AIMS Mathematics

https://www.aimspress.com/journal/Math

AIMS Mathematics, 11(9): 27665–27688.
DOI: 10.3934/math.20261106
Received: 06 November 2025
Revised: 25 February 2026
Accepted: 03 March 2026
Published: 01 September 2026

Research article

Monotone forward expected shortfall paths with joint conformal calibration

Çağlar Sözen*

Görele School of Applied Sciences, Department of Finance and Banking, Giresun University, Giresun, Turkey

* **Correspondence:** Email: caglar.sozen@giresun.edu.tr.

Abstract: We proposed a path-level framework for forecasting forward Expected Shortfall (ES) over multiple horizons using a leak-safe, shape-preserving, and conformal-calibrated design. A single multi-output predictor produced the full ES trajectory, and an isotonic post-projection enforced monotonicity in the *reported multi-horizon ES path representation*, ensuring horizon-wise coherence by construction under our target definition and sign convention. To quantify path-level uncertainty, we attached one-sided conformal safety bands using (i) a global scalar surcharge calibrated on the pathwise residual maximum and (ii) a Šidák-adaptive allocation that redistributed the conformal error budget across horizons. Our empirical analysis focused on two representative exchange-traded funds (ETFs) with contrasting risk profiles, the iShares iBoxx $ Investment Grade Corporate Bond ETF (LQD) and the ARK Innovation ETF (ARKK), under a strictly chronological label-end split with an H-embargo and dependence-aware block calibration in the main specification ($H = 20$). The results showed that joint path coverage under the global safety band was near the nominal target for LQD and conservative for ARKK in the test sample, while the Šidák-adaptive safety band can materially reduce average excess buffer area (AEB; capital intensity) with only limited softening of joint path coverage in some cases. We also documented that marginal (horizon-wise) coverage can remain high even when joint path coverage differs, reinforcing the need to evaluate reliability at the path level. Appendix diagnostics further showed that calibration design choices (block vs. day-wise) can materially affect global-band joint path coverage, and that finite-sample Šidák and Bonferroni implementations may coincide under exact order-statistic quantiles when compared under matched nominal family-wise targets. Overall, the framework provided a transparent and modular way to move from single-horizon tail-risk summaries toward trajectory-level ES monitoring, while keeping assumptions and finite-sample limitations explicit.

Keywords: expected shortfall; multi-horizon forecasting; isotonic regression; conformal prediction; joint path coverage
Mathematics Subject Classification: 62M10, 62P20, 91G70

1. Introduction

Measuring and forecasting financial risk remains a cornerstone of modern asset management and prudential regulation. In practice, downside risk is still commonly summarized at a *single horizon* (e.g., 1-day or 10-day Value-at-Risk (VaR) / Expected Shortfall (ES)), including in regulatory and internal risk-control workflows; see, e.g., [1,2]. While such snapshot measures are useful for daily monitoring, they do not describe how tail risk *evolves across future horizons*. For portfolio managers, treasury desks, and risk committees, the operational question is rarely only "what is tomorrow's risk?", but rather "what does the next month's downside-risk trajectory look like?" We refer to this object as the forward-looking *tail-risk path*.

In this paper, the monotonicity requirement is imposed on the *reported multi-horizon ES path representation* (under our target construction and sign convention), rather than asserted as a universal property of every horizon-specific ES definition. This target-dependent formulation is important because monotonicity can depend on the precise target definition, scaling, and sign convention. This distinction matters most during stress episodes. The 2008 global financial crisis and the COVID-19 shock illustrated that risk accumulation is often nonlinear over time: volatility clustering, contagion, and regime changes can produce markedly different short- and long-horizon risk exposures [3]. Yet many standard implementations remain horizon-by-horizon or one-step-based, which can generate economically implausible path shapes when extrapolated across horizons. In particular, iterated or independently estimated multi-step risk forecasts may exhibit *crossing* behavior (or, more generally, horizon inconsistencies), implying that a longer-horizon forecast may lie below a shorter-horizon forecast in the reported risk-path representation. For reported cumulative tail-risk or capital-envelope representations, such inversions are typically difficult to interpret economically and can hinder coherent capital planning.

A growing literature on multi-horizon forecasting addresses related issues in mean and volatility prediction, including direct multi-step methods designed to avoid error compounding and horizon-specific inefficiencies; e.g., [4, 5]; see also related discussions in [6]. In risk measurement, recent contributions extend tail-risk forecasting beyond one-step settings; e.g., [7]. However, most such approaches remain *horizon-specific*: they deliver separate forecasts (and often separate uncertainty statements) for each horizon, but do not jointly enforce (i) economically sensible path shape constraints and (ii) path-level reliability guarantees for the entire trajectory.

Conformal prediction (CP) offers a complementary route by constructing calibrated prediction sets with finite-sample validity under exchangeability-like assumptions [8–11]. In finance and econometrics, CP-based intervals/bands are increasingly used for returns, volatility, and related forecasting tasks. However, most implementations target *marginal* coverage at individual horizons. For a risk manager, this is not enough: a method can appear "safe" horizon-by-horizon while still failing to protect the *full trajectory* on a substantial fraction of days. Put differently, "90% coverage at each horizon" does not imply "90% of 20- or 30-day ES paths are fully covered"; the latter is the quantity that matters for forward-looking capital planning, internal limits, and supervisory review of pathwise risk buffers.

Throughout, we distinguish the ES tail probability used in target construction from the conformal miscoverage level. Let α_{ES} denote the ES level used to construct the forward ES target, and let $\alpha_{\mathrm{conf}} \in (0, 1)$ denote the conformal miscoverage level (e.g., $\alpha_{\mathrm{conf}} = 0.10$). For notational consistency with the methodology section, let $Y_t = (Y_t(1), \ldots, Y_t(H))^\top$ denote the realized forward ES target path at aligned index t, let $\hat{Y}_t^{\text{raw}}$ denote the raw multi-horizon forecast, and let $\hat{Y}_t$ denote the reported monotone ES path after isotonic post-processing (and nonnegativity clipping) under our target construction and sign convention. Let $B_t(h)$ denote a generic one-sided safety band at horizon h (with specific constructions $B_{\text{glob},t}(h)$, $B_{\text{sid},t}(h)$, etc., defined in Section 4). We define *joint path coverage* as the event

$$Y_t(h) \leq B_t(h) \quad \text{for all } h = 1, \ldots, H,$$

and report the fraction of test dates for which this event occurs. This criterion is strictly stronger than horizon-wise marginal coverage and is directly aligned with the decision problem faced by a regulator or risk committee: whether the *entire projected risk path* remains within an acceptable capital envelope.

This paper addresses the above gaps by proposing a forward ES path forecasting framework that combines multi-horizon prediction, monotone shape regularization, and path-level conformal calibration. Importantly, the design is *modular*: the conformal calibration layer and pathwise safety-band construction are not tied to ridge regression per se, and can in principle be paired with alternative base learners (e.g., generalized autoregressive conditional heteroskedasticity (GARCH)-type models, quantile learners, tree ensembles, or neural networks) without changing the logic of the coverage target, provided the calibration design and assumptions are respected. In the present paper, we use a transparent multi-output ridge baseline to foreground the calibration and pathwise reliability layer.

Specifically, we:

1) **Define** the forward ES path as the vector $\{ES_h\}_{h=1}^{H}$ (up to H=30 trading days), representing the multi-horizon ES target trajectory under a common sign convention.

2) **Impose** an isotonic projection on the reported multi-horizon ES path representation to enforce horizon-wise coherence under the target definition and eliminate path crossings in the reported risk map.

3) **Develop** path-level conformal safety bands that target simultaneous (joint path) coverage of the *entire* ES trajectory, rather than only marginal horizon-wise coverage.

4) **Introduce** a global scalar-surcharge calibration and a Šidák-adaptive alternative, and evaluate them against naive horizon-wise adaptive and Bonferroni-style allocations under matched nominal family-wise targets. The global scalar surcharge is particularly attractive for practice because it is *audit-friendly*: a single interpretable quantity inflates the predicted path into a compliant capital envelope.

5) **Operationalize capital efficiency** as minimizing excess capital buffers subject to a joint path-coverage target. In our empirical evaluation, this is measured using average excess buffer area (AEB), i.e., the average area between the predicted ES path and the safety band (equivalently, the mean one-sided surcharge) under a required joint path-coverage level.

The Šidák-adaptive calibration uses horizon-wise error-budget allocation at level

$$\alpha_h^{\text{Sidak}} = 1 - (1 - \alpha_{\text{conf}})^{1/H},$$

which is less conservative than the Bonferroni allocation $\alpha_h^{\text{Bonf}} = \alpha_{\text{conf}}/H$ under independence. At the same time, we emphasize a practical finite-sample caveat: with discrete order-statistic calibration on a finite calibration set, Šidák and Bonferroni may sometimes map to the same empirical quantile rank, so capital-savings differences should be demonstrated empirically rather than assumed a priori.

Empirically, our main analysis focuses on two representative exchange-traded funds (ETFs) with contrasting risk profiles, the iShares iBoxx $ Investment Grade Corporate Bond ETF (LQD) and the ARK Innovation ETF (ARKK), under a strictly chronological label-end split with an H-embargo and dependence-aware block calibration. In the main specification ($H = 20$), the global one-sided conformal safety band attains joint path coverage close to the nominal target for LQD and is conservative in the test sample for ARKK, while the Šidák-adaptive safety band can materially reduce AEB with only limited softening of joint path coverage in some cases. We also report matched-target Šidák–Bonferroni sensitivity checks and show that exact order-statistic calibration can make the two implementations numerically identical in finite samples, with only minor deviations under interpolated quantile sensitivity checks.

Our contribution is methodological and operational. Methodologically, we unify shape-constrained multi-horizon ES forecasting with distribution-free pathwise uncertainty quantification. Operationally, we deliver forward-looking, horizon-aware capital maps with interpretable scalar or horizon-specific surcharges, suitable for stress testing and risk governance. We also emphasize important scope conditions: conformal guarantees are exact only under exchangeability-type assumptions, which may fail during abrupt regime shifts; monotone projection improves consistency but can introduce bias if genuine path inversions arise under rare mean-reverting or rapidly normalizing market conditions. These limitations motivate the robustness checks and future extensions discussed later in the paper.

2. Related literature

We situate our contribution at the intersection of (i) single-horizon tail-risk forecasting, (ii) multi-horizon forecasting and path-shape consistency, and (iii) distribution-free uncertainty quantification via CP for dependent financial data.

The dominant tradition in financial risk forecasting is horizon-specific. ARCH/GARCH families remain foundational for conditional volatility dynamics [12,13], while realized-measure models exploit high-frequency information to improve daily risk and volatility estimation [14, 15]. Quantile-oriented approaches such as Conditional Autoregressive Value at Risk (CAViaR) directly target VaR without requiring a full conditional distribution [16]. On the risk-measurement side, ES is a coherent alternative to VaR [17] and, together with VaR, is jointly elicitable, enabling principled forecast evaluation and comparative backtesting [18]. Despite these advances, the standard output remains a *single-horizon* forecast (or a collection of horizon-wise forecasts), leaving unresolved the pathwise questions central to multi-horizon capital planning: Are forecasts coherent across horizons, and is the *entire* path reliably covered by a safety band (equivalently, a capital envelope)?

A parallel literature studies multi-step forecasting strategies for means and volatilities, including direct and iterated approaches, with well-known trade-offs in bias, variance, and error propagation [4, 5, 19]. These insights are directly relevant for tail-risk forecasting. Iterated one-step procedures and independently fitted horizon-wise models can produce path crossings or horizon inconsistencies, especially when forecast errors accumulate or when horizons are estimated without shape-aware coupling. This problem is closely related in spirit to the broader crossing-quantile/noncrossing literature, where separately estimated quantiles violate basic ordering constraints and require ex-post rearrangement or shape-constrained corrections. Our framework adopts this perspective for forward ES paths: We generate a full multi-horizon forecast and then apply an isotonic projection to enforce monotonicity as an explicit design goal.

At the same time, we acknowledge an important trade-off of isotonic projection: it can alter the raw forecast by flattening or shifting segments of the path. Here, the shape constraint is imposed on the target representation used for pathwise risk reporting and calibration, and is not claimed as a universal structural property of all horizon-specific ES definitions. While this regularization improves horizon-wise interpretability and helps prevent implausible inversions, it may introduce bias if the unconstrained forecast contains a genuine signal (e.g., under rare conditions with strong expected mean reversion or rapid volatility normalization). For this reason, the isotonic step should be interpreted as a *shape-regularization layer*, and its empirical effect (including distortion diagnostics) should be reported rather than assumed to be innocuous.

Conformal prediction provides finite-sample calibrated prediction sets under weak assumptions, most classically exchangeability of calibration and future conformity scores [8, 9, 11]. Extensions such as conformalized quantile regression [20] broaden applicability to conditional quantile problems, while time-series-oriented variants (including ensemble/bagging and online adaptations) aim to improve practical performance under dependence and nonstationarity [21].

Our contribution is distinct in the calibration target: rather than calibrating horizon-wise marginal intervals, we calibrate *path-level* one-sided conformal safety bands for simultaneous coverage of the full ES trajectory. This distinction is not cosmetic. In a multi-horizon risk application, marginal coverage can materially overstate operational reliability because horizon-wise errors are dependent and family-wise misses accumulate across horizons. Accordingly, we compare global and adaptive error-budget allocations using Šidák and Bonferroni logic under matched nominal family-wise targets, and evaluate them through *joint path coverage* and AEB. We also explicitly acknowledge a limitation highlighted by both reviewers: Šidák-style allocations are motivated by independence-type arguments, whereas financial forecast errors are typically dependent. In practice, positive dependence may preserve conservatism, but the degree of conservatism is empirical and should be quantified.

Relative to the above strands, our framework unifies: (1) direct multi-horizon ES path forecasting, (2) an economically motivated monotone projection to ensure coherent path shape, and (3) path-level conformal calibration that yields trajectory-level reliability criteria and interpretable capital surcharges. We emphasize that this is a *framework* rather than a ridge-only model: the base predictor is swappable, whereas the contribution lies in the pathwise target, shape layer, and calibration logic.

We also delimit the scope. First, our empirical implementation uses a parsimonious feature set (e.g., the five-day realized-volatility proxy (RV5), lags, calendar effects) to isolate the contribution of the path-calibration layer and maintain stability and auditability; this may underperform richer nonlinear models in environments with complex high-frequency microstructure effects or latent macro/option-implied information. Second, exchangeability may fail under structural breaks and crisis onsets, so finite-sample CP guarantees should be interpreted cautiously in such regimes; block-wise calibration and guard-band designs are practical mitigants, not a full solution. Third, the need for dedicated calibration windows and guard bands reduces effective training data, which can matter in rapidly evolving markets or for newer/illiquid assets. These limitations motivate future work on richer features, alternative base learners (including GARCH-style, tree-based, and deep sequence models), regime-aware calibration, and broader asset classes (e.g., illiquid assets, cryptoassets, and derivative portfolios).

Recent domain-adjacent studies have also considered path-level forecasting and conformal uncertainty quantification for financial time-series curves, including forward realized variance/volatility path settings with shape regularization and one-sided band constructions [22, 23]. These studies underscore the practical relevance of trajectory-level reliability analysis in finance. In contrast, the present study targets forward ES paths and one-sided conformal safety bands for joint coverage of risk-capital envelope representations.

3. Data and target construction

Our empirical analysis focuses on two U.S.-listed ETFs: the iShares iBoxx $ Investment Grade Corporate Bond ETF (**LQD**) and the ARK Innovation ETF (**ARKK**), which together provide a deliberately heterogeneous pair in terms of risk profile and return dynamics. This heterogeneity is useful for evaluating the robustness of the proposed one-sided pathwise calibration pipeline under a common implementation.

Daily price data are obtained from *Yahoo Finance* [24] using adjusted close prices via `quantmod`. The sample window is fixed to **2022-01-01 through 2025-12-31** (trading days). For each asset, we compute daily log returns

$$r_t = \log(P_t) - \log(P_{t-1}),$$

where P_t denotes the adjusted close on day t. No forward/backward filling is applied to market price or return data. This statement applies to the Yahoo Finance price/return series only; sentiment preprocessing is handled separately as described below (last observation carried forward (LOCF) within the sentiment series only). All models are estimated separately at the asset level.

The baseline predictor set is intentionally compact and code-aligned:

1) a short-horizon realized-volatility proxy,

$$\mathsf{RV5}_t = \sqrt{\sum_{j=0}^{4} r_{t-j}^2},$$

2) lagged returns (r_{t-1}, r_{t-5}),

3) calendar controls: day-of-week dummies and an end-of-month indicator.

In the implementation, the end-of-month indicator equals one when the next calendar day is the first day of a month (i.e., a trading-day end-of-month flag). This parsimonious feature design is intentional: it keeps the base learner auditable and makes the impact of the conformal calibration layer easier to interpret.

The final pipeline evaluates both a baseline specification and a sentiment-augmented specification. The sentiment series is sourced from the EOD Historical Data (EODHD) sentiment endpoint and cached locally for reproducibility. Let $\mathsf{Sent}_t^{\mathrm{raw}}$ denote the provider's daily normalized sentiment score for the ticker. To avoid timing ambiguity, we impose a one-trading-day lag:

$$\mathsf{Sent}_t = \mathsf{Sent}_{t-1}^{\mathrm{raw}}.$$

In the code, missing sentiment values are carried forward within the sentiment series (LOCF), and any remaining initial missing values are set to zero. For the baseline specification, the sentiment channel is set to zero; for the sentiment specification, the lagged sentiment series is used. The comparison is reported as an ablation-style robustness analysis in the empirical-results section.

For each horizon $h = 1, \dots, H$, we construct the h-day cumulative return ending at date e:

$$C_e^{(h)} = \sum_{j=0}^{h-1} r_{e-j}.$$

Using the resulting horizon-specific cumulative-return series, we compute a rolling nonparametric ES label at tail level $\alpha_{\text{ES}} = 0.01$ with window length $W_{\text{ES}} = 400$:

$$ES_h(e) = \max\left\{0,\ -\frac{1}{|A_e^{(h)}|}\sum_{x\in A_e^{(h)}} x\right\}, \quad A_e^{(h)} = \left\{x \in \mathcal{W}_e^{(h)} : x \le q_{\alpha_{\text{ES}}}\left(\mathcal{W}_e^{(h)}\right)\right\},$$

where $\mathcal{W}_e^{(h)}$ is the trailing window of the most recent W_{ES} values of the horizon-h cumulative-return series, and $q_{\alpha_{\text{ES}}}$ is the empirical quantile (implemented with the same deterministic quantile convention used in the code; type 8). The positive-sign convention means that larger values correspond to larger downside tail risk.

Hence, the response at aligned index e is the forward ES path

$$Y_e = (ES_1(e), ES_2(e), \dots, ES_H(e)),$$

which is modeled jointly. In the notation of Section 1, this path corresponds to the realized target vector, while the model produces a raw forecast vector that is subsequently transformed into a monotone reported path via isotonic post-projection. For notational convenience in Section 4, we relabel the aligned label-end index e as t after preprocessing; this is a notational change only. The main specification uses $H_{\text{main}} = 20$, while Appendix results report the grid $H \in \{10, 20, 30\}$.

For a given horizon H, the modeling dataset is formed after removing the warm-up period required by both multi-horizon labels and the rolling ES window. In the implementation, the usable row index begins at $H + W_{\text{ES}}$, and rows are retained only if all predictors and all ES targets $ES_1, \dots, ES_H$ are finite. This yields the final modeling sample used in the train/calibration/test protocol.

We then apply a strict chronological split with explicit embargo gaps of length H between adjacent slices:

- **Train**: first 50% of usable observations,
- **Calibration**: nominally 30% of usable observations, starting after an embargo of length H,
- **Test**: remaining observations, starting after an additional embargo of length H after calibration.

For the main specification ($H = 20$), the resulting sample counts are reported in Table 1; in our final runs, the two assets have the same split sizes. Hyperparameters are selected using training data only, and conformal calibration quantiles are estimated on the calibration slice only and then frozen for test evaluation.

The main text reports results for the two selected assets under a common pipeline, emphasizing the $H = 20$ main specification and comparing baseline vs sentiment variants (Tables 1–4). Appendix tables report block-calibration sensitivity, Šidák-vs-Bonferroni quantile sensitivity (including interpolated-quantile diagnostics), and the full horizon grid (Tables A1–A3). This structure is designed to be concise in the main text while keeping robustness checks transparent and reproducible.

4. Methodology

We propose a modular three-layer pipeline for forecasting and safeguarding the forward ES path: (i) a multi-horizon regression model for point prediction, (ii) monotone/nonnegative post-processing of the predicted path, and (iii) one-sided conformal upper safety bands calibrated on a disjoint chronological slice [8–10,25,26]. Exact finite-sample validity is stated only under exchangeability; for time-series dependence, we use a practical block-maxima calibration heuristic and report sensitivity analyses [8–10].

To make the modular design explicit, let $\mathcal{B}$ denote an arbitrary base multi-horizon predictor mapping features x_t to a raw ES path $\hat{Y}_t^{\mathrm{raw}} \in \mathbb{R}^H$. In this paper, $\mathcal{B}$ is instantiated as a multi-output ridge model for transparency and stability, but the subsequent isotonic projection and conformal calibration layers are model-agnostic [8, 9, 25–27]. In particular, the reference split-conformal validity statements depend on the calibration score construction and exchangeability assumptions, not on ridge regression per se; thus, $\mathcal{B}$ could be replaced by a GARCH-type, quantile-regression, tree-based, or neural-network predictor without changing the calibration logic [8, 9, 26].

For ease of reference, the boxed notation summary below presents the main symbols used in the methodology, including the distinction between the ES target level α_{ES} and the conformal miscoverage level α_{conf}. A compact assumptions roadmap is provided immediately below as well to clarify the scope of the conformal validity statements in our chronological time-series setting [8–10].

Notation & Assumptions Roadmap (for Sections 3 and 4).

Core notation. $Y_t = (Y_t(1), \dots, Y_t(H))^\top$ is the realized forward ES target path at aligned (label-end indexed) time t; $\hat{Y}_t^{\mathrm{raw}}$ is the raw multi-horizon forecast from the base learner; $\hat{Y}_t$ is the deployed reported path after isotonic post-processing and nonnegativity clipping; s_h is the train-only robust scale (median absolute deviation (MAD)-based) for horizon h; $Z_t^+(h) = \max\{(Y_t(h) - \hat{Y}_t(h))/s_h, 0\}$ is the one-sided normalized underprediction score; $R_t = \max_{1 \le h \le H} Z_t^+(h)$ is the pathwise calibration score; q_{glob} is the OS conformal multiplier for the global safety band; α_{ES} is the ES target-construction level; α_{conf} is the conformal miscoverage level.

Calibration design (implementation-aligned).

1) *Chronological split with embargo.* Train / calibration / test are strictly chronological and separated by an H-length embargo to reduce target overlap from rolling multi-horizon labels.
2) *Train-only fitting and scaling.* Predictor standardization and horizon-specific residual scales s_h are estimated on the training slice only and then frozen.
3) *Deployed predictor calibration.* Conformal calibration is applied to the deployed predictor $\hat{Y}_t$ (after isotonic projection and nonnegativity clipping), not to the raw forecast $\hat{Y}_t^{\mathrm{raw}}$.
4) *Main quantile rule.* The conformal layer uses the order-statistic (OS) split-conformal quantile throughout the main results; interpolated quantiles are reported only as descriptive sensitivity checks.

Validity scope and dependence caveat.

1) Theorem 4.1 is stated as the *reference* split-conformal validity result under exchangeability of calibration and future scores.
2) In our time-series setting, exchangeability is not exact; therefore exact finite-sample validity is *not* claimed in general under dependence.
3) Block-maxima calibration (nonoverlapping blocks, main specification $b = 10$) is used as a practical dependence-mitigation heuristic; its impact is assessed empirically via sensitivity analyses.
4) Šidák and Bonferroni allocations are compared under a matched nominal family-wise target $1 - \alpha_{\mathrm{conf}}$; finite-sample OS quantile discreteness can make them numerically identical at some horizons.

4.1. Forward ES path and shape-constrained reported representation

Fix a horizon $H \in \mathbb{N}$ and an ES labeling level $\alpha_{\text{ES}} \in (0, 1)$. In the empirical implementation, $H \in \{10, 20, 30\}$ and the main specification is $H_{\text{main}} = 20$, with $\alpha_{\text{ES}} = 0.01$ and rolling label window $W_{\text{ES}} = 400$.

Let $Y_t = (Y_t(1), \ldots, Y_t(H))^\top \in \mathbb{R}^H$ denote the aligned forward ES target path at index t (in the code, this index corresponds to the label-end date after preprocessing and warm-up alignment). By construction,

$$Y_t(h) = ES_h(t) \in \mathbb{R}_{\geq 0}, \qquad h = 1, \ldots, H.$$

As emphasized in Section 1, the monotonicity requirement is imposed on the *reported* multi-horizon ES path representation under our target construction and sign convention, rather than asserted as a universal property of every horizon-specific ES definition.

Let $\hat{Y}_t^{\text{raw}} \in \mathbb{R}^H$ be an unconstrained multi-horizon forecast from the base learner. We enforce horizon-wise coherence through isotonic regression across the horizon index using the monotone cone

$$\mathcal{M} \equiv \{y \in \mathbb{R}^H : \ y_1 \leq y_2 \leq \cdots \leq y_H\},$$

which is a closed convex cone [25, 27]. The isotonic post-processed forecast is

$$\hat{Y}_t^{\text{iso}} = \text{Iso}\left(\hat{Y}_t^{\text{raw}}\right) \in \mathcal{M}.$$

In the implementation, isotonic regression is applied row-wise using `isoreg` [25, 27]. Since ES labels are nonnegative by construction, the deployed point forecast used for calibration and evaluation is then clipped componentwise at zero:

$$\hat{Y}_t = \max(\hat{Y}_t^{\text{iso}}, 0).$$

This clipping preserves monotonicity because the map $u \mapsto \max(u, 0)$ is monotone coordinatewise. Hence, the deployed reported path satisfies

$$\hat{Y}_t(1) \leq \hat{Y}_t(2) \leq \cdots \leq \hat{Y}_t(H) \qquad (\forall t). \tag{4.1}$$

Proposition 4.1. *For any $\hat{y}^{\text{raw}} \in \mathbb{R}^H$, the isotonic fit $\hat{y} = \text{Iso}(\hat{y}^{\text{raw}})$ is the Euclidean projection onto $\mathcal{M}$:*

$$\hat{y} = \arg\min_{y \in \mathcal{M}} \|y - \hat{y}^{\text{raw}}\|_2^2.$$

Thus, isotonic post-processing introduces the minimum ℓ_2 distortion among all monotone paths.

Remark 4.1. A hard-constrained parameterization (e.g., nonnegative increments) is possible, but isotonic post-processing is simpler, model-agnostic, and computationally negligible at the horizon lengths used here [25, 27]. Crucially, conformal calibration is performed on the deployed (post-processed and nonnegative-clipped) predictor, so uncertainty quantification targets the actual forecast used in evaluation [8–10].

4.2. *Base learner: multi-output ridge (code-aligned)*

Let $x_t \in \mathbb{R}^p$ denote the predictor vector at aligned index t. In the final implementation,

$$x_t = (\mathsf{RV5}_t,\ r_{t-1},\ r_{t-5},\ \mathsf{EOM}_t,\ \mathsf{DOW\ dummies}_t,\ \mathsf{Sent}_t),$$

where Sent_t is the lagged EODHD sentiment feature in the sentiment specification and is set to zero in the baseline specification [28].

We use multi-output ridge regression via `glmnet` with `family = "mgaussian"` and `alpha = 0` [26]. Predictors are standardized using training-slice means and standard deviations only (with a unit fallback for zero/degenerate scales). Let $\mathcal{T}_{\mathrm{tr}}$ denote the training slice. The coefficient matrix $B \in \mathbb{R}^{p\times H}$ solves

$$\min_B \frac{1}{|\mathcal{T}_{\mathrm{tr}}|} \sum_{t\in\mathcal{T}_{\mathrm{tr}}} \|Y_t - B^\top x_t\|_2^2 + \lambda\|B\|_F^2, \tag{4.2}$$

with ridge penalty $\lambda \geq 0$ selected by blocked/time-ordered 5-fold cross-validation on the training slice only [26]. The raw prediction is

$$\hat{Y}_t^{\mathrm{raw}} = B^\top x_t,$$

followed by isotonic projection and nonnegativity clipping.

Remark 4.2. The base learner is intentionally simple and stable. The primary contribution of the pipeline is the combination of shape control and one-sided pathwise calibration, not aggressive predictive feature engineering or high-capacity model selection [26]. Accordingly, ridge should be read as a transparent baseline instantiation of the modular predictor layer, rather than as a restriction of the framework.

4.3. *Chronological split, embargo, and leakage control*

Let $\mathcal{T}_{\mathrm{tr}}$, $\mathcal{T}_{\mathrm{cal}}$, and $\mathcal{T}_{\mathrm{te}}$ denote the train, calibration, and test slices, respectively, after preprocessing and finite-row filtering. The implementation uses a strict chronological split: 50% train, 30% calibration, and the remaining observations for test, with an embargo (guard band) of length H between train–calibration and calibration–test.

This embargo is important because multi-horizon ES labels are constructed from overlapping rolling cumulative returns. The embargo reduces direct target overlap across adjacent slices and supports a more conservative evaluation design.

4.4. *One-sided split-conformal calibration on normalized residuals*

To distinguish the conformal layer from the ES labeling level, let $\alpha_{\mathrm{conf}} \in (0, 1)$ denote the conformal miscoverage level. In the main specification, $\alpha_{\mathrm{conf}} = 0.10$, so the target joint path coverage is $1-\alpha_{\mathrm{conf}} = 0.90$.

For $t \in \mathcal{T}_{\mathrm{tr}}$, define training residuals

$$E_t^{\mathrm{tr}}(h) = Y_t(h) - \hat{Y}_t(h), \qquad h = 1, \ldots, H.$$

Residual dispersion differs by horizon, so we normalize using robust horizon-specific scales estimated on the training slice only:

$$s_h = 1.4826 \times \mathrm{MAD}\Big(\{E_t^{\mathrm{tr}}(h) : t \in \mathcal{T}_{\mathrm{tr}}\}\Big), \qquad h = 1, \ldots, H.$$

The implementation uses the median-centered MAD and applies a positive floor when the estimate is degenerate. These scales are frozen after training and are not re-estimated on calibration or test.

For $t \in \mathcal{T}_{\mathrm{cal}}$, define calibration residuals

$$E_t(h) = Y_t(h) - \hat{Y}_t(h), \quad Z_t(h) = \frac{E_t(h)}{s_h}, \quad Z_t^+(h) = \max\{Z_t(h), 0\}.$$

Positive values correspond to *underprediction* of ES and therefore represent the relevant direction for one-sided capital protection.

Given a finite score sample $\{u_i\}_{i=1}^n$ and a target conformal miscoverage level $\tilde{\alpha}_{\mathrm{conf}}$, we use the order-statistic split-conformal quantile

$$Q^{\mathrm{OS}}_{1-\tilde{\alpha}_{\mathrm{conf}}}(u_{1:n}) = u_{(k)}, \qquad k = \lceil (1 - \tilde{\alpha}_{\mathrm{conf}})(n + 1) \rceil,$$

with k truncated to the interval $[1, n]$ [8–10]. This is the main quantile rule used throughout the conformal calibration layer [8–10].

Define the pathwise calibration score

$$R_t = \max_{1 \le h \le H} Z_t^+(h), \qquad t \in \mathcal{T}_{\mathrm{cal}}.$$

The global conformal multiplier is

$$q_{\mathrm{glob}} = Q^{\mathrm{OS}}_{1-\alpha_{\mathrm{conf}}}(\{R_t : t \in \mathcal{T}_{\mathrm{cal}}\}),$$

and the global one-sided upper safety band is

$$B_{\mathrm{glob},t}(h) = \hat{Y}_t(h) + q_{\mathrm{glob}}\, s_h, \qquad h = 1, \ldots, H. \tag{4.3}$$

This safety band targets the joint path-coverage event $Y_t(h) \le B_t(h)$ for all $h = 1, \ldots, H$ [8–10].

Theorem 4.1. *Suppose the calibration scores $\{R_t : t \in \mathcal{T}_{\mathrm{cal}}\}$ and the score of a future point are exchangeable, with the predictor and post-processing map fixed before calibration. Then, the global safety band* (4.3) *satisfies*

$$\mathbb{P}\Big(Y_{\mathrm{new}}(h) \le B_{\mathrm{glob,new}}(h) \ \ \forall h = 1, \ldots, H\Big) \ge 1 - \alpha_{\mathrm{conf}}.$$

In our time-series application, exchangeability is not exact. We therefore present Theorem 4.1 as the canonical split-conformal reference result and assess finite-sample performance empirically under chronological testing, including block-based sensitivity analyses [8–10].

4.5. Per-horizon one-sided safety bands: naive horizon-wise adaptive, Šidák, and Bonferroni

For each horizon h, let $\{Z_t^+(h) : t \in \mathcal{T}_{\mathrm{cal}}\}$ denote the horizon-specific one-sided score sample.

Using the same global conformal miscoverage level at each horizon,

$$q_{\mathrm{naive}}(h) = Q^{\mathrm{OS}}_{1-\alpha_{\mathrm{conf}}}(\{Z_t^+(h)\}), \qquad B_{\mathrm{naive},t}(h) = \hat{Y}_t(h) + q_{\mathrm{naive}}(h) s_h.$$

This *naive horizon-wise adaptive* safety band targets marginal horizon-wise coverage and does not provide joint pathwise control [8, 9].

Define

$$\alpha_h^{\text{Sidak}} = 1 - (1 - \alpha_{\text{conf}})^{1/H},$$

and calibrate

$$q_{\text{sid}}(h) = Q^{\text{OS}}_{1-\alpha_h^{\text{Sidak}}}(\{Z_t^+(h)\}), \qquad B_{\text{sid},t}(h) = \hat{Y}_t(h) + q_{\text{sid}}(h)s_h.$$

Define

$$\alpha_h^{\text{Bonf}} = \alpha_{\text{conf}}/H,$$

and calibrate

$$q_{\text{bonf}}(h) = Q^{\text{OS}}_{1-\alpha_h^{\text{Bonf}}}(\{Z_t^+(h)\}), \qquad B_{\text{bonf},t}(h) = \hat{Y}_t(h) + q_{\text{bonf}}(h)s_h.$$

The Šidák and Bonferroni allocations are compared at the same nominal family-wise target $1-\alpha_{\text{conf}}$ (matched target), so any empirical difference reflects the allocation rule and finite-sample quantile mapping rather than a different nominal reliability objective [8, 9, 29]. A finite-sample technical point is important for interpretation: because split conformal uses order-statistic (OS) quantiles, distinct nominal levels α_h^{Sidak} and α_h^{Bonf} can map to the same empirical quantile index k for a given horizon (and this is more likely when the effective calibration sample is small, e.g., after block-maxima compression). In such cases, the resulting Šidák and Bonferroni safety bands are identical at that horizon, so negligible AEB differences are expected despite different nominal formulas [8, 9, 29].

Under independence across horizons, Šidák is exact for the joint target, whereas Bonferroni is conservative [29]. Under serial and cross-horizon dependence, we do not claim a universal dominance ordering; instead, we compare the resulting safety bands empirically using joint path coverage and AEB.

4.6. Dependence mitigation via block-maxima calibration

Because calibration scores are serially dependent in time series, exact split-conformal guarantees need not hold. Our main specification therefore uses a practical dependence-mitigation heuristic based on nonoverlapping block maxima, while avoiding exact finite-sample validity claims under dependence [9, 10].

Given a score sequence $\{u_t\}_{t\in\mathcal{T}_{\text{cal}}}$ (either the global pathwise scores R_t or horizon-specific scores $Z_t^+(h)$), partition the calibration slice into nonoverlapping blocks of size b and compute block maxima

$$\widetilde{u}_j = \max_{t\in\mathcal{B}_j} u_t.$$

Calibration then proceeds by applying the same OS quantile rule to $\{\widetilde{u}_j\}$ instead of the raw score sequence. Any incomplete trailing block is dropped in the implementation when forming nonoverlapping blocks.

In the final main specification, block calibration is enabled with block size $b = 10$. Appendix Table A1 reports the block-vs-non-block sensitivity at $H = 20$.

Remark 4.3. The main conformal results use the OS rule. Because OS quantiles are discrete in finite samples (especially after block-maxima compression), Šidák-vs-Bonferroni differences may be muted or absent even when the nominal levels differ [8, 9, 29]. For descriptive sensitivity only, Appendix Table A2 also reports a block-disabled comparison using interpolated empirical quantiles (type 8). This does not replace the OS-based main calibration.

4.7. Algorithmic summary

Inputs: returns $\{r_t\}$, predictors $\{x_t\}$, horizon H, ES label parameters $(\alpha_{\text{ES}}, W_{\text{ES}})$, conformal miscoverage level α_{conf}, and optional block size b.
Outputs: point forecasts $\hat{Y}_t$ and one-sided upper safety bands $B_{\text{glob}}, B_{\text{naive}}, B_{\text{sid}}, B_{\text{bonf}}$.

1) For each $h = 1, \ldots, H$, compute rolling ES labels from h-day cumulative returns using window length W_{ES}; align labels by the common index used for model fitting.
2) Compute RV5, lagged returns, calendar controls, and (for the sentiment specification) lagged EODHD sentiment; cache sentiment locally and apply one-day lag [28].
3) Remove warm-up and incomplete rows (implementation starts from index $H + W_{\text{ES}}$ and retains only rows with finite predictors and all H labels).
4) Split into train/calibration/test using 50% / 30% / remaining proportions and embargo length H between adjacent slices.
5) Standardize predictors using train-only statistics and fit multi-output ridge with blocked/time-ordered 5-fold CV for λ [26].
6) Apply isotonic regression across horizons and clip negative values to zero [25, 27].
7) Compute train-only robust scales s_h, form normalized positive calibration scores, and estimate OS conformal multipliers (with optional block-maxima calibration) [8–10].
8) Compute point forecast accuracy, marginal coverage, joint path coverage, and capital-efficiency summaries.

4.8. Evaluation metrics and capital-efficiency summaries

We report test-slice performance along three dimensions:

1) **Point forecast accuracy:** root mean squared error (RMSE) over short horizons ($h = 1$–5) and RMSE over all horizons ($h = 1{:}H$).
2) **Coverage:** marginal (horizon-wise) coverage and joint path coverage, estimated empirically on the test slice:
$$\widehat{\text{Cov}}_{\text{joint}} = \frac{1}{|\mathcal{T}_{\text{te}}|} \sum_{t \in \mathcal{T}_{\text{te}}} \mathbf{1}\{Y_t(h) \leq B_t(h)\ \forall h = 1, \ldots, H\}.$$
3) **Capital efficiency:** AEB, computed as the mean nonnegative gap between the upper safety band and the point forecast:
$$\text{AEB}_{1:H} = \frac{1}{|\mathcal{T}_{\text{te}}| H} \sum_{t \in \mathcal{T}_{\text{te}}} \sum_{h=1}^{H} \left(B_t(h) - \hat{Y}_t(h)\right)_+,$$
together with a short-horizon version averaged over $h = 1, \ldots, \min(5, H)$.

To make safety-band comparisons operationally interpretable, the main results table reports the AEB-based percentage reduction when moving from the Global uniform safety band to the Šidák safety band (Global→Šidák "capital saved" %). We treat this as an empirical trade-off summary, not a universal dominance claim. Where relevant, we additionally report Šidák-vs-Bonferroni sensitivity at the same nominal family-wise target, and explicitly note cases of OS-quantile coincidence in finite samples.

Remark 4.4. The main text emphasizes $H = 20$ with block calibration enabled and reports baseline and sentiment results for LQD and ARKK. Appendix tables provide: (i) block-vs-non-block calibration sensitivity, (ii) Šidák-vs-Bonferroni sensitivity under OS and interpolated quantiles (type 8), and (iii) the full grid over $H \in \{10, 20, 30\}$ and both variants.

5. Empirical results

5.1. Design and reporting

We report results for two representative ETFs with contrasting risk profiles: **LQD** (investment-grade corporate bonds) and **ARKK** (high-volatility growth/innovation exposure). The empirical design uses a fixed horizon $H = 20$ for the *main specification*, with strictly chronological, leak-safe splitting and an embargo of length H between train/calibration/test segments (Table 1).

The predictor is a monotone ES-path forecaster (post-processed via row-wise isotonic regression / pool-adjacent-violators (PAV) algorithm), and uncertainty is attached using one-sided conformal upper safety bands. Our primary reporting emphasizes: (i) point accuracy (RMSE over $h = 1{:}5$ and $h = 1{:}H$), (ii) *joint path coverage* (uniform over horizons), and (iii) capital intensity summarized by AEB. To avoid over-interpretation, horizon-wise marginal coverages are shown separately (Figure 3) and are not conflated with joint path coverage. Additional finite-sample sensitivity checks (block vs. day-wise calibration, matched-target Šidák vs. Bonferroni comparisons under exact OS and interpolated quantiles, and horizon/variant grid comparisons) are deferred to Appendix Tables A1–A3.

Unless otherwise stated, results use the baseline feature set, $H = 20$, $\alpha_{\mathrm{ES}} = 0.01$, $\alpha_{\mathrm{conf}} = 0.10$ (target joint path coverage = 0.90), and **block calibration** with block size 10.

5.2. Sample sizes and leak-safe split

Table 1 documents the effective sample sizes under the label-end split with H-embargo. Both assets share the same design counts ($N = 583$, $n_{\mathrm{train}} = 291$, $n_{\mathrm{calib}} = 174$, $n_{\mathrm{test}} = 78$), which facilitates like-for-like comparison of calibration behavior across materially different volatility regimes. The broader horizon grid ($H \in \{10, 20, 30\}$) used in appendix sensitivity summaries is reported in Table A3.

Table 1. Sample sizes and leak-safe split (embargo = H) for the main specification ($H = 20$, baseline variant, block calibration).

		Leak-safe split sizes				ES / Conformal setup				Block calibration	
Asset	H	N	Train	Calib.	Test	α_{ES}	W_{ES}	α_{conf}	Target joint cov.	Used	Size
LQD	20	583	291	174	78	0.01	400	0.10	0.90	Yes	10
ARKK	20	583	291	174	78	0.01	400	0.10	0.90	Yes	10

5.3. Descriptive contrast between LQD and ARKK

Table 2 confirms the intended contrast in scale and tail risk. Relative to LQD, ARKK exhibits substantially larger return volatility, higher short-horizon ES (ES_1), and a much larger H-horizon ES level (ES_H). This scale separation is useful for testing whether the same conformal design behaves sensibly across both low- and high-volatility assets. The same two-asset contrast persists across the extended horizon/variant grid in Appendix Table A3.

Table 2 reports descriptive summaries on the pre-model descriptive sample used for target/predictor diagnostics. Accordingly, its count ($N = 584$) can differ slightly from the final effective modeling sample in Table 1 ($N = 583$), which is obtained after the full joint finite-row filtering used for train/calibration/test estimation.

Table 2. Descriptive statistics (2022-01-01–2025-12-31) for returns, RV5, and ES targets (main specification: $H = 20$). Here, N denotes the descriptive pre-model sample count (before the final joint finite-row filtering used in Table 1).

		Returns		RV5		ES_1		ES_H	
Asset	N	Mean	SD	Mean	SD	Mean	SD	Mean	SD
LQD	584	0.000255	0.004703	0.009587	0.004365	0.015352	0.001780	0.052998	0.013031
ARKK	584	0.001001	0.025103	0.051104	0.023401	0.073866	0.012342	0.265265	0.072129

5.4. Main test results (H=20, baseline, block calibration)

Table 3 reports the core out-of-sample results for the main specification. Sensitivity analyses that contextualize these estimates are reported in Tables A1–A3.

Three findings are most relevant. First, for LQD, the global one-sided safety band attains joint path coverage 0.9103, close to the nominal 0.90 target. For ARKK, global joint path coverage is 1.0000 on the test segment, indicating conservative calibration in this sample. We interpret this as finite-sample conservatism (especially under dependence-aware block calibration), not as a universal guarantee. Table A1 shows that this conclusion is materially affected by the calibration scheme (block vs. day-wise), particularly for the global safety band. Second, for LQD, moving from the Global safety band to the Šidák safety band reduces AEB sharply (about 95.4% on the full-horizon metric), but joint path coverage decreases from 0.9103 to 0.8974 (slightly below the nominal target). For ARKK, the Šidák safety band also reduces AEB materially (about 52.3%) while joint path coverage remains high at 0.9872. This illustrates the expected trade-off: adaptive allocation can improve capital efficiency, but path-level reliability should be checked empirically rather than assumed. Third, we compare Šidák and Bonferroni under the *same nominal family-wise target* (matched target; $1 - \alpha_{\text{conf}} = 0.90$). In the main $H = 20$ block-calibrated specification, Table A1 shows numerically identical Šidák and Bonferroni *joint path coverage* for both assets. In our main runs, the corresponding adaptive-band results are also numerically indistinguishable under exact OS calibration (not separately duplicated in Table 3 to avoid redundancy). Accordingly, the adaptive-band values reported in Table 3 (e.g., LQD: joint path coverage 0.8974, $AEB_{1:H} = 0.000284$; ARKK: joint path coverage 0.9872, $AEB_{1:H} = 0.033033$) can be read as the practical matched-target benchmark for both Šidák and Bonferroni in this finite sample. Table A2 (reported for block = FALSE as a finite-sample quantile-implementation sensitivity check) explains the mechanism: under exact OS calibration, Šidák and Bonferroni often map to the same empirical order statistic, while small differences can reappear under interpolated quantiles (type 8). We report those interpolated differences as descriptive sensitivity rather than as a replacement for the OS-based main calibration.

Table 3. Main test results (baseline variant; block calibration; $H = 20$). Coverage columns report joint path coverage (uniform over horizons); AEB denotes average excess buffer area. "Saved %" reports AEB-based capital saved when moving from the Global safety band to the Šidák safety band. "Over-pen." denotes the code-level asymmetric over-prediction penalty coefficient (η), reported for reproducibility (it does not change the conformal calibration formulas).

		Forecast accuracy		Joint path cov.			AEB		Saved %		Cal./scale/repro.		
Asset	H	$\text{RMSE}_{1:5}$	$\text{RMSE}_{1:H}$	Glob.	Naive	Šidák	Glob.	Šidák	All	1:5	q_{glob}	Med. s_h	Over-pen.
LQD	20	0.0026	0.0144	0.910	0.897	0.897	0.0062	0.0003	95.42	75.68	0.8038	0.0081	0.00
ARKK	20	0.0042	0.0346	1.000	0.987	0.987	0.0693	0.0330	52.32	87.32	2.9497	0.0253	0.05

5.5. Median ES paths and global safety bands

Figure 1 summarizes the median forward ES path together with the global one-sided safety band for both assets. The monotonic shape constraint yields economically coherent (nondecreasing) paths by construction. The contrast across panels is informative: LQD exhibits a relatively smooth and low-amplitude path, whereas ARKK shows a much steeper ES term structure and a wider conformal envelope, consistent with the larger residual/tail scale documented in Tables 2 and 3. The extended horizon/variant patterns are tabulated in Table A3.

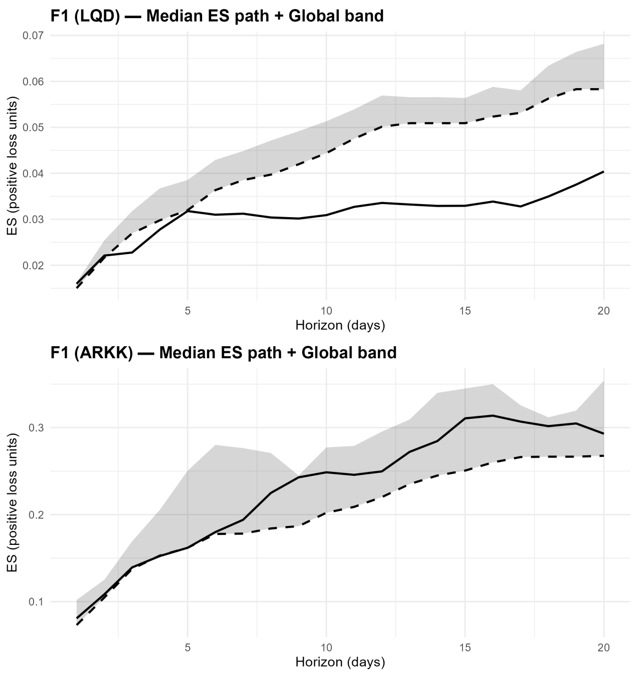

Figure 1. Median forward ES path and global one-sided safety band for LQD and ARKK under the main specification ($H = 20$, baseline variant, block calibration). The panels highlight the contrast in ES scale and safety-band amplitude across a low-volatility bond ETF and a high-volatility innovation ETF.

5.6. *Width-ratio diagnostics: Šidák vs global*

Figure 2 plots the mean width ratio (Šidák / Global) across horizons. Values below 1 indicate thinner Šidák safety bands than the global alternative. For LQD, the ratio is well below 1 for most horizons and is close to zero at several horizons, consistent with the strong AEB savings in Table 3. The near-zero segments should be interpreted as a finite-sample/discrete-quantile effect (rather than a general structural result), which is also consistent with the order-statistic sensitivity discussion in Table A2. For ARKK, the ratio remains below 1 but varies more across horizons, indicating that the adaptive allocation is less uniformly dominant in the higher-volatility setting. We treat these horizon profiles as descriptive finite-sample diagnostics.

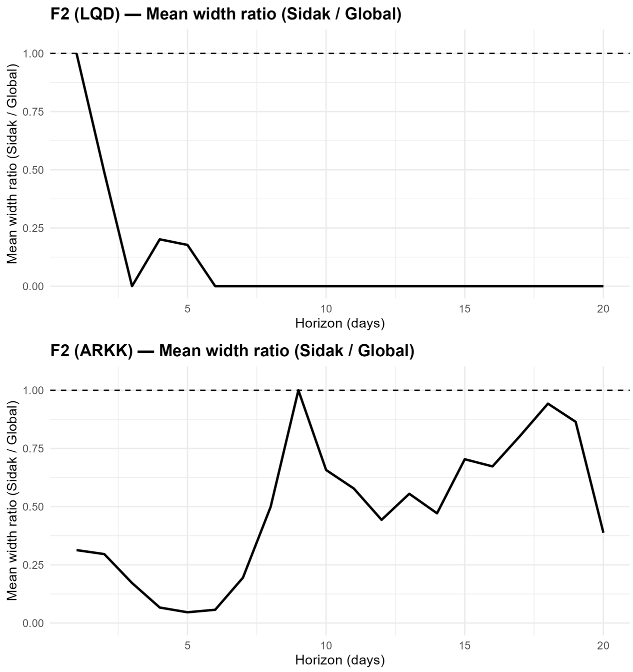

Figure 2. Mean width ratio across horizons (Šidák / Global) for LQD and ARKK under the main specification. The dashed line at 1 marks parity with the global safety band; values below 1 indicate thinner Šidák-adaptive safety bands.

5.7. *Marginal coverage (zoomed) and the joint-vs-marginal distinction*

Figure 3 shows horizon-wise marginal coverages (zoomed around the target region). For both assets, marginal coverage is generally high (often close to 1.00) for both Global and Šidák safety bands. This is fully compatible with the joint path coverage results in Table 3: *high marginal coverage at each horizon does not imply high joint path coverage across all horizons simultaneously*. We therefore prioritize joint path coverage in the main performance summary. Table A1 further shows that calibration dependence handling can affect joint path coverage much more strongly than horizon-wise summaries might suggest.

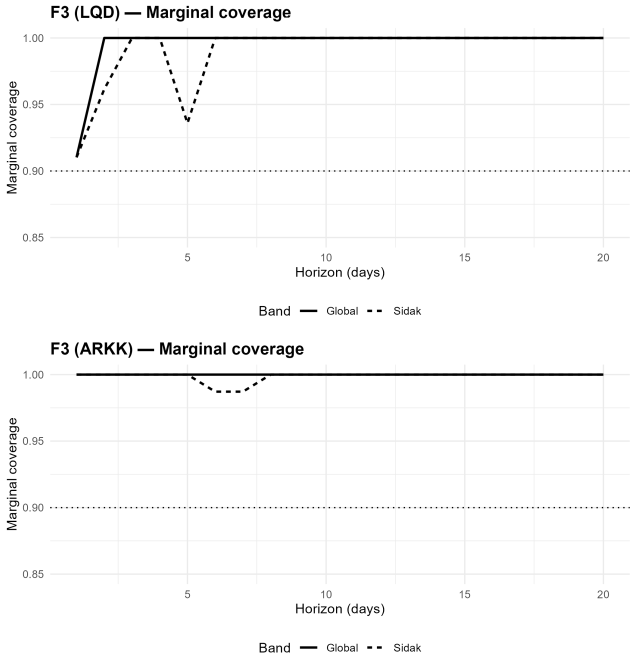

Figure 3. Zoomed horizon-wise marginal coverage for Global vs. Šidák safety bands under the main specification. The dotted line at 0.90 is shown as a visual reference (the nominal *joint-path* target level in the main specification), while the main reliability criterion in the text remains *joint path coverage*.

5.8. Sentiment ablation (H=20)

Table 4 compares the baseline and sentiment-augmented variants under the same block-calibrated design. The corresponding horizon-grid patterns ($H = 10, 20, 30$) are reported in Table A3. The ablation is asset-dependent. For LQD, the sentiment variant is effectively identical to the baseline (same RMSE and coverage up to rounding), suggesting little incremental signal in this specification. For ARKK, the sentiment variant is worse in point accuracy and more capital-intensive (higher AEB and larger normalized global conformal multiplier), while also producing very high coverage. In this sample, the added sentiment signal appears to increase conservatism without improving forecast quality for ARKK. For this reason, we retain the **baseline** variant as the main specification.

Table 4. Sentiment ablation at $H = 20$: baseline vs. sentiment (block calibration).

Asset	Variant	H	$\text{RMSE}_{1:5}$	$\text{RMSE}_{1:H}$	Global cov.	Šidák cov.	AEB (Global)	q_{glob} (norm.)
ARKK	baseline	20	0.00424	0.034641	1	0.987179	0.069286	2.949709
ARKK	sentiment	20	0.052749	0.101894	1	1	0.321641	9.672557
LQD	baseline	20	0.002554	0.014373	0.910256	0.897436	0.006213	0.803752
LQD	sentiment	20	0.002554	0.014373	0.910256	0.897436	0.006213	0.803752

6. Conclusion and future work

This paper develops a path-level framework for forward ES forecasting that combines (i) direct multi-horizon prediction, (ii) isotonic shape enforcement for economically coherent trajectories, and (iii) one-sided conformal calibration for uncertainty quantification at the *joint path* level. Rather than reporting isolated horizon-specific tail-risk values, the method produces a forward ES curve together with pathwise safety bands that can be interpreted as a horizon-indexed capital-buffer map.

In the empirical analysis, we focus on two representative ETFs with clearly different risk profiles (LQD and ARKK) under a leak-safe, strictly chronological design. The main specification uses $H = 20$, label-end splitting with an H-embargo, and dependence-aware block calibration. Within this setup, four findings stand out. **First**, the monotone post-processing step yields nondecreasing ES paths by construction, improving interpretability and preventing economically implausible horizon-order reversals. This shape control is simple to implement and does not require changing the base predictor architecture. **Second**, joint path coverage and marginal (horizon-wise) coverage should be interpreted separately. In our results, marginal coverages are often high for both global and Šidák-style safety bands, yet joint path coverage can differ materially across calibration choices and band constructions. This reinforces that path-level planning should be evaluated using joint criteria rather than horizon-wise summaries alone. **Third**, calibration design matters at least as much as point prediction for reliability. The appendix sensitivity analysis shows that block calibration can materially change global-band joint path coverage relative to day-wise calibration (especially for LQD), which is consistent with the role of serial dependence in pathwise residual maxima. We therefore view dependence-aware calibration as a substantive component of the pipeline, not a cosmetic add-on. **Fourth**, the global scalar-surcharge safety band is especially attractive from an implementation and governance perspective. Because it maps path uncertainty into a single one-sided surcharge, it is easy to document, monitor, and audit in routine risk-control workflows. In this sense, the framework is not only statistically motivated but also operationally *audit-friendly*.

The comparison between global and Šidák-adaptive safety bands highlights a practical trade-off. In our test samples, the Šidák allocation can reduce AEB, sometimes substantially, while joint path coverage may soften slightly in finite samples. We therefore do not treat adaptive allocation as uniformly superior; instead, it is best viewed as a capital-efficiency option that should be checked empirically against joint path coverage targets. Relatedly, any claim of capital savings should be assessed relative to standard error-budget baselines (including Bonferroni-style allocations) under matched joint-coverage targets, since finite-sample order-statistic calibration may attenuate theoretical differences in practice. Likewise, the sentiment ablation illustrates that added predictors need not improve performance: for LQD the sentiment variant is effectively neutral, while for ARKK it increases conservatism and point error without clear gains in forecast quality.

Several limitations should be kept in view. Conformal guarantees are exact only under exchangeability assumptions; in time-series settings with dependence, regime changes, and structural breaks, block calibration is a mitigation strategy rather than a complete solution. The isotonic constraint improves economic coherence but can be restrictive if an asset's true forward ES profile exhibits local flattening or non-monotone behavior over the forecast horizon. The feature set used here is intentionally lightweight and stability-oriented, which supports transparency and auditability but may miss richer nonlinear or cross-asset dynamics. In addition, rolling nonparametric ES target estimation can be data-hungry and potentially unstable for newer or less liquid assets, and the use of dedicated calibration windows and guard bands reduces the amount of data available for model fitting in rapidly evolving markets. These limitations motivate several natural extensions within finance. Future work should benchmark the same leak-safe, shape-preserving conformal wrapper against richer base learners (e.g., GARCH-family scaling models, quantile/tree-based learners, and neural sequence models) and broader feature sets (e.g., options-implied volatility, macro/market-state variables, and cross-asset spillover signals). It is also important to study regime-aware or online conformal updates, multi-asset joint calibration, and applications to less liquid assets, cryptocurrencies, and derivative portfolios, where dependence, nonstationarity, and tail estimation noise are more severe. More broadly, the underlying design principle of this paper—trajectory-level reliability under resource (capital) efficiency constraints—may have value beyond financial risk forecasting. For example, path-level optimization and uncertainty-aware buffer design arise in infrastructure planning and energy-system operations, including long-horizon allocation of electrolysers in transmission planning [30], integrated grid-impact management for e-mobility deployment [31], coordinated aggregation of distributed flexibility resources [32], and risk-aware optimization in supply-chain network design using VaR/CVaR-type criteria [33]. We view these as promising cross-domain directions for translating joint-coverage and path-calibration ideas into other sequential decision settings.

Overall, the proposed framework offers a transparent and modular path from single-horizon tail-risk reporting to trajectory-level ES analysis. Its main contribution is not a claim of universal dominance, but a practical design that makes shape constraints, calibration choices, capital-efficiency trade-offs, and finite-sample reliability limitations explicit for empirical risk monitoring.

Use of Generative-AI tools declaration

The author used a generative AI tool only for language polishing.

Conflict of interest

The author declares no conflicts of interest in this paper.

References

1. A. J. McNeil, R. Frey, P. Embrechts, *Quantitative risk management: Concepts, techniques and tools*, Princeton University Press, 2015.
2. P. Embrechts, M. Hofert, Statistics and quantitative risk management for banking and insurance, *Annu. Rev. Stat. Appl.*, **1** (2014), 493–514. https://doi.org/10.1146/annurev-statistics-022513-115631
3. J. Danielsson, *The illusion of control: Why financial crises happen, and what we can (and can't) do about it*, Yale University Press, 2022.
4. M. Marcellino, J. H. Stock, M. W. Watson, A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series, *J. Econ.*, **135** (2006), 499–526. https://doi.org/10.1016/j.jeconom.2005.07.020

5. E. Ghysels, A. Sinko, R. Valkanov, MIDAS regressions: Further results and new directions, *Econ. Rev.*, **26** (2007), 53–90. https://doi.org/10.1080/07474930600972467
6. E. Ghysels, M. Marcellino, *Applied economic forecasting using time series methods*, Oxford University Press, 2018.
7. J. Baruník, M. Nevrla, Quantile spectral beta: A tale of tail risks, investment horizons, and asset prices, *J. Financ. Econ.*, **21** (2023), 1590–1646. https://doi.org/10.1093/jjfinec/nbac017
8. V. Vovk, A. Gammerman, G. Shafer, *Algorithmic learning in a random world*, Springer, 2005.
9. J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, L. Wasserman, Distribution-free predictive inference for regression, *J. Am. Stat. Assoc.*, **113** (2018), 1094–1111. https://doi.org/10.1080/01621459.2017.1307116
10. G. Shafer, V. Vovk, A tutorial on conformal prediction, *J. Mach. Learn. Res.*, **9** (2008), 371–421.
11. A. N. Angelopoulos, S. Bates, Conformal prediction: A gentle introduction, *Found. Trends Mach. Le.*, **16** (2023), 494–591. https://doi.org/10.1561/2200000101
12. R. F. Engle, Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation, *Econometrica*, **50** (1982), 987–1007. https://doi.org/10.2307/1912773
13. T. Bollerslev, Generalized autoregressive conditional heteroskedasticity, *J. Econ.*, **31** (1986), 307–327. https://doi.org/10.1016/0304-4076(86)90063-1
14. T. G. Andersen, T. Bollerslev, F. X. Diebold, P. Labys, Modeling and forecasting realized volatility, *Econometrica*, **71** (2003), 579–625. https://doi.org/10.1111/1468-0262.00418
15. P. R. Hansen, Z. Huang, H. H. Shek, Realized GARCH: A joint model for returns and realized measures of volatility, *J. Appl. Economet.*, **27** (2012), 877–906. https://doi.org/10.1002/jae.1234
16. R. F. Engle, S. Manganelli, CAViaR: Conditional autoregressive value at risk by regression quantiles, *J. Bus. Econ. Stat.*, **22** (2004), 367–381. https://doi.org/10.1198/073500104000000370
17. C. Acerbi, D. Tasche, Expected shortfall: A natural coherent alternative to value at risk, *Econ. Notes*, **31** (2002), 379–388. https://doi.org/10.1111/1468-0300.00091
18. T. Fissler, J. F. Ziegel, Higher order elicitability and Osband's principle, *Ann. Statist.*, **44** (2016), 1680–1707. https://doi.org/10.1214/16-AOS1439
19. F. X. Diebold, C. Li, Forecasting the term structure of government bond yields, *J. Economet.*, **130** (2006), 337–364. https://doi.org/10.1016/j.jeconom.2005.03.005
20. Y. Romano, E. Patterson, E. J. Candès, Conformalized quantile regression, *Adv. Neural Inf. Process. Syst.*, **32** (2019), 3543–3553.
21. C. Xu, Y. Xie, Conformal prediction interval for dynamic time-series, In: *Proceedings of the 38th International Conference on Machine Learning*, **139** (2021), 11559–11569.
22. Ç. Sözen, F. Kabakcı, Forecasting future realized variance paths with depth-weighted ridge and conformal diagnostics, *AIMS Mathematics*, **10** (2025), 30246–30270. https://doi.org/10.3934/math.20251329
23. Ç. Sözen, Uniform one-sided conformal bands for forward realized volatility curves, *AIMS Mathematics*, **10** (2025), 27314–27337. https://doi.org/10.3934/math.20251201

24. Yahoo Finance, Historical market data for equities and ETFs (Adjusted Close series), 2025.
25. R. E. Barlow, D. J. Bartholomew, J. M. Bremner, H. D. Brunk, *Statistical inference under order restrictions: The theory and application of isotonic regression*, John Wiley & Sons, London, 1972.
26. T. Hastie, R. Tibshirani, J. Friedman, *The elements of statistical learning: Data mining, inference, and prediction*, Springer, 2009. https://doi.org/10.1007/978-0-387-84858-7
27. M. J. Best, N. Chakravarti, Active set algorithms for isotonic regression: A unifying framework, *Math. Program.*, **47** (1990), 425–439. https://doi.org/10.1007/BF01580873
28. T. Loughran, B. McDonald, When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks, *J. Financ.*, **66** (2011), 35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x
29. Z. Šidák, Rectangular confidence regions for the means of multivariate normal distributions, *J. Am. Stat. Assoc.*, **62** (1967), 626–633. https://doi.org/10.1080/01621459.1967.10482935
30. S. Giannelos, I. Konstantelos, D. Pudjianto, G. Strbac, The impact of electrolyser allocation on Great Britain's electricity transmission system in 2050, *Int. J. Hydrogen Energ.*, **202** (2026), 153097. https://doi.org/10.1016/j.ijhydene.2025.153097
31. European Commission, *E-mobility deployment and impact on grids: Impact of EV and charging infrastructure on European T&D grids: innovation needs*, 2022. https://doi.org/10.2833/937755
32. Z. Dong, X. Zhang, L. Zhang, S. Giannelos, G. Strbac, Flexibility enhancement of urban energy systems through coordinated space heating aggregation of numerous buildings, *Appl. Energ.*, **374** (2024), 123971. https://doi.org/10.1016/j.apenergy.2024.123971
33. S. Giannelos, I. Konstantelos, G. Strbac, Optimal supply chain design using machine learning, risk assessment and optimisation applied to coal distribution, *EURO J. Decis. Processes*, **13** (2025), 100062. https://doi.org/10.1016/j.ejdp.2025.100062

A. Appendix: Additional empirical diagnostics

This appendix focuses on additional diagnostics for the two-asset design (LQD, ARKK), complementing the main text with calibration-scheme, quantile-implementation, and horizon/variant sensitivity analyses.

Tables A1–A3 complement the main-text results in Tables 3 and 4 and Figures 2 and 3 by documenting (i) calibration-scheme sensitivity, (ii) finite-sample quantile-implementation sensitivity, and (iii) horizon/variant robustness.

A.1. Block-calibration sensitivity at the main horizon (H = 20)

Table A1 compares block calibration (TRUE) versus day-wise calibration (FALSE) for the baseline variant at $H = 20$. The sensitivity is economically meaningful for the *global* safety band, especially for LQD (joint path coverage rises from 0.526 to 0.910 with block calibration). For ARKK, the Šidák and Bonferroni joint path coverages are unchanged in this comparison, while the global safety band becomes more conservative under block calibration. Overall, the pattern is consistent with dependence in calibration residual maxima affecting the global quantile more strongly than horizon-wise marginal quantiles. We therefore keep block calibration in the main specification.

Table A1. Block-calibration sensitivity at $H = 20$ (baseline variant).

		Forecast accuracy		Global band		Joint path cov.		Joint path cov.
Asset	Calibration	$\text{RMSE}_{1:5}$	$\text{RMSE}_{1:H}$	AEB (Global)	q_{glob}	Global	Šidák	Bonf.
LQD	Block	0.0026	0.0144	0.0062	0.8038	0.910	0.897	0.897
LQD	Day-wise	0.0026	0.0144	0.0042	0.5376	0.526	0.962	0.962
ARKK	Block	0.0042	0.0346	0.0693	2.9497	1.000	0.987	0.987
ARKK	Day-wise	0.0042	0.0346	0.0675	2.8724	0.769	0.987	0.987

Table A2. Šidák vs. Bonferroni under block = FALSE, with exact OS calibration and interpolated-quantile sensitivity (type 8).

		OS quantiles					
Asset	H	AEB Sid.	AEB Bonf.	Saved % (all)	$\text{AEB}_{1:5}$ Sid.	$\text{AEB}_{1:5}$ Bonf.	Saved % (1:5)
ARKK	10	0.0174	0.0174	0	0.0029	0.0029	0
LQD	10	0.0005	0.0005	0	0.0010	0.0010	0
ARKK	20	0.0330	0.0330	0	0.0056	0.0056	0
LQD	20	0.0003	0.0003	0	0.0014	0.0014	0
ARKK	30	0.0312	0.0312	0	0.0104	0.0104	0
LQD	30	0.0003	0.0003	0	0.0019	0.0019	0
		Interpolated quantiles (type 8)					
Asset	H	AEB Sid.	AEB Bonf.	Saved % (all)	$\text{AEB}_{1:5}$ Sid.	$\text{AEB}_{1:5}$ Bonf.	Saved % (1:5)
ARKK	10	0.0172	0.0173	0.09	0.0028	0.0028	0.61
LQD	10	0.0004	0.0004	1.63	0.0008	0.0008	1.63
ARKK	20	0.0328	0.0329	0.11	0.0056	0.0056	0.11
LQD	20	0.0003	0.0003	0.41	0.0013	0.0013	0.41
ARKK	30	0.0312	0.0312	0	0.0104	0.0104	0
LQD	30	0.0003	0.0003	0	0.0019	0.0019	0
		Per-horizon levels					
Asset	H	α_h^{Sid}	α_h^{Bonf}	$1-\alpha_h^{\text{Sid}}$	$1-\alpha_h^{\text{Bonf}}$		
ARKK	10	0.0105	0.0100	0.9895	0.9900		
LQD	10	0.0105	0.0100	0.9895	0.9900		
ARKK	20	0.0053	0.0050	0.9947	0.9950		
LQD	20	0.0053	0.0050	0.9947	0.9950		
ARKK	30	0.0035	0.0033	0.9965	0.9967		
LQD	30	0.0035	0.0033	0.9965	0.9967		

A.2. Šidák vs. Bonferroni in finite samples: OS vs. interpolated quantiles

Table A2 reports a finite-sample sensitivity check comparing Šidák and Bonferroni under block = FALSE, using both exact OS and interpolated quantiles (type 8). With OS quantiles, Šidák and Bonferroni often coincide numerically because both map to the same order statistic in finite calibration samples. Small differences appear only under interpolated quantiles, and these differences are economically minor in our setting. This explains why several rows in the main and appendix results can show identical Šidák- and Bonferroni-style outcomes despite the theoretical distinction.

A.3. Full grid over horizon and variants

Table A3 provides the full grid over $H \in \{10, 20, 30\}$ and baseline vs. sentiment variants for the two selected assets under block calibration. The main qualitative conclusions remain stable: (i) LQD is well-behaved and largely insensitive to the sentiment augmentation; and (ii) ARKK exhibits much higher ES scale and stronger sensitivity to the sentiment variant, which increases both point error and conformal-buffer size without clear predictive gains. The $H = 20$ rows in Table A3 correspond to the main-text comparisons in Tables 3 and 4.

Table A3. Appendix: full grid for the two selected assets ($H = 10, 20, 30$; baseline vs. sentiment; block calibration). "Over-pen." denotes the code-level asymmetric over-prediction penalty coefficient (η), reported for reproducibility.

		Forecast accuracy		Joint path coverage			Global band		Repro.
Asset	Variant–H	$\text{RMSE}_{1:5}$	$\text{RMSE}_{1:H}$	Global	Naive	Šidák	AEB	q_{glob}	Over-pen.
LQD	baseline, 10	0.0029	0.0078	0.900	0.880	0.880	0.0049	0.6874	0.0000
LQD	sentiment, 10	0.0029	0.0078	0.900	0.880	0.880	0.0049	0.6874	0.0000
LQD	baseline, 20	0.0026	0.0144	0.910	0.897	0.897	0.0062	0.8038	0.0000
LQD	sentiment, 20	0.0026	0.0144	0.910	0.897	0.897	0.0062	0.8038	0.0000
LQD	baseline, 30	0.0023	0.0183	0.964	0.964	0.964	0.0129	1.1248	0.0143
LQD	sentiment, 30	0.0023	0.0183	0.964	0.964	0.964	0.0129	1.1248	0.0143
ARKK	baseline, 10	0.0042	0.0256	0.920	0.900	0.900	0.0439	1.5538	0.0000
ARKK	sentiment, 10	0.0540	0.0756	1.000	1.000	1.000	0.2485	6.4301	0.0500
ARKK	baseline, 20	0.0042	0.0346	1.000	0.987	0.987	0.0693	2.9497	0.0500
ARKK	sentiment, 20	0.0527	0.1019	1.000	1.000	1.000	0.3216	9.6726	0.0500
ARKK	baseline, 30	0.0061	0.0325	0.946	0.946	0.946	0.0933	3.1903	0.0000
ARKK	sentiment, 30	0.0567	0.1197	1.000	1.000	1.000	0.4383	11.9184	0.0500