

When a Winning Forecast Does Not Identify an Action: An Evidence-Budget Algorithm for Distributional Decisions

Min Huang

`minhuang1@link.cuhk.edu.cn`

Chronoedge Limited

Mingyan Liu

Chronoedge Limited

Xiaoer Li

Chronoedge Limited

Research Article

Keywords: distribution forecasting, model uncertainty, action resolution, evidence budget, computational economics, value at risk

Posted Date: July 31st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10524572/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

When a Winning Forecast Does Not Identify an Action: An Evidence-Budget Algorithm for Distributional Decisions

Min Huang¹

Mingyan Liu

Xiaoer Li

Chronoedge Limited, Shenzhen, Guangdong, China

Abstract

Distributional forecasting systems often select the model with the best validation score and treat its implied action as resolved. Finite validation data, however, may rank models without excluding other distributions that imply materially different actions. We introduce the Finite-Evidence Decision Identification Procedure (FEDIP), a modular audit that applies a declared competitive-model screen, projects retained distributions into a prespecified action space, and reports the unresolved action range. “Decision identification” is procedural: it means contraction of candidate-implied actions under the declared system, not structural identification. Validation length is supplied evidence rather than an estimated optimal budget. An oracle concentration benchmark shows that evidence demand rises with score noise and candidate-space size and may be unbounded when consequential alternatives are nonseparable under the declared score margin. A Monte Carlo factorial conducted under a locally frozen protocol covers 120 conditions and 48,000 replication-condition runs. Expanding from five to nine models increases the standardized 5% value-at-risk range more at 20 than at 500 validation observations: the stable-process interaction is 0.0526 (95% Monte Carlo interval 0.0440–0.0612), concentrated in mixture and skewed processes. In 2,710 rolling blocks for 11 assets, the screen retains 8.025–8.976 of nine models on average, indicating weak discrimination. A maximum-threshold policy reduces violations from 5.06% to 3.99% but raises the standardized threshold by 9.47% relative to the selected winner and worsens asset-averaged symmetric calibration error. FEDIP therefore diagnoses unresolved action dispersion; final action choice requires an explicit economic loss function.

Keywords: distribution forecasting; model uncertainty; action resolution; evidence budget; computational economics; value at risk

1 Introduction

Many economic analytics systems fit a library of predictive distributions, select the model with the smallest validation loss, and enter a functional of that winner into a decision rule. The numerical action can look precise even when the evidence supporting it is not. A risk system, for example, may select a standardized 5% value-at-risk (VaR) threshold of 1.70 while other distributions not excluded by its own validation rule imply 1.60–1.90. The system has ranked models without resolving the action.

Expanding the candidate space sharpens this problem. Mixtures, skew distributions, nonparametric densities, and local-window models reduce misspecification risk, and in a nested space the best observed score cannot worsen. Each added candidate, however, creates another comparison for finite validation data. Better best-case fit can therefore coexist with greater disagreement about what to do.

Existing tools produce different output objects. Model confidence sets return statistically competitive models (Hansen et al. 2011; Barde 2026); Rashomon analyses return near-optimal predictive sets (Fisher et al. 2019; Marx et al. 2020; Semenova et al. 2022); distributional combinations return an aggregate predictive distribution (Hall and Mitchell 2007); and robust decision methods choose an action under an explicit loss function and uncertainty set. FEDIP instead returns the action correspondence induced by a declared retained set and studies how its width changes jointly with validation evidence and model-

¹ Corresponding author: minhuang1@link.cuhk.edu.cn

universe size. Its novelty is this integrated diagnostic and comparative static, not the max–min operator or the Bonferroni screen. Proper scores remain an inherited discipline for distributional comparison (Gneiting and Raftery 2007; Gneiting and Ranjan 2011; Diks et al. 2011).

We introduce the Finite-Evidence Decision Identification Procedure (FEDIP), an audit layer above an existing forecasting pipeline. A replaceable first module constructs an operationally retained model set; our implementation uses a transparent single-step Bonferroni–HAC screen. The distinctive second module projects every retained distribution through a prespecified economic functional and reports the action range. The methodological contribution is not the max–min operator alone, but the integration of competitive-model retention, action projection, and an evidence-by-universe comparative static in one auditable design.

Validation data are interpreted as a finite evidence budget. Candidate-space expansion increases the demand for discrimination; validation observations supply it. An oracle benchmark shows that the evidence requirement rises with model count and score noise and becomes large when economically consequential candidates approach the declared negligible-score margin. This benchmark rationalizes the comparative static tested by FEDIP; it is not a coverage theorem for the implemented empirical-winner screen.

Study 1 tests this comparative static in a Monte Carlo factorial conducted under a locally frozen protocol, with eight data-generating processes, 400 replications, three nested model spaces, and five validation lengths: 120 conditions and 48,000 valid replication-condition runs. Expanding from five to nine candidates increases the VaR diameter more with 20 than with 500 validation observations. Averaged within replication across four fixed stable processes, the interaction is 0.0526 estimation-sample standard deviations (95% Monte Carlo interval 0.0440–0.0612). The retained-set-cardinality interaction is 0.619 (0.569–0.670). The effect is concentrated in mixture and skew-normal environments and is near zero in Gaussian and Student-*t* environments, consistent with the theoretical prediction that the interaction is weak when consequential separation is absent.

In a financial illustration covering 11 assets and 2,710 sealed decision blocks, the screen retains 8.025–8.976 of nine models on average. A policy that mechanically chooses the maximum retained threshold lowers the violation rate from 5.06% to 3.99% but increases symmetric calibration error and raises the standardized VaR proxy by 9.47% relative to the selected-winner policy. The illustration shows that diagnosing unresolved action dispersion does not determine the optimal response to it.

The paper contributes a problem formulation, a system-level comparative static, and an operational diagnostic. It does not propose a new multiple-testing procedure, a selective-inference confidence set, or a universal law of model expansion.

Section 2 formalizes the problem and algorithm. Section 3 derives the evidence-budget theory and its boundary conditions. Section 4 describes the computational design and reports the factorial evidence. Section 5 interprets the real-asset weak-discrimination result and separates diagnosis from action choice. Section 6 concludes.

2 The Decision-Resolution Problem and FEDIP

2.1 Module I: construct an operationally retained set

The problem is conditional on a declared forecasting system, not a claim about structural or partial identification. We use “decision identification” only at this first definition to mean procedural action

resolution: whether the declared evidence-and-screening system contracts the candidate-implied action set. Let $\mathcal{F}_K = \{F_1, \dots, F_K\}$ be its fitted predictive distributions. Candidate k produces observation-level loss $S(F_k, Y_t)$. We use negative log score because it is strictly proper under standard regularity conditions (Gneiting and Raftery 2007); changing the score changes what the evidence can distinguish.

Given m validation observations, let

$$\bar{S}_{k,m} = m^{-1} \sum_{t=1}^m S(F_k, Y_t), \quad \hat{k}_m = \arg \min_{k \in \{1, \dots, K\}} \bar{S}_{k,m}.$$

For every nonwinner define the paired loss difference

$$d_{k,t} = S(F_k, Y_t) - S(F_{\hat{k}_m}, Y_t).$$

Let $\hat{\text{se}}_{\text{HAC}}(\bar{d}_k)$ estimate the standard error of the mean using a Newey–West long-run variance. For a negligible mean-loss margin δ , model k is excluded only if the multiplicity-adjusted one-sided lower confidence bound for $E[d_{k,t}]$ exceeds δ . The resulting set is

$$\hat{\mathcal{A}}_{K,m}(\delta) = \{k: \bar{d}_k - z_{1-\alpha/(K-1)} \hat{\text{se}}_{\text{HAC}}(\bar{d}_k) \leq \delta\},$$

where $z_p = \Phi^{-1}(p)$ and the empirical winner is included by construction. The screen is an operational rule, not a selective-inference pivot. We set $\delta = 0.02$ nats per observation and $\alpha = 0.05$; the margin corresponds to a geometric-mean realized density ratio of $\exp(-0.02) = 0.9802$. The value was fixed before the primary analysis as a predictive-score tolerance, not a monetary-loss threshold. In applications it should be calibrated on an independent period against the empirical distribution of out-of-sample score differences among established benchmark models, then held fixed for validation and reported over a declared sensitivity grid; here sensitivity analysis varies it from zero to 0.04.

Bonferroni–HAC is used for transparency and auditability, not claimed superiority over model confidence set (MCS) stepdown procedures (Hansen et al. 2011). Module I can instead accept an MCS or another auditable retained set. In every case, retention is procedural: it implies neither coverage for a true model nor economic admissibility.

2.2 Module II: project the set into action space

Let $a(F)$ be a scalar action or decision-relevant functional, such as 5% VaR or expected shortfall. Define the retained-set decision diameter

$$D_{K,m}(a) = \max_{k \in \hat{\mathcal{A}}_{K,m}} a(F_k) - \min_{k \in \hat{\mathcal{A}}_{K,m}} a(F_k).$$

For comparability across replications, Study 1 divides tail diameters by the estimation-sample standard deviation. Cardinality $|\hat{\mathcal{A}}_{K,m}|$ measures how many models remain. Diameter measures whether their disagreement matters for the selected decision. A large set can have a narrow diameter if its members imply similar actions; a two-model set can have a wide diameter if its members disagree in the tail.

The diameter is deliberately a support-width diagnostic: one retained endpoint can determine it. It asks whether any surviving model changes the action range, not whether most retained actions cluster near the winner. It is conditional on the declared system and is not a welfare loss.

2.3 The complete audit algorithm

The preceding objects define a computational diagnostic that can be implemented above an existing forecasting pipeline. Procedure 1 names the diagnostic to make its inputs, operations, and outputs auditable; it is not a new density estimator.

Procedure 1: Finite-Evidence Decision Identification Procedure (FEDIP)

Input: candidate model space; estimation and validation samples; proper score S ; negligible-loss margin δ ; significance level α ; economic decision functional a .

1. Fit every candidate distribution on the estimation sample.
2. Evaluate observation-level proper-score losses on the independent validation sample.
3. Select the empirical validation winner.
4. Apply paired HAC-adjusted, multiplicity-corrected loss comparisons and retain candidates not operationally excluded.
5. Map every retained distribution into the prespecified economic action $a(F)$.
6. Compute the decision diameter D as the maximum retained action minus the minimum retained action.
7. Report the winner, operationally retained set, and continuous decision diameter D .

Output: the winner, retained set, and unresolved action range. FEDIP neither constructs a confidence set nor chooses the final action.

In deployment, the validation function declares the score, screen, and negligible-score margin δ , while the decision owner declares the action functional and a tolerable action width ε . If $D_{K,m}(a) > \varepsilon$, the audit triggers additional validation evidence, model review, or explicit loss-based decision analysis. It does not trigger automatic selection of either retained endpoint.

3 Evidence-Budget Benchmark and Prediction

3.1 The fit option value of a larger model space

For integers $K < K'$, suppose $\mathcal{F}_K \subset \mathcal{F}_{K'}$; that is, the larger universe contains every candidate in the smaller universe.

Proposition 1 (Nested best-fit option value). *If $\mathcal{F}_K \subseteq \mathcal{F}_{K'}$, then for every realized validation sample,*

$$\min_{F \in \mathcal{F}_{K'}} \tilde{S}_m(F) \leq \min_{F \in \mathcal{F}_K} \tilde{S}_m(F).$$

The proof follows immediately because minimization over a superset cannot produce a larger minimum. This is a mechanical benchmark, not a new theorem. It explains why winner-only scoring favors expansion: added candidates can improve the observed best score and cannot worsen it.

No analogous monotonicity holds for retained-set diameter. A new candidate can exceed the evidentiary exclusion margin, can survive but imply the same action, or can survive and expand the action range. Moreover, multiplicity adjustment becomes more demanding as K rises. Thus the same expansion that weakly improves best fit can increase, decrease, or preserve decision disagreement.

3.2 Score separation, action separation, and evidence demand

A simple concentration argument supplies comparative statics. Let candidate k 's expected score gap from the best candidate be $\Delta_k > 0$, and suppose centered paired score differences are sub-Gaussian with variance proxy σ_k^2 . A union bound gives a sufficient condition of the form

$$m \gtrsim \max_{k: \Delta_k > \delta} \frac{2\sigma_k^2}{(\Delta_k - \delta)^2} \log\left(\frac{K-1}{\alpha}\right)$$

for separating every candidate whose gap exceeds δ . The expression is deliberately schematic. Serial dependence, estimated parameters, local models, and score tails complicate its literal use. It is not an optimal sample-size formula. Its value is to isolate three forces: discrimination becomes harder with score noise, harder with the number of comparisons, and sharply harder as consequential candidates approach the negligible-loss boundary.

The statistical and economic dimensions can be linked in an oracle benchmark. Let k^* denote the candidate with the smallest expected score and define

$$\Delta_k = \mathbb{E}[S(F_k, Y) - S(F_{k^*}, Y)], \quad g_k = |a(F_k) - a(F_{k^*})|.$$

Partition action-consequential candidates into those that are nonseparable at the declared score margin and those that are separable relative to that margin:

$$\mathcal{C}_{\varepsilon, \delta}^0 = \{k: g_k > \varepsilon, \Delta_k \leq \delta\},$$

$$\mathcal{C}_{\varepsilon, \delta}^+ = \{k: g_k > \varepsilon, \Delta_k > \delta\}.$$

For this benchmark only, define the oracle retained set

$$\mathcal{A}_{K,m}^{\text{or}}(\delta) = \{k: \bar{d}_{k,m} \leq \delta\}$$

relative to the fixed oracle comparator, and let $D_{K,m}^{\text{or}}(a)$ be the diameter of its action projection. This object is distinct from the implemented empirical-winner HAC retained set.

Proposition 2 (Oracle evidence-budget benchmark). *Assume a fixed oracle comparator, independent sub-Gaussian paired score differences with variance proxies σ_k^2 , and $\mathcal{C}_{\varepsilon, \delta}^0 = \emptyset$. Then $\Pr(D_{K,m}^{\text{or}} > 2\varepsilon) \leq \sum_{k=1}^K \mathbf{1}\{k \in \mathcal{C}_{\varepsilon, \delta}^+\} \exp\left[-\frac{m(\Delta_k - \delta)^2}{2\sigma_k^2}\right]$.*

The proof follows from the event inclusion

$$\{D_{K,m}^{\text{or}} > 2\varepsilon\} \subseteq \bigcup_{k: g_k > \varepsilon} \{k \in \mathcal{A}_{K,m}^{\text{or}}(\delta)\}.$$

For each $k \in \mathcal{C}_{\varepsilon, \delta}^+$,

$$\Pr(k \in \mathcal{A}_{K,m}^{\text{or}}(\delta)) = \Pr(\bar{d}_{k,m} \leq \delta) \leq \exp\left[-\frac{m(\Delta_k - \delta)^2}{2\sigma_k^2}\right],$$

so a union bound gives the result. If $\mathcal{C}_{\varepsilon, \delta}^0$ is nonempty, population score separation under margin δ does not require its exclusion even as $m \rightarrow \infty$. Otherwise the bound contracts exponentially with m , most slowly near the margin. This is an oracle benchmark, not a coverage theorem for the empirical-winner HAC procedure.

This logic yields a conditional prediction. Expanding the universe should create a larger finite-evidence identification cost when added candidates have consequential action gaps and small positive

excess score gaps above the retention margin, so short validation samples often retain them while longer evidence separates them. In that case the expansion-induced diameter should attenuate with m . If added candidates exceed the exclusion margin even at short m , expansion has little cost. If multiple candidates remain effectively equivalent even with long m , evidence may not attenuate the incremental diameter. The mechanism therefore predicts an interaction under consequential separation, not a universal main effect.

Prediction 1 (empirical evidence-by-universe interaction). For two nested universes $U5 \subset U9$, define

$$\Gamma_D = [D_{U9,20} - D_{U5,20}] - [D_{U9,500} - D_{U5,500}].$$

When short validation evidence does not exclude consequential added candidates and long evidence separates them, $\Gamma_D > 0$. When the candidates remain effectively equivalent or are immediately separated, Γ_D can be zero.

This is a falsifiable prediction about a declared system, not an algebraic proposition. We test it under the primary protocol for VaR diameter and retained-set cardinality.

The benchmark supplies comparative statics, not statistical validity for the implemented screen. Proper scores and competitive-model-set construction are inherited components. Model confidence sets target competitive models (Hansen et al. 2011); Rashomon analyses describe near-optimal collections (Fisher et al. 2019; Marx et al. 2020); and pooling constructs one combined forecast (Hall and Mitchell 2007). FEDIP instead measures the action range generated by an operationally retained set. Because that set is procedural, its projection has no asserted coverage guarantee.

4 Computational Design and Results

4.1 Factorial design

The primary protocol was frozen locally before execution but was not externally registered or independently timestamped. Study 1 uses primary seed 2026072602. Each of eight data-generating processes has 400 replications. Every replication contains 1,000 estimation observations, two disjoint 500-observation validation streams, and a sealed 250-observation test block. The retained-set analysis uses nested prefixes of 20, 60, 120, 250, and 500 from the first validation stream. The second stream measures winner reversal; the sealed test measures regret.

The eight processes are stable Gaussian, stable Student- t , stable skew-normal, stable Gaussian mixture, causally scale-adjusted GARCH, a single break, Markov switching, and smooth drift. This inventory separates stable shape misspecification from several forms of change. Results are never pooled selectively across processes. The co-primary inference averages the four stable processes within replication, as specified before execution, and every process-specific contrast is retained.

The nested universes are:

- $U3$: Gaussian, Student- t , and skew-normal;
- $U5$: $U3$ plus generalized normal and a two-component Gaussian mixture;
- $U9$: $U5$ plus kernel density, a three-component Gaussian mixture, recent-60 Student- t , and recent-120 Student- t .

Every model is fitted only with its permitted estimation information. Validation observations do not update model parameters. Negative log score evaluates the fitted densities. The Newey–West lag is

$$\ell_m = \lfloor 4(m/100)^{2/9} \rfloor.$$

The retained-set rule uses the 0.02-nat margin and a Bonferroni-adjusted 95% one-sided lower bound across the $K - 1$ comparisons with the empirical winner. At $m = 20$, the HAC-normal comparison is interpreted as a declared operational screen whose finite-sample behavior is evaluated by simulation, not as a calibrated small-sample test.

The co-primary outcomes are standardized 5% VaR diameter and retained-set cardinality. For outcome Y , the locked primary estimand is

$$\Gamma_Y = [Y_{U9,20} - Y_{U5,20}] - [Y_{U9,500} - Y_{U5,500}].$$

The central hypothesis passes only if both paired 95% intervals exclude zero in the positive direction. Replication is the unit of inference after averaging the four stable-process outcomes within each replication. Prespecified gates also require zero nested best-score violations within numerical tolerance, less than 5% fitting failure, and disclosure of every process-specific estimate. Expected-shortfall diameter, winner reversal across validation streams, and sealed-test regret are secondary outcomes.

The $U9 - U5$ estimand is intentionally a total-system contrast. Expansion changes candidate content, the familywise screening threshold, and potentially the empirical winner used for centering. The primary result therefore evaluates two complete declared computational forecasting systems; it does not identify the isolated causal contribution of the four added models. The intervals are Monte Carlo uncertainty intervals across 400 generated replications for the fixed four-process stable collection. They do not treat the eight processes as a random sample from a population of possible markets.

4.2 Evidence moderates model-space expansion

4.2.1 Co-primary interactions

Within the declared factorial, the primary analysis supports the prespecified evidence-by-universe interaction predicted by the evidence-budget interpretation. Expanding $U5$ to $U9$ increases standardized VaR diameter more at 20 than at 500 validation observations by 0.0526 (95% Monte Carlo interval 0.0440–0.0612). The corresponding interaction for retained-set cardinality is 0.619 (0.569–0.670). All 48,000 replication-condition runs are valid, no fit fails, and the larger universe never has a worse best observed score than the nested smaller universe. The Monte Carlo intervals measure simulation precision conditional on the chosen processes, model libraries, and algorithms; they do not quantify uncertainty over the universe of possible markets or research designs.



Fig. 1 Validation evidence moderates the decision-identification cost of model-universe expansion. Lines show stable-process means across 400 replications for nested $U3$, $U5$, and $U9$; bars are 95% Monte Carlo intervals. The retained set uses paired negative log scores, Newey–West uncertainty, a 0.02-nat margin, and Bonferroni-adjusted one-sided 95% bounds.

The levels behind the interaction show why the contrast matters. At $m = 20$, mean VaR diameter is 0.1886 in $U5$ and 0.3333 in $U9$, a 0.1448 expansion cost. At $m = 500$, the corresponding means are 0.0526 and 0.1447, a 0.0922 cost. Evidence reduces disagreement in both universes, but it reduces the incremental cost of the larger universe by 0.0526. Because the diameter is in estimation-sample standard-deviation units, its applied scale depends on volatility and exposure. At 2% daily volatility, for illustration, 0.0526 standardized units correspond to about 0.105 percentage points in a return threshold, while the $m = 20$ expansion cost of 0.1448 corresponds to about 0.290 percentage points; neither translation is a welfare or regulatory-capital estimate. Retained-set cardinality tells the same story. At $m = 20$, the means are 4.169 for $U5$ and 7.744 for $U9$; at $m = 500$, they are 3.183 and 6.138.

Table 1. Stable-process means by universe and validation evidence

Universe	m	Best score	Retained size	VaR diameter
$U3$	20	1.3527	2.629	0.1546
$U3$	500	1.3594	2.151	0.0694
$U5$	20	1.2648	4.169	0.1886
$U5$	500	1.2695	3.183	0.0526
$U9$	20	1.2546	7.744	0.3333
$U9$	500	1.2680	6.138	0.1447

Means average the four stable data-generating processes within 400 replications. VaR diameter is in estimation-sample standard-deviation units.

The best-score columns reveal the fit–identification tradeoff. At 20 observations, $U9$ ’s best validation score is 1.2546 compared with 1.2648 for $U5$, while its VaR diameter is 77% larger. At 500 observations the best scores are nearly equal, 1.2680 and 1.2695, but $U9$ ’s diameter remains larger because several variants are still acceptable. Expansion delivers option value in fit and a separate cost in decision identification. A single winner score records only the first.

4.2.2 Heterogeneity and consequential separation

The average interaction conceals a sharp boundary. Figure 2 reports every process-specific VaR estimate. Stable mixture data produce the largest interaction, 0.1519 (95% CI 0.1221–0.1817), followed by stable skew-normal data at 0.0576 (0.0445–0.0706). Stable Gaussian data yield 0.0047 (−0.0008, 0.0102), and stable Student- t data yield −0.0038 (−0.0125, 0.0050). The latter intervals include zero.



Fig. 2 Data-generating-process-specific evidence-by-universe interactions for standardized 5% VaR diameter. Points are paired interaction estimates across 400 replications; bars are two-sided 95% confidence intervals. No process was selected or omitted after execution.

The pattern fits the separation prediction. In skew and mixture settings, the expanded universe contains flexible models whose tail conclusions differ materially at short evidence lengths and become more distinguishable as evidence accumulates. In Gaussian and Student- t settings, several parametric, local, and flexible candidates remain effectively similar under the score-decision combination. More evidence narrows diameter within each universe, but it does not selectively remove the added variants enough to produce a positive $U_9 - U_5$ attenuation contrast.

A post-outcome add-one ablation clarifies candidate content. It is used for mechanism interpretation rather than confirmatory inference. The stable-process interaction is 0.0474 when recent-60 Student- t is added to U_5 , 0.0257 for recent-120 Student- t , 0.0046 for the three-component mixture, and −0.0016 for KDE. These contrasts are not additive because each six-model system has its own winner and multiplicity threshold, but they show that the full-system result is driven mainly by the local-window candidates.

The change processes reinforce rather than simplify this conclusion. Smooth drift has a positive interaction of 0.0745 (0.0584–0.0906), and Markov switching has 0.0214 (0.0125–0.0302). The causally scale-adjusted GARCH interaction is 0.0010 (−0.0070, 0.0089), and the single-break interaction is −0.0024 (−0.0086, 0.0037). Therefore, “nonstationary” is not itself an explanation. The relevant property is whether the validation information distinguishes added candidates whose tail actions differ.

These nulls refine the central claim. The study supports: *within the declared nested universes, longer independent validation evidence attenuates the added decision disagreement from expansion on average across the stable processes, with effects concentrated where evidence separates consequential*

distribution families. It does not support: *adding models always makes a distribution unidentifiable*. That stronger statement would contradict the process-specific results.

4.2.3 Screening-rule sensitivity

Post-outcome checks vary δ from 0 to 0.04 and alter the screening rule. The interaction remains positive in all 16 variants. At $\delta = 0.02$, it is 0.0395 with a common $U9$ critical value, 0.0150 without multiplicity adjustment, and 0.0554 with a comparator selected on the disjoint validation stream. The last result bridges the fixed-comparator theory and the implemented procedure; none of these checks establishes post-selection coverage. Full estimates are reported in Online Resource 1 and preserved in the replication archive.

A bounded post-outcome diagnostic directly describes the implemented screen. Among added $U9$ candidates whose score-gap proxy on the disjoint validation stream exceeds $\delta = 0.02$ and whose standardized VaR action gap exceeds 0.05, the empirical-winner HAC screen retains 73.4% at $m = 20$ and 32.2% at $m = 500$. The diagnostic confirms low short-sample discrimination and substantial attenuation with evidence, while also showing that persistence differs by data-generating process. It is descriptive, uses an independent-stream proxy rather than a population gap, and does not confer coverage. Definitions and process-specific results appear in Online Resource 1.

5 Financial Illustration: Measurement and Action

The illustration asks a narrower question than Study 1: does measuring unresolved disagreement justify automatically choosing the maximum retained tail threshold? The same nine-model system is applied without retuning to SPY, QQQ, IWM, EFA, TLT, IEF, LQD, HYG, GLD, EURUSD, and BTC. The preserved adjusted-close snapshots span 1993–2026 for the longest equity series and 2014–2026 for BTC; asset-specific source and evaluation dates are reported in Online Resource 1. Each rolling decision uses 1,000 estimation returns, 120 validation returns, and a non-overlapping sealed block of 20 returns, yielding 2,710 blocks. Returns are standardized by a causal EWMA volatility estimate. The comparison includes Gaussian, selected-winner, score-weighted ensemble, and maximum-retained-threshold policies. Reported intervals use 2,000 two-stage bootstrap draws (seed 2026072603), resampling assets and non-overlapping sealed blocks within sampled assets. This targets an equal-weight asset-and-block resampling population; it does not preserve residual dependence induced by overlapping estimation and validation histories or contemporaneous cross-asset dependence. A post-outcome circular moving-block bootstrap preserves local within-asset dependence more directly and yields the same directional conclusions; its specification and results are included in Online Resource 1.

The principal empirical fact is weak discrimination. The screen retains 8.025–8.976 of nine models on average across assets. This does not establish population equivalence; it means that the declared margin, multiplicity adjustment, and 120-observation validation histories rarely exclude a fitted distribution. The maximum-threshold rule is therefore close to taking the largest tail estimate in the entire library.

Table 2 reports the policy comparison. The maximum-threshold rule reduces the equal-weight violation rate to 3.99%, versus 5.06% for the selected model, but its asset-level absolute distance from the 5% target, averaged equally across assets, is 0.00891 higher (95% bootstrap interval 0.00424–0.01002). This quantity differs from the absolute error of the pooled 3.99% rate because absolute errors are computed before averaging across assets. Relative to the selected-winner policy, its standardized VaR proxy rises by 9.47% (7.91%–11.29%). Under the prespecified symmetric calibration criterion, the conservative shortcut therefore fails.

Table 2. Equal-weight multi-asset policy performance

Policy	Violation rate	Coverage error	Standardized VaR proxy	Standardized ES proxy
Gaussian reference	0.0507	0.00437	1.7069	2.1582
Selected winner	0.0506	0.00119	1.7144	2.3474
Score-weighted ensemble	0.0508	0.00146	1.7076	2.3450
Maximum retained threshold	0.0399	0.01010	1.8767	2.5476

Note. Means give equal weight to the 11 assets. The study contains 2,710 non-overlapping asset-blocks and zero fitting failures. Coverage error is computed within asset as the absolute distance from the 5% violation target and then averaged equally across assets.

The result rejects a policy shortcut, not the diagnostic. An asymmetric loss could rationalize lower violation rates, but that loss was not prespecified and is not retrofitted here. FEDIP measures what the evidence leaves unresolved; choosing among worst-case, pooling, or other set-aware actions requires an explicit objective. The maximum endpoint is therefore not presented as a general robust-optimization solution. The financial exercise is an illustration of weak discrimination and the measurement–action distinction, not evidence that FEDIP improves financial performance.

6 Conclusion

This study shows that expanding a computational distribution-model space has two distinct finite-sample effects. It weakly improves the best attainable validation loss, yet it can widen the range of economic actions supported by models that the available evidence does not exclude. In the factorial, the incremental VaR-diameter effect of moving from five to nine models is materially larger with 20 than with 500 validation observations. The effect concentrates where added distribution families imply different tails and their separation is sensitive to validation length; it is near zero when expansion adds little consequential separation.

The multi-asset experiment establishes a second boundary. A wide decision diameter measures unresolved specification consequences; it does not prescribe the largest threshold. The maximum retained-threshold policy reduces observed violations but worsens the prespecified symmetric calibration criterion and raises the standardized VaR proxy. Set-valued forecast evidence and economic action choice must therefore be connected through an explicit loss function rather than through an automatic worst-case rule.

The evidence-budget interpretation is conditional on the declared model space, score, validation design, tolerance, and decision functional. FEDIP reports unresolved action dispersion at a supplied validation length; it does not estimate an optimal evidence budget. Within that scope, it offers a practical design principle for computational economics: increases in model-space complexity should be evaluated jointly with the validation information available to distinguish economically consequential alternatives. Future work can formalize optimal evidence allocation and compare set-aware actions under asymmetric economic losses.

A Proofs and Implementation Details

A.1 Proof of Proposition 1

Let f_K^* be an empirical score minimizer over the smaller universe. Because the smaller universe is contained in the larger one, the same fitted distribution is feasible in the larger-universe minimization. Hence

$$\min_{f \in \mathcal{F}_{K'}} \bar{S}_m(f) \leq \bar{S}_m(f_K^*) = \min_{f \in \mathcal{F}_K} \bar{S}_m(f).$$

The result concerns observed best loss only. It does not imply that selection from the larger universe has lower expected loss on a sealed sample.

A.2 Finite-evidence separation sketch

For a candidate with expected paired score gap $\Delta_k > \delta$, suppose

$$\Pr(\bar{d}_k - \Delta_k \leq -x) \leq \exp\left(-\frac{mx^2}{2\sigma_k^2}\right).$$

Setting $x = \Delta_k - \delta$ bounds the probability that the candidate appears within the negligible-loss margin. Applying a union bound over $K - 1$ nonbest candidates gives

$$\Pr(\text{at least one nonnegligible candidate survives}) \leq \sum_{k: \Delta_k > \delta} \exp\left[-\frac{m(\Delta_k - \delta)^2}{2\sigma_k^2}\right].$$

Requiring each summand to be at most $\alpha/(K - 1)$ yields the sufficient condition

$$m \geq \max_{k: \Delta_k > \delta} \frac{2\sigma_k^2}{(\Delta_k - \delta)^2} \log\left(\frac{K - 1}{\alpha}\right).$$

The logarithm is essential: this is a union-bound sufficient condition, not a linear penalty in $(K - 1)/\alpha$. The expression motivates the dependence on evidence, score noise, separation, and universe size. The empirical procedure uses HAC standard errors rather than assuming independent sub-Gaussian observations. Accordingly, the sketch supplies comparative statics rather than finite-sample coverage for the same-sample empirical-winner screen.

A.3 Implementation details

All Study 1 paths contain 1,000 estimation observations, two successive 500-observation validation streams, and a 250-observation sealed test. Stable processes are standard Gaussian; variance-standardized Student- t_5 ; standardized skew-normal with shape 7; and a two-component mixture with left weight 0.15, left component $N(-3.2, 0.3^2)$, and a mean-balancing right component. The change processes are GARCH(1,1) with coefficients 0.03, 0.10, and 0.87 followed by causal EWMA adjustment; a break from Gaussian to variance-standardized t_4 at the end of estimation; a two-state Gaussian/skew-normal chain with 0.96 stay probability; and a mixture whose left-tail probability rises linearly from 0.01 to 0.20.

Gaussian parameters use sample mean and standard deviation. Student- t , skew-normal, and generalized-normal parameters are moment matched with bounded shape grids. Gaussian mixtures use full covariance, 10^{-5} covariance regularization, one initialization, and at most 150 EM iterations. The KDE uses 81 empirical-quantile anchors and a Silverman-style bandwidth with a 0.05 floor. Recent models fit Student- t laws to the last 60 or 120 estimation observations. Densities are floored at 10^{-10} .

VaR and ES are numerically integrated on a 4,001-point grid spanning at least eight estimation-sample standard deviations.

The Newey–West estimator uses Bartlett weights, divisor m , no finite-sample correction, and the lag rule stated in Section 4. Study 2 removes invalid or duplicate dates, computes log returns, and applies causal EWMA scaling with decay 0.94. The fixed 11-asset panel and adjusted-close snapshots are not a point-in-time vendor database; source hashes in the replication archive define the analyzed vintage.

References

- Barde, Sylvain. 2026. “Large-Scale Model Comparison with Fast Model Confidence Sets.” *Journal of Econometrics* 253 (January): 106123. <https://doi.org/10.1016/j.jeconom.2025.106123>.
- Diks, Cees, Valentyn Panchenko, and Dick van Dijk. 2011. “Likelihood-Based Scoring Rules for Comparing Density Forecasts in Tails.” *Journal of Econometrics* 163 (2): 215–30. <https://doi.org/10.1016/j.jeconom.2011.04.001>.
- Fisher, Aaron, Cynthia Rudin, and Francesca Dominici. 2019. “All Models Are Wrong, but Many Are Useful: Learning a Variable’s Importance by Studying an Entire Class of Prediction Models Simultaneously.” *Journal of Machine Learning Research* 20 (177): 1–81. <https://jmlr.org/papers/v20/18-760.html>.
- Gneiting, Tilmann, and Adrian E Raftery. 2007. “Strictly Proper Scoring Rules, Prediction, and Estimation.” *Journal of the American Statistical Association* 102 (477): 359–78. <https://doi.org/10.1198/016214506000001437>.
- Gneiting, Tilmann, and Roopesh Ranjan. 2011. “Comparing Density Forecasts Using Threshold- and Quantile-Weighted Scoring Rules.” *Journal of Business & Economic Statistics* 29 (3): 411–22. <https://doi.org/10.1198/jbes.2010.08110>.
- Hall, Stephen G., and James Mitchell. 2007. “Combining Density Forecasts.” *International Journal of Forecasting* 23 (1): 1–13. <https://doi.org/10.1016/j.ijforecast.2006.08.001>.
- Hansen, Peter R., Asger Lunde, and James M. Nason. 2011. “The Model Confidence Set.” *Econometrica* 79 (2): 453–97. <https://doi.org/10.3982/ECTA5771>.
- Marx, Charles T., Flavio du Pin Calmon, and Berk Ustun. 2020. “Predictive Multiplicity in Classification.” *Proceedings of the 37th International Conference on Machine Learning*, Proceedings of machine learning research, vol. 119: 6765–74. <https://proceedings.mlr.press/v119/marx20a.html>.
- Semenova, Lesia, Cynthia Rudin, and Ronald Parr. 2022. “On the Existence of Simpler Machine Learning Models.” *2022 ACM Conference on Fairness Accountability and Transparency*, FAccT 22, 1827–58. <https://doi.org/10.1145/3531146.3533232>.

Statements and Declarations

Funding

The authors received no financial support for the research, authorship, or publication of this article.

Competing interests

The authors have no relevant financial or non-financial interests to disclose.

Author contributions

Min Huang: conceptualization, methodology, software, formal analysis, data curation, visualization, and writing—original draft. Mingyan Liu: methodology, validation, theoretical formalization, and writing—review and editing. Xiaoe Li: investigation, data processing, validation, and writing—review and editing. All authors read and approved the manuscript.

Data and code availability

Online Resource 1, “Supplementary Diagnostics and Data Inventory,” provides the bounded screening diagnostic, Study 2 data inventory, sensitivity details, and dependence check. The replication archive contains the frozen analysis code, preserved numerical outputs, source inventories, cell- and block-level results, asset summaries, bootstrap and leave-one-out results, temporal-integrity audit materials, file manifests, and SHA-256 hashes. The 11 preserved market-data snapshots are supplied for confidential editorial audit and remain subject to applicable redistribution restrictions. Where direct redistribution is not permitted, the authors will provide source inventories, retrieval information, processing code, and file hashes.

Use of generative AI and AI-assisted technologies

OpenAI ChatGPT and Codex were used to support language editing, manuscript organization, consistency checks, and formatting. The authors critically reviewed and edited all outputs, verified the analysis and citations, and take full responsibility for the manuscript.

Acknowledgements

Not applicable.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [ESM1OnlineResource20260729110452.pdf](#)
- [ReplicationArchivePublic20260729110452.zip](#)