Iterated Forecast Combination for Joint VaR–ES Prediction: Stabilizing Performance under Rule-Selection Uncertainty

Burak Alparslan Eroğlu* & Haluk Yener† & Taner Yiğit‡

April 2026

Abstract

The proliferation of methods for forecasting Value-at-Risk (VaR) and Expected Shortfall (ES) motivates forecast combination, which may itself leave practitioners with substantial rule-selection uncertainty. A Monte Carlo study comprising eight scenarios isolates key sources of instability, including smooth drift, structural breaks, state-dependent specialists, and catastrophic failures. Across 96,000 replications, no single method dominates universally; the best-performing rule varies with the scenario, panel size, and tail probability level. Simple-average iterated combination (IC-SA) emerges as a stabilizer: it compresses the upper tail of regret—relative shortfall from the ex post best feasible rule under the Fissler–Ziegel (FZ) loss—rather than acting as a universal champion. We validate these findings empirically using eighty-one individual VaR–ES methods combined by six standard and three iterated (second-stage) rules across five major equity indices at three tail probability levels. IC-SA delivers the strongest aggregate loss-based performance and reduces case-level 95th-percentile regret in the vast majority of cases, while ranking-based leadership remains with simpler stage-1 rules. The evidence supports interpreting iterated combination not as a universal forecast improver, but as a device for limiting the severity of costly rule-selection mistakes.

*Department of Economics, Bakırçay University, 35665, Menemen, Izmir, TURKEY; burak.eroglu@bakircay.edu.tr.

†*Corresponding Author*: International Finance Department, Özyeğin University, Çekmeköy Campus, 34794, Istanbul, TURKEY; haluk.yener@ozyegin.edu.tr.

‡Department of Economics, Bilkent University, 06533, Bilkent, Ankara, TURKEY; tyigit@bilkent.edu.tr.

JEL Classification: C1, C22, C53, G11, G12, G14.
Keywords: Value-at-Risk; Expected Shortfall; Iterated Combination; Forecast Combination; Regret Diagnostics; Rule-Selection Uncertainty.

1 Introduction

In financial risk management, value at risk (VaR) and expected shortfall (ES) are two central risk measures for estimating potential losses in financial positions. Both VaR and ES have been at the focus of financial regulatory agencies because losses may amount to a level that threatens the existence of the financial system, as witnessed during the 2008/2009 global financial crisis; see, for example, the technical report of Basel Committee on Banking Supervision (BCBS, 2016). ES has gained further prominence as the risk measure of choice because VaR is an incoherent risk measure (Artzner et al., 1999): it does not account for the benefits of diversification and therefore suffers from the lack of sub-additivity. Thus, relying solely on VaR may be misleading, despite it providing a quick snapshot regarding the risk of a financial position.

Apart from the regulatory focus, both risk measures have been scrutinized in the academic literature with the development of new methodologies under empirical settings. There is a plethora of research on the topic with various parametric assumptions (parametric, semiparametric, nonparametric), different distributions (normal, Skewed-t, GED, etc.), and dynamic specifications, measured with different scoring functions and calibration criteria. Under all approaches, the calculation of VaR is easier than that of ES as it entails finding the quantile of returns at the lower tail of their distributions, usually corresponding to 1%, 2.5%, or 5% level. While such computations may look straightforward, the selection of an appropriate method is complicated by the concept of elicitability (Gneiting, 2011): VaR is elicitable, meaning that a consistent scoring function for VaR exists, but ES is not elicitable on its own. Nonetheless, as shown by Fissler and Ziegel (2016), ES is jointly elicitable with VaR, and with the use of a strictly consistent joint scoring function one can rank risk measure forecasts obtained from different models.

Given the wide array of available models, finding a unique individual specification that provides the best performance across different scenarios is difficult. To address this issue, a natural approach is to combine forecasts, thereby benefiting

from a pool of models that capture different stylized aspects of financial time series (Bates and Granger, 1969; Stock and Watson, 2004; Timmermann, 2006). In the risk forecasting context, Taylor (2020) applies this logic to joint VaR–ES prediction, showing that combination rules consistently outperform individual models. However, the same uncertainty that applies to the selection of individual models may also apply to the selection of the combination rule itself. Multiple standard combination methods exist—simple average, median, trimmed and winsorized means, and score-based weighting schemes—and there is no guarantee that any single combination rule will dominate across all environments.

To address this second layer of uncertainty, we adopt an iterated combination approach, which we denote by IC, extending the framework of Lin et al. (2018) to VaR–ES prediction. Lin et al. (2018) propose combining a large set of individual predictors in a first stage and then iterating that combination at a second stage to improve corporate bond return forecasts; the method has since been extended to exchange rate forecasting (Alexandridis et al., 2024). We are the first to apply this iterated combination framework to joint VaR–ES prediction. In our study, we apply a richer set of combination techniques at the first stage and then combine the resulting stage-1 forecasts at a second stage. The design is fully recursive: at each forecast origin, both stage-1 and stage-2 weights are estimated using only the information available up to that date, and forecasts are evaluated one step ahead out of sample.

Our study proceeds in two main steps. In the first step, we develop a Monte Carlo framework with eight distinct forecaster scenarios that isolate the instability sources most relevant for empirical forecast combination—smooth drift, abrupt breaks, state-dependent specialists, catastrophic failures, ranking reshuffling, clustered specialists, and stale leaders with delayed adaptation. The simulation uses a controlled pseudo-forecast panel whose size, correlation structure, and cross-sectional heterogeneity are calibrated to reflect the properties of a large and diverse pool of individual models, mirroring the empirical exercise that follows. In both settings, the forecast inputs are combined using six standard methods—sample average (S-SA), median (S-med), trimmed mean (S-tSA), winsorized mean (S-wSA), relative score combining (S-RSC), and minimum score combining (S-MSC)—to form six stage-1 forecasts, from which three iterated combinations are constructed: IC-SA, IC-med, and IC-MSC, yielding nine combination methods in total. Across 96,000 replications, we document three findings: (i) no single combination rule dominates uniformly across scenarios and

design dimensions; (ii) the identity of the best method shifts with the environment, confirming that rule-selection uncertainty is a substantive problem; and (iii) iterated combination, particularly IC-SA, is most attractive when evaluated through regret diagnostics derived from the realized FZ loss, because it systematically limits upper-tail relative shortfall from the ex post best feasible rule even when it does not attain the lowest average loss.

In the second step, we apply the same combination architecture to an out-of-sample empirical exercise. The individual forecast panel comprises eighty-one VaR–ES models spanning non-parametric, parametric, and semi-parametric families, estimated recursively across five major equity indices (CAC 40, DAX 40, FTSE 100, NIKKEI 225, S&P 500) at three tail probability levels ($\alpha \in \{0.01, 0.025, 0.05\}$), yielding fifteen empirical cases in total. We find evidence consistent with the simulation findings. Performance separates along two distinct dimensions: on loss-based criteria, IC-SA leads the aggregate ranking and most effectively limits upper-tail regret; on rank-based criteria, simpler stage-1 rules such as S-med prove more persistently competitive. No method dominates both dimensions simultaneously, confirming that rule-selection uncertainty is an empirical reality rather than a simulation artifact. This divergence is not a contradiction: mean loss rewards consistent small improvements across all cases, while rank-based criteria reward large wins in selected cases, and these two notions of performance need not agree. The data therefore support the interpretation that iterated combination acts as a stabilizer—limiting the severity of rule-selection shortfall—rather than as a universal performance champion.

These findings carry a clear practical message. Rather than searching for a universally superior combination rule, practitioners can reduce rule-selection risk by adding a lightweight second-stage averaging layer. The cost of this additional step is minimal, while its empirical value lies in limiting the upper-tail cost of selecting a rule that turns out not to be the ex post best performer. The evidence rests on a combination of regret-based diagnostics, pairwise equal predictive ability comparisons aggregated into DM–Copeland scores, and joint VaR–ES backtests covering both standard VaR calibration (UC, CC, DQ) and ES-specific procedures (NZ, MF, BD), together providing a multi-dimensional picture of combination method performance that goes beyond average loss rankings alone.

The remainder of the paper is organized as follows. Section 2 describes the model set, score functions, backtesting procedures, and the standard and iterated combina-

tion methods. Section 3 presents the Monte Carlo design and simulation evidence. Section 4 reports the empirical application. Section 5 provides a discussion linking the simulation and empirical findings, and Section 6 concludes. Detailed mathematical descriptions of the individual models are provided in Section S1.5 of the Online Supplement. Additional simulation robustness results and supplementary empirical tables are collected in the Online Supplement.

2 Methodology

2.1 Risk Measures and Score Function

We study the one-step-ahead VaR and ES of financial returns $r_t = \log(P_t/P_{t-1}) \times 100$, where P_t is the closing price of an equity index at date t. At tail probability $\alpha \in (0,1)$, these are defined as

$$\begin{aligned} \mathrm{VaR}_{t+1}^{\alpha} &:= \inf\{r_{t+1} \in \mathbb{R} \mid F(r_{t+1} \mid \mathcal{F}_t) \geq \alpha\}; \\ \mathrm{ES}_{t+1}^{\alpha} &:= \mathbb{E}\left[r_{t+1} \mid r_{t+1} \leq \mathrm{VaR}_{t+1}^{\alpha}, \mathcal{F}_t\right], \end{aligned} \tag{2.1}$$

where $\mathcal{F}_t$ is the information set at time t and $F(\cdot \mid \mathcal{F}_t)$ is the conditional cumulative distribution function. We have $\mathrm{ES}_{t+1}^{\alpha} < \mathrm{VaR}_{t+1}^{\alpha} < 0$.

For the evaluation of forecasting methods, we employ a strictly consistent joint scoring rule. While VaR is elicitable under the quantile loss function (Gneiting and Raftery, 2007; Gneiting, 2011), ES is not elicitable on its own. Nonetheless, as shown by Fissler and Ziegel (2016), ES is jointly elicitable with VaR. We adopt the Fissler–Ziegel (FZ) loss function as our primary evaluation criterion. For a realization r_t, forecasted VaR v_t and ES ξ_t, the loss at tail probability α is

$$\ell_{\alpha}(r_t, v_t, \xi_t) = -\frac{1}{\alpha \xi_t} \mathbf{1}\{r_t \leq v_t\}(v_t - r_t) + \frac{v_t}{\xi_t} + \log(-\xi_t) - 1, \tag{2.2}$$

with the tail restrictions $v_t < 0$ and $\xi_t < v_t$.

Remark 2.1. *The general class of strictly consistent scoring functions for the pair (VaR, ES) developed in Fissler and Ziegel (2016) admits multiple specifications, including the FZG specification of Fissler et al. (2016) and the AL (asymmetric Laplace) score function of Taylor (2019). In additional robustness checks, available upon re-*

quest, evaluating our forecasts under these alternative consistent scoring rules yields qualitatively identical conclusions. We therefore report results based on the FZ loss in (2.2) *throughout for brevity.*

2.2 Model Classes

The individual VaR–ES forecasts are drawn from three broad families spanning non-parametric, parametric, and semi-parametric approaches. Mathematical details are provided in Section S1.5 of the Online Supplement; here we give a brief summary.

The first class consists of the *basic models* (BM): Historical Simulation (HS), Weighted Historical Simulation (WHS), Filtered Historical Simulation (FHS), and RiskMetrics (RM). We estimate HS and FHS with rolling windows of 250 and 500 days, and WHS with weight parameters 0.95 and 0.99. RM assumes a Normal distribution. In total, this class contains *9 specifications*.

The second class comprises the *parametric models*. These specifications depend on both the model structure and the assumed distribution. We estimate them by maximum likelihood (MLE) under the Normal, Skewed-t, and GED distributions. This class has two sub-groups. The first is the *GARCH universe* (GARCH), which contains sixteen GARCH-type volatility models (see Table S2 of the Online Supplement), including GARCH, IGARCH, EGARCH, GJR-GARCH, APARCH, FIGARCH, MS-GARCH, and Robust Bootstrap GARCH. Estimating each model under three distributions yields *48 specifications*. The second is the *HAR-Range* model. It applies the heterogeneous autoregressive realized-volatility specification of Corsi (2009) to intraday range data, following Taylor (2020). Estimating it under three distributions yields *3 specifications*.

The third class encompasses the *semi-parametric models*. These specifications remain model-dependent but are distribution-free. This class has three sub-groups. The first, *FZ-GARCH*, takes the GARCH-type models from the parametric class, excluding MSGARCH and Rob-GARCH-boot, and estimates them by minimizing the FZ loss function of Patton et al. (2019) rather than by MLE, yielding *14 specifications*. The second, *FZ-GAS*, includes two specifications of the one-factor GAS model of Patton et al. (2019). The third, *Asymmetric Laplace* (AsL), contains four CaViaR specifications of Engle and Manganelli (2004) estimated under the asymmetric Laplace distribution of Taylor (2019). Each is paired with simple and dynamic

ES estimation, yielding *8 specifications*; 3 are later dropped because of convergence issues, leaving *5* in practice.

In sum, we initially considered eighty-four individual specifications. Because a subset failed to generate stable recursive forecasts due to numerical issues, the final evaluation is conducted using 81 individual methods. We provide a summary in Table 1.

Table 1: Number of Specifications per Model Class

		Parametric		**Semi-parametric**			
	BM	**GARCH**	**HAR-Range**	**FZ-GARCH**	**FZ-GAS**	**AsL**	**Total**
# of Models	9	16× 3	3	14	2	8	84 (81 estimated)
Distributions	-	Normal, Skewed-t, GED		-	-	-	3

2.3 Backtesting

We employ backtesting procedures to measure the calibration quality of the VaR and ES forecasts; for background see Christoffersen (2010) and Kuester et al. (2006).

The VaR backtests use the demeaned hit sequence $Hit_t = \mathbf{1}\{r_t \leq v_t\} - \alpha$ and comprise three tests. The Unconditional Coverage (UC) test of Christoffersen (1998) checks whether the unconditional exceedance rate equals the nominal level α. The Conditional Coverage (CC) test of Christoffersen (1998) additionally tests whether violations are serially independent. The Dynamic Quantile (DQ) test of Engle and Manganelli (2004) checks whether the demeaned hit process is unpredictable given lagged information and the contemporaneous VaR forecast.

To complement these VaR-only procedures, we also report three joint VaR–ES backtests that explicitly target the calibration of expected shortfall. The McNeil and Frey (2000) (MF) test examines whether the mean exceedance residual $r_t - \widehat{ES}_t$, computed over all dates where $r_t \leq \widehat{VaR}_t$, equals zero under a two-sided test; the null distribution is approximated by a stationary bootstrap (Politis and Romano, 1994) with geometric mean block length $b = \lceil T^{1/3} \rceil$ and 500 replications. The Nolde and Ziegel (2017) (NZ) test uses the moment condition $\mathbb{E}[Z_t] = 0$, where $Z_t = \alpha^{-1}\mathbf{1}\{r_t \leq \widehat{VaR}_t\}r_t - \widehat{ES}_t$, which follows directly from joint elicitability of (VaR, ES) (Fissler and Ziegel, 2016); the test statistic is a heteroskedasticity and autocorrelation consistent (HAC) t-ratio with Newey–West variance (Newey and West, 1987). The Bayer and Dimitriadis (2022) (BD) test jointly tests the bivariate moment condition $\mathbb{E}[\Psi_t] = 0$,

where $\Psi_t = \left(\mathbf{1}\{r_t \leq \widehat{VaR}_t\} - \alpha,\ \alpha^{-1}\mathbf{1}\{r_t \leq \widehat{VaR}_t\}r_t - \widehat{ES}_t\right)^\top$, via a HAC Wald statistic that follows a $\chi^2(2)$ distribution asymptotically. Detailed case-by-case results for all six backtests are reported in the tables in Sections S4.1 and S4.2 of the Online Supplement.

2.4 Standard Combination Methods

We now describe the six standard combination rules used to combine the VaR and ES forecasts from the set of individual models described in Section 2.2. Let $\mathrm{M} \in \{\mathrm{VaR}, \mathrm{ES}\}$ and let $\{\widehat{\mathrm{M}}^i_{t+1}\}_{i=1}^n$ denote the one-step-ahead forecasts from the n individual models.

The first rule, **S-SA** (Sample Average), computes the equally weighted average of the individual forecasts:

$$\widehat{\mathrm{M}}^{\text{S-SA}}_{t+1} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\mathrm{M}}^i_{t+1}. \tag{2.3}$$

The second rule, **S-tSA** (Trimmed Mean), removes the lowest and highest q-fraction of forecasts (with $q = 0.10$) and averages the remainder:

$$\widehat{\mathrm{M}}^{\text{S-tSA}}_{t+1} = \frac{1}{n - 2\lfloor qn \rfloor} \sum_{i=\lfloor qn \rfloor + 1}^{n-\lfloor qn \rfloor} \widehat{\mathrm{M}}^{(i)}_{t+1}, \tag{2.4}$$

where $\widehat{\mathrm{M}}^{(i)}_{t+1}$ denotes the i-th order statistic of the cross-section of forecasts.

The third rule, **S-wSA** (Winsorized Mean), replaces forecasts below the q-th and above the $(1-q)$-th quantile of the cross-section with the corresponding boundary values (with $q = 0.10$), and then averages:

$$\widehat{\mathrm{M}}^{\text{S-wSA}}_{t+1} = \frac{1}{n}\left(\sum_{i=\lfloor qn \rfloor + 1}^{n-\lfloor qn \rfloor} \widehat{\mathrm{M}}^{(i)}_{t+1} + \lfloor qn \rfloor \left[\widehat{\mathrm{M}}^{(q)}_{t+1} + \widehat{\mathrm{M}}^{(1-q)}_{t+1}\right]\right). \tag{2.5}$$

The fourth rule, **S-med** (Median), takes the median of the cross-section:

$$\widehat{\mathrm{M}}^{\text{S-med}}_{t+1} = \text{median}\left\{\widehat{\mathrm{M}}^1_{t+1}, \ldots, \widehat{\mathrm{M}}^n_{t+1}\right\}. \tag{2.6}$$

The fifth rule, **S-MSC** (Minimum Score Combining), follows Taylor (2020) and chooses combination weights $\mathbf{w} = (w_1, \ldots, w_n)'$ by minimizing the in-sample FZ loss

of the combined forecast, subject to $w_i \geq 0$ and $\sum_i w_i = 1$:

$$\begin{aligned} \widehat{\mathrm{VaR}}_{t+1}^{\text{S-MSC}} &= \sum_{i=1}^{n} \mathrm{w}_i \widehat{\mathrm{VaR}}_{t+1}^{i}; \\ \widehat{\mathrm{ES}}_{t+1}^{\text{S-MSC}} &= \widehat{\mathrm{VaR}}_{t+1}^{\text{S-MSC}} + \sum_{i=1}^{n} \mathrm{w}_i \left(\widehat{\mathrm{ES}}_{t+1}^{i} - \widehat{\mathrm{VaR}}_{t+1}^{i} \right). \end{aligned} \tag{2.7}$$

The second equation prevents the crossing of VaR by ES. Weights are updated recursively in a rolling-window fashion.

The sixth rule, **S-RSC** (Relative Score Combining), is also from Taylor (2020) and assigns softmax weights based on recent scores accumulated over a rolling window $\mathcal{W}_{1,t}$ of the most recent L_1 observations:

$$\mathrm{w}_{i,t} = \frac{\exp\left(-\lambda \sum_{s \in \mathcal{W}_{1,t}} \ell_\alpha(r_s, \widehat{\mathrm{VaR}}_s^{i}, \widehat{\mathrm{ES}}_s^{i})\right)}{\sum_{j=1}^{n} \exp\left(-\lambda \sum_{s \in \mathcal{W}_{1,t}} \ell_\alpha(r_s, \widehat{\mathrm{VaR}}_s^{j}, \widehat{\mathrm{ES}}_s^{j})\right)}, \tag{2.8}$$

where ℓ_α is the FZ loss defined in (2.2), $\lambda > 0$ is a tuning parameter selected from a finite grid by recent historical performance, and $\mathcal{W}_{1,t}$ rolls forward at each forecast origin so that only past information enters the weights. As $\lambda \downarrow 0$, S-RSC converges to S-SA; as λ grows, it concentrates weight on the historically best model. The window length L_1 is specified in Section 4.

In sum, we have six standard (stage-1) combination methods: S-SA, S-med, S-tSA, S-wSA, S-RSC, and S-MSC.

2.5 Iterated Combination (IC)

To address the uncertainty over which standard combination rule to select, we introduce an iterated combination (IC) approach, extending the framework of Lin et al. (2018) to joint VaR–ES forecasting.

The application is straightforward. At each forecast origin t, we first produce the six stage-1 combination forecasts described in Section 2.4. We define the set of stage-1 combinations as $\mathcal{C} := \{\text{S-SA}, \text{S-tSA}, \text{S-wSA}, \text{S-med}, \text{S-MSC}, \text{S-RSC}\}$. At the second stage, we combine these six stage-1 outputs using three combination rules. The first,

IC-SA, is the simple average of the six stage-1 forecasts:

$$\widehat{\mathrm{VaR}}_{t+1}^{\text{IC-SA}} = \frac{1}{6}\sum_{k\in\mathcal{C}} \widehat{\mathrm{VaR}}_{t+1}^{k}; \qquad \widehat{\mathrm{ES}}_{t+1}^{\text{IC-SA}} = \frac{1}{6}\sum_{k\in\mathcal{C}} \widehat{\mathrm{ES}}_{t+1}^{k}. \tag{2.9}$$

Because all six stage-1 inputs are constructed from the same underlying pool of 81 individual-model forecasts, the second-stage panel is best viewed as a set of overlapping summaries rather than as six independent forecast sources. Accordingly, any gain from IC-SA should be interpreted as regularization across alternative first-stage combination rules, not as diversification across unrelated information sets.

The second, IC-med, takes the median of the six stage-1 forecasts. The third, IC-MSC, applies minimum score combining to the six stage-1 time series, with weights re-estimated on a rolling window.

In total, the study evaluates nine combination methods: six standard (stage-1) and three iterated (stage-2). Table 2 provides a compact summary.

Table 2: Summary of Combination Methods

Stage	**Label**	**Description**
Stage 1 (Standard)	S-SA	Sample average of 81 individual forecasts
	S-med	Median of 81 individual forecasts
	S-tSA	10% trimmed mean
	S-wSA	10% winsorized mean
	S-RSC	Relative score combining
	S-MSC	Minimum score combining
Stage 2 (Iterated)	IC-SA	Simple average of 6 stage-1 forecasts
	IC-med	Median of 6 stage-1 forecasts
	IC-MSC	Min. score combining of 6 stage-1 forecasts

3 Simulation Study

We conduct a Monte Carlo simulation study to identify the conditions under which iterated combination is most useful. The key advantage of a simulation study is that we can systematically vary the sources of instability while knowing the true data-generating process, which allows us to isolate the genuine contribution of iterated combination from the noise inherent in any single empirical dataset.

The simulation is designed around the instability patterns most relevant for em-

pirical forecast combination. Specifically, we consider environments with strong cross-model dependence, time-varying relative predictive ability, regime shifts, and score-based rules that are slow to adapt after structural change. Throughout, the exercise is fully recursive: at each forecast origin t, combination weights at both stages are estimated using only observations available up to time $t-1$, and the resulting forecasts are evaluated one period ahead out of sample. This design replicates the real-time information constraints that practitioners face.

3.1 Design

We generate returns from the following GARCH$(1,1)$ model with Student-t innovations:

$$r_t = \sqrt{h_t}\, z_t, \quad h_t = \omega + \alpha_g r_{t-1}^2 + \beta h_{t-1}, \tag{3.1}$$

where z_t follows a standardized Student-t distribution with ν degrees of freedom. We set $\omega = 0.02$, $\alpha_g = 0.06$, $\beta = 0.92$, and $\nu = 8$. The selected values generate returns with moderately heavier tails than a Gaussian benchmark (consistent with equity index returns), a moderate response to recent shocks, and high volatility persistence. Thus, the model selected for data generation captures some stylized facts such as zero mean, volatility clustering and heavy tails observed in financial returns. Our aim with this selection is to remain simplistic so as to attribute differences in combination-method performance to scenario design rather than to the complexity of the data generating process.

When running the simulations, we initialize the process at its unconditional variance, $h_0 = \omega/(1-\alpha_g-\beta) = 1$. Moreover, to eliminate dependence on this initialization, we discard the first 500 observations and do our estimations and evaluations after the discarded period.

Next, we denote the true conditional VaR and ES with

$$v_{t,\alpha}^{\star} = \sigma_t c_{\alpha,\nu}, \qquad e_{t,\alpha}^{\star} = \sigma_t d_{\alpha,\nu}, \tag{3.2}$$

where $\alpha \in \{0.01, 0.025\}$ is the left-tail probability level, $\sigma_t = \sqrt{h_t}\sqrt{(\nu-2)/\nu}$ is the standardized conditional standard deviation, $c_{\alpha,\nu}$ is the α-quantile of the Student-t distribution, and $d_{\alpha,\nu} = -f_\nu(c_{\alpha,\nu})(\nu + c_{\alpha,\nu}^2)/[\alpha(\nu-1)]$ is the ES scale factor (see Patton

et al. (2019)). The simulation covers the two regulatory tail levels most relevant for the Basel framework; the 5% level is included in the empirical exercise but is less central to regulatory capital applications and is excluded here to keep the design tractable.

In each simulation run, we generate $K \in \{20, 50, 80\}$ individual pseudo-forecasts constructed as the sum of the true values in (3.2) and noise. That is, the baseline pseudo-forecasts are $\widehat{v}_{k,t} = v^{\star}_{t,\alpha} + s_{v,t}u^{(v)}_{k,t}$ and $\widehat{e}_{k,t} = e^{\star}_{t,\alpha} + s_{e,t}u^{(e)}_{k,t}$, with time-varying noise scales $s_{v,t} = 0.10\sqrt{h_t}$ and $s_{e,t} = 0.12\sqrt{h_t}$. These scales co-move with the volatility model in (3.1), so that forecast dispersion increases during turbulent periods in a manner consistent with empirical estimation uncertainty. The ES noise coefficient is set slightly higher than the VaR coefficient because ES, as an average of tail losses beyond the quantile, is inherently harder to estimate precisely.

To mimic the cross-model dependence observed in a large pool of models, we employ a one (common)-factor structure with correlation parameter $\rho \in \{0.95, 0.98\}$,

$$u^{(v)}_{k,t} = \sqrt{\rho}\, c^{(v)}_t + \sqrt{1-\rho}\, \varepsilon^{(v)}_{k,t}, \qquad u^{(e)}_{k,t} = \sqrt{\rho}\, c^{(e)}_t + \sqrt{1-\rho}\, \varepsilon^{(e)}_{k,t}, \tag{3.3}$$

where $c^{(v)}_t$ and $c^{(e)}_t$ are common Gaussian shocks, $\varepsilon^{(v)}_{k,t}$ and $\varepsilon^{(e)}_{k,t}$ are idiosyncratic Gaussian shocks. After scenario-specific distortions described below, we enforce the tail-consistency restrictions $\widehat{v}_{k,t} < 0$ and $\widehat{e}_{k,t} < \widehat{v}_{k,t}$ for all k and t.

The selection of baseline coefficient values involves some degree of judgment. However, our central findings regarding the relative performance of combination methods are identified through the eight scenario designs $S0$–$S7$ rather than by the specific baseline parameterization. Each scenario introduces a qualitatively distinct and empirically relevant source of instability whose distortions grow increasingly large relative to the baseline noise. Specifically, beginning from the stable benchmark $S0$, the degree of instability increases progressively across scenarios until $S7$, which represents the most challenging environment. Through this design, cross-method performance variation is driven by the scenario structure rather than by the chosen baseline values. The key findings are subsequently validated through the empirical exercise, confirming that the conclusions do not depend on the specific simulation parameter values.

Table 3 summarizes the eight scenarios:

For each simulation run, the sample contains $T = 2,550$ observations split into four periods: a 500-observation burn-in period discarded as described above, a 1250-

Table 3: Simulation Scenario Descriptions

Scenario	Name	Description
$S0$	Stable benchmark	No additional distortion. All models are equally well calibrated on average. Serves as a reference case that tests whether combination methods add unnecessary noise in a benign environment.
$S1$	Smooth drift	Slowly varying sinusoidal bias with model-specific phases. Some models gradually improve while others deteriorate, mimicking the slow evolution of relative model quality over time.
$S2$	Abrupt break	A one-time structural break at 35% of the sample, after which noise increases and a post-break bias is introduced. Tests whether methods can adapt quickly to a new regime.
$S3$	State-dependent specialists	Models specialize in either calm or stressed markets, so that no single model is best in both states. The optimal combination rule must balance performance across regimes.
$S4$	Temporary catastrophic failures	A subset of models experiences rare but severe forecast failures at random times. Tests robustness to occasional extreme mis-predictions.
$S5$	Ranking instability	Recurrent reshuffling of model rankings, where the relative quality of individual models changes frequently and unpredictably.
$S6$	Clustered specialists	Three model clusters take turns performing best. Tests whether combination rules can handle rotating group advantage.
$S7$	Stale leaders	Initially strong models gradually deteriorate after a break while previously mediocre models slowly improve, mimicking delayed adaptation to changing market conditions.

observation pre-training period, a 500-observation second-stage (meta-) training period, and a 300-observation out-of-sample evaluation period. Stage-1 weights are estimated over a rolling window of $L_1 = 1,000$ observations, while stage-2 weights are estimated over a shorter rolling window of $L_2 = 500$ using stage-1 observations. The six stage-1 standard forecast combinations (S-SA, S-med, S-tSA, S-wSA, S-RSC, and S-MSC) and three stage-2 iterated forecast combinations (IC-SA, IC-med, and IC-MSC) are described in Section 2.4 and Section 2.5, respectively.

For forecast evaluation, we use the Fissler–Ziegel (FZ) loss in (2.2) as the primary scoring criterion and complement it with regret-, rank-, and tournament-based summaries. The role of these additional metrics is diagnostic rather than foundational: they do not replace the FZ loss, but characterize different aspects of method behavior under the realized FZ loss outcomes.

Let $\mathcal{M}$ denote the set of feasible non-oracle combination rules considered in the simulation study. For each simulation replication r and method $m \in \mathcal{M}$, let $L_{m,r}$ denote the out-of-sample average FZ loss of method m in replication r. We define replication-level regret as $\text{Regret}_{m,r} := L_{m,r} - \min_{j \in \mathcal{M}} L_{j,r} \geq 0$. Thus, regret is an *ex post opportunity-loss* measure: it records how far method m falls short of the best feasible rule in the same replication under the realized FZ loss. In this sense, regret

is computed from the primary scoring rule rather than serving as an alternative to it. This usage is distinct from the minimax-regret criterion in statistical decision theory, where regret is minimized over worst-case distributions; here, regret is simply a diagnostic quantity derived from realized FZ outcomes.

We summarize the regret distribution by its mean and upper quantiles, with particular emphasis on $p95$ (95th-percentile) regret. Mean regret measures the average relative shortfall from the ex post best feasible rule, whereas $p95$ regret measures the upper-tail severity of that shortfall. Accordingly, mean FZ loss and $p95$ regret answer different questions: the former concerns absolute average predictive performance, while the latter concerns the costliest relative underperformance realizations induced by rule-selection uncertainty.

To complement the regret evidence, we also report two rank-based and two pairwise tournament summaries. The first rank-based measure is average rank across simulation replications, where rank 1 denotes the lowest FZ loss. The second is top-3 probability, namely the frequency with which a method finishes among the three best performers. The first tournament measure is the raw Copeland score (Copeland, 1951), under which a method receives $+1$ for each rival it defeats in a majority of replications within the relevant design cell, -1 for each majority loss, and 0 for a tie. The second is an inference-adjusted sign-Copeland score, constructed in the same way except that a pairwise win or loss is counted only when the majority outcome is statistically significant under a two-sided exact binomial sign test. These summaries distinguish ordinal prominence and broad pairwise dominance from relative-loss control.

The metrics above need not agree in finite samples, and this disagreement is substantively informative rather than contradictory. A method may achieve strong average FZ performance without offering the best upper-tail regret control, while another method may rarely rank first yet still avoid severe relative shortfall. The paper therefore intentionally distinguishes among three notions of success: *average predictive accuracy* (mean FZ loss), *relative shortfall control* (regret, especially its upper tail), and *ordinal or pairwise prominence* (rank-based and Copeland-type measures).

The full design grid consists of 8 scenarios $\times$ 3 panel sizes $\times$ 2 correlation levels $\times$ 2 tail probabilities $\times$ 1,000 replications $= 96,000$ total replications.

3.2 Simulation Evidence

In this section, we summarize the main Monte Carlo findings using four complementary criteria: mean FZ loss, $p95$ regret, raw Copeland score, and inference-adjusted sign-Copeland score. Table 4 reports the scenario winners, while Figures 2–3 illustrate the corresponding patterns for selected methods across the eight scenarios. Mean FZ loss captures average predictive accuracy, $p95$ regret captures upper-tail relative shortfall, and Copeland-type measures capture broad pairwise dominance. Additional scenario-level results are reported in the Online Supplement.

Inspection of the first column of Table 4 immediately reveals that no single method wins everywhere. Under mean FZ loss, S-med performs best in the quiescent scenarios where ranking noise is limited ($S0$, $S3$, $S4$, $S7$), while S-SA leads in the scenarios featuring the most pronounced instability ($S1$, $S2$, $S5$), likely because simple averaging is intrinsically robust to temporary shifts in model quality; S-wSA leads only in $S6$, where rotating cluster advantage favors a winsorized approach. Importantly, this heterogeneity persists even when we further disaggregate by panel size K, correlation ρ, and tail level α (see the Online Supplement). The message is unambiguous: the best method on average depends on the environment, and no single rule dominates uniformly.

Table 4: Scenario-wise winner summary

Scenario	Mean FZ	$p95$ regret	Copeland	Sign-Copeland
S0	S-med	IC-SA	S-tSA	S-tSA
S1	S-SA	IC-SA	S-SA	S-SA
S2	S-SA	IC-SA	S-SA	S-SA / IC-med
S3	S-med	IC-SA	S-med	S-med
S4	S-med	IC-med	S-RSC	S-RSC
S5	S-SA	IC-SA	S-SA	S-SA / S-wSA
S6	S-wSA	IC-SA	S-wSA	S-wSA
S7	S-med	IC-SA	S-med	S-med / S-tSA

Note: For each scenario, the table reports the method with the lowest mean FZ loss, the lowest 95th-percentile regret, the highest raw Copeland score, and the highest inference-adjusted sign-Copeland score. Copeland winners are based on majority pairwise wins across pooled replications within scenario after averaging over the remaining design dimensions (K, ρ, α). Sign-Copeland applies the same tournament logic but counts a pairwise win or loss only when the majority outcome is statistically significant under a two-sided exact binomial sign test; ties are reported explicitly when multiple methods attain the same top score.

When attention shifts from average performance to upper-tail regret, the ordering changes materially. Table 4 shows that IC-SA attains the lowest $p95$ regret in seven

of the eight scenarios, with IC-med leading only in $S4$. Figure 1 confirms the same pattern visually. This does not imply that IC-SA is the mean-loss winner in every environment. It implies something more specific: among the feasible rules, IC-SA most consistently limits the upper-tail cost of selecting a method that turns out not to be the ex post best one. That is precisely the sense in which iterated combination acts as a stabilizer.

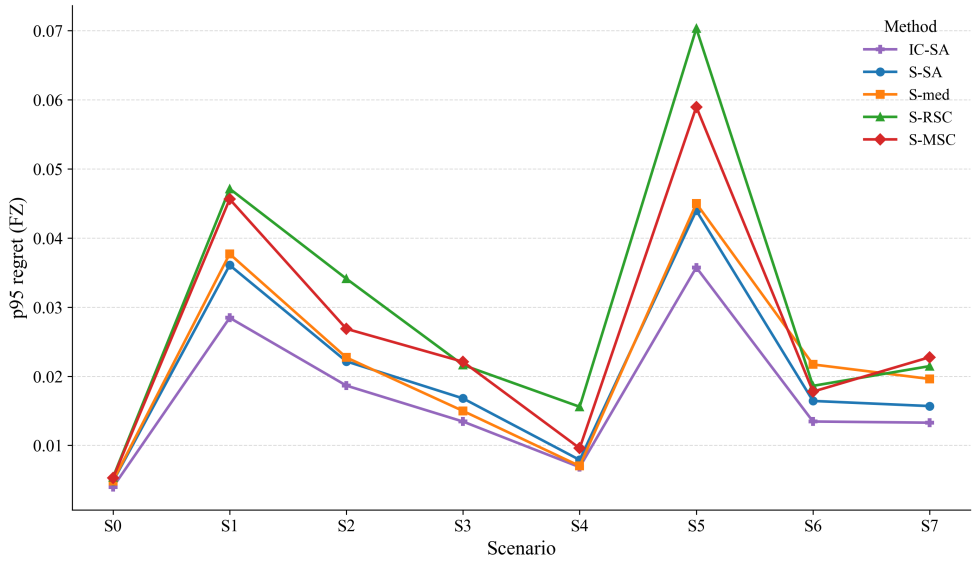


Figure 1: 95th percentile of replication-level FZ regret by scenario for selected methods

Notes: Lower values indicate better upper-tail regret control. The figure highlights that IC-SA is most competitive when performance is evaluated through upper-tail regret.

By contrast, leadership under mean-loss and tournament-style criteria remains dispersed. Figure 2 shows that average FZ loss leadership rotates across methods rather than concentrating on a single rule. Figure 3 delivers the same conclusion for raw Copeland rankings, where stage-1 rules such as S-SA, S-med, and S-wSA continue to dominate in different scenarios. The inference-adjusted sign-Copeland results reported in Table 4 sharpen this conclusion further: there is still no universal pairwise winner once statistical significance is imposed on majority victories.

Taken together, the simulation evidence supports a precise interpretation. No combination rule is a universal champion across all environments and criteria. However, iterated combination—and IC-SA in particular—systematically compresses the upper tail of the regret distribution, while leadership under mean loss and tourna-

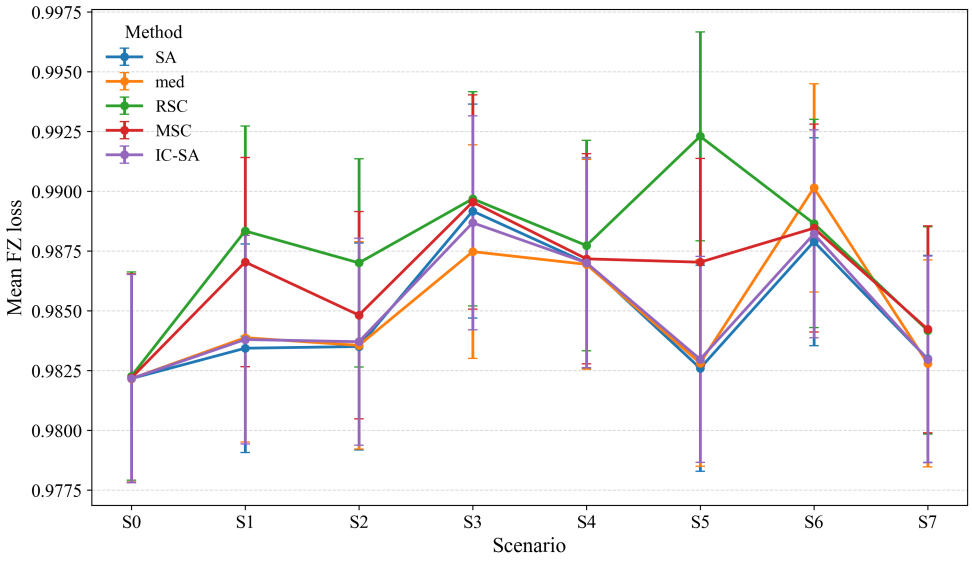


Figure 2: Mean FZ loss across scenarios for selected methods

Notes: Lower values indicate better performance. The figure summarizes the scenario-level mean FZ profiles for the selected combination rules, with intervals showing the associated sampling uncertainty.

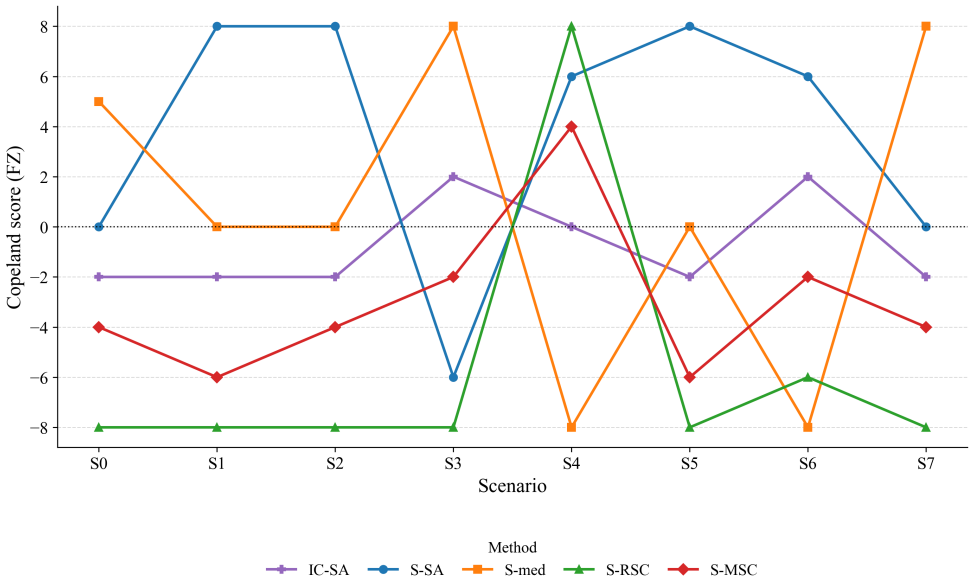


Figure 3: Copeland tournament scores across scenarios

Notes: Higher values indicate broader pairwise dominance. The figure emphasizes that tournament leadership remains heterogeneous even when IC-SA is strongest under regret-based criteria.

ment criteria remains dispersed across stage-1 rules. Section 4 asks whether this

two-dimensional pattern survives in real data.

4 Empirical Application

We next evaluate the same combination rules in real data to assess whether the stabilizer pattern documented in the Monte Carlo study persists out of sample.

Daily Thomson Reuters Datastream data cover five major developed-market equity indices: CAC 40, DAX 40, FTSE 100, NIKKEI 225, and S&P 500. Daily returns are computed as log differences of closing prices, and the HAR-Range specifications use the corresponding daily high–low range. The sample contains 5,281 observations from 5 May 2000 to 4 May 2021: 2,000 observations for initial estimation through 21 April 2008 and 3,281 out-of-sample observations from 22 April 2008 to 4 May 2021. The evaluation period covers the global financial crisis, the European sovereign debt crisis, and the initial COVID-19 shock.

We adopt a rolling-window forecasting framework. At each forecast origin, each of the eighty-one individual models is estimated using a rolling window of 2000 observations. Combination weights at stage 1 are computed over the most recent $L_1 = 1{,}000$ out-of-sample scores, and stage-2 IC weights use a rolling window of $L_2 = 500$ stage-1 scores. After obtaining the parameter estimates, we generate one-step-ahead VaR and ES forecasts. The six standard combination methods and three iterated combinations described in Section 2 are then applied. All computations are performed in R and Python.

Extending the sample beyond 4 May 2021 is not computationally feasible within the present study: each of the eighty-one individual models must be estimated at every recursive forecast origin under a rolling scheme, which amounts to several months of parallel computation even under optimized parallel loops.

We consider fifteen empirical cases obtained by combining the five indices with three tail probability levels, $\alpha \in \{0.01, 0.025, 0.05\}$. Performance is evaluated using the same metrics as in the simulation study: mean FZ loss, mean regret, $p95$ regret, average rank, top-3 probability, and Copeland score. We additionally report pass rates for all six backtests (UC, CC, DQ for VaR; NZ, MF, BD for ES), which jointly assess whether both risk measures are correctly calibrated from a regulatory perspective.

A practical challenge with joint VaR–ES forecasting is that some individual models occasionally produce numerically degenerate forecasts (e.g., near-zero VaR or ES

values, or violations of the ordering $\mathrm{ES}_t < \mathrm{VaR}_t < 0$). When such forecasts enter a combination or a loss calculation, they can create extreme outliers that distort the evaluation. To handle this transparently, we apply a common date filter within each case: a date is removed *for all methods* if, for *any* method, $\mathrm{VaR}_t \geq 0$, $\mathrm{ES}_t \geq 0$, $\mathrm{VaR}_t < \mathrm{ES}_t$, or $\min\{|\mathrm{VaR}_t|, |\mathrm{ES}_t|\} < 10^{-12}$. The filter equalizes the evaluation sample across methods while conditioning only on eight dates in total across the fifteen empirical cases; backtesting results remain based on the raw, unfiltered forecast sequences.

4.1 Aggregate Performance and Cross-Case Heterogeneity

Table 5 reports the aggregate results across fifteen cases. The table presents each method's average performance across all fifteen index–α combinations, sorted by average mean FZ loss from best to worst. Three main patterns emerge from the results.

First, as in the simulation study, no method is uniformly dominant across all evaluation criteria; the identity of the leader depends on what aspect of performance is emphasized. This divergence is substantively informative rather than contradictory. Mean FZ loss evaluates absolute average predictive accuracy under the primary scoring rule, whereas the regret measures evaluate relative shortfall from the best feasible competing rule on the same filtered sample. In particular, mean regret summarizes average relative shortfall, while $p95$ regret summarizes the upper tail of that shortfall within a case by tracking the 95th percentile of the date-level regret distribution. The divergence between loss-based and rank-based criteria has a structural explanation: mean loss rewards methods that achieve consistent small improvements across all cases, while rank-based criteria reward methods that achieve large wins in selected cases. A method optimized for the latter can appear near the top frequently without ever delivering the most stable performance across the full cross-section of cases. This distinction motivates the two-dimensional reading of Table 5: loss-based columns and rank-based columns are measuring different things, and no method leads on both simultaneously.

Second, IC-SA is the strongest aggregate performer on the loss-based criteria. It attains the lowest average mean FZ loss (1.0705), the lowest average mean regret (0.1552), and the lowest average case-level $p95$ regret (0.2570). The last result is especially important for the paper's interpretation. It means that, within a typical empirical case, the upper tail of IC-SA's relative shortfall from the best feasible rule is

smaller than that of any competing method. Put differently, IC-SA does not merely perform well on average; it more effectively limits severe relative underperformance episodes associated with rule-selection uncertainty. IC-SA also achieves the strongest overall backtesting profile across all six procedures: 15/15 UC and CC pass rates and 14/15 DQ pass rates for VaR calibration, and 15/15, 14/15, and 15/15 pass rates under the NZ, MF, and BD joint ES backtests respectively. The single MF rejection for IC-SA occurs in the SP500-0.05 case with a borderline p-value of 0.036, and three MF rejections arise across the full table, all concentrated at the $\alpha = 0.05$ level and all borderline ($p \in [0.036, 0.048]$). No method records any rejection under NZ or BD across all 135 case-method pairs. Taken together, this confirms that IC-SA's favorable loss and regret performance is not obtained at the expense of poor calibration of either risk measure.

Accordingly, the empirical evidence does not suggest that IC-SA is the universal winner under every perspective; rather, it suggests that IC-SA offers the most attractive aggregate trade-off when the objective is to limit the upper-tail cost of selecting a rule that turns out not to be the ex post best performer.

Table 5: Aggregate empirical performance across the 15 cases

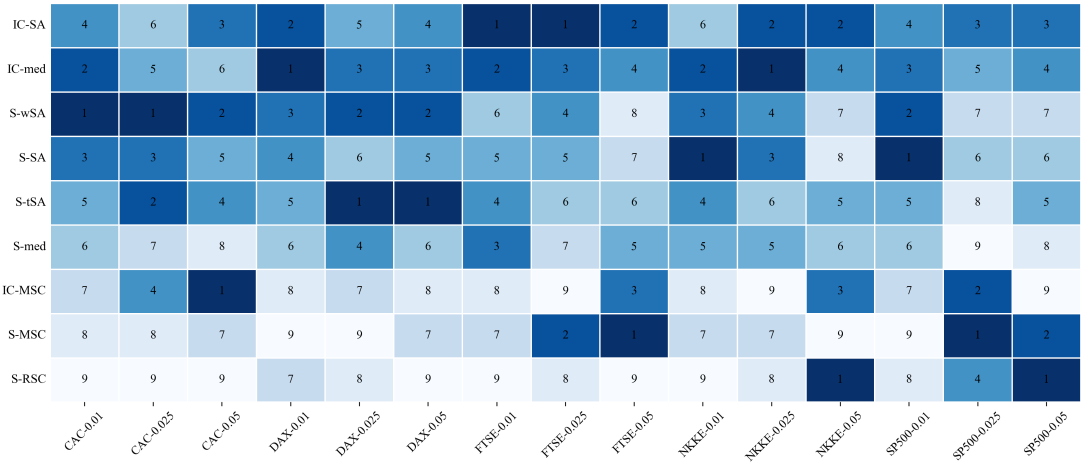
Method	Avg. mean FZ	Avg. mean regret	Avg. $p95$ regret	Avg. rank	Avg. top-3 (%)	Avg. Copeland	UC	CC	DQ	NZ	MF	BD
IC-SA	**1.0705**	**0.1552**	**0.2570**	5.09	14.7	−2.00	**15/15**	**15/15**	**14/15**	**15/15**	14/15	**15/15**
IC-med	1.0718	0.1565	0.2930	5.00	3.4	−1.47	**15/15**	14/15	13/15	**15/15**	**15/15**	**15/15**
S-wSA	1.0721	0.1567	0.3064	5.67	26.3	−5.07	**15/15**	14/15	11/15	**15/15**	**15/15**	**15/15**
S-SA	1.0727	0.1574	0.3199	7.06	12.4	−7.33	**15/15**	14/15	11/15	**15/15**	**15/15**	**15/15**
S-tSA	1.0734	0.1581	0.2939	4.33	42.7	2.13	**15/15**	14/15	12/15	**15/15**	**15/15**	**15/15**
S-med	1.0792	0.1638	0.3008	**3.63**	**57.6**	**6.67**	**15/15**	14/15	13/15	**15/15**	14/15	**15/15**
IC-MSC	1.0828	0.1675	0.2818	4.76	42.2	1.87	13/15	12/15	10/15	**15/15**	**15/15**	**15/15**
S-MSC	1.0904	0.1750	0.3073	4.80	48.8	2.00	**15/15**	**15/15**	10/15	**15/15**	14/15	**15/15**
S-RSC	1.0960	0.1807	0.2831	4.67	52.0	3.20	14/15	14/15	12/15	**15/15**	**15/15**	**15/15**

Notes: Methods are sorted by average mean FZ loss. Lower is better for loss, regret, and rank; higher is better for top-3 probability, Copeland, and pass counts. Loss-based metrics use the common date filter described in the text. Backtest pass counts are out of 15 cases at the 5% significance level. UC, CC, DQ assess VaR calibration; NZ, MF, BD assess joint VaR–ES calibration. Case-by-case results for all six tests are in the tables in Sections S4.1 and S4.2 of the Online Supplement.

Figure 4 shows the within-case rank variation through a heatmap, where each cell shows a method's rank (1–9) within a specific index–α case. This visualization reveals an important pattern: even IC-SA and IC-med, the strongest aggregate performers, do not uniformly achieve rank 1. IC-SA's ranks range from 1 to 7 depending on the case—it is not a universal winner. However, IC-SA and IC-med exhibit the most *stable* near-frontier profiles: they consistently avoid bottom-tail placements (ranks 8–9) and cluster in the upper half of the distribution. By contrast, S-RSC and S-MSC display much more volatile case-level behavior: they occasionally achieve rank

1 but also frequently finish in the bottom third. Such an observation directly mirrors the simulation finding that iterated combination stabilizes performance rather than maximizing it.

Figure 4: Within-case ranks by filtered mean FZ loss across 15 cases



Notes: Each cell reports the rank of a method within a given stock-index–α case, where rank 1 denotes the lowest filtered mean FZ loss.

4.2 Direct Parent–Child Comparisons

To isolate the contribution of the iterated (second-stage) step, we compare each IC method directly against its parent stage-1 rule. Table 6 reports the number of cases (out of 15) in which the iterated version wins, loses, or ties relative to its parent under each criterion. Notice that Table 6 is descriptive rather than inferential. It summarizes, case by case, whether the iterated rule attains a better realized value than its parent under each evaluation criterion. Formal equal predictive ability evidence is reported separately in the Supplementary Material using Diebold–Mariano comparisons under the filtered Fissler–Ziegel loss.

The most striking parent–child comparison concerns IC-SA versus S-SA. IC-SA wins in 11 of the 15 cases for both mean FZ loss and mean regret, indicating a broad improvement in average predictive performance and average relative shortfall. More importantly, IC-SA wins in 14 of the 15 cases for $p95$ regret relative to S-SA. This is the clearest empirical signature of a stabilizing method in the present paper. Relative to its parent S-SA, IC-SA does not merely improve average performance; it much more systematically compresses the upper tail of relative shortfall from the best feasible

competing rule. By combining S-SA with the other stage-1 rules, the second stage preserves the robustness of simple averaging while hedging against the cases in which S-SA is not the most suitable first-stage rule.

The comparison between IC-med and S-med reveals a different trade-off. IC-med dominates S-med in all 15 cases for mean FZ loss and mean regret (15–0 for both), indicating that the second-stage averaging consistently reduces the average losses incurred by the median parent. The p95 regret advantage is more moderate, with IC-med winning in 10 of the 15 cases. However, IC-med loses to S-med on all rank-based criteria (0–15 for average rank, top-3, and Copeland). In other words, IC-med is systematically better when judged by average predictive performance, but worse when judged by how often it finishes near the top. The pattern reinforces the general theme: iteration improves the loss profile of the median parent at the cost of some near-frontier frequency.

Finally, the comparison of IC-MSC versus S-MSC yields mixed evidence. The win–loss record is approximately balanced across criteria, consistent with the simulation finding that the stabilizing benefit of iteration is strongest when the parent rule is S-SA (a robust anchor) rather than S-MSC (which is already an optimization-based approach with its own adaptive weighting mechanism).

Table 6: Head-to-head comparison of iterated combinations against their parent methods

Comparison	Mean FZ	Mean regret	$p95$ regret	Avg. rank	Top-3 prob.	Copeland
IC-SA vs S-SA	11–4–0	11–4–0	14–1–0	15–0–0	7–8–0	14–1–0
IC-med vs S-med	15–0–0	15–0–0	10–5–0	0–15–0	0–15–0	0–15–0
IC-MSC vs S-MSC	7–8–0	7–8–0	8–7–0	9–6–0	6–9–0	7–8–0

Notes: Entries report the number of empirical cases in which the iterated rule outperforms, underperforms, or ties with its parent according to the realized value of each criterion. The Copeland column refers to the descriptive case-level Copeland score, not to inference-adjusted Diebold–Mariano (DM)-Copeland.

Accordingly, the parent–child evidence suggests that the main value of iterated combination is not that it guarantees first place, but that it reduces the severity of costly rule-selection mistakes when the parent rule is no longer the ex post best performer.

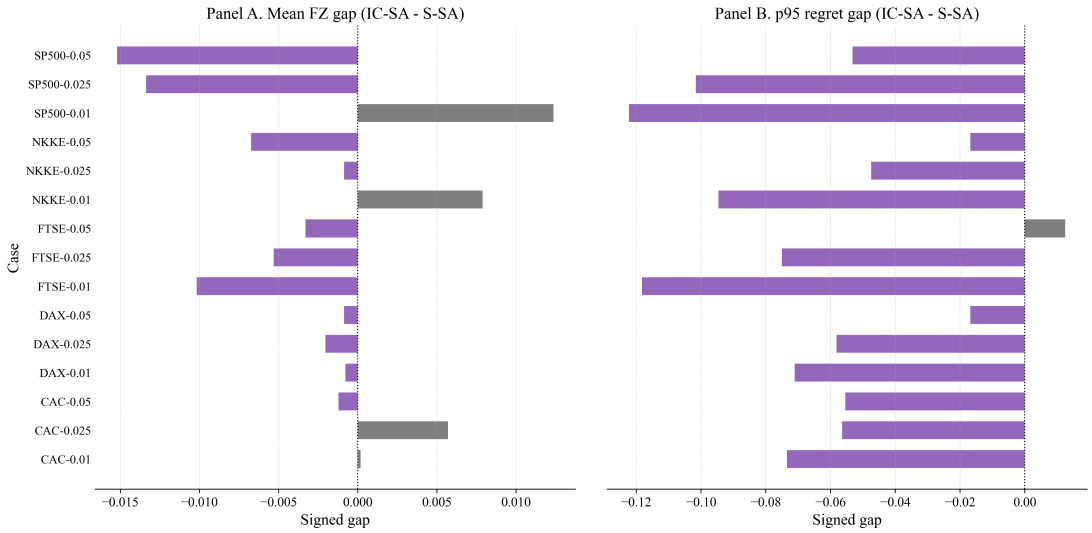
4.3 Magnitude of IC-SA Improvements

While Table 6 establishes the direction of IC-SA's advantage (how often it wins), it is equally important to assess the *magnitude* of the improvement. Figure 5 addresses this question by plotting the case-level signed gaps between IC-SA and S-SA across all fifteen cases.

Panel A of the figure shows the mean FZ loss gap, computed as the IC-SA mean FZ loss minus the S-SA mean FZ loss, so that negative values favor IC-SA. IC-SA improves mean FZ loss in 11 of 15 cases, with typical absolute gaps between 0.005 and 0.015 loss units. The four cases where S-SA leads involve similarly small margins. These are moderate but persistent improvements, consistent with the simulation evidence that IC-SA works through broad-based stabilization rather than dramatic outperformance in any single case.

Panel B reveals a sharper pattern for $p95$ regret. IC-SA improves $p95$ regret relative to S-SA in 14 of 15 cases, with gains that are often larger (in relative terms) than the mean FZ improvements. In the single case where IC-SA's $p95$ regret is higher, the deterioration is small. This asymmetry is the empirical signature of a stabilizer: IC-SA's advantage is most pronounced exactly where it matters most—in limiting the worst-case outcomes that a risk manager would like to avoid.

Figure 5: Case-level signed performance gaps between IC-SA and S-SA



Notes: Panel A: mean FZ gap Mean FZ(IC-SA)−Mean FZ(S-SA). Panel B: $p95$ regret gap. Negative values indicate better performance of IC-SA.

4.4 Formal Equal Predictive Ability Evidence

The loss and regret comparisons discussed above are economically informative, but the differences between methods are often modest in absolute magnitude. To assess whether these differences are statistically distinguishable, Table 7 reports formal equal predictive ability comparisons for IC-SA against each rival method using the filtered Fissler–Ziegel loss, based on pairwise Diebold–Mariano tests (Diebold and Mariano, 1995) with Harvey–Leybourne–Newbold correction. The purpose of this table is not to search for a universal statistically dominant method. Rather, it is to determine whether IC-SA exhibits a favorable inferential profile relative to the main alternatives.

Because the pairwise DM comparisons are used here as a structured inferential complement to the loss and regret evidence rather than as a stand-alone multiple-testing selection device, we report unadjusted case-level 5% significance and interpret the overall pattern jointly with the economic magnitudes and the case-level heterogeneity documented elsewhere in the paper. An alternative would be the Model Confidence Set of Hansen et al. (2011), which controls the familywise error rate across all pairwise comparisons simultaneously. We adopt the pairwise approach because the paper's inferential goal is specifically to characterize IC-SA's bilateral profile against each rival, not to select a single best method; the Copeland-type tournament summaries already provide the multi-method perspective in a descriptive form. Three results stand out from the table. First, pairwise significance is sparse, which is fully consistent with the paper's broader conclusion that the empirical landscape does not contain a universal champion. Second, IC-SA records no significant losses against any rival method across the 15 cases. Third, its significant wins are concentrated against the more adaptive rules—especially S-RSC and S-MSC—while comparisons with the stronger robust anchors such as S-SA, S-tSA, and S-wSA are dominated by ties. This is precisely the inferential pattern one would expect from a stabilizing rule: IC-SA does not statistically overwhelm the field, but it is systematically difficult to beat and most clearly improves on the methods whose performance is more exposed to instability.

We note that the formal DM evidence is not restricted to parent–child pairs. Instead, it compares the full feasible method menu within each case, so the parent–child relations discussed in Section 4.2 are embedded as a subset of a broader inferential comparison set.

Table 7: Pairwise equal predictive ability evidence for IC-SA across the 15 empirical cases

Rival	Avg. loss gap (rival−IC-SA)	DM (W–L–T)	Case wins	Case losses
S-SA	0.0022	1–0–14	11	4
S-med	0.0087	2–0–13	13	2
S-tSA	0.0029	2–0–13	11	4
S-wSA	0.0016	1–0–14	8	7
S-RSC	0.0255	4–0–11	13	2
S-MSC	0.0199	3–0–12	12	3
IC-med	0.0013	1–0–14	7	8
IC-MSC	0.0123	1–0–14	12	3

Notes: The table summarizes pairwise Diebold–Mariano (DM) evidence for IC-SA against each rival method across the 15 main-design empirical cases. The average loss gap is defined as the average across cases of $\bar{L}_{\text{rival}} - \bar{L}_{\text{IC-SA}}$ under the filtered Fissler–Ziegel loss, so positive values favor IC-SA. DM outcomes report significant wins, losses, and ties for IC-SA based on case-by-case two-sided DM tests with Harvey–Leybourne–Newbold correction and Newey–West HAC variance estimation, using automatic lag selection $\lfloor 4(T/100)^{2/9} \rfloor$. Case wins and case losses are based only on the sign of the case-level mean FZ difference and therefore do not require statistical significance.

4.5 Synthesis with Simulation Evidence

Taken together, the economic and inferential evidence point to the same conclusion. No method emerges as a universal champion across all empirical cases and all criteria. At the same time, IC-SA exhibits the strongest overall stabilizer profile: it leads the aggregate loss and regret criteria, improves materially on its parent S-SA, and records no significant pairwise losses against any rival in the formal equal predictive ability comparisons. Performance separates into two distinct components: (i) loss-based leadership, where IC-SA is the best aggregate performer under mean FZ loss, mean regret, and $p95$ regret and also achieves the strongest backtesting pass rates, confirming that its loss improvements are not obtained at the expense of poor calibration; and (ii) rank-based leadership, where S-med is the most consistently near-frontier method, finishing in the top three more often and accumulating the highest Copeland score. The data therefore reject a "one-rule-dominates-all" view while supporting the claim that iterated combination on top of S-SA yields the most coherent loss-based improvement, especially in the adverse states that risk managers care most about.

This pattern closely mirrors the simulation evidence in three respects. First, the absence of a universal winner persists from simulation to real data—environmental

heterogeneity is an empirical fact, not an artifact of the simulation design. Second, IC-SA's advantage concentrates in regret-based criteria rather than in average performance, exactly as predicted by the Monte Carlo study. Third, the head-to-head asymmetry (IC-SA strongly dominates its parent S-SA but not S-med) replicates the simulation finding that iterated combination provides the largest uplift when its parent is a robust averaging rule. What the simulation predicted—a two-dimensional performance structure where loss-based and rank-based leadership diverge—the data confirms through precisely the channel the theory identifies: IC-SA limits rule-selection shortfall not by dominating every criterion, but by being the method most difficult to beat when the cost of being wrong is highest.

5 Discussion

5.1 Why There Is No Universal Winner

A natural question arising from our results is why one should not expect a single universally dominant rule for combining risk forecasts. The answer lies in the nature of the forecasting problem itself. Joint VaR–ES forecasting is inherently a multi-dimensional and unstable prediction problem. Different individual models are designed to capture different stylized facts—some excel at modeling heavy tails, others at capturing leverage effects, time-varying persistence, or semiparametric flexibility. Because the relative importance of these features varies across assets, time periods, and tail probability levels, no single model provides the best forecast everywhere. By extension, no single method of combining these models can be universally optimal, because the relative quality of the inputs changes with the environment.

This observation is consistent with a broader theme in the forecast-evaluation literature. Giacomini and White (2006) and Giacomini and Rossi (2010) emphasize that predictive superiority is often environment-specific rather than globally fixed, especially when models are estimated on rolling or recursive windows. Pesaran and Timmermann (2007) show that when structural breaks are present, even the optimal estimation window may change over time, let alone the optimal forecasting method. Rossi (2013) provides a comprehensive survey of these instability issues. In our setting, the problem is even more demanding because VaR and ES summarize different features of the left tail, while modern evaluation relies on strictly consistent joint

scoring rules rather than on a single calibration criterion (Fissler and Ziegel, 2016; Patton et al., 2019). The practical consequence is that the absence of a universal winner should not be interpreted as a weakness of the empirical design. Instead, it is a *structural property* of tail-risk forecasting under instability, misspecification, and multiple economically relevant notions of performance.

5.2 Why Iterated Combination Works as a Stabilizer

Against this background, iterated combination should not be interpreted too literally as diversification across six independent forecast sources. The six stage-1 rules are all functions of the same underlying panel of 81 individual-model forecasts, so the second-stage menu is inherently overlapping. The role of the iterated layer is therefore not to create a new information set, but to regularize exposure across alternative first-stage *rule mappings* from a common forecast pool into a final VaR–ES pair.

This qualification sharpens, rather than weakens, the stabilizer interpretation. The collinearity concern is real at the level of forecast construction, but it does not imply that IC-SA is useful only because it mechanically suppresses the adaptive rules. The supplementary diagnostics show instead that a restricted four-rule average built only from the robust parents does not match full IC-SA. Hence the empirical gain from IC-SA is not reducible to simply shutting out S-RSC and S-MSC.

A more accurate interpretation is that IC-SA combines a robust core with limited but nonzero exposure to adaptive rules whose relative usefulness changes across environments. When the environment is stable, score-adaptive rules can concentrate weight efficiently; when the environment is unstable, they may overreact to recent successes. Equal weighting at the second stage imposes a disciplined shrinkage device across these competing first-stage logics. In that sense, IC-SA acts as a stabilizer under rule-selection uncertainty: it does not attempt to identify the universally best parent rule, but instead limits the cost of being too concentrated in the wrong one.

Remark 5.1. *The stabilizing mechanism is more transparent in forecast space than in raw-loss space. Let $x_{m,t} = (v_{m,t}, \xi_{m,t})'$ denote the VaR–ES forecast produced by stage-1 rule $m \in \mathcal{C}$, and let $\bar{x}_t = |\mathcal{C}|^{-1} \sum_{m \in \mathcal{C}} x_{m,t}$ denote the IC-SA forecast. For the conditional expected FZ score $Q_t(x) = E_{t-1}\{\ell_\alpha(r_t, x)\}$, suppose that Q_t is locally twice differentiable around the true conditional pair $x_t^\star = (v_t^\star, \xi_t^\star)'$, with positive-definite*

curvature matrix $H_t = \nabla^2 Q_t(x_t^\star)$*. A second-order expansion then implies*

$$Q_t(\bar{x}_t) \approx \frac{1}{|\mathcal{C}|}\sum_{m\in\mathcal{C}} Q_t(x_{m,t}) - \frac{1}{2|\mathcal{C}|}\sum_{m\in\mathcal{C}}(x_{m,t}-\bar{x}_t)' H_t (x_{m,t}-\bar{x}_t).$$

The final term is non-negative and is strictly positive whenever the parent rules produce locally distinct VaR–ES forecasts. Hence IC-SA receives a Jensen-type curvature gain: it reduces exposure to the idiosyncratic forecast error of any single parent rule by averaging forecasts before the nonlinear FZ score is realized. This mechanism does not require the raw parent losses to be weakly correlated. High loss correlations mainly reflect common market difficulty; regret subtracts the best feasible rule in the same evaluation unit and therefore focuses on the active shortfall induced by rule selection. IC-SA can consequently compress the upper tail of regret by pooling across parent rule mappings, even when the corresponding realized loss series are highly correlated.

This interpretation fits the broader forecast-combination literature, which has long emphasized that simple or regularized combinations often perform well not because they are formally optimal in every environment, but because they limit estimation error, instability, and overreaction to recent performance (Timmermann, 2006; Giacomini and Rossi, 2010; Rossi, 2013). Here the same logic operates one level higher. The object being regularized is not an individual forecast, but the choice among overlapping first-stage combination rules. Accordingly, the practical value of iterated combination lies not in universal dominance, but in delivering a more stable loss profile when the identity of the best parent rule is itself unstable.

Sub-period evidence in the supplementary material reinforces the same conclusion. IC-SA is not a crisis-only winner, but method leadership remains regime-dependent even within the empirical sample, with IC-SA strongest before 2020 and simpler stage-1 averages gaining ground in the COVID/post-COVID period. This again points to stabilization, rather than universal superiority, as the relevant interpretation.

5.3 Why IC-MSC Does Not Inherit the Advantages of IC-SA

A natural question raised by the empirical results is why the second-stage minimum-score rule, IC-MSC, does not systematically improve on the simpler iterated average. The most natural interpretation is that the second-stage optimization problem offers relatively little additional stabilization benefit once the analysis has already moved

from individual models to first-stage combination rules. The six parent rules entering the second stage are highly dependent summaries of the same underlying forecast universe, so the incremental gain from estimating flexible second-stage weights is limited. At the same time, the cost of doing so can be nontrivial: rolling-window score minimization introduces additional estimation noise, and the joint VaR–ES loss is sufficiently nonlinear that small changes in recent performance can induce unstable reweighting. In that environment, IC-MSC can overreact to short-run score differences without accessing much new cross-sectional information, whereas IC-SA applies a strong shrinkage device by pooling evenly across the first-stage rules.

The empirical evidence is consistent with exactly this interpretation. IC-MSC does not emerge as a robust winner in the main design, its relative weakness persists under shorter rolling windows, and its backtesting profile is generally less stable than that of IC-SA. Taken together, these patterns suggest that the second stage is valuable primarily as a device for stabilizing rule-selection uncertainty, not as an additional layer of aggressive score optimization. This interpretation also helps explain why IC-SA and IC-med remain competitive while IC-MSC does not: at the second stage, shrinkage across parent rules appears more valuable than re-optimization across a small and highly overlapping rule set.

5.4 The Role of S-med's Rank Dominance

The empirical role of S-med is best understood as one of ordinal robustness rather than mean-loss optimality. S-med does not dominate the aggregate FZ loss criteria, nor does it emerge as the main regret-minimizing rule. Its comparative advantage instead appears in criteria that reward persistent near-frontier placement, such as average rank, top-3 frequency, and Copeland-type summaries. This interpretation is strengthened by the rank-stability diagnostics in the supplement: S-med combines the lowest average rank with a high probability of finishing near the top and a relatively low incidence of poor rank outcomes.

In other words, S-med is not the rule that most aggressively pursues the lowest average loss, but it is the rule that most reliably avoids drifting far from the competitive set. That distinction is important for the paper's broader message. Different criteria reward different notions of robustness, and S-med's strength lies in ordinal stability, whereas IC-SA's strength lies in reducing loss-based regret under unstable

rule choice. Consequently, S-med should be viewed not as the universal loss winner, but as the most ordinally stable rule in the empirical menu.

5.5 Empirical Scope and External Validity

One limitation of the empirical exercise is the breadth of the asset universe. The analysis is conducted on five major developed-market equity indices and therefore does not directly establish that the same performance patterns carry over to other asset classes such as bonds, exchange rates, commodities, or cryptocurrencies. Moreover, the five indices are substantially correlated in returns and share similar volatility dynamics, so the fifteen empirical cases—while treated as distinct—are not statistically independent. This means the aggregate evidence reflects depth of model competition within a homogeneous asset class rather than broad cross-market generalizability. Accordingly, the empirical results should be read as strong evidence that the stabilizer interpretation of IC-SA holds in a demanding, high-correlation equity-index environment rather than as proof of universal dominance across all financial markets.

The stabilizer mechanism—compressing rule-selection regret by averaging over behaviorally distinct combination rules—is not specific to equity indices. It operates at the level of the combination layer, which depends on the statistical structure of forecast errors rather than on the asset class per se. The simulation study provides the asset-class-agnostic evidence; the empirical exercise validates the mechanism in one concrete setting.

A second scope limitation is the sample endpoint. The recursive estimation of eighty-one individual models is computationally prohibitive to extend beyond 4 May 2021 within the present study, and the evaluation period therefore excludes later stress episodes, including the 2022 inflation shock and the 2023 banking stress episode. However, the out-of-sample period already contains a sequence of heterogeneous volatility regimes—including the global financial crisis and the COVID-19 collapse and recovery—which together constitute a demanding test environment. The sample endpoint is therefore a practical constraint rather than a selective choice that favors the findings.

6 Conclusion

In this paper, we evaluate eighty-one individual VaR–ES forecasting methods combined through six standard rules and three iterated (second-stage) rules, asking whether there exists a universally superior combination rule and, if not, how iterated combination can address the resulting rule-selection uncertainty.

Our Monte Carlo study, consisting of eight distinct scenarios and 96,000 replications, delivers a clear answer. No single combination rule dominates uniformly. The identity of the best method varies with the scenario, correlation structure, panel size, and tail probability level. However, when performance is assessed through regret-based criteria—especially the upper tail of the regret distribution—iterated combination, and IC-SA in particular, emerges as the most consistently attractive alternative. Its value lies not in producing the lowest average loss in every environment, but in systematically limiting the cost of selecting an inferior rule.

The empirical exercise, covering five equity indices at three tail probability levels, confirms the simulation findings out of sample. IC-SA delivers the strongest aggregate loss-based performance and reduces $p95$ regret relative to its parent S-SA in 14 of the 15 empirical cases. At the same time, ranking-based leadership remains with simpler stage-1 rules such as S-med. Both layers of evidence therefore support the same structural interpretation: iterated combination is best understood as a regret-reducing stabilizer, or combination insurer, under rule-selection uncertainty.

These findings carry direct practical implications. Rather than attempting to identify the single best combination rule—a task our evidence suggests is inherently unstable—practitioners can reduce rule-selection risk by adding a lightweight second-stage averaging layer. The additional computational cost is negligible, while the insurance against particularly unfavorable outcomes is systematic and robust.

A final clarification concerns the role of regret. Mean FZ loss remains the primary evaluation criterion; regret is not an alternative elicitable loss but a derived diagnostic that quantifies the cost of rule-selection uncertainty relative to the ex post best feasible rule. This distinction is central: the paper does not show that IC-SA is a universal winner under every criterion, but that it is especially attractive when the objective is to limit the severity of costly rule-selection mistakes—since it repeatedly compresses the upper tail of regret across all environments examined.

Several extensions remain for future work. First, the present design applies each

combination rule to all eighty-one individual forecasts; investigating whether pre-screening models based on their historical performance can further improve iterated combination is an interesting direction. Second, while we focused on the FZ loss as the primary evaluation criterion, extending the analysis to a broader class of consistent scoring rules may reveal additional insight into the mechanism behind iterated stabilization. Third, extending the window-sensitivity analysis to expanding-window schemes, adaptive learning-window selection, or broader grids of first- and second-stage lookback lengths would complement the shorter-window robustness exercise already reported in the Online Supplement. Applying the iterated combination framework to a wider asset universe—including exchange rates, bond indices, commodities, and emerging market equities—would allow a more direct assessment of cross-asset generalizability; this is the most substantive open direction left by the present study. Finally, updating the recursive evaluation to cover post-May-2021 volatility regimes, including the 2022 inflation shock and the 2023 banking stress episode, would provide further stress-test evidence once the computational infrastructure permits.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the authors used the generative AI tools ChatGPT and Claude to improve language clarity, flow, and editorial consistency. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the manuscript.

References

Alexandridis, A.K., Panopoulou, E., Souropanis, I., 2024. Forecasting exchange rates: An iterated combination constrained predictor approach. Journal of Forecasting 43, 983–1017. doi:10.1002/for.3067.

Artzner, P., Delbaen, F., Eber, J.M., Heath, D., 1999. Coherent measures of risk. Mathematical finance 9, 203–228.

Bates, J.M., Granger, C.W.J., 1969. The combination of forecasts. Journal of the Operational Research Society 20, 451–468. doi:10.1057/jors.1969.103.

Bayer, S., Dimitriadis, T., 2022. Regression-based expected shortfall backtesting. Journal of Business & Economic Statistics 40, 90–104.

BCBS, 2016. Minimum capital requirements for market risk. Technical report, Basel Committee on Banking Supervision .

Christoffersen, P., 2010. Backtesting. Encyclopedia of Quantitative Finance .

Christoffersen, P.F., 1998. Evaluating interval forecasts. International economic review 39, 841–862.

Copeland, A.H., 1951. A "Reasonable" Social Welfare Function. Mimeographed Notes. University of Michigan, Seminar on Applications of Mathematics to the Social Sciences. Ann Arbor, MI.

Corsi, F., 2009. A simple approximate long-memory model of realized volatility. Journal of Financial Econometrics 7, 174–196.

Diebold, F.X., Mariano, R.S., 1995. Comparing predictive accuracy. Journal of Business & Economic Statistics 13, 253–263. doi:10.1080/07350015.1995.10524599. reprinted 2002, vol. 20.

Engle, R.F., Manganelli, S., 2004. Caviar: Conditional autoregressive value at risk by regression quantiles. Journal of business & economic statistics 22, 367–381.

Fissler, T., Ziegel, J.F., 2016. Higher order elicitability and osband's principle. The Annals of Statistics 44, 1680–1707.

Fissler, T., Ziegel, J.F., Gneiting, T., 2016. Expected shortfall is jointly elicitable with value at risk-implications for backtesting. Risk Magazine .

Giacomini, R., Rossi, B., 2010. Forecast comparisons in unstable environments. Journal of Applied Econometrics 25, 595–620. doi:10.1002/jae.1177.

Giacomini, R., White, H., 2006. Tests of conditional predictive ability. Econometrica 74, 1545–1578. doi:10.1111/j.1468-0262.2006.00718.x.

Gneiting, T., 2011. Making and evaluating point forecasts. Journal of the American Statistical Association 106, 746–762.

Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. Journal of the American statistical Association 102, 359–378.

Hansen, P.R., Lunde, A., Nason, J.M., 2011. The model confidence set. Econometrica 79, 453–497.

Kuester, K., Mittnik, S., Paolella, M.S., 2006. Value-at-risk prediction: A comparison of alternative strategies. Journal of Financial Econometrics 4, 53–89.

Lin, H., Wu, C., Zhou, G., 2018. Forecasting corporate bond returns with a large set of predictors: An iterated combination approach. Management Science 64, 4218–4238.

McNeil, A.J., Frey, R., 2000. Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach. Journal of empirical finance 7, 271–300.

Newey, W.K., West, K.D., 1987. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica 55, 703–708.

Nolde, N., Ziegel, J.F., 2017. Elicitability and backtesting: perspectives for banking regulation. Annals of Applied Statistics 11, 1833–1874.

Patton, A.J., Ziegel, J.F., Chen, R., 2019. Dynamic semiparametric models for expected shortfall (and value-at-risk). Journal of econometrics 211, 388–413.

Pesaran, M.H., Timmermann, A., 2007. Selection of estimation window in the presence of breaks. Journal of Econometrics 137, 134–161. doi:10.1016/j.jeconom.2006.03.010.

Politis, D.N., Romano, J.P., 1994. The stationary bootstrap. Journal of the American Statistical Association 89, 1303–1313.

Rossi, B., 2013. Advances in forecasting under instability, in: Elliott, G., Timmermann, A. (Eds.), Handbook of Economic Forecasting. Elsevier. volume 2, pp. 1203–1324. doi:10.1016/B978-0-444-62731-5.00021-X.

Stock, J.H., Watson, M.W., 2004. Combination forecasts of output growth in a seven-country data set. Journal of forecasting 23, 405–430.

Taylor, J.W., 2019. Forecasting value at risk and expected shortfall using a semiparametric approach based on the asymmetric laplace distribution. Journal of Business & Economic Statistics 37, 121–133.

Taylor, J.W., 2020. Forecast combinations for value at risk and expected shortfall. International Journal of Forecasting 36, 428–441.

Timmermann, A., 2006. Forecast combinations, in: Elliott, G., Granger, C.W.J., Timmermann, A. (Eds.), Handbook of Economic Forecasting. Elsevier. volume 1, pp. 135–196. doi:10.1016/S1574-0706(05)01004-9.