

Variance reduction and interval efficiency in forecast reconciliation

Jufeng Li

q3529416243@gmail.com

China University of Petroleum, East China

Wenjie Li

China University of Petroleum, East China

Junhao Zhan

China University of Petroleum, East China

Research Article

Keywords:

Posted Date: October 29th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11530753/v1>

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

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Variance reduction and interval efficiency in forecast reconciliation

Jufeng Li^{1*}, Wenjie Li² and Junhao Zhan³

¹College of Science, China University of Petroleum (East China), Qingdao, 266580, China.

²College of Mechanical and Electronic Engineering, China University of Petroleum (East China), Qingdao, 266580, China.

³Qingdao Institute of Software, College of Computer Science and Technology, China University of Petroleum (East China), Qingdao, 266580, China.

*Corresponding author(s). E-mail(s): q3529416243@gmail.com;

Abstract

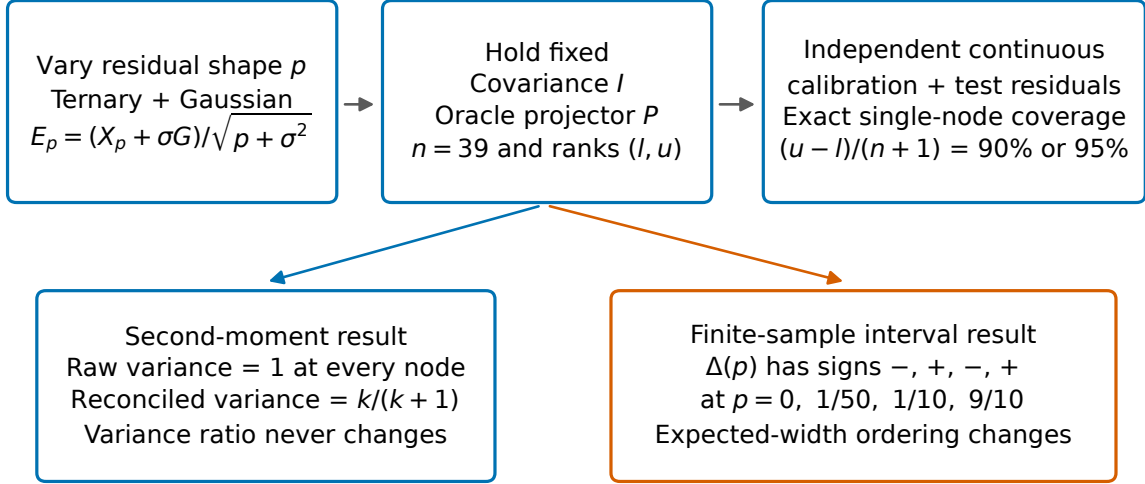
Forecast reconciliation is commonly motivated by variance reduction, whereas finite-sample interval widths depend on order statistics. We construct a nonelliptical residual family for two- and three-leaf hierarchies in which the full covariance, oracle minimum-trace projector, calibration size and actual marginal coverage remain fixed, yet the expected ordinary-width ordering reverses at least three times. Exact rational sign witnesses at calibration size 39 and a uniform Gaussian smoothing bound establish the result for continuous laws with exact 90% or 95% single-node coverage. The construction identifies a boundary of variance-based efficiency arguments outside elliptical residual families. We also compare existing covariance estimators on two related Australian retail partitions and a tourism hierarchy. In 72-month sequential evaluations with annual expanding-window refits, past-only covariance fitting, rolling calibration and fixed normalization scales, centered diagonal-target shrinkage reduces mean normalized interval scores by 2.0%, 3.3% and 0.8% relative to raw forecasts at nominal 90%. Realized coverage remains below nominal, and centered shrinkage is not uniformly best. Historical comparisons additionally show why cross-design scores require common normalization. The study introduces no estimator and makes no temporal-dependence coverage guarantee.

Keywords: hierarchical forecasting; conformal prediction; order statistics; covariance shrinkage; finite-sample efficiency; uncertainty quantification

1 Introduction

Hierarchical and grouped forecasting impose linear aggregation constraints on a vector of predictions. Forecast reconciliation maps independently constructed base forecasts onto the coherent subspace. Minimum-trace reconciliation chooses a projection using the error covariance, and its practical performance depends on how that covariance is estimated (Hyndman et al. 2011; Wickramasuriya et al. 2019; Athanasopoulos et al. 2024).

Prediction intervals introduce an additional objective. For signed intervals defined by two calibration order statistics, expected width depends on the residual distribution beyond its covariance. Under a common elliptical residual shape with finite covariance, the marginal span distribution scales with standard deviation. Outside that family, a variance comparison alone need not order expected interval widths. We isolate this distinction by varying residual shape while holding fixed the entire covariance matrix, the oracle projector, the calibration sample size, the ranks and the actual unconditional single-node coverage. The resulting expected-width difference changes sign at least three times.



Existence construction: $k = 2, 3$; fixed positive σ within certified bounds. No new estimator.

Fig. 1 Fixed-covariance counterexample framework. Residual shape varies while covariance, oracle projection, calibration size, ranks and actual marginal coverage remain fixed. The variance ratio stays constant while the ordinary expected-width ordering changes sign. The diagram describes proof objects and does not introduce a forecasting algorithm.

The theoretical contribution is an exactly certified construction and its same-covariance consequence (Figure 1). Atomic calculations supply rational sign witnesses; Gaussian smoothing gives continuous marginals and exact rank coverage without losing those signs. The construction uses small star hierarchies and a specific residual family. Its purpose is to identify a limit of a second-moment justification, rather than to characterize the prevalence of the phenomenon in forecasting data.

The empirical component compares existing estimators on public retail and tourism hierarchies. We distinguish isotropic-target Ledoit-Wolf shrinkage from diagonal-target shrinkage, and report both centered and uncentered versions of the latter. These recipes differ in shrinkage target and intensity, and in some comparisons also in centering; their observed differences are not a target-only intervention. We retain historical frozen-model stress tests, correct their cross-design normalization, and add 72-month sequential evaluations with annual model updates. Score, width and realized coverage are reported together because an empirical score comparison under temporal dependence has a different interpretation from the equal-coverage theoretical result.

2 Relation to existing work

2.1 Reconciliation and trace minimization

Wickramasuriya et al. (2019) derive the minimum-trace projection under an error covariance matrix. A practical introduction is provided by Hyndman and Athanasopoulos (2021). Panagiotelis et al. (2021) give a geometric treatment of projection and bias correction, and Panagiotelis et al. (2023) develop probabilistic reconciliation and score optimization. Athanasopoulos et al. (2024) review the broader literature. The author implementation in `hts` uses a diagonal shrinkage target that retains unequal node-specific second moments (Hyndman et al. n.d.). In contrast, an isotropic target replaces part of each diagonal element by a common scale. We keep these estimator labels separate throughout.

2.2 Conformal intervals in hierarchies

Principato et al. (2026) establish component-wise efficiency results under elliptical residual assumptions. Their formulation includes expected squared interval lengths; under a common elliptical shape, ordinary marginal lengths also inherit scale comparisons. They already report instability of estimated minimum-trace configurations and identify temporal dependence as an extension. Our result concerns ordinary expected width outside that distributional class, with covariance and actual marginal coverage held fixed. It does not contradict their efficiency theorem.

2.3 Finite-sample size and calibration

Dhillon et al. (2024) study expected conformal set size at finite calibration sample size. The order-statistic identity and rank-coverage argument used below are classical tools. Gibbs and Candès (2024) develop conformal inference under distribution shift, and Hong et al. (2026) combine reconciliation with per-node adaptive calibration in load forecasting. Here frozen and past-only rolling calibration are evaluation factors, not new calibration algorithms.

2.4 Scope of the contribution

The specific statement combines fixed covariance, a fixed oracle projector, fixed finite calibration size, ordinary expected width, exact single-node coverage and repeated sign reversal. Our comparison with the cited literature did not identify the same conjunction. This bounded comparison does not establish priority over all published work. The new content is the certified construction, rather than the order-statistic formula, scale normalization or existing estimators used in the case studies.

3 Setup

3.1 Hierarchy, projection and residuals

Let $S \in \mathbb{R}^{m \times k}$ have full column rank. Coherent responses satisfy $Y = Sb$ for a bottom-level vector b . A reconciliation matrix P is a projection with $PS = S$ and image $\text{Im}(S)$. For a fixed predictor f , let $E = Y - f(X)$; coherence of Y gives reconciled residual PE . For a positive-definite covariance W , the minimum-trace projection is

$$P_W = S(S^\top W^{-1}S)^{-1}S^\top W^{-1}.$$

Comparisons include unreconciled forecasts ($P = I$, which is not a projection onto the coherent subspace), OLS ($W = I$), bottom-up, diagonal WLS, isotropic-target Ledoit-Wolf and diagonal-target shrinkage. For the last estimator let $M = x^\top x/n$ for residual matrix x , after column centering when specified, and let $D = \text{diag}(\text{diag}(M))$. The covariance recipe is $W = (1 - \lambda)M + \lambda D$, with λ estimated from the off-diagonal sample correlations. The centered and uncentered recipes are reported separately; Online Resource 1 gives their formulas and numerical conventions.

3.2 Signed intervals and coverage

Let $E_1, \dots, E_n$ be independent copies of the residual vector and let $E_{(r),j}$ denote the r th calibration order statistic at node j . The signed interval for an independent new observation is

$$I_j(X_{\text{new}}) = [f_j(X_{\text{new}}) + E_{(l),j}, f_j(X_{\text{new}}) + E_{(u),j}].$$

For continuous marginal residuals its unconditional single-node coverage is exactly $(u-l)/(n+1)$. At $n = 39$, ranks (2, 38) yield 90% and ranks (1, 39) yield 95%. These statements concern the independent continuous model used in the theorem. They do not automatically extend to atomic laws with ties or to serially dependent forecasting residuals.

3.3 Efficiency functional and empirical score

Define the ordinary expected total-width difference

$$\Delta(P, F) = \sum_{j=1}^m \mathbb{E}[L_j^P] - \sum_{j=1}^m \mathbb{E}[L_j], \quad L_j = E_{(u),j} - E_{(l),j}.$$

A negative value favors reconciliation. This functional is ordinary width, rather than squared width or joint-region volume. The empirical analysis uses the interval score

$$\begin{aligned} \text{IS}_\alpha(l, u; y) &= (u - l) + \frac{2}{\alpha}(l - y)\mathbf{1}\{y < l\} \\ &\quad + \frac{2}{\alpha}(y - u)\mathbf{1}\{y > u\}. \end{aligned}$$

We divide by each node's training-period mean absolute seasonal difference and average equally over nodes and predictors. Within a design, all methods use the same node scales. Comparisons across

historical training designs are additionally recomputed with a common reference scale (Section 5.5). The sequential benchmark fixes its scales before the test period. Coverage is reported alongside scores; lower score at unequal realized coverage is not an equal-coverage width theorem.

4 A fixed-covariance construction

4.1 Model and width functionals

Fix $k \in \{2, 3\}$, $m = k + 1$ and $S = [\mathbf{1}_k^\top; I_k]$. The OLS projector $P = S(S^\top S)^{-1}S^\top$ has diagonal entries $k/(k+1)$. Let X_p have independent coordinates with probabilities $p/2, 1-p, p/2$ on $-1, 0, 1$, and let G be an independent standard Gaussian m -vector. For $\sigma > 0$, define

$$Z_p = X_p + \sigma G, \quad E_p = \frac{Z_p}{\sqrt{p + \sigma^2}}.$$

The unnormalized model has covariance $(p + \sigma^2)I_m$. Its oracle projector is already P , because multiplying a covariance by a positive scalar leaves the minimum-trace projector unchanged. Normalization fixes the covariance itself: $\mathbb{E}[E_p] = 0$ and $\text{Cov}(E_p) = I_m$ for every $p \in [0, 1]$. The atomic specialization $\sigma = 0$ is used to compute exact sign witnesses; its normalized form is used only at $p > 0$.

Write the total expected-width differences in the two scales as

$$\begin{aligned} \delta_{k,\sigma}(p) &= \sum_j \mathbb{E}[L_j^P(Z)] - \sum_j \mathbb{E}[L_j(Z)], \\ \Delta_{k,\sigma}(p) &= \sum_j \mathbb{E}[L_j^P(E)] - \sum_j \mathbb{E}[L_j(E)]. \end{aligned}$$

Positive homogeneity of order statistics gives

$$\Delta_{k,\sigma}(p) = \frac{\delta_{k,\sigma}(p)}{\sqrt{p + \sigma^2}}. \quad (1)$$

Proposition 1 (ordinary-width multi-crossing). For each $k \in \{2, 3\}$ and calibration size $n = 39$:

- (i) The atomic difference $\delta_{k,0}(p)$ has signs positive, negative, positive at $p = 1/50, 1/10, 9/10$, for each of the rank pairs $(2, 38)$ and $(1, 39)$.
- (ii) For $0 < \sigma \leq 1/200$ at ranks $(2, 38)$, or $0 < \sigma \leq 1/500$ at ranks $(1, 39)$, $\delta_{k,\sigma}$ is continuous on $[0, 1]$, retains those three signs and satisfies $\delta_{k,\sigma}(0) < 0$. Hence it has a zero in each of $(0, 1/50)$, $(1/50, 1/10)$ and $(1/10, 9/10)$.
- (iii) The constant-covariance difference $\Delta_{k,\sigma}$ is continuous on $[0, 1]$ and has the same negative, positive, negative, positive sign sequence at $0, 1/50, 1/10, 9/10$. In particular,

$$\Delta_{k,\sigma}(0) = (k+1)C \left(\sqrt{\frac{k}{k+1}} - 1 \right) < 0,$$

where C is the expected span of the corresponding standard-normal order statistics.

For every admissible **positive** σ , both raw and projected intervals have exact unconditional single-node coverage, 90% or 95% according to the rank pair. No such exact-coverage claim is attached to the atomic witnesses in part (i), which have ties. The proposition concerns ordinary expected width.

Proposition 2 (same covariance, same variance gain). Along the continuous family of Proposition 1, the covariance is I_m , the oracle projector is P , each projected coordinate has variance $k/(k+1)$ against raw variance 1, and the sample size and actual coverage are fixed. Nevertheless, the sign of Δ changes at least three times. Thus identical second-moment information does not determine the expected-width ordering on this family.

Proposition 2 follows from the normalization identity (1) and isolates the interpretation of the construction.

4.2 Exact finite-support evaluation

For a scalar law with ordered finite support $v_1 < \dots < v_s$, cumulative probabilities F_i and gaps $d_i = v_{i+1} - v_i$,

$$\mathbb{E}[X_{(u)} - X_{(l)}] = \sum_{i=1}^{s-1} d_i \sum_{r=l}^{u-1} \binom{n}{r} F_i^r (1 - F_i)^{n-r}.$$

The interval crosses the i th gap precisely when the number of calibration values at or below v_i is between l and $u - 1$. At rational p and $\sigma = 0$, all support values and probabilities are rational. The certificate evaluates the identity in exact arithmetic and checks it against a separate expected-rank calculation, with exhaustive enumeration for small samples.

Every row of the star projector has the same multiset of absolute coefficients, $\{k/(k+1), 1/(k+1), \dots, 1/(k+1)\}$. Independence, identical distribution and symmetry of the coordinates therefore make all projected marginal spans identically distributed. Consequently,

$$\delta_{k,0}(p) = (k+1)\{W_{\text{proj}}(p) - W_{\text{raw}}(p)\},$$

where W_{proj} and W_{raw} are scalar expected spans. Table 1 gives the twelve witnesses. The atomic values certify signs; equal actual coverage belongs to the positive-noise family obtained in Section 4.3.

Table 1. Atomic sign witnesses, rounded to six decimals. The rank-based target labels identify the intervals whose continuous versions have exact coverage. Here δ is unnormalized and $\Delta = \delta/\sqrt{p}$; $14m\sigma$ is the smoothing bound at the stated maximum positive noise level.

Rank-based target	k	$p = 1/50$: δ, Δ	$p = 1/10$: δ, Δ	$p = 9/10$: δ, Δ	Bound $14m\sigma$
90%	2	+0.400540, +2.832245	-0.427006, -1.350313	+1.707260, +1.799610	0.21
90%	3	+0.660523, +4.670600	-0.336604, -1.064434	+2.465241, +2.598592	0.28
95%	2	+0.086035, +0.608356	-1.153665, -3.648210	+1.943699, +2.048839	0.084
95%	3	+0.293268, +2.073719	-1.010199, -3.194530	+3.482433, +3.670807	0.112

4.3 Smoothing, boundary value and continuity

Order statistics are 1-Lipschitz in the sample maximum norm. Coupling an atomic calibration sample to its Gaussian perturbation gives $|L_\sigma - L_0| \leq 2\sigma M$, where M is the largest absolute perturbation among the n scalar observations. Raw and projected perturbation variances are at most one. Gaussian tail integration yields $\mathbb{E}[M^2] \leq 2\log(2n) + 2 < 12$ at $n = 39$, so $\mathbb{E}[M] < 7/2$. Summing the raw and projected expectations over nodes gives the uniform bound

$$|\delta_{k,\sigma}(p) - \delta_{k,0}(p)| \leq 14m\sigma. \quad (2)$$

Each bound in Table 1 is strictly smaller than all corresponding absolute witnesses, including the tightest case $0.084 < 0.0860345706$ at 95%, $k = 2$. Thus the three positive-parameter signs persist. In the normalized scale the comparison uses a common denominator:

$$\left| \Delta_{k,\sigma}(p) - \frac{\delta_{k,0}(p)}{\sqrt{p + \sigma^2}} \right| \leq \frac{14m\sigma}{\sqrt{p + \sigma^2}}.$$

At $p = 0$, raw coordinates are Gaussian with variance σ^2 and projected coordinates have variance $\sigma^2 k/(k+1)$. Hence

$$\delta_{k,\sigma}(0) = (k+1)\sigma C \left(\sqrt{\frac{k}{k+1}} - 1 \right) < 0.$$

Here $C = 3.4826$ at ranks (2, 38) and $C = 4.3012$ at ranks (1, 39). Dividing by σ gives Proposition 1(iii). For fixed positive σ , the residual laws are finite Gaussian mixtures with continuous mixture probabilities in p and uniformly dominated first moments. Their expected spans, and therefore δ and Δ , are continuous on $[0, 1]$. The intermediate value theorem gives the three distinct crossing intervals. Exact rational signs and the smoothing bound establish this claim independently of the numerical resolution of a plotted curve.

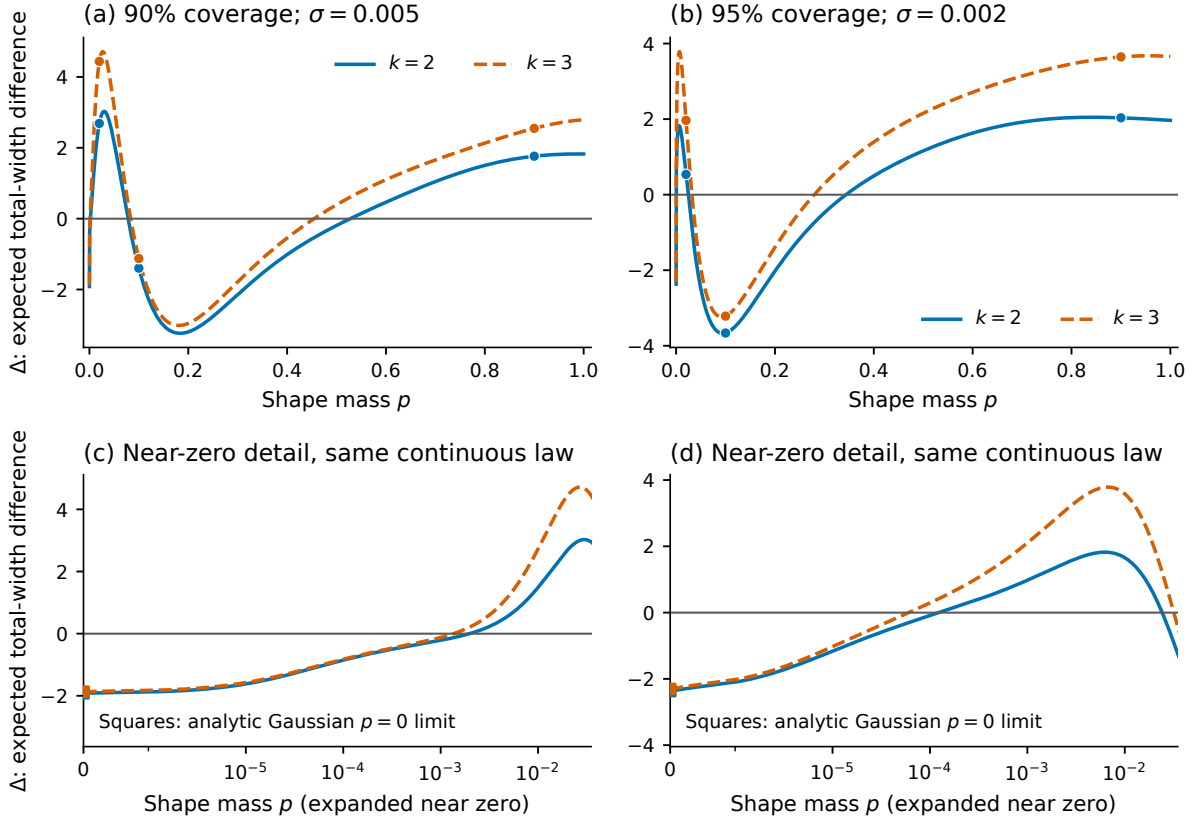


Fig. 2 Expected total ordinary-width difference in the continuous Gaussian-smoothed, constant-covariance family. Curves show normalized $\Delta_{k,\sigma}(p)$ computed by deterministic Gaussian-mixture quadrature, for $k = 2, 3$ and $n = 39$. The upper panels show the full parameter range; lower panels resolve the near-zero region and its negative Gaussian boundary. Noise is $\sigma = 1/200$ at 90% coverage and $\sigma = 1/500$ at 95%. Witness points have certified signs. Numerical zero locations are illustrative and do not assert that exactly three zeros exist.

4.4 Realisability

The residual construction is compatible with coherent responses. For an existence argument let the feature vector be E , take $Y = 0 \in \text{Im}(S)$ and fix $f(E) = -E$. Then $Y - f(E) = E$ and $Y - Pf(E) = PE$, with independent calibration and test copies. This realization establishes admissibility, rather than a realistic forecasting mechanism. Scope limits are collected in Section 8.

5 Controlled reanalysis of retail trade

5.1 Source and partitions

The retail data are the `aus_retail` series in `tsibbledata` 0.4.1 (O’Hara-Wild et al. 2022), originating from the Australian Bureau of Statistics (ABS). The source archive is hash-locked, and its records were checked against an independent R read (Section 5.7).

The source contains 152 series: eight states or territories crossed with twenty industry categories, including overlapping parent and child categories. We use the published ABS group-subgroup definitions to construct disjoint partitions (Australian Bureau of Statistics 2024). The primary partition retains fifteen mutually exclusive subcategories; requiring a complete panel from April 1988 to December 2018 leaves 110 state-by-subcategory cells. Adding eight state aggregates, fifteen subcategory aggregates and one total gives 134 nodes and 369 months. A broad-category sensitivity partition has 44 retained cells and 59 nodes. The excluded-cell inventory is provided in Online Resource 2. Aggregates sum retained cells and need not reproduce the complete official national total.

5.2 Roles, models and configurations

The historical evaluation has two chronological folds, each separating model fitting, a 24-vector covariance window, a reserved 79-observation pilot segment, a 79-observation calibration pool and a 24-month test window. The final-fold base model is trained through October 2001 and tested in 2017-2018. This

long-gap schedule is retained as a drift stress test. The later-training sensitivity in Section 5.5 reduces unused buffers, while Section 5.10 evaluates annually updated models. These analyses were developed after inspection of earlier project results and are not preregistered confirmations.

Three predictors are used: seasonal naive (lag 12), ridge regression and random forest with a fixed seed rule. Hyperparameters are not selected using test performance. Seasonal-naive forecasts are coherent by construction and every valid projection leaves them unchanged; reconciliation comparisons are driven by ridge and random forest. The implementation and manifests record the lag features, fitting boundaries and seeds.

Seven candidates are evaluated: raw, OLS, bottom-up, diagonal WLS, isotropic-target Ledoit-Wolf, and uncentered and centered diagonal-target shrinkage. Frozen calibration uses a fixed pool; rolling calibration uses only residuals available before each origin. With two budgets (39, 79), two nominal levels (90%, 95%), three predictors and two folds, the two-policy comparison comprises 336 configurations per partition. The archived grid has 840 configurations because it additionally includes three experimental adaptive policies described in Online Resource 1. Those policies are not used in the main comparisons. All configurations reuse data and forecasts and are not independent experiments.

5.3 Outcome and uncertainty

The primary outcome is the equal-node mean of normalised interval scores, averaged equally over the three predictors, at $n = 39$, 90% and rolling calibration. Coverage is reported alongside. For the historical folds, paired differences between algorithms are summarised by circular block resampling of the 24 test months with block lengths 3, 6 and 12, using 5,000 replications and a fixed seed; all nodes and algorithms are aggregated before resampling so that within-month correlation is preserved. With 24 months these ranges are descriptive sensitivity summaries, not confidence intervals and not independent replication.

5.4 Historical stress-test results

Table 2 and Figure 3 report the historical primary comparison. Relative to raw forecasts, the uncentered diagonal-target estimator reduces the mean normalised score from 10.872 to 10.653, a reduction of 2.0%, and the centered variant reduces it to 10.531, a reduction of 3.1%. Diagonal WLS and bottom-up are almost indistinguishable from raw (10.829 and 10.871). The isotropic-target Ledoit-Wolf estimator increases the score to 11.597 (+6.7%) and OLS increases it to 12.156 (+11.8%).

Table 2. Mean normalised interval score, coverage and mean normalised width; final fold, $n = 39$, target 90%, rolling calibration, equal weight over three predictors.

Method	134 nodes:			59 nodes:		
	score	coverage	width	score	coverage	width
raw	10.872	0.893	8.529	10.674	0.891	8.443
bottom-up	10.871	0.894	8.527	10.673	0.894	8.431
OLS	12.156	0.885	9.496	11.627	0.881	9.155
diagonal WLS	10.829	0.894	8.489	10.645	0.892	8.398
LW isotropic	11.597	0.889	9.077	11.216	0.885	8.864
Diagonal uncentered	10.653	0.889	8.308	10.701	0.890	8.451
Diagonal centered	10.531	0.888	8.206	10.385	0.886	8.133

Coverage is not identical across rows. In the final-fold 134-node comparison, raw coverage is 0.8927 and centered-shrinkage coverage is 0.8885, while their normalized scores are 10.8723 and 10.5314. The score comparison evaluates both width and missed observations; it does not isolate width at matched realized coverage. The small coverage difference alone also does not identify how much of the score change is due to narrower intervals.

The picture changes once all six design cells are read together (Tables 3 and S5). The centred diagonal-target estimator is better than raw forecasts in every cell: -1.85% and -3.14% with 134 nodes in folds 0 and 1, -1.29% and -2.71% with 59 nodes, -2.97% and -4.59% in the recent-training variant. The uncentred reference-style estimator sits close to raw and its sign is not stable: it is better in four cells (-0.158, -0.219, -0.105 and -0.083 score units) and worse in two (+0.027 in the 59-node final fold, +0.076 in the 134-node recent-training cell). The two shrinkage variants never exchange their ordering: the centred version is better than the uncentred one in all six cells. What changes across partitions and designs is whether the uncentred variant beats raw. The isotropic-target Ledoit-Wolf estimator and the

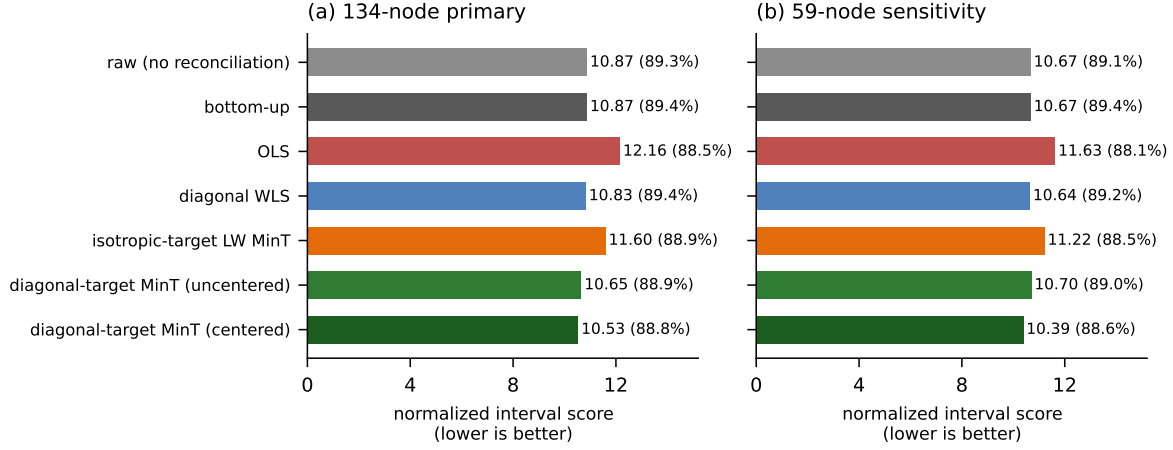


Fig. 3 Historical final-fold mean normalized interval scores for the two non-overlapping retail partitions, $n = 39$, nominal 90%, rolling calibration. Coverage is printed with each score; all nodes and three predictors have equal weight. These are long-gap stress-design results. Different realized coverages prevent interpreting score differences as matched-coverage width differences.

OLS projection are worse than raw in all six cells, by 0.54–1.21 and 0.79–2.34 score units respectively, and diagonal WLS and bottom-up are within ± 0.18 everywhere.

The historical results support estimator-specific conclusions on these designs. They do not identify a single effect of reconciliation independent of the covariance recipe, nor do they isolate shrinkage target from shrinkage-intensity estimation and centering.

Table 3. Historical fold-0 and later-training sensitivity scores, with the same configuration as Table 2. The recent columns use their own training-period normalization and must not be subtracted directly from the earlier-design columns; common-scale comparisons are reported in Section 5.5. Coverage is in Table S5.

Method	134 nodes, f0	59 nodes, f0	134 nodes, recent	59 nodes, recent
raw	13.058	12.530	7.006	7.015
bottom-up	13.062	12.490	6.895	6.837
diagonal WLS	13.034	12.510	6.960	6.989
Diagonal centered	12.816	12.369	6.798	6.693
Diagonal uncentered	12.899	12.425	7.082	6.932
LW isotropic	13.739	13.136	8.214	7.768
OLS	14.088	13.322	9.349	8.742

Per-model results (Table S3) show that centered shrinkage improves on raw for both non-coherent predictors: ridge moves from 6.319 to 6.072 and random forest from 18.644 to 17.868. The isotropic-target estimator scores 7.101 and 20.037, respectively. Bottom-up is almost neutral in the three-model mean but worse for ridge and better for random forest. Seasonal naive remains 7.654 for every projection. Random forest’s much larger score gives it a substantial role in the equal-model mean, so the mean should be read together with these disaggregated results.

The main shrinkage comparisons remain directionally similar across the examined calibration budgets and nominal levels. With $n = 79$ and nominal 90%, the mean score for the 134-node partition moves from 12.698 (raw) to 12.554 (uncentered, -1.1%) and 12.440 (centered, -2.0%), while the isotropic-target estimator reaches 13.613 (+7.2%) and OLS 14.090 (+11.0%). At nominal 95% with $n = 39$, raw scores 12.377 and the same pair scores 12.202 and 12.071, against 13.260 for the isotropic target and 13.879 for OLS; with $n = 79$ the ordering is unchanged again. Coverage in these configurations stays within 0.014 of nominal for all seven candidates. The principal shrinkage contrasts retain their directions in these 134-node configurations, although their magnitudes change.

5.5 Later-training sensitivity and common normalization

The later-training design retains a 24-month covariance window, uses a 39-month calibration pool and removes the unused pilot buffer. Model fitting ends in September 2011; testing remains in 2017-2018.

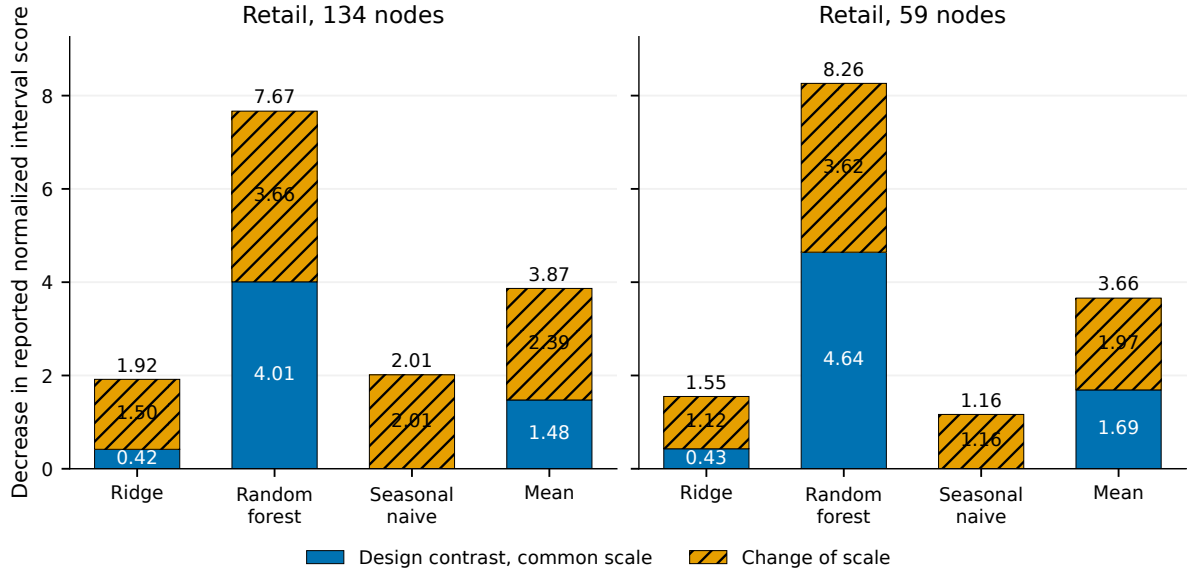


Fig. 4 Original final-fold raw forecasts, $n = 39$, nominal 90%, rolling calibration. The decrease in reported score is decomposed algebraically into a design contrast at the common old scale (blue) and replacing that scale for the later-training intervals (hatched orange). This ordered decomposition is descriptive and does not isolate a causal effect of refitting. Seasonal-naïve intervals are identical across designs; their entire score change is due to normalization. The source table gives all components.

The change was documented before that particular rerun, but followed earlier inspection of the dataset and does not constitute independent preregistration.

Under each design’s own training-period scales, the raw score changes from 10.872 to 7.006 with 134 nodes, and from 10.674 to 7.015 with 59 nodes. These are not comparable units because the denominator changes with the training window. Re-evaluating later-training intervals with the original design’s scales gives 9.397 and 8.983, respectively. Figure 4 separates this scale change from the common-scale contrast. The common-scale changes are -1.475 and -1.691, rather than the larger differences obtained by subtracting design-specific scores. This contrast still changes the training schedule and cannot identify a general causal effect of the training gap.

For the 134-node seasonal-naïve predictor, old and later-training test outcomes and interval endpoints are identical, yet the normalized score changes from 7.654 to 5.640 solely because of the denominator. Its later-training score is again 7.654 under the old scales. Online Resource 1 reports common-normalization method averages; Online Resource 2 contains the complete predictor-level values.

Within the later-training design, centered shrinkage scores below raw in both partitions (-2.97% and -4.59%), with equal or slightly higher realized coverage. Uncentered shrinkage is worse than raw with 134 nodes and better with 59 nodes; isotropic-target shrinkage and OLS are worse in both. The historical ordering is therefore less sensitive than the score level to these particular design choices. Neither frozen-model schedule represents routine updating; the sequential benchmark in Section 5.10 addresses that limitation directly.

5.6 Paired temporal sensitivity (final fold, long-gap design)

Table 4 and Figure 5 summarise paired differences over the 24 test months of the final fold. With 134 nodes the uncentered contrasted with raw gives -0.219 [-0.294, -0.143] at block 6 and the centered contrast gives -0.341 [-0.447, -0.234]; the isotropic-target contrast is positive at +0.725 [0.631, 0.827]. With 59 nodes the uncentered contrast straddles zero (+0.027 [-0.059, 0.111]) while the centered contrast remains negative (-0.289 [-0.396, -0.179]).

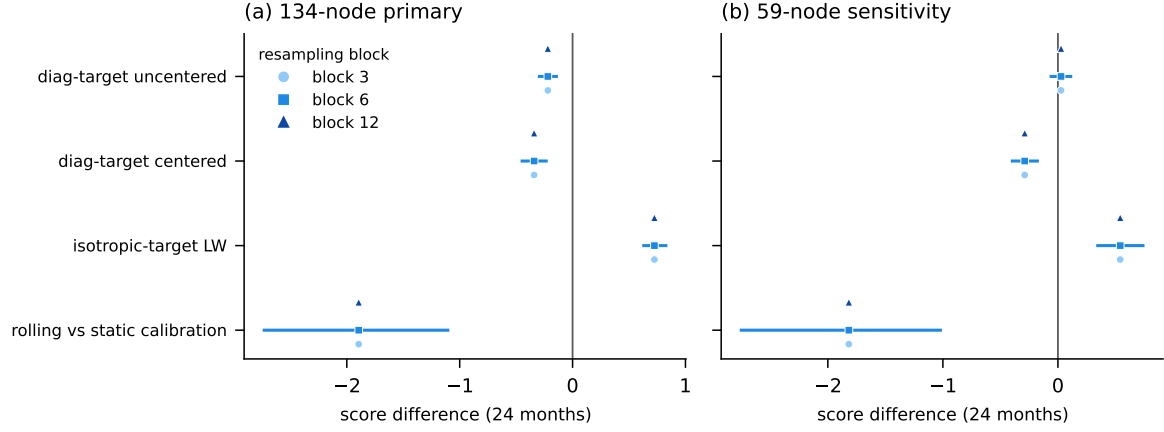


Fig. 5 Historical final-fold paired differences in mean normalized interval score over 24 test months, using circular block resampling with block lengths 3, 6 and 12 months. All nodes and predictors are aggregated before resampling. The displayed ranges describe temporal sensitivity; they are not confidence intervals or independent replications.

Table 4. Paired differences in mean normalised score, 24 test months, $n = 39$, target 90%; block-6 resampling with 5,000 replications.

Contrast	134 nodes: mean [range]	59 nodes: mean [range]
Diagonal uncentered - raw	-0.219 [-0.294, -0.143]	+0.027 [-0.059, 0.111]
Diagonal centered - raw	-0.341 [-0.447, -0.234]	-0.289 [-0.396, -0.179]
LW isotropic - raw	+0.725 [0.631, 0.827]	+0.541 [0.346, 0.739]
rolling - frozen calibration (raw forecasts)	-1.895 [-2.732, -1.105]	-1.817 [-2.753, -1.022]

In this historical 134-node cell, switching raw forecasts from frozen to rolling calibration reduces score by 1.895 units: 5.6 times the centered-shrinkage improvement of 0.341, and 8.7 times the uncentered improvement of 0.219. The centered contrast is the larger of these two reconciliation improvements. These ratios describe this design and should not be transferred to a model-updating regime without evaluation.

5.7 Source verification

The `aus_retail` archive inside the hash-locked source package was decoded twice. The packaged Python decoder (`rddata`) and an independent in-package re-implementation agree on all 64,532 records. Separately, R 4.3.1 loaded the same `.rda` and exported every record; the comparison against the Python decoding matches all 64,532 rows in file order, with identical State, Industry and Series ID labels, identical stored month values (maximum absolute difference 0.0) and identical Turnover values (maximum absolute difference 0.0). The export reports 152 series, 8 states, 20 industries and months from April 1982 to December 2018, with 32 to 441 observations per series. The result is stored in `RESULTS/official_reader_crosscheck_full.json`.

5.8 Covariance diagnostics

The retail covariance window has 24 residual vectors against 134 nodes; the centered sample covariance has rank at most 23, as recorded in the diagnostics. Shrinkage intensities range from 0.06 to 0.32 for the isotropic target and from 0.11 to 0.45 for diagonal-target shrinkage, depending on model and fold. An isotropic target moves node variances toward a common scale, whereas a diagonal target retains the node-specific diagonal. This distinction helps interpret the estimator recipes, but their empirical score gap cannot be attributed exclusively to the target: intensity formulas also differ, and uncentered comparisons additionally differ in centering.

5.9 Second source

The 525-node representation removes 30 exactly duplicated aggregate aliases from the source’s 555 series while retaining all 304 bottom-level series. The provenance record identifies every removed alias and its identical retained series.

An exploratory historical comparison uses a 525-node TourismLarge hierarchy with 228 monthly observations and 304 bottom-level series, the same seven projections and three predictors, and a single 24-month test window with rolling calibration. Its covariance window has only 16 residual vectors; diagonal-target shrinkage intensities of 0.84-0.93 indicate strong regularization. This small covariance window limits what can be inferred from the estimated correlation structure. Section 5.10 evaluates the same source with an updated sequential protocol.

In this historical tourism comparison, diagonal WLS, both diagonal-target variants and bottom-up change the two non-coherent predictors’ scores by approximately -0.11 to $+0.04$. Isotropic-target shrinkage is worse by 1.44 and 1.39, and OLS by 0.32 and 0.47 (Table S4 in Online Resource 1). Seasonal naive is invariant across projections. A zero-variance covariance-window coordinate for seasonal naive is handled by the explicit, recorded numerical floor specified in Online Resource 1. This is an additional dataset illustration, with limited test length and covariance information.

5.10 Sequential evaluation with annual model updates

To evaluate a more routinely updated forecasting setting, we use a 72-month prequential test: each prediction and calibration interval uses only information available before its outcome. Expanding-window point models are refitted every 12 months. The retail test runs from January 2013 through December 2018, and the tourism test from January 2011 through December 2016. The first test forecasts use models last fitted through September 2012 and September 2010, respectively; later annual fits use only preceding observations. Lag features use positive lags. Hyperparameters and deterministic seed rules are held fixed.

At origin t , a 24-vector covariance window uses out-of-sample residuals indexed $[t - 63, t - 39]$ and a disjoint 39-vector calibration window uses $[t - 39, t)$. The current projection is fitted from the covariance window, then applied to the historical out-of-sample residual vectors used for calibration. Those residuals retain the annually updated point-model sequence that generated them. We evaluate seven reconciliation candidates and three predictors at 90% and 95% nominal levels. Node scales are computed once from seasonal differences before the test window and remain fixed across all methods and origins within that dataset. This is a chronological evaluation under temporal dependence; exact exchangeable rank coverage is not asserted.

Figure 6 displays score changes relative to raw with realized coverage.

Table 5. Mean normalized interval score and realized coverage in the 72-month sequential benchmark, nominal 90%, calibration size 39, equal weight over predictors. Each dataset has a fixed pre-test normalization; absolute scores should not be compared across datasets.

Method	Retail 134: score / coverage	Retail 59: score / coverage	Tourism 525: score / coverage
Raw	6.2867 / 0.8833	5.7203 / 0.8774	7.1622 / 0.8955
Bottom-up	6.2437 / 0.8846	5.6447 / 0.8763	7.1656 / 0.8958
OLS	8.3997 / 0.8768	7.4619 / 0.8755	7.3808 / 0.8914
Diagonal WLS	6.2300 / 0.8826	5.6767 / 0.8745	7.1004 / 0.8952
LW isotropic	7.5028 / 0.8760	6.5983 / 0.8731	7.7448 / 0.8902
Diagonal uncentered	6.1921 / 0.8797	5.5867 / 0.8739	7.1007 / 0.8948
Diagonal centered	6.1619 / 0.8797	5.5306 / 0.8754	7.1035 / 0.8949

Centered diagonal-target shrinkage improves on raw by 1.99% and 3.32% in the two retail partitions and by 0.82% in tourism. The uncentered variant also improves on raw, but has a slightly lower score than the centered variant in tourism; diagonal WLS is slightly lower again there. Hence centered shrinkage is not a uniformly best estimator. Isotropic-target shrinkage and OLS score above raw in all three evaluations. Realized 90% coverage remains below nominal for the raw and centered methods in each dataset, so these gains are not equal-coverage width comparisons. The longer sequential evaluation reduces dependence on the historical frozen-model schedule but remains a retrospective comparison of

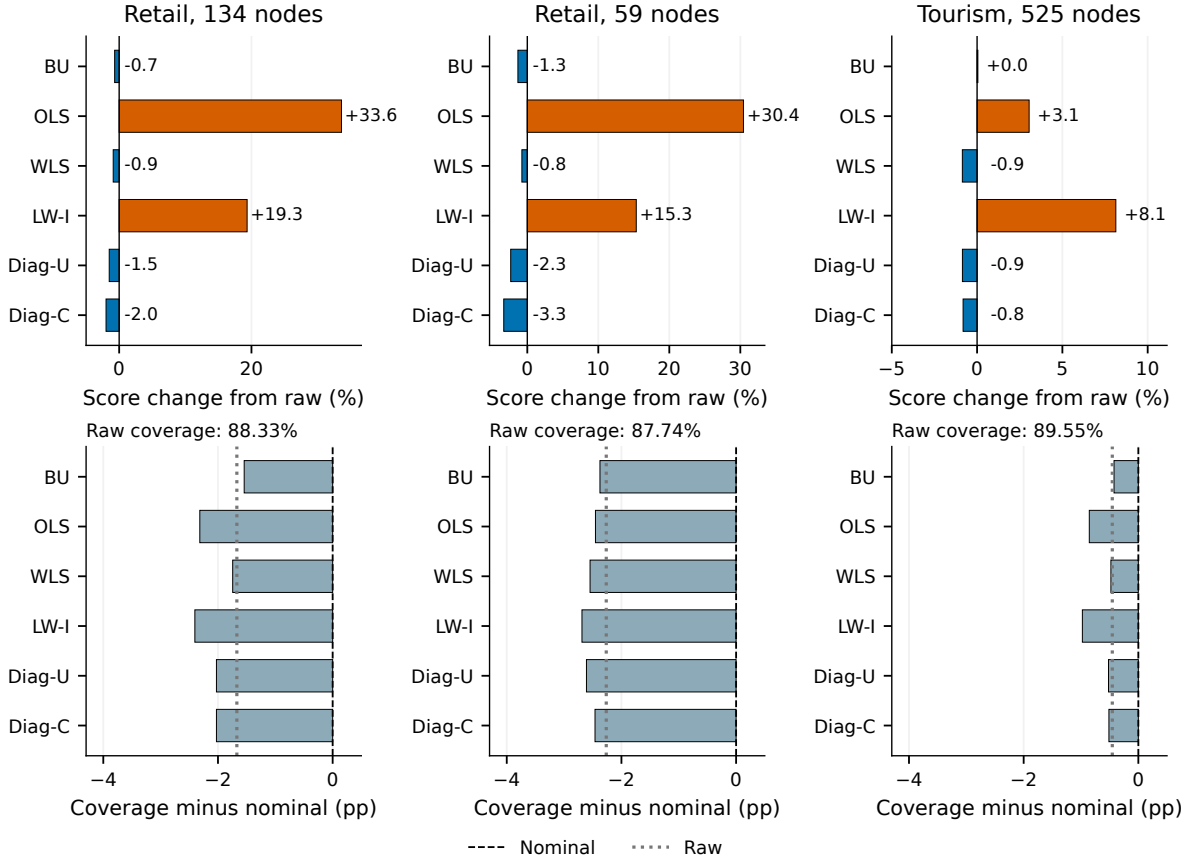


Fig. 6 Seventy-two-month prequential validation at nominal 90%, $n=39$. Upper panels show equal-model/equal-node mean normalized interval-score changes from raw forecasts; negative is better. Lower panels show coverage gaps in percentage points with raw forecasts as dotted references. Model refits are annual and expanding, covariance is fitted to 24 prior out-of-sample residual vectors preceding the 39-vector calibration window, and calibration is past-only. The scale is fixed before the first test month. The two retail partitions share a source and are not independent replications. Abbreviations: BU, bottom-up; WLS, diagonal weighted least squares; LW-I, isotropic-target Ledoit-Wolf; Diag-U/Diag-C, uncentered/centered diagonal-target shrinkage. Coverage is empirical, not guaranteed.

two data sources, including related retail partitions. Complete 90% and 95% results and origin-level boundaries are in Online Resources 1 and 2.

6 Reproducibility and use of AI-assisted tools

The computations use Python 3.11 on a Linux server with NumPy, SciPy, pandas, scikit-learn, statsmodels, threadpoolctl and rdata; numerical-library threading is restricted in the reproduction environment. The archived historical runs include manifests and per-configuration intervals. A server replay from copied historical forecast caches reproduces all 840 configuration rows for each retail partition; this check does not constitute new historical model retraining. The sequential benchmark records every point-model fitting boundary, covariance and calibration window, interval endpoint and score. These records allow the reported summaries to be recomputed independently of the fitting step. For the new 72-month evaluation, two runs from empty prediction caches on the same server and environment reproduce all forecasts, interval arrays and scientific tables exactly. Independent checks cover the 126 interval-endpoint files and 216 monthly role schedules, and perturbing current or future labels does not change the corresponding predictions or intervals. These are numerical and information-flow checks, not independent scientific replication.

Exact rational certificates supply the twelve theoretical witnesses. An independent convolution implementation checks the displayed values, and the continuous curve is numerically evaluated for illustration; the proof depends on the exact signs and the uniform smoothing bound. Online Resource 2 contains the source archive references, implementation, environment specification, validation outputs and file hashes.

Generative AI tools assisted code implementation, numerical analysis, literature organization and manuscript drafting. The numerical results are accompanied by executable code and verification records,

rather than relying on generated prose as evidence. AI systems are not authors and do not assume responsibility for the work. Scientific interpretation, verification and the final text remain the responsibility of the human authors.

7 Discussion

The construction shows that second-moment information alone cannot determine ordinary finite-sample interval-width efficiency on the stated nonelliptical family. Under a common elliptical residual shape with finite covariance, scale-based comparisons remain available. Extending a positive efficiency result beyond that class would require additional shape control or an explicitly different loss criterion. The construction does not describe how frequently its sign pattern occurs in forecasting data.

The empirical comparisons concern existing covariance recipes and chronological evaluation choices. In both the historical stress tests and the sequential benchmark, isotropic-target shrinkage and OLS score above raw forecasts in the configurations reported here. Diagonal-target shrinkage can yield modest improvements, but centering is not uniformly preferable: the tourism sequential comparison slightly favors the uncentered version and diagonal WLS. Estimator labels should therefore specify the target, centering and intensity rule instead of presenting a single undifferentiated minimum-trace result.

Evaluation design also changes interpretation. The large frozen-versus-rolling contrast in the historical long-gap design is design-specific and should not be extrapolated to updated-model settings. Absolute normalized scores across training schedules additionally change when their reference scales change. Keeping a fixed denominator in cross-design comparisons avoids crediting a rescaling effect to the forecast model. These observations support transparent reporting of role boundaries, model updates, scales, fitted covariance rank, shrinkage intensity and coverage.

The theoretical and empirical results answer different questions. The first is an existence result at exact single-node coverage. The second measures score and realized coverage in dependent time series. Their conjunction motivates measuring interval performance directly rather than inferring it from variance reduction; it does not turn the empirical score improvements into a distribution-free efficiency guarantee.

8 Limitations

The theory treats two- and three-leaf stars, a ternary-plus-Gaussian residual family, $n = 39$ and two rank pairs. The smoothing thresholds are sufficient bounds, and the coherent-response realization is artificial. The result establishes existence, not prevalence, and does not characterize general hierarchies, conditional coverage, simultaneous coverage, serial dependence or interval-score optimality.

The empirical analyses are retrospective. Earlier results informed subsequent design repairs and sensitivities. The historical folds have 24-month tests and stale point-model fits; their block-resampling ranges describe temporal sensitivity rather than a confirmatory sampling distribution. The sequential tests cover 72 months per dataset but include only two data sources, with retail partitions sharing observations. Their rolling residuals come from changing point models, and exchangeability is not assumed. Realized coverage differs across methods and can fall below nominal.

The comparisons evaluate established estimators under specified hyperparameters and aggregation choices. They do not isolate shrinkage target from every other estimator feature, establish a universally best method, or generalize modest score gains to other forecasting tasks. The R source-data read is checked against the Python decoding; the original `hts` reconciliation package was not numerically executed, so its formula-level reimplement is not presented as an R-to-Python numerical replication.

9 Data and code availability

Online Resource 1 (`ESM_1.pdf`) contains the proof details, numerical conventions, historical sensitivity tables, common-normalization checks and complete sequential summary tables. Online Resource 2 (`ESM_2.zip`) contains the analysis code, source and data provenance, result arrays, configuration manifests, environment and reproduction instructions, validation records and a hash inventory. The retail source is `aus_retail` in `tsibbledata` 0.4.1, based on ABS aggregate statistics; the exploratory second source is the TourismLarge hierarchy identified in the provenance record. The full Tourism input is reconstructed from the recorded public source by a hash-checked download script rather than redistributed

in the archive. Third-party data and source archives retain their original attributions and licenses. The package identifies which files support each reported result.

Statements and Declarations

Funding. No funding was received for this study.

Competing interests. The authors declare no competing interests relevant to this work.

Ethics approval and consent. The study uses public aggregate time series and does not involve human participants or animals; ethics approval and participant consent are not applicable.

AI-assisted work. Generative AI assistance is described in Section 6. The human authors are responsible for the scientific content and final text.

References

- Athanasopoulos G, Hyndman RJ, Kourentzes N, Panagiotelis A (2024) Forecast reconciliation: a review. *International Journal of Forecasting* 40(2):430-456. <https://doi.org/10.1016/j.ijforecast.2023.10.010>
- Australian Bureau of Statistics (2024) Retail Trade, Australia methodology, January 2024: industry group and subgroup definitions. <https://www.abs.gov.au/methodologies/retail-trade-australia-methodology/jan-2024>. Accessed 27 October 2026
- Dhillon GS, Deligiannidis G, Rainforth T (2024) On the expected size of conformal prediction sets. In: *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics, Proceedings of Machine Learning Research* 238:1549-1557. <https://proceedings.mlr.press/v238/dhillon24a.html>
- Gibbs I, Candès EJ (2024) Conformal inference for online prediction with arbitrary distribution shifts. *Journal of Machine Learning Research* 25(162):1-36. <https://www.jmlr.org/papers/v25/22-1218.html>
- Hong H, Xu S, Yang Y, Hu J, Chen Y, Wu J (2026) Hierarchical reconciliation and per-bus adaptive conformal calibration for distribution-level residential load forecasting: a critical assessment of topology priors. *Energy Informatics*. <https://doi.org/10.1186/s42162-026-00690-1>
- Hyndman RJ, Ahmed RA, Athanasopoulos G, Shang HL (2011) Optimal combination forecasts for hierarchical time series. *Computational Statistics & Data Analysis* 55(9):2579-2589. <https://doi.org/10.1016/j.csda.2011.03.006>
- Hyndman RJ, Athanasopoulos G (2021) *Forecasting: Principles and Practice*, 3rd edn. OTexts. <https://otexts.com/fpp3/hierarchical-forecasting.html>. Accessed 27 October 2026
- Hyndman RJ et al. (n.d.) hts: hierarchical and grouped time series, author reference implementation, R/MinT.R. <https://github.com/robjhyndman/hts/blob/master/R/MinT.R>. Accessed 27 October 2026
- O'Hara-Wild M, Hyndman R, Wang E, Godahewa R (2022) tsibbledata: diverse datasets for tsibble, version 0.4.1. <https://tsibbledata.tidyverts.org/>. Accessed 27 October 2026
- Panagiotelis A, Athanasopoulos G, Gamakumara P, Hyndman RJ (2021) Forecast reconciliation: a geometric view with new insights on bias correction. *International Journal of Forecasting* 37(1):343-359. <https://doi.org/10.1016/j.ijforecast.2020.06.004>
- Panagiotelis A, Gamakumara P, Athanasopoulos G, Hyndman RJ (2023) Probabilistic forecast reconciliation: properties, evaluation and score optimisation. *European Journal of Operational Research* 306(2):693-706. <https://doi.org/10.1016/j.ejor.2022.07.040>
- Principato G, Stoltz G, Amara-Ouali Y, Goude Y, Hamrouche B, Poggi JM (2026) Conformal prediction for hierarchical data. *Transactions on Machine Learning Research*. <https://arxiv.org/html/2411.13479v4>
- Wickramasuriya SL, Athanasopoulos G, Hyndman RJ (2019) Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. *Journal of the American Statistical Association* 114(526):804-819. <https://doi.org/10.1080/01621459.2018.1448825>