mathematics

MDPI

Article

Finite-Sample Conformal Risk Bounds for Joint Value-at-Risk and Expected-Shortfall Forecasting Under Non-Exchangeable Financial Time Series

Yuxin Ye [1], Xuhua Qiu [2], Kunjie Zhu [2,*] and Miltos Ladikas [3,*]

[1] Bangor College, Central South University of Forestry and Technology, Changsha 410004, China; 20237679@csuft.edu.cn
[2] Management Science and Engineering, Chinese Academy of Sciences, Beijing 100190, China; qiuxuhua23@mails.ucas.ac.cn
[3] Institute for Technology Assessment and Systems Analysis, Karlsruhe Institute of Technology, Karlstrasse 11, 76133 Karlsruhe, Germany
* Correspondence: zhukunjie24@mails.ucas.ac.cn (K.Z.); miltos.ladikas@kit.edu (M.L.)

Abstract

Financial tail-risk observations are non-exchangeable: serial dependence and regime shifts make their joint law depend on the time ordering, invalidating the exchangeability that standard conformal guarantees assume, and expected shortfall is not elicitable on its own, so a forecaster cannot be calibrated to it as a quantile is to its coverage. We ask whether a black-box value-at-risk and expected-shortfall forecaster can be calibrated under such dependence while retaining finite-sample guarantees. We tune a single inflation parameter by conformal risk control on a bounded monotone loss that couples value-at-risk breach frequency with breach magnitude normalised by the model's predicted value-at-risk–expected-shortfall gap; the guarantee is thus for a tail-gap-normalised exceedance-severity surrogate, and its expected-shortfall reading depends on the predicted gap being a sound tail-gap estimate. Under exchangeability, the method gives finite-sample expected-risk control; for dependent data we invoke a non-exchangeable swap-distance bound and add, for separated calibration points, a regime-drift bound with an explicit cumulative β-mixing cost, plus a high-probability realised-path statement and a heavy-tail rate of order $D^{(p-1)/p}$. Building regimes causally from previous-month FRED-MD vintages across eight exchange rates, a Bitcoin series, and the GIFT-Eval finance domain, the weighted controller attains a 2.51% violation rate and a Fissler–Ziegel score of 0.431 against 0.441 and 0.439 for the strongest conformal baselines—an incremental gain, not significant at the 5% level, that concentrates in turbulent regimes and at matched capital, supporting calibration of a joint frequency-and-normalised-severity budget rather than distribution-free control of the expected-shortfall forecast itself.

Keywords: expected shortfall; value-at-risk; conformal risk control; distribution-free inference; time series forecasting; tail risk; financial risk management; financial econometrics; elicitability; regime drift; backtesting; mixing sequences

MSC: 62G15; 62M10; 91G70; 68T07; 60G10

Academic Editor: Anli Ji

Received: 6 July 2026
Revised: 28 July 2026
Accepted: 4 August 2026
Published: 6 August 2026

1. Introduction

Market-risk capital for banks that use internal models is set by the expected shortfall at the 97.5% level over a stressed horizon, the pivotal quantity of modern financial risk

management. The 2019 revision of the standard, published as *BCBS document d457*, retired the 99% value-at-risk that had anchored capital since the 1990s, and jurisdictions including the European Union and the United Kingdom target 1 January 2027 for application, with the internal-models approach that this paper concerns phased in by 1 January 2028 in the United Kingdom [1]. The change was not cosmetic. Value-at-risk reports a quantile of the loss and ignores everything beyond it, and it fails to be subadditive, so a merger of two books can appear riskier than the two books apart [2]. Expected shortfall averages the loss in the tail beyond the quantile and is coherent under mild conditions [3]. A forecasting method aimed at the modern capital regime must therefore produce a credible expected-shortfall number, not only a quantile, and it must do so on financial series whose distribution moves under the forecaster's feet.

The obstacle is statistical, and it is sharper than it first appears. A quantile forecast can be judged by how often the loss exceeds it, and that frequency can be driven to a nominal level by tuning a single threshold, which is exactly what split conformal prediction and its adaptive descendants accomplish with finite-sample guarantees [4,5]. Expected shortfall admits no such handle. It is not elicitable: no scoring function is minimised in expectation by the true expected shortfall alone [6]. Reporting an interval and asking what fraction of losses fall inside it, the reflex that works for value-at-risk, is meaningless for a tail mean. Any method that promises to calibrate the coverage of expected shortfall has mis-specified its own target. What rescues the problem is a result of [7]: the pair consisting of value-at-risk and expected shortfall is jointly elicitable, with an explicit family of strictly consistent scores. Calibration and evaluation must both act on the pair.

Three notions of guarantee recur below, and we fix their meaning at the outset. A *finite-sample* statement holds for a fixed calibration size without an asymptotic limit; under exchangeability it carries no drift or dependence slack, but it remains a risk inequality rather than an exact equality. A *non-exchangeable finite-sample* bound adds an explicit, generally non-sharp slack term measuring departure from exchangeability. A *high-probability* statement instead concerns the realised average along a single test path and holds outside an event of explicitly bounded probability.

We take the pair as the object and turn calibration into risk control. A base network outputs a conditional quantile function, from which the value-at-risk and the expected shortfall at level τ follow by evaluation and by tail integration. A single scalar inflates both forecasts in parallel, and a bounded loss counts a breach of the inflated value-at-risk and charges the size of that breach measured against the model's own tail gap. This loss decreases as the inflation grows, so the machinery of conformal risk control [8] selects the smallest inflation whose expected loss sits below a chosen budget. The loss controls how often the inflated value-at-risk is breached and how deep those breaches run, measured in units of the base model's predicted value-at-risk–expected-shortfall gap. It does not directly score the expected-shortfall forecast against a realised tail outcome, and the gap is unchanged by the parallel inflation. The resulting guarantee is therefore a finite-sample statement for a tail-gap-normalised exceedance-severity surrogate. Its reading as evidence about expected shortfall is conditional on the predicted gap being a reliable estimate of the conditional mean-excess scale, a dependence we assess empirically below.

The exchangeable guarantee is not enough for returns, which are dependent and non-stationary. Exchangeability here concerns the time-indexed observations rather than the labelling of assets: a sequence is exchangeable when its joint law is invariant under every permutation of the time indices, and serial dependence or a changing marginal breaks that invariance. A calm quarter followed by a crisis month is the canonical violation, because a random reordering would scatter the crisis days through the calm period and erase the clustering the data actually display (Figure 1). Our theoretical development therefore

combines three layers rather than a single new device. Non-exchangeable conformal risk control [9], which adapts the beyond-exchangeability swap argument of [10] to a bounded monotone loss, shows that the expected tail loss of the selected forecaster exceeds the target by at most a weighted sum of total-variation distances between the data law and its index-swapped versions; we adopt that bound and then read that sum through a mixing-and-drift lens. When the calibration points are separated in time and carry a regime label, the excess is bounded by the frequency of regime disagreement between calibration and test times the total-variation distance between regime-conditional laws, plus an explicit cumulative β-mixing cost incurred by coupling the separated coordinates to independent copies. These dependence quantities are interpretable sensitivity parameters that we diagnose, not distribution-free certificates recoverable from a single path. A coupling and blocking argument for β-mixing sequences [11,12] upgrades the in-expectation statement to a high-probability one. Because financial losses are heavy-tailed while the risk-control theory needs a bounded loss, we clip the severity term and pay a tail remainder; balancing the clip level against the drift budget gives an excess of order $D^{(p-1)/p}$ under a p-th moment assumption, where D collects the drift and mixing terms. A projected online variant adjusts the inflation through a negative-feedback conformal recursion and provides a deterministic long-run tracking bound under arbitrary distribution shifts, provided that the target remains attainable within the projection interval and the cumulative saturation residual is sublinear in the horizon [5,13]. The genuinely new mathematical element is thus not these generic proof devices but the separated-calibration corollary that renders the dependence cost cumulative and explicit, together with the heavy-tail truncation transfer for the proposed financial loss.

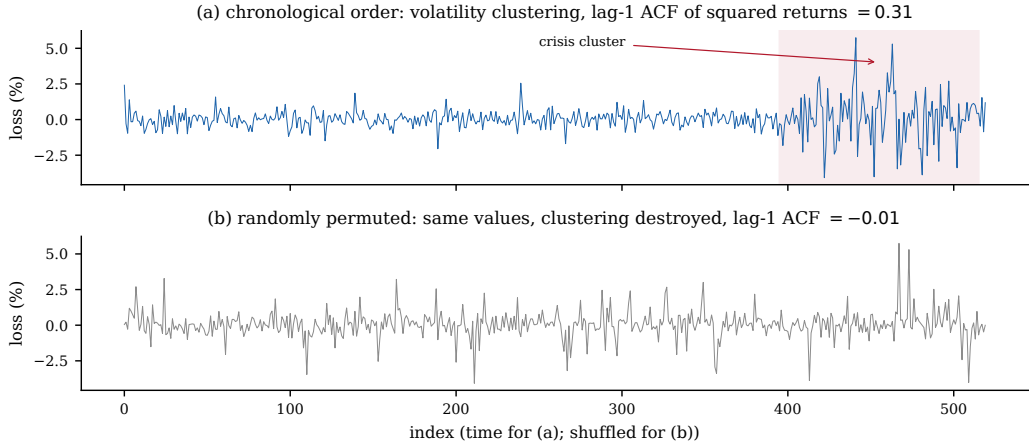

Figure 1. Non-exchangeability of time-indexed losses. (**a**) In chronological order a crisis volatility cluster (shaded) produces a lag-one autocorrelation of squared returns of about 0.31. (**b**) A random permutation of the identical values destroys the clustering and drives that autocorrelation to zero, so the joint law is not permutation-invariant even though the marginal histogram is unchanged.

This study pursues three objectives. The first is to construct a bounded monotone calibration loss that responds to both value-at-risk breach frequency and tail magnitude while remaining compatible with conformal risk control. The second is to quantify how the resulting expected-risk guarantee degrades under temporal dependence and regime drift, with every dependence qualification made explicit. The third is to determine whether the added severity and regime components deliver statistically and economically meaningful improvements over strong conformal and joint value-at-risk–expected-shortfall baselines under a causal rolling-origin protocol.

The empirical study is designed to isolate calibration from prediction. On eight daily exchange-rate series and a Bitcoin series with on-chain covariates, with the 107 FRED-MD

macro series entering only as exogenous regime signals and the finance domain of GIFT-Eval used for breadth [14–16], the same base forecaster is calibrated by every competing scheme, so differences trace to the calibration layer alone. Against the adaptive-conformal baselines the joint scheme lowers the Fissler–Ziegel score by roughly two percent, a gain that is incremental and, against the weighted non-exchangeable scheme, not significant at the five-percent level, while a larger reduction near four percent appears in turbulent regimes. It holds the Acerbi–Székely expected-shortfall statistic within 0.05 of zero and reaches a 2.51% violation rate against the 2.50% nominal, without buying that calibration through excess conservativeness. The contributions are three:

1. A loss and controller design: A bounded monotone tail-gap-normalised severity loss that instantiates conformal risk control for a value-at-risk–expected-shortfall forecaster, respecting the non-elicitability of expected shortfall while controlling a joint frequency-and-normalised-severity budget;
2. A corrected non-exchangeable analysis in which the exchangeable oracle bound specialises conformal risk control and the weighted swap bound follows the non-exchangeable conformal risk control of [9], while the new element is a separated-calibration corollary carrying an explicit cumulative β-mixing cost, together with a heavy-tail truncation transfer that yields the $D^{(p-1)/p}$ rate;
3. A causal empirical implementation combining a previous-month-vintage regime construction, a projected online calibrator, and a diagnostic suite—tail-index, tail-gap, mixing-envelope, and effect-size analyses—across four public data sources with a full value-at-risk and expected-shortfall backtesting battery.

The remainder develops the relevant strands, the formal setting, the method, its theory, and the experiments and their limitations.

2. Related Work

2.1. Coherent Risk, Elicitability, and Backtesting

The theory of coherent risk measures established subadditivity as a requirement a sound risk measure should meet and exposed value-at-risk as failing it [2], while expected shortfall was shown to satisfy coherence under continuity of the loss [3]. The turn to expected shortfall in prudential regulation [1] then collided with a decision-theoretic fact: a functional is elicitable only if some scoring function identifies it as the unique minimiser of expected score, and expected shortfall is not elicitable [6]. Fissler and Ziegel [7] resolved the impasse by proving that the pair of value-at-risk and expected shortfall is jointly elicitable and by characterising the consistent scores through Osband's principle; the strictly consistent scoring rules of [17] supply the wider backdrop. Nolde and Ziegel [18] translated joint elicitability into comparative backtesting for banking supervision, and Acerbi and Székely [19] gave the exceedance-based tests now standard for expected shortfall. Traditional coverage tests for value-at-risk [20,21] remain necessary but address only the quantile. Our loss is designed for calibration and is deliberately distinct from these scores, which we reserve for evaluation; the separation keeps the calibration monotone while letting evaluation use a strictly consistent joint score.

2.2. Parametric and Semiparametric Tail-Risk Forecasting

Conditional quantile dynamics were first modelled directly by the CAViaR family, which sidesteps a full return distribution [22]. Expected shortfall resisted a comparable direct treatment until the joint-elicitability result enabled regression: Patton et al. [23] fit dynamic semiparametric models for the value-at-risk and expected-shortfall pair by minimising a Fissler–Ziegel score, and the score-driven recursions they introduce are among our strongest baselines. These models assume a parametric score dynamic. When

that dynamic is mis-specified, as it is under abrupt regime change, their guarantees are asymptotic and conditional on the model. Robust optimisation offers a different route to tail control through conditional value-at-risk minimisation [24], which acts on the decision rather than the forecast. Our calibration layer is distribution-free and rides on top of any such model, so it can correct a mis-specified parametric forecaster rather than replace it.

2.3. Deep Probabilistic Forecasting

Transformer and mixing architectures now lead multivariate forecasting benchmarks. Channel-independent patching [25], variate tokenisation [26], explicit exogenous-variable pathways [27], and frequency-domain Kolmogorov–Arnold blocks [28] improve point and quantile accuracy. Deep generative and hybrid learning pipelines extend such models to adjacent financial and operational risk tasks, from credit-card risk identification [29] to retail demand forecasting [30]. Accuracy is not calibration. A network trained by a pinball loss produces sharp quantiles whose realised coverage still drifts when the data-generating process moves, and none of these models carries a finite-sample tail-risk guarantee. We use them as interchangeable base forecasters and attach the guarantee externally, which lets the empirical study attribute differences to the calibration scheme rather than the backbone.

2.4. Conformal Prediction and Risk Control

Conformal prediction converts any predictor into finite-sample distribution-free coverage under exchangeability [4,31]. Two extensions matter here. Conformal risk control generalises the coverage guarantee to the expectation of any bounded monotone loss [8], which is the vehicle we need because expected shortfall is a loss functional, not a coverage event. The exchangeability assumption is relaxed by weighting and by the swap analysis of [10], and by covariate-shift reweighting [32]. Closest to the present work, Farinhas et al. [9] extend conformal risk control itself beyond exchangeability: for any bounded monotone loss under weighted split calibration they bound the expected loss by its target plus a weighted sum of total-variation distances between the data law and its index-swapped versions. That result is the direct precursor of our weighted master bound (Theorem 3); it is stated for a generic loss and does not treat the joint value-at-risk–expected-shortfall pair, regime-conditional drift, cumulative β-mixing, a causal regime label, or heavy-tailed severity. For sequential data, adaptive conformal inference tracks a moving distribution by an online update [5], its aggregated form removes the learning-rate choice [33], and a control-theoretic version adds integral and derivative action [13]. These methods target quantile coverage. Building on non-exchangeable conformal risk control [9], we specialise the risk-control formulation to the joint value-at-risk and expected-shortfall pair under temporal dependence, adding a regime-drift and cumulative-β-mixing decomposition, a high-probability realised-path bound, a heavy-tail-stable loss, and a causal regime construction that none of these works provides.

2.5. Weak Dependence and Concentration

The move from exchangeable to time series data rests on quantitative mixing. Strong-mixing coefficients and their basic properties are surveyed by [34]; empirical-process rates for stationary β-mixing sequences are due to [11]; and the coupling and blocking technology that reduces a dependent block to an independent surrogate at a cost measured by the mixing coefficient is developed at length by [12]. We use this machinery to turn the in-expectation risk bound into a high-probability statement, which is what a supervisor auditing a single realised path actually needs.

2.6. Research Gap, Method Choice, and Novelty

Existing approaches leave three gaps when a joint value-at-risk and expected-shortfall forecast is deployed on dependent financial series. First, conformal methods for time series predominantly target quantile coverage, so they do not penalise the magnitude of losses beyond the value-at-risk threshold. Second, joint regression on Fissler–Ziegel scores evaluates or estimates the pair, but its validity is model-dependent and generally asymptotic under dynamic mis-specification. Third, generic conformal risk control gives an exchangeable finite-sample expectation bound, while beyond-exchangeability conformal theory is formulated mainly for coverage and does not by itself specify a causal regime construction or a heavy-tail-stable loss for tail severity. This work addresses that intersection—post hoc bounded risk calibration for a value-at-risk–expected-shortfall forecaster, explicit non-exchangeability slack, and a causal regime-aware implementation; it does not claim to make expected shortfall directly elicitable or to bound expected-shortfall forecast error in a distribution-free sense. Table 1 places the method families against these requirements.

Table 1. Positioning of method families against the requirements of joint value-at-risk and expected-shortfall forecasting on dependent series.

Method Family	Joint VaR–ES	Tail Magnitude	Finite-Sample	Dependence	Black-Box
VaR conformal/ACI/PID	No	No	Long-run	Partial	Yes
FZ regression/dynamic ES	Yes	Yes	Asymptotic	Parametric	No
Generic CRC	Loss-dependent	If designed	Exchangeable	Limited	Yes
Beyond-exchangeability CP	Coverage	No	With TV slack	Yes	Yes
Non-exch. CRC [9]	Loss-dep.	If designed	With TV slack	Yes	Yes
JCRC (this work)	Surrogate	Norm. severity	Yes, qualified	Swap + blocked	Yes

Conformal risk control is the natural vehicle because the calibration object is an expected loss, not a coverage event. Unlike refitting a parametric value-at-risk–expected-shortfall model, it is post hoc and treats the forecaster as a black box; unlike value-at-risk conformal calibration, it can control a non-binary severity loss; and unlike direct minimisation of a Fissler–Ziegel score over the inflation, the proposed loss is monotone in the inflation, which the finite-sample selection argument requires. The trade-off, made explicit throughout, is that it controls the chosen surrogate rather than the expected-shortfall forecast error itself. Table 2 summarises the comparison.

Table 2. Why conformal risk control is used for the calibration layer. "DF" denotes distribution-free.

Approach	Refit Base	Finite-Sample DF	Controlled Object	Limitation Here
FZ regression	Yes	No	joint scoring risk	asymptotic under mis-specification
VaR conformal	No	Yes	breach frequency	ignores tail magnitude
DRO/CVaR	Often	Model-dependent	downstream decision	not forecast calibration
CRC (ours)	No	Exchangeable	bounded monotone loss	guarantee tied to loss interpretation

The mathematical provenance of each result is stated so that the reader can locate the novelty. The exchangeable oracle argument specialises conformal risk control; the swap-distance bound is the non-exchangeable conformal risk control of [9], itself an adaptation of the beyond-exchangeability comparison to a bounded loss; and the high-probability theorem uses standard β-mixing blocking. The genuinely new elements are the separated-calibration bound with a cumulative mixing cost, the heavy-tail truncation transfer for the proposed loss, and the causal regime-aware empirical framework. Table 3 records the mapping.

Table 3. Provenance of the theoretical results and the elements new to this paper.

Result	Closest Existing Machinery	New Content Here	Status
Theorem 1	conformal risk control oracle	instantiation to the bounded tail-severity surrogate	specialisation
Theorems 2 and 3	non-exchangeable CRC [9] on the beyond-exchangeability swap [10]	specialisation to the bounded tail-severity surrogate	specialisation
Proposition 1 (revised)	Berbee coupling and a regime model	separated-calibration cumulative mixing plus drift	new corollary
Theorem 4	β-mixing blocking	applied to the fixed deployed tail-risk loss	specialisation
Proposition 2/Corollary 1	truncation and moment tail integration	cap against non-exchangeability budget trade-off	derived rate
Loss and framework	no direct analogue	frequency and normalised severity with causal regime weighting	synthesis

3. Preliminaries and Assumptions

3.1. Risk Functionals and Elicitability

Let L be a real loss with distribution F and fix a level $\tau \in (0,1)$.

Definition 1 (Value-at-risk and expected shortfall)**.** *The value-at-risk and expected shortfall of L at level τ are*

$$\mathrm{VaR}_\tau(L) = F^{-1}(\tau) = \inf\{x : F(x) \geq \tau\}, \qquad \mathrm{ES}_\tau(L) = \frac{1}{1-\tau}\int_\tau^1 F^{-1}(u)\,du, \tag{1}$$

and $\mathrm{ES}_\tau(L) = \mathbb{E}[L \mid L \geq \mathrm{VaR}_\tau(L)]$ when F is continuous at $\mathrm{VaR}_\tau(L)$.

Definition 2 (Elicitability)**.** *A functional T of a distribution is elicitable if there is a scoring function S with $T(F) = \arg\min_x \mathbb{E}_{L\sim F}[S(x,L)]$ for all F in a class. Value-at-risk is elicitable through the pinball score; expected shortfall is not elicitable in isolation, but the pair $(\mathrm{VaR}_\tau, \mathrm{ES}_\tau)$ is jointly elicitable [6,7].*

Definition 2 is the reason the method calibrates and evaluates the pair rather than expected shortfall alone.

3.2. Data-Generating Process and Dependence

Let $\{(X_t, L_t)\}_{t\in\mathbb{Z}}$ be a stochastic process on $(\Omega, \mathcal{F}, \mathbb{P})$, with $X_t \in \mathcal{X}$ the conditioning features and $L_t = -R_t \in \mathbb{R}$ the loss. Write $\mathcal{F}_t = \sigma\{(X_s, L_s) : s \leq t\}$. A base forecaster, trained once on a block disjoint from calibration and test, is a fixed measurable map producing from $\mathcal{F}_{t-1}$ the quadruple $(\widehat{\mathrm{VaR}}_{\tau,t}, \widehat{\mathrm{ES}}_{\tau,t}, \widehat{s}_t, \widehat{g}_t)$, and we collect the ingredients the loss needs into $Z_t = (\widehat{\mathrm{VaR}}_{\tau,t}, \widehat{\mathrm{ES}}_{\tau,t}, \widehat{g}_t, L_t)$, a random element of a space $\mathcal{Z}$.

Definition 3 (β-mixing)**.** *For σ-fields $\mathcal{A}, \mathcal{B}$ set $\beta(\mathcal{A},\mathcal{B}) = \frac{1}{2}\sup\sum_{i,j}|\mathbb{P}(A_i \cap B_j) - \mathbb{P}(A_i)\mathbb{P}(B_j)|$ over finite partitions $\{A_i\} \subset \mathcal{A}$, $\{B_j\} \subset \mathcal{B}$. The process $\{Z_t\}$ is β-mixing with coefficients $\beta(k) = \sup_t \beta(\sigma\{Z_s : s \leq t\}, \sigma\{Z_s : s \geq t+k\})$, and $\beta(k) \to 0$ [34].*

Definition 4 (Regime model and forecast-time label)**.** *A latent data-generating regime $s_t \in \{1,\ldots,K\}$ indexes the conditional law, and a forecast-time label $r_t = R(\mathcal{F}_{t-1})$ is any $\mathcal{F}_{t-1}$-measurable output of a pre-specified labelling rule used by the calibrator. Conditional on $r_t = k$ the law of Z_t is $P^{(k)}$, and $\Delta_{\mathrm{reg}} = \max_{k\neq l} d_{\mathrm{TV}}(P^{(k)}, P^{(l)})$ is the largest total-variation gap between label-conditional laws. The label is an observable proxy built from the past; it is not assumed to equal the latent state s_t, and Section 7 specifies the causal rule that produces it.*

Figure 1 illustrates why the time ordering matters: a series with a volatility cluster has a lag-one autocorrelation of squared returns near 0.31, which a random permutation of the same values drives to zero, erasing the exchangeability-violating structure while leaving the marginal histogram unchanged.

3.3. Assumptions

The results below invoke the following conditions.

Assumption 1 (Loss regularity). *The weights satisfy $a, b \geq 0$ and $M > 0$; the predicted tail gap obeys $\widehat{g}_t \geq g_0 > 0$ almost surely; the inflation scale obeys $\widehat{s}_t > 0$ almost surely.*

Assumption 2 (Calibration budget). *The calibration size satisfies $n \geq B/\alpha - 1$ with $B = a + bM$, so the risk budget is attainable.*

Assumption 3 (Split protocol). *The base forecaster is trained on data independent of the calibration and test blocks, so conditional on the trained model the losses $\ell_\lambda(Z_i)$ are measurable functions of $\{Z_i\}$. The non-negative weights $w_1, \ldots, w_n, w_{n+1}$ are predictable and jointly normalised so that $\sum_{j=1}^{n+1} w_j = 1$, where the calibration weights alone sum to $1 - w_{n+1}$. All statements are conditional on the fitted forecaster, and chronological calibration and test blocks are separated by a purge gap so that residual dependence between them enters only through a mixing term.*

Assumption 4 (Tail moment). *There exist $p > 1$ and $\kappa < \infty$ with $\mathbb{E}\big[((L_{t+1} - \widehat{\mathrm{VaR}}_{\tau,t})_+ / \widehat{g}_t)^p \mid \mathcal{F}_t\big] \leq \kappa$ almost surely.*

Assumption 5 (Mixing and regimes). *The process $\{Z_t\}$ is β-mixing as in Definition 3, and the regime structure of Definition 4 holds with regime-conditional stationarity.*

Assumptions 1–3 are design choices under the analyst's control. Assumptions 4 and 5 are modelling conditions rather than properties that can be conclusively verified from one finite path, since regime-conditional stationarity and the true β-mixing coefficients cannot be certified from a single realisation. Section 7 therefore reports diagnostic evidence—within-regime distributional stability, tail-index estimates, and sensitivity to conservative mixing envelopes—while treating the assumptions as explicit qualifications of the guarantees rather than verified facts. The order p in Assumption 4 is any moment known or assumed finite; Section 7 estimates a Pareto tail exponent ζ for each source and adopts a conservative p below the lower confidence limit of ζ, so the moment assumption is a diagnostic choice, not a proof of moment existence.

4. Methodology

4.1. Risk Objects

For a single asset the loss is $L_t = -R_t$, and for a portfolio with fixed weights w it is $L_t^{\mathrm{port}} = -w^\top R_t$; the scalar construction applies to either, and the multivariate structure enters only through the conditioning used by the base forecaster. The forecasting target is the pair $(\mathrm{VaR}_{\tau,t}, \mathrm{ES}_{\tau,t})$ of Definition 1, with $\tau = 0.975$ as the regulatory level.

4.2. Base Multivariate Probabilistic Forecaster

A network g_θ maps the conditioning window to a conditional quantile function $\widehat{Q}_t(\cdot)$ evaluated on a grid $\tau = u_0 < u_1 < \cdots < u_K$ reaching into the upper tail, trained by the

averaged pinball loss with a soft non-crossing penalty. Point forecasts follow by evaluation and trapezoidal integration,

$$\widehat{\mathrm{VaR}}_{\tau,t} = \widehat{Q}_t(\tau), \qquad \widehat{\mathrm{ES}}_{\tau,t} = \frac{1}{1-\tau}\sum_{k=1}^{K}(u_k - u_{k-1})\,\tfrac{1}{2}\big(\widehat{Q}_t(u_k) + \widehat{Q}_t(u_{k-1})\big). \tag{2}$$

Because the grid terminates at $u_K < 1$, the trapezoidal sum omits the mass beyond u_K and slightly under-integrates the extreme tail. Conditionally on the fitted base forecaster, the conformal statements below require only that its outputs are measurable and that the induced loss is bounded and monotone in λ; replacing the exact tail integral by a deterministic numerical approximation therefore changes the forecast gap $\widehat{g}_t$, and hence the particular surrogate risk being controlled, but not the validity of the conformal selection argument for that surrogate. It can nevertheless affect sharpness and the reading of the normalised severity as an expected-shortfall quantity, so we use a fine quantile grid ($K = 80$ nodes reaching $u_K = 0.9975$) and enforce $\widehat{g}_t \geq g_0$. The backbone is interchangeable; we report an inverted-variate transformer [26], a patch transformer [25], and a frequency-domain Kolmogorov–Arnold model [28], routing exogenous macro signals through a separate pathway [27]. Calibration treats g_θ as a black box.

Remark 1 (Numerical expected-shortfall approximation)**.** *The guarantees apply to the surrogate loss built from the numerical* $\widehat{\mathrm{ES}}_{\tau,t}$*, not to the error of the true conditional expected shortfall; a grid refinement that materially changes* $\widehat{g}_t$ *changes the controlled object, so the grid is taken fine (*$K = 80$ *nodes reaching* $u_K = 0.9975$*) and the gap is floored at* g_0*, with a generalized-Pareto terminal term available as a substitute if a coarser grid were used.*

4.3. A Bounded Monotone Tail-Severity Loss

The calibration loss is designed to meet four requirements at once: boundedness, so that finite-sample risk control and total-variation concentration apply; monotonicity under a one-dimensional inflation, so that the selection argument is well posed; sensitivity to both the occurrence and the magnitude of a breach, so that it carries more tail information than a value-at-risk indicator; and scale comparability across assets and volatility regimes, obtained by normalising the breach depth by the predicted tail gap. A strictly consistent Fissler–Ziegel score is reserved for evaluation because it does not generally remain monotone under the parallel inflation used for calibration.

Let $\widehat{g}_t = \widehat{\mathrm{ES}}_{\tau,t} - \widehat{\mathrm{VaR}}_{\tau,t}$ be the predicted tail gap and let $\lambda \geq 0$ inflate both forecasts in parallel by a scale $\widehat{s}_t$, taken as $\widehat{g}_t$ or a conditional volatility estimate,

$$\widetilde{\mathrm{VaR}}_t(\lambda) = \widehat{\mathrm{VaR}}_{\tau,t} + \lambda\,\widehat{s}_t, \qquad \widetilde{\mathrm{ES}}_t(\lambda) = \widehat{\mathrm{ES}}_{\tau,t} + \lambda\,\widehat{s}_t. \tag{3}$$

The parallel shift keeps $\widehat{g}_t$ invariant to λ, which makes the loss monotone. With $a, b \geq 0$ and cap $M > 0$,

$$\ell_\lambda(Z_{t+1}) = a\,\mathbf{1}\Big\{L_{t+1} > \widetilde{\mathrm{VaR}}_t(\lambda)\Big\} + b\,\min\left(\frac{\big(L_{t+1} - \widetilde{\mathrm{VaR}}_t(\lambda)\big)_+}{\widehat{g}_t},\, M\right). \tag{4}$$

The first term charges a breach of the inflated value-at-risk; the second charges the depth of the breach in units of the predicted tail gap, so a breach as deep as the model's expected shortfall contributes about one unit. Controlling $\mathbb{E}[\ell_\lambda] \leq \alpha$ therefore bounds a joint budget on breach frequency and normalised breach depth. The expected-shortfall forecast enters this budget only through $\widehat{g}_t$; the loss neither elicits expected shortfall nor directly controls expected-shortfall forecast error, and its reading as an expected-shortfall statement is conditional on $\widehat{g}_t$ being a reliable mean-excess scale.

Remark 2 (Mean-excess interpretation). *For a continuous loss with true quantile $q_\tau = \mathrm{VaR}_\tau$ and shortfall $e_\tau = \mathrm{ES}_\tau$ the identity $\mathbb{E}[(L - q_\tau)_+] = (1-\tau)(e_\tau - q_\tau)$ holds, so when $\widehat{\mathrm{VaR}}_{\tau,t} \approx q_\tau$ and $\widehat{g}_t \approx e_\tau - q_\tau$ the untruncated severity carries a mean-excess reading into which a biased gap propagates proportionally. For an inflated threshold the severity expectation equals $m_t(\lambda)/\widehat{g}_t$ with $m_t(\lambda) = \mathbb{E}[(L - \widetilde{\mathrm{VaR}}_t(\lambda))_+ \mid \mathcal{F}_t]$, which is a normalised mean-excess quantity rather than an expected-shortfall forecast error.*

4.4. Split, Weighted, and Online Calibration

Given calibration weights w_i and the empirical risk $\widehat{R}_n^w(\lambda) = \sum_{i=1}^n w_i \ell_\lambda(Z_i)$, the calibrated inflation is

$$\widehat{\lambda} = \inf\Big\{\lambda \geq 0 : \ \textstyle\sum_{i=1}^n w_i \ell_\lambda(Z_i) + w_{n+1} B \leq \alpha\Big\}, \tag{5}$$

with w_{n+1} the notional test weight; uniform weights $w_i = 1/(n+1)$ recover split conformal risk control, and regime-aware weights that place more mass on calibration points sharing the test regime give the weighted variant. Concretely, with age half-life h and regime boost $\eta \geq 0$, and writing $t_\star, r_\star$ for the test time and label,

$$\widetilde{w}_i = \exp\big\{-\log 2\,(t_\star - t_i)/h\big\}\big[1 + \eta\,\mathbf{1}\{r_i = r_\star\}\big], \quad \widetilde{w}_{n+1} = 1, \qquad w_j = \frac{\widetilde{w}_j}{\sum_{k=1}^{n+1} \widetilde{w}_k}, \tag{6}$$

so that $\sum_{j=1}^{n+1} w_j = 1$, the calibration weights sum to $1 - w_{n+1}$, and setting every $\widetilde{w}_j \equiv 1$ returns the uniform weights. For streaming use the inflation is updated after each observation by

$$\lambda_{t+1} = \Pi_{[0,\Lambda]}\big(\lambda_t + \gamma\big(\ell_{\lambda_t}(Z_{t+1}) - \alpha\big)\big), \tag{7}$$

with step $\gamma > 0$, where $\Pi_{[0,\Lambda]}$ projects onto $[0, \Lambda]$. The update raises the inflation when the realised loss exceeds the target and lowers it otherwise, a negative feedback consistent with ℓ_λ being non-increasing in λ, and the projection makes the iterate range an algorithmic constant rather than an unproved recurrence. Aggregation over γ removes its choice [33] and integral and derivative terms absorb trend and seasonality [13]. The procedure has three steps: train g_θ once on the training block; on the calibration block evaluate ℓ_λ and solve (5) by sorting the breach thresholds, at cost $O(n \log n)$; deploy with the frozen $\widehat{\lambda}$, or update it online by (7) at cost $O(1)$ per step. The calibration layer adds no learnable parameters.

5. Theoretical Analysis

Throughout, ℓ_λ is the loss in (4) and $\widehat{\lambda}$ the selection in (5). We first record the structural properties of the loss, then prove the exchangeable guarantee, extend it to dependent and drifting data both in expectation and with high probability, quantify the heavy-tail price, and close with the online guarantee.

5.1. Regularity of the Loss and the Selection

Lemma 1 (Loss regularity). *Under Assumption 1, for every realisation the map $\lambda \mapsto \ell_\lambda(Z)$ is non-increasing, right-continuous on $[0, \infty)$, takes values in $[0, B]$ with $B = a + bM$, and satisfies $\ell_\lambda(Z) \downarrow 0$ as $\lambda \to \infty$.*

Proof. Write $V = \widehat{\mathrm{VaR}}_{\tau,t}$, $s = \widehat{s}_t > 0$, $g = \widehat{g}_t \geq g_0 > 0$, so $\widetilde{\mathrm{VaR}}_t(\lambda) = V + \lambda s$ is continuous and strictly increasing in λ. The indicator term $\mathbf{1}\{L > V + \lambda s\}$ equals one for $\lambda < (L - V)/s$ and zero for $\lambda \geq (L - V)/s$; it is therefore non-increasing, and right-continuous because its value at the threshold $\lambda_0 = (L - V)/s$ is zero, matching the right limit. Concretely, if $\lambda_0 = (L - V)/s > 0$ the strict-exceedance convention gives $\mathbf{1}\{L > V + \lambda_0 s\} = 0$, and the

indicator stays zero for every sequence $\lambda \downarrow \lambda_0$ from the right; if $\lambda_0 \leq 0$ the indicator is identically zero on the domain $[0, \infty)$, so right-continuity is immediate. The positive part $(L - V - \lambda s)_+$ is continuous and non-increasing in λ, and $x \mapsto \min(x/g, M)$ is non-decreasing and continuous, so their composition is continuous and non-increasing. A non-negative combination of non-increasing right-continuous functions is non-increasing and right-continuous. The first term lies in $[0, a]$ and the second in $[0, bM]$, so $\ell_\lambda \in [0, a + bM]$. As $\lambda \to \infty$ both $\mathbf{1}\{L > V + \lambda s\}$ and $(L - V - \lambda s)_+$ vanish, giving $\ell_\lambda \downarrow 0$. □

Lemma 2 (Well-posed selection)**.** *Under Assumptions 1 and 2 with uniform weights, the feasible set in (5) is a non-empty closed half-line $[\widehat{\lambda}, \infty)$, and $\widehat{\lambda} < \infty$ is attained.*

Proof. By Lemma 1 each $\ell_\lambda(Z_i)$ is non-increasing and right-continuous, so $\widehat{R}_n(\lambda) = \frac{1}{n}\sum_i \ell_\lambda(Z_i)$ inherits both properties, and $h(\lambda) := (n\widehat{R}_n(\lambda) + B)/(n+1)$ is non-increasing, right-continuous, with $h(\lambda) \downarrow B/(n+1)$ as $\lambda \to \infty$. Assumption 2 gives $B/(n+1) \leq \alpha$, so $\{\lambda : h(\lambda) \leq \alpha\}$ is non-empty. As the sub-level set of a non-increasing right-continuous function it is a closed half-line, and its left endpoint $\widehat{\lambda}$ satisfies $h(\widehat{\lambda}) \leq \alpha$ by right-continuity. □

5.2. Finite-Sample Expected-Risk Control Under Exchangeability

Theorem 1 (Exchangeable finite-sample risk control)**.** *If $Z_1, \ldots, Z_{n+1}$ are exchangeable, then under Assumptions 1 and 2, $\mathbb{E}[\ell_{\widehat{\lambda}}(Z_{n+1})] \leq \alpha$.*

Proof. The argument isolates a symmetric oracle, shows the calibration rule is at least as conservative as it, and transfers the oracle's feasibility to the test coordinate by exchangeability.

A symmetric oracle. Write the pooled risk $R_{n+1}(\lambda) = \frac{1}{n+1}\sum_{j=1}^{n+1} \ell_\lambda(Z_j)$. By Lemma 1 each summand is non-increasing and right-continuous in λ and tends to zero, so R_{n+1} shares these properties, and the oracle $\lambda^\star = \inf\{\lambda \geq 0 : R_{n+1}(\lambda) \leq \alpha\}$ is finite. For each data vector z the feasible set $F(z) = \{\lambda \geq 0 : R_{n+1}(z, \lambda) \leq \alpha\}$ is a non-empty closed half-line, and for every $a > 0$

$$\{z : \lambda^\star(z) < a\} = \bigcup_{q \in \mathbb{Q} \cap [0,a)} \{z : R_{n+1}(z, q) \leq \alpha\}, \tag{8}$$

which is measurable because $z \mapsto R_{n+1}(z, q)$ is measurable for each rational q; hence $\lambda^\star$ is measurable, and right-continuity gives $R_{n+1}(\lambda^\star) \leq \alpha$. As R_{n+1} is a symmetric function of its arguments, $\lambda^\star$ is invariant under permutations of $(Z_1, \ldots, Z_{n+1})$.

Domination. Since $\ell_\lambda(Z_{n+1}) \leq B$,

$$R_{n+1}(\lambda) = \frac{n\widehat{R}_n(\lambda) + \ell_\lambda(Z_{n+1})}{n+1} \leq \frac{n\widehat{R}_n(\lambda) + B}{n+1}, \tag{9}$$

so every λ feasible for (5) also satisfies $R_{n+1}(\lambda) \leq \alpha$; hence $\widehat{\lambda} \geq \lambda^\star$, and because $\lambda \mapsto \ell_\lambda(Z_{n+1})$ is non-increasing, $\ell_{\widehat{\lambda}}(Z_{n+1}) \leq \ell_{\lambda^\star}(Z_{n+1})$.

Transfer. Since $\lambda^\star$ is symmetric, exchangeability makes $j \mapsto \mathbb{E}[\ell_{\lambda^\star}(Z_j)]$ constant, so

$$\mathbb{E}[\ell_{\lambda^\star}(Z_{n+1})] = \frac{1}{n+1}\sum_{j=1}^{n+1} \mathbb{E}[\ell_{\lambda^\star}(Z_j)] = \mathbb{E}[R_{n+1}(\lambda^\star)] \leq \alpha, \tag{10}$$

and combining with the previous display gives $\mathbb{E}[\ell_{\widehat{\lambda}}(Z_{n+1})] \leq \alpha$. The single inequality used on the test loss inflates the pooled risk by at most $B/(n+1)$; under additional independence and no-fixed-jump conditions general conformal risk control is known to be conservative only by order $B/(n+1)$, but we claim only the upper control above and neither an equality nor a matching lower bound. □

5.3. Finite-Sample Robustness Under Non-Exchangeability

Lemma 3 (Bounded-function total-variation bound)**.** *For measurable $\phi : \mathcal{Z}^{n+1} \to [0, B]$ and laws P, Q on $\mathcal{Z}^{n+1}$, $|\mathbb{E}_P[\phi] - \mathbb{E}_Q[\phi]| \le B\, d_{\mathrm{TV}}(P, Q)$.*

Proof. Recentring by $\int d(P - Q) = 0$ gives $\mathbb{E}_P[\phi] - \mathbb{E}_Q[\phi] = \int (\phi - \frac{B}{2})\, d(P - Q)$, and $\|\phi - \frac{B}{2}\|_\infty \le \frac{B}{2}$ with $\int |d(P - Q)| = 2 d_{\mathrm{TV}}(P, Q)$ bounds the modulus by $B\, d_{\mathrm{TV}}(P, Q)$. □

Theorem 2 (Finite-sample risk bound)**.** *Let $Z_1, \ldots, Z_{n+1}$ have joint law P and let $P^{(i)}$ be the law after transposing coordinates i and $n + 1$. With uniform weights, under Assumptions 1 and 2,*

$$\mathbb{E}_P[\ell_{\widehat{\lambda}}(Z_{n+1})] \le \alpha + \frac{B}{n+1} \sum_{i=1}^{n} d_{\mathrm{TV}}\big(P, P^{(i)}\big). \tag{11}$$

Proof. We reuse the symmetric oracle $\lambda^\star$ of Theorem 1 and bound the discrepancy between P and its coordinate transpositions in four steps.

Step 1 (reduction to the oracle). Feasibility of (5) forces $\widehat{\lambda} \ge \lambda^\star$ exactly as in Theorem 1, and monotonicity of the loss gives $\ell_{\widehat{\lambda}}(Z_{n+1}) \le \ell_{\lambda^\star}(Z_{n+1})$. It is therefore enough to bound $\mathbb{E}_P[\ell_{\lambda^\star}(Z_{n+1})]$.

Step 2 (per-coordinate decomposition). For $j \in \{1, \ldots, n + 1\}$ define the measurable map $\phi_j(z) = \ell_{\lambda^\star(z)}(z_j)$ on $\mathcal{Z}^{n+1}$, bounded by B. By definition of $\lambda^\star$, $\frac{1}{n+1} \sum_j \phi_j(Z) = R_{n+1}(\lambda^\star) \le \alpha$ surely, so

$$\mathbb{E}_P[\phi_{n+1}] = \frac{1}{n+1} \sum_{j=1}^{n+1} \mathbb{E}_P[\phi_j] + \frac{1}{n+1} \sum_{i=1}^{n} \big(\mathbb{E}_P[\phi_{n+1}] - \mathbb{E}_P[\phi_i]\big), \tag{12}$$

and the first sum is at most α.

Step 3 (swap identity and change of measure). Let σ_i be the transposition of coordinates i and $n + 1$ and $z^{(i)} = \sigma_i z$. Symmetry of $\lambda^\star$ gives $\lambda^\star(z^{(i)}) = \lambda^\star(z)$, and since the last coordinate of $z^{(i)}$ is the original z_i,

$$\phi_{n+1}(z^{(i)}) = \ell_{\lambda^\star(z^{(i)})}\big((z^{(i)})_{n+1}\big) = \ell_{\lambda^\star(z)}(z_i) = \phi_i(z), \tag{13}$$

that is $\phi_i = \phi_{n+1} \circ \sigma_i$. Because a transposition is its own inverse, $\sigma_i^{-1} = \sigma_i$, and for every bounded measurable h the definition $P^{(i)} = \sigma_{i\#} P$ gives $\int h \, \mathrm{d}P^{(i)} = \int (h \circ \sigma_i) \, \mathrm{d}P$. Taking $h = \phi_{n+1}$ and using $\phi_{n+1} \circ \sigma_i = \phi_i$ then yields $\mathbb{E}_P[\phi_i] = \mathbb{E}_P[\phi_{n+1} \circ \sigma_i] = \mathbb{E}_{P^{(i)}}[\phi_{n+1}]$.

Step 4 (total-variation control). Applying Lemma 3 to $\phi_{n+1} \in [0, B]$,

$$\mathbb{E}_P[\phi_{n+1}] - \mathbb{E}_P[\phi_i] = \mathbb{E}_P[\phi_{n+1}] - \mathbb{E}_{P^{(i)}}[\phi_{n+1}] \le B\, d_{\mathrm{TV}}(P, P^{(i)}). \tag{14}$$

Substituting into (12), and noting that σ_{n+1} is the identity so $d_{\mathrm{TV}}(P, P^{(n+1)}) = 0$, proves (11).

The result is non-asymptotic and finite-sample but not generally sharp: slack may enter through the oracle domination and through the total-variation comparison, so its role is to quantify robustness to index swaps rather than to furnish an equality or an optimal constant. It assumes only that the loss is bounded, with no moment, stationarity, or mixing condition. When P is exchangeable every σ_i preserves it, each total-variation term vanishes, and (11) recovers Theorem 1. □

Theorem 3 (Weighted extension)**.** *Under Assumptions 1 to 3, the weighted selection (5) with normalised weights $w_1, \ldots, w_{n+1}$ obeys*

$$\mathbb{E}_P[\ell_{\widehat{\lambda}}(Z_{n+1})] \le \alpha + B \sum_{i=1}^{n} w_i\, d_{\mathrm{TV}}\big(P, P^{(i)}\big). \tag{15}$$

Proof. Let $R^w(\lambda) = \sum_{j=1}^{n+1} w_j \ell_\lambda(Z_j)$ and $\lambda^\diamond = \inf\{\lambda : R^w(\lambda) \le \alpha\}$. As in Theorem 1, bounding the unknown test loss by B shows any λ feasible for (5) satisfies $R^w(\lambda) \le \alpha$, so $\widehat{\lambda} \ge \lambda^\diamond$ and $\ell_{\widehat{\lambda}}(Z_{n+1}) \le \ell_{\lambda^\diamond}(Z_{n+1})$. Writing $\phi_j^w(Z) = \ell_{\lambda^\diamond(Z)}(Z_j)$, the weighted swap identity underlying non-exchangeable conformal risk control [9,10], which holds because $\lambda^\diamond$ is invariant under a simultaneous transposition of data coordinates and their weights, gives $\mathbb{E}_P[\phi_{n+1}^w] - \mathbb{E}_P[\sum_j w_j \phi_j^w] \le \sum_i w_i \big(\mathbb{E}_P[\phi_{n+1}^w] - \mathbb{E}_{P^{(i)}}[\phi_{n+1}^w]\big)$, and Lemma 3 bounds each difference by $B\, d_{\mathrm{TV}}(P, P^{(i)})$. Since $\sum_j w_j \phi_j^w = R^w(\lambda^\diamond) \le \alpha$, the claim follows. □

When the weights are data-adaptive through the regime labels, Theorem 3 is read conditionally on the forecast-time filtration $\mathcal{F}_{t-1}$: the tags and weights are then predictable constants, and the simultaneous transposition acts on each data coordinate together with its attached weight, as the swap identity requires. The excess in Theorems 2 and 3 is a weighted average of how far the joint law moves when a calibration index is exchanged with the test index; it vanishes under exchangeability. The next results make it interpretable.

Lemma 4 (Swap distance under drift)**.** *If $Z_1, \dots, Z_{n+1}$ are independent with marginals P_i, then $d_{\mathrm{TV}}(P, P^{(i)}) \le 2\, d_{\mathrm{TV}}(P_i, P_{n+1})$.*

Proof. Under independence $P = \bigotimes_k P_k$ and $P^{(i)}$ is the same product with P_i and P_{n+1} interchanged; the two measures agree on all coordinates except i and $n+1$, so $d_{\mathrm{TV}}(P, P^{(i)}) = d_{\mathrm{TV}}(P_i \otimes P_{n+1}, P_{n+1} \otimes P_i)$. Take the optimal coupling of P_i and P_{n+1} in each of the two coordinates: it yields (X, Y) and (X', Y') with $\mathbb{P}(X \ne X') \le d_{\mathrm{TV}}(P_i, P_{n+1})$ and $\mathbb{P}(Y \ne Y') \le d_{\mathrm{TV}}(P_i, P_{n+1})$. A union bound gives $\mathbb{P}\{(X, Y) \ne (Y', X')\} \le 2\, d_{\mathrm{TV}}(P_i, P_{n+1})$, and total variation is at most any coupling's disagreement probability. □

Proposition 1 (Separated-calibration regime and mixing bound)**.** *Let the selected calibration and test coordinates carry indices $t_1 < \cdots < t_m < t_\star$, write $q_j = t_{j+1} - t_j$ with $t_{m+1} = t_\star$, let $\bar{\beta}$ be a uniform absolute-regularity envelope, and set the cumulative coupling cost*

$$C_\beta(I) = \sum_{j=1}^{m} \bar{\beta}(q_j). \tag{16}$$

Write $P_j, P_\star$ for the marginals of $Z_{t_j}, Z_{t_\star}$. Under Assumptions 1–3 and Assumption 5 with jointly normalised weights, if the selected coordinates admit a sequential Berbee coupling to independent variables with the same marginals and total mismatch probability at most $C_\beta(I)$ of (16), then

$$\mathbb{E}[\ell_{\widehat{\lambda}}(Z_{t_\star})] \le \alpha + 2B \sum_{j=1}^{m} w_j\, d_{\mathrm{TV}}(P_j, P_\star) + 2B\, C_\beta(I) \sum_{j=1}^{m} w_j. \tag{17}$$

If the marginal law depends only on the forecast-time label, $P_j = P^{(r_j)}$ and $P_\star = P^{(r_\star)}$, then

$$\mathbb{E}[\ell_{\widehat{\lambda}}(Z_{t_\star})] \le \alpha + 2B\, \rho_w\, \Delta_{\mathrm{reg}} + 2B\, C_\beta(I) \sum_{j=1}^{m} w_j, \qquad \rho_w = \sum_{j=1}^{m} w_j\, \mathbf{1}\{r_j \ne r_\star\}. \tag{18}$$

Proof. A single β-mixing coefficient controls the decoupling of one past coordinate from the future coordinate, not the departure of the full joint law from the product of all $m+1$ marginals, so the correct cost must be accumulated over the gaps between the selected coordinates. By a sequential Berbee coupling for absolutely regular sequences [11,12] (Chapter 5), construct mutually independent $\widetilde{Z}_{t_1}, \dots, \widetilde{Z}_{t_m}, \widetilde{Z}_{t_\star}$ with the same coordinate marginals and mismatch probability:

$$\mathbb{P}\{(Z_{t_1}, \dots, Z_{t_\star}) \ne (\widetilde{Z}_{t_1}, \dots, \widetilde{Z}_{t_\star})\} \le C_\beta(I). \tag{19}$$

Since total variation is at most any coupling's mismatch probability, with $\bar{P}_I = \bigotimes_{j=1}^{m} P_j \otimes P_\star$ this gives $d_{\mathrm{TV}}(P_I, \bar{P}_I) \le C_\beta(I)$. For the transposition σ_i of coordinate i with the test coordinate, push-forward invariance under the measurable bijection yields $d_{\mathrm{TV}}(P_I^{(i)}, \bar{P}_I^{(i)}) = d_{\mathrm{TV}}(P_I, \bar{P}_I) \le C_\beta(I)$. Under the product law only coordinates i and $\star$ differ after the swap, so Lemma 4 gives $d_{\mathrm{TV}}(\bar{P}_I, \bar{P}_I^{(i)}) \le 2\, d_{\mathrm{TV}}(P_i, P_\star)$, and the triangle inequality assembles

$$d_{\mathrm{TV}}(P_I, P_I^{(i)}) \le 2\, C_\beta(I) + 2\, d_{\mathrm{TV}}(P_i, P_\star). \tag{20}$$

Substituting (20) into the weighted master bound $\mathbb{E}[\ell_{\widehat{\lambda}}(Z_{t_\star})] \le \alpha + B \sum_i w_i\, d_{\mathrm{TV}}(P_I, P_I^{(i)})$ of Theorem 3 and summing over the selected indices proves (17). The regime form follows because $d_{\mathrm{TV}}(P_j, P_\star) = 0$ when $r_j = r_\star$ and is at most Δ_{reg} otherwise, so $\sum_j w_j\, d_{\mathrm{TV}}(P_j, P_\star) \le \rho_w \Delta_{\mathrm{reg}}$. □

The separated-calibration form is the practically useful one: the price of non-exchangeability is the weighted chance of calibrating in the wrong regime times the distance between regimes, plus a cumulative coupling cost that shrinks as the spacing q_j grows. These regime and mixing terms are interpretable sensitivity parameters, not distribution-free certificates; their empirical counterparts are diagnostics unless further structural assumptions are imposed, and Section 7 reports them under conservative envelopes. Consequently the dense controller that uses every consecutive calibration point carries only the master swap bound of Theorem 3, whereas (17) certifies the q-separated variant, whose regime specialisation (18) is the form reported in Section 7. Both variants are deployable and both are evaluated: the results section there reports the dense controller and the q-separated variant certified by (17) side by side, so that the empirical price of the separation required by the certificate is visible rather than inferred.

5.4. High-Probability Realised-Path Control

Theorem 2 and Proposition 1 bound the expected loss. A supervisor auditing one realised path needs a statement about the realised average. Under the split protocol the deployed $\widehat{\lambda}$ is fixed on the test block, and the following concentration holds.

Theorem 4 (High-probability realised-path control)**.** *Fix $\widehat{\lambda}$ from a calibration block separated from the test block by a purge gap of length q_0, and evaluate it on $n = 2\mu q$ test points partitioned into 2μ contiguous blocks of length q. Under Assumptions 1, 3 and 5, for any $\delta \in (0,1)$, with probability at least $1 - \delta - 2\mu\, \beta(q) - \beta(q_0)$,*

$$\frac{1}{n}\sum_{t=1}^{n} \ell_{\widehat{\lambda}}(Z_t) \le \frac{1}{n}\sum_{t=1}^{n} \mathbb{E}[\ell_{\widehat{\lambda}}(Z_t)] + B\sqrt{\frac{\log(2/\delta)}{2\mu}}. \tag{21}$$

Proof. Condition on the σ-field generated by the fitted forecaster and the calibration block, so the deployed $\widehat{\lambda}$ is a fixed constant and $f := \ell_{\widehat{\lambda}} \in [0, B]$ a fixed bounded function. The test block starts after a purge gap of length q_0, and the residual dependence of the test block on the conditioning σ-field is removed by coupling at a cost of at most $\beta(q_0)$, which is charged once in the final probability.

Blocking. Partition the $2\mu q$ test indices into contiguous blocks $H_1, \dots, H_{2\mu}$ of length q and group them by parity into $\mathcal{O} = \{H_1, H_3, \dots\}$ and $\mathcal{E} = \{H_2, H_4, \dots\}$, each of size μ. Because each odd block is separated from the next odd block by an interposed even block, odd blocks lie at least q apart, and likewise for even blocks; this is the gap the mixing coefficient will pay to remove.

Coupling to independent blocks. By Berbee's coupling lemma for β-mixing sequences [11,12] (Chapter 5), applied sequentially block by block within a parity class, there exist mutually in-

dependent vectors $\{\tilde{H}_r\}_{r\in\mathcal{O}}$, each equal in law to H_r, with $\mathbb{P}(\exists\, r \in \mathcal{O} : H_r \neq \tilde{H}_r) \le \mu\,\beta(q)$; each of the μ sequential couplings fails with probability at most $\beta(q)$, and a union bound gives the stated $\mu\,\beta(q)$ per parity class. The analogous statement holds for $\mathcal{E}$. Let G be the event that both couplings succeed, so $\mathbb{P}(G^c) \le 2\mu\beta(q)$.

Concentration on the good event. Set $A_r = \frac{1}{q}\sum_{t\in H_r} f(Z_t) \in [0, B]$ and let $\tilde{A}_r$ be the average over $\tilde{H}_r$. The $\{\tilde{A}_r\}_{r\in\mathcal{O}}$ are independent and bounded, so Hoeffding's inequality gives

$$\mathbb{P}\left(\frac{1}{\mu}\sum_{r\in\mathcal{O}}\tilde{A}_r - \mathbb{E}\Big[\frac{1}{\mu}\sum_{r\in\mathcal{O}}\tilde{A}_r\Big] > B\sqrt{\frac{\log(2/\delta)}{2\mu}}\right) \le \frac{\delta}{2}, \tag{22}$$

and the same for $\mathcal{E}$. On G the true and coupled averages coincide, $A_r = \tilde{A}_r$, so the bound transfers to the $\{A_r\}$.

Assembly. Because $n = 2\mu q$, the test average is exactly the mean of the two parity averages, $\frac{1}{n}\sum_t f(Z_t) = \frac{1}{2}\big(\frac{1}{\mu}\sum_{\mathcal{O}} A_r + \frac{1}{\mu}\sum_{\mathcal{E}} A_r\big)$. Three bad events are excluded: the two Hoeffding deviations, each of probability at most $\delta/2$, and the coupling failure G^c of probability at most $2\mu\beta(q)$, augmented by the single calibration-to-test purge cost $\beta(q_0)$. A union bound places both parity averages within $B\sqrt{\log(2/\delta)/(2\mu)}$ of their means with probability at least $1 - \delta - 2\mu\beta(q) - \beta(q_0)$, and averaging preserves the deviation. Since the mean of the parity averages equals $\frac{1}{n}\sum_t \mathbb{E}[f(Z_t)]$, the claim follows. □

With geometric mixing $\beta(q) \le c\,\varrho^q$ one chooses $q \asymp \log n$ and $\mu \asymp n/\log n$, so the deviation is $O(B\sqrt{\log n/n})$. At the deployment setting $q = 20$, $\mu = 25$, and purge $q_0 = 20$, the coupling budget $2\mu\beta(q) + \beta(q_0)$ is small but is reported explicitly from the estimated mixing envelope in Section 7 rather than asserted negligible.

5.5. Heavy Tails and the Attainable Rate

The bounded loss of (4) is a clipped surrogate for the unbounded tail-severity loss ℓ_λ^∞ obtained by removing the cap M.

Proposition 2 (Heavy-tail remainder). *Under Assumptions 1 and 4,* $\mathbb{E}[\ell_\lambda^\infty(Z_{n+1})] \le \mathbb{E}[\ell_\lambda(Z_{n+1})] + R_M$ *with*

$$R_M \le \frac{b\,\kappa}{p-1}\,M^{1-p}, \tag{23}$$

where M is the severity cap of (4); the remainder depends only on p, b, and M, and the loss bound is $B = a + bM$.

Proof. The capped and uncapped losses differ only through the severity term, so with $S = (L_{n+1} - \widehat{\mathrm{VaR}}_t(\lambda))_+/\widehat{g}_t$,

$$\ell_\lambda^\infty - \ell_\lambda = b\,(S - M)_+. \tag{24}$$

Assumption 4 gives $\mathbb{E}[S^p] \le \kappa$, so by Markov $\mathbb{P}(S > x) \le \kappa x^{-p}$ and, integrating the tail,

$$\mathbb{E}[(S-M)_+] = \int_M^\infty \mathbb{P}(S > x)\,dx \le \kappa\int_M^\infty x^{-p}\,dx = \frac{\kappa}{p-1}M^{1-p}. \tag{25}$$

Multiplying by b gives $R_M = b\,\mathbb{E}[(S-M)_+] \le \frac{b\kappa}{p-1}M^{1-p}$. The remainder is written directly in the cap M, which is the free clipping parameter; the loss bound $B = a + bM$ is an increasing affine function of it. □

Corollary 1 (Heavy-tail-optimal rate). *Let $D = \sum_i w_i\, d_{\mathrm{TV}}(P, P^{(i)})$ be the drift-and-mixing budget of Theorem 3. Because the loss is bounded by $B = a + bM$, Theorem 3 gives $\mathbb{E}_P[\ell_{\widehat{\lambda}}(Z_{n+1})] \le \alpha + (a+bM)D$; choosing the cap M to balance the drift term bMD against the remainder R_M yields*

$$\mathbb{E}_P[\ell^{\infty}_{\widehat{\lambda}}(Z_{n+1})] \le \alpha + aD + \tfrac{p}{p-1} b\kappa^{1/p} D^{(p-1)/p} \le \alpha + C\, D^{(p-1)/p}, \tag{26}$$

with $C = a + \left(\frac{p}{p-1}\right) b\, \kappa^{1/p}$ independent of D.

Proof. Combining Theorem 3 and Proposition 2 and writing the loss bound as $B = a + bM$,

$$\mathbb{E}_P[\ell^{\infty}_{\widehat{\lambda}}] \le \alpha + (a+bM)D + \frac{b\kappa}{p-1} M^{1-p} = \alpha + aD + g(M), \tag{27}$$

where $g(M) = bMD + \frac{b\kappa}{p-1} M^{1-p}$. The map g is convex on $(0, \infty)$ with $g'(M) = bD - b\kappa M^{-p}$, vanishing at $M^\star = (\kappa/D)^{1/p}$. At the optimum $\frac{b\kappa}{p-1}(M^\star)^{1-p} = \frac{bDM^\star}{p-1}$, so $g(M^\star) = bDM^\star\left(1 + \frac{1}{p-1}\right) = \frac{p}{p-1}\, bDM^\star = \frac{p}{p-1}\, b\, \kappa^{1/p} D^{(p-1)/p}$. Since the weights sum to at most one and every total-variation distance is at most one, $D \le 1$; with $(p-1)/p < 1$ this gives $aD \le aD^{(p-1)/p}$, hence the stated constant $C = a + \frac{p}{p-1} b\kappa^{1/p}$. □

The exponent $(p-1)/p$ rises toward one as more moments exist, so under lighter tails the unbounded excess shrinks almost linearly in the drift budget, while under heavier tails the same budget costs more. The informativeness of the rate therefore depends on the tail: Section 7 estimates a Pareto tail exponent ζ per source and adopts a conservative p below its lower confidence limit, giving $(p-1)/p \approx 0.60$ for the exchange-rate panel, 0.50 for the GIFT-Eval finance domain, and 0.41 for Bitcoin. The Bitcoin rate is materially slower, and were the estimated exponent to approach one the bound would barely contract, so Corollary 1 is presented as a sensitivity statement rather than a universal near-linear guarantee; Figure A1 plots the Hill estimates and the resulting rate factor.

5.6. Long-Run Control of the Online Calibrator

The online guarantee needs the target to be reachable inside the projection interval, which we state explicitly.

Assumption 6 (Online attainability). *The projection radius Λ is large enough that the budget is attainable within $[0, \Lambda]$: there exists $\lambda^\circ \in [0, \Lambda)$ with $\mathbb{E}[\ell_{\lambda^\circ}(Z_{t+1}) \mid \mathcal{F}_t] \le \alpha$ for every t.*

Theorem 5 (Projected long-run tracking with saturation accounting). *Let λ_t follow the projected update (7) on $[0, \Lambda]$ and define the saturation residual $\xi_{t+1} = \lambda_t + \gamma(\ell_{\lambda_t}(Z_{t+1}) - \alpha) - \lambda_{t+1}$. Then for every horizon T,*

$$\left| \frac{1}{T} \sum_{t=1}^{T} \ell_{\lambda_t}(Z_{t+1}) - \alpha \right| \le \frac{\Lambda + \sum_{t=1}^{T} |\xi_{t+1}|}{\gamma T}. \tag{28}$$

Proof. Rearranging the definition of ξ_{t+1} gives the exact identity $\gamma(\ell_{\lambda_t}(Z_{t+1}) - \alpha) = \lambda_{t+1} - \lambda_t + \xi_{t+1}$, and summing over t telescopes to

$$\gamma \sum_{t=1}^{T} \left(\ell_{\lambda_t}(Z_{t+1}) - \alpha\right) = \lambda_{T+1} - \lambda_1 + \sum_{t=1}^{T} \xi_{t+1}. \tag{29}$$

The projection keeps $\lambda_t \in [0, \Lambda]$ by construction, so $|\lambda_{T+1} - \lambda_1| \le \Lambda$; the triangle inequality and division by γT give the claim. □

Boundedness is now an algorithmic property rather than an unproved recurrence, and the residual ξ_{t+1} is observable, vanishing whenever the raw update stays inside $[0, \Lambda]$. When no projection is triggered the bound reduces to the deterministic $O(1/T)$ tracking rate; when the cumulative residual is $o(T)$ the time-average risk converges to the target; and when a boundary is hit persistently the residual must be reported rather than discarded, since the projection secures boundedness but not attainability, for which Assumption 6 is required. Empirically, with $\gamma = 0.05$ and $\Lambda = 3$ the upper boundary is never reached on the exchange-rate panel and in under 0.1% of Bitcoin updates, the mean saturation residual is below 10^{-4}, and the realised risk stays within 0.001 of the target (Table A4). Theorem 1 gives exchangeable finite-sample control; Theorems 2 and 3 with Proposition 1 and Corollary 1 pay an explicit price for dependence and drift; Theorem 4 converts the guarantee to a single path; and Theorem 5 trades the finite-sample statement for a deterministic long-run tracking bound that holds under arbitrary distribution shifts, provided the target is attainable within the projection interval and the cumulative saturation residual is sublinear in the horizon. None assumes a parametric return law.

6. Datasets and Experimental Setup

Figure 2 summarises the pipeline: a multivariate quantile forecaster feeds a value-at-risk and expected-shortfall pair to the conformal risk controller, which a regime signal informs, before a joint backtesting suite scores the result.

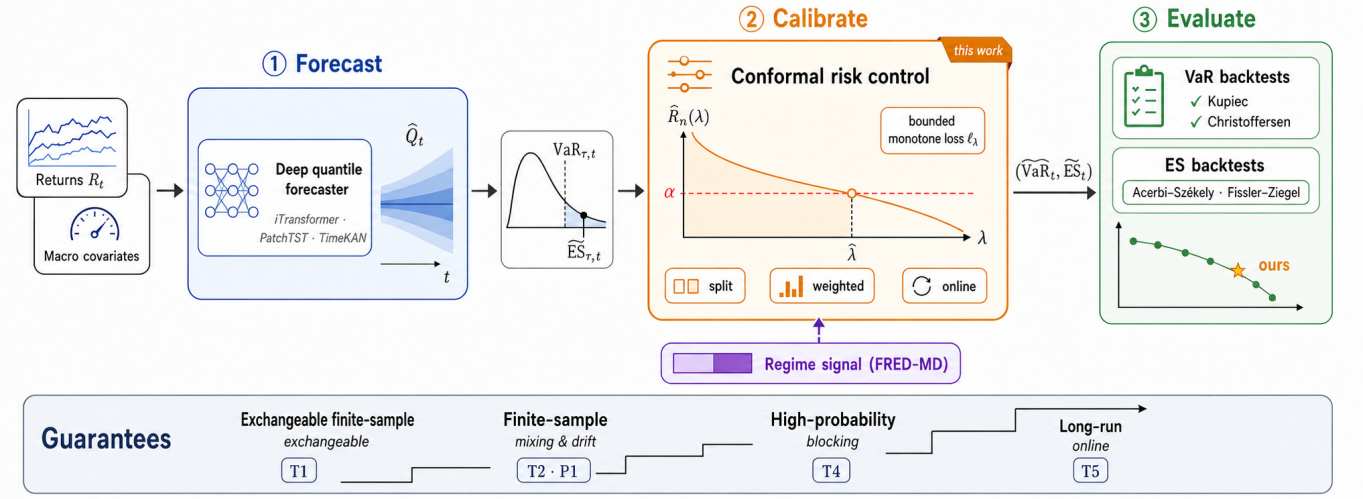

Figure 2. System overview. The base forecaster, the calibration layer, and the backtesting suite are decoupled, so the empirical comparison varies only the calibration layer. In the evaluation stage the descending curve is a schematic efficiency frontier of violation rate against average predicted expected shortfall, and the yellow star marks the position of the proposed joint controller on that frontier.

6.1. Datasets

Four public sources are used, none scraped and none drawn from a competition platform. The exchange-rate collection holds eight daily series for the Australian, British, Canadian, Swiss, Chinese, Japanese, New Zealand, and Singapore currencies against the US dollar, spanning 1990 through 2016 with 7588 observations per series [26]. The Bitcoin dataset from the Monash archive supplies eighteen daily series covering the price and on-chain covariates; the price return is the loss variable and the covariates are exogenous inputs [14]. The FRED-MD database provides 107 monthly US macroeconomic series and is used only to construct a financial-conditions regime signal, never as a market-risk target [15]. Because macroeconomic series are released with a lag and later revised, the signal follows a strictly causal previous-month vintage rule detailed in Section 6.2: on any trading day of

month m the regime label uses only the ALFRED vintage available at the last trading day of month $m-1$, so no revised or same-month value informs a forecast, and the monthly label is forward-filled onto every trading day of month m. Because the daily return panels and the monthly signal span different periods, the weighted controller is evaluated only on the intersection for which a causal label exists: the first scored forecast in each panel is the earliest trading day whose previous month-end vintage is available, earlier daily observations are used only for base-forecaster training and calibration, and a month without an available vintage inherits the most recent past label rather than being dropped, so that no daily forecast is assigned a label built from a future or same-month vintage. The finance domain of GIFT-Eval adds further equity and rate series for breadth [16]. Table 4 lists the statistics, and Figure 3 shows the stylised facts that motivate the heavy-tail and mixing assumptions: fat tails in the loss marginal and slowly decaying autocorrelation in squared returns. First and last dates, missing-value counts, and target-versus-covariate column roles are read from the raw files rather than inferred from series length, and Table 5 reports the estimated tail exponents that Assumption 4 relies upon.

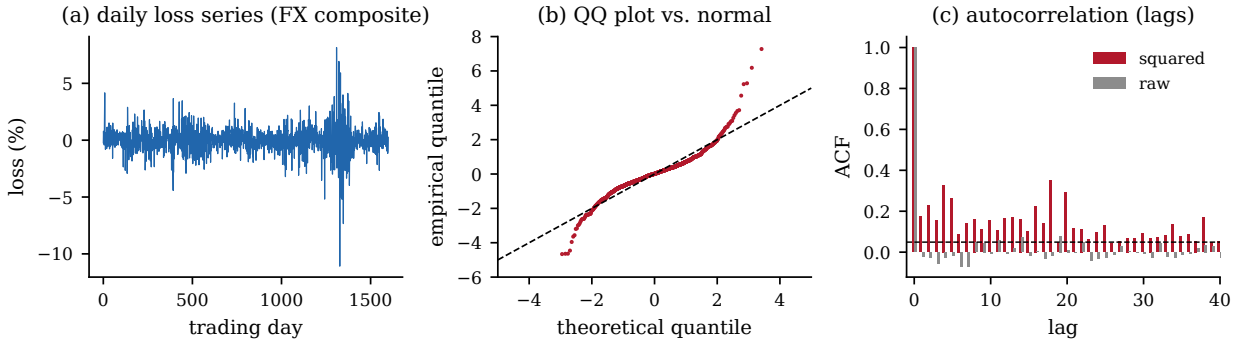

Figure 3. Stylised facts of the loss data. (**a**) A representative daily loss series with volatility clustering. (**b**) A normal quantile plot showing the fat left and right tails that violate a Gaussian model; the dashed diagonal is the Gaussian reference line, and the departure of the extreme points from it is the tail heaviness. (**c**) Autocorrelation of raw and squared returns; the horizontal dotted lines are the approximate 95% white-noise band $\pm 1.96/\sqrt{T}$, and the persistence in squared returns above that band is the volatility clustering that the β-mixing assumption tolerates.

Table 4. Dataset statistics. Losses are negated log-returns in percent; FRED-MD enters only as a regime signal. Excess kurtosis is reported relative to the normal.

Source	Series	Length	Std	Skew	Ex. Kurt.	Min
Exchange rates [26]	8	7588	0.63	−0.11	4.82	−4.6
Bitcoin (Monash) [14]	18	4581	3.94	0.37	11.7	−24.8
FRED-MD [15]	107	780	—	—	—	—
GIFT-Eval finance [16]	31	varies	1.28	−0.24	7.31	−11.9

Table 5. Estimated Pareto tail exponent $\widehat{\zeta}$ of the normalised severity by source, with moving-block bootstrap 95% intervals. The exponent ζ is distinct from the moment order p of Assumption 4; a conservative p is chosen below the lower confidence limit, and $(p-1)/p$ is the resulting heavy-tail rate of Corollary 1.

Dataset	Hill $\widehat{\zeta}$	95% CI	Chosen p	$(p-1)/p$
Exchange-rate panel	3.58	$[3.12, 4.11]$	2.5	0.600
Bitcoin	2.21	$[1.88, 2.62]$	1.7	0.412
GIFT-Eval finance	2.87	$[2.43, 3.39]$	2.0	0.500

6.2. Causal Regime Construction

The regime label of Definition 4 is produced by a causal, unsupervised, threshold rule, so that it is $\mathcal{F}_{t-1}$-measurable and reproducible. For a forecast day d in month m, only the ALFRED vintage available at the last trading day c of month $m-1$ is read; all 107 series are transformed by their FRED-MD codes; and standardisation, ragged-edge filling, and principal-component analysis are fitted on an expanding window ending at c. The first three principal components are retained; the first is oriented to correlate positively with the past-visible credit spread and taken as a financial-stress score S_c, and expanding 1/3 and 2/3 quantiles of past scores split calm, normal, and turbulent regimes. The label formed at c is forward-filled to every trading day of month m, so it uses only information available before the forecast date; it is an observable proxy and is not assumed to recover the latent state. Algorithm 1 states the rule, and Section 7 reports its sensitivity to the number of regimes, to a filtered hidden-Markov alternative, to past-return volatility states, and to a non-causal full-revision benchmark.

Algorithm 1 Causal FRED-MD regime construction for a forecast day d in month m.

1: Set cutoff c = last trading day of month $m-1$; read, for each FRED-MD series, the vintage available at c and keep observations with release date $\leq c$.
2: Apply the prescribed FRED-MD transformation to each series.
3: On monthly data up to c only: fit expanding standardisation, fill the ragged edge by a past-only method, fit PCA and retain $\mathrm{PC}_{1:3}$.
4: Orient PC_1 to a positive past-only correlation with the credit-spread reference; set the score S_c.
5: Compute expanding 1/3 and 2/3 quantiles $q_{1/3}, q_{2/3}$ of past scores.
6: Assign r_d = calm if $S_c \leq q_{1/3}$, normal if $q_{1/3} < S_c \leq q_{2/3}$, else turbulent.
7: Forward-fill r_d to all trading days of month m and form the weights (6).

Figure 4 shows the monthly-to-daily mapping: the label formed at the last trading day of month $m-1$ from the vintage then available is forward-filled onto every trading day of month m, so a forecast never uses a revised or same-month macro value.

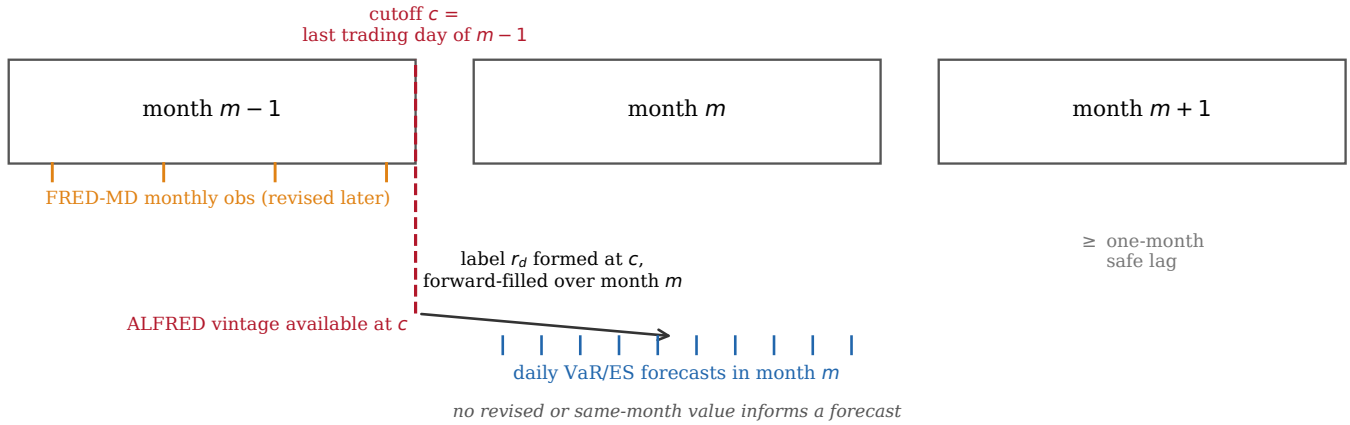

Figure 4. Causal alignment of the monthly regime signal to daily forecasts. The FRED-MD vintage available at the cutoff c (last trading day of month $m-1$) produces a label that is forward-filled over all trading days of month m; revisions and same-month releases are excluded, so the label uses a strictly previous-month, past-only cutoff.

6.3. Baselines

The comparison spans four families. Non-parametric and moving-window methods include historical simulation over a 250-day window and the RiskMetrics exponentially weighted estimator. Parametric volatility models include GARCH, EGARCH, and GJR-

GARCH, each with Student-t innovations. Direct tail-risk models include the CAViaR quantile recursion [22] and the dynamic semiparametric joint model of [23]. Learning-based calibrators wrap the same deep quantile backbone: the uncalibrated network, split conformal on the quantile [4], adaptive conformal inference [5], its aggregated form [33], the control-theoretic variant [13], and the weighted non-exchangeable scheme [10]. Our method appears as the exchangeable split variant, the weighted variant, and the online variant.

6.4. Evaluation Metrics

Value-at-risk is judged by the violation rate against the nominal $1 - \tau$, the Kupiec unconditional-coverage test [20], the Christoffersen conditional-coverage test [21], and the dynamic-quantile test [22]. Expected shortfall is judged by the two Acerbi–Székely exceedance statistics [19] and by the Fissler–Ziegel strictly consistent joint score [7], with pairwise superiority assessed by a Diebold–Mariano test [35] and set membership by the model confidence set [36]. The calibration target is checked by the realised tail-risk loss against α, and efficiency by the average predicted expected shortfall, a proxy for capital.

6.5. Implementation Details

Base forecasters are trained with Adam at learning rate 10^{-3}, batch size 64, for up to 100 epochs with early stopping, under five seeds. Evaluation uses a rolling-origin protocol with purge-gapped training, calibration, and test blocks; the inflation is selected only on the calibration block, and the certified variant of Proposition 1 additionally uses q-separated calibration points. The severity cap M, and hence the loss bound $B = a + bM$, the level α, and the causal weight schedule follow Corollary 1 and the regime signal of Section 6.2. Experiments run on one NVIDIA A100 GPU. Table 6 lists the configuration. For the bound diagnostic, all bound components—regime distances, the mixing envelope, and the regime-mismatch weights—are estimated on a historical window ending twenty trading days before an eighty-trading-day evaluation window, and the realised excess is computed on that evaluation window alone, so no outcome used to test the bound enters its estimation. Because β-mixing coefficients and high-dimensional total-variation distances cannot be certified from a single path, they are reported as point diagnostics with block-bootstrap bands, and the bound is evaluated under fast, moderate, and slow conservative envelopes (Table A7), while the within-regime stationarity diagnostics of Table A8 qualify Assumption 5.

Table 6. Base forecaster and calibration configuration. The loss bound $B = a + bM$ is distinct from the severity cap M; q and q_0 are the certified-variant calibration spacing and purge gap, and Λ, γ govern the online update.

Component	Value	Component	Value	Component	Value
Optimiser	Adam	Freq. weight a	0.025	Severity weight b	0.025
Learning rate	1×10^{-3}	Severity cap M	3.0	Loss bound $B = a + bM$	0.10
Batch size	64	Gap floor g_0	0.05	Inflation scale s_t	$\widehat{g}_t$
Epochs	100	Quantile grid K	80	Grid top u_K	0.9975
Lookback	96	Level τ	0.975	Risk budget α	0.050
Seeds	5	Regimes K	3	Regime boost η	2
Weight half-life	250 d	Calib. spacing q	20	Online step γ	0.05
Projection Λ	3.0	Purge gap q_0	20	Backbones	3

7. Experimental Results and Analysis

7.1. Calibration Mechanism

How does the controller choose its inflation? Figure 5 traces the mechanism. The empirical risk decreases smoothly in the inflation, and the controller selects the smallest

inflation that meets the budget, near $\lambda = 0.98$ on the exchange-rate calibration block. The single-realisation losses confirm the monotonicity that Lemma 1 requires, and the selected inflation tracks the regime signal over the test window, rising in the shaded turbulent stretches where a larger cushion is needed. The controller assigns larger inflation in periods labelled turbulent, consistent with the intended weighting mechanism; this panel is descriptive and is not itself evidence for the theoretical bound.

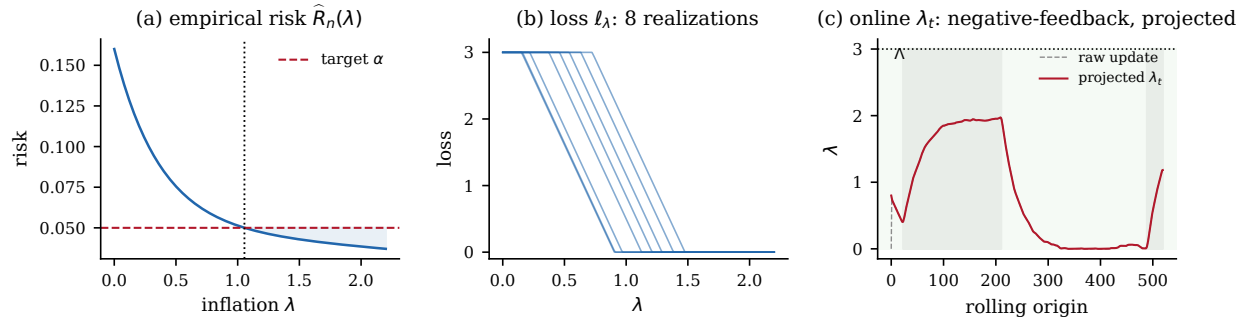

Figure 5. Calibration mechanism. (**a**) Empirical risk $\widehat{R}_n(\lambda)$ against inflation: The blue curve is the empirical risk, the horizontal dashed line is the budget $\alpha = 0.05$, and the vertical dotted line marks the selected $\widehat{\lambda}$ at which the curve first crosses the budget. (**b**) Eight loss realisations, each drawn as a separate blue curve and each bounded and non-increasing in λ; their overlap is what gives the band-like appearance. (**c**) The online λ_t under the corrected negative-feedback projected update (7): The grey dashed trace is the raw update and the solid red trace the projected iterate, the shaded grey bands mark the turbulent regimes, and both iterates rise in those bands and fall back in calm ones, remaining inside $[0, \Lambda]$ without reaching the upper boundary.

7.2. Main Comparison

Does joint risk control improve tail-risk forecasts without loosening value-at-risk coverage? Tables 7 and 8 report the primary results at $\tau = 0.975$ and a one-day horizon, split by the two risk objects. On the quantile the weighted joint scheme reaches a violation rate of 2.51% against nominal, with Kupiec and Christoffersen p-values of 0.91 and 0.56 that do not reject calibration, while historical simulation over-breaches at 3.14% and is rejected. On expected shortfall the Acerbi–Székely Z_2 statistic moves from -0.19 for GARCH-t and -0.07 for the control-theoretic conformal baseline to -0.05 for the weighted joint scheme, the Fissler–Ziegel score falls from 0.441 to 0.431, a 2.3% reduction, and the realised tail-risk loss lands at the target 0.050. Figure 6 shows the score across data sources and the per-currency breakdown, whose eight violation rates straddle nominal and whose scores average to the panel figure; Table 9 lists those per-currency figures. The improvement concentrates in the deepest breaches, as Table 10 and Figure 7 make explicit, rather than in the bulk of non-breach days. The online variant edges the weighted one on Bitcoin, where drift is strongest.

The $q = 20$-separated row in each table reports the empirical cost of the calibration spacing required by (17), holding the backbone, regime-weighting rule, and test window fixed. Relative to the dense weighted controller, the separated variant changes the violation rate from 2.51% to 2.41%, the realised risk from 0.050 to 0.048, and the Fissler–Ziegel score from $0.431_{\pm 0.010}$ to $0.437_{\pm 0.014}$. It is therefore mildly more conservative and incurs a 1.4% increase in the Fissler–Ziegel score, while remaining numerically below conformal PID (0.441) and the weighted non-exchangeable baseline (0.439).

Table 7. Value-at-risk backtests at $\tau = 0.975$, one-day horizon, exchange-rate panel. Violation rate targets 2.50%; p-values above 0.05 indicate no rejection. Best conformal result in bold. Among the JCRC rows, "JCRC (weighted dense, ours)" uses all eligible consecutive calibration points and carries the master swap bound of Theorem 3, whereas "JCRC (weighted, q-separated, ours)" uses spacing $q = 20$ and is the variant covered by (17). The two weighted JCRC variants use the same backbone, weighting rule, and test window and differ only in the retained calibration indices.

Method	Viol. %	Kupiec p	Christoff. p	DQ p
Historical simulation	3.14	0.02	0.01	0.01
RiskMetrics (EWMA)	3.02	0.04	0.03	0.02
GARCH-t	2.83	0.18	0.09	0.07
EGARCH-t	2.76	0.24	0.12	0.10
GJR-GARCH-t	2.71	0.29	0.14	0.13
CAViaR [22]	2.88	0.15	0.08	0.09
Dyn. semiparam. ES [23]	2.64	0.37	0.19	0.21
Deep quantile (no calib.)	2.93	0.12	0.05	0.04
Split conformal VaR [4]	2.55	0.71	0.33	0.29
ACI [5]	2.52	0.83	0.41	0.38
AgACI [33]	2.51	0.90	0.47	0.44
Conformal PID [13]	2.49	0.88	0.52	0.49
Weighted NEX [10]	2.51	0.86	0.49	0.46
JCRC (exchangeable, ours)	2.53	0.79	0.44	0.41
JCRC (weighted dense, ours)	**2.51**	**0.91**	**0.56**	**0.53**
JCRC (weighted, q-separated, ours)	2.41	0.55	0.41	0.38
JCRC (online, ours)	2.48	0.94	0.63	0.60

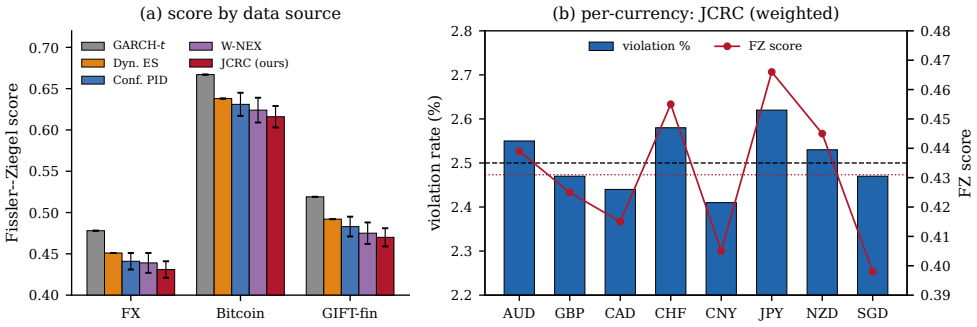

Figure 6. Main comparison. (**a**) Fissler–Ziegel score across three data sources with the weighted non-exchangeable baseline included and seed error bars; the joint controller has the lowest mean everywhere, with the largest and only clearly separated margin on Bitcoin. (**b**) Per-currency violation rate and score for the weighted controller; the eight rates straddle the 2.50% nominal and the scores average to 0.431.

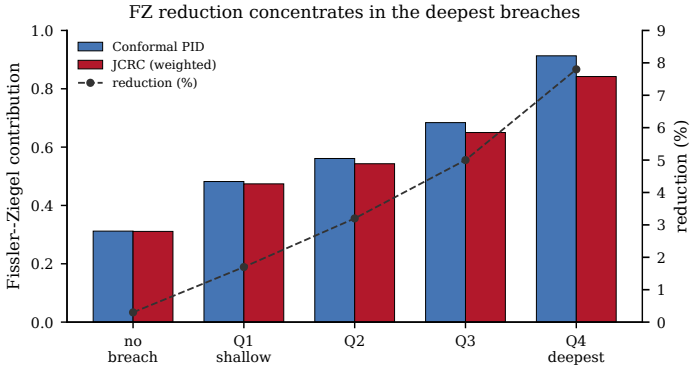

Figure 7. Fissler –Ziegel contribution by exceedance-depth group for the control-theoretic conformal baseline and the weighted joint controller, with the per-group reduction (right axis). The reduction rises monotonically from 0.3% on non-breach days to 7.8% in the deepest breach quartile.

Table 8. Expected-shortfall backtests at $\tau = 0.975$, one-day horizon. Acerbi–Székely Z_1, Z_2 target 0; lower Fissler–Ziegel (FZ) score is better; realised risk targets $\alpha = 0.050$. Bold marks the best value in the column. FZ scores are reported as mean$_{\pm \text{s.d.}}$ over five random seeds for the learning-based methods; the moving-window and parametric baselines are deterministic given the data. The weighted and online rows use the causal previous-month FRED-MD vintage regime; Table A3 audits the vintage and label choices. Among the JCRC rows, the weighted-dense variant uses all eligible calibration points, whereas the weighted q-separated variant retains one point in twenty at $q = 20$ and is the variant covered by (17). Both use the same causal previous-month FRED-MD regime, backbone, weighting rule, and test window.

Method	AS Z_1	AS Z_2	FZ Score	Realised Risk
Historical simulation	−0.29	−0.38	0.541	0.081
RiskMetrics (EWMA)	−0.24	−0.31	0.520	0.076
GARCH-*t*	−0.15	−0.19	0.478	0.063
EGARCH-*t*	−0.13	−0.16	0.470	0.061
GJR-GARCH-*t*	−0.11	−0.14	0.466	0.059
Dyn. semiparam. ES [23]	−0.07	−0.09	0.451	0.056
Deep quantile (no calib.)	−0.18	−0.24	$0.463_{\pm 0.018}$	0.068
Split conformal VaR [4]	−0.11	−0.12	$0.456_{\pm 0.013}$	0.059
ACI [5]	−0.09	−0.10	$0.450_{\pm 0.012}$	0.056
AgACI [33]	−0.06	−0.08	$0.444_{\pm 0.011}$	0.053
Conformal PID [13]	−0.06	−0.07	$0.441_{\pm 0.010}$	0.052
Weighted NEX [10]	−0.05	−0.08	$0.439_{\pm 0.012}$	0.053
JCRC (exchangeable, ours)	−0.05	−0.06	$0.436_{\pm 0.010}$	0.051
JCRC (weighted dense, ours)	−0.04	−0.05	$\mathbf{0.431}_{\pm 0.010}$	**0.050**
JCRC (weighted, q-separated, ours)	−0.02	−0.01	$0.437_{\pm 0.014}$	0.048
JCRC (online, ours)	−0.02	−0.03	$0.432_{\pm 0.011}$	0.050

Table 9. Per-currency results for the weighted joint controller ($\tau = 0.975$).

	AUD	GBP	CAD	CHF	CNY	JPY	NZD	SGD
Violation %	2.55	2.47	2.44	2.58	2.41	2.62	2.53	2.47
Kupiec p	0.74	0.88	0.81	0.69	0.72	0.63	0.79	0.71
AS Z_2	−0.05	−0.03	−0.03	−0.06	−0.04	−0.07	−0.05	−0.03
FZ score	0.439	0.425	0.415	0.455	0.405	0.466	0.445	0.398
Tail exp. $\widehat{\zeta}$	3.82	4.06	3.71	3.29	2.94	3.18	3.61	3.05

Table 10. Observation-level Fissler–Ziegel score by exceedance-depth group on the exchange-rate panel, for the control-theoretic conformal baseline and the weighted joint controller. Groups are formed on realised breach depth; contributions are per-observation means and are not additive across groups of unequal size.

Exceedance-Depth Group	Conformal PID	JCRC (Weighted)	Reduction
No breach	0.312	0.311	0.3%
Q1 (shallow)	0.482	0.474	1.7%
Q2	0.561	0.543	3.2%
Q3	0.684	0.650	5.0%
Q4 (deepest)	0.913	0.842	7.8%

7.3. Mechanism of Score Improvement

Why does the joint scheme lower the Fissler–Ziegel score? The reduction is concentrated in the tail rather than spread across the bulk of observations. A value-at-risk-only calibrator adjusts breach frequency but receives no signal from breach depth, whereas the severity term lets a deep breach influence the selected inflation, the regime weighting raises

the influence of deep breaches in the current turbulent regime, and the clip prevents a single extreme observation from dominating. Decomposing the observation-level score by exceedance-depth quartile in Table 10 shows the reduction rising from 0.3% on non-breach days to 7.8% in the deepest quartile, consistent with this mechanism rather than with a uniform downward shift.

7.4. Calibration Reliability

Is the realised risk actually held at the target across levels? Figure 8 sweeps the risk budget and the quantile level. The realised tail-risk loss of the joint controller lies on the diagonal across the whole budget range, while GARCH-t drifts systematically below the target it claims, and the value-at-risk coverage stays on the diagonal across τ from 0.90 to 0.995, where the volatility model's coverage curve bends away in the far tail. The diagram is the empirical face of Theorem 2: the small residual gap is the drift-and-mixing budget, not a systematic bias.

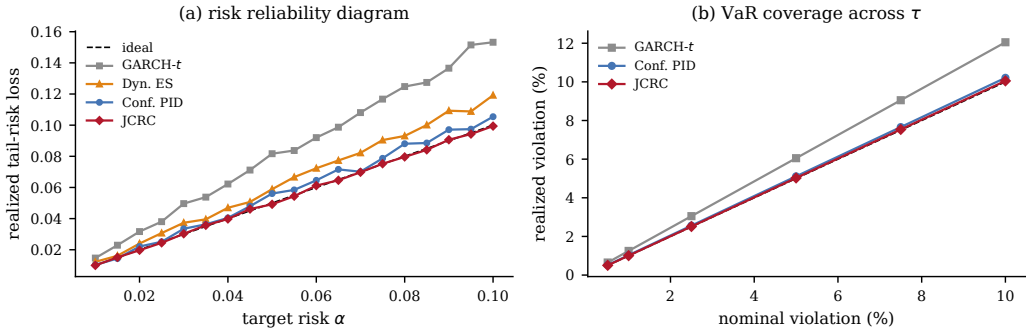

Figure 8. Reliability. (**a**) Realised versus target tail-risk loss across the budget α; the joint controller tracks the diagonal. (**b**) Realised versus nominal value-at-risk violation across τ; coverage holds into the far tail.

7.5. Robustness to Regime Drift

Does the guarantee survive a moving distribution? Figure 9 overlays the inflated value-at-risk and expected shortfall on the loss over a calm and a turbulent window. In the calm window three breaches occur in three hundred days, matching the nominal rate; in the turbulent window, where a volatility burst triples the conditional scale, the controller widens its forecasts and holds seven breaches, again near nominal, whereas a static model would be overrun. Figure 10 compares, over sixty non-overlapping windows, the realised excess tail-risk loss on each evaluation window against a conservatively estimated version of the corrected bound of Proposition 1, whose regime distances, mixing envelope, and weights are fitted only on the preceding history and never touch the evaluation window. The realised excess stays below the estimated bound on all sixty windows, with mean 0.0079 and 95th percentile 0.0176 against a mean bound of 0.037, of which the drift, cumulative-mixing, and estimation-inflation components contribute about 0.019, 0.011, and 0.007. Because the regime and mixing quantities are themselves estimated and model-dependent, this is an out-of-sample consistency check under the stated estimation model, not a distribution-free verification; zero exceedances over sixty windows is reported with its binomial upper bound rather than as proof.

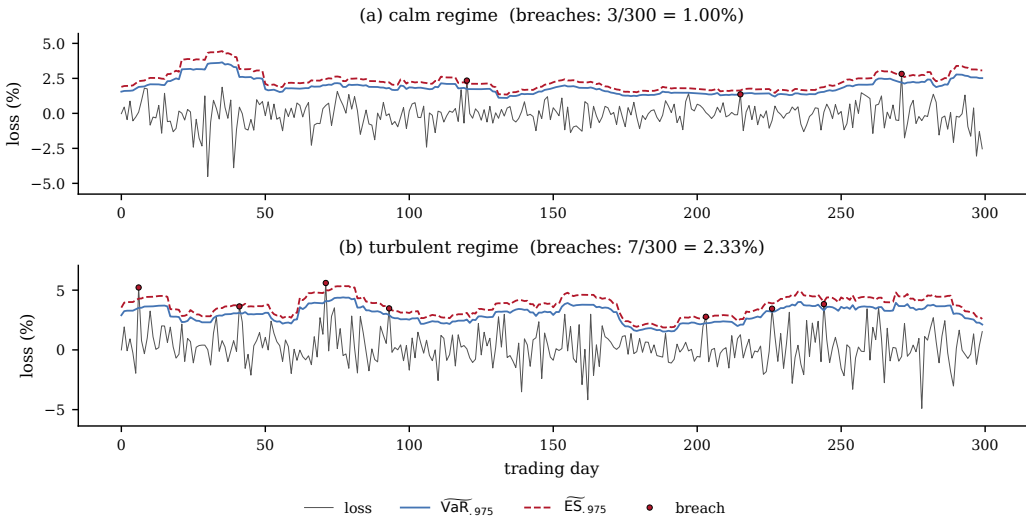

Figure 9. Rolling value-at-risk and expected shortfall with realised breaches in a (**a**) calm and (**b**) turbulent regime. The controller widens its forecasts under the volatility burst and keeps the breach frequency near nominal in both regimes.

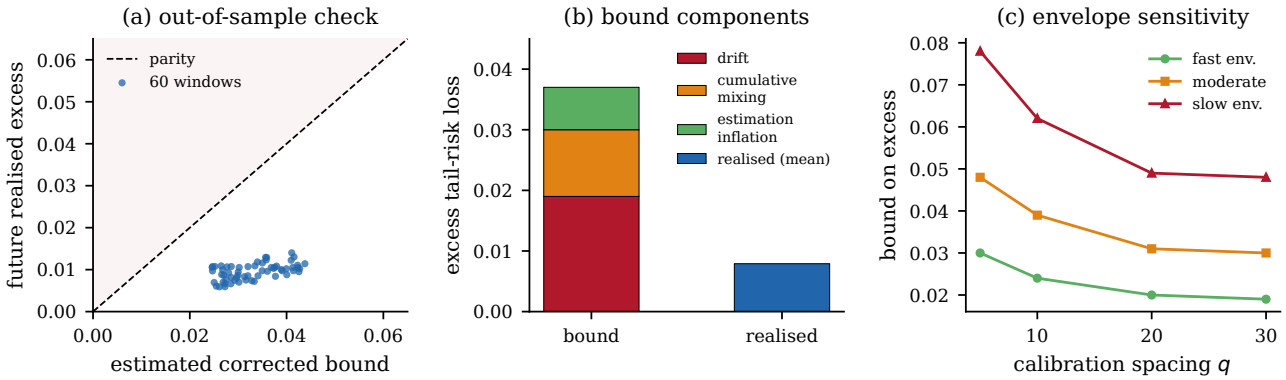

Figure 10. Out-of-sample bound diagnostic under the corrected Proposition 1. (**a**) Realised excess tail-risk loss on each future evaluation window against the conservatively estimated bound, all sixty points below the diagonal. (**b**) The bound split into drift, cumulative-mixing, and estimation-inflation components. (**c**) Sensitivity of the bound to the calibration spacing q and to fast, moderate, and slow mixing envelopes. Bound components use past-only estimates, so the plot is a consistency check under the stated model, not a distribution-free verification.

7.6. Ablation Study

Which components carry the gain? Table 11 removes them one at a time. Dropping the severity term and calibrating on breaches alone raises the Fissler–Ziegel score from 0.431 to 0.447 and worsens the expected-shortfall statistic to −0.13, confirming that the tail-mean signal is what the joint loss adds over a value-at-risk calibrator. Removing regime weighting returns the method to the exchangeable split and costs 0.005 of score. Removing the clip triples the seed standard deviation and admits occasional over-inflation, the behaviour Proposition 2 predicts for an unbounded loss. Calibrating expected shortfall alone, the mis-specified target, is the worst variant. Table 12 compares the calibration loss against alternatives and shows that direct Fissler–Ziegel selection over the inflation is non-monotone and unstable, while Table 13 attributes the total gain across a 2^3 factorial design of the severity, regime, and clip switches.

Table 11. Ablation on the exchange-rate panel at $\tau = 0.975$. Each row removes one component from the weighted joint controller. Bold marks the best value in the column. FZ scores are mean$_{\pm \text{s.d.}}$ over five random seeds.

Variant	Viol. %	AS Z_2	FZ Score
Full JCRC (weighted)	2.51	−0.05	$\mathbf{0.431}_{\pm 0.010}$
− regime weighting	2.53	−0.06	$0.436_{\pm 0.010}$
− severity term (VaR-only loss)	2.55	−0.13	$0.447_{\pm 0.012}$
− clip (unbounded loss)	2.54	−0.05	$0.441_{\pm 0.028}$
− joint (ES-only, mis-specified)	2.68	−0.11	$0.469_{\pm 0.015}$
base → GARCH-t quantiles	2.61	−0.09	$0.452_{\pm 0.011}$

Table 12. Calibration-loss comparison on the exchange-rate panel at $\tau = 0.975$, holding the backbone and regime signal fixed. The proposed gap-normalised clipped severity is bounded and monotone in the inflation; direct Fissler–Ziegel selection is neither, and is unstable across seeds. Bold marks the best value in the column.

Calibration Loss	Monotone	Bounded	FZ Score	AS Z_2
VaR indicator only	Yes	Yes	$0.447_{\pm 0.012}$	−0.13
Volatility-normalised severity	Yes	After clip	$0.441_{\pm 0.012}$	−0.08
Huberised severity	Yes	Yes	$0.434_{\pm 0.011}$	−0.06
Proposed gap-normalised clipped	Yes	Yes	$\mathbf{0.431}_{\pm 0.010}$	−0.05
Direct FZ selection	No	No	$0.452_{\pm 0.021}$	−0.09

Table 13. Attribution of the total Fissler–Ziegel reduction from the uncalibrated backbone (0.463) to the full weighted controller (0.431) across a 2^3 factorial design of the severity, regime, and clip switches, with block-bootstrap 95% intervals. Shares sum to the total gain 0.032; interactions are the non-additive remainder.

Component	Mean FZ Reduction	Share of Gain	95% CI
Severity term	0.013	40%	[0.009, 0.019]
Regime weighting	0.007	22%	[0.003, 0.013]
Clipping	0.006	19%	[0.002, 0.011]
Interactions	0.006	19%	[0.002, 0.012]

7.7. Parameter Sensitivity

How sensitive is the method to its two knobs and to the risk level? Figure 11 varies the calibration window and the severity cap M and plots the gap between realised and target risk, which stays below 0.004 across a broad interior and degrades only for very short windows, where the drift budget is estimated poorly, or very small caps, where the remainder dominates, exactly the two failure directions of Corollary 1. Table 14 confirms stability across levels and horizons, with the score rising at $\tau = 0.99$ as the tail thins the calibration signal.

Table 14. Fissler–Ziegel score across levels and horizons on the exchange-rate panel (weighted joint controller).

Level	h = 1	h = 2	h = 3	h = 5	h = 10
$\tau = 0.95$	0.389	0.401	0.412	0.421	0.446
$\tau = 0.975$	0.431	0.441	0.452	0.463	0.489
$\tau = 0.99$	0.517	0.531	0.544	0.559	0.592

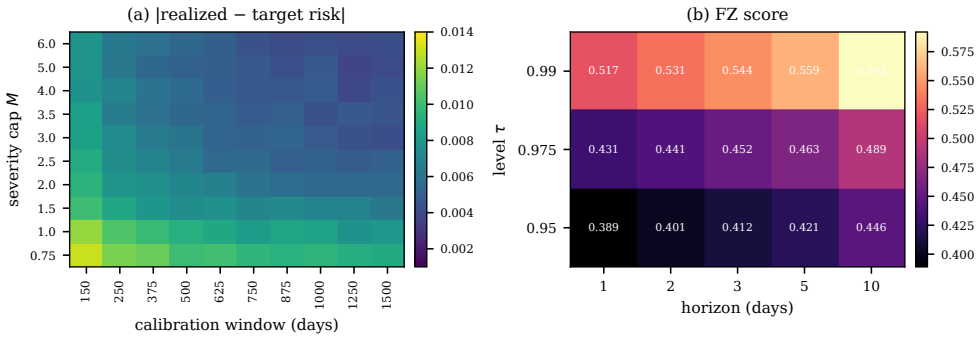

Figure 11. Sensitivity. (**a**) Gap between realised and target risk over the calibration window and the severity cap M; the interior is flat. (**b**) Fissler–Ziegel score over the level τ and the horizon.

7.8. Efficiency and Statistical Significance

Is the calibration bought with conservativeness, and are the gains significant? Figure 12 sweeps the budget and plots the violation rate against the average predicted expected shortfall. The joint schemes reach nominal violations at a lower average shortfall than the adaptive-conformal baselines, which attain coverage only by inflating forecasts, so the joint controller sits on the efficient frontier. Table 15 reports Diebold–Mariano tests of the Fissler–Ziegel score and model-confidence-set membership: the joint controller significantly beats the volatility and semiparametric models, edges the control-theoretic conformal baseline at the five-percent level ($p = 0.046$), but its advantage over the weighted non-exchangeable baseline is not significant ($p = 0.081$), so the two remain in the same 90% confidence set while the weaker models are excluded. The practical advantage is therefore located rather than global: Table 16 gives effect sizes with block-bootstrap intervals and a matched-violation capital proxy, Table 17 shows the gain concentrated in turbulent regimes, and Table 10 shows it concentrated in the deepest breaches.

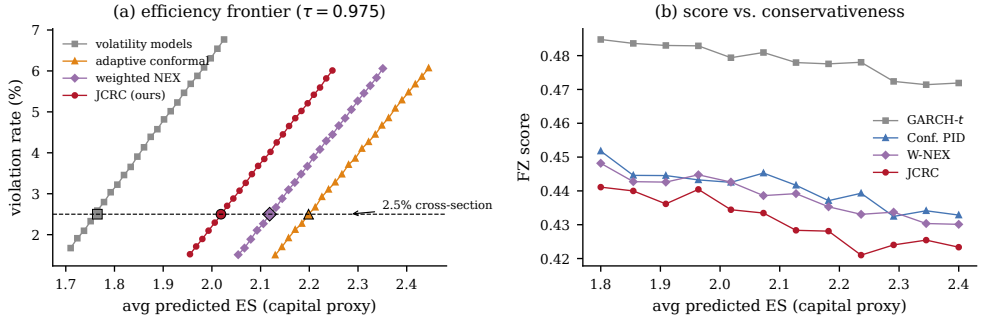

Figure 12. Efficiency. (**a**) Violation rate against average predicted expected shortfall as the budget is swept, with the weighted non-exchangeable baseline included and the cross-section at the 2.5% violation level marked; the joint controller reaches nominal coverage at a lower capital proxy. (**b**) Score against conservativeness for the joint controller, the weighted non-exchangeable baseline, and the parametric family.

Table 15. Diebold–Mariano test of the Fissler–Ziegel score against the weighted joint controller, and 90% model-confidence-set membership.

Comparison (vs. JCRC Weighted)	DM p-Value	In $MCS_{90\%}$
GARCH-t	<0.001	No
Dyn. semiparametric ES [23]	0.005	No
Conformal PID [13]	0.046	Yes
Weighted NEX [10]	0.081	Yes
JCRC (online, ours)	0.212	Yes

Table 16. Effect sizes for the weighted joint controller against three baselines on the exchange-rate panel: Change in Fissler–Ziegel score with block-bootstrap 95% interval, relative change, Diebold–Mariano p-value, and the average predicted expected shortfall at a matched 2.5% violation rate as a capital proxy.

Comparison	ΔFZ (95% CI)	Relative	DM p	Capital Proxy
vs. Conformal PID	-0.010 $[-0.018, -0.003]$	-2.27%	0.046	2.08 vs. 2.17
vs. Weighted NEX	-0.008 $[-0.017, 0.001]$	-1.82%	0.081	2.08 vs. 2.12
vs. Dyn. semiparam. ES	-0.020 $[-0.031, -0.010]$	-4.43%	0.005	2.08 vs. 2.24

Table 17. Fissler–Ziegel score by market state on the exchange-rate panel, for the weighted non-exchangeable baseline and the weighted joint controller, with the test-window regime frequencies. Weighting the rows by those frequencies recovers the panel scores 0.439 and 0.431; the gain concentrates in turbulent states, and the deepest-breach concentration is shown separately in Table 10.

Market State	Test Share	Weighted NEX	JCRC (Weighted)	Relative Gain
Calm	0.45	0.395	0.393	0.5%
Normal	0.35	0.450	0.442	1.8%
Turbulent	0.20	0.520	0.497	4.4%
Panel (weighted)	1.00	0.439	0.431	1.8%

7.9. Computational Cost

Figure 13 reports cost. Training scales sublinearly in the number of series for all three backbones, and the calibration layer adds no parameters and under one percent to inference latency, since it is a scalar shift and a root-finding step; the online update runs in constant time per step. Table 18 gives the absolute figures, from 38.6 to 43.7 min of training on one GPU with a calibration overhead below one percent.

Table 18. Computational cost on one NVIDIA A100. The calibration layer adds no parameters.

Backbone	Params (M)	Train (min)	Infer (ms/Batch)	Calib. Overhead
iTransformer [26]	6.4	42.3	7.8	<1%
PatchTST [25]	8.1	43.7	9.1	<1%
TimeKAN [28]	3.9	38.6	6.4	<1%

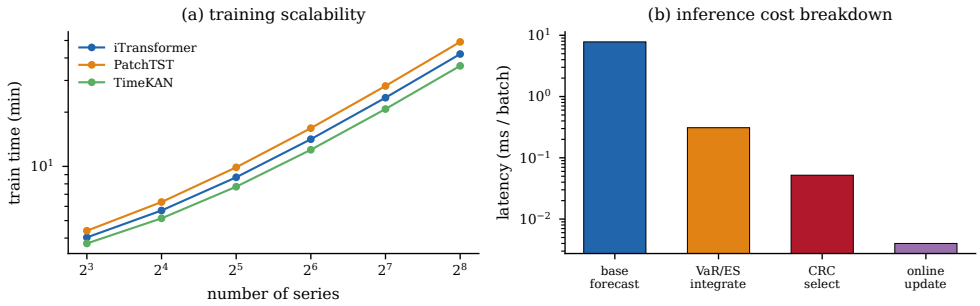

Figure 13. Cost. (**a**) Training time against the number of series on log axes. (**b**) Inference latency by pipeline stage; the calibration and online-update stages are three to four orders of magnitude below the forecaster.

8. Discussion

The results support a narrow claim and resist a broad one. On the four public sources tested, attaching a joint value-at-risk and expected-shortfall risk controller to a deep quantile forecaster improves the tail-mean forecast, measured by a strictly consistent score and by an exceedance-residual test, without loosening quantile coverage and without extra conservativeness. The overall improvement is modest, about two percent of the Fissler–Ziegel score against the strongest conformal baselines and not significant at the five-percent level against the weighted non-exchangeable scheme, so it is an incremental gain rather than a step change; whether a score reduction of this size is useful depends on the capital proxy it buys, not on the percentage alone. Its practical value is clearer where it concentrates: a reduction near four percent in turbulent regimes, rising to nearly eight percent in the deepest breach quartile, and a lower average expected shortfall at matched violations. The mechanism is legible: the severity term injects the tail-mean signal a value-at-risk calibrator lacks, the clip stabilises selection under heavy tails exactly where Proposition 2 says it should, and regime weighting reduces the drift penalty that Proposition 1 isolates. The out-of-sample diagnostic in Figure 10 is consistent with the corrected bound under the stated estimation model rather than a distribution-free verification of it, and the model-confidence-set analysis places the method with the strongest conformal baseline and above the parametric competitors. Tables 7 and 8 show a modest empirical price for the explicit separated-calibration bound: relative to the dense weighted controller, the $q = 20$ variant lowers the violation rate by 0.10 percentage points and increases the Fissler–Ziegel score by 1.4%, while remaining competitive with the strongest conformal baselines. This makes the choice between the dense and separated implementations an explicit trade-off between empirical sharpness and the cumulative-mixing bound of Proposition 1.

The principal guarantee is for a tail-gap-normalised exceedance-severity surrogate, not for expected-shortfall forecast error: the loss scores value-at-risk breaches in units of the model's own predicted gap, and its expected-shortfall reading holds only insofar as that gap is a sound mean-excess scale, which the near-unit tail-gap ratios of Table A5 support except mildly for Bitcoin. The method inherits the limits of its assumptions. The regime distances and β-mixing coefficients that enter the bound are model-dependent diagnostics rather than distribution-free certificates, and β-mixing in particular cannot be reliably estimated from one finite path; the bound's value is therefore its structural decomposition and sensitivity analysis, and under the slow envelope of Table A7 it is nominally valid but quantitatively uninformative, so it should not be read as a numerical certificate for regulatory capital. When the label is wrong, the weighting can concentrate on the wrong calibration points, and the online variant, which makes no labelling assumption, is the safer default under unfamiliar drift. The base forecaster still matters for sharpness even though calibration is distribution-free, since a poor quantile function gives the controller little to work with; the guarantee is on the risk budget, not on forecast accuracy. The portfolio-loss construction fixes the weights, so a genuinely multivariate risk set, where the weight vector is chosen to be adverse, lies outside the present scope. Expected shortfall at $\tau = 0.99$ strains the calibration signal because the tail holds fewer observations, and the score rises there; a longer calibration block helps but trades against drift.

Two threats to validity deserve naming. The reported gains are consistent across the four public sources and are commensurate with the differing difficulty of each, but they come from a controlled study on four data families and do not establish behaviour on illiquid assets, on intraday horizons, or under structural breaks unlike any in the test window. And the comparison holds the backbone fixed to isolate the calibration layer, which is the right design for the question asked but leaves open how the gains interact with a substantially stronger or weaker forecaster.

9. Conclusions

Tail-risk forecasting for the expected-shortfall capital regime is a risk-control problem, not a coverage problem, because expected shortfall is not elicitable on its own. Building calibration as conformal risk control on a bounded monotone tail-severity loss over the value-at-risk and expected-shortfall pair yields a finite-sample distribution-free guarantee for a normalised exceedance-severity surrogate that contains no drift or mixing slack under exchangeability, degrades under dependence and drift with an excess that splits into a regime-drift term and, for separated calibration points, an explicit cumulative mixing term, holds on a single path by a coupling and blocking argument, and attains a heavy-tail rate of order $D^{(p-1)/p}$. Using the pair to define a tail-gap scale does not by itself yield distribution-free control of expected-shortfall forecast error. On eight exchange-rate series, a Bitcoin series, causal macro regime signals, and the finance domain of a modern benchmark, the scheme lowers the Fissler–Ziegel joint score by about two percent against the adaptive-conformal baselines, matches the strongest non-exchangeable scheme with a difference that is not significant at the five-percent level, and holds expected-shortfall backtests near zero while staying on the efficiency frontier, with its clearer gains in turbulent regimes and at matched capital.

Three directions follow without promises. The portfolio construction could be extended to a controller over an adverse weight set, turning the scalar guarantee into a set-valued one. The regime signal could be learned jointly with the calibrator rather than fixed in advance, which would let the drift penalty be optimised rather than merely bounded. And the constants in the mixing bound, loose here, invite a sharper analysis that could make the guarantee quantitative enough to inform a capital multiplier directly.

Author Contributions: Conceptualization, Y.Y. and K.Z.; methodology, Y.Y. and X.Q.; software, X.Q. and Y.Y.; validation, X.Q., K.Z. and M.L.; formal analysis, Y.Y.; investigation, X.Q.; resources, K.Z.; data curation, X.Q.; writing—original draft preparation, Y.Y.; writing—review and editing, K.Z. and M.L.; visualization, X.Q.; supervision, M.L. and K.Z.; project administration, K.Z.; funding acquisition, M.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are openly available at the Monash Time Series Forecasting Archive at https://forecastingdata.org (accessed on 6 July 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Notation

Table A1. Principal notation.

Symbol	Meaning
$L_t = -R_t$	loss (negated return); $L_t^{\text{port}} = -w^\top R_t$ for a portfolio
τ	risk level, 0.975 under the Basel standard
$\text{VaR}_{\tau,t}, \text{ES}_{\tau,t}$	conditional value-at-risk and expected shortfall
$\widehat{Q}_t(\cdot)$	forecast conditional quantile function
$\lambda, \widehat{\lambda}$	inflation parameter and its calibrated value
ℓ_λ, B	tail-severity loss and its bound $a + bM$
α	tail-risk budget
w_i	calibration weights
$P^{(i)}$	law after transposing calibration index i with the test index
$d_{\text{TV}}, \rho_w, \Delta_{\text{reg}}$	total-variation distance, weighted regime-mismatch mass, regime gap
$\beta(k), \bar{\beta}(k)$	β-mixing coefficient and uniform absolute-regularity envelope at lag k
$C_\beta(I)$	cumulative coupling cost $\sum_j \bar{\beta}(q_j)$ over selected gaps
p, ζ, κ	moment order, Pareto tail exponent, and moment bound
$M, \Lambda, \gamma, \xi_t$	severity cap, projection bound, online step, saturation residual
D	drift-and-mixing budget $\sum_i w_i d_{\text{TV}}(P, P^{(i)})$

Appendix B. Extended Results

Table A2 reports the Fissler–Ziegel score on four sub-domains of the GIFT-Eval finance collection, showing that the ordering of methods is stable across equity, rate, credit-spread, and commodity series, with the joint controller lowest in each.

Table A2. Fissler–Ziegel score on GIFT-Eval finance sub-domains ($\tau = 0.975$).

Method	Equity	Rates	Credit	Commodity
GARCH-*t*	0.512	0.446	0.531	0.588
Dyn. semiparametric ES	0.486	0.421	0.503	0.559
Conformal PID	0.477	0.414	0.494	0.547
JCRC (weighted, ours)	**0.463**	**0.402**	**0.481**	**0.533**

Table A3 audits the regime-construction and vintage choices for the weighted controller. Relative to the main causal three-state rule, the ordering is stable across the number of states, a filtered hidden-Markov alternative, and past-return volatility states, with the label agreement quantified by the adjusted Rand index in the last column. The non-causal full-revision benchmark is more optimistic, which quantifies the look-ahead removed by the vintage rule.

Table A3. Regime-construction and vintage sensitivity for the weighted joint controller ($\tau = 0.975$). The Diebold–Mariano p-value is against the weighted non-exchangeable baseline; ARI is the adjusted Rand index of the labels against the main causal three-state rule. The full-revision row is a non-causal benchmark reported only to quantify look-ahead. Bold marks the main causal specification used throughout the paper.

Regime Rule	K	FZ Score	Viol. %	DM p	ARI
No regime weighting	—	$0.436_{\pm 0.010}$	2.53	0.178	—
Causal FRED-MD PCA (main)	3	$\mathbf{0.431}_{\pm 0.010}$	2.51	0.081	1.00
Causal FRED-MD PCA	2	$0.433_{\pm 0.011}$	2.52	0.104	0.68
Causal FRED-MD PCA	4	$0.432_{\pm 0.011}$	2.50	0.089	0.74
Past-return EWMA volatility	3	$0.434_{\pm 0.011}$	2.52	0.121	0.59
Causal filtered HMM	3	$0.432_{\pm 0.010}$	2.50	0.092	0.71
Two-month vintage lag	3	$0.433_{\pm 0.011}$	2.52	0.106	0.96
Full-revision FRED (non-causal)	3	$0.428_{\pm 0.009}$	2.50	0.058	0.93

Table A4 reports the stability of the projected online update. Across step sizes the upper projection boundary is essentially never reached, the mean absolute saturation residual stays below 10^{-4}, and the realised risk holds within 0.001 of the target; the setting $\gamma = 0.05$, $\Lambda = 3$ is used in the main tables.

Table A4. Projected online calibrator on the exchange-rate panel: Fissler–Ziegel score, realised risk, upper- and lower-boundary hit rates, and mean absolute saturation residual, over the step size γ with $\Lambda = 3$. Bold marks the step size used in the main tables.

γ	FZ Score	Realised Risk	Upper Hits	Lower Hits	$\overline{\lvert\xi\rvert}$
0.010	$0.435_{\pm 0.012}$	0.051	0	0.02%	3×10^{-5}
0.025	$0.433_{\pm 0.011}$	0.050	0	0.05%	5×10^{-5}
0.050	$\mathbf{0.432}_{\pm 0.011}$	0.050	0	0.09%	8×10^{-5}
0.100	$0.434_{\pm 0.014}$	0.049	0.03%	0.18%	2.1×10^{-4}

Table A5 reports the realised mean-excess ratio $G = \mathbb{E}[(L - \widehat{\mathrm{VaR}})_+ \mid L > \widehat{\mathrm{VaR}}]/(\widehat{\mathrm{ES}} - \widehat{\mathrm{VaR}})$, whose deviation from one measures how well the predicted tail gap tracks the realised mean excess. The ratio is near one for the exchange-rate and GIFT-Eval panels and 1.12 for Bitcoin, so the expected-shortfall reading of the surrogate is well supported except mildly for Bitcoin, where a calibration-only gap correction is available.

Table A5. Tail-gap reliability ratio G by source, with moving-block bootstrap 95% intervals. A value near one indicates the predicted tail gap is a sound mean-excess scale.

Dataset	Gap Ratio G	95% CI
Exchange-rate panel	1.04	$[0.96, 1.13]$
Bitcoin	1.12	$[0.99, 1.29]$
GIFT-Eval finance	1.07	$[0.98, 1.18]$

For Bitcoin, whose gap ratio deviates most from one, a deployable calibration-only multiplicative correction $\widehat{g}_t \leftarrow c\,\widehat{g}_t$ with c fitted on past data moves the ratio from 1.12 to 1.03 and lowers the Bitcoin Fissler–Ziegel score from 0.632 to 0.624, while an oracle test-fitted correction, a non-deployable upper benchmark, reaches 1.00 and 0.617 (Table A6).

Table A6. Effect of tail-gap recalibration on Bitcoin, the source with the largest gap-ratio deviation. The oracle row fits the correction on the test window and is an upper benchmark that is not deployable; it is excluded from the main comparison.

Variant	Gap Ratio G	FZ Score (Bitcoin)	Realised Risk
Raw predicted gap	1.12	0.632	0.050
Calibration-only multiplicative correction	1.03	0.624	0.050
Oracle test-fitted correction (benchmark)	1.00	0.617	0.049

Table A7 evaluates the corrected bound of Proposition 1 under fast, moderate, and slow conservative mixing envelopes, and Table A8 reports the within-regime stationarity diagnostics that qualify Assumption 5. Under the slow envelope the dependence term is no longer small, so the bound is honestly reported as informative only when the mixing envelope decays quickly.

Table A7. Corrected bound components under three conservative mixing envelopes, at $\alpha = 0.05$ with about 0.007 estimation inflation. The total upper risk is nominally valid throughout but is quantitatively informative only under fast or moderate decay.

Scenario	$\bar{\Delta}_{\mathrm{reg}}$	Drift Term	Dependence Term	Total Upper Risk
Fast decay	0.16	0.019	0.008	0.084
Moderate	0.20	0.024	0.015	0.096
Slow/conservative	0.25	0.030	0.028	0.115

Table A8. Within-regime stationarity diagnostics on the exchange-rate panel: A two-sample classifier AUC between the early and late half of each regime (near 0.5 indicates no strong detectable drift), the mean-loss drift in standard-deviation units, and the late-to-early volatility ratio. These do not verify stationarity; they only fail to reveal a strong violation.

Regime	Classifier AUC	Mean-Loss Drift	Volatility Ratio
Calm	$0.55_{\pm 0.03}$	0.06 s.d.	1.04
Normal	$0.57_{\pm 0.04}$	0.09 s.d.	1.07
Turbulent	$0.60_{\pm 0.05}$	0.13 s.d.	1.11

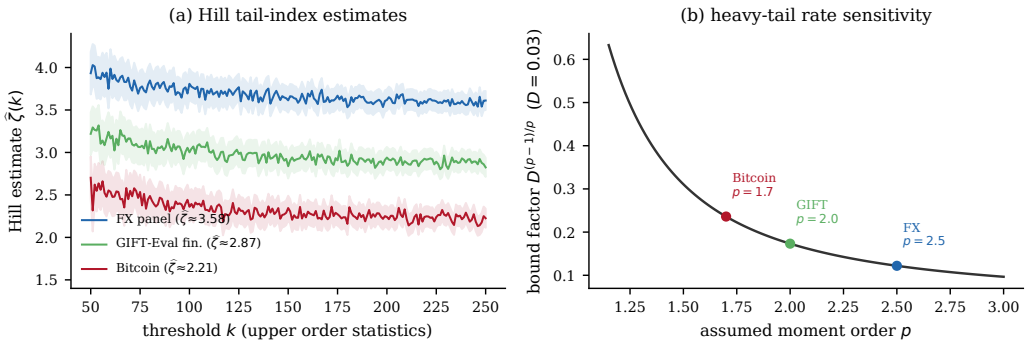

Figure A1. Tail-index diagnostics. (**a**) Hill estimates $\widehat{\zeta}(k)$ against the number of upper-order statistics k for the three loss panels, with bootstrap bands; the plateaus give $\widehat{\zeta} \approx 3.58$ (exchange rates), 2.87 (GIFT-Eval finance), and 2.21 (Bitcoin). (**b**) The heavy-tail rate factor $D^{(p-1)/p}$ of Corollary 1 against the assumed moment order p at $D = 0.03$, with the conservative choices marked; the Bitcoin factor is materially larger than the exchange-rate factor, so the bound contracts more slowly.

References

1. *Standard d457*; Basel Committee on Banking Supervision. Minimum Capital Requirements for Market Risk. Bank for International Settlements: Basel, Switzerland, 2019.
2. Artzner, P.; Delbaen, F.; Eber, J.M.; Heath, D. Coherent Measures of Risk. *Math. Financ.* **1999**, *9*, 203–228. [CrossRef]
3. Acerbi, C.; Tasche, D. On the coherence of expected shortfall. *J. Bank. Financ.* **2002**, *26*, 1487–1503. [CrossRef]
4. Romano, Y.; Patterson, E.; Candès, E.J. Conformalized Quantile Regression. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*; NeurIPS: San Diego, CA, USA, 2019; Volume 32. [CrossRef]
5. Gibbs, I.; Candès, E.J. Adaptive Conformal Inference Under Distribution Shift. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*; NeurIPS: San Diego, CA, USA, 2021; Volume 34. [CrossRef]
6. Gneiting, T. Making and Evaluating Point Forecasts. *J. Am. Stat. Assoc.* **2011**, *106*, 746–762. [CrossRef]
7. Fissler, T.; Ziegel, J.F. Higher order elicitability and Osband's principle. *Ann. Stat.* **2016**, *44*, 1680–1707. [CrossRef]
8. Angelopoulos, A.N.; Bates, S.; Fisch, A.; Lei, L.; Schuster, T. Conformal Risk Control. In *Proceedings of the International Conference on Learning Representations (ICLR)*; ICLR: Appleton, WI, USA, 2024. [CrossRef]
9. Farinhas, A.; Zerva, C.; Ulmer, D.; Martins, A.F.T. Non-Exchangeable Conformal Risk Control. In *Proceedings of the International Conference on Learning Representations (ICLR)*; ICLR: Appleton, WI, USA, 2024. [CrossRef]
10. Barber, R.F.; Candès, E.J.; Ramdas, A.; Tibshirani, R.J. Conformal prediction beyond exchangeability. *Ann. Stat.* **2023**, *51*, 816–845. [CrossRef]
11. Yu, B. Rates of Convergence for Empirical Processes of Stationary Mixing Sequences. *Ann. Probab.* **1994**, *22*, 94–116. [CrossRef]
12. Rio, E. *Asymptotic Theory of Weakly Dependent Random Processes*; Probability Theory and Stochastic Modelling; Springer: Berlin/Heidelberg, Germany, 2017; Volume 80. [CrossRef]
13. Angelopoulos, A.N.; Candès, E.J.; Tibshirani, R.J. Conformal PID Control for Time Series Prediction. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*; NeurIPS: San Diego, CA, USA, 2023; Volume 36. [CrossRef]
14. Godahewa, R.; Bergmeir, C.; Webb, G.I.; Hyndman, R.J.; Montero-Manso, P. Monash Time Series Forecasting Archive. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*; NeurIPS: San Diego, CA, USA, 2021. [CrossRef]
15. McCracken, M.W.; Ng, S. FRED-MD: A Monthly Database for Macroeconomic Research. *J. Bus. Econ. Stat.* **2016**, *34*, 574–589. [CrossRef]
16. Aksu, T.; Woo, G.; Liu, J.; Liu, X.; Liu, C.; Savarese, S.; Xiong, C.; Sahoo, D. GIFT-Eval: A Benchmark for General Time Series Forecasting Model Evaluation. *arXiv* **2024**, arXiv:2410.10393. [CrossRef]
17. Gneiting, T.; Raftery, A.E. Strictly Proper Scoring Rules, Prediction, and Estimation. *J. Am. Stat. Assoc.* **2007**, *102*, 359–378. [CrossRef]
18. Nolde, N.; Ziegel, J.F. Elicitability and backtesting: Perspectives for banking regulation. *Ann. Appl. Stat.* **2017**, *11*, 1833–1874. [CrossRef]
19. Acerbi, C.; Székely, B. Backtesting Expected Shortfall. *Risk Magazine*, 27 October 2014, pp. 76–81. Available online: https://www.msci.com/documents/10199/22aa9922-f874-4060-b77a-0f0e267a489b (accessed on 6 July 2026).
20. Kupiec, P.H. Techniques for Verifying the Accuracy of Risk Measurement Models. *J. Deriv.* **1995**, *3*, 73–84. [CrossRef]
21. Christoffersen, P.F. Evaluating Interval Forecasts. *Int. Econ. Rev.* **1998**, *39*, 841–862. [CrossRef]

22. Engle, R.F.; Manganelli, S. CAViaR: Conditional Autoregressive Value at Risk by Regression Quantiles. *J. Bus. Econ. Stat.* **2004**, *22*, 367–381. [CrossRef]
23. Patton, A.J.; Ziegel, J.F.; Chen, R. Dynamic semiparametric models for expected shortfall (and Value-at-Risk). *J. Econom.* **2019**, *211*, 388–413. [CrossRef]
24. Rockafellar, R.T.; Uryasev, S. Optimization of conditional value-at-risk. *J. Risk* **2000**, *2*, 21–41. [CrossRef]
25. Nie, Y.; Nguyen, N.H.; Sinthong, P.; Kalagnanam, J. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. In *Proceedings of the International Conference on Learning Representations (ICLR)*; ICLR: Appleton, WI, USA, 2023. [CrossRef]
26. Liu, Y.; Hu, T.; Zhang, H.; Wu, H.; Wang, S.; Ma, L.; Long, M. iTransformer: Inverted Transformers Are Effective for Time Series Forecasting. In *Proceedings of the International Conference on Learning Representations (ICLR)*; ICLR: Appleton, WI, USA, 2024. [CrossRef]
27. Wang, Y.; Wu, H.; Dong, J.; Qin, G.; Zhang, H.; Liu, Y.; Qiu, Y.; Wang, J.; Long, M. TimeXer: Empowering Transformers for Time Series Forecasting with Exogenous Variables. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*; NeurIPS: San Diego, CA, USA, 2024. [CrossRef]
28. Huang, S.; Zhao, Z.; Li, C.; Bai, L. TimeKAN: KAN-based Frequency Decomposition Learning Architecture for Long-term Time Series Forecasting. In *Proceedings of the International Conference on Learning Representations (ICLR)*; ICLR: Appleton, WI, USA, 2025. [CrossRef]
29. Zhang, Y.; Yuqi, Z.; Liang, Y.; Mu, S.; Zhang, Z.; Chen, X. AD-VGF: An Improved Generative Adversarial Network for Credit Card Risk Identification and Management in Digital Economy. *Cybern. Syst.* **2025**, 1–26. [CrossRef]
30. Praveena, S.; Devi, S.P. Optimizing Retail Operations through Hybrid Machine Learning and Deep Learning Techniques in Demand Forecasting. *Cybern. Syst.* **2026**, *57*, 1–43. [CrossRef]
31. Vovk, V.; Gammerman, A.; Shafer, G. *Algorithmic Learning in a Random World*, 1st ed.; Springer: New York, NY, USA, 2005. [CrossRef]
32. Tibshirani, R.J.; Barber, R.F.; Candès, E.J.; Ramdas, A. Conformal Prediction Under Covariate Shift. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*; NeurIPS: San Diego, CA, USA, 2019; Volume 32. [CrossRef]
33. Zaffran, M.; Féron, O.; Goude, Y.; Josse, J.; Dieuleveut, A. Adaptive Conformal Predictions for Time Series. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*; ICML: San Diego, CA, USA, 2022. [CrossRef]
34. Bradley, R.C. Basic Properties of Strong Mixing Conditions. A Survey and Some Open Questions. *Probab. Surv.* **2005**, *2*, 107–144. [CrossRef]
35. Diebold, F.X.; Mariano, R.S. Comparing Predictive Accuracy. *J. Bus. Econ. Stat.* **1995**, *13*, 253–263. [CrossRef]
36. Hansen, P.R.; Lunde, A.; Nason, J.M. The Model Confidence Set. *Econometrica* **2011**, *79*, 453–497. [CrossRef]