# Risk Governance Through Long-Term Risk Modelling: An Enhanced Filtered Historical Simulation Approach for Financial Institutions

G. Barone-Adesi* M. Bonollo† V. Damato‡ F. Luce§

January 22, 2026

**Abstract**

Financial institutions must produce coherent tail-risk measures across multiple regulatory horizons—from short-term market-risk monitoring to longer-term capital and solvency assessments—under stringent model-risk governance expectations. We develop a governance-oriented framework within the historical-simulation family that preserves empirical cross-sectional dependence through multivariate resampling while remaining operationally scalable for large portfolios. We compare standard historical bootstrap engines with filtered historical simulation architectures and introduce LT-FHS, a long-horizon extension that constrains cumulative shocks using historically observed levels and volatility-conditioned dynamic buffers. The empirical assessment combines portfolio-level out-of-sample backtesting at market-risk horizons with a risk-factor-level tail-quantile accuracy exercise that benchmarks simulated tails against an empirical resampling proxy across horizons. Covering interest-rate, credit-spread, and equity risk classes, filtered architectures are consistently competitive at market-risk horizons, while LT-FHS delivers the best long-horizon tail-quantile accuracy and the closest match to empirical tail benchmarks, with gains concentrated in stressed regimes. These results provide a practical, regulator-aligned playbook for selecting scenario engines over different holding periods in banking and insurance internal-model settings.

**Keywords:** Risk Governance; Long-term Risk Modelling; Economic Capital; ICAAP; Solvency; Filtered Historical Simulation.
**JEL classifications:** C14; C22; C53; G17; G21; G32.

## 1 Introduction

Financial institutions operate under a regulatory and supervisory environment that increasingly demands *multi-horizon* and *internally consistent* risk measurements. Moreover, supervisory frameworks require institutions to reconcile *short-horizon* market-risk monitoring with *long-horizon* capital-adequacy and solvency requirements. For banking institutions, market-risk metrics and model governance are anchored in the Basel framework (from the 1996 market-risk amendment to the 2019 Fundamental Review of the Trading Book rules), which compels coherent risk measurement across multiple holding periods (see Basel Committee on Banking Supervision [1996, 2019]). Furthermore, the Internal Capital Adequacy Assessment Process (ICAAP) and Economic Capital (ECAP) frameworks establish a forward-looking assessment of capital needs that supervisors expect to be anchored to a *one-year* horizon for capital planning, stress design, and Pillar 2 determinations, in line with the European Central Bank (ECB) *Guide*

*Swiss Finance Institute, Università della Svizzera italiana (USI), Lugano. Email: giovanni.baroneadesi@usi.ch
†Iason Ltd., Risk Management Senior Consultant. Email: michele.bonollo@iasonltd.com
‡Banco BPM Group, Head of Market and Counterparty Risk Unit. Email: vito.damato@bancobpm.it
§Banco BPM Group, Market Risk Specialist. Email: federico.luce@bancobpm.it

*to the ICAAP* [European Central Bank, 2018a]. This supervisory expectation complements the broader market-risk rulebook, ensuring that the long-horizon ICAAP view remains aligned with the short-horizon metrics used for regulatory market risk. Similarly, for insurance undertakings, the annual horizon is a formal regulatory requirement: under Solvency II, the Solvency Capital Requirement (SCR) is defined as a *one-year* Value-at-Risk at the 99.5% confidence level [European Commission, 2015], making annual-horizon tail-risk measurement a core pillar of prudential oversight in the insurance sector. As a result, financial institutions must adopt model-governance practices that remain testable and informative at short horizons, yet deliver empirically plausible and regulator-defensible tail metrics over an annual horizon, despite the empirical scarcity and regime uncertainty inherent in long holding periods. This tension further heightens the need for transparent scenario-generation, robust tail calibration, and auditable aggregation mechanisms within a unified multi-horizon framework.

In light of the regulatory and supervisory landscape outlined above, a modern risk-governance system, as articulated in the European regulatory framework, places risk quantification at the core of internal capital adequacy and supervisory assessment. Article 73 of the Capital Requirements Directive [crd, 2013] requires institutions to implement an Internal Capital Adequacy Assessment Process (ICAAP) capable of identifying, measuring, aggregating, and monitoring all material risks on a forward-looking basis. This requirement is operationalized through the Supervisory Review and Evaluation Process (SREP), as specified in the EBA Guidelines on SREP [Authority, 2014, 2018], under which supervisors assess the soundness of a bank's risk-management framework, the credibility of its capital planning, and the robustness of its stress-testing and scenario-generation methodologies. The ECB's supervisory expectations—as detailed in the *Guide to the ICAAP* and reinforced through on-site inspections (OSI) and thematic reviews [European Central Bank, 2018b]—emphasize that risk measurement must be transparent, reproducible, and internally consistent across risk classes and horizons, and must be embedded within the institution's decision-making architecture through its Risk Appetite Framework (RAF) [European Banking Authority, 2021], limit system, and internal capital policy. Within this governance structure, risk quantification serves as the quantitative pivot linking risk identification to capital adequacy and management oversight, functioning as the mechanism through which supervisory expectations are rendered into measurable, auditable outputs across organizational layers. A governance-aligned measurement architecture must therefore deliver (i) methodological soundness and empirical credibility, including realistic tail behavior; (ii) supervisory defensibility, with assumptions, parameters, and scenario engines that withstand audit and OSI scrutiny; (iii) operational sustainability and scalability in high-dimensional portfolios; and (iv) horizon consistency, ensuring that short-horizon metrics used for market-risk monitoring remain coherent with long-horizon assessments required for capital planning under ICAAP and SREP. It is against this governance backdrop that we assess scenario-generation paradigms and their suitability for multi-horizon, multi–risk-class applications.

Operationally, a natural starting point for the implementation of a risk-quantification framework embedded within a modern risk-governance system is the set of standard scenario-generation paradigms: Gaussian-parametric approximations, Monte Carlo simulation, and historical simulation. These three families crystallize distinct modelling philosophies—analytic tractability, stochastic completeness, and empirical grounding—each offering different trade-offs between statistical realism, computational complexity, and governance defensibility. From an institutional perspective, however, the critical question is not only statistical accuracy, but also whether a methodology remains operationally scalable, transparent, and auditable while maintaining tail fidelity across heterogeneous risk classes. In this paper, we therefore adopt a governance lens that evaluates engines jointly along (i) tail realism, (ii) validation feasibility across horizons, and (iii) operational efficiency in high-dimensional portfolios. Our focus is the historical-simulation family. We study two baseline historical bootstrap engines (HS-iid and HS-$\lambda$) and a filtered historical simulation (FHS) family that combines a common econometric filtering layer with mul-

tivariate residual-vector resampling. Within FHS, we consider the seminal recursive-volatility formulation of Barone-Adesi et al. (denoted *FHS-Recursive*) and a static-volatility rescaling variant (*FHS-Static*). To address the annual-horizon tail-risk problem, we introduce *LT-FHS*, an extension of FHS-Static that applies a dynamic bounding mechanism calibrated on historical level constraints and realized-volatility buffers. Empirically, we implement two complementary validation blocks in order to test the performance of the models under consideration. First, we perform portfolio-level out-of-sample backtesting for 5, 10, and 21 business days holding-periods using canonical exceedance-based tests, and synthesize the results across engines, risk classes, and horizons through the introduction of a Composite Risk Governance Score (CRGS) that aggregates multiple test outcomes into a single, governance-oriented ranking signal. Second, we conduct a risk-factor-level tail calibration exercise for horizons 5, 10, 21, and 252 trading days by benchmarking simulated cumulative shocks against an empirical block-bootstrap proxy and ranking engines via a tail-quantile loss. To provide a governance-oriented, quantitative measure of tail plausibility across horizons, we introduce the Tail Risk Governance Score (TRGS), which measures the tail-quantile calibration performance of each engine. This structure aligns the paper with the institutional need to maintain a single, coherent scenario-generation architecture while acknowledging that the feasible validation tools differ sharply across horizons.

The remainder of the paper is organized as follows. Section 2 reviews the most widely used risk quantification paradigms and their standard implementations through an institutional governance lens. Section 3 defines the HS and FHS engines under a common multivariate bootstrap architecture and introduces LT-FHS. Section 4 details the empirical design and validation framework, while Section 5 reports backtesting and tail-calibration results. Section 6 concludes.

## 2 Literature and Standard Implementations

The development of quantitative market-risk measurement has historically followed a progression from parametric, to semi-parametric, to non-parametric scenario-generation frameworks. Early Value-at-Risk (VaR) engines were grounded in variance–covariance methods and linear portfolio approximations, which gained rapid traction due to their compatibility with modern portfolio theory and their ease of implementation [Barone-Adesi and Giannopoulos, 2001]. These models relied on the assumption that risk-factor changes were jointly normal and that portfolio risk could be fully characterized by first and second moments. Their operational simplicity made them attractive for institutions building the first generation of internal models under the Basel 1996 amendment. However, empirical evidence accumulated during major market disruptions—including the Global Financial Crisis and the Lehman default (2008), the euro-area sovereign debt crisis (2010–2012), the Brexit referendum shock (2016), the U.S. presidential election surprise (2016), the COVID-19 crash (2020), the Russia–Ukraine war and the ensuing European energy crisis (from 2022), and, more broadly, the global inflation surge and the most synchronized monetary tightening in decades (2021–2023)—revealed that Gaussian-based VaR models systematically understated tail risk, failed to capture synchronous cross-asset crashes, and were operationally burdensome when large covariance structures had to be maintained. These limitations motivated the development of simulation-based approaches. Monte Carlo methods relaxed the Normality assumption by introducing explicit—and potentially far more sophisticated—stochastic dynamics for risk factors, whereas historical simulation (HS) adopted an empirical, non-parametric philosophy by replaying past shocks directly. As emphasized by Barone-Adesi and Giannopoulos [2001], these techniques were intended to address the inadequacies of purely parametric VaR models: HS preserved nonlinearities and contemporaneous co-movements without requiring covariance estimation, whereas Monte Carlo simulation allowed richer dynamics and complex dependence structures. Nonetheless, both approaches introduced new challenges: Monte Carlo models inherit the misspecification risk of the chosen stochastic structure, while HS is highly sensitive to non-stationarity, window selection, and the absence

of unprecedented shocks in the historical sample. Against this backdrop, the standard toolkit for risk measurement converged around three canonical paradigms: Gaussian Parametric models (GP), Monte Carlo simulation (MCS), and Historical Simulation (HS). Each embodies a distinct modelling philosophy—analytic tractability, stochastic flexibility, and empirical grounding—and each poses governance-relevant trade-offs in scalability, transparency, stability, and tail calibration. We review these paradigms below, highlighting the structural features most relevant for multi-risk-class and long-horizon applications.

**Gaussian Parametric Models (GP).** Canonical references include RiskMetrics and VaR practice-oriented treatments, as well as the foundational literature on coherent risk measures and expected shortfall (see Longerstaey and Spencer [1996], Artzner et al. [1999], Acerbi and Tasche [2002], Jorion [2007], Acerbi and Szekely [2017]). Gaussian parametric models (also referred to as variance–covariance approaches) assume that risk factor returns (or portfolio PnL) are normally distributed, with risk fully characterized by first and second moments. The primary advantage of GP methods lies in their computational efficiency and their transparency: once the mean vector and the variance–covariance matrix are specified, $VaR$ can be computed quickly, enabling high-frequency risk monitoring. However, GP methods are subject to two fundamental limitations from a long-horizon tail-risk and governance perspective. First, the normality assumption is often inconsistent with empirical return features such as fat tails, skewness, and volatility clustering. This mismatch tends to be most severe in the tail region, precisely where capital-relevant risk measures concentrate. Second, GP approaches typically require the identification, estimation, and maintenance of a large variance–covariance (or correlation) matrix across the risk factor universe. In realistic institutional settings, this matrix can be high dimensional and structurally unstable, particularly when risk factors span multiple heterogeneous risk classes. As a result, sampling error, regime sensitivity, and frequent recalibration introduce material model risk governance burdens and ongoing monitoring requirements.

**Monte Carlo Simulation (MCS).** Standard references include Glasserman [2003] and modern quantitative risk management texts (see McNeil et al. [2015]). Monte Carlo simulation replaces the Gaussian distributional shortcut with an explicit stochastic model for the joint dynamics of underlying risk factors or portfolio values. In principle, MCS can incorporate complex nonlinearities, path-dependence, and instrument-specific valuation features. It also offers a natural framework for conditioning and stress analysis, which may be relevant when institutions embed forward-looking narratives within capital planning. In practice, the accuracy of MCS depends critically on model specification and calibration quality: misspecification of dynamics, dependence structures, or parameters translates directly into biased tail estimates. Moreover, large-scale tail risk measurement typically requires very large scenario sets and repeated valuation or revaluation steps, which can be computationally expensive. The resulting operational cost, together with the need for parameter monitoring, stability checks, sensitivity analysis, change management, and independent validation, often leads to a high model governance overhead. These features make MCS powerful but frequently challenging to deploy as a standard engine for multi-risk-class tail estimation at scale, especially when extended holding periods are required.

**Historical Simulation (HS).** Historical simulation and its widely used enhancements (including weighted and hybrid variants) are discussed in, among others, Hendricks [1996], Hull and White [1998], Boudoukh et al. [1998], Pritsker [2001]. Historical simulation is a nonparametric approach that constructs the loss distribution by re-applying realized historical shocks to current positions, thereby reducing reliance on strong parametric assumptions. This feature makes HS appealing from a governance standpoint: the method is transparent, the link between observed market moves and simulated PnL outcomes is direct, and cross-sectional dependence

is embedded through joint historical scenarios without requiring an explicit correlation matrix. Nevertheless, classical HS relies on the implicit assumption that the past provides a representative proxy for the future. Consequently, HS is sensitive to the selection of the historical window and may be fragile under non-stationarity and regime changes. Moreover, standard HS does not explicitly address volatility clustering unless enhanced, and it cannot represent unprecedented shocks that are absent from the historical sample. These limitations become particularly acute when the holding period is extended, as tail estimation becomes increasingly sensitive to non-stationarity and to the scarcity of extreme observations.

Table 1: Pros and cons of canonical risk quantification approaches.

| **Approach** | **Pros** | **Cons** |
|---|---|---|
| Gaussian Parametric (GP) | • Very fast computation; suitable for high-frequency monitoring.<br>• High transparency and interpretability (moment-based).<br>• Straightforward implementation and reporting. | • Strong distributional assumptions; weak tail realism (fat tails/skewness).<br>• Reliance on large covariance/correlation estimation at scale (instability, sampling error).<br>• Limited capability to represent conditional heteroskedasticity unless extended. |
| Monte Carlo Simulation (MCS) | • High modelling flexibility; can represent nonlinearities and complex instruments.<br>• Natural framework for dynamic risk drivers and scenario conditioning.<br>• Extensible to stress scenarios and structural what-if analyses. | • Heavy computational burden for tail measures (large scenario sets + repricing).<br>• High model risk exposure (specification, calibration, parameter instability).<br>• Governance-intensive: validation, change management, and sensitivity monitoring. |
| Historical Simulation (HS) | • Nonparametric; reduced reliance on distributional assumptions.<br>• Transparent link to realized historical market moves.<br>• Naturally embeds cross-sectional dependence within joint historical scenarios. | • Window-length sensitivity; limited representativeness under regime shifts.<br>• Classical HS does not address volatility clustering unless enhanced.<br>• Limited ability to represent unprecedented shocks absent in the sample. |

**Pros and Cons: A Synthesis for Governance and Implementation.** The governance discussion is framed by regulatory holding-period requirements and internal capital processes. Tables 1 and 2 summarize the key strengths and limitations of the three canonical approaches. The first table provides a concise pros/cons overview. The second table emphasizes governance-relevant considerations, including model risk drivers, validation feasibility, dependence handling at scale, and operational efficiency. Together, these syntheses motivate the focus on the historical

simulation family adopted in the remainder of this paper, and they clarify why extensions such as filtered historical simulation are particularly relevant when institutions must jointly satisfy tail fidelity, scalability, and sustainable model governance in a high-dimensional, multi-risk-class environment.

Table 2: Governance-oriented comparison: model risk burden, validation feasibility, and operational efficiency.

| **Dimension** | **GP** | **MCS** | **HS** |
|---|---|---|---|
| Primary model risk drivers | Distributional assumptions; covariance/correlation estimation risk | Model specification; calibration and parameter instability; dependence modelling choices | Representativeness of historical window; non-stationarity; sampling scarcity in tails |
| Validation feasibility | High on mechanics; weaker on tail realism | Conceptually strong but practically demanding; many degrees of freedom | Moderate: transparent, but sensitive to window choice and regime changes |
| Dependence handling at scale | Explicit correlation matrix required; high-dimensional estimation burden | Explicit dependence specification/calibration required (often high-dimensional) | Implicit via joint historical moves; constrained by sample size and regimes |
| Computational efficiency | High | Low to moderate (tail risk often expensive) | High to moderate (depends on repricing and scenario set size) |
| Governance overhead (change/monitoring) | Moderate to high (covariance maintenance, assumption monitoring) | High (calibration cycles, sensitivity checks, model changes) | Moderate (window policy, stability monitoring, overlays/stress complements) |

**Governance rationale.** Table 2 compares the three canonical scenario engines discussed in this paper through the dimensions that typically determine whether a model is implementable and maintainable under a risk governance process: (i) the primary sources of model risk, (ii) the practical feasibility of validation (especially for tail quantities), (iii) dependence handling at scale, (iv) computational and operational efficiency, and (v) governance overhead associated with change and monitoring. In production, the binding constraint is rarely "maximum flexibility"; rather, it is the ability to provide repeatable evidence to Model Risk Management, internal audit, and supervisors under finite data and tight change-control. GP and MCS can be theoretically appealing, but they move the governance burden onto distributional assumptions, high-dimensional covariance/dependence modelling, and recurring calibration choices—each difficult to defend in the tails and expensive to monitor as the risk universe grows. HS, by contrast, derives scenarios from observed joint moves, reducing discretionary degrees of freedom, embedding cross-asset dependence mechanically, and supporting pragmatic validation via backtesting and empirical benchmarking. The governance problem then becomes an explicit, testable policy choice: how the historical window is defined, how non-stationarity and regime shifts are monitored, and which controlled overlays (e.g., stress complements or filtered/volatility-rescaled variants) are permitted to address tail scarcity and volatility clustering. For these reasons, when the framework must be implemented end-to-end in the "world of facts", the HS family typically offers the most defensible trade-off between model risk burden, validation feasibility, and operational efficiency.

# 3 Methodology: Scenario-Generation Models

This section formalizes the scenario-generation methodology for the historical-simulation engines adopted in this paper. We consider five engines within a common multivariate resampling architecture: (i) i.i.d. bootstrap historical simulation (**HS-iid**), (ii) exponentially weighted historical simulation (**HS-$\lambda$**), (iii) filtered historical simulation with recursive volatility propagation (**FHS-Recursive**, following Barone-Adesi et al.), (iv) filtered historical simulation with static volatility rescaling (**FHS-Static**), and (v) a long-term extension of FHS-Static with dynamic bounding (**LT-FHS**). All engines preserve empirical cross-sectional dependence by resampling *joint* historical vectors rather than relying on explicitly estimated correlation matrices.

## 3.1 Historical Simulation: direct bootstrap of historical shocks

### 3.1.1 Historical database and rolling information set

Let $\{\mathbf{x}_t\}_{t=1}^{T}$ denote the historical series of multivariate daily shocks, where

$$\mathbf{x}_t = \big(x_{t,1}, \ldots, x_{t,d}\big)^\top \in \mathbb{R}^d, \qquad t = 1, \ldots, T, \tag{1}$$

and $d$ is the total number of risk factors in scope (across heterogeneous risk classes). At a generic computation date $\tau$, we define the dynamic information set

$$\mathcal{H}_\tau := \{\mathbf{x}_t : t \le \tau^-\}, \tag{2}$$

and let $n = n(\tau)$ be the number of observations available in $\mathcal{H}_\tau$. For notational simplicity, we re-index the available history as $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, where $\mathbf{x}_n$ is the most recent observation prior to $\tau$.

### 3.1.2 Common bootstrap path architecture

Fix a holding period (path length) $H \in \mathbb{N}$ and a number of scenarios $M \in \mathbb{N}$. For each scenario $m = 1, \ldots, M$, we generate an $H$-step path by sampling (with replacement) indices

$$I_1^{(m)}, \ldots, I_H^{(m)} \in \{1, \ldots, n\}, \tag{3}$$

according to a discrete probability mass function $\{w_i\}_{i=1}^{n}$. The resulting multivariate path of shocks is

$$\mathbf{X}^{(m)} := \begin{bmatrix} \mathbf{x}_{I_1^{(m)}}^\top \\ \mathbf{x}_{I_2^{(m)}}^\top \\ \vdots \\ \mathbf{x}_{I_H^{(m)}}^\top \end{bmatrix} \in \mathbb{R}^{H \times d}. \tag{4}$$

All HS-family engines share (4) and differ *only* in the specification of the sampling weights $\{w_i\}$.

### 3.1.3 i.i.d. bootstrap historical simulation

In the i.i.d. bootstrap historical simulation framework (HS-iid), sampling is governed by the discrete uniform distribution over historical dates:

$$w_i^{\text{iid}} = \frac{1}{n}, \qquad i = 1, \ldots, n, \tag{5}$$

so that, for each scenario $m$ and step $h$

$$I_h^{(m)} \sim \text{Categorical}\Big(w_1^{\text{iid}}, \ldots, w_n^{\text{iid}}\Big) \equiv \text{Uniform}\{1, \ldots, n\}. \tag{6}$$

### 3.1.4 Exponentially weighted bootstrap historical simulation

Building on the idea originally proposed by Boudoukh et al. [1998], and harmonizing the methodology with multiperiod bootstrap-based risk measurement models, in order to increase responsiveness to recent market conditions, the exponentially weighted bootstrap historical simulation (HS-$\lambda$) framework assigns probabilities proportional to an exponential time-decay. Let $\lambda \in (0,1)$ be the decay factor and let $i = n$ index the most recent observation. Define unnormalized time-decay weights

$$\tilde{w}_i(\lambda) = (1-\lambda)\lambda^{\,n-i}, \qquad i = 1, \ldots, n, \tag{7}$$

and normalize them to obtain a probability mass function:

$$w_i^{\lambda} = \frac{\tilde{w}_i(\lambda)}{\sum_{j=1}^{n} \tilde{w}_j(\lambda)} = \frac{(1-\lambda)\lambda^{\,n-i}}{1-\lambda^n}, \qquad i = 1, \ldots, n. \tag{8}$$

Trajectory construction follows a multinomial/categorical resampling logic: for each scenario $m$ and step $h$,

$$I_h^{(m)} \sim \text{Categorical}\Big(w_1^{\lambda}, \ldots, w_n^{\lambda}\Big), \qquad \text{independently across } h \text{ and } m. \tag{9}$$

Equivalently, conditional on the probability vector $(w_1^{\lambda}, \ldots, w_n^{\lambda})$, each $H$-step path is the realization of $H$ independent draws whose probabilities are proportional to the exponential time-decay profile.

## 3.2 Filtered Historical Simulation (FHS): filtering, residual bootstrap, and volatility rescaling

### 3.2.1 FHS family common filtering layer

While HS engines resample raw shocks $\mathbf{x}_t$, FHS engines resample *standardized* shocks and then rescale them to a target conditional volatility state. For each risk factor $j \in \{1, \ldots, d\}$, consider the mean–volatility decomposition

$$x_{t,j} = \mu_{t,j} + \varepsilon_{t,j}, \qquad \varepsilon_{t,j} = \sigma_{t,j}\, z_{t,j}, \tag{10}$$

where $\mu_{t,j}$ denotes the conditional mean, $\sigma_{t,j} > 0$ the conditional volatility, and $z_{t,j}$ is the standardized residual with zero mean and unit variance.

In this paper, all filtered engines (FHS-Static, FHS-Recursive, and LT-FHS) share the same econometric filtering layer. Following the foundational filtered-historical-simulation framework of Barone-Adesi and co-authors, we specify, for each risk factor $j$, an AR(1) mean equation without intercept coupled with a GJR-GARCH(1,1) conditional volatility dynamics. The volatility specification captures leverage and asymmetry in a parsimonious form and is standard in risk applications (see Engle [1982], Bollerslev [1986], Glosten et al. [1993], Barone-Adesi et al. [1999]). We write the common filtering layer explicitly as

$$x_{t,j} = \phi^{(j)} x_{t-1,j} + \varepsilon_{t,j}, \quad \mathbb{E}[\varepsilon_{t,j} \mid \mathcal{F}_{t-1}] = 0, \quad \mathbb{V}ar(\varepsilon_{t,j} \mid \mathcal{F}_{t-1}) = (\sigma_{t,j})^2, \tag{11}$$

$$(\sigma_{t,j})^2 = \omega^{(j)} + \alpha^{(j)}\,(\varepsilon_{t-1,j})^2 + \beta^{(j)}\,(\sigma_{t-1,j})^2 + \gamma^{(j)}\,(\varepsilon_{t-1,j})^2\,\mathbb{I}_{\{\varepsilon_{t-1,j}<0\}}. \tag{12}$$

where $\mathcal{F}_{t-1}$ is the information set and $\hat{z}_{t,j} := \varepsilon_{t,j}/\sigma_{t,j}$ are the standardized residuals used for multivariate residual-vector resampling. Let $\{\hat{z}_{t,j}\}_{t=1}^{n}$ be the estimated standardized residuals and define the multivariate residual vectors

$$\hat{\mathbf{z}}_t := \big(\hat{z}_{t,1}, \ldots, \hat{z}_{t,d}\big)^{\top} \in \mathbb{R}^d, \qquad t = 1, \ldots, n. \tag{13}$$

Resampling the *joint* vectors $\hat{\mathbf{z}}_t$ preserves empirical cross-sectional dependence across risk factors and risk classes, thereby mitigating the need to specify and maintain high-dimensional correlation matrices.

### 3.2.2 Bootstrap of standardized residual vectors (shared by FHS variants)

FHS scenario paths are generated by resampling indices exactly as in HS-iid, but applied to $\{\hat{\mathbf{z}}_t\}_{t=1}^n$ instead of $\{\mathbf{x}_t\}_{t=1}^n$. For each scenario $m$ and step $h$, sample

$$I_h^{(m)} \in \{1,\dots,n\} \quad \text{from } \{w_i\}_{i=1}^n \text{ with } w_i = \frac{1}{n}\ \forall i, \qquad \mathbf{z}_h^{(m)} := \hat{\mathbf{z}}_{I_h^{(m)}} \in \mathbb{R}^d. \tag{14}$$

Let $\boldsymbol{\sigma}_0(\tau) := (\sigma_{0,1},\dots,\sigma_{0,d})^\top$ denote the target initial conditional volatility vector at date $\tau$ (e.g., the last estimated conditional volatility prior to $\tau$).

### 3.2.3 FHS-Recursive (recursive volatility propagation)

This is the seminal filtered-historical-simulation formulation proposed in Barone-Adesi et al. [1999] and extended in subsequent work Barone-Adesi et al. [2002, 2017]. In the recursive-volatility FHS engine, conditional volatility is updated along the simulated path via a factor-wise volatility recursion. Let $\sigma_{1,j}^{(m)} = \sigma_{0,j}$. For $h \geq 2$ and each risk factor $j$,

$$(\sigma_{h,j}^{(m)})^2 = \omega^{(j)} + \alpha^{(j)}\big(\varepsilon_{h-1,j}^{(m)}\big)^2 + \beta^{(j)}\big(\sigma_{h-1,j}^{(m)}\big)^2 + \gamma^{(j)}\,\big(\varepsilon_{h-1,j}^{(m)}\big)^2\,\mathbb{I}_{\{\varepsilon_{h-1,j}^{(m)}<0\}}, \tag{15}$$

$$\varepsilon_{h,j}^{(m)} = \sigma_{h,j}^{(m)}\, z_{h,j}^{(m)}. \tag{16}$$

In vector form, with $\boldsymbol{\sigma}_h^{(m)} = (\sigma_{h,1}^{(m)},\dots,\sigma_{h,d}^{(m)})^\top$ and $\boldsymbol{\varepsilon}_h^{(m)} = \mathrm{diag}(\boldsymbol{\sigma}_h^{(m)})\mathbf{z}_h^{(m)}$, a state propagation representing risk-factor shocks/returns, under $\boldsymbol{\Phi} = \mathrm{diag}(\phi_1,\dots,\phi_d)$ and zero intercept, can be written as

$$\mathbf{r}_1^{(m)} = \boldsymbol{\varepsilon}_1^{(m)}, \qquad \mathbf{r}_h^{(m)} = \boldsymbol{\Phi}\mathbf{r}_{h-1}^{(m)} + \boldsymbol{\varepsilon}_h^{(m)}, \qquad h = 2,\dots,H. \tag{17}$$

For short holding periods, recursive volatility propagation remains particularly relevant from a risk-management perspective. Financial return series exhibit well-documented stylized facts, most notably pronounced volatility clustering and substantial excess kurtosis, that materially affect short-term tail behaviour. By updating conditional volatility along the simulated path, FHS-Recursive preserves these dynamics within each trajectory, ensuring that shock sequences reflect the persistence and clustering of large moves typically observed in turbulent markets. As a consequence, for $H$ in the market-risk range, the recursive scheme provides a prudential representation of short-horizon risk by maintaining the heteroscedastic structure that dominates the empirical distribution of short-term financial risk factors shocks.

### 3.2.4 FHS-Static (constant volatility rescaling)

This variant follows the same filtering layer but fixes the target volatility state over the simulated pathway. In the static-volatility FHS engine, conditional volatility is fixed at $\boldsymbol{\sigma}_0(\tau)$ throughout the horizon. Define the simulated innovations component-wise as

$$\varepsilon_{h,j}^{(m)} := \sigma_{0,j}\, z_{h,j}^{(m)}, \qquad j = 1,\dots,d,\ \ h = 1,\dots,H, \tag{18}$$

or, in vector form,

$$\boldsymbol{\varepsilon}_h^{(m)} := \mathrm{diag}\big(\boldsymbol{\sigma}_0(\tau)\big)\,\mathbf{z}_h^{(m)} \in \mathbb{R}^d. \tag{19}$$

If a mean component is retained (e.g., factor-wise AR(1)), it can be propagated consistently along the simulated path; again, under $\boldsymbol{\Phi} = \mathrm{diag}(\phi_1,\dots,\phi_d)$ and zero intercept

$$\mathbf{r}_1^{(m)} = \boldsymbol{\varepsilon}_1^{(m)}, \qquad \mathbf{r}_h^{(m)} = \boldsymbol{\Phi}\mathbf{r}_{h-1}^{(m)} + \boldsymbol{\varepsilon}_h^{(m)}, \qquad h = 2,\dots,H, \tag{20}$$

where $\{\mathbf{r}_h^{(m)}\}$ denotes the simulated filtered shocks.

Beyond its operational simplicity, the FHS-Static mechanism exhibits two notable governance relevant properties. First, rescaling standardized residuals with a *flat* volatility state along the entire simulated trajectory introduces a natural efficiency–prudence trade-off in economic capital deployment. When the compute-date volatility $\boldsymbol{\sigma}_0(\tau)$ is high, all simulated daily shocks are drawn from a distribution with a correspondingly large scale, producing prudential long-horizon risk metrics. Conversely, when the prevailing volatility is low, the same mechanism aligns capital requirements with calmer market regimes by avoiding unnecessary amplification of long-horizon losses. Second, because cumulative multi-period returns tend to normalize as the holding period $H$ increases (see Section 4.2.1), the construction of long-horizon trajectories using shocks drawn from a single volatility-regime distribution promotes convergence toward the aggregation effects predicted by the Lindeberg–Feller central limit theorem. In other words, the homogeneous rescaling embedded in FHS-Static reinforces the regularizing force of long-horizon aggregation, yielding cumulative shock distributions whose tail behavior is consistent with the empirical CLT-type regularities observed for long-term financial risk factors shocks.

### 3.3 Proposal: The Long-Term FHS Model

This paper introduces the long-term FHS model, denoted LT-FHS, as our novel modelling contribution within the filtered historical simulation family. LT-FHS builds upon FHS-Static, naturally aligned with short-horizon market-risk applications, by incorporating an innovative dynamic bounding mechanism that constrains admissible cumulative shocks to historically observed levels, complemented by a volatility-conditioned buffer. This design ensures empirical consistency at short horizons while delivering plausibility and stability for long-horizon tails, addressing a critical governance challenge in multi-horizon risk modelling.

Let $L_t^{(j)}$ denote the historical level series for risk factor $j$. Define $L_{\min}^{(j)} = \min_t L_t^{(j)}$, $L_{\max}^{(j)} = \max_t L_t^{(j)}$, and a reference level $L_{\text{cmp}}^{(j)}$ (e.g. a compute-date anchor or a stable central tendency level).

For each bootstrap trial $m = 1, \ldots, M$, consider a simulated static path $\{r_{k,m}^{(j)}\}_{k=1}^{H}$ and compute the raw cumulative shock

$$C_m^{(j)} \;=\; \sum_{k=1}^{H} r_{k,m}^{(j)}. \tag{21}$$

To construct a regime-consistent buffer, define a *realized-volatility* proxy computed by driving the GJR recursion with the *static* innovations $\varepsilon_{k-1,m}^{(j)} = \sigma_0^{(j)} z_{k-1,m}^{(j)}$, where $\sigma_0^{(j)}$ is the conditional volatility level of risk factor $j$ estimated at the compute date of the risk metrics:

$$(\tilde{\sigma}_{1,m}^{(j)})^2 = (\sigma_0^{(j)})^2, \tag{22}$$

$$(\tilde{\sigma}_{k,m}^{(j)})^2 = \omega^{(j)} + \alpha^{(j)}(\varepsilon_{k-1,m}^{(j)})^2 + \beta^{(j)}(\tilde{\sigma}_{k-1,m}^{(j)})^2 + \gamma^{(j)}(\varepsilon_{k-1,m}^{(j)})^2 \mathbb{I}_{\{\varepsilon_{k-1,m}^{(j)}<0\}}, \qquad k \geq 2, \tag{23}$$

$$RV_{H,m}^{(j)} = \left( \sum_{k=1}^{H} (\tilde{\sigma}_{k,m}^{(j)})^2 \right)^{1/2}, \tag{24}$$

$$\overline{RV}_H^{(j)} = \frac{1}{M} \sum_{m=1}^{M} RV_{H,m}^{(j)}. \tag{25}$$

The admissible band is then defined as

$$LB^{(j)} \;=\; \big(L_{\min}^{(j)} - L_{\text{cmp}}^{(j)}\big) - \overline{RV}_H^{(j)}, \qquad UB^{(j)} \;=\; \big(L_{\max}^{(j)} - L_{\text{cmp}}^{(j)}\big) + \overline{RV}_H^{(j)}, \tag{26}$$

so that the final LT-FHS cumulative shock for scenario $m$ is a clipped version of the static shock:

$$S_{H,\text{LT},m}^{(j)} \;=\; \min\Big\{ \max\big\{C_m^{(j)},\, LB^{(j)}\big\},\, UB^{(j)} \Big\}. \tag{27}$$

Beyond preserving all the desirable properties of FHS-Static, LT-FHS introduces two additional governance-enhancing features that are especially relevant for fixed-income risk factors. First, by enforcing admissible bounds through a level-anchored map, LT-FHS is able to capture mean-reverting tendencies that naturally characterize interest rate and credit spread dynamics, without invoking parametric mean-reversion structures (e.g., Ornstein–Uhlenbeck or affine term-structure dynamics). This avoids importing the modelling burden and validation complexity typically associated with fully parametric Monte Carlo engines, while retaining a realistic long-horizon behaviour for fixed-income shock paths. Second, the dynamic buffer based on $\overline{RV}_H^{(j)}$ ensures principled governance at the boundaries of the historical level range: when the compute-date volatility is elevated but the factor level is already near its historical minimum or maximum, the buffer allows LT-FHS to project shocks toward previously unobserved but volatility-consistent regions. This prevents artificial saturation of simulated trajectories at boundary levels and enables scenario generation that remains coherent with the prevailing volatility regime, thereby ensuring prudential yet economically efficient long-horizon tail modelling.

We note, for completeness, that the control of simulated paths introduced in LT-FHS differs in nature and timing from the trajectory control implicitly embedded in the GARCH option-pricing framework with FHS proposed by Barone-Adesi et al. [2008]. In the pricing context, path control is enforced *on-the-run*, at each step of the simulation, through volatility consistent rescaling and martingale-preserving corrections applied recursively along the underlying price process. By contrast, the LT-FHS framework operates an *ex-post* control on cumulative shocks, implemented at the horizon level through admissible-band constraints that are explicitly designed for risk measurement rather than valuation. While an *on-the-run* control is natural and necessary in arbitrage-free pricing environments, the *ex-post* trajectory governance adopted here is more aligned with long-horizon risk modeling objectives, where plausibility, stability and capital interpretability dominate exact martingale propagation. This distinction allows FHS-Static and LT-FHS to coexist within a single, coherent multi-horizon framework, preserving methodological continuity across short- and long-term risk assessments while adapting the trajectory-control mechanism to horizon-specific governance requirements.

**Short-horizon inertness of the bounding map.** Two ingredients jointly explain why bounding is typically negligible at market-risk horizons. First, uncertainty—and therefore the dispersion of cumulative shocks—is increasing in the holding period $H$. Second, the dynamic cap is constructed from full-sample historical level extremes $(L_{\min}^{(j)}, L_{\max}^{(j)})$ and a regime-conditioned buffer $\overline{RV}_H^{(j)}(q)$, which itself is increasing in $H$ and in the volatility state through the rescaling $\sigma_0^{(j)}(q)$. As a result, for $H \leq 21$ the simulated cumulative shocks rarely breach the admissible band, and LT-FHS behaves *operationally* like FHS-Static in the corresponding VaR forecasting exercises.

We document this activation property empirically through the clipping fraction, which measures the percentage of simulated cumulative shocks replaced by the dynamic cap. Table 3 shows that clipping is essentially negligible for $H \in \{5, 10, 21\}$ and becomes material at $H = 252$, with further amplification in stressed volatility regimes. For the empirical analyses conducted in this paper, the reference level $L_{\text{cmp}}^{(j)}$ used in the LT-FHS bounding map is taken to be the sample-average level of each risk factor, which serves as a neutral starting point in place of the compute-date level employed in production settings.

### 3.4 Advantages of the FHS family within Historical Simulation approach

The HS-family engines adopted in this paper are built on a common multivariate bootstrap architecture, yet they differ materially in how tightly scenario sets are anchored to the empirical record. To clarify both implementation and interpretation, the following remarks highlight three

Table 3: LT-FHS activation diagnostics via clipping fraction. Panel A pools across datasets, risk factors, and volatility regimes; Panel B conditions on the highest volatility regime ($q = 100\%$). Entries are in percentage points.

| $H$ | Mean | Median | 95th pct. |
|---|---|---|---|
| **Panel A: pooled** | | | |
| 5 | 0.121 | 0.000 | 0.197 |
| 10 | 0.213 | 0.000 | 0.754 |
| 21 | 0.389 | 0.000 | 2.284 |
| 252 | 2.710 | 1.117 | 12.023 |
| **Panel B: high-volatility regime ($q = 100\%$)** | | | |
| 5 | 1.276 | 0.139 | 5.373 |
| 10 | 2.085 | 0.597 | 7.733 |
| 21 | 3.308 | 2.011 | 10.235 |
| 252 | 12.640 | 14.582 | 20.630 |

methodological aspects that are central to their use in a multi–risk-class setting and to their suitability for multi-horizon tail risk metrics: (i) how cross-sectional dependence is carried across heterogeneous risk factors, (ii) what temporal structure is (and is not) imposed within an $H$-step holding period, and (iii) the distinction between exogenous historical resampling (HS) and endogenous scenario generation enabled by filtering and rescaling (FHS).

**Cross-sectional dependence and multi–risk-class consistency.** All engines considered in this paper resample *joint* vectors (either raw shocks $\mathbf{x}_t$ in HS, or standardized residual vectors $\hat{\mathbf{z}}_t$ in FHS). In either case, a sampled index selects an entire cross-sectional state across all $d$ risk factors, so the scenario generator preserves the empirical dependence structure embedded in historical co-movements across heterogeneous risk classes. This property reduces the practical reliance on explicitly specifying, estimating, and maintaining large correlation matrices across thousands of risk factors, and it provides a consistent dependence carrier when aggregating scenarios across risk classes within a unified risk metric.

**Temporal structure within the holding period.** In HS-iid and HS-$\lambda$, the $H$ steps of a trajectory are generated by independent draws with replacement from the historical set, with either uniform probabilities or exponentially time-decayed probabilities. Therefore, these engines do not impose an explicit time-series dependence across the $H$ steps beyond what is implicitly induced by the empirical shock distribution and the sampling scheme. Extensions that preserve serial dependence (e.g., block bootstrap) are conceptually compatible with the same architecture but are not considered here, as the focus is on baseline HS-family benchmarks and on isolating the impact of weighting and filtering choices.

**Exogenous HS versus endogenous scenario generation in FHS.** HS-iid and HS-$\lambda$ are *exogenous* historical engines in the sense that their scenario sets are obtained by resampling realized historical shock vectors (with uniform or time-decayed probabilities). As a result, their forward-looking content is largely constrained by the support of the empirical record: while bootstrap resampling can generate new *paths* through recombination of past daily shocks, the cross-sectional realizations at each step remain anchored to historically observed joint states. Consequently, these engines are particularly sensitive to (i) the representativeness of the chosen historical window and (ii) non-stationarity and regime shifts that alter the relevance of past observations.

In contrast, FHS-Static and FHS-Recursive introduce an *endogenous* generative layer through volatility filtering and rescaling. By decomposing shocks into conditional scale and standardized residuals, and by resampling (approximately) i.i.d. residual vectors rather than raw returns, FHS decouples scenario generation from the level and clustering properties of historical volatility. The rescaling step maps resampled residuals to a target conditional volatility state, and in the recursive variant the conditional volatility is updated along the simulated path. This mechanism yields scenario sets that are less tightly bound to historical realized shock magnitudes and are therefore able, in a statistically coherent manner, to produce tail-relevant innovations that were not explicitly observed in the historical record (or only observed under very different volatility regimes), while still preserving empirical cross-sectional dependence through joint residual-vector resampling.

Accordingly, the primary limitations of HS-iid and HS-$\lambda$ relate to historical representativeness and regime sensitivity, whereas the governance focus for FHS engines shifts toward the specification, stability, and validation of the filtering layer (mean and volatility dynamics), as well as the monitoring of parameter drift and volatility-state consistency across holding periods.

**FHS scenario generation vs GenAI approach** Beyond (or before) the risk measurement goal, FHS methodology is a very powerful tool to generate risk factor scenarios by collecting data (with the proper data quality and data filling steps), normalizing them and finally being able (by bootstrapping and concatenating) to achieve market data scenarios with useful features: multidimensional, large size, long horizon, reliance on a compact set of parameters (no correlations to manage). Due to the recent trends in the market, also the so called *generative AI* (GenAI for short) techniques received much attention by the financial institutions. The topic is out of scope in this work. We just point out two aspects of these approaches. First, most of the proposals that are labeled as AI are based on quite popular (traditional or innovative) quantitative techniques, such as the *autoencoders* (a sort of non linear principal component analysis), *transformers* (neural network that adopt *self-attention* methodology to infer the dependencies), *gaussian mixture models* (GMM) and many others. See Acciaio et al. [2024] for a detailed recent review. Second, at the current state of the art the application of AI to scenario generation appears to be quite hard due to huge computational effort and data needs. FHS augments the available data base by filtering through econometric models. AI uses richer nonlinear models, in theory superior. GenAI models require, however, massive data bases for learning. Such databases are much bigger than the ones required for most econometric modeling and are not generally easily available in financial markets for the necessary time lengths. The direct application to the risk modeling of AI is therefore problematic.

# 4 Application: Empirical Design and Validation Framework

This section describes the empirical design and validation framework, organized into two complementary blocks that mirror the institutional governance workflow. Block I covers portfolio-level out-of-sample backtesting at typical market-risk horizons ($H \in \{5, 10, 21\}$), where exceedance-based tests are statistically feasible and directly actionable for monitoring. Block II focuses on risk-factor-level tail calibration across multiple horizons ($H \in \{5, 10, 21, 252\}$), designed for settings where long-term exceedance backtesting is structurally underpowered and where plausibility and stability of multi-periodal shocks tails become the binding constraints. While the tail validation framework is applicable to all holding periods, its governance utility becomes critical in scenarios where real-world observations are scarce and validation must rely on simulation-based evidence—such as the annual horizon ($H = 252$) typically required under ICAAP and ECAP prescriptions.

To operationalize these two validation blocks, this paper introduces two governance-oriented scoring tools: the **Composite Risk Governance Score (CRGS)**, a backtesting-driven metric

that synthesizes multiple exceedance-based tests into a single governance signal for short-medium term horizon validation; and the **Tail Risk Governance Score (TRGS)**, designed to assess the quality of tail reconstruction in simulated risk-factor distributions under multi-horizon constraints, with particular relevance for long-horizon capital frameworks.

## 4.1 Block I: Portfolio-level backtesting at market-risk horizons

### 4.1.1 Out-of-sample design with widening (expanding) information sets

Let $\tau_k$ denote the $k$-th compute date for a given holding period $H$, with non-overlapping scheduling $\tau_{k+1} \approx \tau_k + H$ business days. At each compute date $\tau_k$, the risk engine delivers a forecast of the portfolio loss distribution over the next $H$ business days, conditional on the historical information set

$$\mathcal{H}_{\tau_k} := \{\mathbf{x}_t : t \leq \tau_k^-\}, \tag{28}$$

where $\mathbf{x}_t \in \mathbb{R}^d$ is the multivariate risk-factor shock vector defined in Section 3. The widening-window principle means that the estimation set expands over time:

$$\mathcal{H}_{\tau_k} \subset \mathcal{H}_{\tau_{k+1}} \qquad \text{for all } k, \tag{29}$$

i.e., each forecast at $\tau_k$ uses all information available up to $\tau_k$ (and only that information).

Let $\mathrm{PnL}_{\tau_k}^{(H)}$ denote the *realized* $H$-day profit-and-loss in euro units at backtesting date $\tau_k$, so that losses correspond to negative values, as described in Section 5.1. For a confidence level $\mathrm{CL} \in (0,1)$, define $\alpha := 1 - \mathrm{CL}$. The Value-at-Risk forecast at level $\alpha$ is computed from the distribution of *simulated* $H$-day profit-and-loss values $\{\mathrm{PnL}_{\tau_k,\mathrm{sim}}^{(H,m)}\}_{m=1}^{M}$, where $m$ indexes the scenario, as

$$q_\alpha(\tau_k) := \inf\left\{ x \in \mathbb{R} : \frac{1}{M}\sum_{m=1}^{M} \mathbb{I}\left\{\mathrm{PnL}_{\tau_k,\mathrm{sim}}^{(H,m)} \leq x\right\} \geq \alpha \right\}, \tag{30}$$

with $M$ the number of scenarios. Consistently with the convention adopted throughout the paper, the Value-at-Risk is treated as the (positive) loss magnitude,

$$\widehat{\mathrm{VaR}}_{\tau_k}^{(H)}(\alpha) := -q_\alpha(\tau_k) > 0, \qquad q_\alpha(\tau_k) = -\widehat{\mathrm{VaR}}_{\tau_k}^{(H)}(\alpha) < 0. \tag{31}$$

A VaR exceedance (violation) occurs when the *realized* PnL falls below the forecasted quantile threshold $-\widehat{\mathrm{VaR}}$:

$$I_{\tau_k}(\alpha) := \mathbb{I}\Big\{\mathrm{PnL}_{\tau_k}^{(H)} < -\widehat{\mathrm{VaR}}_{\tau_k}^{(H)}(\alpha)\Big\}. \tag{32}$$

Under correct calibration, the sequence $\{I_{\tau_k}(\alpha)\}$ should behave as an (approximately) i.i.d. Bernoulli process with mean $\alpha$.

### 4.1.2 Models under comparison (portfolio-level backtesting)

The backtesting comparison covers four scenario-generation engines that share the same multivariate bootstrap architecture (i.e., joint-vector resampling to preserve cross-sectional dependence across all risk factors), and differ only in the weighting and/or filtering layers used to transform the resampled information into $H$-day risk scenarios. For each engine and each holding period $H \in \{5, 10, 21\}$, risk measures are computed from a set of $M = 100{,}000$ bootstrap pathways generated at a daily time step and aggregated over $H$ business days.

- **HS-iid**: i.i.d. bootstrap historical simulation, where historical joint shock vectors are sampled with uniform probabilities $w_i^{\mathrm{iid}} = 1/n$.

- **HS-$\lambda$**: exponentially weighted historical simulation, where sampling probabilities follow the time-decay scheme in (8) with decay factor fixed at $\lambda = 0.99$. This choice of $\lambda$ reflects a trade-off: it ensures a sufficiently long effective memory to allow HS-$\lambda$ to compete with filtered historical simulation engines calibrated on widening windows, while maintaining a visible decay effect that prioritizes recent observations and preserves responsiveness to regime changes. With $\lambda = 0.99$, the effective half-life of the weighting scheme is approximately 69 days, which balances historical representativeness and recency in the scenario set.
- **FHS-Recursive**: filtered historical simulation with recursive volatility propagation. Daily shocks are decomposed factor-wise via an AR(1) mean specification without intercept and a GJR-GARCH$(1,1)$ volatility recursion. Standardized residual vectors are bootstrapped jointly and then rescaled by a time-varying conditional volatility path propagated along each simulated trajectory.
- **FHS-Static**: filtered historical simulation with static (constant) volatility rescaling. The same AR(1) (no intercept) and GJR-GARCH$(1,1)$ filtering structure is used to obtain standardized residuals, which are bootstrapped jointly; however, rescaling is performed using a fixed conditional volatility state (anchored at the compute-date) rather than a recursively updated volatility path. As discussed in Section 3.3, LT-FHS extension is not considered here, since it is designed for longer holding periods and would not be meaningful when competing over short holding-period horizons.

#### 4.1.3 Backtesting toolkit: UC, CC, and DQ tests

For each tuple (risk class, $H, \alpha$, engine), the validation toolkit comprises UC (unconditional coverage), CC (conditional coverage), and the Dynamic Quantile (DQ) test.

**Unconditional coverage test (UC).** Following Kupiec [1995], let $N$ be the number of out-of-sample observations and $n_1 = \sum_{k=1}^{N} I_{\tau_k}(\alpha)$ the number of exceedances. Define $\hat{\pi} = n_1/N$. The UC likelihood-ratio statistic tests $H_0 : \pi = \alpha$ via

$$\mathrm{LR}_{\mathrm{UC}} = -2\log\left[\frac{(1-\alpha)^{N-n_1}\alpha^{n_1}}{(1-\hat{\pi})^{N-n_1}\hat{\pi}^{n_1}}\right], \qquad \mathrm{LR}_{\mathrm{UC}} \sim \chi^2(1) \text{ under } H_0. \tag{33}$$

**Conditional coverage test (CC).** CC extends UC by also testing independence of exceedances. Following Christoffersen [1998], let $n_{ij}$ denote the number of transitions $I_{\tau_{k-1}} = i \to I_{\tau_k} = j$ for $i, j \in \{0,1\}$. With $\hat{\pi}_0 = n_{01}/(n_{00}+n_{01})$, $\hat{\pi}_1 = n_{11}/(n_{10}+n_{11})$, and $\hat{\pi} = (n_{01}+n_{11})/(n_{00}+n_{01}+n_{10}+n_{11})$, the independence statistic is

$$\mathrm{LR}_{\mathrm{IND}} = -2\log\left[\frac{(1-\hat{\pi})^{n_{00}+n_{10}}\hat{\pi}^{n_{01}+n_{11}}}{(1-\hat{\pi}_0)^{n_{00}}\hat{\pi}_0^{n_{01}}(1-\hat{\pi}_1)^{n_{10}}\hat{\pi}_1^{n_{11}}}\right], \qquad \mathrm{LR}_{\mathrm{IND}} \sim \chi^2(1) \text{ under } H_0. \tag{34}$$

The CC statistic is then $\mathrm{LR}_{\mathrm{CC}} = \mathrm{LR}_{\mathrm{UC}} + \mathrm{LR}_{\mathrm{IND}}$, with $\mathrm{LR}_{\mathrm{CC}} \sim \chi^2(2)$ under $H_0$.

**Dynamic Quantile test (DQ).** As proposed by Engle and Manganelli [2004], we define the de-meaned hit process

$$\mathrm{Hit}_{\tau_k}(\alpha) := I_{\tau_k}(\alpha) - \alpha. \tag{35}$$

Under correct specification, $\mathbb{E}[\mathrm{Hit}_{\tau_k}(\alpha)] = 0$ and the process has no predictable dynamics with respect to $\mathcal{H}_{\tau_k}$. The DQ test operationalizes this requirement via the linear specification

$$\mathrm{Hit}_{\tau_k}(\alpha) = \delta_0 + \sum_{\ell=1}^{L} \delta_\ell \,\mathrm{Hit}_{\tau_{k-\ell}}(\alpha) + \delta_{L+1}\,\widehat{\mathrm{VaR}}^{(H)}_{\tau_{k-1}}(\alpha) + \varepsilon_k, \tag{36}$$

where $L$ is the number of lags (set to $L = 4$ in the following empirical analysis) and $\widehat{\mathrm{VaR}}^{(H)}_{\tau_{k-1}}(\alpha)$ is the one-period-lagged VaR forecast (loss magnitude). Let $\mathbf{h} \in \mathbb{R}^N$ collect $\mathrm{Hit}_{\tau_k}(\alpha)$ and let $\mathbf{X} \in \mathbb{R}^{N\times(L+2)}$ collect the regressors in (36) (constant, $L$ lagged hits, and lagged VaR). The DQ statistic is the Wald form

$$\mathrm{DQ} = \frac{\mathbf{h}^\top \mathbf{X}\big(\mathbf{X}^\top \mathbf{X}\big)^{-1}\mathbf{X}^\top \mathbf{h}}{\alpha(1-\alpha)}, \qquad \mathrm{DQ} \overset{a}{\sim} \chi^2(L+2) \text{ under } H_0 : \delta_0 = \cdots = \delta_{L+1} = 0. \tag{37}$$

The DQ test therefore provides a joint diagnostic of (i) correct unconditional exceedance frequency and (ii) absence of dynamic misspecification.

All tests are evaluated at significance level $\gamma = 5\%$ and are reported as PASS/FAIL decisions.

#### 4.1.4 Composite Risk Governance Score

When a risk organization runs large grids of backtests—across competing engines, multiple holding periods $H$, heterogeneous risk classes, and several confidence levels CL—the output quickly becomes a high-dimensional set of binary decisions and scattered $p$-values. In that setting, governance is rarely served by looking at tests one-by-one: model selection, challenger monitoring, and management escalation require a *stable, comparable, and auditable* ranking signal that can be aggregated and summarized without losing the essential statistical content. A robust Composite Risk Governance Score (CRGS) [1] provides exactly this: it compresses a large backtesting dashboard into a single ordering criterion for each holding–period / risk–class / confidence–level triplet, enabling (i) cross-engine ranking, (ii) consistency checks across horizons and portfolios, (iii) trend monitoring over time, and (iv) transparent communication to the functions responsible for Model Risk Management (MRM) and Internal Validation (IVU). In practice, it also reduces discretion: rather than "picking winners" based on selective slices, the score forces a uniform decision logic across the whole validation grid.

Methodologically, the score is designed to reflect two complementary governance needs. First, it must respect the hard pass/fail nature of regulatory-style backtesting (acceptance regions driven by a test size $\gamma$). Second, it must still discriminate among "passes" and among "fails" by capturing *margin-to-rejection*, because for ongoing monitoring the distance from the threshold is an early-warning signal (a model that barely passes is more fragile under regime change than one that passes comfortably). This motivates a construction that blends an indicator of acceptance with the slack of the $p$-value over the threshold, producing an ordinal measure that is monotone in the validation evidence while remaining easy to audit.

To support ranking over engines and dimensions $(H, \text{risk class}, \text{CL})$, we compute a composite score that blends statistical decisions and margin-to-rejection. Let $p_{\mathrm{UC}}$, $p_{\mathrm{CC}}$, $p_{\mathrm{DQ}}$ be the test $p$-values and define

$$\mathbb{I}_{\mathrm{UC}} = \mathbb{I}\{p_{\mathrm{UC}} > \gamma\}, \quad \mathbb{I}_{\mathrm{CC}} = \mathbb{I}\{p_{\mathrm{CC}} > \gamma\}, \quad \mathbb{I}_{\mathrm{DQ}} = \mathbb{I}\{p_{\mathrm{DQ}} > \gamma\}. \tag{38}$$

We set component scores

$$s_{\mathrm{UC}} = \mathbb{I}_{\mathrm{UC}} + (p_{\mathrm{UC}} - \gamma), \qquad s_{\mathrm{CC}} = \mathbb{I}_{\mathrm{CC}} + (p_{\mathrm{CC}} - \gamma), \qquad s_{\mathrm{DQ}} = \mathbb{I}_{\mathrm{DQ}} + (p_{\mathrm{DQ}} - \gamma), \tag{39}$$

and define the baseline Composite Risk Governance Score as a weighted average of UC and CC,

$$\mathrm{CRGS}(\mathrm{w}_{\mathrm{UC}}, \mathrm{w}_{\mathrm{CC}}) = w_{\mathrm{UC}} s_{\mathrm{UC}} + w_{\mathrm{CC}} s_{\mathrm{CC}}, \qquad (w_{\mathrm{UC}}, w_{\mathrm{CC}}) = (0.6, 0.4). \tag{40}$$

[1] This paper introduces the concept of a composite risk governance score in a heuristic manner, rather than through an axiomatic framework such as those proposed for coherent risk measures by Artzner et al. [1999]. The construction is pragmatic and designed for operational governance rather than theoretical completeness. However, this perspective naturally raises a research question: *what axioms should a composite risk governance score satisfy to be considered internally consistent and regulator-defensible?*

The choice to weight UC more than CC reflects a governance rationale: unconditional coverage is the first-order requirement (correct exceedance frequency), whereas conditional coverage adds a second-order refinement (independence/serial structure of exceedances) that is informative but can be less stable in small samples and at extreme CL. The weights therefore encode a conservative preference for frequency correctness while still rewarding models that also control violations clustering.

We add a DQ bonus only when the exceedance count is sufficiently large to make dynamic testing informative:

$$\text{CRGS} \leftarrow \text{CRGS} + b_{\text{DQ}} s_{\text{DQ}} \text{ if } n_1 \geq 20, \qquad b_{\text{DQ}} = 0.1. \tag{41}$$

This rule is particularly relevant at extreme confidence levels, where expected exceedances can be structurally small.

The portfolio-level backtests above provide direct evidence on short-horizon calibration and are the natural governance tool for market-risk monitoring. We now shift to the long-horizon setting, where the validation problem changes: at $H = 252$ exceedance-based backtests become intrinsically low-power, and tail realism must be assessed through benchmark-based long-horizon tail diagnostics.

## 4.2 Block II: Long-horizon validation of filtered engines

### 4.2.1 Long-horizon shocks aggregation and weak CLT regularities

Let $\{x_t^{(j)}\}_{t=1}^{T}$ denote the daily shock series (in bps) for risk factor $j$ (*e.g.*, a single curve pillar/maturity) within a given risk class. For a generic holding period $H$—expressed in trading days—define the cumulative long-horizon shock for a single generic simulation (one bootstrap trial) drawn from the scenario-generation models analyzed in this paper:

$$S_H^{(j)} \;=\; \sum_{k=1}^{H} r_k^{(j)}, \tag{42}$$

where $\{r_k^{(j)}\}$ are daily shocks generated by a given engine based on the information contained in the empirical series $\{x_t^{(j)}\}$. This definition provides the building block for tail-validation metrics, as it allows comparing simulated cumulative shocks against empirical benchmarks obtained via dependence-preserving resampling. The cumulative shock $S_H^{(j)}$ is the fundamental input for the Tail Risk Governance Score (TRGS), which evaluates the accuracy of tail reconstruction across simulated distributions under multi-horizon constraints.

A stylized fact that is critical for ICAAP/ECAP/Solvency governance is that, as $H$ increases, long-horizon returns/shocks tend to *normalize* in the sense of weak Central Limit Theorem (CLT)-type aggregation effects[2], higher moments, and in particular excess kurtosis, decline

[2]The Lindeberg–Feller Central Limit Theorem is a weaker version of the classical CLT because it relaxes the i.i.d. assumption and applies to independent (not necessarily identically distributed) variables with finite variance. Formally, if $\{X_i\}$ are independent with $\mathbb{E}[X_i] = 0$ and $\text{Var}(X_i) = \sigma_i^2$, and if the Lindeberg condition

$$\forall \varepsilon > 0, \quad \frac{1}{v_n^2} \sum_{i=1}^{n} \mathbb{E}\big[X_i^2 \mathbf{1}_{\{|X_i| > \varepsilon v_n\}}\big] \to 0, \quad v_n^2 = \sum_{i=1}^{n} \sigma_i^2,$$

holds, then $S_n / v_n \xrightarrow{d} \mathcal{N}(0,1)$ as $n \to \infty$. This result is practically relevant for financial risk-factor shocks, which are not i.i.d. and exhibit conditional heteroskedasticity, but typically satisfy weak dependence and finite variance. Consequently, as the holding period $H$ increases, the normalized sum of shocks

$$Z_H^{(j)} = \frac{\sum_{k=1}^{H} r_k^{(j)}}{\sqrt{\sum_{k=1}^{H} \mathbb{V}ar(r_k^{(j)})}}$$

materially with $H$ even when daily shocks are non-Gaussian and conditionally heteroskedastic. This effect is not a mere academic curiosity: at large horizons it becomes a hard constraint for model plausibility, because engines that produce *increasing* leptokurtosis at large $H$ tend to over-amplify tail quantiles and distort capital-relevant metrics.

Table 4 reports the empirical long-horizon behaviour obtained through dependence-preserving block bootstrap (Section 4.2.2). Consistent with weak CLT regularities, mean kurtosis collapses when moving from weekly horizons to the annual horizon, in both risk classes; simultaneously, dispersion and extreme quantiles grow in magnitude with $H$.

Table 4: Empirical long-horizon behaviour under dependence-preserving block-bootstrap aggregation. Statistics are averaged across risk factors within each risk class (interest rate (IR) and credit spread (CS)).

| **Risk class** | $H$ **(business days)** | **Mean Std Dev (bps)** | **Mean Excess Kurtosis** | **Mean** $q_{0.1\%}$ **(bps)** | **Mean** $q_{99.9\%}$ **(bps)** |
|---|---|---|---|---|---|
| IR | 5 | 9.6 | 4.5 | -45.8 | 46.5 |
| IR | 10 | 13.7 | 4.2 | -62.0 | 64.1 |
| IR | 21 | 20.4 | 3.7 | -84.5 | 89.7 |
| IR | 252 | 74.7 | 0.8 | -262.7 | 266.4 |
| CS | 5 | 14.8 | 9.7 | -86.8 | 91.7 |
| CS | 10 | 19.6 | 8.8 | -94.8 | 122.4 |
| CS | 21 | 27.3 | 6.6 | -134.0 | 148.9 |
| CS | 252 | 93.8 | 1.0 | -323.2 | 359.9 |

### 4.2.2 Empirical benchmark via circular moving block bootstrap

Long-horizon validation requires a benchmark distribution for $S_H^{(j)}$ that (i) preserves serial dependence in daily shocks and (ii) avoids imposing strong parametric assumptions at long horizons. Following Politis and Romano [1992], we therefore construct an empirical benchmark via a circular moving block bootstrap (C-MBB), applied factor-wise to the observed daily shocks.

Fix a block length $\ell$ and let $m = \lceil H/\ell \rceil$. For each replicate, draw block start indices $U_1, \ldots, U_m$ i.i.d. uniform on $\{1, \ldots, T\}$. Each block is formed with circular wrapping:

$$B(U_r) \;=\; \big(x^{(j)}_{(U_r) \bmod T},\; x^{(j)}_{(U_r+1) \bmod T},\; \ldots,\; x^{(j)}_{(U_r+\ell-1) \bmod T}\big).$$

Concatenate blocks and retain the first $H$ elements to obtain a bootstrap pseudo-path

$$\{x_1^{*(j)}, \ldots, x_H^{*(j)}\},$$

and define the empirical cumulative shock

$$S^{(j)}_{H,\mathrm{emp}} \;=\; \sum_{k=1}^{H} x_k^{*(j)}. \tag{43}$$

Repeating this procedure $N_{\mathrm{obs}}$ times yields an empirical sample for each $(j, H)$ and thus an empirical quantile function $Q^{(j,H)}_{\mathrm{emp}}(p)$. In our analysis, the number of replications is set to $N_{\mathrm{obs}} = 100{,}000$ to ensure stable estimation of extreme quantiles and reduce simulation error

tends to approximate a normal distribution, explaining why higher-order moments (e.g., excess kurtosis) decline with $H$ under realistic conditions. Engines that fail to respect this normalization property—by producing increasing kurtosis at large $H$—violate a fundamental plausibility constraint for long-horizon risk metrics.

in tail calibration. The block length for the circular moving block bootstrap is set to $\ell = 50$, which represents an optimal trade-off between block independence, intra-block clustering, and the ability to generate sufficient variability in the simulated trajectories, especially for long holding periods.

### 4.2.3 Why the long-term analysis focuses on the filtered family

The restriction of the long-term analysis to the filtered family is further motivated by the structural deficiencies of historical simulation methods documented, among others, by Pritsker [2001]. Pritsker shows that both standard historical simulation (HS-iid) and its exponentially weighted extension (HS-$\lambda$) are intrinsically slow in recognizing changes in conditional risk, as VaR updates are mechanically tied to realized tail losses rather than to shifts in the latent volatility state. Consequently, increases in true risk remain undetected with high probability whenever they are not immediately accompanied by extreme negative realizations, even under standard GARCH data-generating processes.

Simulation evidence in Pritsker [2001] further shows that HS-iid and HS-$\lambda$ generate persistent and highly autocorrelated VaR errors, with economically material episodes of underestimation following volatility regime changes.

Against this backdrop, Pritsker [2001] identifies Filtered Historical Simulation as a conceptually superior alternative within the historical-simulation envelope, as it decouples tail estimation from the realization of extreme portfolio losses by conditioning on the evolving volatility state. Thus, the filtered framework uniquely addresses the responsiveness and asymmetry defects of HS-iid and HS-$\lambda$, providing a structurally more reliable basis for long-term risk measurement and governance.

The backtesting results reported in Section 5.1 corroborate this diagnosis. Across market-risk horizons, HS-iid and HS-$\lambda$ display delayed tail adjustment and violation clustering, while filtered specifications achieve materially faster convergence to target coverage and significantly weaker error persistence. These findings confirm that the limitations identified by Pritsker [2001] remain relevant in multi-horizon, multi-risk-class settings.

Accordingly, and in line with a governance objective that simultaneously targets *prudential coverage* and *efficiency in capital deployment*, the long-term analysis is restricted to the filtered engines:

- **FHS-Recursive** (standard FHS-style recursion);
- **FHS-Static** (flat volatility rescaling at the compute-date state);
- **LT-FHS**: a long-horizon extension of **FHS-Static** with dynamic level-consistent bounding, as presented in Section 3.3.

To probe regime dependence, the initial volatility state is anchored via a quantile of the fitted conditional volatilities:

$$\sigma_0^{(j)}(q) \;=\; \mathrm{Quantile}\big(\{\hat{\sigma}_t^{(j)}\}_{t=1}^{T},\, q\big), \qquad q \in \mathcal{Q}. \tag{44}$$

### 4.2.4 Tail Risk Governance Score

The Tail Risk Governance Score (TRGS) is introduced in this paper as a governance-oriented metric for evaluating the ability of scenario-generation engines to reproduce the empirical behavior of risk-factor tails across multiple holding periods. Building on the governance rationale discussed in previous sections, the TRGS provides a quantitative, model-agnostic measure of tail plausibility and stability, which is particularly relevant when direct exceedance-based backtesting is statistically underpowered or infeasible due to data scarcity. The TRGS framework is applicable to all holding periods, but its governance utility becomes critical in long-horizon

settings—such as those required by ICAAP and ECAP—where the validation of simulated distributions must rely on the quality of tail reconstruction rather than on direct out-of-sample exceedance counts.

Concretely, the TRGS is computed by comparing the quantile functions of simulated cumulative shocks to those obtained from an empirical benchmark, using a tail-focused Root Mean Squared Error (RMSE) metric that aggregates discrepancies in both the left and right tails according to governance-driven weights. This approach enables a transparent and auditable ranking of scenario engines with respect to their ability to deliver plausible and stable tail risk metrics under multi-horizon constraints.[3]

Long-horizon governance is thus driven by tail realism. Accordingly, we validate engines by comparing their tail quantile functions to the empirical benchmark.

Fix tail mass $\alpha \in (0, 1/2)$ and an extreme-probability floor $p_{\min}$. Define uniform grids in each tail:

$$\mathcal{P}_L \subset [p_{\min}, \alpha], \qquad \mathcal{P}_R \subset [1-\alpha, 1-p_{\min}]. \tag{45}$$

For each $(j, H, q)$, let $Q_{\text{emp}}^{(j,H)}(p)$ denote the empirical quantile function from (43) and $Q_{\text{mod}}^{(j,H,q)}(p)$ the engine-implied quantile function.

Define tail mean-squared errors (units: $\text{bps}^2$):

$$\text{MSE}_L^{(j,H,q)} = \frac{1}{|\mathcal{P}_L|} \sum_{p \in \mathcal{P}_L} \left( Q_{\text{mod}}^{(j,H,q)}(p) - Q_{\text{emp}}^{(j,H)}(p) \right)^2, \tag{46}$$

$$\text{MSE}_R^{(j,H,q)} = \frac{1}{|\mathcal{P}_R|} \sum_{p \in \mathcal{P}_R} \left( Q_{\text{mod}}^{(j,H,q)}(p) - Q_{\text{emp}}^{(j,H)}(p) \right)^2. \tag{47}$$

We report both tails RMSE in *bps*:

$$\text{RMSE}_L^{(j,H,q)} = \sqrt{\text{MSE}_L^{(j,H,q)}}, \qquad \text{RMSE}_R^{(j,H,q)} = \sqrt{\text{MSE}_R^{(j,H,q)}}. \tag{48}$$

To obtain the engine-level Tail Risk Governance Score (TRGS), we blend left and right tails with weights $w_L, w_R$:

$$\text{TRGS}^{(j,H,q)}(w_L, w_R) = \sqrt{w_L \, \text{MSE}_L^{(j,H,q)} + w_R \, \text{MSE}_R^{(j,H,q)}}, \qquad w_R \in [0, 1]. \tag{49}$$

All results presented in this paper set $w_L = 0.3$ and $w_R = 0.7$, and report averages over risk factors and volatility regimes $q$. This asymmetric choice reflects two governance considerations: first, financial institutions typically hold long positions in bonds within non-trading portfolios (e.g., banking book), which are exposed to losses when interest rates or credit spreads rise; second, empirical evidence shows that most fixed-income risk factors exhibit positive skewness, as market turmoil is generally associated with upward moves in these parameters.

# 5 Results

## 5.1 Portfolio backtesting results at market-risk horizons

The out-of-sample backtesting exercise is performed on a set of hypothetical portfolios designed to represent typical exposures across three principal risk classes: equity (EQ), interest rate (IR), and credit spread (CS). For interest rate and credit spread risks, portfolio exposures are

[3] Technical settings for the computation of the Tail Risk Governance Score (TRGS) are as follows: the score is computed as the Root Mean Squared Error (RMSE) between model and empirical quantiles, evaluated over a uniform grid of 2,000 probability points in each tail, focusing on the lower 5% and upper 5% of the distribution (with a minimum probability threshold $p_{\min} = 10^{-5}$).

constructed via sensitivities to shifts in the relevant risk factors, expressed in euro per basis point (€/bps). The interest rate portfolio consists of standardized positions on the ESTR Overnight Indexed Swap (OIS) curve, with maturities spanning from 0.25 to 30 years and uniform sensitivities of −1000 €/bps. The credit spread risk portfolio is constructed from standardized exposures to the Italian Asset Swap Spread (ASW) curve [4], extracted from the Bloomberg BVAL BVIS0575 ASW curve, with maturities ranging from 1 to 30 years and the same sensitivity profile of −1000 €/bps. Equity risk exposures are specified as notional positions of €1,000,000 per instrument, diversified across broad indices and sector baskets. Specifically, the equity portfolio includes FTSE MIB, EURO STOXX 50, S&P 500, and a comprehensive set of MSCI European and MSCI Italian sector indices spanning consumer discretionary, consumer staples, energy, financials, industrials, telecommunication services, and utilities.

Because the primary focus of this paper is on the distributional assumptions embedded in historical and filtered historical simulation methods, the analysis abstracts from other potential sources of error in Value-at-Risk estimation. Specifically, VaR is evaluated for standardized linear portfolios in underlying risk factors, for which the mapping between factor shocks and portfolio profit-and-loss is exact. As a result, the only source of estimation error considered in this study arises from the statistical assumptions and dynamic features of the risk-generation models themselves.

The OOS backtesting exercise was conducted over a comprehensive historical window spanning from March 5, 2010, to April 24, 2025, thus covering more than 15 years of market data and encompassing multiple market regimes. Because the compute dates are scheduled at non-overlapping intervals of $H$ business days, the effective sample size declines sharply as $H$ increases. Table 5 reports the effective number of out-of-sample observations $N$ for each horizon. These sample sizes motivate the restriction of classical exceedance backtesting to market-risk horizons ($H \in \{5, 10, 21\}$) and the use of benchmark-based tail validation at the annual horizon.

Table 5: Backtesting horizons and effective sample sizes (non-overlapping observations).

| $H$ (business days) | Interpretation | No. OOS observations ($N$) |
|---|---|---|
| 5 | 1 week (managerial horizon) | 769 |
| 10 | 2 weeks (bank trading-book) | 384 |
| 21 | 1 month (asset managers) | 183 |

Throughout the backtesting window, portfolio compositions are held constant. Consequently, differences in VaR performance across engines and horizons reflect solely the evolution of market shocks and the dynamics induced by the scenario-generation methodologies, rather than changes in exposures. This design choice supports a transparent, controlled, and risk-class-consistent comparison of model calibration and stability across market-risk horizons.

Let $RC$ denote a generic risk class (e.g., IR, CS, EQ), and let $j \in RC$ index the risk factors within that class. At each backtesting date $\tau_k$, the realized $H$-day profit-and-loss for the portfolio associated with risk class $RC$ is computed as

$$\mathrm{PnL}_{\tau_k}^{(H,RC)} = \mathbf{Sens}_{RC}^{\top} \mathbf{S}_{H,\tau_k}^{(RC)}, \tag{50}$$

where $\mathbf{Sens}_{RC} \in \mathbb{R}^{d_{RC}}$ is the vector of sensitivities for the $d_{RC}$ risk factors in class $RC$, and $\mathbf{S}_{H,\tau_k}^{(RC)}$ collects the realized, non-overlapping, $H$-day cumulative shocks (or multi-period returns) for each risk factor in $RC$.

[4] Asset swap spreads are used as a standard market proxy for sovereign credit spread risk, as they isolate the credit component from risk-free rate movements.

**IR and CS risk classes.** For interest-rate and credit-spread risk factors, daily shocks are additive. Thus the $H$-day vector of cumulative shocks is

$$\mathbf{S}_{H,\tau_k}^{(RC)} = \left( \sum_{h=1}^{H} x_{t_k+h}^{(j_1)},\ \ldots,\ \sum_{h=1}^{H} x_{t_k+h}^{(j_{d_{RC}})} \right)^{\top}, \qquad RC \in \{\mathrm{IR}, \mathrm{CS}\}, \tag{51}$$

with $x_t^{(j)}$ denoting the observed daily shock for risk factor $j$ on trading day $t$.

**EQ risk class.** For equity risk factors, multi-period aggregation follows the compounding convention. Daily arithmetic returns are first mapped into gross returns and then compounded across the $H$-day window. The resulting $H$-day multi-period return vector is

$$\mathbf{S}_{H,\tau_k}^{(\mathrm{EQ})} = \left( \prod_{h=1}^{H} (1 + x_{t_k+h}^{(j_1)}) - 1,\ \ldots,\ \prod_{h=1}^{H} (1 + x_{t_k+h}^{(j_{d_{\mathrm{EQ}}})}) - 1 \right)^{\top}. \tag{52}$$

In all cases, the aggregation window consists of the $H$ consecutive trading days immediately following the compute date $\tau_k$, with no overlap between holding periods.

For each risk class and backtesting date, the Value-at-Risk and related backtesting excercises are computed by generating, for each engine, a set of simulated $H$-day cumulative shocks or multi-period returns $\{\mathbf{S}_{H,\tau_k,\mathrm{sim}}^{(RC,m)}\}_{m=1}^{M}$ and evaluating the corresponding scenario PnL as

$$\mathrm{PnL}_{\tau_k,\mathrm{sim}}^{(H,RC,m)} = \mathbf{Sens}_{RC}^{\top} \mathbf{S}_{H,\tau_k,\mathrm{sim}}^{(RC,m)}, \tag{53}$$

where $m = 1, \ldots, M$ indexes the simulated scenarios. The empirical quantile of the simulated PnL distribution is then used to compute the Value-at-Risk and to evaluate exceedances as described in Section 4.1.

This procedure is applied to all risk classes analyzed in this study; portfolio-level results are then aggregated and compared across engines and horizons using the Composite Risk Governance Score as described in Section 4.1.4.

### 5.1.1 Backtesting results: aggregate dashboards across risk classes

Tables 6–8 provide an aggregate "dashboard" view of the four competing engines, organized by holding period $H$ and confidence level (CL). For each engine and CL, we report: (i) the Composite Risk Governance Score averaged across risk classes (**Avg. CRGS**), (ii) the same score decomposed by risk class (Equity–**EQ**, Interest Rate–**IR**, Credit Spread–**CS**), and (iii) UC and CC pass rates computed across the three risk classes (**UC pass**, **CC pass**). DQ inclusion and DQ pass rates are computed analogously across risk classes; **DQ incl.** measures the share of risk classes for which the DQ component is included in the score according to the exceedance-count rule.

Table 6: Aggregate backtesting dashboard at $H = 5$ (weekly horizon). CRGS are reported as Avg. Score and by risk class; pass rates are computed across the three risk classes.

| **Engine** | **CL** | **Avg. CRGS** | **EQ CRGS** | **IR CRGS** | **CS CRGS** | **UC pass** | **CC pass** | **DQ incl.** | **DQ pass** |
|---|---|---|---|---|---|---|---|---|---|
| HS-iid | 99.9% | -0.035 | -0.044 | -0.031 | -0.031 | 0% | 0% | 0% | 67% |
| HS-$\lambda$ | 99.9% | -0.049 | -0.050 | -0.050 | -0.049 | 0% | 0% | 0% | 67% |
| FHS-Static | 99.9% | -0.049 | -0.050 | -0.047 | -0.050 | 0% | 0% | 0% | 100% |
| FHS-Recursive | 99.9% | 0.770 | 1.297 | 1.044 | -0.031 | 67% | 67% | 0% | 100% |
| HS-iid | 99.0% | 0.882 | 1.412 | -0.045 | 1.280 | 67% | 67% | 0% | 0% |
| HS-$\lambda$ | 99.0% | 0.858 | -0.055 | 1.067 | 1.562 | 67% | 67% | 33% | 67% |
| FHS-Static | 99.0% | 0.964 | 1.396 | -0.063 | 1.562 | 67% | 67% | 0% | 100% |
| FHS-Recursive | 99.0% | 1.366 | 1.158 | 1.659 | 1.280 | 100% | 100% | 0% | 100% |
| HS-iid | 97.5% | 0.894 | -0.036 | 1.647 | 1.071 | 100% | 33% | 33% | 0% |
| HS-$\lambda$ | 97.5% | 1.038 | 1.351 | 1.756 | 0.006 | 67% | 67% | 100% | 67% |
| FHS-Static | 97.5% | 1.472 | 1.756 | 1.694 | 0.966 | 100% | 100% | 67% | 100% |
| FHS-Recursive | 97.5% | 1.558 | 1.756 | 1.760 | 1.156 | 100% | 100% | 67% | 100% |
| HS-iid | 95.0% | 0.303 | -0.027 | -0.043 | 0.979 | 33% | 0% | 100% | 33% |
| HS-$\lambda$ | 95.0% | 1.326 | 0.022 | 1.340 | 1.329 | 100% | 100% | 100% | 100% |
| FHS-Static | 95.0% | 1.589 | 1.990 | 1.542 | 1.235 | 100% | 100% | 100% | 100% |
| FHS-Recursive | 95.0% | 1.302 | 1.371 | 1.299 | 1.235 | 100% | 100% | 100% | 100% |

Table 7: Aggregate backtesting dashboard at $H = 10$ (two-week trading-book horizon). CRGS are reported as Avg. Score and by risk class; pass rates are computed across the three risk classes.

| **Engine** | **CL** | **Avg. CRGS** | **EQ CRGS** | **IR CRGS** | **CS CRGS** | **UC pass** | **CC pass** | **DQ incl.** | **DQ pass** |
|---|---|---|---|---|---|---|---|---|---|
| HS-iid | 99.9% | 0.465 | 1.479 | -0.048 | -0.034 | 33% | 33% | 0% | 0% |
| HS-$\lambda$ | 99.9% | 0.323 | 0.030 | 0.969 | -0.031 | 33% | 33% | 0% | 67% |
| FHS-Static | 99.9% | 0.328 | 0.006 | 1.009 | -0.031 | 33% | 33% | 0% | 67% |
| FHS-Recursive | 99.9% | 1.183 | 1.479 | 1.479 | 0.590 | 100% | 67% | 0% | 67% |
| HS-iid | 99.0% | 0.750 | 0.022 | 1.479 | 0.749 | 67% | 33% | 0% | 67% |
| HS-$\lambda$ | 99.0% | 1.005 | 1.014 | 1.105 | 0.896 | 67% | 67% | 0% | 100% |
| FHS-Static | 99.0% | 1.526 | 1.371 | 1.610 | 1.598 | 100% | 100% | 0% | 100% |
| FHS-Recursive | 99.0% | 1.505 | 1.545 | 1.457 | 1.513 | 100% | 100% | 0% | 100% |
| HS-iid | 97.5% | 0.596 | -0.043 | 1.412 | 0.418 | 67% | 33% | 0% | 67% |
| HS-$\lambda$ | 97.5% | 0.969 | 1.371 | 1.291 | 0.247 | 67% | 67% | 0% | 100% |
| FHS-Static | 97.5% | 1.275 | 1.371 | 1.634 | 0.820 | 100% | 100% | 0% | 100% |
| FHS-Recursive | 97.5% | 1.278 | 1.625 | 1.791 | 0.418 | 67% | 100% | 0% | 100% |
| HS-iid | 95.0% | 0.262 | -0.048 | -0.034 | 0.868 | 33% | 0% | 33% | 67% |
| HS-$\lambda$ | 95.0% | 1.585 | 1.559 | 1.921 | 1.275 | 100% | 100% | 67% | 100% |
| FHS-Static | 95.0% | 1.314 | 1.926 | -0.049 | 2.066 | 67% | 100% | 33% | 100% |
| FHS-Recursive | 95.0% | 0.965 | 1.489 | 1.425 | -0.020 | 67% | 67% | 33% | 100% |

Table 8: Aggregate backtesting dashboard at $H = 21$ (one-month asset-manager horizon). CRGS are reported as Avg. Score and by risk class; pass rates are computed across the three risk classes.

| **Engine** | **CL** | **Avg. CRGS** | **EQ CRGS** | **IR CRGS** | **CS CRGS** | **UC pass** | **CC pass** | **DQ incl.** | **DQ pass** |
|---|---|---|---|---|---|---|---|---|---|
| HS-iid | 99.9% | 0.385 | -0.043 | 1.479 | -0.280 | 33% | 33% | 0% | 67% |
| HS-$\lambda$ | 99.9% | -0.021 | -0.025 | -0.034 | -0.005 | 0% | 0% | 0% | 100% |
| FHS-Static | 99.9% | 0.929 | 1.225 | 1.610 | -0.047 | 67% | 67% | 0% | 100% |
| FHS-Recursive | 99.9% | 1.353 | 1.225 | 1.610 | 1.225 | 100% | 100% | 0% | 100% |
| HS-iid | 99.0% | 1.003 | 1.567 | -0.063 | 1.506 | 67% | 67% | 0% | 67% |
| HS-$\lambda$ | 99.0% | 0.875 | -0.049 | 1.479 | 1.194 | 67% | 67% | 0% | 67% |
| FHS-Static | 99.0% | 1.216 | 0.006 | 1.479 | 2.164 | 100% | 67% | 0% | 100% |
| FHS-Recursive | 99.0% | 1.511 | 1.567 | 1.483 | 1.483 | 100% | 100% | 0% | 100% |
| HS-iid | 97.5% | 1.342 | 1.371 | 1.352 | 1.303 | 100% | 67% | 0% | 67% |
| HS-$\lambda$ | 97.5% | 1.233 | 0.022 | 1.425 | 2.252 | 100% | 67% | 0% | 100% |
| FHS-Static | 97.5% | 1.319 | 0.006 | 1.256 | 2.696 | 100% | 67% | 0% | 100% |
| FHS-Recursive | 97.5% | 1.194 | 0.424 | 1.389 | 1.770 | 67% | 100% | 0% | 100% |
| HS-iid | 95.0% | 0.955 | -0.049 | 1.479 | 1.435 | 67% | 67% | 0% | 67% |
| HS-$\lambda$ | 95.0% | 1.278 | 1.371 | 1.421 | 1.042 | 100% | 100% | 0% | 100% |
| FHS-Static | 95.0% | 1.640 | 1.776 | 1.513 | 1.632 | 100% | 100% | 0% | 100% |
| FHS-Recursive | 95.0% | 1.313 | 1.274 | 1.391 | 1.274 | 100% | 100% | 0% | 100% |

*Takeaways (Tables 6–8).* Across horizons, the filtered engines dominate the extreme tail: at CL=99.9% the highest Avg. Score and the strongest UC/CC acceptance are consistently delivered by FHS-Recursive, with residual weakness concentrated in Credit at $H = 5$. Moving to CL=99% and CL=97.5%, calibration improves across all engines and leadership becomes more diversified; nevertheless, filtered engines remain systematically robust across Equity and IR and maintain high UC/CC pass rates at the longer horizons. In the central-tail region (CL=95%), time-decay weighting materially improves HS performance and can become competitive in specific segments, yet the static filtered rescaling frequently achieves the most stable cross–risk-class profile, particularly at $H = 21$.

### 5.1.2 Risk-class lens: winner maps by horizon and confidence level

While Tables 6–8 summarize performance *on average* across risk classes, Tables 9–12 adopt a "winner map" perspective. For each holding period $H$ and risk class, we report the engine with the highest CRGS at a fixed confidence level. Alongside the score, we report the observed number of VaR exceedances (**Viol**), the expected number under correct calibration (**Exp**$= N\alpha$), and UC/CC PASS/FAIL outcomes. When all engines fail UC/CC in a given cell, the reported winner should be interpreted as the *least adverse* performer (highest score).

Table 9: Winner map at CL=99.9% ($\alpha = 0.001$): best engine (highest CRGS) by risk class and holding period.

| $H$ | **Risk class** | **Best engine** | **Score** | **Viol** | **Exp** | **UC** | **CC** |
|---|---|---|---|---|---|---|---|
| 5 | EQ | FHS-Recursive | 1.297 | 2 | 0.77 | PASS | PASS |
| 5 | IR | FHS-Recursive | 1.044 | 3 | 0.77 | PASS | PASS |
| 5 | CS | HS-iid | -0.031 | 4 | 0.77 | FAIL | FAIL |
| 10 | EQ | HS-iid | 1.479 | 1 | 0.38 | PASS | PASS |
| 10 | IR | FHS-Recursive | 1.479 | 1 | 0.38 | PASS | PASS |
| 10 | CS | FHS-Recursive | 1.479 | 1 | 0.38 | PASS | PASS |
| 21 | EQ | FHS-Static | 1.225 | 1 | 0.18 | PASS | PASS |
| 21 | IR | FHS-Static | 1.610 | 0 | 0.18 | PASS | PASS |
| 21 | CS | FHS-Recursive | 1.225 | 1 | 0.18 | PASS | PASS |

Table 10: Winner map at CL=99.0% ($\alpha = 0.01$): best engine (highest CRGS) by risk class and holding period.

| $H$ | **Risk class** | **Best engine** | **Score** | **Viol** | **Exp** | **UC** | **CC** |
|---|---|---|---|---|---|---|---|
| 5 | EQ | HS-iid | 1.412 | 9 | 7.69 | PASS | PASS |
| 5 | IR | FHS-Static | 1.659 | 9 | 7.69 | PASS | PASS |
| 5 | CS | HS-$\lambda$ | 1.562 | 8 | 7.69 | PASS | PASS |
| 10 | EQ | FHS-Recursive | 1.545 | 4 | 3.84 | PASS | PASS |
| 10 | IR | FHS-Static | 1.610 | 5 | 3.84 | PASS | PASS |
| 10 | CS | HS-$\lambda$ | 1.893 | 4 | 3.84 | PASS | PASS |
| 21 | EQ | FHS-Static | 1.567 | 1 | 1.83 | PASS | PASS |
| 21 | IR | FHS-Recursive | 1.483 | 3 | 1.83 | PASS | PASS |
| 21 | CS | HS-$\lambda$ | 1.483 | 3 | 1.83 | PASS | PASS |

Table 11: Winner map at CL=97.5% ($\alpha = 0.025$): best engine (highest CRGS) by risk class and holding period.

| $H$ | **Risk class** | **Best engine** | **Score** | **Viol** | **Exp** | **UC** | **CC** |
|---|---|---|---|---|---|---|---|
| 5 | EQ | FHS-Recursive | 1.756 | 22 | 19.23 | PASS | PASS |
| 5 | IR | FHS-Recursive | 1.760 | 22 | 19.23 | PASS | PASS |
| 5 | CS | HS-$\lambda$ | 1.756 | 20 | 19.23 | PASS | PASS |
| 10 | EQ | FHS-Recursive | 1.625 | 9 | 9.60 | PASS | PASS |
| 10 | IR | FHS-Recursive | 1.791 | 10 | 9.60 | PASS | PASS |
| 10 | CS | HS-$\lambda$ | 1.247 | 7 | 9.60 | PASS | PASS |
| 21 | EQ | FHS-Static | 1.770 | 4 | 4.58 | PASS | PASS |
| 21 | IR | FHS-Recursive | 1.389 | 6 | 4.58 | PASS | PASS |
| 21 | CS | HS-$\lambda$ | 1.806 | 5 | 4.58 | PASS | PASS |

Table 12: Winner map at CL=95.0% ($\alpha = 0.05$): best engine (highest CRGS) by risk class and holding period.

| $H$ | **Risk class** | **Best engine** | **Score** | **Viol** | **Exp** | **UC** | **CC** |
|---|---|---|---|---|---|---|---|
| 5 | EQ | FHS-Static | 1.990 | 38 | 38.45 | PASS | PASS |
| 5 | IR | FHS-Static | 1.542 | 45 | 38.45 | PASS | PASS |
| 5 | CS | HS-$\lambda$ | 1.329 | 31 | 38.45 | PASS | PASS |
| 10 | EQ | FHS-Static | 1.926 | 19 | 19.20 | PASS | PASS |
| 10 | IR | HS-$\lambda$ | 1.921 | 20 | 19.20 | PASS | PASS |
| 10 | CS | HS-$\lambda$ | 1.216 | 14 | 19.20 | PASS | PASS |
| 21 | EQ | FHS-Static | 1.776 | 9 | 9.15 | PASS | PASS |
| 21 | IR | FHS-Static | 1.513 | 10 | 9.15 | PASS | PASS |
| 21 | CS | FHS-Static | 1.632 | 8 | 9.15 | PASS | PASS |

*Takeaways (Tables 9–12).* At CL=99.9%, the filtered recursive engine dominates almost all risk-class/horizon cells, confirming superior extreme-tail calibration when model dynamics include both residual resampling and an endogenous time-varying volatility path; the only exception is Credit at $H = 5$, where all engines fail UC/CC and the reported winner is therefore the least adverse performer rather than an acceptable model. At CL=99%, all winners satisfy UC/CC and leadership becomes more diversified: time-decay weighting is structurally competitive in Credit, while filtered engines retain strong performance in IR and at longer horizons. At CL=97.5% and CL=95%, the ranking further reflects the interaction between horizon and tail depth: FHS-Recursive remains dominant in Equity and IR at intermediate tails, whereas HS-$\lambda$ often leads in Credit at shorter horizons; in the central-tail region, FHS-Static frequently emerges as the most stable winner across risk classes, especially at $H = 21$.

## Implications for risk governance and global model ordering

Overall, the backtesting evidence supports a clear governance-oriented ordering of the competing engines once performance is assessed jointly across risk classes, horizons, and confidence levels. The filtered historical simulation family represents a material step forward with respect to purely historical resampling engines. In particular, FHS-Recursive (the standard filtered implementation) provides the most reliable calibration in the extreme tail (CL=99.9%) and retains robust UC/CC acceptance at longer horizons, which are precisely the regimes where market-risk governance is most sensitive to underestimation and where the effective backtesting sample is structurally smaller. FHS-Static remains highly competitive and often dominant in the central-tail region, supporting its role as a stable benchmark when a constant volatility-state rescaling is sufficient.

The comparison against HS-$\lambda$ is especially informative for governance. Despite exponentially weighted HS improving responsiveness by prioritizing recent observations, this recency weighting can mechanically down-weight (and, in practice, "forget") stress episodes that lie far back in the sample. By contrast, FHS engines do not require direct recency weights on realized shocks; instead, they integrate a parametric conditional-volatility layer that continuously maps resampled standardized residuals to the volatility regime prevailing at the compute date. This design preserves the information content of past stress scenarios while still adapting the scale of innovations to current market conditions, thereby avoiding the trade-off between responsiveness and memory that characterizes purely weighted historical schemes.

Finally, the recursive filtered engine provides an additional structural advantage: by propagating the conditional volatility state along each simulated trajectory, it can generate extreme scenarios in which volatility clustering is endogenously amplified along the path. This mech-

anism is consistent with stress-period stylized facts, where volatility persistence coexists with increases in effective dependence and tail thickness, and it provides a coherent explanation for the observed outperformance of FHS-Recursive in the most conservative confidence regions. From a risk-governance perspective, these properties translate into an engine that is both scalable across heterogeneous risk classes and materially more reliable for capital-relevant tail quantification than exogenous historical resampling approaches, including those augmented with time-decay weighting.

## 5.2 Risk-factor tail calibration results across horizons via TRGS

### 5.2.1 Results: long-horizon tail calibration and governance ranking

**Empirical vs simulated tail surfaces: shock quantiles across volatility regimes.** Before turning to summary tail scores and inferential testing about TRGS, Tables 13–14 compare the empirical annual-horizon shock quantiles to the corresponding simulated quantiles produced by each engine at $H = 252$. For each risk class we report the same tail probability levels used in the market-risk horizon analysis ($0.1\%, 1\%, 2.5\%, 5\%$ and symmetrically $95\%, 97.5\%, 99\%, 99.9\%$), stratified by conditional initial-volatility regimes $q_{\sigma_0} \in \mathcal{Q} = (75\%, 80\%, 85\%, 90\%, 95\%, 100\%)$. As an optional compact diagnostic, the last column reports the mean absolute deviation from the empirical benchmark across the eight reported tail levels, $\overline{|e|} = \frac{1}{8}\sum_{p\in\mathcal{P}} |\hat{q}_p - q_p^{\text{emp}}|$.

Prior to the governance implications, the tabulated evidence clarifies the mechanism at work. The C-MBB benchmark, by construction, provides the closest empirical proxy available for long-horizon shock distributions, as it fully exploits the information content of the historical sample through all admissible recombinations of past shocks. Against this realistic empirical reference, the behaviour of the filtered engines reflects their endogenous nature (see Paragraph 3.4): because volatility rescaling conditions on the compute-date volatility state, FHS-based engines amplify the scale of simulated multi-period innovations when prevailing volatility is high, and dampen it when volatility is low. When the holding period is sufficiently long for the risk factor to evolve and for dispersion to accumulate, this conditioning enables filtered engines to generate tail scenarios that expand prudently with increasing volatility—producing, in stressed regimes, tail quantiles that become materially larger than those implied by the empirical C-MBB proxy. This divergence is therefore the structural consequence of embedding current volatility information into scenario generation, which aligns filtered engines with governance expectations for long-horizon risk metrics. The tail-surface comparison presented in the following tables brings the risk governance message into sharp focus. While the empirical C-MBB benchmark remains the closest nonparametric approximation to long-horizon tail behavior, the recursive filtered engine departs too rapidly from this empirical reference: its tail divergence grows disproportionately with conditional volatility, reflecting the instability of unconstrained volatility propagation under long-horizon aggregation. By contrast, LT-FHS displays a more disciplined response across regimes. It stays close to the static baseline in benign volatility environments, consistent with its design as a level-preserving extension of FHS-Static, and becomes materially closer to the empirical tails than static filtering in stressed regimes—precisely where long-horizon tails determine economic capital and risk appetite. In this sense, LT-FHS captures the prudential expansion implied by elevated volatility without losing empirical coherence, thereby delivering a balanced and governance-aligned representation of long-horizon tail risk.

Table 13: IR (ESTR OIS Rate in bps)–annual horizon $H = 252$: empirical benchmark vs simulated shock quantiles by conditional volatility regime. Columns report the tail probability levels used in market-risk validation ($0.1\%, 1\%, 2.5\%, 5\%$ and symmetrically $95\%, 97.5\%, 99\%, 99.9\%$). All entries are averaged across risk factors within the risk class. Rows group results by the conditional initial-volatility quantile $q_{\sigma_0}$ (75%–100% in steps of 5%). The last column reports the mean absolute deviation from the empirical quantiles across the eight reported probability levels, $\overline{|e|} = \frac{1}{8}\sum_{p\in\mathcal{P}} |\hat{q}_p - q_p^{\text{emp}}|$, as an optional compact error summary.

| $q_{\sigma_0}$ | Series | 0.1% | 1% | 2.5% | 5% | 95% | 97.5% | 99% | 99.9% | $\overline{\lvert e\rvert}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 75% | Empirical | -269.9 | -193.9 | -160.6 | -132.2 | 116.5 | 147.2 | 183.0 | 263.7 | 0.0 |
| | FHS-Static | -246.8 | -190.2 | -163.0 | -139.0 | 109.6 | 133.5 | 161.6 | 219.9 | 22.4 |
| | LT-FHS | -241.2 | -190.2 | -163.0 | -139.0 | 109.6 | 133.5 | 161.6 | 219.9 | 23.1 |
| | FHS-Recursive | -763.3 | -292.4 | -198.2 | -146.3 | 106.9 | 145.5 | 210.7 | 509.2 | 172.8 |
| 80% | Empirical | -269.9 | -193.9 | -160.6 | -132.2 | 116.5 | 147.2 | 183.0 | 263.7 | 0.0 |
| | FHS-Static | -267.5 | -206.2 | -176.7 | -150.8 | 118.7 | 144.6 | 175.0 | 238.2 | 15.3 |
| | LT-FHS | -255.1 | -206.2 | -176.7 | -150.8 | 118.7 | 144.6 | 175.0 | 238.2 | 17.0 |
| | FHS-Recursive | -842.2 | -317.5 | -213.7 | -156.9 | 113.9 | 155.8 | 227.0 | 555.5 | 192.7 |
| 85% | Empirical | -269.9 | -193.9 | -160.6 | -132.2 | 116.5 | 147.2 | 183.0 | 263.7 | 0.0 |
| | FHS-Static | -310.4 | -239.1 | -205.9 | -177.3 | 140.4 | 170.0 | 204.5 | 277.3 | 32.9 |
| | LT-FHS | -287.4 | -239.1 | -205.9 | -177.3 | 140.4 | 170.0 | 204.5 | 277.3 | 25.7 |
| | FHS-Recursive | -990.8 | -362.9 | -241.5 | -175.7 | 132.7 | 184.0 | 273.2 | 681.2 | 266.1 |
| 90% | Empirical | -269.9 | -193.9 | -160.6 | -132.2 | 116.5 | 147.2 | 183.0 | 263.7 | 0.0 |
| | FHS-Static | -396.4 | -306.3 | -263.8 | -227.1 | 179.4 | 217.3 | 261.4 | 354.2 | 89.7 |
| | LT-FHS | -338.9 | -306.3 | -263.8 | -227.1 | 179.4 | 217.3 | 261.4 | 339.4 | 76.6 |
| | FHS-Recursive | -1255.3 | -443.5 | -297.3 | -215.6 | 151.1 | 222.9 | 340.2 | 862.0 | 389.7 |
| 95% | Empirical | -269.9 | -193.9 | -160.6 | -132.2 | 116.5 | 147.2 | 183.0 | 263.7 | 0.0 |
| | FHS-Static | -578.5 | -447.2 | -385.1 | -331.5 | 262.8 | 318.4 | 383.0 | 519.0 | 201.2 |
| | LT-FHS | -433.0 | -404.6 | -383.9 | -331.5 | 262.8 | 318.4 | 383.0 | 440.6 | 154.3 |
| | FHS-Recursive | -1720.3 | -586.7 | -399.1 | -285.0 | 174.0 | 270.1 | 433.3 | 1153.0 | 611.3 |
| 100% | Empirical | -269.9 | -193.9 | -160.6 | -132.2 | 116.5 | 147.2 | 183.0 | 263.7 | 0.0 |
| | FHS-Static | -836.4 | -646.6 | -557.1 | -479.3 | 380.4 | 461.1 | 555.0 | 752.6 | 433.1 |
| | LT-FHS | -425.1 | -425.1 | -425.1 | -425.1 | 582.6 | 582.6 | 582.6 | 582.6 | 419.1 |
| | FHS-Recursive | -3835.2 | -1130.7 | -703.1 | -496.7 | 269.4 | 451.6 | 874.3 | 3594.2 | 1564.7 |

Table 14: CS (Italy Asset Swap Spread in bps)–annual horizon $H = 252$: empirical benchmark vs simulated shock quantiles by conditional volatility regime. Columns report the tail probability levels used in market-risk validation ($0.1\%, 1\%, 2.5\%, 5\%$ and symmetrically $95\%, 97.5\%, 99\%, 99.9\%$). All entries are averaged across risk factors within the risk class. Rows group results by the conditional initial-volatility quantile $q_{\sigma_0}$ (75%–100% in steps of 5%). The last column reports the mean absolute deviation from the empirical quantiles across the eight reported probability levels.

| $q_{\sigma_0}$ | Series | 0.1% | 1% | 2.5% | 5% | 95% | 97.5% | 99% | 99.9% | $\overline{\lvert e \rvert}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 75% | Empirical | -398.9 | -277.7 | -222.0 | -164.8 | 179.8 | 225.6 | 282.2 | 415.0 | 0.0 |
| | FHS-Static | -300.2 | -222.4 | -184.8 | -152.2 | 188.8 | 222.5 | 261.7 | 344.3 | 38.4 |
| | LT-FHS | -234.0 | -213.5 | -184.8 | -152.2 | 188.8 | 222.5 | 261.7 | 344.3 | 47.8 |
| | FHS-Recursive | -769.3 | -264.1 | -181.2 | -138.2 | 219.4 | 328.2 | 546.2 | 1962.5 | 300.6 |
| 80% | Empirical | -398.9 | -277.7 | -222.0 | -164.8 | 179.8 | 225.6 | 282.2 | 415.0 | 0.0 |
| | FHS-Static | -339.3 | -251.4 | -208.9 | -172.1 | 213.6 | 251.7 | 296.1 | 389.5 | 25.7 |
| | LT-FHS | -247.3 | -235.5 | -208.9 | -172.1 | 213.6 | 251.7 | 296.1 | 389.5 | 38.9 |
| | FHS-Recursive | -835.0 | -281.5 | -191.4 | -145.3 | 230.8 | 348.0 | 584.1 | 2126.0 | 334.6 |
| 85% | Empirical | -398.9 | -277.7 | -222.0 | -164.8 | 179.8 | 225.6 | 282.2 | 415.0 | 0.0 |
| | FHS-Static | -380.8 | -282.4 | -234.6 | -193.5 | 239.3 | 282.0 | 331.8 | 436.3 | 38.5 |
| | LT-FHS | -266.7 | -262.4 | -238.1 | -200.8 | 249.3 | 293.8 | 345.6 | 454.6 | 47.9 |
| | FHS-Recursive | -925.0 | -307.1 | -207.4 | -156.3 | 248.7 | 377.6 | 640.6 | 2389.0 | 377.0 |
| 90% | Empirical | -398.9 | -277.7 | -222.0 | -164.8 | 179.8 | 225.6 | 282.2 | 415.0 | 0.0 |
| | FHS-Static | -493.0 | -365.5 | -303.6 | -250.2 | 310.8 | 366.4 | 430.9 | 566.9 | 123.1 |
| | LT-FHS | -300.3 | -300.3 | -287.1 | -250.2 | 310.8 | 366.4 | 430.9 | 558.7 | 107.5 |
| | FHS-Recursive | -1087.2 | -355.5 | -237.2 | -176.9 | 281.5 | 433.2 | 748.6 | 2858.4 | 469.5 |
| 95% | Empirical | -398.9 | -277.7 | -222.0 | -164.8 | 179.8 | 225.6 | 282.2 | 415.0 | 0.0 |
| | FHS-Static | -698.9 | -518.4 | -430.6 | -354.8 | 441.9 | 520.7 | 612.6 | 806.0 | 287.5 |
| | LT-FHS | -372.5 | -372.5 | -369.6 | -344.0 | 441.9 | 520.7 | 612.6 | 691.0 | 229.3 |
| | FHS-Recursive | -1500.4 | -471.6 | -307.6 | -225.4 | 360.2 | 566.9 | 1011.0 | 3966.0 | 831.8 |
| 100% | Empirical | -398.9 | -277.7 | -222.0 | -164.8 | 179.8 | 225.6 | 282.2 | 415.0 | 0.0 |
| | FHS-Static | -2683.4 | -1995.0 | -1655.5 | -1364.8 | 1709.6 | 2013.6 | 2371.2 | 3120.0 | 1915.6 |
| | LT-FHS | -1081.2 | -1081.2 | -1081.2 | -1081.2 | 1427.6 | 1427.6 | 1427.6 | 1427.6 | 1211.1 |
| | FHS-Recursive | -5949.3 | -1753.3 | -1086.5 | -764.6 | 1234.5 | 2050.8 | 3863.5 | 16576.0 | 4656.2 |

**Long-horizon calibration scores across holding periods.** Table 15 reports the TRGS outputs across risk classes and horizons, averaged over risk factors and initial-volatility regimes. While all engines become less accurate as $H$ increases, the annual-horizon ranking is decisive: LT-FHS dominates the filtered family at $H = 252$, whereas the recursive filtered engine deteriorates sharply and becomes orders of magnitude worse.

Table 15: TRGS (in bps) averaged across risk factors and volatility regimes. The weighted score emphasizes the right tail.

| **Risk class** | $H$ | **Engine** | **TRGS** | **Left-tail RMSE** | **Right-tail RMSE** |
|---|---|---|---|---|---|
| IR | 5 | FHS-Recursive | 12.5 | 12.9 | 12.3 |
| IR | 5 | FHS-Static | 11.9 | 12.3 | 11.6 |
| IR | 5 | LT-FHS | 11.9 | 12.0 | 11.9 |
| IR | 10 | FHS-Recursive | 17.0 | 18.8 | 16.1 |
| IR | 10 | FHS-Static | 15.7 | 17.3 | 14.8 |
| IR | 10 | LT-FHS | 15.8 | 17.2 | 15.0 |
| IR | 21 | FHS-Recursive | 25.2 | 28.5 | 23.4 |
| IR | 21 | FHS-Static | 22.4 | 24.9 | 21.1 |
| IR | 21 | LT-FHS | 22.2 | 24.2 | 21.2 |
| IR | 252 | FHS-Recursive | 390.0 | 559.9 | 273.6 |
| IR | 252 | FHS-Static | 75.0 | 95.0 | 63.2 |
| IR | 252 | LT-FHS | 69.2 | 79.6 | 61.1 |
| CS | 5 | FHS-Recursive | 49.3 | 38.0 | 52.9 |
| CS | 5 | FHS-Static | 44.1 | 39.6 | 45.7 |
| CS | 5 | LT-FHS | 43.5 | 35.7 | 46.2 |
| CS | 10 | FHS-Recursive | 74.3 | 51.1 | 81.8 |
| CS | 10 | FHS-Static | 62.7 | 55.1 | 65.5 |
| CS | 10 | LT-FHS | 59.9 | 47.0 | 64.3 |
| CS | 21 | FHS-Recursive | 124.3 | 73.8 | 139.6 |
| CS | 21 | FHS-Static | 87.6 | 81.9 | 89.7 |
| CS | 21 | LT-FHS | 83.0 | 67.3 | 88.3 |
| CS | 252 | FHS-Recursive | 1617.7 | 490.2 | 1895.1 |
| CS | 252 | FHS-Static | 300.2 | 255.2 | 314.8 |
| CS | 252 | LT-FHS | 230.2 | 175.6 | 245.2 |

*Takeaways (Table 15).* At $H = 252$, both FHS-Static and LT-FHS remain in a controlled range, whereas FHS-Recursive becomes dramatically worse in both risk classes, confirming that recursion is not governance-defensible as a standalone annual-horizon tail engine. LT-FHS improves upon the static baseline in both classes, and the improvement is economically large for Credit Spreads.

**Inferential evidence on score differences.** To conclude decisively on whether the observed improvements are systematic rather than descriptive, we perform paired inferential tests on score differences across the full grid of *risk factor* × *initial-volatility regime* (paired unit, $n = 66$ per dataset and holding period). Table 16 reports the Holm-corrected [5] two-sided $p$-values for (i)

[5]To control the family-wise error rate (FWER) in inferential testing of Tail Risk Governance Score (TRGS) differences, Holm step-down correction [Holm, 1979] is applied to the set of two-sided $p$-values for all pairwise

paired $t$-tests and (ii) Wilcoxon signed-rank tests, together with permutation and sign-flip tests for robustness. All three pairwise relations are shown: LT-FHS vs Static, Static vs Recursive, and LT-FHS vs Recursive.

Table 16: Annual horizon ($H = 252$): paired inference on TRGS differences (bps), with Holm-corrected two-sided $p$-values. Each row tests the mean/median difference $\Delta = \text{Score}(A) - \text{Score}(B)$ over paired units (risk factor $\times$ volatility regime). Negative $\Delta$ indicates that engine $A$ outperforms engine $B$.

| Dataset | Comparison ($A - B$) | $n$ | $\Delta$ (bps) | $p_{\text{t,Holm}}$ | $p_{\text{W,Holm}}$ | $p_{\text{perm,Holm}}$ |
|---|---|---|---|---|---|---|
| IR | LT-FHS – FHS-Static | 66 | -5.9 | 0.037 | 0.012 | 0.006 |
| IR | FHS-Static – FHS-Recursive | 66 | -315.0 | 0.020 | $< 10^{-4}$ | $< 10^{-4}$ |
| IR | LT-FHS – FHS-Recursive | 66 | -320.9 | 0.020 | $< 10^{-4}$ | $< 10^{-4}$ |
| CS | LT-FHS – FHS-Static | 66 | -70.0 | 0.018 | 0.090 | 0.0012 |
| CS | FHS-Static – FHS-Recursive | 66 | -1317.5 | 0.005 | $< 10^{-4}$ | $< 10^{-4}$ |
| CS | LT-FHS – FHS-Recursive | 66 | -1387.5 | 0.005 | $< 10^{-4}$ | $< 10^{-4}$ |

*Takeaways (Table 16).* Across both risk classes, Recursive is statistically and economically far from both Static and LT-FHS at $H = 252$. For LT-FHS vs Static, the improvement is supported by Holm-corrected inference; in CS, evidence is strongest under permutation testing even when the Wilcoxon test is conservative, reflecting the heavy-tailed and skewed nature of differences.

### 5.2.2 Annual-horizon diagnostics: tail error, kurtosis, and relative performance

Table 17 isolates the annual horizon and augments tail-error metrics with mean kurtosis computed over simulated annual distributions. This provides a direct governance diagnostic: engines that generate exploding kurtosis at $H = 252$ violate long-horizon plausibility constraints and become unreliable for ICAAP/ECAP/Solvency tail quantification.

Table 17: Annual-horizon ($H = 252$) tail calibration diagnostics. "Rel. vs Static" reports the percentage difference in TRGS relative to FHS-Static. Kurtosis is excess kurtosis of the simulated annual distribution averaged over risk factors and volatility regimes.

| **Risk class** | **Engine** | **TRGS** | **Left-tail RMSE** | **Right-tail RMSE** | **Mean kurtosis** | **Rel. vs Static (%)** |
|---|---|---|---|---|---|---|
| IR | FHS-Recursive | 390.0 | 559.9 | 273.6 | 1131.4 | +419.7 |
| IR | FHS-Static | 75.0 | 95.0 | 63.2 | 0.0 | 0.0 |
| IR | LT-FHS | 69.2 | 79.6 | 61.1 | -0.1 | -7.8 |
| CS | FHS-Recursive | 1617.7 | 490.2 | 1895.1 | 3906.5 | +438.8 |
| CS | FHS-Static | 300.2 | 255.2 | 314.8 | 0.0 | 0.0 |
| CS | LT-FHS | 230.2 | 175.6 | 245.2 | -0.3 | -23.3 |

*Takeaways (Table 17).* At $H = 252$, LT-FHS improves both tails relative to the static filtered baseline while maintaining a controlled (near-normal) annual kurtosis, consistent with long-horizon normalization constraints. Conversely, FHS-Recursive exhibits severe kurtosis explosion and a correspondingly unstable right tail, explaining its large score deterioration.

engine comparisons within each risk class and holding period. Specifically, for each aggregation unit (risk class $\times$ holding period), the three pairwise differences in TRGS (tail-focused RMSE) between engines are tested using paired $t$-tests, Wilcoxon signed-rank, and permutation tests. For each test type, the corresponding $p$-values are sorted in increasing order $p_{(1)} \le p_{(2)} \le p_{(3)}$ and adjusted as $p_{(k)}^{\text{Holm}} = \max_{j \le k} \{(3 - j + 1) \cdot p_{(j)}\}$ for $k = 1, 2, 3$, ensuring strong FWER control within each family of engine comparisons. Only Holm-corrected $p$-values for the TRGS metric are reported in the tables.

### 5.2.3 Granular governance: pillar-level ranking at the annual horizon

Long-horizon governance requires maturity granularity: tail miscalibration is not uniform along the curve. Tables 18–19 report pillar-level TRGS at $H = 252$ averaged over volatility regimes.

Table 18: IR (ESTR OIS Rate) pillar-level results at $H = 252$, TRGS (in bps) averaged over volatility regimes. Improvements are relative to FHS-Static.

| Risk factor | FHS-Static | LT-FHS | LT vs Static (%) | FHS-Recursive | Rec vs Static (%) |
|---|---|---|---|---|---|
| ESTR OIS 1Y | 94.9 | 83.9 | -11.6 | 1801.9 | 1798.0 |
| ESTR OIS 3Y | 76.1 | 69.5 | -8.7 | 190.5 | 150.3 |
| ESTR OIS 5Y | 61.8 | 59.3 | -4.1 | 106.5 | 72.3 |
| ESTR OIS 10Y | 59.8 | 57.7 | -3.4 | 64.3 | 7.7 |
| ESTR OIS 20Y | 72.5 | 68.0 | -6.1 | 79.8 | 10.2 |
| ESTR OIS 30Y | 85.2 | 76.4 | -10.2 | 96.9 | 13.8 |

Table 19: CS (Italy Asset Swap Spread) pillar-level results at $H = 252$, TRGS (in bps) averaged over volatility regimes. Improvements are relative to FHS-Static.

| Risk factor | FHS-Static | LT-FHS | LT vs Static (%) | FHS-Recursive | Rec vs Static (%) |
|---|---|---|---|---|---|
| Italy ASW 1Y | 462.0 | 348.3 | -24.6 | 4192.3 | 807.4 |
| Italy ASW 3Y | 394.3 | 287.5 | -27.1 | 2592.9 | 557.6 |
| Italy ASW 5Y | 328.2 | 246.0 | -25.1 | 1405.3 | 328.2 |
| Italy ASW 10Y | 227.0 | 183.8 | -19.1 | 650.1 | 186.3 |
| Italy ASW 20Y | 196.3 | 157.3 | -19.9 | 478.9 | 144.0 |
| Italy ASW 30Y | 193.4 | 158.3 | -18.1 | 386.5 | 99.8 |

*Takeaways (Tables 18–19).* Recursive propagation is consistently far from both Static and LT-FHS, with the largest instability at the IR short end (1Y) and the CS front pillars (1Y–5Y). LT-FHS improvements over Static are maturity-dependent and concentrate where annual tails are most difficult and most capital-relevant.

### 5.2.4 Volatility regime dependence: LT-FHS gains increase with volatility

To assess regime sensitivity, we evaluate performance across initial volatility states $\sigma_0^{(j)}(q)$. Tables 20–21 report mean TRGS at $H = 252$ by initial-volatility quantile. The key governance message is that LT-FHS may be close to (or slightly more conservative than) Static in benign regimes, but the incremental benefit increases sharply as volatility rises, becoming decisive in stressed states where annual tails drive capital.

Table 20: IR (ESTR OIS Rate) regime dependence at $H = 252$: mean TRGS (in bps) by initial volatility quantile.

| **Initial $\sigma_0$ quantile** | **FHS-Static** | **LT-FHS** | **LT vs Static (%)** | **FHS-Recursive** | **Rec vs Static (%)** |
|---|---|---|---|---|---|
| 75% | 24.6 | 25.3 | +2.9 | 142.7 | +480.2 |
| 85% | 35.2 | 35.9 | +1.9 | 246.8 | +600.4 |
| 90% | 48.8 | 47.9 | -1.9 | 258.4 | +429.3 |
| 95% | 78.3 | 74.7 | -4.6 | 593.8 | +658.2 |
| 100% | 346.7 | 288.2 | -16.9 | 1430.0 | +312.4 |

Table 21: CS (Italy Asset Swap Spread) regime dependence at $H = 252$: mean TRGS (in bps) by initial volatility quantile.

| **Initial $\sigma_0$ quantile** | **FHS-Static** | **LT-FHS** | **LT vs Static (%)** | **FHS-Recursive** | **Rec vs Static (%)** |
|---|---|---|---|---|---|
| 75% | 50.1 | 57.0 | +13.9 | 748.2 | +1394.3 |
| 85% | 83.0 | 87.6 | +5.6 | 684.3 | +724.8 |
| 90% | 135.9 | 133.4 | -1.8 | 941.0 | +592.4 |
| 95% | 285.1 | 262.4 | -8.0 | 1437.4 | +404.2 |
| 100% | 1800.8 | 1094.4 | -39.2 | 7783.4 | +332.2 |

*Takeaways (Tables 20–21).* The incremental value of LT-FHS is concentrated in stressed regimes: the improvement over Static increases with initial volatility and becomes decisive at the top regime. This is precisely the governance-relevant region where annual-horizon tails drive capital and risk appetite. By contrast, the recursive engine remains excessively distant from both Static and LT-FHS in all regimes, with extreme amplification in high volatility.

### Long-horizon risk governance implications

The long-horizon evidence yields a clear governance ordering within the filtered HS family. FHS-Static is a stable baseline with controlled annual tails, but it can become excessively sensitive to stressed initial volatility in long-horizon conditioning. LT-FHS closes this gap by enforcing long-horizon admissibility through level-anchored bounds and a regime-consistent buffer, thereby preventing implausible drift and stabilizing annual tails. Quantitatively, LT-FHS improves long-horizon tail goodness-of-fit relative to Static, and the improvement is supported by paired inferential testing with Holm correction on score differences. Conversely, FHS-Recursive is systematically too distant from both Static and LT-FHS and exhibits severe kurtosis explosion at $H = 252$, making it unsuitable as a standalone long-term capital engine without explicit stabilizers. The failure of the FHS-Recursive engine at long horizons admits a clear theoretical interpretation. As evidenced by the results, recursive volatility updating induces a compounding of conditional scale shocks that is not asymptotically regularized in the tails. In the presence of recursive volatility propagation, bootstrap-generated clusters of extreme standardized shocks induce persistent high-volatility regimes that do not dissipate within finite horizons; consequently, large contributions to aggregated risk fail to become asymptotically negligible. This mechanism leads to a finite-horizon violation of the Lindeberg regularization underlying the Feller–Lindeberg CLT, despite the stationarity of the underlying GJR-GARCH dynamics, as higher-order moments of the aggregated shocks are not uniformly controlled along extreme paths. By contrast, both FHS-Static and LT-FHS operate under a fixed volatility anchor at the compute date, so that long-horizon aggregation effectively occurs at constant conditional scale. In this setting, regularization is achieved by summing shocks drawn from the empirically ob-

served innovation distribution, which acts as a benchmark for long-run aggregation and restores the conditions required for Gaussian-domain convergence. This constant-variance aggregation reinforces asymptotic convergence, restores Gaussian-domain regularization, and results in a systematic decline of excess kurtosis at annual horizons. Building on this stabilized aggregation, the horizon-level bounding mechanism of LT-FHS provides an additional layer of control by constraining cumulative shocks within historically and volatility-consistent admissible ranges, further smoothing extreme realizations. Crucially, this architecture is particularly suited to very long holding periods—such as those required in ICAAP, ECAP and Solvency frameworks.

# 6 Conclusions

This paper has pursued a deliberately governance-oriented objective: to make tail-risk measurement coherent across horizons in the "world of facts," where institutions must simultaneously run short-horizon market-risk monitoring and one-year capital frameworks under strict constraints of scalability, auditability, and defensible validation, in line with the supervisory expectations highlighted in the regulatory framework. Building from the canonical trade-offs reviewed in the literature, we argued that the binding constraint for real institutions is not theoretical flexibility but governability: controlling model-risk degrees of freedom, embedding dependence consistently at scale, and sustaining an evidence-based validation program that can be repeated, explained, and challenged. This logic motivates the focus on the historical-simulation family and, within it, on filtered architectures that preserve empirical cross-sectional dependence through joint vector resampling while addressing volatility clustering through an econometric layer, thereby providing a direct assist to risk governance by supporting efficient capital deployment and ensuring scalability over heterogeneous risk factors and multiple holding periods. Methodologically, we then formalized a common multivariate bootstrap architecture spanning HS-iid and HS-$\lambda$ benchmarks and a filtered family (FHS-Static, FHS-Recursive), and introduced LT-FHS as a long-horizon extension of FHS-Static that adds an admissibility map anchored to historical levels with a volatility-conditioned buffer. The empirical program was designed to respect the temporal structure of the governance problem: at market-risk horizons ($H \in \{5, 10, 21\}$, i.e. 1 week, 2 weeks, 1 month), where sufficient out-of-sample non-overlapping observations exist, we evaluated engines with standard UC/CC/DQ backtests and summarized the high-dimensional decision grid via a Composite Risk Governance Score (CRGS) that (i) enforces a uniform pass/fail logic across ($H$, risk class, CL) while (ii) preserving margin-to-rejection as an early-warning monitoring signal and (iii) gating dynamic diagnostics when exceedances are structurally scarce. Within our risk governance framework, we complement portfolio-level backtesting with a risk-factor-level analysis of tail behavior. Specifically, we introduce the Tail Risk Governance Score (TRGS) to benchmark the plausibility of simulated tail quantiles for each risk factor, across all relevant holding periods. This approach enables a granular assessment of tail risk calibration, allowing engines to be ranked and validated not only at the aggregate portfolio level but also in terms of their ability to reproduce empirical tail features at the risk-factor level—an essential requirement for robust, multi-horizon risk governance. Importantly, our analysis extends up to the 1 year horizon ($H = 252$), in line with supervisory expectations for risk assessment, ensuring that the framework remains applicable and defensible for both short-term risk monitoring and long-term capital-adequacy and solvency evaluations. The resulting evidence supports a horizon-dependent architecture that is straightforward to govern: filtered historical simulation (FHS) engines are consistently competitive for short-horizon risk monitoring across all risk classes. In contrast, effective long-horizon risk governance requires explicit control of tail plausibility at the level of aggregated, multi-period risk-factors shocks, ensuring that scenario engines remain robust and defensible when assessing extreme outcomes over extended time frames. In particular, unconstrained recursive volatility propagation (the standard Filtered Historical Simulation design) is not defensible as a standalone long-horizon

engine: it produces tails that are systematically too distant from the empirical benchmark, with severe kurtosis explosion and volatility-regime amplification, precisely in the states that dominate capital and risk appetite. FHS-Static (flat rescaling at the compute-date volatility state) provides a stable baseline with controlled long-horizon tails, but it can become overly sensitive to stressed initial volatility. The Long-Term Filtered Historical Simulation (LT-FHS), introduced in this paper, closes this gap by enforcing admissibility through level-anchored bounds and a volatility regime-consistent buffer, thereby preventing implausible drift and stabilizing the long-horizon tail while preserving short-horizon behavior. Quantitatively, LT-FHS delivers the best annual-horizon tail-quantile accuracy and the closest match to long-horizon empirical tail benchmarks, with improvements that are statistically supported where governance scrutiny is highest and where prudence matters most. To summarize, the paper yields a practical, regulator-aligned playbook: adopt a historical-simulation foundation for high-dimensional scalability and auditability; use a filtered layer to improve joint coverage and loss-severity diagnostics at market-risk horizons; complement short-horizon backtesting with a risk governance score that enables consistent ranking across a large validation grid; and, for annual holding periods, replace the illusion of "short-term backtesting sufficiency" with benchmark-driven tail diagnostics, score-based inference, and regime-stratified monitoring — culminating in LT-FHS as the preferred engine within the filtered family when institutions must deliver coherent, prudential tail metrics under stressed-market conditions and efficient economic-capital estimation, avoiding undue overestimation of capital requirements due to model risk at extended horizons.

## References


Directive 2013/36/eu (crd iv), article 73: Internal capital adequacy assessment process, 2013. Official Journal of the European Union L 176/338.

Beatrice Acciaio, Stephanirk Eckstein, and Songyan Hou. Time-causal vae: Robust financial time series generator. *ArXiv.org*, (Papers 2411.02947), 2024.

Carlo Acerbi and Balázs Szekely. General properties of backtestable statistics. *Mathematics and Financial Economics*, 11(2):181–202, 2017.

Carlo Acerbi and Dirk Tasche. Expected shortfall: A natural coherent alternative to value at risk. *Economic Notes*, 31(2):379–388, 2002.

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

European Banking Authority. Eba guidelines on common procedures and methodologies for the supervisory review and evaluation process (srep), 2014.

European Banking Authority. Revised eba guidelines on srep and supervisory stress testing, 2018.

Giovanni Barone-Adesi and Kostas Giannopoulos. A simplified approach to the conditional estimation of value at risk. Futures and Options World, 1996. October, pp. 68–72.

Giovanni Barone-Adesi and Kostas Giannopoulos. Non-parametric VaR techniques: Myths and realities. *Economic Notes*, 30(2):167–181, 2001.

Giovanni Barone-Adesi, Frederick Bourgoin, and Kostas Giannopoulos. Don't look back. *Risk*, 11:100–104, August 1998.

Giovanni Barone-Adesi, Kostas Giannopoulos, and Les Vosper. VaR without correlations for portfolios of derivative securities. *Journal of Futures Markets*, 19(5):583–602, August 1999.

Giovanni Barone-Adesi, Kostas Giannopoulos, and Les Vosper. Backtesting derivative portfolios with filtered historical simulation (fhs). *European Financial Management*, 8(1):31–58, 2002.

Giovanni Barone-Adesi, Robert F. Engle, and Loriano Mancini. A GARCH option pricing model with filtered historical simulation. *The Review of Financial Studies*, 21(3):1223–1258, 2008.

Giovanni Barone-Adesi, Kostas Giannopoulos, and Les Vosper. Estimating the joint tail risk under the filtered historical simulation. an application to the ccp's default and waterfall fund. *European Journal of Finance*, 23(11):1009–1036, 2017.

Basel Committee on Banking Supervision. Amendment to the capital accord to incorporate market risks. Technical Report 23, BIS, 1996.

Basel Committee on Banking Supervision. Guidelines for computing capital for incremental risk in the trading book. Technical Report 159, BIS, 2009.

Basel Committee on Banking Supervision. Minimum capital requirements for market risk. Technical Report 457, BIS, 2019.

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

Michele Bonollo, Martino Grasselli, Gianmarco Mori, and Havva Oz Nilsu. Informative risk measures in the banking industry: A proposal based on the magnitude-propensity approach, 2025. Available at SSRN.

Jacob Boudoukh, Matthew Richardson, and Robert Whitelaw. The best of both worlds: A hybrid approach to calculating value at risk. *Risk*, 11(5):64–67, May 1998.

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

Paul E. Desmier and David W. Maybury. Quantifying and managing foreign exchange risk in the canadian department of national defence. In *RTO-MP-SAS-093: Analytical Support to Defence Transformation*. NATO Science and Technology Organization, 2011.

Bradley Efron and Robert Tibshirani. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1(1):54–75, 1986.

Bradley Efron and Robert Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, 1993.

Robert F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 50(4):987–1007, 1982.

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

European Banking Authority. Final report on guidelines on internal governance under directive 2013/36/eu. Technical Report EBA/GL/2021/05, EBA, 2021.

European Central Bank. Ecb guide to the internal capital adequacy assessment process (icaap), 2018a.

European Central Bank. Ecb supervisory manual: On-site inspections and internal model investigations, 2018b.

European Commission. Solvency ii delegated regulation, 2015. (EU) 2015/35.

European Parliament. (eu) 2024/1623 amending regulation (eu) no 575/2013 as regards requirements for credit risk, cva risk, operational risk, market risk and the output floor, 2024. (EU) 2024/1623.

European Union. Directive 2009/138/ec (solvency ii), 2009. Directive 2009/138/EC.

Olivier Faugeras and Gilles Pagès. Risk quantization by magnitude and propensity. *Insurance: Mathematics and Economics*, 116:134–147, 2024.

Mark Garman. Taking var to pieces. *Risk*, 10(10), 1997.

Kostas Giannopoulos, Ramzi Nekhili, and Christos Christodoulou-Volos. Estimating tail risk in ultra-high-frequency cryptocurrency data. *International Journal of Financial Studies*, 12(4): 99, 2024.

Paul Glasserman. *Monte Carlo Methods in Financial Engineering*. Springer, 2003.

Lawrence R. Glosten, Ravi Jagannathan, and David E. Runkle. 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, 1993.

Darryll Hendricks. Evaluation of value-at-risk models using historical data. *Economic Policy Review*, 2(1):39–69, 1996.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

John Hull and Alan White. Incorporating volatility updating into the historical simulation method for value at risk. *Journal of Risk*, 1(1):5–19, 1998.

Philippe Jorion. *Value at Risk: The New Benchmark for Managing Financial Risk*. McGraw-Hill, 3 edition, 2007.

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

Noureddine Lehdili and Arshia Givi. Efficient computation of value-at-risk and expected shortfall in large and heterogeneous credit portfolios: Application to default risk charge. *Risk and Decision Analysis*, 7(3-4):91–105, 2018.

Jacques Longerstaey and Martin Spencer. Riskmetrics technical document. Technical report, J.P. Morgan, 1996. 4th edition.

Alexander J. McNeil, Rüdiger Frey, and Paul Embrechts. *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton University Press, 2 edition, 2015.

Serguei Y. Novak. *Extreme Value Methods with Applications to Finance*. Chapman & Hall / CRC Press, 2011.

Dimitris N. Politis and Joseph P. Romano. A circular block resampling procedure for stationary data. In *Exploring the Limits of Bootstrap*, pages 263–270. Wiley, 1992.

Matthew Pritsker. The hidden dangers of historical simulation. Finance and Economics Discussion Series 2001-27, Board of Governors of the Federal Reserve System, 2001.

Andrea Resti and Andrea Sironi. *Risk Management and Shareholders' Value in Banking: From Risk Measurement Models to Capital Allocation Policies*. Wiley, 2012.

Indrajit Roy. Estimating value at risk (var) using filtered historical simulation in the indian capital market. *Reserve Bank of India Occasional Papers*, 32(2):81–98, 2011.

Sasa Zikovic and Randall K. Filer. Ranking of var and es models: Performance in developed and emerging markets. Technical report, CESifo Working Paper No. 3980, 2012.

## Appendix: Notations

Notation used throughout the paper with symbols introduced according to the order of their appearance in the paper.

| Symbol | Meaning |
|---|---|
| $t = 1, \ldots, T$ | Time index for the full historical sample; $T$ is the total number of daily observations. |
| $d$ | Number of risk factors (dimension of the multivariate shock vector). |
| $\mathbf{x}_t = \big(x_{t,1}, \ldots, x_{t,d}\big)^\top \in \mathbb{R}^d$ | Multivariate daily shock (or return) vector observed at time $t$. |
| $x_{t,j}$ | Shock (or return) of risk factor $j$ at time $t$ (component of $\mathbf{x}_t$). |
| $\tau$ (or $\tau_k$) | Computation date; $\tau_k$ denotes the $k$-th out-of-sample compute date for horizon $H$. |
| $\mathcal{H}_\tau := \{\mathbf{x}_t : t \le \tau^-\}$ | Information set (historical database) available at computation date $\tau$. |
| $n = n(\tau)$ | Number of observations available in $\mathcal{H}_\tau$ (re-indexed as $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$). |
| $H \in \mathbb{N}$ | Holding period / pathway length in business days (e.g., $5, 10, 21, 252$). |
| $M \in \mathbb{N}$ | Number of simulated scenarios (bootstrap pathways) generated at each compute date. |
| $I_1^{(m)}, \ldots, I_H^{(m)} \in \{1, \ldots, n\}$ | Resampled historical indices used to build the $H$-step path for scenario $m$. |
| $\{w_i\}_{i=1}^n$ | Sampling probabilities over historical indices (engine-specific). |
| $w_i^{\text{iid}} = \frac{1}{n}$ | Uniform sampling probabilities (HS-iid). |
| $\lambda \in (0, 1)$ | Exponential decay factor (HS-$\lambda$). |
| $\tilde{w}_i(\lambda) = (1-\lambda)\lambda^{n-i}$ | Unnormalized exponential time-decay weights (HS-$\lambda$). |
| $w_i^\lambda = \dfrac{(1-\lambda)\lambda^{n-i}}{1-\lambda^n}$ | Normalized exponential sampling probabilities (HS-$\lambda$). |
| $\mathbf{X}^{(m)} \in \mathbb{R}^{H \times d}$ | $H$-step multivariate path of shocks obtained by resampling joint historical vectors. |
| $\mu_{t,j}$ | Conditional mean of factor $j$ at time $t$ in the filtering decomposition. |
| $\varepsilon_{t,j}$ | Innovation (mean-adjusted shock): $x_{t,j} = \mu_{t,j} + \varepsilon_{t,j}$. |
| $\sigma_{t,j}$ | Conditional volatility of factor $j$ at time $t$. |
| $z_{t,j}$ | Standardized residual with $\varepsilon_{t,j} = \sigma_{t,j} z_{t,j}$. |
| $\hat{z}_{t,j} := \varepsilon_{t,j}/\sigma_{t,j}$ | Estimated standardized residual used for residual bootstrap in FHS engines. |
| $\hat{\mathbf{z}}_t = \big(\hat{z}_{t,1}, \ldots, \hat{z}_{t,d}\big)^\top$ | Multivariate standardized residual vector resampled jointly to preserve cross-sectional dependence. |
| $\mathcal{F}_{t-1}$ | Information set (filtration) conditioning mean/variance at time $t-1$. |
| $\phi^{(j)}$ | AR(1) coefficient (no intercept) for factor $j$ in the mean equation. |

*Continued on next page.*

| Symbol | Meaning |
|---|---|
| $\omega^{(j)}, \alpha^{(j)}, \beta^{(j)}, \gamma^{(j)}$ | GJR-GARCH(1,1) parameters for factor $j$ (intercept/ARCH/GARCH/leverage). |
| $\mathbf{\Phi} = \mathrm{diag}(\phi_1, \ldots, \phi_d)$ | Diagonal matrix of AR(1) coefficients used in vector propagation. |
| $\boldsymbol{\sigma}_0(\tau) = (\sigma_{0,1}, \ldots, \sigma_{0,d})^\top$ | Target initial conditional volatility vector at date $\tau$ (e.g., last fitted volatility prior to $\tau$). |
| $\sigma_{h,j}^{(m)}$ | Path-dependent conditional volatility at step $h$ for factor $j$ in scenario $m$ (FHS-Recursive). |
| $\varepsilon_{h,j}^{(m)} = \sigma_{h,j}^{(m)}\, z_{h,j}^{(m)}$ | Simulated innovation at step $h$ for factor $j$ in scenario $m$ (recursive rescaling). |
| $\varepsilon_{h,j}^{(m)} = \sigma_{0,j}\, z_{h,j}^{(m)}$ | Static-rescaled innovation using fixed $\sigma_{0,j}$ (FHS-Static; baseline for LT-FHS). |
| $\mathbf{r}_h^{(m)}$ | Simulated filtered shock vector at step $h$ (after AR(1) propagation). |
| $L_t^{(j)}$ | Historical level series (in levels) for risk factor $j$ used in LT-FHS bounds. |
| $L_{\min}^{(j)} = \min_t L_t^{(j)}$ | Historical minimum level for factor $j$. |
| $L_{\max}^{(j)} = \max_t L_t^{(j)}$ | Historical maximum level for factor $j$. |
| $L_{\mathrm{cmp}}^{(j)}$ | Reference/anchor level in the LT-FHS bounding map (paper uses sample-average level as a neutral anchor). |
| $C_m^{(j)} = \sum_{k=1}^{H} r_{k,m}^{(j)}$ | Raw cumulative shock over horizon $H$ in scenario $m$ (pre-bounding). |
| $\tilde{\sigma}_{k,m}^{(j)}$ | Pathwise volatility proxy driven by static innovations along scenario $m$. |
| $RV_{H,m}^{(j)} = \left(\sum_{k=1}^{H} (\tilde{\sigma}_{k,m}^{(j)})^2\right)^{1/2}$ | Realized-volatility proxy over horizon $H$ for scenario $m$. |
| $\overline{RV}_H^{(j)} = \frac{1}{M}\sum_{m=1}^{M} RV_{H,m}^{(j)}$ | Average realized-volatility proxy across $M$ scenarios (buffer magnitude). |
| $LB^{(j)}$ | Lower admissible bound for cumulative shocks (level-anchored bound minus buffer). |
| $UB^{(j)}$ | Upper admissible bound for cumulative shocks (level-anchored bound plus buffer). |
| $S_{H,\mathrm{LT},m}^{(j)}$ | LT-FHS bounded cumulative shock (clipped version of $C_m^{(j)}$ within $[LB^{(j)}, UB^{(j)}]$). |
| $RC \in \{\mathrm{EQ}, \mathrm{IR}, \mathrm{CS}\}$ | Risk class: Equity (EQ), Interest Rate (IR), Credit Spread (CS). |
| $d_{RC}$ | Number of risk factors within risk class $RC$. |
| $\mathbf{Sens}_{RC} \in \mathbb{R}^{d_{RC}}$ | Portfolio sensitivity vector for risk class $RC$ (e.g., €/bps for IR/CS). |
| $\mathbf{S}_{H,\tau_k}^{(RC)}$ | Realized non-overlapping $H$-day cumulative shock vector for risk class $RC$ after $\tau_k$ (additive aggregation for IR/CS; compounded aggregation for EQ). |
| $\mathbf{S}_{H,\tau_k,\mathrm{sim}}^{(RC,m)}$ | Simulated non-overlapping $H$-day cumulative shock (or multi-period return) vector for risk class $RC$ in scenario $m$ at compute date $\tau_k$. |

*Continued on next page.*

| Symbol | Meaning |
|---|---|
| $\mathrm{PnL}_{\tau_k}^{(H,RC)} = \mathbf{Sens}_{RC}^{\top}\mathbf{S}_{H,\tau_k}^{(RC)}$ | Realized $H$-day portfolio profit-and-loss at backtesting date $\tau_k$ for risk class $RC$. |
| $\mathrm{PnL}_{\tau_k,\mathrm{sim}}^{(H,RC,m)} = \mathbf{Sens}_{RC}^{\top}\mathbf{S}_{H,\tau_k,\mathrm{sim}}^{(RC,m)}$ | Simulated $H$-day portfolio PnL under scenario $m$ at compute date $\tau_k$ for risk class $RC$. |
| $\mathrm{CL} \in (0,1)$ | Confidence level. |
| $\alpha := 1 - \mathrm{CL}$ | Tail probability corresponding to CL. |
| $q_\alpha(\tau_k)$ | $\alpha$-quantile of the simulated $H$-day PnL distribution at date $\tau_k$. |
| $\widehat{\mathrm{VaR}}_{\tau_k}^{(H)}(\alpha) := -q_\alpha(\tau_k) > 0$ | VaR forecast, treated as a positive loss magnitude. |
| $I_{\tau_k}(\alpha)$ | VaR exceedance (violation) indicator at date $\tau_k$. |
| $N$ | Number of out-of-sample (non-overlapping) observations in the backtest for a given $H$. |
| $n_1 = \sum_{k=1}^{N} I_{\tau_k}(\alpha)$ | Number of exceedances; $\hat{\pi} = n_1/N$ is the observed exceedance frequency. |
| $\mathrm{LR}_{\mathrm{UC}}$ | Kupiec unconditional coverage likelihood-ratio statistic. |
| $n_{ij}$ | Counts of transitions $I_{\tau_{k-1}} = i \to I_{\tau_k} = j$, $i, j \in \{0,1\}$. |
| $\mathrm{LR}_{\mathrm{IND}}$ | Christoffersen independence likelihood-ratio statistic. |
| $\mathrm{LR}_{\mathrm{CC}} = \mathrm{LR}_{\mathrm{UC}} + \mathrm{LR}_{\mathrm{IND}}$ | Christoffersen conditional coverage likelihood-ratio statistic. |
| $\mathrm{Hit}_{\tau_k}(\alpha) = I_{\tau_k}(\alpha) - \alpha$ | De-meaned hit process used in the Dynamic Quantile (DQ) test. |
| $L$ | Number of lags in the DQ regression (paper uses $L = 4$). |
| $\delta_0, \ldots, \delta_{L+1}$ | Coefficients in the DQ regression specification. |
| DQ | Engle–Manganelli Dynamic Quantile Wald statistic. |
| $\gamma$ | Significance level (paper uses $\gamma = 5\%$). |
| CRGS | Composite Risk Governance Score aggregating backtesting diagnostics into a ranking signal. |
| $p_{\mathrm{UC}}, p_{\mathrm{CC}}, p_{\mathrm{DQ}}$ | $p$-values for UC, CC, and DQ tests. |
| $s_{\mathrm{UC}}, s_{\mathrm{CC}}, s_{\mathrm{DQ}}$ | Component scores combining pass/fail and margin-to-rejection: $s = \mathbb{I}\{p > \gamma\} + (p - \gamma)$. |
| $w_{\mathrm{UC}}, w_{\mathrm{CC}}$ | CRGS weights (paper uses $w_{\mathrm{UC}} = 0.6$, $w_{\mathrm{CC}} = 0.4$). |
| $b_{\mathrm{DQ}}$ | DQ bonus weight when included (paper uses $b_{\mathrm{DQ}} = 0.1$ with inclusion rule $n_1 \geq 20$). |
| $S_H^{(j)} = \sum_{k=1}^{H} r_k^{(j)}$ | Generic cumulative $H$-day shock for factor $j$ (model-generated). |
| $\ell$ | Block length in C-MBB (paper uses $\ell = 50$). |
| $N_{\mathrm{obs}}$ | Number of bootstrap replications for the empirical benchmark (paper uses 100,000). |
| $Q_{\mathrm{emp}}^{(j,H)}(p)$ | Empirical quantile function of $S_H^{(j)}$ from the C-MBB benchmark. |
| $Q_{\mathrm{mod}}^{(j,H,q)}(p)$ | Model-implied quantile function at horizon $H$ and initial-volatility regime $q$. |
| $\mathcal{Q}$ | Set of initial-volatility quantiles defining regimes (paper uses $q \in \{75\%, 80\%, 85\%, 90\%, 95\%, 100\%\}$). |
| $\sigma_0^{(j)}(q)$ | Initial volatility state for factor $j$ anchored at quantile $q$ of fitted $\hat{\sigma}_t^{(j)}$. |

*Continued on next page.*

| Symbol | Meaning |
|---|---|
| $\alpha \in (0, 1/2)$ (tail mass) | Tail mass defining left and right tail regions for TRGS evaluation. |
| $p_{\min}$ | Extreme-probability floor for tail grids (paper uses $p_{\min} = 10^{-5}$). |
| $\mathcal{P}_L \subset [p_{\min}, \alpha]$ | Left-tail probability grid used for tail error evaluation. |
| $\mathcal{P}_R \subset [1-\alpha, 1-p_{\min}]$ | Right-tail probability grid used for tail error evaluation. |
| $\mathrm{MSE}_L^{(j,H,q)}$, $\mathrm{MSE}_R^{(j,H,q)}$ | Tail mean-squared errors between model and empirical quantiles in left and right tails. |
| $\mathrm{RMSE}_L^{(j,H,q)}$, $\mathrm{RMSE}_R^{(j,H,q)}$ | Root mean-squared errors in left and right tails (units: bps). |
| $\mathrm{TRGS}^{(j,H,q)}(w_L, w_R)$ | Tail Risk Governance Score combining left/right tail MSE with weights $w_L, w_R$. |
| $w_L, w_R$ | Tail weights (paper sets $w_L = 0.3$, $w_R = 0.7$). |