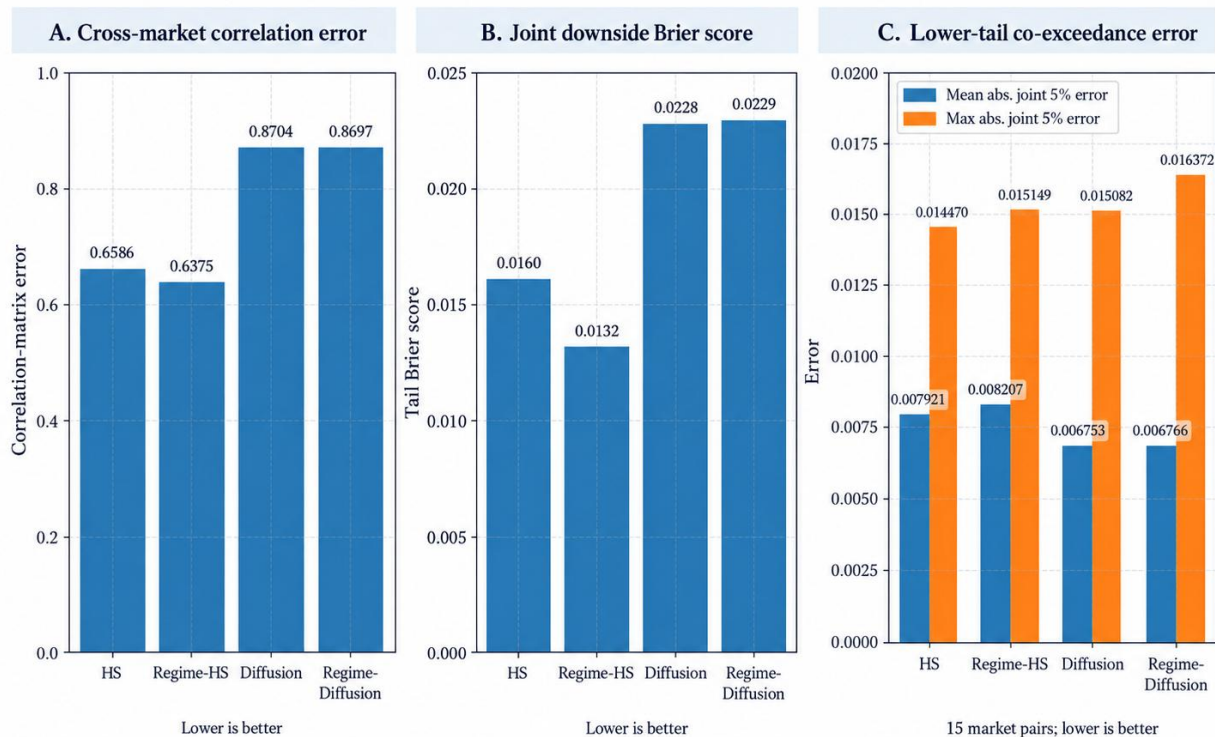
Model	Mean Absolute Joint 5% Error	Maximum Joint 5% Error	Tail Co-Exceedance Brier Score
HS	0.00792	0.01447	0.01596
Regime-HS	0.00821	0.01515	0.01316
Diffusion	0.00675	0.01508	0.02276
Regime-Diffusion	0.00677	0.01637	0.02290

The tail-dependence results are more nuanced than the marginal-tail findings. When only the **average unconditional pairwise frequency** of joint 5% downside events is considered, the diffusion models produce slightly smaller mean absolute errors than the historical methods. Diffusion records an error of approximately 0.00675 and Regime-Diffusion approximately 0.00677 .

However, the dynamic tail co-exceedance Brier score yields the opposite ordering. Regime-HS produces the lowest score at approximately 0.0132 , followed by HS at 0.0160 , while both diffusion models produce values of approximately 0.0228 – 0.0229 . This indicates that matching the average frequency of simultaneous downside events does not imply accurate **time-varying probabilistic calibration** of those events.

This distinction is important for Expected Shortfall. Portfolio losses become most severe when individual market drawdowns occur simultaneously. A model that approximately reproduces the unconditional number of joint tail events, but assigns their probabilities poorly across time, may still produce inaccurate conditional portfolio-tail forecasts.

Thus, the dependence evidence should not be interpreted as showing that diffusion fails on every joint-tail metric. Instead, it indicates that its dependence performance is mixed: average lower-tail co-exceedance frequencies are reproduced reasonably well, but overall correlation fidelity and dynamic tail-event calibration remain weaker than those of the strongest historical benchmark.

Figure 7. Realized and Generated Cross-Market/Tail Dependence

Notes: Metrics are computed relative to realized held-out dependence. Panel A reports correlation-matrix error, Panel B reports lower-tail Brier score, and Panel C reports lower-tail co-exceedance error at the 5% threshold. Lower values indicate better dependence fidelity.

5.11 Interpretation of the Main Findings

The empirical results provide three broad conclusions.

First, generative scenario modelling offers some improvement over basic Historical Simulation according to the joint FZ0 scoring criterion, but this improvement does not extend uniformly to VaR calibration. Both diffusion specifications generate substantially more VaR exceedances than the nominal 2.5% level.

Second, explicit HMM regime conditioning does not improve diffusion-based Expected Shortfall forecasting. Regime-Diffusion produces a slightly worse mean FZ0 score than Diffusion, and the difference is statistically insignificant. By contrast, regime conditioning produces a clearer numerical improvement when applied to Historical Simulation, suggesting that regime information is more useful when the underlying method does not already incorporate continuous market-state variables.

Third, and most importantly, the scenario-fidelity analysis provides an explanation for the forecasting results. The diffusion models generate conditional risk distributions that are considerably less negatively skewed and less fat-tailed than the realized data. They

also reproduce the overall cross-market correlation structure less accurately and exhibit weaker dynamic calibration of simultaneous downside events. Explicit regime probabilities do not materially correct these deficiencies.

The results therefore suggest that the principal challenge for diffusion-based market-risk models is not their ability to produce large numbers of conditional scenarios. Rather, it is their ability to allocate sufficient probability mass to the **extreme, asymmetric, and jointly adverse outcomes** that determine portfolio Expected Shortfall.

In summary, diffusion models generate responsive conditional risk distributions, but the additional flexibility of the generative framework does not by itself ensure accurate representation of the extreme joint left tail. Explicit HMM regime conditioning does not overcome this limitation in the present specification. These findings motivate the robustness tests and stress-scenario analysis presented in the following section.

Figure 8 provides a visual comparison of Expected Shortfall forecasts and realized portfolio losses over the held-out evaluation period.

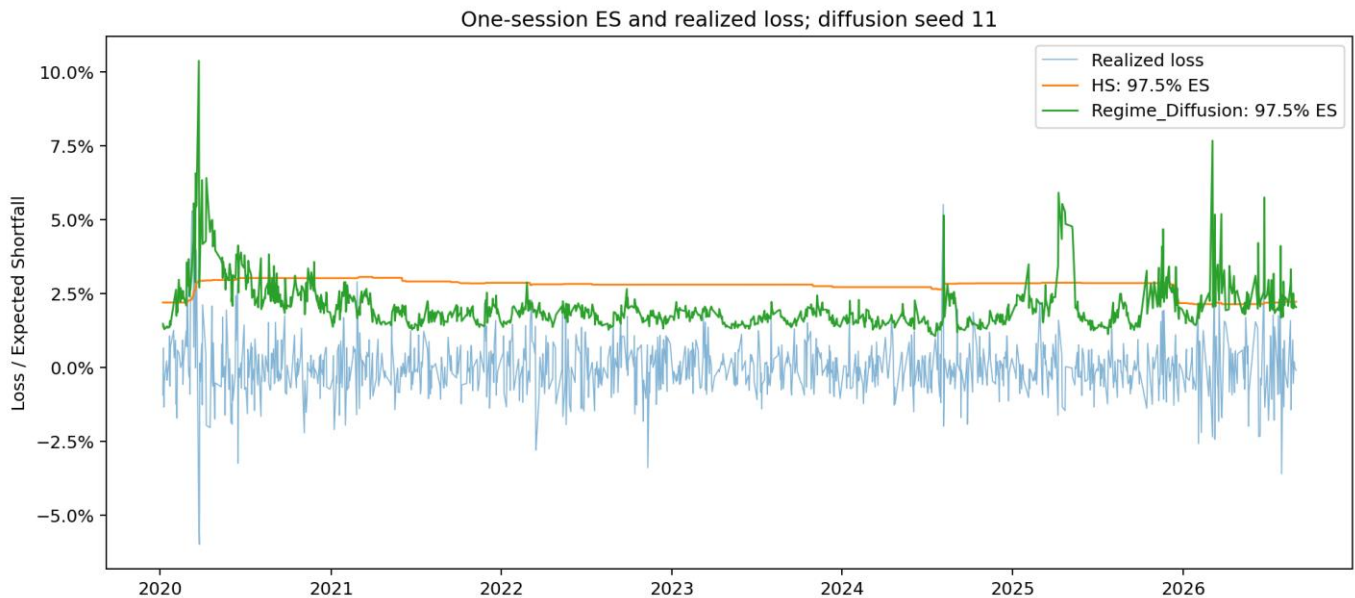


Figure 8. Expected Shortfall Forecasts and Realized Portfolio Losses

6. Robustness and Extensions

The main results indicate that GARCH-FHS provides the strongest numerical forecasting performance, while explicit HMM regime conditioning does not improve the diffusion model. This section evaluates whether these conclusions depend on the selected ES confidence level, Historical Simulation window, portfolio composition, diffusion initialization, or number of HMM states. An additional multi-day stress exercise examines whether the regime-conditioned diffusion model can generate losses comparable with particularly severe realized market episodes.

6.1 Alternative Expected Shortfall Confidence Levels

The baseline analysis uses a 97.5% confidence level. To determine whether the model ranking depends on this choice, the analysis is repeated at 95% and 99%.

At the more extreme 99% confidence level, the principal ranking remains broadly unchanged:

GARCH-FHS → Regime-HS → Diffusion →
Regime-Diffusion → HS.

The corresponding mean FZ0 scores are:

Model	97.5% FZ0	99% FZ0	99% VaR Violation Rate
HS	-3.4896	-3.0999	1.30%
Regime-HS	-3.6153	-3.3239	0.96%
GARCH-FHS	-3.7746	-3.5718	0.96%
Diffusion	-3.6060	-3.2743	2.21%
Regime-Diffusion	-3.5913	-3.2657	1.98%

The 99% results strengthen two features of the baseline analysis. First, GARCH-FHS remains numerically strongest and achieves a violation frequency close to the nominal 1% rate. Second, Regime-Diffusion again fails to improve upon the non-regime Diffusion model. Both diffusion specifications also continue to exhibit excessive VaR violations, particularly at this more extreme confidence level.

The 95% analysis requires an important technical qualification. For HS, Regime-HS, and GARCH-FHS, mean FZ0 scores are well defined and equal to -3.7433 , -3.8363 , and -3.9323 , respectively. For each diffusion specification, however, one of the 1,151 forecast dates violates the strict admissibility conditions required by the implemented FZ0 scoring function. Rather than silently removing the affected observation and reporting a potentially biased average, the aggregate 95% FZ0 score for the diffusion models is treated as undefined in the strict robustness pipeline.

Historical Simulation Window Robustness

Specification	HS FZ0	Regime-HS FZ0
500 observations	-3.4752	-3.6096
1,000 observations	-3.4896	-3.6153
1,500 observations	-3.5371	-3.6699
Expanding window	-3.5259	-3.6982

Regime-HS produces a lower FZ0 score than conventional HS under every window specification. The difference is therefore not an artifact of selecting a particular 1,000-observation lookback.

The results also illustrate the trade-off inherent in historical-window selection. A short window responds more rapidly to chang-

The corresponding 95% VaR violation rates can nevertheless be reported: approximately 6.38% for Diffusion and 6.19% for Regime-Diffusion, compared with a nominal rate of 5%. The isolated scoring-domain issue is therefore disclosed as a diagnostic limitation rather than being resolved through ex-post deletion. Overall, the alternative-confidence-level analysis does not provide evidence that explicit regime conditioning materially improves the diffusion model.

6.2 Alternative Historical Simulation Windows

The main HS and Regime-HS specifications use the most recent 1,000 eligible observations. To assess sensitivity to this choice, the analysis is repeated using 500 observations, 1,500 observations, and an expanding historical window.

ing conditions but contains fewer extreme observations, whereas a longer or expanding window provides a larger tail sample at the cost of incorporating observations from potentially less relevant historical environments. No single window eliminates this trade-off.

Importantly, however, changing the window does not overturn the broader conclusion that incorporating regime information is numerically more useful for the relatively simple Historical Simulation framework than for the diffusion model.

6.3 Alternative Portfolio Composition

The baseline portfolio assigns an equal weight of one-sixth to each of the six markets. As a robustness test, a more India-concentrated portfolio is examined, assigning 40% to the Nifty 50 and 12% to each of the other five indices:

$$w_{\text{India}} = 0.40,$$

and

$$w_i = 0.12 \text{ for each of the remaining five markets.}$$

This alternative specification changes the portfolio's exposure without introducing portfolio optimization or estimating weights using the test sample.

For the India-concentrated portfolio, the mean 97.5% FZ0 scores are:

$$\text{GARCH-FHS} = -3.8122, \text{Diffusion} = -3.7083, \text{Regime-Diffusion} = -3.6868, \text{Regime-HS} = -3.5925,$$

and

$$\text{HS} = -3.4685.$$

The relative position of some intermediate models changes, but the two central findings remain intact. GARCH-FHS again records the strongest numerical score, and Regime-Diffusion again performs slightly worse than Diffusion. The difference between the two diffusion specifications is therefore not specific to the equal-weight portfolio.

6.4 Diffusion-Seed Robustness

Because neural-network estimation depends on initialization and stochastic optimization, the diffusion results are evaluated across five independent random seeds:

$$11, 29, 47, 71, 101.$$

Across these estimations, Diffusion produces a mean FZ0 score of

$$-3.6060 \pm 0.0724,$$

while Regime-Diffusion produces

$$-3.5913 \pm 0.0475,$$

where the reported dispersion is the standard deviation across seeds.

The corresponding mean VaR violation rates are

$$3.86\% \pm 0.35\%$$

for Diffusion and

$$3.82\% \pm 0.25\%$$

for Regime-Diffusion.

More importantly, the distributional-fidelity weaknesses reported in Section 5 are also stable across seeds. Realized out-of-sample returns exhibit average skewness of approximately -1.34 , whereas five-seed average generated skewness is only -0.090 for both diffusion specifications. Realized excess kurtosis is approximately 9.80 , compared with 3.12 for Diffusion and 3.37 for Regime-Diffusion.

The mean correlation-matrix errors are also very similar:

$$0.8704$$

for Diffusion and

$$0.8697$$

for Regime-Diffusion. Likewise, the mean dynamic tail co-exceedance Brier scores are approximately 0.02276 and 0.02290 , respectively.

Thus, neither the tail underrepresentation nor the lack of improvement from regime conditioning can be attributed to one unusually poor neural-network initialization. The central diffusion findings are stable across five independently trained models.

6.5 Number of HMM States

The primary HMM uses three states because the Calm-Normal-Stress structure was selected ex ante for economic interpretability. A separate robustness exercise estimates two-, three-, and four-state HMMs.

The information-criterion evidence favors the more complex specification. Across the seven annual out-of-sample estimation windows from 2020 through 2026, the four-state model has the lowest BIC in every case. Mean BIC values are approximately:

$$13,507.2$$

for two states,

$$11,205.5$$

for three states, and

$$9,731.2$$

for four states.

However, improved in-sample statistical fit does not translate into better tail-risk forecasting. The corresponding Regime-HS FZ0 scores are:

$$-3.6307$$

for two states,

$$-3.6153$$

for three states, and

$$-3.5573$$

for four states.

Thus, the four-state specification, despite having the best BIC, produces the weakest Regime-HS FZ0 score among the three alternatives. HAC comparisons also fail to establish statistically significant differences at the 5% level. The two-state versus four-state comparison is closest to significance, with a p -value of approximately 0.0503 , but remains marginally above the conventional threshold.

This result reinforces an important broader conclusion:

$$\text{greater model complexity} \Rightarrow \text{better Expected Shortfall forecasting.}$$

The four-state result was obtained as a robustness exercise after the primary specification had been frozen. The diffusion models were therefore **not retroactively retrained using four regimes**, as doing so after observing held-out test results would introduce a post-hoc model-selection concern.

6.6 Illustrative Multi-Day Stress Test

The final extension examines the three largest realized 10-session losses in the portfolio and asks whether a Regime-Diffusion model conditioned persistently on the Stress state can generate losses of comparable magnitude. The exercise uses 5,000 simulated paths for each episode.

It is important to distinguish this exercise from the one-session ES backtest. It is an **ex-post conditional stress scenario**, not a formally calibrated regulatory 10-day ES forecast.

Table 8. Multi-Day Stress-Scenario Extension

Episode Origin	Actual 10-Session Loss	Regime-Diffusion 97.5% ES	Worst Simulated Loss	Simulated Paths Worse than Actual
22 Jul 2024	8.22%	3.09%	6.65%	0.00%
11 Jan 2022	3.16%	3.81%	7.07%	2.00%
8 Feb 2023	2.61%	5.73%	9.65%	16.02%

The most severe episode is especially informative. The realized portfolio loss over the subsequent 10 sessions is approximately

$$8.22\%,$$

equivalent to a portfolio return of approximately

$$-8.22\%.$$

The worst Regime-Diffusion path produces a loss of only approximately 6.65%. None of the 5,000 simulated paths is more severe than the realized event. The model therefore fails to reproduce the magnitude of the largest selected stress episode.

The other two cases yield a different conclusion. Their realized losses of approximately 3.16% and 2.61% fall within the generated distributions, with 2.00% and 16.02% of simulated paths, respectively, producing still larger losses.

The stress evidence is therefore **mixed rather than uniformly negative**. The regime-conditioned diffusion model can generate losses exceeding two of the three selected events, but it fails to span the most severe realized episode. This result is consistent with the broader evidence that diffusion-based scenario generation captures substantial conditional variation while remaining vulnerable to underrepresenting the most extreme tail outcomes.

7. Discussion

7.1 Why Does Regime Conditioning Not Add Value?

One of the central findings of this study is that adding HMM regime probabilities to the diffusion model does not improve out-of-sample Expected Shortfall forecasting. The Regime-Diffusion specification produces a slightly weaker mean FZ0 score than the baseline Diffusion model, and the difference is not statistically significant. This result is important because it distinguishes between the economic relevance of market regimes and their **incremental predictive value** within an already conditional forecasting model.

The baseline diffusion model is not unconditional. It already receives information on the most recent returns across the six Asia-Pacific equity markets and their recent volatility. These variables contain substantial information about prevailing market conditions. Large negative returns, elevated volatility, and changes in cross-market behaviour are precisely the observable features from which the HMM derives its probabilities of Calm, Normal, and Stress states.

Consequently, the HMM probabilities may largely act as a compressed representation of information already supplied to the neural model in continuous form. Conceptually, if π_t denotes the regime-probability vector and c_t contains recent returns and volatility, the incremental predictive information contained in π_t , conditional on c_t , may be limited:

$$I(L_{t+1}; \pi_t | c_t) \approx 0,$$

where L_{t+1} denotes the subsequent portfolio loss. This expression is conceptual rather than an estimated information-theoretic statistic in the present study.

The contrast with Regime-HS is informative. Introducing regime information into Historical Simulation produces a clearer numerical improvement because conventional HS otherwise has relatively little mechanism for adapting the historical scenario distribution to the current market environment. In the diffusion framework, however, continuous state variables already provide a mechanism for conditional adaptation. The discrete regime variables may therefore summarize information that the network has already received.

The finding should not be interpreted as evidence that financial regimes are irrelevant. The regime analysis shows clear economic differences between Calm, Normal, and Stress states. Rather, the result suggests that **economically meaningful information is not necessarily incrementally useful when added to an already information-rich forecasting model**.

7.2 Why Does GARCH-FHS Perform Well?

The strong numerical performance of GARCH-FHS provides a second important insight. Despite its substantially simpler architecture, GARCH-FHS achieves the lowest mean FZ0 score and the

closest VaR violation frequency to the nominal 2.5% level in the primary experiment.

This performance is economically plausible because GARCH-FHS is directly constructed around several characteristics that matter for short-horizon financial tail risk. First, the GARCH component explicitly models **conditional heteroskedasticity**. Financial-market volatility varies considerably over time, and the empirical analysis shows pronounced volatility clustering across the six Asia-Pacific markets. By scaling historical shocks according to forecast conditional volatility, GARCH-FHS reacts directly to changes in the current risk environment.

Second, the method retains standardized empirical residuals rather than imposing a purely Gaussian innovation distribution. Historical extreme shocks therefore remain represented in the scenario pool. The model combines this empirical tail information with dynamic volatility adjustment.

This combination is particularly well aligned with the forecasting objective. Expected Shortfall is a specific functional of the extreme-loss distribution rather than a general measure of distributional realism. A model designed to capture changing volatility and preserve historical innovations may therefore perform strongly even if it lacks the architectural flexibility of a deep generative model.

The broader implication is that:

$$\text{greater architectural flexibility} \Rightarrow \\ \text{better tail-risk forecasting.}$$

When the target is a narrowly defined extreme-tail functional, a relatively parsimonious model designed around the dominant properties of financial returns can remain highly competitive.

At the same time, the statistical evidence requires caution. Although GARCH-FHS produces the strongest numerical ranking and calibration profile, the pairwise HAC tests do not establish its superiority over all competing models at the conventional 5% significance level. The appropriate conclusion is therefore that GARCH-FHS is the **strongest numerical benchmark in this application**, not that it has been proven universally superior.

7.3 Why Do the Diffusion Models Struggle in the Tail?

The scenario-fidelity analysis provides a possible explanation for the diffusion models' comparatively weak tail calibration.

The realized out-of-sample return distributions exhibit average skewness of approximately

$$-1.34,$$

whereas the corresponding diffusion-generated distributions exhibit average skewness close to

$$-0.09.$$

Thus,

$$|\text{Skewness}_{\text{Realized}}| > |\text{Skewness}_{\text{Diffusion}}|.$$

The synthetic distributions are substantially more symmetric than the observed returns.

An even larger difference appears in kurtosis. Realized excess kurtosis is approximately

$$9.80,$$

compared with approximately

$$3.12$$

for Diffusion and

$$3.37$$

for Regime-Diffusion. Hence,

$$\text{Kurtosis}_{\text{Realized}} > \text{Kurtosis}_{\text{Diffusion}}.$$

The diffusion models therefore place insufficient probability mass in very extreme outcomes.

Several mechanisms could potentially contribute to this behaviour, although the present analysis does not establish any one of them causally. The diffusion training objective is based on prediction error across the entire distribution. Because extreme observations are rare, they contribute relatively little to an average mean-squared denoising objective. The model may therefore learn the

high-density central region more accurately than the sparsely populated extreme tail.

The limited number of genuine crisis observations may reinforce this problem. Synthetic generation can produce many scenarios, but it cannot create reliable information about extreme states that is absent or weakly represented in the training data.

The Gaussian forward-noising construction may also influence how efficiently highly asymmetric tail features are learned, although this possibility requires dedicated experimentation before any causal conclusion can be drawn. Similarly, the model's conditioning architecture may place greater emphasis on recent volatility and central conditional behaviour than on rare joint-tail transitions.

Most importantly, neither diffusion specification is trained using an explicit VaR-, ES-, or tail-weighted objective. The neural network is therefore not directly penalized for failing to allocate sufficient probability to the extreme left tail. The findings suggest that future financial diffusion models intended specifically for risk measurement may need objectives that give greater weight to extreme outcomes.

7.4 Implications for Generative Modelling in Finance

The results should not be interpreted as evidence that diffusion models are unsuitable for financial applications. The models successfully generate large numbers of multivariate scenarios and adapt their conditional distributions to current market information. Such characteristics may remain useful for scenario exploration, synthetic data generation, conditional simulation, and other financial applications in which the entire distribution is of interest.

The more specific conclusion is that **the ability to generate plausible synthetic scenarios does not automatically guarantee accurate tail-risk calibration.**

This distinction matters because generative models are often evaluated according to overall distributional similarity. Market-risk applications impose a much stricter requirement. A relatively small subset of extreme observations determines VaR and especially ES. Small distributional errors in the centre may be relatively unimportant, whereas seemingly modest underrepresentation of negative skewness, kurtosis, or simultaneous market declines can materially affect the estimated tail.

For applications specifically targeting ES, generative models may therefore require more explicit **tail-aware design**. Potential directions include tail-weighted loss functions, joint training objectives linked to VaR and ES, extreme-value components, or mechanisms that oversample and learn stressed observations more effectively.

7.5 Implications for Risk Management Practice

The findings also provide a practical comparison of the four modelling approaches represented in the study.

Conventional HS retains substantial appeal because it is transparent, easily explained, and preserves actual historical scenarios. Its main weakness is its slow response to changing conditions and its dependence on a finite historical window.

Regime-HS preserves much of this interpretability while allowing the relevance of historical observations to depend on the current market state. The empirical results indicate that such conditioning can improve numerical performance relative to ordinary HS, although violation clustering remains a concern.

GARCH-FHS emerges as an especially important practical benchmark. It combines empirical shocks with a direct mechanism for adapting to current volatility and performs strongly without requiring a computationally intensive neural architecture. For short-horizon daily market-risk forecasting, the results support retaining volatility-filtered historical methods as serious benchmarks even when evaluating substantially more sophisticated machine-learning models.

Diffusion models provide much greater scenario-generation flexibility, but the present evidence does not support replacing established risk models solely on the basis of that flexibility. Their use for formal tail-risk measurement should depend on demonstrated tail calibration rather than synthetic realism alone.

7.6 Model Complexity and Forecasting Performance

A broader lesson emerges from two separate findings in the study.

First, augmenting the diffusion model with HMM regime probabilities increases the information and complexity of the specification but does not improve Expected Shortfall forecasting.

Second, increasing the HMM from three to four states produces a substantially lower BIC across the annual estimation windows, indicating better in-sample statistical fit. Nevertheless, the four-state Regime-HS model produces a weaker out-of-sample FZ0 score than the two- and three-state alternatives.

Together, these results highlight the distinction between **model fit** and **forecast usefulness**:

better in-sample fit $\Rightarrow$
better out-of-sample tail forecasts,

and more generally,

greater model complexity $\Rightarrow$
better Expected Shortfall.

For risk modelling, complexity should therefore be justified through genuine out-of-sample improvement rather than superior in-sample likelihood or greater representational capacity alone.

7.7 Limitations

The results should be interpreted in light of several limitations.

1. **Market coverage.** The analysis covers six major Asia-Pacific equity indices and should not be interpreted as representing every market in the region.
2. **Data frequency.** The core analysis uses daily observations. Results may differ at intraday, weekly, or other frequencies.
3. **Portfolio specification.** The primary portfolio is equally weighted. Although an India-concentrated portfolio is examined as a robustness test, other institutional portfolios may generate different tail exposures.
4. **Currency treatment.** Equity indices are analysed in their local-market return series. The study therefore does not explicitly model FX-hedged or unhedged returns from the perspective of an investor with a common base currency.
5. **Trading-calendar alignment.** The common-session methodology avoids artificial zero returns but necessarily discards some market-specific trading sessions.
6. **Regime variables.** The HMM is based on returns and volatility. Additional information such as liquidity, macroeconomic variables, credit conditions, interest rates, or option-implied indicators could identify market states differently.
7. **Extreme-event scarcity.** Genuine crisis observations remain limited even in a long daily sample. This is particularly challenging for data-intensive generative methods.
8. **Diffusion architecture.** The study evaluates one primary family of conditional diffusion architecture. The findings should therefore not be generalized to every possible diffusion specification.
9. **Forecast horizon.** The central experiment concerns one-session-ahead VaR and ES. Longer-horizon dynamics may require different architectures and evaluation procedures.
10. **Stress extension.** The multi-day stress exercise is exploratory and ex post; it is not presented as a formally calibrated 10-day regulatory ES model.
11. **Data source.** The empirical analysis relies on publicly available market-index series. Institutional-grade databases may differ in adjustments, historical corrections, or market-close conventions.
12. **Computational requirements.** Diffusion training and repeated scenario generation require substantially greater computational resources than HS or GARCH-FHS.
13. **Architecture-dependent synthetic tails.** The observed tail weakness reflects the implemented diffusion architecture and training objective. Alternative tail-aware generative designs may produce materially different results.

These limitations define the boundaries of the present evidence while also identifying clear directions for future research.

8. Conclusion

This study examines whether diffusion-based generative scenario models can improve Expected Shortfall forecasting for a diversified portfolio of six major Asia-Pacific equity markets and, in particular, whether explicit market-regime information adds incremental value to a conditional diffusion framework. Using daily data covering India, Japan, Hong Kong, South Korea, Australia, and Singapore from 2007 to August 2026, the study evaluates 1,151 genuine out-of-sample forecasts between January 2020 and August 2026. Five approaches are compared within a common forecasting framework: Historical Simulation, Regime-Historical Simulation, GARCH-Filtered Historical Simulation, a non-regime conditional diffusion model, and a regime-conditioned diffusion model.

The first research question asks whether diffusion-generated scenarios improve 97.5% Expected Shortfall forecasts relative to conventional Historical Simulation. The results provide a qualified answer. The Diffusion model achieves a lower mean FZ0 score than basic Historical Simulation, indicating that flexible conditional scenario generation can improve upon a relatively static empirical-resampling benchmark. However, this advantage does not extend to the stronger volatility-adaptive benchmark. GARCH-FHS records the best numerical FZ0 performance and produces a VaR violation rate close to the nominal 2.5% level. The diffusion models, by contrast, experience violation rates of approximately 3.8%, indicating that their forecast loss distributions are insufficiently conservative in the extreme tail. Formal pairwise tests do not establish GARCH-FHS as statistically superior to all competing models at conventional significance levels, but its combination of numerical performance and calibration demonstrates that greater generative flexibility does not automatically dominate well-established econometric approaches.

The second research question concerns whether explicit HMM regime conditioning improves diffusion-based tail-risk forecasts. The evidence does not support this proposition. The Regime-Diffusion model produces a slightly weaker mean FZ0 score than the baseline Diffusion model, and the difference is statistically insignificant. Its VaR calibration is also only marginally different. Thus, explicit Calm, Normal, and Stress probabilities do not provide statistically significant incremental forecasting value once the diffusion model already conditions on recent returns and volatility.

This finding does not imply that market regimes are unimportant. The HMM identifies economically distinct states characterized by substantial differences in volatility, downside quantiles, cross-market dependence, and realized losses. Rather, the results suggest that the information contained in HMM probabilities may already be largely represented by the continuous return and volatility variables supplied to the diffusion model. Consistent with this interpretation, regime conditioning generates a clearer improvement when introduced into conventional Historical Simulation, where the underlying framework has less ability to adapt dynamically to prevailing market conditions.

The third research question asks whether differences in Expected Shortfall performance can be related to the ability of the models to reproduce the distributional characteristics of extreme returns. The scenario-fidelity analysis provides strong evidence in this direction. Realized returns exhibit average skewness of approximately -1.34 and excess kurtosis of approximately 9.80 . By comparison, the diffusion-generated scenarios exhibit average skewness close to -0.09 and excess kurtosis of only about 3.1 to 3.4 . The resulting synthetic distributions are therefore substantially more symmetric and less heavy-tailed than the observed data.

The diffusion models also reproduce the overall cross-market correlation structure less accurately than the strongest historical approaches and show weaker dynamic calibration of simultaneous downside events. These deficiencies are economically important because portfolio Expected Shortfall depends not only on extreme losses within individual markets but also on whether those losses

occur simultaneously. The empirical findings therefore indicate that the limitation of the diffusion framework lies less in its ability to generate large numbers of conditional scenarios than in whether those scenarios assign sufficient probability to rare, asymmetric, and jointly adverse market outcomes.

Two broader conclusions follow from the study:

Generative flexibility $\neq$
tail-forecasting accuracy

and

Greater model complexity $\neq$
superior out-of-sample risk prediction.

The second conclusion is reinforced by the regime-robustness analysis. Although a four-state HMM produces lower BIC values than the baseline three-state specification, it does not improve out-of-sample Regime-HS performance. Better statistical fit and greater representational complexity should therefore not be treated as substitutes for genuine out-of-sample validation.

For practitioners, the results suggest that diffusion models should presently be viewed as complements to, rather than automatic replacements for, established market-risk methods. Their flexibility may be valuable for scenario exploration, synthetic-data generation, and conditional simulation, but formal tail-risk applications require explicit evidence that extreme-loss behaviour is accurately reproduced. GARCH-FHS, in particular, remains a strong benchmark because it directly addresses volatility clustering while preserving empirical innovations.

Future research should therefore focus less on increasing generative complexity alone and more on making generative models explicitly **tail-aware**. Promising directions include tail-weighted diffusion objectives, loss functions directly targeted toward VaR and Expected Shortfall, hybrid extreme-value-theory and diffusion models, copula-diffusion structures for modelling joint downside dependence, and richer regime-conditioning variables incorporating macroeconomic, liquidity, or option-implied information. Further extensions could examine mixed equity, foreign-exchange, and interest-rate portfolios, transformer-diffusion architectures, higher-frequency risk factors, and explicitly multi-horizon scenario generation.

Overall, the evidence suggests that the central challenge for generative market-risk modelling is not simply to create more plausible scenarios, but to generate the **right extreme scenarios at the right probabilities**. For Expected Shortfall forecasting, accurate representation of the joint left tail remains more important than model complexity itself.

9. References

- Acerbi, C., Szekely, B. (2014). Backtesting Expected Shortfall. *MSCI Research Insight*. MSCI Inc.
- Acerbi, C., Tasche, D. (2002). On the coherence of expected shortfall. *Journal of Banking & Finance*, **26**(7), 1487–1503. [https://doi.org/10.1016/S0378-4266\(02\)00283-2](https://doi.org/10.1016/S0378-4266(02)00283-2)
- Ang, A., Bekaert, G. (2002). International asset allocation with regime shifts. *The Review of Financial Studies*, **15**(4), 1137–1187. <https://doi.org/10.1093/rfs/15.4.1137>
- Ang, A., Chen, J. (2002). Asymmetric correlations of equity portfolios. *Journal of Financial Economics*, **63**(3), 443–494. [https://doi.org/10.1016/S0304-405X\(02\)00068-5](https://doi.org/10.1016/S0304-405X(02)00068-5)
- Ang, A., Timmermann, A. (2012). Regime changes and financial markets. *Annual Review of Financial Economics*, **4**, 313–337. <https://doi.org/10.1146/annurev-financial-110311-101808>
- Artzner, P., Delbaen, F., Eber, J.M., Heath, D. (1999). Coherent measures of risk. *Mathematical Finance*, **9**(3), 203–228. <https://doi.org/10.1111/1467-9965.00068>
- Bae, K.H., Karolyi, G.A., Stulz, R.M. (2003). A new approach to measuring financial contagion. *The Review of Financial Studies*, **16**(3), 717–763. <https://doi.org/10.1093/rfs/hhg012>

- Barone-Adesi, G., Giannopoulos, K., Vosper, L. (1999). VaR without correlations for portfolios of derivative securities. *Journal of Futures Markets*, **19**(5), 583–602. [https://doi.org/10.1002/\(SICI\)1096-9934\(199908\)19:5%3C583::AID-FUT5%3E3.0.CO;2-S](https://doi.org/10.1002/(SICI)1096-9934(199908)19:5%3C583::AID-FUT5%3E3.0.CO;2-S)
- Basel Committee on Banking Supervision. (2026). *Basel Framework: Minimum capital requirements for market risk*. Bank for International Settlements.
- Bayer, S., Dimitriadis, T. (2022). Regression-based Expected Shortfall backtesting. *Journal of Financial Econometrics*, **20**(3), 437–471. <https://doi.org/10.1093/jfinc/nbaa013>
- Bekaert, G., Harvey, C.R., Ng, A. (2005). Market integration and contagion. *The Journal of Business*, **78**(1), 39–69.
- Berkowitz, J., O'Brien, J. (2002). How accurate are Value-at-Risk models at commercial banks? *The Journal of Finance*, **57**(3), 1093–1111. <https://doi.org/10.1111/1540-6261.00455>
- Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, **31**(3), 307–327. [https://doi.org/10.1016/0304-4076\(86\)90063-1](https://doi.org/10.1016/0304-4076(86)90063-1)
- Boudoukh, J., Richardson, M., Whitelaw, R.F. (1998). The best of both worlds: A hybrid approach to calculating Value at Risk. *Risk*, **11**(5), 64–67.
- Brauneis, A., Sahiner, M. (2026). Crypto volatility forecasting: Mounting a HAR, sentiment, and machine learning horseshoe. *Asia-Pacific Financial Markets*, **33**, 379–411. <https://doi.org/10.1007/s10690-024-09510-6>
- Brier, G.W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, **78**(1), 1–3.
- Chiang, T.C., Jeon, B.N., Li, H. (2007). Dynamic correlation analysis of financial contagion: Evidence from Asian markets. *Journal of International Money and Finance*, **26**(7), 1206–1228. <https://doi.org/10.1016/j.jimonfin.2007.06.005>
- Christoffersen, P.F. (1998). Evaluating interval forecasts. *International Economic Review*, **39**(4), 841–862. <https://doi.org/10.2307/2527341>
- Diebold, F.X., Gunther, T.A., Tay, A.S. (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review*, **39**(4), 863–883. <https://doi.org/10.2307/2527342>
- Engle, R.F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, **50**(4), 987–1007. <https://doi.org/10.2307/1912773>
- Engle, R.F., Manganelli, S. (2004). CAViaR: Conditional autoregressive Value at Risk by regression quantiles. *Journal of Business & Economic Statistics*, **22**(4), 367–381. <https://doi.org/10.1198/073500104000000370>
- Fissler, T., Ziegel, J.F. (2016). Higher order elicibility and Osband's principle. *The Annals of Statistics*, **44**(4), 1680–1707. <https://doi.org/10.1214/16-AOS1439>
- Fissler, T., Ziegel, J.F., Gneiting, T. (2016). Expected Shortfall is jointly elicitable with Value at Risk: Implications for backtesting. *Risk*, January, 58–61.
- Forbes, K.J., Rigobon, R. (2002). No contagion, only interdependence: Measuring stock market comovements. *The Journal of Finance*, **57**(5), 2223–2261. <https://doi.org/10.1111/0022-1082.00494>
- Fong, W.M., Koh, S.K. (2002). The political economy of volatility dynamics in the Hong Kong stock market. *Asia-Pacific Financial Markets*, **9**(3), 259–282. <https://doi.org/10.1023/A:1024133632104>
- Glosten, L.R., Jagannathan, R., Runkle, D.E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *The Journal of Finance*, **48**(5), 1779–1801. <https://doi.org/10.1111/j.1540-6261.1993.tb05128.x>
- Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, **106**(494), 746–762. <https://doi.org/10.1198/jasa.2011.r10138>
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, **27**, 2672–2680.
- Gray, S.F. (1996). Modeling the conditional distribution of interest rates as a regime-switching process. *Journal of Financial Economics*, **42**(1), 27–62. [https://doi.org/10.1016/0304-405X\(96\)00875-6](https://doi.org/10.1016/0304-405X(96)00875-6)
- Guidolin, M., Timmermann, A. (2007). Asset allocation under multivariate regime switching. *Journal of Economic Dynamics and Control*, **31**(11), 3503–3544. <https://doi.org/10.1016/j.jedc.2006.12.004>
- Hamilton, J.D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, **57**(2), 357–384. <https://doi.org/10.2307/1912559>
- Ho, J., Jain, A., Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, **33**, 6840–6851.
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, **6**(2), 65–70.
- Hull, J., White, A. (1998). Incorporating volatility updating into the historical simulation method for Value-at-Risk. *Journal of Risk*, **1**(1), 5–19. <https://doi.org/10.21314/JOR.1998.001>
- Kingma, D.P., Welling, M. (2014). Auto-encoding variational Bayes. *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*.
- Kupiec, P.H. (1995). Techniques for verifying the accuracy of risk measurement models. *The Journal of Derivatives*, **3**(2), 73–84.
- Liao, S., Ni, H., Sabate-Vidales, M., Szpruch, L., Wiese, M., Xiao, B. (2024). Sig-Wasserstein GANs for conditional time series generation. *Mathematical Finance*, **34**(2), 622–670. <https://doi.org/10.1111/mafi.12423>
- Longin, F., Solnik, B. (2001). Extreme correlation of international equity markets. *The Journal of Finance*, **56**(2), 649–676. <https://doi.org/10.1111/0022-1082.00340>
- Mar'I, M., Seraj, M. (2025). The tail dependence and lead-lag relationship in financial markets. *Asia-Pacific Financial Markets*, **32**(3), 1021–1048. <https://doi.org/10.1007/s10690-024-09479-2>
- McNeil, A.J., Frey, R. (2000). Estimation of tail-related risk measures for heteroscedastic financial time series: An extreme value approach. *Journal of Empirical Finance*, **7**(3–4), 271–300. [https://doi.org/10.1016/S0927-5398\(00\)00012-8](https://doi.org/10.1016/S0927-5398(00)00012-8)
- Nelson, D.B. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica*, **59**(2), 347–370. <https://doi.org/10.2307/2938260>
- Newey, W.K., West, K.D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, **55**(3), 703–708. <https://doi.org/10.2307/1913610>
- Nichol, A.Q., Dhariwal, P. (2021). Improved denoising diffusion probabilistic models. *Proceedings of the 38th International Conference on Machine Learning*, PMLR **139**, 8162–8171.
- Patton, A.J. (2006). Modelling asymmetric exchange rate dependence. *International Economic Review*, **47**(2), 527–556. <https://doi.org/10.1111/j.1468-2354.2006.00387.x>
- Patton, A.J., Ziegel, J.F., Chen, R. (2019). Dynamic semiparametric models for expected shortfall (and Value-at-Risk). *Journal of Econometrics*, **211**(2), 388–413. <https://doi.org/10.1016/j.jeconom.2018.10.008>
- Politis, D.N., Romano, J.P. (1994). The stationary bootstrap. *Journal of the American Statistical Association*, **89**(428), 1303–1313. <https://doi.org/10.1080/01621459.1994.10476870>
- Poon, S.H., Rockinger, M., Tawn, J. (2004). Extreme value dependence in financial markets: Diagnostics, models, and financial implications. *The Review of Financial Studies*, **17**(2), 581–610. <https://doi.org/10.1093/rfs/hhg058>
- Pritsker, M. (2006). The hidden dangers of historical simulation. *Journal of Banking & Finance*, **30**(2), 561–582. <https://doi.org/10.1016/j.jbankfin.2005.04.013>
- Rasul, K., Seward, C., Schuster, I., Vollgraf, R. (2021). Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. *Proceedings of the 38th International Conference on Machine Learning*, PMLR **139**, 8857–8868.

- Rockafellar, R.T., Uryasev, S. (2000). Optimization of conditional Value-at-Risk. *The Journal of Risk*, **2**(3), 21–41. <https://doi.org/10.21314/JOR.2000.038>
- Rockafellar, R.T., Uryasev, S. (2002). Conditional Value-at-Risk for general loss distributions. *Journal of Banking & Finance*, **26**(7), 1443–1471. [https://doi.org/10.1016/S0378-4266\(02\)00271-6](https://doi.org/10.1016/S0378-4266(02)00271-6)
- Song, J., Meng, C., Ermon, S. (2021). Denoising diffusion implicit models. *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B. (2021). Score-based generative modeling through stochastic differential equations. *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Su, F., Wang, F., Xu, Y. (2025). Economic policy uncertainty and volatility spillovers among international stock market indices during the COVID-19 outbreak. *Asia-Pacific Financial Markets*, **32**(1), 237–266. <https://doi.org/10.1007/s10690-024-09452-z>
- Taamouti, A. (2009). Analytical Value-at-Risk and Expected Shortfall under regime-switching. *Finance Research Letters*, **6**(3), 138–151. <https://doi.org/10.1016/j.frl.2009.03.004>
- Tashiro, Y., Song, J., Song, Y., Ermon, S. (2021). CSDI: Conditional score-based diffusion models for probabilistic time series imputation. *Advances in Neural Information Processing Systems*, **34**.
- Wiese, M., Knobloch, R., Korn, R., Kretschmer, P. (2020). Quant GANs: Deep generation of financial time series. *Quantitative Finance*, **20**(9), 1419–1440. <https://doi.org/10.1080/14697688.2020.1730426>
- Yoon, J., Jarrett, D., van der Schaar, M. (2019). Time-series generative adversarial networks. *Advances in Neural Information Processing Systems*, **32**.
- Yuan, X., Qiao, Y. (2024). Diffusion-TS: Interpretable diffusion for general time series generation. *Proceedings of the International Conference on Learning Representations (ICLR)*.

10. Appendix / Supplementary Material

Appendix A. Data Details

A.1 Market Series and Ticker Identifiers

Daily closing price-index data were downloaded from Yahoo Finance using the yfinance interface. The analysis uses closing index levels rather than total-return indices; dividends are therefore not incorporated.

Table A1. Market Series and Data Coverage

Market	Index	Yahoo Finance Ticker	First Price Observation	First Return Observation	Last Observation	Price Observations
India	Nifty 50	^NSEI	17 Sep 2007	18 Sep 2007	31 Aug 2026	4,650
Japan	Nikkei 225	^N225	4 Jan 2007	5 Jan 2007	31 Aug 2026	4,806
Hong Kong	Hang Seng Index	^HSI	2 Jan 2007	3 Jan 2007	31 Aug 2026	4,840
South Korea	KOSPI	^KS11	2 Jan 2007	3 Jan 2007	31 Aug 2026	4,842
Australia	S&P/ASX 200	^AXJO	2 Jan 2007	3 Jan 2007	31 Aug 2026	4,966
Singapore	Straits Times Index	^STI	3 Jan 2007	4 Jan 2007	31 Aug 2026	4,914

The requested data window was 1 January 2007 through 31 August 2026. Because Nifty 50 observations begin only in September 2007, the effective multivariate sample begins later. Following return construction and synchronization, the first usable common one-session forecast target occurs on **17 October 2007**.

The final matched dataset contains **3,258 common target observations** through 31 August 2026.

A.2 Missing-Data Audit

The six markets operate on different trading calendars. Missing values in the union of all trading dates therefore predominantly reflect legitimate exchange closures rather than missing observations within a common global trading calendar.

Table A2. Data-Quality and Calendar Audit

Market	Dates Missing from Union Calendar	Maximum Calendar Gap	Absolute Log Returns >20%
Nifty 50	473	6 days	0
Nikkei 225	317	11 days	0
Hang Seng	283	6 days	0
KOSPI	281	11 days	0
S&P/ASX 200	157	6 days	0
Straits Times	209	6 days	0

No observations were removed merely because the realized return was large. The audit identified no absolute one-session log return above 20%, reducing concern that obvious price-data errors were driving the extreme-tail results.

A.3 Trading-Calendar Treatment

Returns were first calculated independently on each exchange's native trading calendar:

$$r_{i,t} = \ln \left(\frac{P_{i,t}}{P_{i,t-1}} \right).$$

Non-trading dates were not converted into artificial zero-return observations, and index levels were not mechanically forward-filled for return calculation. Native-session returns were subsequently aligned to form the common multivariate sample.

This procedure avoids artificially lowering measured volatility on exchange holidays and reduces the possibility of inducing spurious cross-market correlations through calendar imputation.

The frozen data snapshot was retrieved on **15 September 2026**, and the final price observation included in the analysis is **31 August 2026**. The frozen data file has SHA-256 hash:

426c16cc946ae86d6a8d19f1cada9d3ba8d5f84a3d6c072ce527609448e116b9.

Appendix B. Diffusion-Model Architecture and Training

B.1 Prediction Target and Conditioning Variables

The diffusion models generate the next-session six-dimensional log-return vector

$$\mathbf{r}_{t+1} = [r_{1,t+1} \quad \cdots \quad r_{6,t+1}]^\top.$$

The non-regime Diffusion model is conditional rather than unconditional. Its conditioning vector contains:

- the most recently observed return for each of the six markets; and
- the logarithm of the corresponding trailing 20-session volatility for each market.

The baseline conditioning vector therefore has 12 dimensions.

For Regime-Diffusion, the three HMM filtered probabilities are added:

$$\boldsymbol{\pi}_t = [\pi_{C,t} \quad \pi_{N,t} \quad \pi_{S,t}]^\top,$$

giving a 15-dimensional conditioning vector.

All continuous inputs and prediction targets are standardized using parameters estimated from the corresponding training sample only.

B.2 Forward Diffusion Process

For standardized target vector x_0 , the forward diffusion process at diffusion step m is

$$x_m = \sqrt{\bar{\alpha}_m}x_0 + \sqrt{1 - \bar{\alpha}_m}\epsilon,$$

where

$$\epsilon \sim \mathcal{N}(0, I).$$

A cosine noise schedule is used to determine the sequence of $\bar{\alpha}_m$.

The complete forward process contains

$$M = 100$$

diffusion steps.

B.3 Denoising Network

The denoising network is a residual multilayer perceptron.

Table B1. Final Diffusion Architecture

Component	Final Setting
Prediction dimension	6 market returns
Baseline context dimension	12
Regime context dimension	15
Time-embedding dimension	32
Hidden width	128
Residual blocks	2
Activation	SiLU
Output dimension	6
Diffusion parameterization	Velocity prediction
Diffusion steps	100
Sampling method	Deterministic DDIM
Sampling steps	20
Noise schedule	Cosine

Each residual block contains two fully connected transformations with SiLU nonlinear activation. The residual connection is scaled to stabilize propagation through the network.

Candidate hidden widths of 64 and 128 were evaluated using the 2017–2019 validation period. A width of 128 was retained before the final out-of-sample test.

B.4 Optimization

Table B2. Training Hyperparameters

Parameter	Value
Optimizer	AdamW
Learning rate	0.001
Weight decay	0.0001
Batch size	128
Maximum epochs	150
Early-stopping patience	15 epochs
Diffusion steps	100
DDIM sampling steps	20
Forecast scenarios per date	5,000
Validation scenarios	2,000

Within each annual estimation sample, the earliest 85% of observations are initially used for parameter training and the latest 15% for early-stopping selection. After the training length is selected, the network is reinitialized and fitted using the complete information set available at the corresponding annual cutoff.

B.5 Random Seeds

Five independently initialized diffusion estimations are used:

$$11, 29, 47, 71, 101.$$

Seeds are treated as alternative neural-network estimations and are **not** treated as additional financial observations when conducting statistical inference.

Appendix C. Hidden Markov Model Details

C.1 State Specification

The baseline regime model is a three-state Gaussian HMM. The state-variable vector is

$$x_t = \begin{bmatrix} 100R_{p,t} \\ \ln(\sigma_{20,t}) \end{bmatrix},$$

where $R_{p,t}$ is the equal-weight synchronized basket return and $\sigma_{20,t}$ is the cross-market 20-session volatility proxy.

Conditional on latent state $S_t = k$,

$$x_t | S_t = k \sim \mathcal{N}(\mu_k, \Sigma_k),$$

where Σ_k is diagonal.

The latent state follows a first-order Markov process:

$$P(S_t = j | S_{t-1} = i) = p_{ij}.$$

The model is estimated by maximum likelihood using the expectation-maximization algorithm. Three separate initialization attempts are used for each annual estimation, with the highest-likelihood non-degenerate solution retained. Maximum iterations are set to 120.

C.2 Filtered State Probabilities

Forecasting uses the filtered distribution

$$P(S_t = k | \mathcal{F}_t),$$

rather than the smoothed distribution

$$P(S_t = k | \mathcal{F}_T), T > t.$$

This distinction is essential because smoothed probabilities would incorporate information unavailable at the forecast origin.

The probability vector is

$$\pi_t = \begin{bmatrix} P(S_t = C | \mathcal{F}_t) \\ P(S_t = N | \mathcal{F}_t) \\ P(S_t = S | \mathcal{F}_t) \end{bmatrix}.$$

C.3 State Labelling

HMM state identifiers have no intrinsic economic ordering. Following estimation, states are ranked according to the fitted volatility coordinate.

The state with the lowest estimated volatility is labelled **Calm**, the intermediate state **Normal**, and the highest-volatility state **Stress**.

The following values summarize the fitted emission means averaged across the annual 2020–2026 three-state models.

Table C1. Average Fitted HMM State Coordinates

State	Basket-Return Coordinate	20-Session Volatility Coordinate
Calm	0.0496%	0.7120%
Normal	0.0099%	0.9809%
Stress	-0.0392%	1.6936%

These are **fitted HMM emission means**, not the realized out-of-sample regime statistics reported in Table 2 of the main manuscript.

C.4 Transition Probabilities

Averaging the annual three-state transition matrices over 2020–2026 yields:

Table C2. Average HMM Transition Matrix

From / To	Calm	Normal	Stress
Calm	0.9745	0.0237	0.0018
Normal	0.0183	0.9675	0.0142
Stress	<0.0001	0.0206	0.9794

The large diagonal elements indicate substantial state persistence. In particular, once the model enters the Stress state, it assigns a high probability to remaining in that state during the following observation.

C.5 Number of States and Information Criteria

The three-state Calm-Normal-Stress specification was chosen ex ante for economic interpretability. It was not selected because it minimized BIC.

Table C3. HMM State-Count Robustness, 2020–2026

Number of States	Mean BIC	Mean Δ BIC from Best	Years BIC-Best	Regime-HS FZ0
2	13,507.18	3,775.99	0/7	-3.6307
3	11,205.49	1,474.30	0/7	-3.6153
4	9,731.19	0.00	7/7	-3.5573

The four-state specification has the lowest BIC in every out-of-sample annual estimation window, but its Regime-HS forecast score is weaker.

Pairwise HAC comparisons of Regime-HS FZ0 scores give:

- two states versus three states: $p = 0.2971$;
- two states versus four states: $p = 0.0503$;
- three states versus four states: $p = 0.1065$.

None is statistically significant at the conventional 5% level.

The four-state result was obtained as a robustness exercise. Regime-Diffusion was not subsequently redesigned and retrained with four states, because doing so after examining held-out test performance would constitute post-hoc model modification.

Appendix D. Risk Forecasting and Statistical-Testing Equations

D.1 Value-at-Risk

Let $F_{L,t}$ denote the conditional distribution of portfolio loss L at forecast origin t . For confidence level α ,

$$VaR_{\alpha,t} = F_{L,t}^{-1}(\alpha).$$

Equivalently,

$$VaR_{\alpha,t} = \inf\{\ell: F_{L,t}(\ell) \geq \alpha\}.$$

The baseline confidence level is

$$\alpha = 0.975.$$

D.2 Expected Shortfall

Expected Shortfall is defined by the upper-tail quantile integral

$$ES_{\alpha,t} = \frac{1}{1-\alpha} \int_{\alpha}^1 F_{L,t}^{-1}(u) du.$$

For a continuous loss distribution, this corresponds to

$$ES_{\alpha,t} = E[L_t | L_t \geq VaR_{\alpha,t}].$$

For empirical or weighted scenario distributions, the implementation integrates the empirical upper quantile function and permits fractional probability mass at the VaR threshold. This avoids ambiguity arising from tied observations or finite tail samples.

D.3 Kupiec Unconditional-Coverage Test

Define the violation indicator

$$I_t = 1(L_t > VaR_{\alpha,t}).$$

With n forecasts and x violations,

$$\hat{p} = \frac{x}{n},$$

while the theoretical violation probability is

$$p = 1 - \alpha.$$

The Kupiec likelihood-ratio statistic is

$$LR_{UC} = -2 \ln \left[\frac{p^x (1-p)^{n-x}}{\hat{p}^x (1-\hat{p})^{n-x}} \right].$$

Under the null hypothesis of correct unconditional coverage,

$$LR_{UC} \chi_1^2.$$

D.4 Christoffersen Independence Test

Let n_{ij} denote the number of occasions on which violation state i is followed by state j , for $i, j \in \{0, 1\}$.

Define

$$\hat{\pi}_0 = \frac{n_{01}}{n_{00} + n_{01}}, \hat{\pi}_1 = \frac{n_{11}}{n_{10} + n_{11}},$$

and

$$\hat{\pi} = \frac{n_{01} + n_{11}}{n_{00} + n_{01} + n_{10} + n_{11}}.$$

The independence statistic is

$$LR_{IND} = -2 \ln \left[\frac{(1-\hat{\pi})^{n_{00}+n_{10}} \hat{\pi}^{n_{01}+n_{11}}}{(1-\hat{\pi}_0)^{n_{00}} \hat{\pi}_0^{n_{01}} (1-\hat{\pi}_1)^{n_{10}} \hat{\pi}_1^{n_{11}}} \right].$$

Under the null of independent violations,

$$LR_{IND} \chi_1^2.$$

Conditional coverage can additionally be assessed through

$$LR_{CC} = LR_{UC} + LR_{IND},$$

with

$$LR_{CC} \chi_2^2.$$

D.5 FZ0 Joint VaR-ES Score

Under the positive-loss convention used in this study, the FZ0 score is

$$S_{\alpha}(L_t, q_t, e_t) = \frac{q_t}{e_t} + \ln(e_t) - 1 + \frac{(L_t - q_t)_+}{(1-\alpha)e_t},$$

where

$$(x)_+ = \max(x, 0),$$

q_t is the VaR forecast, and e_t is the ES forecast.

The score is evaluated only where

$$e_t > 0$$

and

$$e_t \geq q_t.$$

Most importantly,

Lower FZ0 scores indicate better joint VaR-ES forecasts.

D.6 HAC Predictive-Accuracy Inference

For models A and B , define the paired score differential

$$d_t = S_{A,t} - S_{B,t}.$$

The average differential is

$$\bar{d} = \frac{1}{n} \sum_{t=1}^n d_t.$$

Since lower scores are preferred,

$$\bar{d} < 0$$

favours model A .

The long-run variance is estimated using a Bartlett-weighted HAC estimator with maximum lag $L = 20$:

$$\hat{\Omega} = \hat{\Gamma}_0 + 2 \sum_{\ell=1}^L \left(1 - \frac{\ell}{L+1}\right) \hat{\Gamma}_\ell,$$

where

$$\hat{\Gamma}_\ell = \frac{1}{n} \sum_{t=\ell+1}^n (d_t - \bar{d})(d_{t-\ell} - \bar{d}).$$

The standard error of the mean differential is

$$SE(\bar{d}) = \sqrt{\frac{\hat{\Omega}}{n}}.$$

The corresponding statistic is

$$z = \frac{\bar{d}}{SE(\bar{d})}.$$

D.7 Stationary Bootstrap

As a non-parametric robustness procedure, score differentials are also evaluated using a stationary bootstrap.

A total of

$$B = 2,000$$

bootstrap samples are generated. Blocks have geometrically distributed lengths with expected block length

$$20.$$

The entire paired forecast-score timeline is resampled jointly, ensuring that score differences and regime indicators remain aligned. (Diebold et al., 1998).

If $\hat{\theta}$ denotes the observed mean score difference and $\hat{\theta}^{*(b)}$ the b -th bootstrap estimate, the centered two-sided p -value is calculated from the proportion of bootstrap deviations at least as large as the observed difference.

The bootstrap conditions on the fitted forecasts. It does **not** retrain the HMM, GARCH, or diffusion models inside every bootstrap replication.

Holm adjustment is used for the pre-specified primary pairwise comparison family.

Appendix E. Additional Robustness Results

E.1 Alternative ES Confidence Levels

Table E1. Confidence-Level Robustness

Model	95% FZ0	95% Violation Rate	97.5% FZ0	97.5% Violation Rate	99% FZ0	99% Violation Rate
HS	-3.7433	6.26%	-3.4896	3.04%	-3.0999	1.30%
Regime-HS	-3.8363	4.95%	-3.6153	2.52%	-3.3239	0.96%
GARCH-FHS	-3.9323	5.39%	-3.7746	2.52%	-3.5718	0.96%
Diffusion	NA*	6.38%	-3.6060	3.86%	-3.2743	2.21%
Regime-Diffusion	NA*	6.19%	-3.5913	3.82%	-3.2657	1.98%

*One forecast date for each diffusion specification violates the strict implemented FZ0 scoring-domain condition at 95%. The observation was not silently deleted; consequently, the aggregate mean is reported as unavailable.

E.2 Historical Simulation Window Robustness

Table E2. Alternative Historical Windows

Specification	Mean FZ0	Mean ES	VaR Violation Rate
HS — 500	-3.4752	2.68%	2.95%
HS — 1,000	-3.4896	2.77%	3.04%
HS — 1,500	-3.5371	2.64%	3.13%
HS — Expanding	-3.5259	3.40%	1.56%
Regime-HS — 500	-3.6096	2.73%	2.69%
Regime-HS — 1,000	-3.6153	2.81%	2.52%
Regime-HS — 1,500	-3.6699	2.84%	2.35%
Regime-HS — Expanding	-3.6982	3.06%	1.74%

Regime-HS produces lower FZ0 scores than the corresponding HS specification under every historical-window choice.

E.3 Alternative Portfolio Composition

The robustness portfolio assigns

$$w_{\text{India}} = 0.40$$

and

$$w_i = 0.12$$

to each of the remaining five markets.

Table E3. India-Concentrated Portfolio Robustness

Model	Mean FZ0	Mean ES	VaR Violation Rate
HS	-3.4685	2.80%	2.69%
Regime-HS	-3.5925	2.83%	2.00%
GARCH-FHS	-3.8122	2.44%	2.43%
Diffusion	-3.7083	2.01%	3.09%
Regime-Diffusion	-3.6868	2.05%	2.99%

Regime-Diffusion again fails to outperform the corresponding Diffusion specification.

E.4 Seed-Level Forecast Results

Table E4. Diffusion Results by Random Seed

Model	Seed	Mean FZ0	Mean ES	VaR Violation Rate
Diffusion	11	-3.6256	1.88%	4.08%
Diffusion	29	-3.5975	2.01%	3.91%
Diffusion	47	-3.6058	2.14%	3.65%
Diffusion	71	-3.7018	2.16%	3.39%
Diffusion	101	-3.4995	1.88%	4.26%
Regime-Diffusion	11	-3.6356	2.04%	3.65%
Regime-Diffusion	29	-3.6336	2.27%	3.65%
Regime-Diffusion	47	-3.5765	1.90%	4.26%
Regime-Diffusion	71	-3.5202	2.10%	3.74%
Regime-Diffusion	101	-3.5907	2.05%	3.82%

Across seeds,

$$\bar{\text{FZ0}}_{\text{Diffusion}} = -3.6060 \pm 0.0724,$$

and

$$\bar{\text{FZ0}}_{\text{Regime-Diffusion}} = -3.5913 \pm 0.0475.$$

E.5 Five-Seed Scenario-Fidelity Robustness

Table E5. Scenario Fidelity across Diffusion Seeds

Metric	Diffusion	Regime-Diffusion	Realized Benchmark
Mean generated skewness	-0.0905	-0.0906	-1.3401
Mean excess kurtosis	3.1239	3.3691	9.7981
Mean Wasserstein distance	0.00210	0.00206	—
Mean absolute 2.5% quantile error	0.00450	0.00425	—
Correlation Frobenius error	0.8704	0.8697	—
Tail co-exceedance Brier score	0.02276	0.02290	—

The underrepresentation of negative skewness and excess kurtosis is therefore not driven by a single neural-network initialization.

E.6 HMM State-Count Robustness

Table E6. Regime-HS Performance under Alternative Numbers of HMM States

States	Mean FZ0	Mean ES	Violation Rate	Mean BIC	BIC-Best Years
2	-3.6307	2.74%	2.78%	13,507.18	0/7
3	-3.6153	2.81%	2.52%	11,205.49	0/7
4	-3.5573	2.76%	2.35%	9,731.19	7/7

The specification with the best statistical fit does not provide the best out-of-sample FZ0 performance.

Appendix F. Multi-Day Stress-Scenario Analysis

The stress extension examines the three largest realized 10-session losses selected mechanically from the observed portfolio series.

For each episode, Regime-Diffusion is simulated under persistent Stress-state conditioning using

$$5,000$$

paths over a 10-session horizon.

The exercise is a **conditional what-if analysis**. It is not a calibrated 10-day VaR/ES backtest and should not be interpreted as a regulatory market-risk calculation.

Table F1. Stress-Test Summary

Epi- sode	Forecast Origin	Realized Period	Actual 10- Session Loss	Regime-Diffusion VaR 97.5%	Regime-Diffusion ES 97.5%	Worst Generated Loss	Probability of Generat- ed Loss Exceeding Actual
1	22 Jul 2024	23 Jul–5 Aug 2024	8.22%	2.30%	3.09%	6.65%	0.00%
2	11 Jan 2022	12–25 Jan 2022	3.16%	2.99%	3.81%	7.07%	2.00%
3	8 Feb 2023	9–22 Feb 2023	2.61%	4.79%	5.73%	9.65%	16.02%

For Episode 1,

$$L_{\text{actual}} = 8.22\%$$

exceeds the worst Regime-Diffusion simulated loss,

$$L_{\text{max}}^{\text{sim}} = 6.65\%.$$

None of the 5,000 generated paths is more severe than the observed outcome.

By contrast, Episodes 2 and 3 both fall within the simulated support. The evidence is therefore mixed: the model is capable of generating losses more severe than two selected realized episodes but fails to reproduce the severity of the largest event.

Episode 1: conditional 10-session stress application

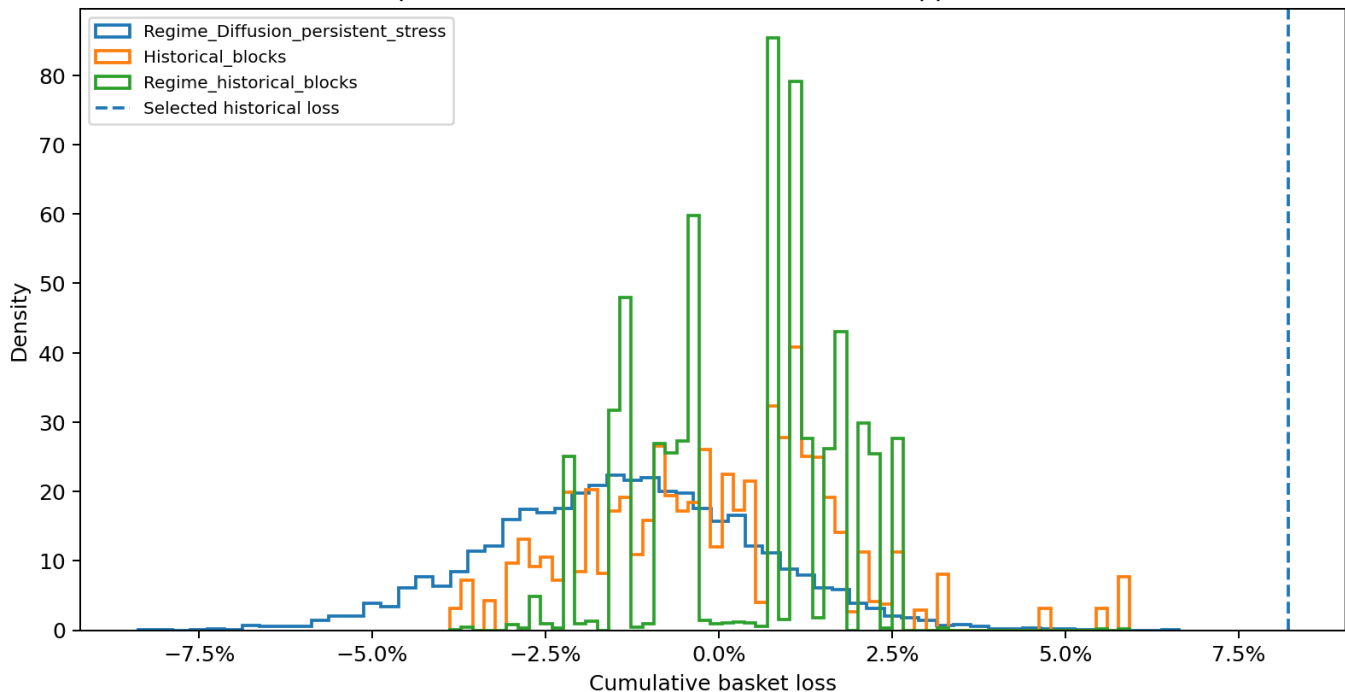


Figure F1. Episode 1 — Distribution of Simulated 10-Session Losses

Episode 1: 25 persistent-stress scenario paths

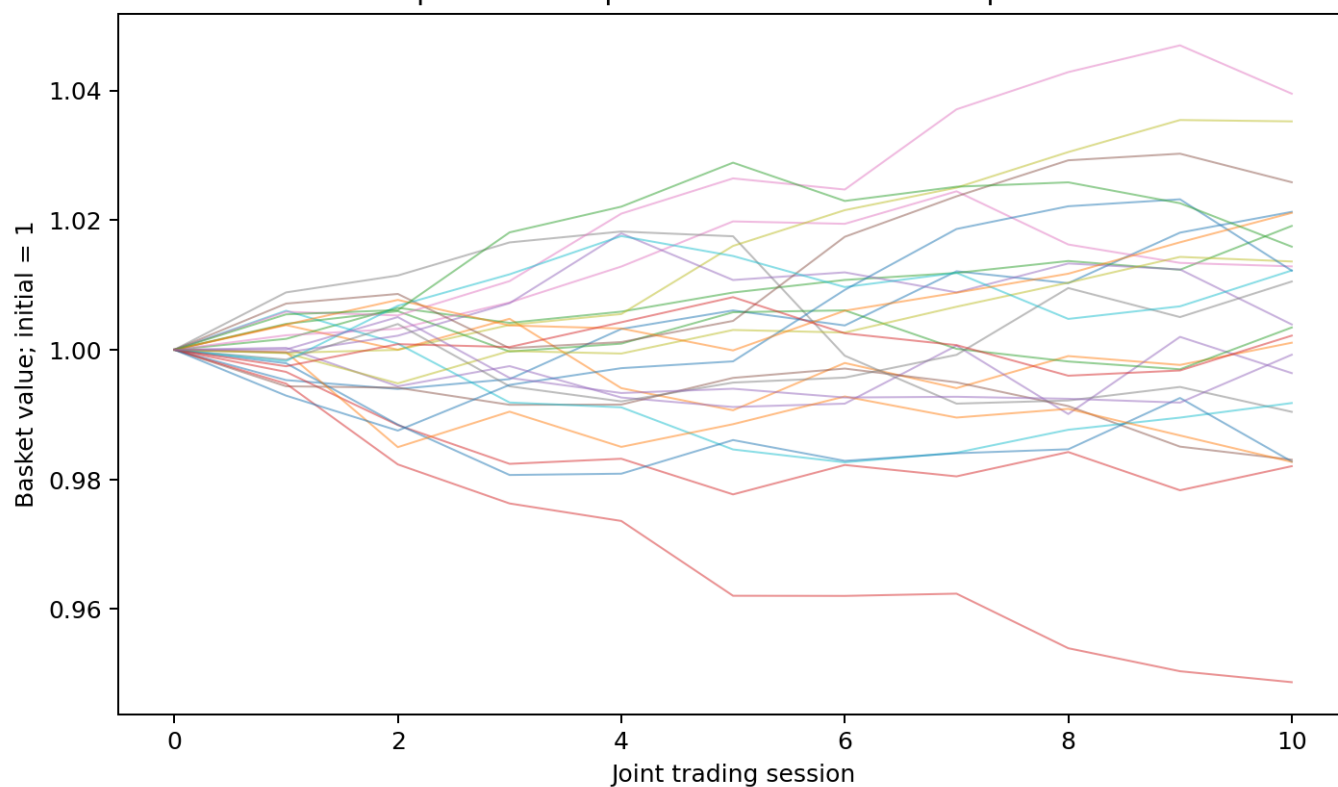


Figure F2. Episode 1 — Simulated Stress Paths and Realized Outcome

Episode 2: conditional 10-session stress application

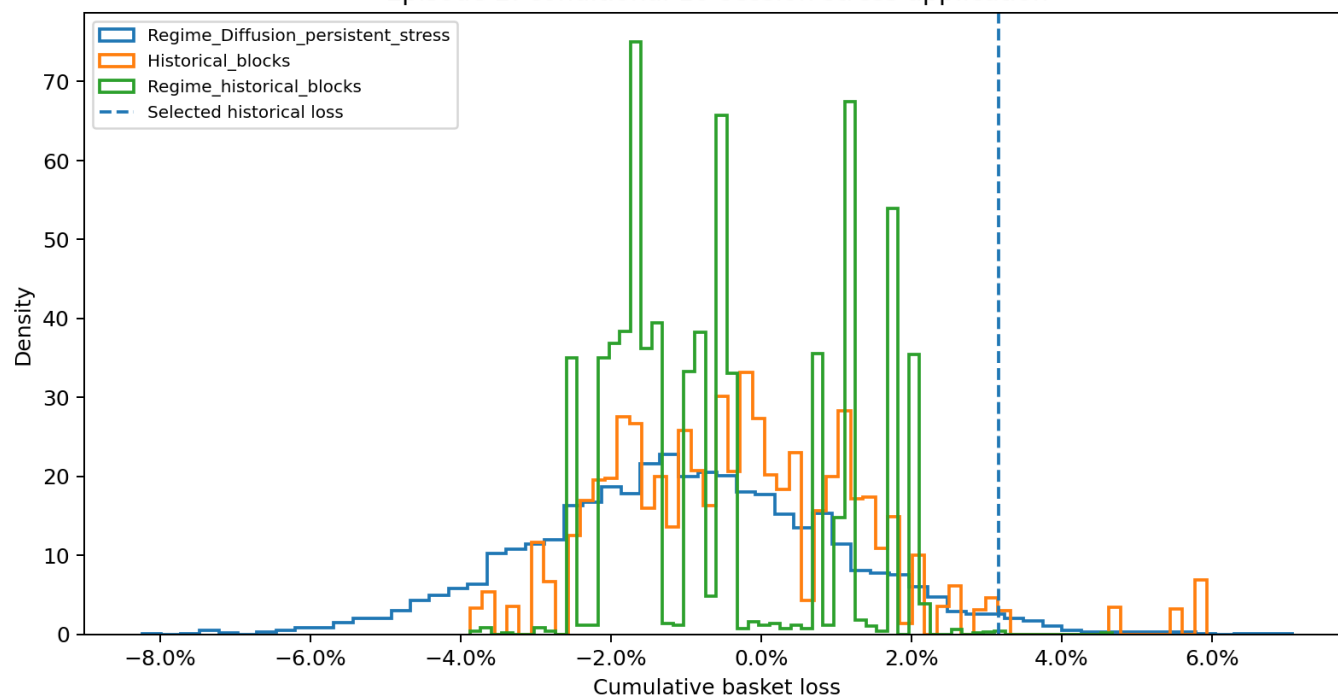


Figure F3. Episode 2 — Distribution of Simulated 10-Session Losses

Episode 2: 25 persistent-stress scenario paths

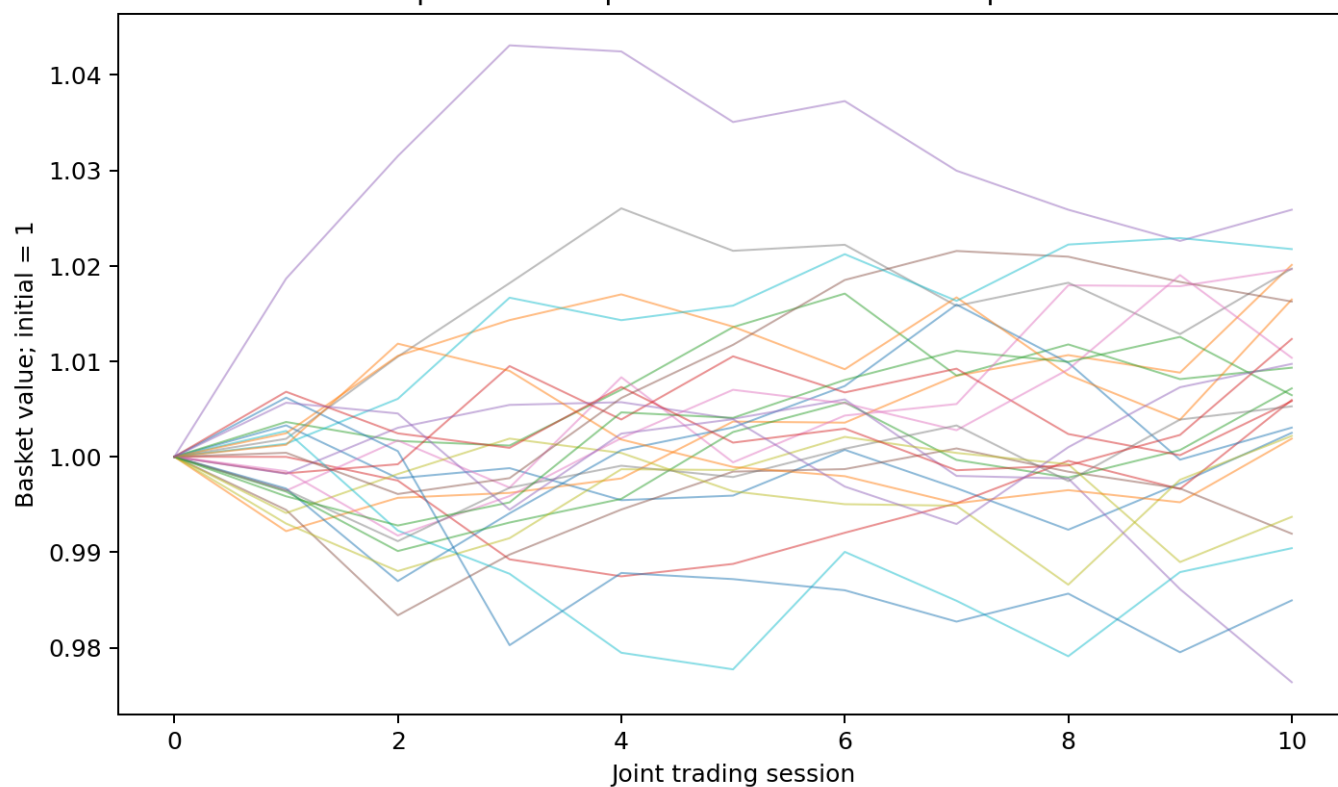


Figure F4. Episode 2 — Simulated Stress Paths and Realized Outcome

Episode 3: conditional 10-session stress application

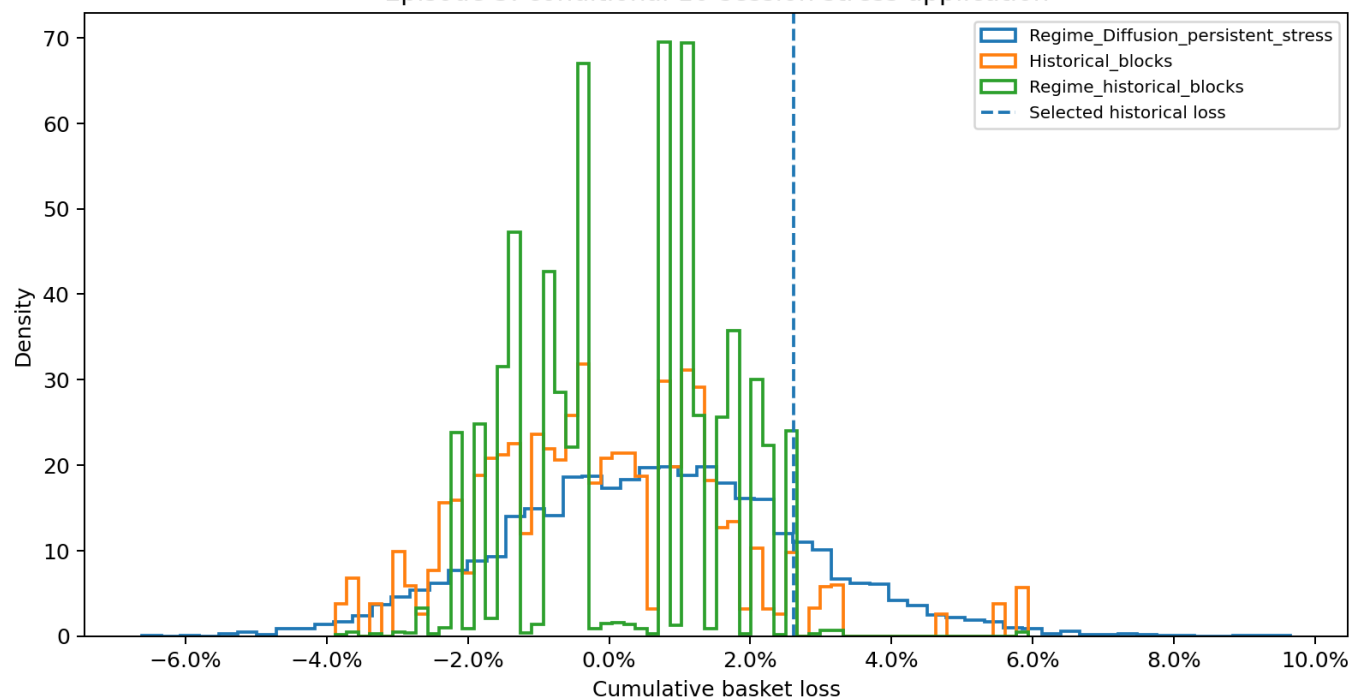


Figure F5. Episode 3 — Distribution of Simulated 10-Session Losses

Episode 3: 25 persistent-stress scenario paths

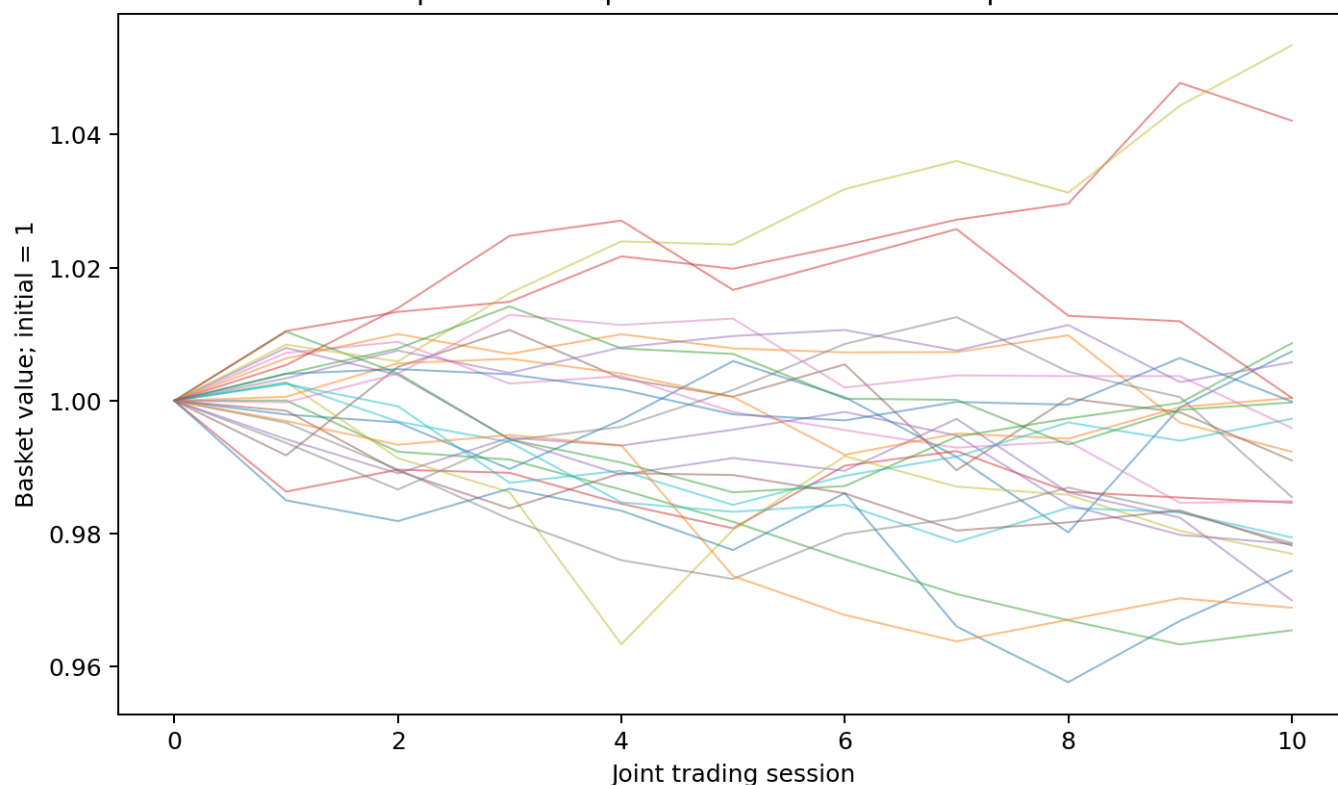


Figure F6. Episode 3 — Simulated Stress Paths and Realized Outcome

The figures should remain in the supplement rather than the main manuscript because the stress exercise is illustrative rather than part of the primary forecast-validation design.

Appendix G. Reproducibility and Computational Environment

G.1 Frozen Research Configuration

The final empirical analysis uses the following configuration.

Table G1. Core Research Parameters

Parameter	Final Value
Requested sample	1 Jan 2007–31 Aug 2026
Validation period	2017–2019
Final test period	7 Jan 2020–31 Aug 2026
OOS forecast dates	1,151
Primary ES confidence	97.5%
Alternative confidence levels	95%, 99%
HS window	1,000
Rolling volatility window	20 sessions
HMM states	3
HMM restarts	3
HMM maximum iterations	120
Regime-HS uniform shrinkage	5%
Diffusion scenarios	5,000 per date
Diffusion steps	100
DDIM sampling steps	20
Random seeds	11, 29, 47, 71, 101
Bootstrap replications	2,000
Expected bootstrap block length	20
Stress horizon	10 sessions
Stress scenarios	5,000
Stress episodes	3

The HMM and diffusion models are **refitted annually**. For forecast year y , only data dated prior to 1 January of year y are used for estimation.

G.2 Computational Environment

The original completed run recorded the following environment.

Table G2. Software Environment

Component	Version
Python	3.13.15
Operating platform	Linux x86-64
NumPy	2.1.3
pandas	2.2.3
SciPy	1.16.3
scikit-learn	1.6.1
PyTorch	2.11.0+cpu
Numba	0.61.2
Matplotlib	3.10.0
openpyxl	3.1.5
yfinance	0.2.66
nbformat	5.11.1
CUDA available	No

The original research run was therefore executed using CPU-based PyTorch.

G.3 Data and Result Freezing

The source data were retrieved on **15 September 2026**, while the final observation permitted in the empirical dataset is **31 August 2026**.

The frozen source-data hash is

426c16cc946ae86d6a8d19f1cada9d3ba8d5f84a3d6c072ce527609448e116b9.

The final research package contains:

- processed market data;
- matched forecast targets;
- filtered regime probabilities;
- daily model forecasts;
- diffusion seed results;
- scenario-fidelity diagnostics;
- HMM robustness results;
- confidence-level robustness;
- historical-window robustness;
- alternative-portfolio results;
- stress-test results;
- manuscript tables and figures; and
- the analysis code and configuration files.