

Decision-Equivalent Volatility Forecasts: Statistical Loss, Portfolio Value, and Allocation Commonality

Fuli Yang

GARCH Institute

Harvard Square, Cambridge, MA 02138, USA

ORCID: 0009-0004-1255-2483

Abstract

Predictive accuracy and decision quality can rank volatility forecasts differently because portfolio maps are invariant to some forecast errors and sensitive to others. This paper formalizes decision equivalence: statistically distinct volatility forecasts are decision-equivalent when they induce the same allocation under a specified rule. For global minimum variance, inverse-volatility, and equal-risk-contribution portfolios, common positive rescaling of marginal volatility forecasts is an exact equivalence transformation. A frozen replay on IVV, EFA, EEM, AGG, and GSG from August 2011 through July 2026 perturbs six prespecified error directions while holding correlations fixed. The common-scale placebo changes mean QLIKE by as much as 1.0024 while maximum weight deviation is $2.22e-16$. The strongest cross-sectional distortion worsens QLIKE by 0.406 and raises annualized certainty-equivalent return by 4.19 percentage points, although no cell survives the frozen 120-test Holm family at 5%. Similar loss deterioration can produce zero, modest, or large portfolio-value changes, and the strongest privately useful distortion reduces cross-rule cosine similarity. A supplementary daily-return variance check on 178 complete months preserves the leading loss-deterioration signs. The results separate statistical loss, portfolio value, and allocation commonality as distinct evaluation dimensions.

Keywords: volatility forecasting; decision equivalence; predict-then-optimize; portfolio allocation; QLIKE; certainty equivalent; allocation commonality

JEL: C53; C58; G11; G17

1. Introduction

Financial forecasts are commonly evaluated in two spaces. The first is statistical: a volatility forecast is compared with a realized proxy using a loss such as QLIKE. The second is decisional: the forecast enters a portfolio map, and the resulting allocation is judged by realized risk, return, utility, turnover, or costs. These evaluations need not agree because an optimizer does not respond to every component of forecast error. Some errors cancel under normalization; others rotate relative risk and therefore the portfolio.

Recent evidence on large-language-model volatility estimates makes this distinction especially visible. Lim and Kim (2026) report that LLM-based volatility estimates can generate attractive portfolios and concentrated covariance structures. Yatawara (2026) shows that such portfolio gains need not reflect superior volatility forecasting: in that setting, the apparent edge is associated with covariance geometry and a standing exposure tilt, is regime-dependent, and weakens under realistic costs. Those findings establish that downstream portfolio performance is not a certificate of forecast skill. They do not, however, identify the forecast-error directions to which common portfolio maps are exactly invariant, nor do they ask whether privately useful error directions make distinct allocation rules more or less alike.

This paper studies those two questions. The first contribution is formal. Two forecasts are decision-equivalent for a specified portfolio map if they generate identical relative weights when all other decision inputs are held fixed. Common scaling of marginal volatility forecasts is an exact decision-equivalence direction for global minimum variance (GMV), inverse volatility (IVOL), and equal risk contribution (ERC). The same statistical forecast can therefore be made substantially better or worse under QLIKE without changing the downstream allocation at all.

The second contribution is empirical. A frozen historical replay perturbs prespecified cross-sectional directions in a five-ETF volatility vector while holding the correlation estimate fixed. The design separates common scale, dispersion, equity versus bonds, U.S. versus international equity, emerging versus developed equity, and commodities versus financial assets. The treatment grid and inference family were fixed before the exact-data replay. This is deliberately not an exercise in searching for a profitable bias vector.

The results sharpen the distinction between prediction and decision. The common-scale placebo is exact to machine precision: across η from -0.50 to +0.50, maximum absolute weight deviation is $2.22e-16$, even though mean QLIKE changes by as much as 1.0024. Cross-sectional distortions can create economically large forecast-loss/utility wedges. The largest, equity versus bonds at $\eta=-0.50$ under GMV, worsens mean QLIKE by 0.4063 and log-squared error by 1.1457 yet increases annualized certainty-equivalent return by 4.19 percentage points. The 95% moving-block interval for the

utility gain is 1.68 to 6.75 percentage points with 12-month blocks and 1.62 to 6.55 points with 24-month blocks. The gain remains positive after 5 and 15 basis-point one-way turnover costs and in both pre-2020 and 2020-onward subsamples.

The confirmatory result is deliberately more restrained. The frozen H2 family contains 120 geometry-direction by nonzero-eta by portfolio-rule tests. With 1,999 bootstrap repetitions, the strongest intersection-union p-values attain the minimum grid value of 0.0005, but Holm adjustment produces a minimum familywise p-value of 0.060. The paper therefore does not claim 5% familywise confirmation of H2. This inference boundary is retained rather than repaired by redefining the family, replacing the loss function, or increasing bootstrap repetitions after inspecting the outcome.

The third contribution concerns allocation commonality. A convenient story would be that privately useful forecast errors also make different portfolio rules converge. The data reject that as a general proposition. The largest private-benefit cell reduces cross-rule weight cosine by 0.0676 and raises effective rank by 0.1075, indicating less rather than more level commonality. Other directions, especially negative U.S.-versus-international distortions, modestly increase similarity. Private portfolio value and allocation commonality are therefore distinct downstream outcomes.

The paper contributes to predict-then-optimize research by identifying equivalence and sensitivity classes in a volatility-to-portfolio map rather than training predictions directly for a downstream objective. It also contributes to volatility evaluation by showing that statistical loss, realized private utility, and cross-rule allocation diversity should be reported separately. The analysis stops at allocation geometry: greater commonality is not interpreted as evidence of realized market impact, liquidity deterioration, or endogenous volatility.

2. Related literature and positioning

2.1 Forecast loss and economic value

Volatility forecasting has long distinguished statistical accuracy from economic usefulness. The conditional-variance tradition established the role of dynamic volatility models (Engle, 1982; Bollerslev, 1986), while realized-volatility research provides higher-frequency measures against which latent variance forecasts can be assessed (Andersen et al., 2003; Corsi, 2009). Fleming, Kirby, and Ostdiek (2001, 2003) evaluate volatility information through portfolio utility and realized covariance applications, while Clements and Silvennoinen (2013) show that economic value need not track forecasting sophistication one-for-one. Ledoit and Wolf (2004) further illustrate why covariance estimation and portfolio construction cannot be separated in finite samples. Patton (2011) develops forecast-comparison losses that remain meaningful when volatility is observed through an imperfect proxy. The present paper does not choose between statistical and economic criteria. It treats them as different objects and asks which forecast perturbations can move one without moving the other.

2.2 Predict-then-optimize

Elmachtoub and Grigas (2022) formalize the idea that prediction quality should be evaluated with respect to the downstream optimization problem. Decision-focused applications in finance make the same point operationally: Lee et al. (2025), for example, show that mean-variance decision-focused training can create systematic prediction biases that improve the downstream portfolio objective. The broader portfolio-choice literature also emphasizes that the mapping from estimated risk inputs to weights is itself an economically consequential object (Markowitz, 1952; Ledoit and Wolf, 2004). The present design differs in three respects. The predicted object is marginal volatility rather than expected return, the errors are externally controlled rather than learned, and the analysis explicitly separates decision equivalence from decision-sensitive rotation and from cross-rule commonality.

2.3 LLM volatility and covariance geometry

Lim and Kim (2026) and Yatawara (2026) provide the closest financial motivation. Their results show why a covariance or portfolio outcome can look attractive even when marginal volatility estimates do not dominate conventional forecasting benchmarks. This paper takes that mechanism as prior evidence rather than a result to rediscover. Its incremental object is a forecast-error map: which directions are exact null directions for the allocation rule, which directions change private utility, and whether those same directions align or separate otherwise distinct portfolio rules.

3. Decision equivalence and error geometry

3.1 Setup and definitions

Let $s(t)$ be an n -vector of positive marginal-volatility forecasts and $R(t)$ a correlation forecast. Let $D(t)$ be the diagonal matrix containing $s(t)$, so the covariance forecast is $\Sigma(t) = D(t) R(t) D(t)$. A controlled treatment replaces each $s(i,t)$ by $s(i,t) \exp[\eta b(i)]$, where b is a prespecified direction and η is treatment magnitude. Except for the common-scale placebo, each b is centered to cross-sectional mean zero and scaled to unit cross-sectional standard deviation.

Definition 1 (Decision equivalence). For a portfolio map A , two positive volatility-forecast vectors are decision-equivalent if they produce the same portfolio weights when the correlation input and feasible set are held fixed. Decision equivalence is rule-specific and does not imply equality of statistical loss.

Definition 2 (Decision-sensitive direction). A direction b is decision-sensitive at a forecast vector if a sufficiently small nonzero treatment along b changes the resulting portfolio weights, holding all other decision inputs fixed.

Definition 3 (Allocation commonality). Given several portfolio rules, allocation commonality is the degree to which their weight vectors, or their rebalancing vectors, align in weight space. The primary level metrics are mean pairwise cosine similarity and the participation-ratio effective rank of the rule-by-asset weight matrix.

3.2 Proposition 1: common-scale decision equivalence

Proposition 1. Multiplying every marginal-volatility forecast by the same positive constant leaves relative GMV, normalized inverse-volatility, and ERC weights unchanged, provided the feasible set is itself scale independent.

The invariance follows because a common positive scale factor cancels under weight normalization for GMV and inverse-volatility portfolios, while it scales all ERC risk contributions proportionally. Appendix A provides the formal proof.

3.3 Proposition 2: pairwise inverse-volatility rotation

Proposition 2. Under inverse-volatility allocation, multiplying asset i volatility by $\exp[\eta b(i)]$ changes each pairwise log weight ratio according to

$$\log[w(i,\eta)/w(j,\eta)] = \log[w(i,0)/w(j,0)] - \eta [b(i) - b(j)].$$

The treatment parameter therefore has a direct relative-risk interpretation: only cross-sectional differences in b rotate inverse-volatility weights. For GMV and ERC the map is nonlinear in correlations, while the common-scale null direction remains exact. Appendix A gives the derivation.

4. Data and frozen replay

4.1 Data

The primary universe contains IVV, EFA, EEM, AGG, and GSG, representing U.S. large-cap equity, developed international equity, emerging-market equity, U.S. aggregate bonds, and broad commodities. The full-precision monthly-return panel is extracted from the independent public GitHub repository HerdingAI/stock-price-returns (<https://github.com/HerdingAI/stock-price-returns>; accessed 5 September 2026). The repository is publicly visible and constructs monthly returns from Yahoo Finance data obtained through yfinance using auto-adjusted closes. The source file is output/monthly_returns_by_ticker.csv; the file's Git blob SHA is ee6594efc62cff087fb23cad748c5f114029cbe8, and the latest commit affecting that file at extraction was 6a44fca9fb8ec3e92cf136dc982a819db6f77362. The frozen five-asset extract contains 240 complete month-end observations from August 2006 through July 2026, with no missing values; its SHA-256 is 3126ec75ce29d01933ad4b093024a74aadde4763637471858e57334093f5e9f4.

For the supplementary measurement check, daily simple returns come from output/daily_returns_by_ticker.parquet in the same repository. The downloaded snapshot ends on 17 July 2026; its SHA-256 is 2d79cb14384fbc04e08cc93d7c507bc5cb03e002e37c0936f9841ddeb60ca36. We require a finite daily return for every expected XNYS trading session for all five ETFs, using exchange_calendars 4.13.2 and checking the 2026 holiday schedule against NYSE. July 2026 is partial (12 of 22 sessions), and May 2026 lacks 26 May (19 of 20 sessions); both are excluded from this check. The resulting 178 out-of-sample observations span August 2011 to June 2026, excluding May 2026. Monthly realized variance is the sum of squared daily returns. For retained months, daily-compounded and source monthly returns differ by at most 0.00000211 in absolute return units. The monthly forecast history and treatment maps are retained; evaluation of both loss proxies uses the same 178 months.

4.2 Timing and portfolio maps

A uniform 60-month warm-up is imposed. The first portfolio is formed at the end of July 2011 and evaluated on August 2011 returns; the out-of-sample evaluation therefore contains 180 months through July 2026. At formation month t , marginal volatility is the trailing 36-month standard deviation and correlations use the trailing 60 months. Correlations are held fixed across treatments within each month. Weights are long-only and sum to one. The three primary maps are GMV, IVOL, and ERC.

4.3 Frozen treatments and outcomes

The eta grid is $\{-0.50, -0.35, -0.20, -0.10, 0, 0.10, 0.20, 0.35, 0.50\}$. Directions are common scale, contemporaneous dispersion, equity versus bonds, U.S. versus international equity, emerging versus developed equity, and commodities versus financial assets. No direction was added after pilot or exact-result inspection. Forecast loss is canonical QLIKE, $y/f - \log(y/f) - 1$, averaged across assets. Portfolio value is annualized certainty equivalent with primary risk aversion $\gamma=3$, alongside annualized return, volatility, Sharpe ratio, turnover, and 5/15bp one-way trading-cost deductions. Allocation commonality is measured by cross-rule weight cosine, effective rank, and rebalancing-vector similarity.

4.4 Inference

Treatment effects are paired monthly differences from $\eta=0$. The primary inference uses a 12-month moving-block bootstrap with 1,999 repetitions; 24-month blocks are a prespecified robustness check. H2 requires both worse QLIKE and higher certainty-equivalent return under an intersection-union test. Holm adjustment is applied across the complete 5 geometry directions by 8 nonzero eta values by 3 rules, for 120 confirmatory tests. Portfolio rules are not treated as independent replications. State dependence is pre-specified as a secondary analysis. The primary stress indicator equals one for out-of-sample months in the top decile of absolute IVV monthly returns; 85th- and 95th-percentile thresholds are robustness checks. These state labels are constructed only for evaluation and never enter forecasts, treatment directions, or portfolio formation. All stochastic procedures use seed 20260905: the primary 12- and 24-month moving-block bootstraps and the supplementary daily-RV bootstrap. Portfolio optimization and coverage screening are deterministic conditional on the frozen inputs.

5. Results

5.1 Exact decision equivalence

Table 1 is the empirical counterpart of Proposition 1. Across the entire common-scale grid, maximum absolute relative-weight deviation is $2.22\text{e-}16$, effectively machine precision. Statistical loss is not invariant. At $\eta=-0.50$, mean QLIKE rises by 1.0024; at $\eta=+0.10$ it falls by 0.0112; and at $\eta=+0.50$ it rises by 0.2634. Thus the same downstream decision can correspond to volatility forecasts that are materially better or worse under the statistical loss.

Table 1. Common-scale placebo: allocation invariance and forecast-loss changes

eta	Max abs. weight deviation	Delta mean QLIKE
-0.50	2.22e-16	+1.0024
-0.35	2.22e-16	+0.4814
-0.20	2.22e-16	+0.1732
-0.10	2.22e-16	+0.0580
0.00	0.00e+00	+0.0000
0.10	2.22e-16	-0.0112
0.20	2.22e-16	+0.0158
0.35	2.22e-16	+0.1133
0.50	2.22e-16	+0.2634

Notes: Common-scale treatment relative to $\eta=0$. Weight deviation is the maximum across dates and the three portfolio maps.

5.2 Private-benefit wedges

Among the 120 confirmatory geometry cells, 57 have the economic sign pattern of worse QLIKE and higher certainty-equivalent return. Table 2 reports the leading cells. The largest is equity versus bonds at $\eta=-0.50$ under GMV: mean QLIKE worsens by 0.4063 while annualized certainty equivalent rises by 4.19 percentage points. The 95% moving-block interval for the CE gain is 1.68 to 6.75 percentage points. U.S. versus international equity at $\eta=-0.50$ under GMV raises CE by 1.19 points with a 0.8047 QLIKE deterioration. EM versus developed equity at $\eta=+0.50$ under IVOL raises CE by 0.73 points with a 0.4683 QLIKE deterioration.

Table 2. Selected private-benefit wedges under the frozen treatment grid

Direction	eta	Rule	Delta QLIKE	Delta CE (pp)	95% CI CE (pp)	Raw IUT p	Holm p
equity vs bond	-0.50	GMV	+0.4063	+4.19	[+1.68, +6.75]	0.0005	0.060
us vs international	-0.50	GMV	+0.8047	+1.19	[+0.68, +1.74]	0.0005	0.060
us vs international	-0.50	IVOL	+0.8047	+0.99	[+0.43, +1.42]	0.0010	0.107
em vs developed	+0.50	IVOL	+0.4683	+0.73	[+0.38, +1.16]	0.0005	0.060
em vs developed	+0.50	ERC	+0.4683	+0.55	[+0.27, +0.90]	0.0005	0.060
equity vs bond	-0.20	GMV	+0.0515	+0.62	[+0.14, +1.05]	0.0415	1.000

Notes: CE is annualized with gamma=3. The frozen confirmatory family contains 120 tests. No adjusted p-value is below 0.05.

The multiplicity result places a clear boundary on interpretation. Several leading cells attain the minimum attainable one-sided bootstrap p-value of 0.0005, but the 120-test Holm family yields a minimum adjusted value of 0.060. The exact-data replay therefore provides strong cell-level evidence of economically meaningful wedges without establishing a 5% familywise rejection of the confirmatory hypothesis.

5.3 Exposure attribution

The largest wedge is readily interpretable as an exposure rotation rather than hidden volatility information. Under equity-versus-bond eta=-0.50, the GMV portfolio increases average IVV weight by 32.60 percentage points, EFA by 6.96 points, and EEM by 8.32 points, while reducing AGG by 43.92 points and GSG by 3.96 points. Annualized realized return rises by 5.01 points, while mean turnover rises by 0.0315 under the frozen drift-adjusted definition. The statistical forecast becomes worse; the portfolio benefits because the distortion changes what the investor owns.

Table 3. Exposure attribution for selected distortions

Direction	eta	Rule	IVV pp	EFA pp	EEM pp	AGG pp	GSG pp	Delta turnover
equity vs bond	-0.50	GMV	+32.6	+7.0	+8.3	-43.9	-4.0	+0.0315
us vs international	-0.50	GMV	+12.2	-0.0	+0.0	-10.2	-2.0	+0.0081
em vs developed	+0.50	IVOL	+6.3	+5.8	-7.0	-4.2	-0.9	-0.0003

Notes: Average treatment-minus-baseline weight differences across the 180 evaluation months.

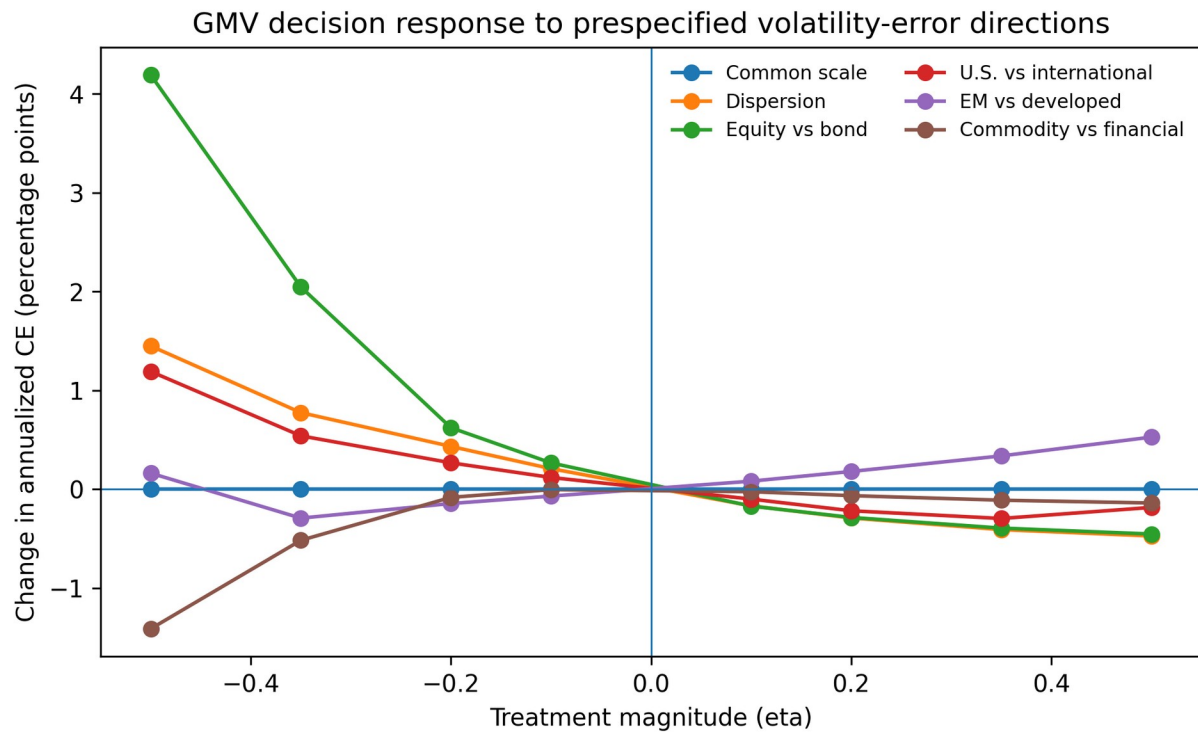


Figure 1. GMV certainty-equivalent response to the prespecified error directions

Notes: Annualized treatment-minus-baseline CE in percentage points. The common-scale line is identically zero. The signed response is descriptive of the frozen treatment grid; it is not a newly selected direction.

5.4 Private value does not imply allocation convergence

Allocation commonality produces the paper's most important negative mechanism result. The strongest private-benefit treatment does not make the three portfolio rules converge. Equity versus bonds at $\eta = -0.50$ lowers mean cross-rule weight cosine by 0.0676, with a 95% interval from -0.0808 to -0.0576, and raises effective rank by 0.1075. Negative dispersion treatments behave similarly. By contrast, U.S.-versus-international $\eta = -0.50$ raises cosine by only 0.0071 and modestly lowers effective rank. The direction of a private portfolio gain therefore does not identify the direction of cross-rule commonality.

Table 4. Allocation commonality under selected distortions

Direction	eta	Delta cosine	95% CI cosine	Delta eff. rank	95% CI rank	Delta trade cosine
dispersion	-0.50	-0.0786	[-0.0945, -0.0658]	+0.1213	[+0.1016, +0.1491]	+0.0748
equity vs bond	-0.50	-0.0676	[-0.0808, -0.0576]	+0.1075	[+0.0935, +0.1278]	+0.0459
us vs international	-0.50	+0.0071	[+0.0017, +0.0112]	-0.0090	[-0.0149, -0.0006]	+0.0262
em vs developed	+0.50	-0.0207	[-0.0271, -0.0140]	+0.0313	[+0.0192, +0.0431]	-0.0219
commodity vs financial	+0.50	+0.0024	[+0.0006, +0.0034]	-0.0043	[-0.0049, -0.0029]	-0.1046
dispersion	+0.50	+0.0419	[+0.0283, +0.0540]	-0.0583	[-0.0780, -0.0368]	-0.0597
equity vs bond	+0.50	+0.0411	[+0.0283, +0.0532]	-0.0571	[-0.0762, -0.0367]	+0.0124

Notes: Commonality metrics are common to the three rules at a given treatment/date. Higher cosine and lower effective rank indicate greater level similarity.

5.5 Costs, subsamples, and state dependence

The leading equity-versus-bond GMV wedge survives the prespecified economic robustness checks. Gross annualized CE rises by 4.19 percentage points; after 5bp and 15bp one-way turnover costs, the gains are 4.15 and 4.08 points. The effect is 3.34 points before 2020 and 5.27 points from 2020 onward. In the top decile of absolute IVV monthly returns, the CE gain is 7.60 points, compared with 3.79 points outside those months. The wedge also remains positive for gamma values of 2, 5, and 10. These checks support an exposure interpretation while showing that the magnitude is state-dependent.

Table 5. Costs, subperiods, and state dependence for the strongest private-benefit treatment

Subset	Cost bp	gamma	Delta ann. return pp	Delta CE pp	Delta Sharpe	Delta turnover
full	0	3	+5.01	+4.19	+0.274	+0.0315
full	5	3	+4.97	+4.15	+0.270	+0.0315
full	15	3	+4.90	+4.08	+0.264	+0.0315
pre_2020	0	3	+3.66	+3.34	+0.120	+0.0117
2020_plus	0	3	+6.74	+5.27	+0.411	+0.0566
full	0	2	+5.01	+4.46	+0.274	+0.0315
full	0	5	+5.01	+3.64	+0.274	+0.0315
full	0	10	+5.01	+2.28	+0.274	+0.0315
stress90	0	3	+11.48	+7.60	+0.512	+0.0353
nonstress90	0	3	+4.29	+3.79	+0.295	+0.0311

Notes: Equity-versus-bond $\eta = -0.50$, GMV. Stress follows the pre-specified definition in Section 4.4.

5.6 Matched-loss and alternative-loss diagnostics

A forecast-loss change does not identify a portfolio-value change even when loss magnitudes are closely matched. For GMV, three prespecified treatments generate QLIKE deterioration in a narrow range: common scale at $\eta = -0.35$ raises QLIKE by 0.4814 but leaves CE exactly unchanged; dispersion at $\eta = -0.50$ raises QLIKE by 0.4196 and CE by 1.45 percentage points; and equity versus bonds at $\eta = -0.50$ raises QLIKE by 0.4063 and CE by 4.19 points. Because these treatments differ mainly in error direction rather than in the order of forecast-loss deterioration, the comparison isolates the economic importance of geometry more sharply than a ranking by QLIKE alone. The matched-loss comparison is an exploratory diagnostic and is not part of the frozen H2 family. The common-scale row is especially instructive: QLIKE rises while log-squared error falls. QLIKE is asymmetric in the variance ratio y/f , so downward scaling that makes forecasts too small can receive a larger penalty than the corresponding upward error. Log-squared error is symmetric in log space and can fall when the rescaling moves forecasts closer to the geometric center of the realized targets. The two losses therefore encode different forecast-comparison geometries; the result is not an arithmetic inconsistency.

Table 6. Exploratory matched-loss diagnostic under GMV

Direction	η	Delta QLIKE	Delta logSE	Delta CE (pp)	Delta cosine
common scale	-0.35	0.4814	-1.3839	0.00	0.0000
dispersion	-0.50	0.4196	1.0239	1.45	-0.0786
equity vs bond	-0.50	0.4063	1.1457	4.19	-0.0676

Notes: The three rows were chosen after the exact replay solely because their QLIKE deteriorations are similar; this is an exploratory geometry diagnostic, not confirmatory inference. CE is annualized and expressed in percentage points.

The strongest equity-versus-bond GMV cell also worsens the prespecified log-squared-error loss by 1.1457, so its forecast deterioration is not specific to QLIKE. Its 24-month-block CE interval is 1.62 to 6.55 percentage points, close to the 12-month-block interval of 1.68 to 6.75. These checks address loss-function and dependence-length sensitivity while retaining the squared-monthly-return proxy. Section 5.7 separately evaluates sensitivity to the realized proxy using daily returns.

5.7 Daily realized-variance robustness

The corrected daily-return variance check retains 178 complete out-of-sample months. On this common evaluation sample, the leading geometry treatments retain positive loss deterioration: equity versus bonds at $\eta = -0.50$: daily-RV delta QLIKE +0.6542, 95% interval [+0.3791, +0.9401], versus +0.4033 with squared monthly returns on the same months; U.S. versus international equity at $\eta = -0.50$: daily-RV delta QLIKE +1.0426, 95% interval [+0.6292, +1.6265], versus +0.8078 with squared monthly returns on the same months; emerging versus developed equity at $\eta = +0.50$: daily-RV delta QLIKE +0.6478, 95% interval [+0.3813, +1.0456], versus +0.4746 with squared monthly returns on the same

months. The common-scale placebo at $\eta = -0.35$ raises daily-RV QLIKE by 0.7319, with unchanged relative weights by Proposition 1. These selected treatments were identified in the primary results; their intervals are supplementary, not multiplicity-adjusted confirmatory evidence. The 12-month moving-block bootstrap uses 1,999 shared resamples (seed 20260905; the primary replay uses the same seed) on the calendar-month grid, preserving May 2026 as missing rather than joining April and June as adjacent observations. Missing values are omitted within each resampled mean. This checks the loss side only: the primary 180-month CE estimates are not estimates for this reduced sample, and no joint loss/utility test is claimed for the RV sample.

6. Discussion

The exact placebo and the geometry treatments together imply that forecast accuracy has no one-to-one decision interpretation. A loss difference can be economically irrelevant if it lies in a decision-equivalent direction, while a statistically poor forecast can be privately valuable if its cross-sectional error rotates the portfolio toward rewarded exposures. The latter result does not rehabilitate the forecast. It attributes the portfolio outcome to the decision map.

The multiplicity result is substantively useful rather than an inconvenience. Economically large cell-level wedges coexist with a failure to clear the frozen familywise 5% threshold. A positive-results-only workflow would be tempted to report the strongest direction as if it were selected in advance or to reduce the testing family. The frozen protocol prevents that. The resulting claim is correspondingly narrower: controlled wedges exist in the realized sample and several are individually strong, but the broad confirmatory family is not rejected at 5%.

The matched-loss diagnostic clarifies why this is not merely a story about larger forecast mistakes creating larger portfolio changes. Common scaling, dispersion, and equity-versus-bond treatments can produce QLIKE deterioration of roughly four to five tenths while generating, respectively, zero, modest, and large GMV utility changes. The optimizer is therefore responding to where the forecast error points in risk space, not simply to how large the scoring-rule penalty becomes.

The commonality evidence also changes the interpretation of systematic forecast error. A common error direction need not make heterogeneous portfolio rules converge. In the largest private-benefit treatment, GMV, IVOL, and ERC become less alike in level weights. This occurs because each map translates the same relative-risk distortion through a different nonlinear allocation rule. Common forecast error and common portfolio geometry are therefore separate mechanisms.

The analysis has limitations. Squared monthly returns are noisy variance proxies; the supplementary daily-return variance check supports the leading loss-deterioration signs on 178 complete months, but daily sampling is still coarser than intraday realized variance. Two months fail the daily-coverage requirement, and small differences between monthly-source returns and daily compounding remain. Portfolio utility has not been re-estimated on the reduced RV sample, so this check does not independently confirm a joint private-benefit wedge. The five-ETF universe limits external validity to single stocks, rates portfolios, options, and high-frequency systems. Holding correlations fixed isolates marginal-volatility geometry but omits joint errors in correlations and marginal forecasts. Finally, allocation commonality is not evidence of equilibrium price impact, market volatility, or liquidity effects.

7. Conclusion

Volatility forecasts can be statistically different yet decision-equivalent, and statistically worse yet privately more valuable, because portfolio maps are selective about forecast error. In the exact common-scale placebo, QLIKE changes materially while GMV, IVOL, and ERC weights remain identical to machine precision. Cross-sectional error directions break that equivalence and can generate large realized utility wedges, but the frozen 120-test family does not achieve 5% familywise significance. An exploratory matched-loss diagnostic shows that comparable forecast-loss deterioration can map into zero, modest, or large utility changes depending on error direction. More importantly, the largest privately useful distortion makes different portfolio rules less, not more, similar. Statistical loss, private portfolio value, and allocation commonality are therefore separate rankings. Forecast evaluation in automated portfolio systems should report all three rather than treating downstream performance as validation of predictive skill.

Declarations

Declaration of Interest. The author declares no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Ethics Approval. This study uses historical financial market data and involves no human participants, animals, personal data, or interventions. Ethics approval and informed consent were therefore not required.

Use of Generative AI. Generative AI tools were used only as computational aids for organizing analysis materials and for limited language editing. The author retains responsibility for the research design, theoretical development, code, data processing, statistical analysis, interpretation, citation verification, and final manuscript. No AI system served as an author or as an independent source of empirical evidence.

Data and Code Availability. The frozen protocol, exact-data extraction audit, replay code, treatment outputs, and table-generating files are retained in a versioned replication archive. The source monthly-return panel is publicly available at <https://github.com/HerdingAI/stock-price-returns>, with the exact source file, Git blob SHA, commit SHA, extraction date, and derived-panel SHA-256 reported in Section 4.1. The replication archive contains the five-ETF extract used in the analysis together with the replay code, subject to the source repository's redistribution terms. The corrected daily-RV supplement contains the coverage audit, extracted daily returns, matched-month loss series, bootstrap summaries, and executable correction script. NYSE calendar source: <https://www.nyse.com/markets/hours-calendars>.

Appendix A. Proof details

Proof of Proposition 1. For GMV, replacing Sigma by c squared Sigma multiplies the inverse by c to the power minus two, and this factor cancels when the inverse-covariance weight vector is normalized to sum to one. For IVOL, each unnormalized inverse-volatility weight is multiplied by $1/c$, which again cancels on normalization. For ERC, both total variance and each component risk contribution scale by c squared, leaving all relative risk-contribution equalities unchanged. Scale-independent long-only constraints preserve the same solution set.

Proof of Proposition 2. Under IVOL, the unnormalized weight on asset i is proportional to $1/[s(i) \exp(\eta b(i))]$. Taking the ratio of assets i and j cancels the common normalization term and gives $[s(j)/s(i)] \exp\{-\eta[b(i)-b(j)]\}$. Taking logarithms yields the stated result.

Appendix B. Inference-resolution note

The moving-block bootstrap uses 1,999 repetitions as frozen before the exact-data replay. One-sided empirical p-values therefore lie on a grid with minimum positive value $1/(1,999+1)=0.0005$. Under Holm correction across 120 H2 tests, a smallest raw p-value of 0.0005 maps to an adjusted value of 0.060 at the first step. The paper reports that finite-resolution consequence directly and does not increase the repetition count after observing the boundary result. The same seed, 20260905, is used for the supplementary daily-RV bootstrap; no additional stochastic procedure is used.

References

- Markowitz, H. (1952). Portfolio selection. *Journal of Finance*, 7(1), 77–91. <https://doi.org/10.2307/2975974>.
- Ledoit, O., & Wolf, M. (2004). Honey, I shrunk the sample covariance matrix. *Journal of Portfolio Management*, 30(4), 110–119. <https://doi.org/10.3905/jpm.2004.110>.
- Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50(4), 987–1007. <https://doi.org/10.2307/1912773>.
- Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3), 307–327. [https://doi.org/10.1016/0304-4076\(86\)90063-1](https://doi.org/10.1016/0304-4076(86)90063-1).
- Andersen, T. G., Bollerslev, T., Diebold, F. X., & Labys, P. (2003). Modeling and forecasting realized volatility. *Econometrica*, 71(2), 579–625. <https://doi.org/10.1111/1468-0262.00418>.
- Clements, A., & Silvennoinen, A. (2013). Volatility timing: How best to forecast portfolio exposures. *Journal of Empirical Finance*, 24, 108–115. <https://doi.org/10.1016/j.jempfin.2013.09.004>.
- Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2), 174–196. <https://doi.org/10.1093/jjfinec/nbp001>.
- Elmachtoub, A. N., & Grigas, P. (2022). Smart “Predict, then Optimize”. *Management Science*, 68(1), 9–26. <https://doi.org/10.1287/mnsc.2020.3922>.

- Fleming, J., Kirby, C., & Ostdiek, B. (2001). The economic value of volatility timing. *Journal of Finance*, 56(1), 329-352. <https://doi.org/10.1111/0022-1082.00327>.
- Fleming, J., Kirby, C., & Ostdiek, B. (2003). The economic value of volatility timing using “realized” volatility. *Journal of Financial Economics*, 67(3), 473-509. [https://doi.org/10.1016/S0304-405X\(02\)00259-3](https://doi.org/10.1016/S0304-405X(02)00259-3).
- Lee, J., Jeon, H., Bae, H., & Lee, Y. (2025). Return prediction for mean-variance portfolio selection: How decision-focused learning shapes forecasting models. *Proceedings of the 6th ACM International Conference on AI in Finance*, 114-122. <https://doi.org/10.1145/3768292.3770423>.
- Lim, T., & Kim, J. H. (2026). Measuring volatility with LLMs: Implications for covariance structure and portfolio construction. *Finance Research Letters*, 106, 110293. <https://doi.org/10.1016/j.frl.2026.110293>.
- Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, 160(1), 246-256. <https://doi.org/10.1016/j.jeconom.2010.03.034>.
- Yatawara, A. (2026). Good portfolios from bad forecasts: The anatomy of LLM volatility estimates. *Finance Research Letters*, 109 (forthcoming October 2026), 110572. <https://doi.org/10.1016/j.frl.2026.110572>. Earlier working-paper version: SSRN 6999349.