CoFlow: Leaf-Space Factorization for Point and Probabilistic Forecast Reconciliation

Anonymous authors
Paper under double-blind review

Abstract

Forecast reconciliation is usually posed as a point-adjustment problem. When forecasts must also quantify uncertainty, coherence must extend to the joint predictive law, while its center may not optimize the target point-forecast loss. We introduce CoFlow, a post-hoc leaf-space framework that separates a coherent anchor, a joint error law, and a separately optimized point forecast. Mapping both error draws and point corrections through the summing matrix guarantees coherence while allowing predictive shape and point accuracy to be selected independently. This factorization also exposes aggregate dispersion as a central calibration problem. For Gaussian node errors with any predictive-to-true variance ratio, we theoretically derive exact expressions for coverage and expected CRPS and interval-score regret, together with quadratic lower bounds. Under attenuated cross-leaf covariance, these penalties increase with the share of aggregate variance carried by dependence. CoFlow combines Gaussian, copula, count, empirical, and flow-based leaf laws with coherent attention-based and sparse constraint-routing point heads. Across 29 hierarchies, three rolling origins, and five seeds, Tweedie with routing improves relative MAE by 9.8%, nCRPS by 6.2%, and normalized interval score by 16.0% over MinT. An RBIG-Gaussian mixture improves nCRPS on all 29 datasets and both nMIS and coverage gap on 28. Ablations isolate gains from dependence, marginal shape, and point estimation.

1 Introduction

Many forecasting problems carry a *hierarchy*: individual series aggregate into higher-level totals through fixed linear constraints, such as products into categories or stores into regions. A set of forecasts is *coherent* when it respects these constraints exactly. Base forecasts produced independently per series are almost never coherent, so a *reconciliation* step maps them onto the coherent subspace (Hyndman et al., 2011; Wickramasuriya et al., 2019; Athanasopoulos et al., 2024; Wang et al., 2023). The dominant map, minimum-trace reconciliation (MinT), is the best linear unbiased projection onto that subspace, and its Gaussian predictive is governed by the covariance of the base-forecast errors (Wickramasuriya et al., 2019). Probabilistic coherence has since been pursued through temporal hierarchies (Athanasopoulos et al., 2017), score-optimised evaluation (Panagiotelis et al., 2023), and end-to-end trainable coherent models (Rangapuram et al., 2021; 2023; Olivares et al., 2024; Das et al., 2023).

An empirical limitation of the projection approach concerns its *predictive dispersion at aggregate levels* when W is estimated. Each aggregate node is a sum of many leaves, and the variance of that sum is set by the leaf variances together with all pairwise leaf covariances. Estimation error or distributional misspecification can therefore produce a nodewise variance ratio below one even when the predictive mean is accurate. We derive the resulting coverage and proper-score costs exactly and show that covariance attenuation is amplified by the share of aggregate variance carried by cross-leaf dependence. Shrinkage toward a diagonal target (Schäfer & Strimmer, 2005; Ledoit & Wolf, 2004) can affect that share. Empirically, the estimated MinT predictive often has narrow aggregate intervals and coverage below nominal (Panagiotelis et al., 2023; Wickramasuriya, 2024). Section 3.2 locates this problem precisely within the Gaussian MinT covariance.

We address this estimation problem with CoFlow, a post-hoc factorization into three independently configurable parts: a coherent leaf anchor, a joint leaf-error law, and a separately optimized coherent

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

point forecast. Given anchor $\hat{b}$ and error draw η, the predictive sample is $y = S(\hat{b} + \eta)$. This wraps any base forecaster, and coherence holds by construction because $HS = 0$. Centering the leaf errors retains the MinT predictive location, while a separate differentiable point head trained through S can optimize the point forecast without moving a predictive sample. The main contributions of the paper are:

- We formulate reconciliation in leaf space, factorizing the coherent anchor, joint leaf-error law, and point forecast. This post-hoc construction wraps any base forecaster and guarantees coherence of every sample and point forecast.
- We derive a dispersion-regret theorem connecting nodewise variance error to coverage, CRPS, and interval score under variance misspecification, and show that covariance attenuation is amplified by its share of aggregate variance.
- We introduce coherent attention-based and sparse constraint-routing point heads that redistribute aggregate discrepancies through S, with message conservation, and metric-aligned losses.
- Across 29 hierarchies, controlled ablations isolate the effects of dependence, marginal shape, residual source, anchor, flow, mixture, point architecture, and loss effects under one sample-based evaluation. The resulting compositions achieve the best point and probabilistic scores among the 72 evaluated methods.

2 Background and notation

A hierarchy of n series with m leaves is encoded by a full-column-rank summing matrix $S \in \mathbb{R}^{n\times m}$. With leaves ordered last, $S = [A^\top \; I_m]^\top$ and every coherent vector has the form $y = Sb$. Hence $\mathcal{C} = \mathrm{range}(S)$ has dimension m. Let $H \in \mathbb{R}^{(n-m)\times n}$ have orthonormal rows spanning $\mathcal{C}^\perp$, so $HS = 0$ and coherence is equivalent to $Hy = 0$. A base forecaster provides h-step forecasts $\hat{y} \in \mathbb{R}^{n\times h}$ and in-sample residuals $E = [e_1, \ldots, e_T]$. We write E_b for the leaf rows of E.

Under base-error covariance $W \succ 0$, MinT is the trace-optimal unbiased linear reconciler (Wickramasuriya et al., 2019):

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

$$\tilde{y} = P_W \hat{y}, \qquad \widetilde{\Sigma} = P_W W P_W^\top = S(S^\top W^{-1} S)^{-1} S^\top. \tag{2}$$

Our static reference is MinT-shrink, with $P_\lambda = P_{W_\lambda}$ and

$$W_\lambda = (1-\lambda)\widehat{W} + \lambda \,\mathrm{diag}(\widehat{W}), \qquad \widehat{W} = T^{-1} E E^\top, \tag{3}$$

where $\lambda \in [0, 1]$ is the closed-form shrinkage intensity (Ledoit & Wolf, 2004; Schäfer & Strimmer, 2005). We evaluate point accuracy with geometric relative MAE, WMAPE, and MASE, and predictive accuracy with sample-based nCRPS, central interval score and coverage, and multivariate energy score (Gneiting & Raftery, 2007; Gneiting et al., 2008).

3 Coherent generative reconciliation in leaf space

3.1 Method overview

CoFlow separates the coherent anchor, predictive law, and coherent point forecast. We use the MinT-shrink leaf anchor $\hat{b} = J P_\lambda \hat{y}$, fit a joint law ν to centered raw leaf residuals, and draw K errors $\eta^{(k)}$. Their finite-sample mean is removed before aggregation. The predictive samples and point forecast are

$$y^{(k)} = S(\hat{b} + \eta^{(k)}), \qquad y_{\mathrm{point}} = S\{\hat{b} + \delta_\theta(f)\}. \tag{4}$$

Both outputs lie in $\mathrm{range}(S)$ for every leaf law and every point head. Centering fixes the ensemble mean at $S\hat{b}$, while δ_θ changes only the point forecast. Thus predictive shape and point accuracy can be studied and composed independently. Figure 1 shows the two branches. Full model specifications and centering properties are given in Appendix A.1.

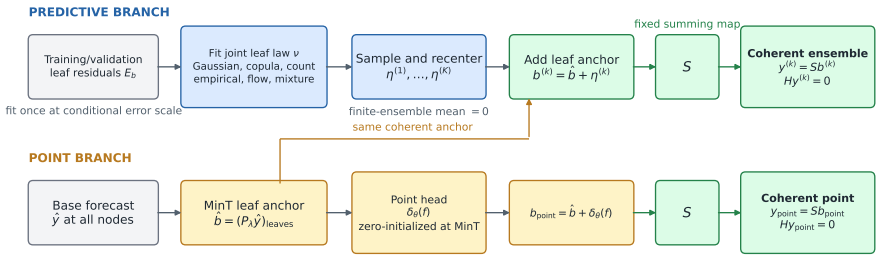


Figure 1: Overview of CoFlow. The predictive branch fits a joint leaf law at the conditional error scale, recenters its samples, and adds the MinT leaf anchor. The point branch learns a separate correction to the same anchor. Both branches pass through the fixed summing matrix S, so every predictive sample and the point forecast are coherent by construction.

3.2 Leaf-space structure and Gaussian MinT

Because S has full column rank, it gives one-to-one coordinates for coherent laws. In particular, every coherent law is $S_{\#}\nu$ for a unique leaf law ν. Appendix A.1 states this fact formally and Appendix A.5 gives its Wasserstein, divergence, coverage-event, and energy-score consequences.

MinT in leaf coordinates. Under a Gaussian base law, conditioning on the linear constraints is equivalent to Gaussian MinT (Corani et al., 2020; Wickramasuriya, 2024; Zambon et al., 2024a). Let $J = [\,0\ I_m\,]$ select the leaf rows, $W_{bb} = JWJ^\top$, $R_W = HWH^\top$, and $B_W = JWH^\top$. The unique leaf covariance of the MinT predictive is

$$C_W = (S^\top W^{-1} S)^{-1} = W_{bb} - \underbrace{B_W R_W^{-1} B_W^\top}_{D_W \succeq 0}. \tag{5}$$

Hence MinT is the leaf pushforward $S_{\#}\mathcal{N}(JP_W\hat{y}, C_W)$. The PSD term D_W is the standard Gaussian conditional-covariance reduction. When the Gaussian model and W are correct, this reduction is calibrated. Our comparison concerns the estimated predictive: a direct leaf law uses its own covariance $\widehat{\Sigma}_b$ and can represent non-Gaussian shape. The complete decomposition and calibrated-model argument are given in Appendix A.1.

3.3 From covariance attenuation to score regret

The conditioning identity alone does not imply miscalibration. We therefore separate two questions. First, what is the score cost of using the wrong Gaussian variance at one node? Second, when does attenuating cross-leaf covariance make that variance error larger under aggregation? Let $\Sigma_b \in \mathbb{R}^{m\times m}$ be a leaf-error covariance and let $s_a \in \{0,1\}^m$ be row a of S. Coherent propagation gives

$$\mathrm{Var}(y_a) \;=\; s_a^\top \Sigma_b\, s_a \;=\; \sum_{i\in L_a} (\Sigma_b)_{ii} \;+\; \sum_{i\neq j\in L_a} (\Sigma_b)_{ij}, \tag{6}$$

where $L_a = \{i : (s_a)_i = 1\}$. The next result depends only on the scalar variance ratio, so it applies regardless of how that ratio arose.

Theorem 1 (Gaussian dispersion regret)**.** *Fix $0 < \alpha < 1$. Let the true node error be $Y \sim \mathcal{N}(0, v)$ and let the predictive law be $F = \mathcal{N}(0, \widehat{v})$, where $v, \widehat{v} > 0$. Define $r = \widehat{v}/v$, $z = z_{1-\alpha/2}$, and $G_\alpha(c) = 2c + 4\{\phi(c) - c[1-\Phi(c)]\}/\alpha$. Then*

$$\mathrm{Cov}(r, \alpha) = 2\Phi(z\sqrt{r}) - 1, \tag{7}$$

$$\frac{\Delta_{\mathrm{CRPS}}(r)}{\sqrt{v}} = \frac{\sqrt{2(1+r)} - \sqrt{r} - 1}{\sqrt{\pi}}, \tag{8}$$

$$\frac{\Delta_{\mathrm{MIS}}(r)}{\sqrt{v}} = G_\alpha(z\sqrt{r}) - G_\alpha(z). \tag{9}$$

Both regrets are nonnegative and vanish only at $r = 1$. Coverage is below nominal exactly when $r < 1$. On $0 < r \leq 1$, both normalized regrets increase strictly as r decreases, and

$$\Delta_{\mathrm{CRPS}}(r) \geq \frac{\sqrt{v}}{4\sqrt{\pi}}(1 - \sqrt{r})^2, \tag{10}$$

$$\Delta_{\mathrm{MIS}}(r) \geq \sqrt{v}\frac{2\phi(z)z^2}{\alpha}(1 - \sqrt{r})^2. \tag{11}$$

The proof is in Appendix A.1.3.

Corollary 1 (Covariance-weighted attenuation under aggregation)**.** *Suppose a Gaussian predictive preserves leaf variances and attenuates nonnegative within-node cross-covariances by factors $\kappa_{ij} \in [0, 1]$. Let d_a and c_a be the diagonal and cross-covariance contributions to node a, let $\theta_a = c_a/(d_a + c_a)$, and let $\bar{\kappa}_a$ be the covariance-weighted mean of the κ_{ij}. Then*

$$r_a = 1 - (1 - \bar{\kappa}_a)\theta_a. \tag{12}$$

Coverage and both normalized regrets worsen as θ_a increases or $\bar{\kappa}_a$ decreases. For q exchangeable leaves with correlation $\rho > 0$ and common attenuation $\kappa < 1$,

$$r_q = \frac{1 + \kappa\rho(q - 1)}{1 + \rho(q - 1)}. \tag{13}$$

Thus r_q and coverage decrease strictly with q, while both normalized regrets increase.

The proof, exact definitions, and limits of the support-size interpretation are in Appendix A.1.3. The key quantity is the covariance share θ_a, not depth alone. Our dependence ablation tests this mechanism in Section 4.2.

3.4 Leaf laws

The leaf law controls two distinct features: marginal shape and cross-leaf dependence. Gaussian and diagonal Gaussian laws isolate the dependence contribution. Gaussian and Student-t copulas, empirical resampling, RBIG (Laparra et al., 2011), and RealNVP (Dinh et al., 2017) add increasing flexibility in marginals or dependence. For intermittent nonnegative series, Tweedie and zero-inflated negative-binomial marginals represent zeros and positive skew, with dependence imposed by rank matching. Fixed mixtures such as RBIG+Gaussian combine models with complementary errors.

Continuous laws are fitted to centered raw leaf residuals. Count dispersion is matched to the conditional residual scale so that variation already explained by the base forecaster is not added back as uncertainty. The default anchor remains MinT-shrink. Anchor, residual-source, dependence, marginal, flow-objective, and mixture variants change one part of the factorization at a time. Full definitions and fitting details are in Appendix A.1.4.

3.5 Coherent point estimation

Centering keeps the predictive ensemble mean at the MinT anchor $S\hat{b}$, but the point forecast can target a different decision loss. We therefore learn a leaf correction $\delta_\theta(f)$ and report $S\{\hat{b} + \delta_\theta(f)\}$. Every parameter value is coherent, and higher-level errors backpropagate through the fixed map S.

We study four information patterns (Vaswani et al., 2017). A shared MLP uses local leaf features. LeafAttn lets leaves exchange information and redistribute corrections across series. Cross-level attention adds aggregate tokens so leaves can use direct forecasts and residual scales from every hierarchy level. Its full, tree-masked, and leaf-clique variants control which node pairs communicate. The fourth pattern, RouteAttn, encodes the hierarchy directly and scales with its nonzero edges.

The token features distinguish level, uncertainty, structure, and disagreement. Leaf tokens contain the transformed anchor level, residual scale, historical series size, normalized displacement from the direct leaf forecast, and recent forecast bias. Cross-level tokens add direct all-node forecasts, node scale, and a level embedding. Routing aggregate tokens contain the normalized discrepancy, transformed direct forecast, log support size, level, and horizon. Leaf keys use the corresponding

Table 1: Information paths in the coherent point heads. Complexity assumes fixed hidden width. Attention masks restrict information flow but do not sparsify the dense Transformer computation.

Head	Tokens	Permitted information path	Complexity
Shared MLP	Leaves	Local features only	$O(m)$
LEAFATTN	Leaves	Every leaf pair	$O(m^2)$
FULLATTN	All nodes	Every node pair	$O(n^2)$
TREEATTN	All nodes	Lineage and same-parent siblings	$O(n^2)$
LEAFTREEATTN	All nodes	Lineage plus a complete leaf subgraph	$O(n^2)$
ROUTEATTN	Aggregates and leaves	Aggregate to descendant leaves only	$O(\mathrm{nnz}(S))$

leaf features. Every normalization is estimated on the training origins. Thus an aggregate forecast can influence a leaf only through an attention edge or an explicit descendant route, rather than by being treated as another unconstrained output.

For aggregate node a, routing forms the discrepancy $g_a = \hat{y}_a - s_a^\top \hat{b}$ between its direct base forecast and the coherent anchor. Shared encoders produce aggregate queries q_a, leaf keys k_i, and bounded gains c_a. A softmax is evaluated only over descendants of a:

$$\pi_{ai} = \frac{\exp(q_a^\top k_i)}{\sum_{j:S_{aj}=1} \exp(q_a^\top k_j)}, \tag{14}$$

$$\delta_i^{\mathrm{route}} = \sum_{a:S_{ai}=1} \pi_{ai} c_a g_a + \ell_i. \tag{15}$$

The local term ℓ_i handles leaf-specific adjustment. Since the descendant weights sum to one, aggregate message $c_a g_a$ is distributed once rather than copied to every leaf. This conservation gives an interpretable redistribution rule and $O(\mathrm{nnz}(S))$ cost for fixed hidden width. Nested aggregates can send different messages to the same leaf, and their sum remains a leaf correction whose image under S is coherent. The bounded gain controls whether a direct aggregate forecast contributes positively or negatively and determines its magnitude. The simplex weights allocate its discrepancy among descendant leaves.

The discrepancy is useful because reconciliation compresses the n direct forecasts to a coherent anchor represented by m leaf coordinates. Conflicts between a direct aggregate forecast and that anchor may still reveal which level is locally reliable. Routing retains this signal without reporting the incoherent direct forecast: a shared gain can suppress, accept, or reverse each aggregate message, and the descendant simplex allocates the accepted amount. In contrast, an unrestricted attention head must learn both hierarchy membership and redistribution from data. Table 1 therefore separates increasing communication capacity from an explicit structural inductive bias.

The head families contain the MinT point. Dense readouts are zero-initialized and routing gains are initialized near zero. Routing is supervised with recursive rolling-origin examples in which the base forecasts and MinT anchors are recomputed using only the history available at each past origin. Chronologically later origins provide early stopping, so the head learns to correct the same forecasting and reconciliation procedure used at the evaluation origins. The metric-aligned variants minimize

$$\mathcal{L}_\lambda = \lambda \frac{\mathrm{gRelMAE}(S b_\theta, y)}{\mathrm{gRelMAE}(S\hat{b}, y)} + (1-\lambda)\frac{\mathrm{WMAPE}(S b_\theta, y)}{\mathrm{WMAPE}(S\hat{b}, y)}, \qquad 0 \le \lambda \le 1. \tag{16}$$

The two terms are normalized by the corresponding MinT loss. The fixed value λ controls the balance between volume-weighted and per-series accuracy. Suffixes .L25 and .L50 denote $\lambda = 0.25$ and 0.50.

After fitting any head, an optional scalar fallback interpolates its correction toward MinT. On a calibration block, let $e_t = S\hat{b}_t - y_t$, $d_t = S\delta_t$, and $\omega_t = \|y_t\|_1^{-1}$. The exact minimizer of $\sum_t \omega_t \|e_t + \gamma d_t\|_1$ over $0 \le \gamma \le 1$ is

$$\gamma^* = \Pi_{[0,1]}\, \mathrm{wmed}\left(\left\{ -\frac{e_{at}}{d_{at}} \right\}, \{\omega_t |d_{at}|\} \right), \tag{17}$$

where terms with $d_{at} = 0$ are omitted. Since $\gamma = 0$ is feasible, the calibrated point $S(\hat{b} + \gamma^* \delta)$ cannot increase this calibration loss relative to MinT. This is an in-sample safeguard, not an out-

Table 2: Main comparison over 29 datasets. Each score is a geometric-mean ratio to the unreconciled Base, with lower values better, followed by its average within-dataset rank. The $\pm$ term is the standard deviation of per-dataset log ratios. Bold marks the column best. Brackets report the number of datasets on which that best method wins and the two-sided paired p against it, with bold $p < 0.05$.

Method	relMAE / Base	relMAE rank	WMAPE / Base	WMAPE rank	nCRPS / Base	nCRPS rank	nMIS / Base	nMIS rank
Base (Ke et al., 2017)	1.000 ± 0.000 (21/29, .180)	7.55 ± 3.10 (21/29, **.001**)	1.000 ± 0.000 (25/29, <**.001**)	7.69 ± 3.97 (25/29, <**.001**)	1.000 ± 0.000 (20/29, .057)	6.41 ± 4.37 (20/29, **.004**)	1.000 ± 0.000 (18/29, .094)	4.76 ± 3.54 (18/29, .068)
Bottom-up (Hyndman et al., 2011)	1.001 ± 0.039 (18/29, .188)	7.00 ± 3.89 (18/29, **.011**)	0.961 ± 0.263 (20/29, **.004**)	6.31 ± 3.47 (20/29, **.042**)	0.971 ± 0.274 (23/29, .108)	6.10 ± 3.29 (23/29, **.003**)	1.000 ± 0.322 (23/29, **.034**)	5.03 ± 2.47 (23/29, **.006**)
WLS-struct (Wickramasuriya et al., 2019)	1.389 ± 0.555 (28/29, <**.001**)	9.76 ± 2.54 (28/29, <**.001**)	0.997 ± 0.094 (25/29, <**.001**)	7.21 ± 2.94 (25/29, <**.001**)	1.013 ± 0.096 (25/29, **.004**)	7.17 ± 3.22 (25/29, <**.001**)	1.063 ± 0.112 (24/29, **.002**)	6.00 ± 2.39 (24/29, **.001**)
WLS-var (Wickramasuriya et al., 2019)	0.989 ± 0.050 (19/29, .308)	5.28 ± 3.00 (19/29, .217)	0.941 ± 0.239 (18/29, **.007**)	5.14 ± 2.79 (18/29, .264)	0.964 ± 0.249 (24/29, .064)	5.93 ± 2.60 (24/29, **.001**)	1.037 ± 0.293 (25/29, **.002**)	6.90 ± 2.57 (25/29, <**.001**)
MinT-shrink (Wickramasuriya et al., 2019)	1.058 ± 0.338 (21/29, **.004**)	6.29 ± 1.88 (21/29, **.002**)	0.957 ± 0.241 (19/29, <**.001**)	6.12 ± 2.13 (19/29, **.032**)	0.987 ± 0.250 (27/29, <**.001**)	7.03 ± 1.97 (27/29, <**.001**)	1.084 ± 0.285 (27/29, <**.001**)	8.31 ± 2.16 (27/29, <**.001**)
Normality-MinT (Panagiotelis et al., 2023)	1.058 ± 0.338 (21/29, **.004**)	6.29 ± 1.88 (21/29, **.002**)	0.957 ± 0.241 (19/29, <**.001**)	6.12 ± 2.13 (19/29, **.032**)	0.987 ± 0.249 (27/29, <**.001**)	7.62 ± 2.13 (27/29, <**.001**)	1.086 ± 0.284 (27/29, <**.001**)	8.90 ± 2.29 (27/29, <**.001**)
Bootstrap-MinT (Panagiotelis et al., 2023)	1.058 ± 0.338 (21/29, **.004**)	6.29 ± 1.88 (21/29, **.002**)	0.957 ± 0.241 (19/29, <**.001**)	6.12 ± 2.13 (19/29, **.032**)	0.991 ± 0.249 (27/29, <**.001**)	7.83 ± 2.05 (27/29, <**.001**)	1.078 ± 0.285 (28/29, <**.001**)	8.00 ± 2.27 (28/29, <**.001**)
Conformal-MinT (Olivares et al., 2022)	1.058 ± 0.338 (21/29, **.004**)	6.29 ± 1.88 (21/29, **.002**)	0.957 ± 0.241 (19/29, <**.001**)	6.12 ± 2.13 (19/29, **.032**)	0.992 ± 0.249 (27/29, <**.001**)	8.28 ± 2.25 (27/29, <**.001**)	1.079 ± 0.284 (28/29, <**.001**)	8.24 ± 2.59 (28/29, <**.001**)
HINT-BU (Olivares et al., 2024)	1.317 ± 1.070 (17/29, **.008**)	6.41 ± 4.90 (17/29, **.032**)	1.872 ± 1.219 (24/29, <**.001**)	9.72 ± 3.48 (24/29, <**.001**)	1.714 ± 1.284 (18/29, **.001**)	7.38 ± 4.72 (18/29, **.003**)	1.717 ± 1.239 (19/29, **.003**)	6.93 ± 4.72 (19/29, **.005**)
HINT-WLS (Olivares et al., 2024)	2.300 ± 1.554 (23/29, <**.001**)	8.10 ± 4.43 (23/29, **.002**)	1.809 ± 1.239 (24/29, <**.001**)	8.83 ± 3.88 (24/29, **.001**)	1.695 ± 1.272 (19/29, **.001**)	7.38 ± 4.69 (19/29, **.002**)	1.746 ± 1.243 (20/29, **.002**)	8.03 ± 4.48 (20/29, <**.001**)
CoFlow-Tweedie-RouteAttn.L50	**0.954 ± 0.182**	**3.97 ± 3.22**	0.892 ± 0.227 (13/29, .154)	4.38 ± 3.59 (13/29, .782)	**0.926 ± 0.250**	**3.00 ± 2.66**	**0.911 ± 0.325**	**2.97 ± 2.81**
CoFlow-RBIG+Gauss-RouteAttn.L25	0.975 ± 0.201 (23/29, <**.001**)	4.76 ± 2.82 (23/29, **.003**)	**0.881 ± 0.231**	**4.24 ± 3.11**	0.962 ± 0.249 (24/29, <**.001**)	3.86 ± 1.66 (24/29, **.006**)	0.976 ± 0.299 (23/29, **.018**)	3.93 ± 1.46 (23/29, **.008**)

of-sample dominance claim. Appendix A.1.5 gives the full architectures and proofs of coherence, conservation, and equation 17.

3.6 Sampling and evaluation

We center K leaf-error draws, add $\hat{b}$, and map through S. Centering preserves ranks, pairwise differences, and interval widths (Appendix A.6). Linear baselines are sampled from $\mathcal{N}(P\hat{y}, PWP^\top)$ using an eigenvalue square root. Every method is scored by the same sample-based CRPS, interval-score, coverage, and energy estimators.

4 Experiments

Setup. We evaluate 72 methods on 29 hierarchies: Australian Labour, Tourism Small, Tourism Large, Traffic, Wiki2, M5, and 23 anonymized proprietary datasets across retail, manufacturing, pharmaceutical, automotive, and other industries. A shared LightGBM forecaster (Ke et al., 2017) supplies the base forecasts for all post-hoc methods. Comparisons include standard reconcilers, Gaussian, bootstrap, and conformal MinT predictives, and two HINT NHITS-GMM configurations. Point metrics are geometric relative MAE, WMAPE, and MASE. Predictive metrics are nCRPS, normalized central-90% interval score, coverage gap, and energy score, using $K = 200$ draws.

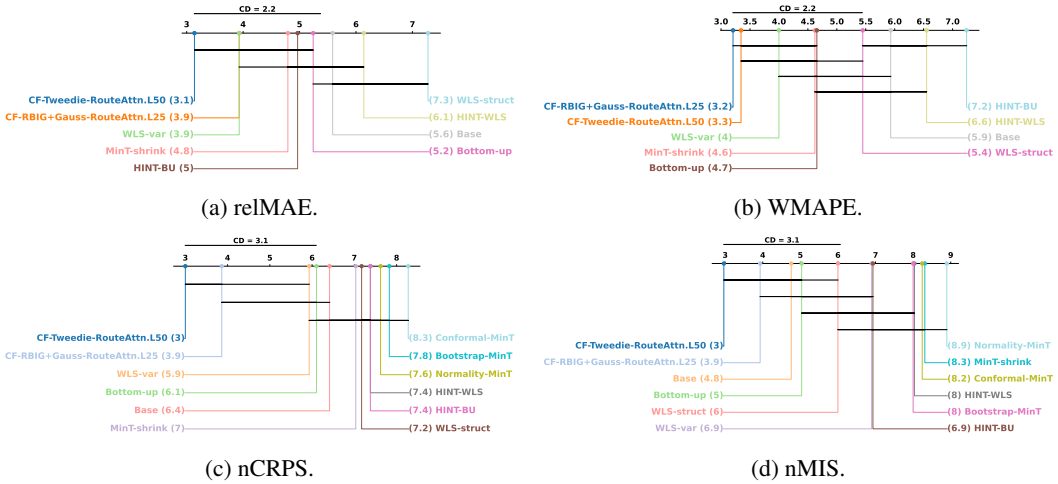
Each dataset contributes three adjacent rolling origins and each stochastic method uses five seeds. Seeds are averaged within origin, leaving 87 repeated observations in 29 independent dataset blocks. Pairwise tests flip all origins of a dataset together, bootstrap intervals resample complete datasets, and Friedman-Nemenyi comparisons use dataset means. Level analysis first averages named levels into Total, Middle, and Bottom within each dataset-origin. Appendix A.2 gives the full protocol, method list, and additional statistical summaries.

4.1 Main comparison

Table 2 reports geometric-mean ratios to the unreconciled base, average within-dataset ranks, cross-dataset dispersion, and paired comparisons with the column leader. Figure 2 gives the corresponding within-dataset rank comparisons.

The benefit of reconciliation depends on the method and metric. Relative to the unreconciled base, MinT-shrink raises relMAE by 5.8% and nMIS by 8.4%, and every other linear reconciler degrades at least one of the four metrics. Both CoFlow configurations improve all four. CoFlow-Tweedie-RouteAttn.L50 leads on relMAE (0.954), nCRPS (0.926), and nMIS (0.911), while CoFlow-RBIG+Gauss-RouteAttn.L25 has the lowest WMAPE (0.881). These gains come from distinct parts of the model. The routing head controls point accuracy, while the leaf law controls the predictive scores. Their composition therefore improves point and predictive accuracy at once rather than trading one for the other.

Against MinT-shrink, the Tweedie-routing composition improves relMAE by 9.7%, WMAPE by 6.8%, nCRPS by 6.0%, and nMIS by 15.8%, winning on 21, 18, 27, and 27 datasets. The Holm-adjusted dataset-clustered sign-flip tests give $p = 0.016$, 0.008, 10^{-4}, and 10^{-4}. The RBIG-Gaussian composition is also significant on WMAPE, nCRPS, and nMIS. Dataset ranks, all-pairs



(a) relMAE. (b) WMAPE.

(c) nCRPS. (d) nMIS.

Figure 2: Average ranks over 29 datasets for the main comparison. Lower is better, and connected methods are not separated by the Nemenyi test at 0.05. Point panels omit probabilistic MinT variants that share the MinT point. CoFlow is abbreviated CF-.

Table 3: Predictive-law ablation over 29 datasets. Entries are geometric-mean ratios to MinT-shrink $\pm$ the standard deviation of per-dataset log ratios, with lower values better. CovGap is the mean signed difference from MinT-shrink, where negative values are closer to nominal coverage. Bold marks the column best, and brackets give the two-sided Wilcoxon p against it.

Family	Method	nCRPS / MinT	nCRPS rank	nMIS / MinT	nMIS rank	CovGap − MinT	CovGap rank
Reconciliation baseline	MinT-shrink	1.000 ± 0.000 (<.001)	16.59 ± 4.83 (<.001)	1.000 ± 0.000 (<.001)	18.62 ± 5.43 (<.001)	0.000 ± 0.000 (<.001)	18.03 ± 4.98 (<.001)
Cross-leaf dependence	CoFlow-Diag-MinT	0.995 ± 0.023 (<.001)	13.55 ± 5.18 (<.001)	0.978 ± 0.104 (<.001)	15.52 ± 5.03 (<.001)	-0.047 ± 0.028 (<.001)	8.88 ± 3.50 (.045)
Cross-leaf dependence	CoFlow-Gauss-MinT	0.983 ± 0.013 (.005)	8.10 ± 4.31 (<.001)	0.925 ± 0.048 (<.001)	8.38 ± 3.64 (.002)	-0.047 ± 0.028 (<.001)	8.62 ± 4.40 (.066)
Cross-leaf dependence	CoFlow-Gauss.reconres-MinT	0.994 ± 0.011 (<.001)	14.28 ± 4.19 (<.001)	0.975 ± 0.052 (<.001)	16.07 ± 4.57 (<.001)	-0.016 ± 0.031 (<.001)	16.14 ± 5.91 (<.001)
Non-Gaussian marginals	CoFlow-Cop-MinT	0.995 ± 0.011 (<.001)	15.22 ± 4.13 (<.001)	0.957 ± 0.059 (<.001)	12.71 ± 4.58 (<.001)	-0.000 ± 0.061 (<.001)	17.34 ± 4.84 (<.001)
Non-Gaussian marginals	CoFlow-tCop-MinT	0.996 ± 0.012 (<.001)	15.31 ± 4.80 (<.001)	0.960 ± 0.065 (<.001)	14.48 ± 5.16 (<.001)	0.001 ± 0.062 (<.001)	18.48 ± 4.83 (<.001)
Non-Gaussian marginals	CoFlow-tMarg-MinT	0.986 ± 0.015 (.004)	9.48 ± 4.44 (<.001)	0.932 ± 0.053 (<.001)	10.24 ± 4.42 (<.001)	-0.036 ± 0.028 (<.001)	12.90 ± 3.85 (<.001)
Non-Gaussian marginals	CoFlow-Gate-MinT	0.995 ± 0.012 (<.001)	14.09 ± 4.02 (<.001)	0.958 ± 0.058 (<.001)	13.67 ± 4.44 (<.001)	0.000 ± 0.061 (<.001)	17.97 ± 4.71 (<.001)
Non-Gaussian marginals	CoFlow-Emp-MinT	0.996 ± 0.014 (<.001)	15.24 ± 5.80 (<.001)	0.944 ± 0.066 (<.001)	11.07 ± 5.24 (<.001)	-0.001 ± 0.062 (<.001)	16.90 ± 5.23 (<.001)
Count marginals	CoFlow-Tweedie-MinT	0.938 ± 0.074 (.221)	**4.34 ± 4.89**	0.840 ± 0.143 (.004)	5.00 ± 5.61 (.081)	-0.016 ± 0.091 (<.001)	15.24 ± 7.67 (<.001)
Count marginals	CoFlow-ZINB-MinT	**0.935 ± 0.139**	6.38 ± 8.41 (.946)	**0.745 ± 0.315**	**4.24 ± 6.53**	-0.053 ± 0.113 (<.001)	9.90 ± 8.88 (.004)
Normalising flows	CoFlow-RBIG-MinT	0.987 ± 0.025 (<.001)	10.31 ± 5.25 (<.001)	0.914 ± 0.131 (<.001)	7.72 ± 5.35 (.002)	-0.077 ± 0.118 (.011)	6.53 ± 5.52 (.135)
Normalising flows	CoFlow-NVP-MinT	0.993 ± 0.015 (<.001)	14.24 ± 5.07 (<.001)	0.966 ± 0.064 (<.001)	16.17 ± 5.36 (<.001)	-0.039 ± 0.029 (<.001)	11.52 ± 4.05 (.006)
Normalising flows	CoFlow-NVP.energy-MinT	1.004 ± 0.023 (<.001)	17.72 ± 4.90 (<.001)	1.011 ± 0.103 (<.001)	20.52 ± 4.05 (<.001)	-0.019 ± 0.030 (<.001)	16.19 ± 4.34 (<.001)
Hedged flows	CoFlow-RBIG+Gauss-MinT	0.975 ± 0.016 (.010)	5.14 ± 2.08 (.010)	0.901 ± 0.079 (<.001)	6.24 ± 2.34 (.004)	-0.073 ± 0.075 (.006)	6.52 ± 3.39 (.205)
Hedged flows	CoFlow-NVP+Gauss-MinT	0.987 ± 0.011 (.002)	10.45 ± 4.08 (<.001)	0.939 ± 0.051 (<.001)	11.86 ± 4.36 (<.001)	-0.044 ± 0.027 (<.001)	10.12 ± 3.80 (.013)
Hedged flows	CoFlow-RBIG+NVP-MinT	0.981 ± 0.019 (.003)	8.79 ± 4.03 (<.001)	0.924 ± 0.089 (<.001)	10.59 ± 4.52 (.001)	-0.070 ± 0.075 (.002)	7.45 ± 3.01 (.063)
Bottom-up centering	CoFlow-Gauss-BU	0.985 ± 0.114 (.029)	12.03 ± 7.93 (.001)	0.924 ± 0.151 (.001)	10.31 ± 6.29 (.002)	-0.063 ± 0.100 (<.001)	9.79 ± 5.15 (.008)
Bottom-up centering	CoFlow-RBIG-BU	0.992 ± 0.126 (.014)	12.97 ± 8.74 (<.001)	0.920 ± 0.246 (.002)	9.00 ± 7.69 (.001)	-0.090 ± 0.156 (.009)	8.59 ± 7.84 (.041)
Bottom-up centering	CoFlow-NVP-BU	0.996 ± 0.113 (.014)	15.72 ± 8.25 (<.001)	0.970 ± 0.148 (<.001)	16.38 ± 6.99 (<.001)	-0.055 ± 0.099 (<.001)	12.09 ± 5.30 (.002)
External predictive	Normality-MinT	1.001 ± 0.001 (<.001)	17.48 ± 5.02 (<.001)	1.001 ± 0.002 (<.001)	19.28 ± 5.46 (<.001)	0.001 ± 0.001 (<.001)	18.78 ± 4.86 (<.001)
External predictive	Bootstrap-MinT	1.004 ± 0.013 (<.001)	18.72 ± 4.37 (<.001)	0.995 ± 0.063 (<.001)	17.97 ± 4.36 (<.001)	0.017 ± 0.044 (<.001)	21.62 ± 2.87 (<.001)
External predictive	Conformal-MinT	1.005 ± 0.013 (<.001)	19.10 ± 4.78 (<.001)	0.995 ± 0.064 (<.001)	18.52 ± 4.62 (<.001)	0.019 ± 0.044 (<.001)	22.59 ± 2.26 (<.001)
External predictive	HINT-BU	1.737 ± 1.184 (.016)	14.72 ± 11.06 (.003)	1.584 ± 1.134 (.006)	14.17 ± 10.30 (<.001)	**-0.146 ± 0.168**	**5.52 ± 7.04**
External predictive	HINT-WLS	1.717 ± 1.169 (.016)	15.00 ± 11.00 (.002)	1.610 ± 1.140 (.006)	16.28 ± 9.82 (<.001)	-0.125 ± 0.162 (.006)	7.31 ± 7.19 (.006)

tests, bootstrap intervals, and the distinction between control and all-pairs comparisons are reported in Appendix A.3.

4.2 Factor ablations

Tables 3 and 4 change one factor at a time. Full Gaussian dependence lowers nMIS by 7.5% relative to MinT, compared with 2.2% for independent leaves. Using residuals after MinT projection retains only a 2.5% gain, consistent with the covariance reduction in equation 5. Marginal shape adds a separate benefit. Tweedie lowers nCRPS by 6.2% and nMIS by 16.0%, while the RBIG-Gaussian mixture gives the largest and most consistent coverage improvement.

The point-head ablation shows a corresponding separation. Balanced leaf attention concentrates its gains in relMAE and MASE, whereas routing primarily improves volume-weighted accuracy. RouteAttn.L0 lowers WMAPE by 8.5%, RouteAttn.L50 lowers relMAE by 10.1%, and RouteAttn.L25 lies between them. Complete predictive and point ablations, including architecture, loss, anchor, residual-source, and flow studies, appear in Appendices A.8 and A.9. Appendix A.4 provides detailed mechanism interpretation, model-selection guidance, calibration options, and scope.

Table 4: Point-head ablation over 29 datasets. Entries are geometric-mean ratios to MinT-shrink ± the standard deviation of per-dataset log ratios, with lower values better. Bold marks the column best, and brackets give the two-sided Wilcoxon p against it.

Family	Method	relMAE / MinT	relMAE rank	WMAPE / MinT	WMAPE rank	MASE / MinT	MASE rank
Reconciliation baseline	MinT-shrink	1.000 ± 0.000 (<.001)	12.07 ± 4.57 (<.001)	1.000 ± 0.000 (.004)	9.83 ± 4.46 (.034)	1.000 ± 0.000 (<.001)	11.45 ± 4.37 (<.001)
Per-leaf head	CoFlow-Gauss-MLP	0.985 ± 0.044 (<.001)	11.14 ± 4.64 (<.001)	0.994 ± 0.021 (.005)	8.93 ± 4.53 (.150)	0.995 ± 0.018 (<.001)	11.48 ± 3.88 (<.001)
Cross-leaf attention	CoFlow-Gauss-LeafAttn	0.967 ± 0.070 (<.001)	9.93 ± 3.68 (<.001)	0.996 ± 0.034 (.003)	10.14 ± 4.79 (.049)	0.983 ± 0.046 (.002)	10.48 ± 4.39 (.003)
Cross-leaf attention	CoFlow-Gauss-LeafAttn.L25	0.952 ± 0.081 (.009)	7.10 ± 3.77 (.042)	0.996 ± 0.031 (.003)	10.14 ± 4.24 (.033)	0.963 ± 0.067 (.026)	7.41 ± 3.79 (.138)
Cross-leaf attention	CoFlow-Gauss-LeafAttn.L50	0.934 ± 0.096 (.325)	5.17 ± 3.70 (.392)	0.992 ± 0.042 (.003)	9.55 ± 4.35 (.030)	0.938 ± 0.118 (.550)	5.97 ± 4.33 (.392)
Cross-level attention	CoFlow-Gauss-FullAttn	0.975 ± 0.052 (<.001)	10.34 ± 2.86 (<.001)	0.995 ± 0.019 (.007)	8.62 ± 3.18 (.138)	0.983 ± 0.040 (.001)	10.76 ± 2.61 (<.001)
Cross-level attention	CoFlow-Gauss-TreeAttn	0.974 ± 0.057 (<.001)	10.62 ± 2.51 (<.001)	0.994 ± 0.033 (.003)	9.48 ± 3.55 (.038)	0.983 ± 0.043 (.002)	10.41 ± 2.67 (<.001)
Cross-level attention	CoFlow-Gauss-LeafTreeAttn	0.974 ± 0.052 (<.001)	10.10 ± 2.64 (<.001)	0.994 ± 0.022 (.006)	8.69 ± 3.12 (.124)	0.983 ± 0.039 (.001)	10.34 ± 2.48 (<.001)
Cross-level attention	CoFlow-Gauss-FullAttn.L25	0.965 ± 0.066 (<.001)	8.69 ± 2.78 (<.001)	0.994 ± 0.018 (.006)	8.07 ± 3.23 (.197)	0.974 ± 0.050 (.005)	8.10 ± 2.81 (.052)
Cross-level attention	CoFlow-Gauss-FullAttn.L50	0.955 ± 0.074 (.015)	6.48 ± 3.03 (.035)	0.995 ± 0.022 (.005)	8.45 ± 3.47 (.201)	0.960 ± 0.064 (.121)	6.17 ± 2.93 (.238)
Constraint routing	CoFlow-Gauss-RouteAttn.L0	0.973 ± 0.102 (<.001)	10.45 ± 4.88 (<.001)	**0.915 ± 0.143**	6.93 ± 5.15 (.246)	0.964 ± 0.152 (.004)	9.83 ± 5.27 (.001)
Constraint routing	CoFlow-Gauss-RouteAttn.L25	0.922 ± 0.225 (<.001)	8.00 ± 4.52 (<.001)	0.920 ± 0.137 (.247)	**6.38 ± 5.00**	0.944 ± 0.105 (.051)	7.17 ± 4.72 (.016)
Constraint routing	CoFlow-Gauss-RouteAttn.L50	0.899 ± 0.285 (<.001)	6.41 ± 4.80 (<.001)	0.932 ± 0.114 (.831)	6.90 ± 5.53 (.757)	**0.930 ± 0.136**	5.83 ± 4.50 (.038)
Constraint routing	CoFlow-Gauss-RouteAttn.L75	**0.899 ± 0.248**	**4.38 ± 3.68**	0.949 ± 0.088 (.670)	7.31 ± 4.80 (.292)	0.931 ± 0.140 (.121)	**5.17 ± 4.46**
Constraint routing	CoFlow-Gauss-RouteAttn.L100	0.906 ± 0.192 (.247)	4.45 ± 3.84 (.875)	0.968 ± 0.085 (.230)	8.62 ± 5.52 (.089)	0.933 ± 0.149 (.314)	5.21 ± 4.51 (.798)
Rolling supervision	CoFlow-Gauss-LeafAttn.ro	0.987 ± 0.107 (<.001)	10.66 ± 5.92 (<.001)	0.958 ± 0.129 (.275)	7.97 ± 6.30 (.404)	0.973 ± 0.091 (<.001)	10.21 ± 5.75 (<.001)

5 DISCUSSION

What the ablations identify. The factorization exposes three different failure modes. First, dependence sets aggregate scale: replacing the full Gaussian covariance by a diagonal one changes nMIS from 0.925 to 0.978 while preserving the fitted leaf variances. The gap is largest above the bottom level, consistent with equation 6 and Corollary 1. This comparison isolates one mechanism but does not imply that every empirical score difference is caused by one scalar variance ratio. Second, marginal shape remains visible at the leaves. Tweedie's atom at zero and positive tail retain a large Bottom-level gain, while the RBIG-Gaussian mixture is the most stable coverage hedge across datasets. Third, the point forecast is optimized separately. Leaf attention primarily improves per-series metrics. Routing converts disagreement at high-volume aggregates into descendant corrections and therefore improves WMAPE. Varying λ traces the expected tradeoff without refitting the predictive law. The composition results support the architectural claim that a point head and a leaf law can be selected independently and their gains retained together.

Model-selection implications. A full Gaussian leaf law provides a low-complexity diagnostic because it preserves estimated cross-leaf covariance. Comparing it with the diagonal Gaussian tests whether dependence is material, especially at Total and Middle levels. Marginal flexibility should then be chosen from rolling proper scores rather than likelihood alone. Tweedie is a natural candidate for nonnegative intermittent demand, empirical or copula marginals for irregular continuous residuals, and the RBIG-Gaussian mixture when no single residual regime dominates. Flow capacity is warranted only when the available residual history supports it. Point selection can proceed separately. Leaf attention emphasizes equal treatment of series, whereas routing emphasizes aggregate volume accuracy. The fixed λ should be chosen before the final test according to the stated evaluation objective. No single weight is claimed to be uniformly best.

Calibration without losing coherence. The leaf law must represent error at the conditional scale left by the base forecaster. Raw base residuals provide a common, comparatively stable source in our study. Fitting after MinT projection changes Gaussian nMIS from 0.925 to 0.975, because the projected residuals already contain the conditional-covariance reduction in equation 5. This does not make raw residuals universally correct. With enough origins, cross-fitted residuals from the chosen anchor directly target anchor-specific calibration. Remaining dispersion error should be adjusted jointly in leaf space, or by a shared transformation of coherent draws followed by S. Independent nodewise quantile corrections generally leave $\mathrm{range}(S)$ and can destroy sample coherence. Coverage should also be interpreted beside CRPS or interval score: wider intervals can improve coverage without improving the predictive distribution.

Computational scaling. Dense covariance estimation and attention are at least quadratic in the number of leaves or nodes, whereas routing follows the nonzero edges of S. Wide hierarchies therefore motivate structured covariance models and sparse point architectures.

Scope and limitations. The dispersion theorem is exact for centered Gaussian node errors. The covariance corollary further assumes nonnegative attenuated cross-covariances, and its monotone

support-size statement requires an exchangeable special case. These results diagnose one mechanism, not skewness, bias, negative dependence, or arbitrary misspecification. Atomic count marginals are coupled by ranks because the copula is not identified inside tied zero blocks. Finite-sample centering guarantees the chosen anchor mean and coherence but may move individual count-law draws off their original support. The point fallback guarantees only its calibration objective, and learned heads require enough historical origins to estimate corrections reliably. Finally, the corpus has only three test origins per dataset, includes 23 non-releasable industry hierarchies, uses one shared base-learner family, and emphasizes retail and intermittent demand. Dataset-clustered inference avoids treating origins and seeds as independent, but broader public corpora, base forecasters, dynamic hierarchies, and very wide sparse systems remain important tests. Appendix A.4 expands these qualifications and provides additional model-selection analysis.

6 Related work

Forecast reconciliation projects incoherent base forecasts onto the coherent subspace, with MinT as the trace-optimal linear map (Hyndman et al., 2011; Wickramasuriya et al., 2019; Athanasopoulos et al., 2024). Probabilistic reconciliation studies score optimization and uncertainty quantification (Panagiotelis et al., 2023; Wickramasuriya, 2024; Møller et al., 2024). Conditioning-based methods construct a coherent law from an all-node base predictive (Corani et al., 2020; Zambon et al., 2024a;b). Other work treats the reconciliation covariance as uncertain, placing an inverse-Wishart prior on a static W to obtain a multivariate-t predictive (Carrara et al., 2025) or guarding W by robust optimization (Aikawa et al., 2025).

Joint bottom-up methods train the bottom series and their dependence jointly before aggregation (Bertani et al., 2025). End-to-end neural methods learn coherent forecasts through projection layers, reconciliation-consistent training, or bottom-up sampling (Rangapuram et al., 2021; 2023; Olivares et al., 2024; Das et al., 2023). CoFlow occupies a different operating point. It is post-hoc and base-forecaster agnostic, models the error law around a coherent anchor rather than retraining observation dynamics, and exposes a separate coherent point-forecast branch. This factorization supports a broad family of continuous, count, empirical, and flow laws under one scorer, while the dispersion-regret theorem quantifies when a variance error becomes a proper-score cost at an aggregate. The leaf models draw on normalizing flows (Laparra et al., 2011; Dinh et al., 2017; Rezende & Mohamed, 2015; Papamakarios et al., 2021), the point heads on attention (Vaswani et al., 2017), and the evaluation on proper scoring rules (Gneiting & Raftery, 2007; Gneiting et al., 2008).

7 Conclusion

We presented CoFlow, which reframes coherent forecasting as two distinct tasks around a shared anchor: modeling a joint predictive law and optimizing a point forecast. Both are performed in leaf coordinates and mapped through S, making coherence a property of the representation rather than of a particular distribution or point estimator. Our Gaussian analysis quantifies how nodewise variance error translates into coverage and proper-score regret, and when cross-leaf covariance attenuation amplifies these costs at aggregates. The experiments show the same separation: dependence and marginal shape improve predictive scores, while attention and constraint routing improve point accuracy under different objectives. Across 29 hierarchies, Tweedie with routing improves relative MAE, nCRPS, and nMIS over MinT by 9.8%, 6.2%, and 16.0%. The RBIG-Gaussian mixture improves nCRPS on all 29 datasets and both nMIS and coverage gap on 28. Together, these results support leaf-space factorization as a flexible alternative to coupling coherence, uncertainty, and point estimation in a single reconciliation map. Promising extensions include cross-fitted residual laws, conformal dispersion calibration, and structured covariance for wide hierarchies.

AI use statement

In this work, we used generative AI tools for language editing, coding, and experiment orchestration. We have not used generative AI tools for literature discovery and idea generation. The authors developed and verified the ideas, theory, experimental design, and analysis. All the LLM-generated code and assisted content was verified and tested for correctness by the authors. We take responsi-

bility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Ethics statement

The data are public benchmarks or anonymized aggregate series, with no individual records or sensitive attributes. Forecasts used for planning and inventory decisions require domain oversight.

Reproducibility statement

Six datasets are public benchmarks and 23 are anonymized proprietary hierarchies that cannot be distributed. The public evaluation is reproducible from released code and data. All post-hoc methods share one LightGBM base forecaster and one sample-based scorer. Results use three rolling origins and five seeds. Inference averages seeds, clusters origins by dataset, and balances level classes as specified in Section 4. Sections 3-3.6 define the models and evaluation. The appendix provides proofs, complete method specifications, and the experimental protocol.

References

Shuhei Aikawa, Aru Suzuki, Kei Yoshitake, Kanata Teshigawara, Akira Iwabuchi, Ken Kobayashi, and Kazuhide Nakata. Hierarchical time series forecasting with robust reconciliation. *arXiv preprint arXiv:2510.20383*, 2025.

George Athanasopoulos, Rob J Hyndman, Nikolaos Kourentzes, and Fotios Petropoulos. Forecasting with temporal hierarchies. *European Journal of Operational Research*, 262(1):60–74, 2017.

George Athanasopoulos, Rob J Hyndman, Nikolaos Kourentzes, and Anastasios Panagiotelis. Forecast reconciliation: A review. *International Journal of Forecasting*, 40(2):430–456, 2024.

Nicolò Bertani, Shane T Jensen, and Ville A Satopää. Joint bottom-up method for probabilistic forecasting of hierarchical time series. *Operations Research*, 73(6):3260–3277, 2025.

Chiara Carrara, Dario Azzimonti, Giorgio Corani, and Lorenzo Zambon. Modeling the uncertainty on the covariance matrix for probabilistic forecast reconciliation. *arXiv preprint arXiv:2506.19554*, 2025.

Giorgio Corani, Dario Azzimonti, Joao PSC Augusto, and Marco Zaffalon. Probabilistic reconciliation of hierarchical forecast via bayes' rule. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 211–226. Springer, 2020.

Abhimanyu Das, Weihao Kong, Biswajit Paria, and Rajat Sen. Dirichlet proportions model for hierarchically coherent probabilistic forecasting. pp. 518–528, 2023.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=HkpbnH9lx.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Tilmann Gneiting, Larissa I Stanberry, Eric P Grimit, Leonhard Held, and Nicholas A Johnson. Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *Test*, 17(2):211–235, 2008.

Rob J Hyndman, Roman A Ahmed, George Athanasopoulos, and Han Lin Shang. Optimal combination forecasts for hierarchical time series. *Computational statistics & data analysis*, 55(9): 2579–2589, 2011.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. volume 30, 2017.

Valero Laparra, Gustavo Camps-Valls, and Jesús Malo. Iterative gaussianization: from ica to random rotations. *IEEE transactions on neural networks*, 22(4):537–549, 2011.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411, 2004.

Jan Kloppenborg Møller, Peter Nystrup, Poul G Hjorth, and Henrik Madsen. Optimal forecast reconciliation with uncertainty quantification. *arXiv preprint arXiv:2402.06480*, 2024.

Kin G Olivares, Azul Garza, David Luo, Cristian Challú, Max Mergenthaler, Souhaib Ben Taieb, Shanika L Wickramasuriya, and Artur Dubrawski. Hierarchicalforecast: A reference framework for hierarchical forecasting in python. *arXiv preprint arXiv:2207.03517*, 2022.

Kin G Olivares, O Nganba Meetei, Ruijun Ma, Rohan Reddy, Mengfei Cao, and Lee Dicker. Probabilistic hierarchical forecasting with deep poisson mixtures. *International Journal of Forecasting*, 40(2):470–489, 2024.

Anastasios Panagiotelis, Puwasala Gamakumara, George Athanasopoulos, and Rob J Hyndman. Probabilistic forecast reconciliation: Properties, evaluation and score optimisation. *European Journal of Operational Research*, 306(2):693–706, 2023.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

Syama Sundar Rangapuram, Lucien D Werner, Konstantinos Benidis, Pedro Mercado, Jan Gasthaus, and Tim Januschowski. End-to-end learning of coherent probabilistic forecasts for hierarchical time series. In *International Conference on Machine Learning*, pp. 8832–8843. PMLR, 2021.

Syama Sundar Rangapuram, Shubham Kapoor, Rajbir Singh Nirwan, Pedro Mercado, Tim Januschowski, Yuyang Wang, and Michael Bohlke-Schneider. Coherent probabilistic forecasting of temporal hierarchies. In *International Conference on Artificial Intelligence and Statistics*, pp. 9362–9376. PMLR, 2023.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pp. 1530–1538. PMLR, 2015.

Juliane Schäfer and Korbinian Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, 4(1), 2005.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Xiaoqian Wang, Rob J Hyndman, Feng Li, and Yanfei Kang. Forecast combinations: An over 50-year review. *International Journal of Forecasting*, 39(4):1518–1547, 2023.

Shanika L Wickramasuriya. Probabilistic forecast reconciliation under the gaussian framework. *Journal of Business & Economic Statistics*, 42(1):272–285, 2024.

Shanika L Wickramasuriya, George Athanasopoulos, and Rob J Hyndman. Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. *Journal of the American Statistical Association*, 114(526):804–819, 2019.

Lorenzo Zambon, Arianna Agosto, Paolo Giudici, and Giorgio Corani. Properties of the reconciled distributions for gaussian and count forecasts. *International Journal of Forecasting*, 40(4):1438–1448, 2024a.

Lorenzo Zambon, Dario Azzimonti, and Giorgio Corani. Efficient probabilistic reconciliation of forecasts for real-valued and count time series: L. zambon et al. *Statistics and Computing*, 34(1): 21, 2024b.

A Appendix

A.1 Extended method details and proofs

A.1.1 Leaf coordinates of coherent laws

The following result formalizes the coordinate statement used in Section 3.2. Related bottom-level formulations appear in Bertani et al. (2025).

Lemma 1 (Leaf coordinates of coherent laws)**.** *Let S have full column rank and let $S^+ = (S^\top S)^{-1}S^\top$. The map $b \mapsto Sb$ is a linear isomorphism from $\mathbb{R}^m$ onto $\mathcal{C}$, with inverse S^+ on $\mathcal{C}$. Consequently, $\nu \mapsto S_\#\nu$ is a bijection between probability laws on $\mathbb{R}^m$ and coherent probability laws on $\mathbb{R}^n$, with inverse $\mu \mapsto S^+_\#\mu$. Every draw from $S_\#\nu$ is coherent. If ν has a finite first moment, then its pushforward has mean $S\mathbb{E}_\nu[b]$. Weak convergence of leaf laws implies weak convergence of their pushforwards.*

Proof. If $\mu = S_\#\nu$, then μ is supported on $\mathrm{range}(S) = \mathcal{C}$. Conversely, let μ be coherent. The restriction $S : \mathbb{R}^m \to \mathcal{C}$ has inverse S^+, and $S^+S = I_m$. Thus a coherent draw $y \sim \mu$ satisfies $y = S(S^+y)$ almost surely, which gives $S_\#(S^+_\#\mu) = \mu$. Linearity gives the expectation identity whenever the first moment exists. Continuity of S and the continuous mapping theorem give the weak-convergence statement. □

A.1.2 Gaussian MinT decomposition

Let $C_W = (S^\top W^{-1}S)^{-1}$, $R_W = HWH^\top$, $B_W = JWH^\top$, and $W_{bb} = JWJ^\top$. Conditioning $Y \sim \mathcal{N}(\hat{y}, W)$ on $HY = 0$ gives

$$Y \mid HY = 0 \sim \mathcal{N}\big(P_W\hat{y}, W - WH^\top R_W^{-1}HW\big) = \mathcal{N}\big(P_W\hat{y}, P_W W P_W^\top\big)\,. \tag{18}$$

Together with equation 5, this yields

$$\mathcal{N}(P_W\hat{y}, P_W W P_W^\top) = S_\#\mathcal{N}(JP_W\hat{y}, C_W), \tag{19}$$

$$s_a^\top W_{bb}s_a - s_a^\top C_W s_a = \left\| R_W^{-1/2}B_W^\top s_a \right\|_2^2. \tag{20}$$

The covariance reduction $D_W = B_W R_W^{-1}B_W^\top$ is positive semidefinite and has rank at most $n - m$. If a direct leaf model uses covariance $\widehat{\Sigma}_b$, then

$$s_a^\top \widehat{\Sigma}_b s_a - s_a^\top C_W s_a = s_a^\top(\widehat{\Sigma}_b - W_{bb})s_a + \left\| R_W^{-1/2}B_W^\top s_a \right\|_2^2. \tag{21}$$

This decomposition compares two fitted covariances. If the conditional MinT-point error is exactly $P_W e$ with $e \sim \mathcal{N}(0, W)$ and known W, then $\mathcal{N}(0, P_W W P_W^\top)$ is calibrated and optimal under every strictly proper score.

A.1.3 Proofs for Gaussian dispersion regret

Proof of Theorem 1. Standardizing the interval $[-z\sqrt{\widehat{v}}, z\sqrt{\widehat{v}}]$ by $\sqrt{v}$ gives equation 7. For independent $X, X' \sim F$ and $Y \sim \mathcal{N}(0, v)$,

$$\mathbb{E}|X - Y| - \frac{1}{2}\mathbb{E}|X - X'| = \frac{\sqrt{v}}{\sqrt{\pi}}\left\{\sqrt{2(1+r)} - \sqrt{r}\right\}.$$

Subtracting its value $\sqrt{v/\pi}$ at $r = 1$ proves equation 8. Its derivative with respect to r is

$$\frac{1}{\sqrt{2(1+r)}} - \frac{1}{2\sqrt{r}},$$

which is negative for $r < 1$, zero at $r = 1$, and positive for $r > 1$. Put $t = \sqrt{r}$. The normalized regret is $f(t) = \{\sqrt{2(1+t^2)} - t - 1\}/\sqrt{\pi}$, where $f(1) = f'(1) = 0$ and the second derivative before division by $\sqrt{\pi}$ is $\sqrt{2}(1+t^2)^{-3/2} \geq 1/2$ on $0 \leq t \leq 1$. This proves the CRPS lower bound.

For interval score, put $c = z\sqrt{r}$ and standardize the truth to $Z \sim \mathcal{N}(0,1)$. Symmetry and $\mathbb{E}[(Z - c)\mathbf{1}\{Z > c\}] = \phi(c) - c[1 - \Phi(c)]$ give expected score $\sqrt{v}G_\alpha(c)$. Direct differentiation gives

$$G'_\alpha(c) = 2 - \frac{4}{\alpha}[1 - \Phi(c)], \qquad G''_\alpha(c) = \frac{4}{\alpha}\phi(c) > 0.$$

The unique minimizer is $c = z$. On $0 \le c \le z$, $G''_\alpha(c) \ge 4\phi(z)/\alpha$, so strong convexity gives the interval-score lower bound in equation 10. Monotonicity follows because G_α decreases on $[0, z]$. □

Proof of Corollary 1. Write

$$d_a = s_a^\top \operatorname{diag}(\Sigma_b) s_a, \qquad c_a = s_a^\top \{\Sigma_b - \operatorname{diag}(\Sigma_b)\} s_a,$$

and, when $c_a > 0$,

$$\bar{\kappa}_a = \frac{\sum_{i \ne j \in L_a} \kappa_{ij} \Sigma_{b,ij}}{c_a}, \qquad \theta_a = \frac{c_a}{d_a + c_a}.$$

Separating diagonal and off-diagonal terms gives $v_a = d_a + c_a$, $\widehat{v}_a = d_a + \bar{\kappa}_a c_a$, and $r_a = 1 - (1 - \bar{\kappa}_a)\theta_a$. The monotonicity claims follow from Theorem 1. In the exchangeable case, $d_a = q\sigma^2$ and $c_a = q(q-1)\rho\sigma^2$. Hence equation 13 follows, and

$$\frac{d}{dx}\frac{1 + \kappa\rho x}{1 + \rho x} = \frac{\rho(\kappa - 1)}{(1 + \rho x)^2} < 0, \qquad x = q - 1.$$

□

The covariance share θ_a, rather than depth alone, controls the general result. Heterogeneous leaf variances can prevent θ_a from increasing with support size. The explicit monotonic law in q therefore belongs to the exchangeable case.

A.1.4 Leaf-law specifications

All continuous laws are fitted to centered raw leaf residuals. *Gaussian* uses their shrinkage covariance, while *Diagonal* retains the marginal variances and removes cross-leaf covariance. *Copula* combines Gaussian dependence on normal scores with empirical marginals. *t-copula* changes the score dependence to Student-t, and *Student-t marginals* use Gaussian-copula dependence with degrees of freedom chosen per leaf by kurtosis matching. *Gate* uses Gaussian marginals for low-kurtosis leaves and empirical marginals otherwise. *Empirical* resamples complete historical residual vectors.

For intermittent nonnegative leaves, *Tweedie* uses compound Poisson-Gamma marginals and *ZINB* uses zero-inflated negative-binomial marginals. Count dispersion is matched to residual variation along the fitted mean. We draw correlated Gaussian scores, sample each count marginal, and reorder the marginal draws according to the score ranks. This preserves every finite-sample marginal and remains defined when many samples are tied at zero.

The gradient-free *RBIG* flow (Laparra et al., 2011) alternates marginal Gaussianization and random orthogonal rotation. The gradient-trained *RealNVP* flow (Dinh et al., 2017) composes affine coupling layers

$$y_{\text{free}} = x_{\text{free}} \odot \exp\{s(x_{\text{mask}})\} + t(x_{\text{mask}}), \qquad y_{\text{mask}} = x_{\text{mask}}. \tag{22}$$

It is trained by likelihood or by the leaf-space energy score. Fixed and validation-weighted mixtures draw prescribed fractions from two component laws. The suffix BU changes the anchor, RECONRES changes the residual source, and NVP.ENERGY changes the flow objective.

A.1.5 Point architectures and scalar fallback

The per-leaf head shares an MLP across leaves. LEAFATTN applies self-attention over leaf tokens. Cross-level variants add aggregate tokens carrying direct forecasts, residual scales, series sizes, and level embeddings. FULLATTN allows every node pair to communicate, TREEATTN keeps lineage and sibling links, and LEAFTREEATTN adds a complete leaf subgraph to the lineage graph. All

variants read out one correction per leaf and are trained through S. Rolling-origin variants recompute base forecasts and MinT anchors at historical origins.

For routing, define $g_a = \hat{y}_a - s_a^\top \hat{b}$. Aggregate and leaf encoders produce the queries, keys, gains, and local terms in equation 14 and equation 15.

Lemma 2 (Coherent heads and conservative routing)**.** *For every leaf readout $\delta_\theta(f)$, the point $S\{\hat{b} + \delta_\theta(f)\}$ is coherent. If $\delta_0 = 0$, the parameterization contains and starts at $S\hat{b}$. For routed message $m_{ai} = \pi_{ai} c_a g_a$,*

$$\sum_{i:S_{ai}=1} m_{ai} = c_a g_a. \tag{23}$$

For fixed hidden dimensions, routing costs $O(\mathrm{nnz}(S))$.

Proof. Every output belongs to $\mathrm{range}(S)$, and the initialization claim follows from $\delta_0 = 0$. Message conservation follows from $\sum_{i:S_{ai}=1} \pi_{ai} = 1$. The sparse computation performs a constant number of operations per nonzero edge of S. □

The basic all-node training objective is

$$\min_\theta \sum_t \frac{\|S\{\hat{b}_t + \delta_\theta(f_t)\} - y_t\|_1}{\|y_t\|_1}. \tag{24}$$

Residual history supplies uncertainty, scale, and bias features. Cross-level and routing heads also receive direct aggregate forecasts, support size, hierarchy level, and horizon context. Held-out blocks provide early stopping.

Proposition 1 (Exact scalar fallback)**.** *Fix calibration cases with $\|y_t\|_1 > 0$ and learned leaf corrections δ_t. Define $e_t = S\hat{b}_t - y_t$, $d_t = S\delta_t$, and $\omega_t = \|y_t\|_1^{-1}$. Along*

$$b_{\gamma,t} = \hat{b}_t + \gamma\delta_t, \qquad \mathcal{L}(\gamma) = \sum_t \omega_t \|e_t + \gamma d_t\|_1, \qquad 0 \le \gamma \le 1, \tag{25}$$

let $r_{at} = -e_{at}/d_{at}$ and $u_{at} = \omega_t |d_{at}|$ whenever $d_{at} \neq 0$. If γ_0 is any weighted median of the r_{at} with weights u_{at}, then

$$\gamma^* = \Pi_{[0,1]}(\gamma_0) \tag{26}$$

globally minimizes equation 25. If all $d_{at} = 0$, take $\gamma^* = 0$. In every case,

$$\mathcal{L}(\gamma^*) \le \mathcal{L}(0), \tag{27}$$

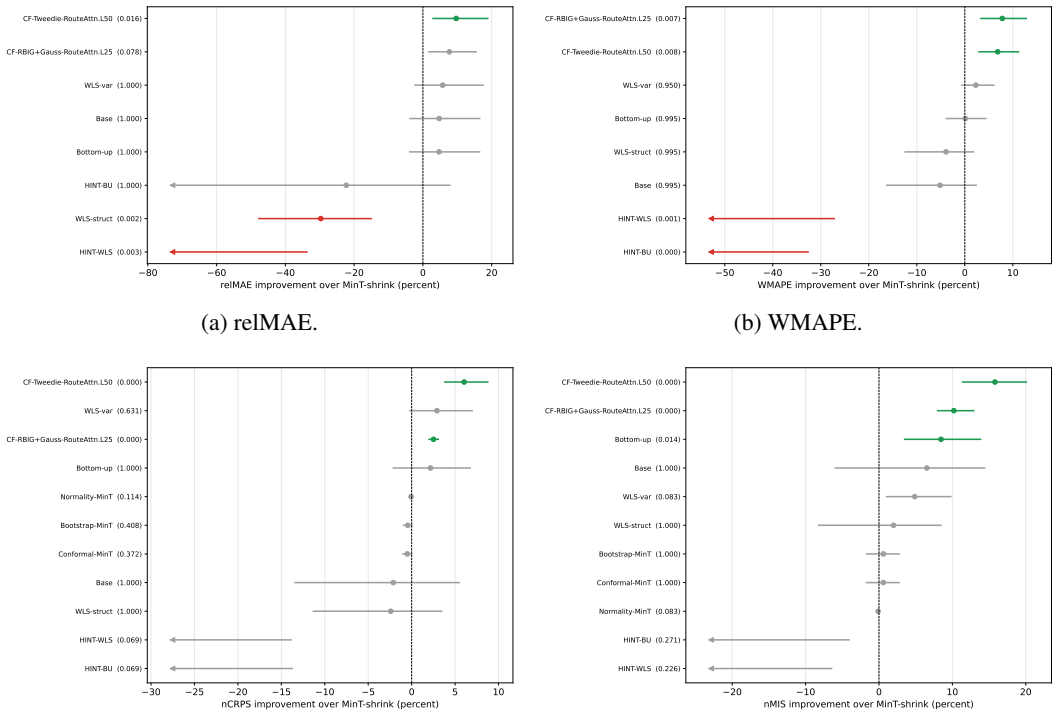
and the resulting point is coherent.

Proof. For $d_{at} \neq 0$, $\omega_t |e_{at} + \gamma d_{at}| = u_{at} |\gamma - r_{at}|$. All other terms are constant in γ. Thus $\mathcal{L}$ is a weighted absolute-deviation objective whose unconstrained minimizers are the weighted medians. Projection onto $[0, 1]$ gives a constrained minimizer. Since zero is feasible, equation 27 follows. Coherence follows because every point lies in $\mathrm{range}(S)$. □

The guarantee applies to the calibration cases used to select γ^*. It provides an exact empirical fallback without changing the head architecture.

A.2 Experimental protocol and inference

The corpus contains six public benchmarks and 23 anonymized industry hierarchies. The public sets are Australian Labour ($n = 57$), Tourism Small ($n = 89$), Tourism Large ($n = 555$), Traffic ($n = 207$), Wiki2 ($n = 199$), and a product-level M5 hierarchy ($n \approx 3060$). A shared LightGBM forecaster produces the base forecasts for every post-hoc reconciler. We evaluate standard bottom-up, middle-out, OLS, WLS, MinT-sample, and MinT-shrink methods, three maintained probabilistic MinT baselines, two HINT configurations, and all CoFlow variants. Top-down proportions are undefined when a period has zero total, and PERMBU requires a strict tree, so those two methods do not enter the common 29-dataset analysis set.



(a) relMAE. (b) WMAPE.

(c) nCRPS. (d) nMIS.

Figure 3: Improvement over MinT-shrink with dataset-cluster bootstrap 95% intervals. Positive values favor the method. Labels give Holm-adjusted dataset-clustered sign-flip p values. Arrows mark clipped intervals.

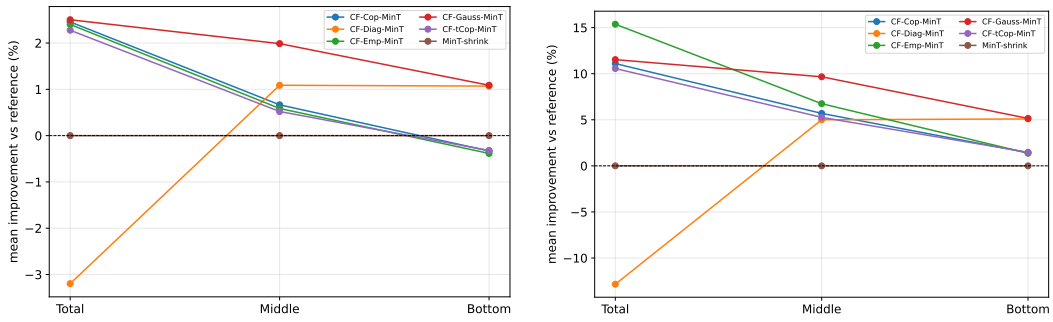
Each dataset is evaluated at three adjacent rolling origins, each holding out one complete dataset-specific horizon. Every stochastic method uses five seeds. Seeds are averaged within origin because they share one observed future and quantify optimization or Monte Carlo variation rather than independent forecast tasks. The resulting 87 observations are repeated measurements within 29 independent dataset blocks.

Pairwise inference uses a dataset-clustered sign-flip test that flips all three origin effects of one dataset together. Dataset-cluster bootstrap intervals, origin and dataset win counts, and dataset-level Wilcoxon tests provide complementary summaries. Friedman and Nemenyi comparisons use the 29 dataset means. For level-wise analysis, named levels are first averaged into Total, Middle, and Bottom within each dataset-origin, so hierarchy levels are never treated as independent replicates.

A.3 Full main comparison and statistical summaries

The two CoFlow rows occupy the top two positions on every metric in Figure 2. On nCRPS and nMIS, CoFlow-Tweedie-RouteAttn.L50 is separated from MinT, all three probabilistic MinT variants, and both HINT configurations. The point panels give the more conservative all-pairs view.

Figure 3 reports prespecified comparisons with MinT-shrink. The labels use Holm adjustment over the control comparisons in each panel. The all-pairs rank procedure and the control analysis answer different questions, which explains their different point-metric conclusions. The predictive conclusions agree under both analyses.



(a) nCRPS. (b) nMIS.

Figure 4: Improvement over MinT-shrink by level class for the dependence ablation. Positive values favor the method.

A.4 Extended discussion and model selection

A.4.1 Dependence and marginal shape act at different levels

The leaf models vary along two conceptually separate axes. The diagonal and full Gaussian laws hold the fitted leaf variances and marginal family fixed while changing cross-leaf covariance. The copula variants retain a score-space dependence model while changing the leaf marginals. Student-t dependence additionally permits joint tail movement. Empirical resampling preserves historical error vectors, and the flows change marginal shape and dependence together. These comparisons should not be collapsed into a single ordering of "distributional flexibility," because each axis enters an aggregate forecast differently.

At a leaf, a univariate score depends only on that leaf's marginal law. An aggregate score also depends on cross-leaf association because the scale of node a contains every covariance in $s_a^\top \Sigma_b s_a$. The full-versus-diagonal Gaussian contrast in Table 5 isolates this term: removing covariance changes nCRPS from 0.983 to 0.995 and nMIS from 0.925 to 0.978. The level breakdown in Figure 4 shows that the difference is concentrated above the Bottom level. This pattern is consistent with Corollary 1: removing covariance matters most when covariance contributes a large share of aggregate variance. It is not evidence that hierarchy depth by itself is causal. The general quantity is the node-specific covariance share θ_a. Support size is monotone only under the exchangeable positive-dependence assumptions stated in the corollary.

Marginal shape remains important where aggregation cannot average it away. In Table 6, Tweedie and ZINB produce the largest pooled improvements, and the level results show that Tweedie retains a substantial Bottom-level interval-score gain. A central-limit argument gives useful but incomplete intuition: summing many leaves can make standardized aggregate shape more Gaussian, while the aggregate scale still depends on the complete covariance. Heterogeneous, strongly skewed, or common-shock leaves need not approach that regime quickly. The level-wise evidence should consequently be read as a diagnostic of this corpus rather than as a universal ranking of dependence over marginal shape.

A.4.2 Capacity and finite-sample behavior

Greater distributional capacity creates estimation variance as well as approximation power. A joint residual law is fitted from one historical vector per time point, so increasing the number of leaves does not create more independent joint observations. When m is comparable to or larger than the residual history, a dense flow or empirical joint law can be data-limited even though each leaf has many scalar residuals. Shrinkage gives the Gaussian law a strong finite-sample inductive bias, and RBIG adds flexible marginal Gaussianization without training a large gradient-based network.

Table 8 illustrates the resulting bias–variance tradeoff. Likelihood-trained RealNVP attains nCRPS 0.993 and nMIS 0.966, whereas the energy-trained variant reaches 1.004 and 1.011. This comparison concerns the implemented objectives, architecture, and sample sizes. It does not imply

that likelihood universally dominates score training. RBIG is stronger on nMIS but more variable across datasets. The fixed RBIG–Gaussian hedge reduces that sensitivity, attaining nCRPS 0.975 and nMIS 0.901 with smaller cross-dataset dispersion than RBIG alone. The hedge is effective because its components have complementary errors. This result does not imply that a mixture must outperform its members on every dataset.

Capacity should therefore be increased only when rolling validation resolves an improvement and the gain is stable across origins. Evidence against increased capacity includes a large average gain accompanied by much larger log-ratio dispersion or strong dependence on a small subset of datasets. Parameter count, residual-history length, hierarchy width, and the number of validation origins should be considered together. For very wide hierarchies, block, factor, sparse, or low-rank dependence models are more credible extensions than an unrestricted dense law fitted to a short residual history.

A.4.3 Conditional scale, anchor, and residual source

The coherent anchor and the error source answer different questions. The anchor fixes predictive location after centering. The residual source supplies scale, shape, and dependence. This distinction is visible in Table 7: changing the Gaussian anchor from MinT to bottom-up has little effect on pooled nMIS (0.925 versus 0.924), while fitting Gaussian errors after MinT projection raises nMIS to 0.975. The projected residuals already contain the covariance reduction in equation 5. Using them to fit a second predictive law can therefore carry that reduction forward. The result identifies a consequential modeling choice in this experiment, not a general theorem that raw residuals are always preferable.

Residuals should represent the uncertainty left after conditioning on the information available to the base forecaster. Unconditional variation in the observed series also contains predictable level, seasonality, and promotion effects and can therefore be much wider than forecast error. For the continuous laws, we use centered base residuals as a stable shared source. For Tweedie and ZINB, the dispersion is matched along the fitted mean so that predictable variation is not added back as uncertainty. If residual scale changes materially with horizon, season, or forecast magnitude, a single static law is insufficient. Stratified scale estimates or a conditional leaf law should then be validated.

Raw base residuals are attractive when the reconciled anchor has too few out-of-sample historical origins because they are plentiful and common to all post-hoc methods. Their limitation is mismatch. The fitted errors are not errors around the chosen future anchor. Rolling cross-fitting removes that mismatch by refitting the base forecaster, estimating the reconciliation map, forming the anchor, and recording its future error at each historical origin. These errors target the selected forecasting and reconciliation procedure but are fewer and more expensive to obtain. The resulting choice is a bias–variance tradeoff. Cross-fitted anchor residuals are preferable when enough origins are available. Otherwise, the stable raw-residual fit can be compared with a regularized cross-fitted alternative on held-out origins.

A.4.4 Nonnegative and atomic marginals

Tweedie and ZINB address related but different support assumptions. For power $1 < p < 2$, Tweedie is a compound Poisson–Gamma law with an atom at zero and a continuous positive tail. It is suitable for intermittent nonnegative amounts but is not an integer-count model. ZINB is discrete, adds explicit zero inflation, and is appropriate when integer support is meaningful. Both models set their future mean from the coherent leaf anchor and estimate dispersion or zero inflation from historical fitted values and residuals.

The dependence construction is deliberately described as *rank coupling*. Independent marginal draws are sorted and assigned the ranks of correlated Gaussian scores. Before the final centering step, this preserves each finite-sample marginal draw multiset exactly and transfers monotone rank association. It also avoids requiring a closed-form Tweedie quantile. For an atomic marginal, however, tied zeros correspond to an interval of latent ranks. Permutations inside that block produce the same observed values. A unique continuous copula is therefore not identified inside the atom. Terms such as "Gaussian-copula ZINB" should be understood as the specified rank-coupling algorithm, not as identification of latent dependence among tied zeros.

The mean ratios alone favor ZINB in Table 6: its nCRPS and nMIS ratios are 0.935 and 0.745, compared with 0.938 and 0.840 for Tweedie. Stability changes the decision. ZINB's per-dataset log-ratio standard deviations are 0.139 and 0.315, versus 0.074 and 0.143 for Tweedie, and its advantage is concentrated in part of the corpus. Tweedie also has the better average nCRPS rank. We therefore use Tweedie in the main composition and retain ZINB as evidence that explicit count structure can be valuable when its assumptions match the dataset.

There is a second support tradeoff. Finite-sample centering makes the empirical draw mean exactly equal to the chosen anchor, but subtracting a sample mean can move ZINB draws off integer and nonnegative support and can move Tweedie draws below zero. When exact sample support is required by the statistical setting, finite-K centering can be omitted for the count branch. Every uncentered leaf draw and its aggregate remain coherent and on support, while the Monte Carlo mean approaches the anchor as K grows. The experiments use exact finite-sample location for a common comparison. This choice makes explicit the tradeoff between anchor equality and support preservation.

A.4.5 Separate predictive laws and point forecasts

The predictive law and coherent point forecast are separate both algebraically and by construction. Changing the leaf law changes the samples $S(\hat{b} + \eta)$ but not the point $S\{\hat{b} + \delta_\theta(f)\}$. Changing the head changes the latter without touching any predictive draw. Table 12 provides direct empirical evidence. Within the Gaussian law, all five heads retain nCRPS 0.983 and nMIS 0.925. Within RBIG, they retain nCRPS 0.987 and nMIS 0.914. Conversely, the same MLP has relMAE 0.985 under either law, and leaf attention gives 0.967 and 0.968. Table 13 extends the same pattern to the observed combinations.

The point architectures emphasize different loss structures. Leaf attention permits redistribution across series and is more stable when each series receives similar weight. Routing uses discrepancies from direct aggregate forecasts and is strongest under volume-weighted accuracy. Table 11 makes this tradeoff explicit. Pure WMAPE training ($\lambda = 0$) attains WMAPE 0.915 but relMAE 0.973. Setting $\lambda = 0.25$ retains WMAPE 0.920 while improving relMAE to 0.922. Setting $\lambda = 0.50$ reaches relMAE 0.899 with WMAPE 0.932. The weight should encode the target evaluation objective and be fixed before final evaluation rather than selected retrospectively from the test frontier.

The scalar fallback in Proposition 1 is useful when the learned correction is unstable. It can only improve the specified weighted absolute calibration loss because MinT is the $\gamma = 0$ endpoint. It does not guarantee improvement on a future origin, on a different metric, or after repeatedly selecting among many heads. A distinct chronological block should be used for this gate whenever the residual history permits one.

A.4.6 Model-selection considerations

The following sequence separates the principal modeling choices and limits unnecessary search.

1. **Experimental design.** Rolling origins, horizon, base forecaster, hierarchy, residual window, and evaluation metrics should be specified before fitting a leaf law. All learned inputs are recomputed at each origin. Seeds quantify Monte Carlo or optimization variation and are averaged within an origin rather than treated as additional forecast cases.
2. **Dependence diagnostic.** Full and diagonal Gaussian leaf errors fitted from the same residuals isolate the contribution of cross-leaf dependence. Proper scores are compared by Total, Middle, and Bottom classes. A material full-versus-diagonal gap above the leaves supports further dependence modeling. A small gap instead shifts attention toward marginal shape or conditional scale.
3. **Marginal support.** The Gaussian law is the baseline for signed, approximately symmetric residuals. Student-t, empirical, or Gaussian-copula marginals address continuous heavy tails. Tweedie addresses nonnegative intermittent amounts. ZINB addresses counts with excess zeros, for which stability should be considered together with the pooled mean. A flow is supported only when rolling validation shows a repeatable gain over these structured candidates.

4. **Predictive-law comparison.** Coverage gap alone rewards excessive width and is therefore interpreted beside nCRPS and nMIS. Energy score measures joint dependence. Per-dataset log ratios, ranks, and level classes supplement the pooled average. A fixed mixture is justified when components have complementary validation errors, not solely because it expands the candidate set.
5. **Point-head comparison.** If MinT point accuracy is adequate, no head is required. Otherwise, local or leaf-attention heads address per-series loss and routing addresses volume-weighted loss. The value of λ follows the stated metric tradeoff or nested rolling validation. Separate held-out blocks are used for early stopping and, when applicable, the scalar fallback.
6. **Robustness analysis.** Coverage and proper scores are examined by level, horizon, and dataset rather than only through a pooled mean. Atomic laws require an explicit choice between exact support and exact finite-sample centering. Wide hierarchies additionally motivate comparisons with structured dependence and sparse point architectures.

A sequential analysis reduces the risk that a joint search over anchors, residual sources, laws, mixtures, heads, and losses overfits a small number of origins. It also preserves the interpretation of each ablation. Each branch is selected from a targeted comparison, and the final composition is evaluated once on untouched origins.

A.4.7 Calibration while preserving coherence

Calibration problems should first be classified as location, scale, marginal-shape, or dependence errors. Persistent signed residuals indicate a location problem and belong in the point branch. Approximately correct central shape with systematic undercoverage suggests scale. Poor leaf scores with acceptable aggregate coverage indicate marginal mismatch, while aggregate undercoverage that grows with covariance share points toward dependence. These diagnoses are imperfect, but they map more directly to the components of equation 4 than a single pooled coverage number.

A scalar or horizon-specific multiplier applied to the complete leaf-error vector preserves coherence: samples remain $S(\hat{b} + c_h \eta_h)$. More general corrections can also be learned in leaf space and then passed through S. In contrast, independently rescaling or shifting each node after aggregation generally moves samples outside $\mathrm{range}(S)$. The same limitation applies to separate nodewise conformal intervals: marginal intervals can be reported, but independently altered sample paths are not a coherent joint predictive unless the correction is constructed through leaf coordinates or another coherent parameterization.

Calibration parameters must be estimated out of sample with respect to the base forecaster and reconciler. Reusing in-sample fitted residuals can hide model adaptation and make intervals appear too narrow. When only a few rolling origins are available, low-dimensional adjustments are preferable, and their variability should be reported. Empirical coverage should always be paired with interval score or CRPS. The HINT comparison in Appendix A.11 shows why coverage alone is insufficient.

A.5 Distributional transfer through the summing map

Lemma 1 is topological. The following standard Lipschitz consequence is quantitative. Let $\mathcal{P}_p(\mathbb{R}^d)$ denote the probability measures on $\mathbb{R}^d$ with finite pth moment, and let W_p denote the p-Wasserstein distance induced by the Euclidean norm.

Lemma 3 (Bi-Lipschitz Wasserstein transfer)**.** *Let $S \in \mathbb{R}^{n \times m}$ have full column rank, with smallest and largest singular values $\sigma_{\min}(S)$ and $\sigma_{\max}(S)$. For every $p \geq 1$ and every $\nu, \nu' \in \mathcal{P}_p(\mathbb{R}^m)$,*

$$\sigma_{\min}(S) W_p(\nu, \nu') \leq W_p(S_{\#}\nu, S_{\#}\nu') \leq \sigma_{\max}(S) W_p(\nu, \nu'). \tag{28}$$

For the marginal law at node a,

$$W_p((s_a^\top)_{\#}\nu, (s_a^\top)_{\#}\nu') \leq \|s_a\|_2 W_p(\nu, \nu'). \tag{29}$$

For a binary summing row with q_a leaves, $\|s_a\|_2 = \sqrt{q_a}$. Consequently, density of a leaf model class in W_p is equivalent to density of its pushforwards among coherent laws in W_p.

Proof. Let π be any coupling of ν and ν', with $(X, X') \sim \pi$. Its image under $(x, x') \mapsto (Sx, Sx')$ is a coupling of the two pushforwards. Therefore

$$W_p^p(S_\#\nu, S_\#\nu') \le \mathbb{E}_\pi \|S(X - X')\|_2^p \le \sigma_{\max}(S)^p \mathbb{E}_\pi \|X - X'\|_2^p.$$

Taking the infimum over π proves the upper bound. Conversely, S^+ is the inverse of S on $\mathcal{C}$, and $\|S^+\|_2 = 1/\sigma_{\min}(S)$. Applying the upper bound to S^+ and to any coupling of the coherent pushforwards gives

$$W_p(\nu, \nu') \le \|S^+\|_2 W_p(S_\#\nu, S_\#\nu'),$$

which proves the lower bound. Applying the same coupling argument to the scalar linear map $x \mapsto s_a^\top x$ proves equation 29. The density statement follows from the two-sided bound. □

Wasserstein control bounds transportation error. Calibration requires control of event probabilities. The relevant fact is that an injective pushforward preserves statistical divergences at the full-law level, while taking a node marginal can only reduce them.

Lemma 4 (Divergence preservation and coverage transfer)**.** *Let ν_* be the true leaf-error law for the chosen leaf anchor and let $\widehat{\nu}$ be a predictive leaf law. Define $\mu_* = S_\#\nu_*$ and $\widehat{\mu} = S_\#\widehat{\nu}$. Then*

$$\mathrm{TV}(\mu_*, \widehat{\mu}) = \mathrm{TV}(\nu_*, \widehat{\nu}), \qquad \mathrm{KL}(\mu_* \| \widehat{\mu}) = \mathrm{KL}(\nu_* \| \widehat{\nu}), \tag{30}$$

where the KL identity holds whenever $\nu_* \ll \widehat{\nu}$. For node a, let $\mu_{*,a}$ and $\widehat{\mu}_a$ be the corresponding scalar marginals. Then

$$\mathrm{TV}(\mu_{*,a}, \widehat{\mu}_a) \le \mathrm{TV}(\nu_*, \widehat{\nu}), \qquad \mathrm{KL}(\mu_{*,a} \| \widehat{\mu}_a) \le \mathrm{KL}(\nu_* \| \widehat{\nu}). \tag{31}$$

If an interval I_a has predictive probability $\widehat{\mu}_a(I_a) = p_a$, then

$$|\mu_{*,a}(I_a) - p_a| \le \mathrm{TV}(\nu_*, \widehat{\nu}). \tag{32}$$

In particular, if $p_a = 1 - \alpha$ and $\mathrm{KL}(\nu_* \| \widehat{\nu}) \le \varepsilon_{\mathrm{KL}}$, then

$$|\mu_{*,a}(I_a) - (1 - \alpha)| \le \sqrt{\varepsilon_{\mathrm{KL}}/2}. \tag{33}$$

The same bound holds for any measurable joint event in the coherent space.

Proof. The restriction $S : \mathbb{R}^m \to \mathcal{C}$ is a continuous bijection with continuous inverse S^+. Hence measurable events in leaf space and in $\mathcal{C}$ correspond in both directions. Taking the supremum of probability differences over corresponding events proves the total-variation equality in equation 30. Invariance of the Radon-Nikodym derivative under a measurable bijection gives the KL equality. Equivalently, one may apply the KL data-processing inequality first to S and then to its inverse S^+, which gives both inequalities and hence equality.

The node marginal is the image under the measurable map $b \mapsto s_a^\top b$. The data-processing inequalities for total variation and KL prove equation 31. Applying the definition of total variation to the event I_a proves equation 32. Pinsker's inequality gives

$$\mathrm{TV}(\nu_*, \widehat{\nu}) \le \sqrt{\tfrac{1}{2}\mathrm{KL}(\nu_* \| \widehat{\nu})},$$

which proves equation 33. A joint coherent event is handled directly by the full-law total-variation equality. □

Remark 1 (Role of the anchor). The coverage result is conditional on ν_* being the error law of the same anchor used at forecast time. This condition separates support coherence from calibration. Coherence holds for every leaf law by Lemma 1, while the calibration bound is only as accurate as the estimated law around the chosen point forecast.

A.6 Finite-ensemble centering

Each finite ensemble is recentered to preserve the MinT ensemble mean exactly. This operation is a common translation of all draws, which allows its effect to be characterized without an asymptotic approximation.

Lemma 5 (Exact effect of ensemble centering). *Let $\eta^{(1)}, \dots, \eta^{(K)} \in \mathbb{R}^m$ be arbitrary draws, let $\bar{\eta} = K^{-1} \sum_k \eta^{(k)}$, and define*

$$\widehat{\nu}_K = \frac{1}{K} \sum_{k=1}^{K} \delta_{\eta^{(k)}}, \qquad \widetilde{\nu}_K = \frac{1}{K} \sum_{k=1}^{K} \delta_{\eta^{(k)} - \bar{\eta}}.$$

For every $p \geq 1$,

$$W_p(\widehat{\nu}_K, \widetilde{\nu}_K) = \|\bar{\eta}\|_2, \qquad W_p(S_{\#}\widehat{\nu}_K, S_{\#}\widetilde{\nu}_K) = \|S\bar{\eta}\|_2. \tag{34}$$

Centering preserves all pairwise sample differences and every coordinatewise rank. At node a, it translates every sample and every empirical quantile by $-s_a^\top \bar{\eta}$, so every interval width and the pairwise term of the empirical energy score are unchanged.

If the draws are independent with mean zero and covariance Σ_b, then

$$\mathbb{E}\|S\bar{\eta}\|_2^2 = \frac{1}{K} \operatorname{tr}(S\Sigma_b S^\top), \qquad \mathbb{E}|s_a^\top \bar{\eta}|^2 = \frac{1}{K} s_a^\top \Sigma_b s_a. \tag{35}$$

If each scalar $s_a^\top \eta$ is sub-Gaussian with variance proxy τ_a^2, then with probability at least $1 - \delta$,

$$|s_a^\top \bar{\eta}| \leq \tau_a \sqrt{\frac{2\log(2n/\delta)}{K}} \quad \text{for every } a \in \{1, \dots, n\}. \tag{36}$$

Proof. Coupling each $\eta^{(k)}$ with $\eta^{(k)} - \bar{\eta}$ gives the upper bound $W_p(\widehat{\nu}_K, \widetilde{\nu}_K) \leq \|\bar{\eta}\|_2$. For any coupling (X, Y) of the two empirical measures, Jensen's inequality gives

$$(\mathbb{E}\|X - Y\|_2^p)^{1/p} \geq \|\mathbb{E}X - \mathbb{E}Y\|_2 = \|\bar{\eta}\|_2.$$

This proves equality. Applying the same argument after multiplication by S proves the second identity in equation 34. The identities

$$(\eta^{(k)} - \bar{\eta}) - (\eta^{(\ell)} - \bar{\eta}) = \eta^{(k)} - \eta^{(\ell)}$$

prove preservation of pairwise differences. Subtracting one constant per coordinate preserves ranks and translates all quantiles, which proves the deterministic claims.

Under independence and zero mean, $\operatorname{Cov}(\bar{\eta}) = \Sigma_b / K$. The trace identity $\mathbb{E}\|X\|_2^2 = \operatorname{tr}\operatorname{Cov}(X)$ for a zero-mean vector proves equation 35. Finally, $s_a^\top \bar{\eta}$ is sub-Gaussian with proxy τ_a^2 / K. Its two-sided tail bound followed by a union bound over the n nodes proves equation 36. □

When the original draws are independent and identically distributed, the centered draws are exchangeable but not independent because they share $\bar{\eta}$. The lemma shows that their departure from the original empirical law is exactly a translation of order $K^{-1/2}$ under the stated moment conditions.

A.7 Hierarchy-weighted energy scoring

The energy score used for evaluation lives in the full n-dimensional hierarchy. Its exact leaf-space counterpart is obtained by replacing the Euclidean leaf norm with the norm induced by S.

Lemma 6 (Energy score commutes with the coherent pushforward). *For $x \in \mathbb{R}^m$, define $\|x\|_S = \|Sx\|_2$. This is a norm because S has full column rank. Let $\nu \in \mathcal{P}_1(\mathbb{R}^m)$ and let $b \in \mathbb{R}^m$. If X, X' are independent with law ν, then*

$$\operatorname{ES}(S_{\#}\nu, Sb) = \mathbb{E}\|X - b\|_S - \frac{1}{2}\mathbb{E}\|X - X'\|_S =: \operatorname{ES}_S(\nu, b). \tag{37}$$

If the true leaf law is $\nu_* \in \mathcal{P}_1(\mathbb{R}^m)$, then

$$\mathbb{E}_{B \sim \nu_*} \operatorname{ES}_S(\nu, B) - \mathbb{E}_{B \sim \nu_*} \operatorname{ES}_S(\nu_*, B) = \frac{1}{2}\mathcal{E}_S(\nu, \nu_*), \tag{38}$$

where

$$\mathcal{E}_S(\nu, \nu_*) = 2\mathbb{E}\|X - Y\|_S - \mathbb{E}\|X - X'\|_S - \mathbb{E}\|Y - Y'\|_S \tag{39}$$

for independent $X, X' \sim \nu$ and $Y, Y' \sim \nu_*$. The quantity $\mathcal{E}_S$ is nonnegative and equals zero if and only if $\nu = \nu_*$. Thus ES_S is strictly proper on leaf laws.

Table 5: Leaf-dependence ablation over 29 datasets. Entries are geometric-mean ratios to MinT-shrink ± the standard deviation of per-dataset log ratios, with lower values better. Bold marks the column best, and brackets give the two-sided Wilcoxon p against it.

Family	Method	nCRPS / MinT	nCRPS rank	nMIS / MinT	nMIS rank
Gaussian anchor	CoFlow-Gauss-MinT	**0.983 ± 0.013**	**1.62 ± 1.27**	**0.925 ± 0.048**	**1.90 ± 1.26**
No dependence	CoFlow-Diag-MinT	0.995 ± 0.023 (<.001)	3.34 ± 1.74 (<.001)	0.978 ± 0.104 (<.001)	3.90 ± 1.76 (<.001)
Copula dependence	CoFlow-Cop-MinT	0.995 ± 0.011 (<.001)	3.52 ± 0.95 (<.001)	0.957 ± 0.059 (<.001)	3.21 ± 1.24 (.006)
Copula dependence	CoFlow-tCop-MinT	0.996 ± 0.012 (<.001)	3.90 ± 1.37 (<.001)	0.960 ± 0.065 (<.001)	4.14 ± 1.27 (<.001)
Empirical joint	CoFlow-Emp-MinT	0.996 ± 0.014 (<.001)	4.24 ± 1.64 (<.001)	0.944 ± 0.066 (.076)	2.83 ± 1.42 (.059)
Reconciliation	MinT-shrink	1.000 ± 0.000 (<.001)	4.38 ± 1.68 (<.001)	1.000 ± 0.000 (<.001)	5.03 ± 1.43 (<.001)

Proof. If $Z = SX$ and $Z' = SX'$, substitution into the definition of the full-space energy score gives equation 37. Taking expectation over an independent truth $B \sim \nu_*$ yields

$$\mathbb{E}\|X - B\|_S - \frac{1}{2}\mathbb{E}\|X - X'\|_S.$$

The corresponding expression for the true forecast ν_* is $\frac{1}{2}\mathbb{E}\|Y - Y'\|_S$. Subtracting proves equation 38. Finally, $\mathcal{E}_S(\nu, \nu_*)$ is the ordinary Euclidean energy distance between $S_\#\nu$ and $S_\#\nu_*$. It is nonnegative and vanishes exactly when the two pushforwards agree. Injectivity of S then implies $\nu = \nu_*$. □

The ordinary Euclidean energy score on leaves replaces $\|\cdot\|_S$ by $\|\cdot\|_2$. It remains strictly proper, but it is not identical to the all-node evaluation score unless $S^\top S$ is a scalar multiple of the identity. Equation 37 therefore identifies the hierarchy-weighted objective that aligns leaf-model training exactly with full-space evaluation.

A.8 Predictive ablations

The appendix tables isolate one modeling choice at a time. Unless stated otherwise, entries are geometric-mean ratios to MinT-shrink over 29 datasets, with values below one indicating an improvement. The ± term is the standard deviation of the per-dataset log ratios. Bold marks the best value in a column. Brackets give the two-sided Wilcoxon p against that best entry, and a bold p falls below 0.05. Thus a small p indicates that a row is measurably behind the column winner, not that the row is desirable.

Critical-difference diagrams rank the 29 dataset means. Significance matrices instead use all 87 origins in the dataset-clustered sign-flip test, with green indicating that the row ranks better and red that it ranks worse. CoFlow is abbreviated CF- in these figures.

A.8.1 Leaf dependence

Table 5 changes only the dependence model. The full Gaussian covariance leads on both scores and both ranks. Removing dependence has the largest cost: the nMIS ratio rises from 0.925 to 0.978, and the average rank falls from 1.90 to 3.90. This is the empirical analogue of Corollary 1. Discarding the cross terms removes the covariance contribution c_a in equation 12.

More elaborate dependence models do not recover this loss. Gaussian and Student-t copulas differ by only 0.003 on nMIS (0.957 versus 0.960), indicating little benefit from tail co-movement once the marginals are fixed. The empirical joint law is the closest alternative at 0.944 (p=0.076), but does not improve the rank. These results favor second-order dependence estimated directly from raw leaf residuals. The pattern is consistent with replacing the conditioned covariance in equation 5 by a direct raw-leaf estimate rather than requiring a more elaborate copula.

A.8.2 Leaf marginal family

Table 6 fixes the Gaussian copula and varies the marginals. Count laws give the largest mean improvements. ZINB attains 0.935 on nCRPS and 0.745 on nMIS, while Tweedie attains 0.938 and 0.840. The corresponding Gaussian ratios are 0.983 and 0.925. Every continuous alternative lies between the Gaussian law and MinT-shrink, including the nonparametric marginals at 0.995 on



(a) nCRPS. (b) nMIS.

Figure 5: Average ranks for the dependence ablation over 29 datasets. Lower ranks are better, and connected methods are not separated by the Nemenyi test at 0.05.

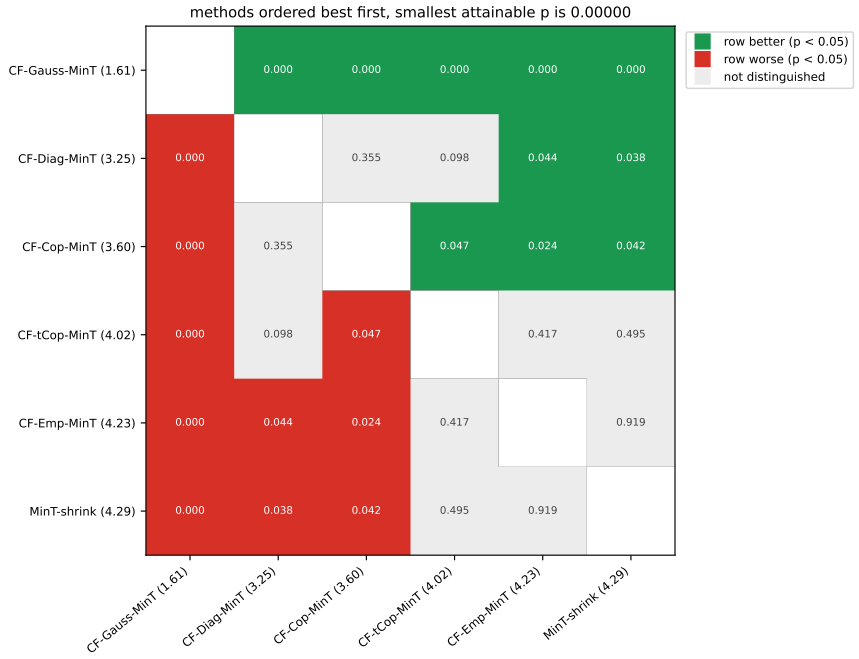


Figure 6: Pairwise dataset-clustered sign-flip decisions for nCRPS over 87 origins. Green indicates that the row ranks better and red that it ranks worse.

nCRPS. For intermittent data, neither a heavier continuous tail nor an empirical marginal fully substitutes for a distribution with an atom at zero.

The ranks add an important qualification. Tweedie has the best nCRPS rank, 2.31, while ZINB has the best mean ratio. The two are not separated on either statistic (p=0.22 for the ratio and p=0.97 for the rank), suggesting that ZINB's larger mean gain is concentrated in a subset of datasets. ZINB also has substantially larger cross-dataset dispersion than Tweedie. We therefore use Tweedie in the main comparison as the more stable count law, rather than selecting only by the best pooled mean. The flow models also differ: RBIG reaches 0.914 on nMIS, whereas RealNVP reaches 0.967. This motivates the hedged variants studied below.

A.8.3 Point anchor and residual source

Table 7 distinguishes the forecast anchor from the residual source. The anchor has no measurable effect. Replacing the MinT anchor with raw leaf forecasts changes nMIS from 0.925 to 0.924 for the Gaussian law, from 0.914 to 0.920 for RBIG, and from 0.967 to 0.970 for RealNVP. Neither Gaussian anchor nor bottom-up-anchored RBIG is separated from the best entry ($p \geq 0.24$). This supports the separation in Section 3.5: location and predictive shape can be chosen independently.

The residual source matters substantially. When the Gaussian law is fitted to residuals after MinT projection, nCRPS moves from 0.983 to 0.994 and nMIS from 0.925 to 0.975 ($p < 10^{-3}$ for both).

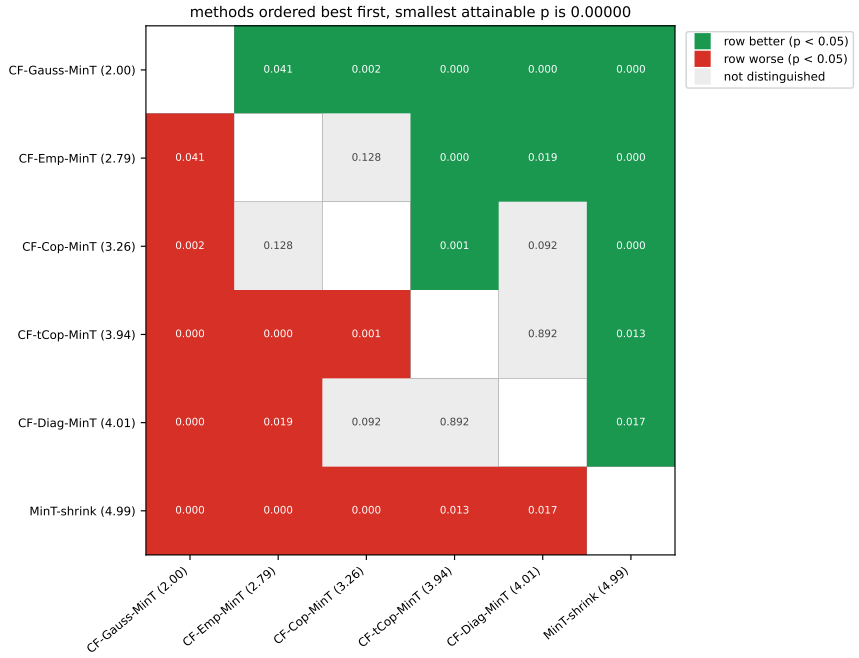


Figure 7: Pairwise dataset-clustered sign-flip decisions for nMIS over 87 origins. Green indicates that the row ranks better and red that it ranks worse.

Table 6: Leaf-marginal ablation over 29 datasets. Entries are geometric-mean ratios to MinT-shrink $\pm$ the standard deviation of per-dataset log ratios, with lower values better. Bold marks the column best, and brackets give the two-sided Wilcoxon p against it.

Family	Method	nCRPS / MinT	nCRPS rank	nMIS / MinT	nMIS rank
Gaussian anchor	CoFlow-Gauss-MinT	0.983 ± 0.013 (.005)	4.24 ± 2.20 (<.001)	0.925 ± 0.048 (<.001)	4.76 ± 2.15 (.004)
Nonparametric marginals	CoFlow-Cop-MinT	0.995 ± 0.011 (<.001)	7.78 ± 1.58 (<.001)	0.957 ± 0.059 (<.001)	6.47 ± 1.91 (<.001)
Nonparametric marginals	CoFlow-tCop-MinT	0.996 ± 0.012 (<.001)	8.03 ± 2.24 (<.001)	0.960 ± 0.065 (<.001)	7.86 ± 2.34 (<.001)
Nonparametric marginals	CoFlow-Emp-MinT	0.996 ± 0.014 (<.001)	8.17 ± 2.62 (<.001)	0.944 ± 0.066 (<.001)	6.10 ± 2.32 (<.001)
Nonparametric marginals	CoFlow-Gate-MinT	0.995 ± 0.012 (<.001)	6.81 ± 1.52 (<.001)	0.958 ± 0.058 (<.001)	7.36 ± 1.80 (<.001)
Heavy-tailed marginals	CoFlow-tMarg-MinT	0.986 ± 0.015 (.004)	4.72 ± 2.02 (<.001)	0.932 ± 0.053 (<.001)	5.79 ± 2.34 (<.001)
Count marginals	CoFlow-ZINB-MinT	**0.935 ± 0.139**	3.21 ± 3.86 (.973)	**0.745 ± 0.315**	**2.41 ± 3.02**
Count marginals	CoFlow-Tweedie-MinT	0.938 ± 0.074 (.221)	**2.31 ± 2.07**	0.840 ± 0.143 (.004)	3.03 ± 2.56 (.064)
Normalising flows	CoFlow-RBIG-MinT	0.987 ± 0.025 (<.001)	5.21 ± 2.72 (<.001)	0.914 ± 0.131 (<.001)	4.45 ± 2.72 (.002)
Normalising flows	CoFlow-NVP-MinT	0.993 ± 0.015 (<.001)	7.00 ± 2.51 (<.001)	0.966 ± 0.064 (<.001)	8.28 ± 2.74 (<.001)
Reconciliation	MinT-shrink	1.000 ± 0.000 (<.001)	8.52 ± 2.64 (<.001)	1.000 ± 0.000 (<.001)	9.48 ± 2.29 (<.001)

Those residuals have already lost the conditioning term in equation 5. Centering can therefore change without consequence, but estimating the leaf law after projection carries the attenuation into the new predictive.

A.8.4 Flow objective and hedging

Table 8 compares the flow objective and the hedges. Maximum likelihood is more effective than the ordinary leaf-space energy objective: the respective ratios are 0.993 and 1.004 on nCRPS, and 0.967 and 1.011 on nMIS. The hierarchy-weighted score ES_S in Lemma 6, rather than ordinary Euclidean leaf geometry, is the objective aligned with all-node evaluation. The observed advantage of likelihood is consistent with this mismatch, although we do not train the flow with ES_S here.

Among the mixtures, RBIG with the Gaussian anchor leads both scores and both ranks, with ratios 0.975 and 0.901. RBIG alone remains tied with the hedge on nMIS (0.914, p=0.75), but not on nCRPS. The Gaussian component therefore improves sharpness while preserving RBIG's interval-score advantage. This specific complementarity, rather than hedging in the abstract, motivates the recommendation in Section A.4.6.



(a) nCRPS. (b) nMIS.

Figure 8: Average ranks for the marginal-family ablation over 29 datasets. Lower ranks are better, and connected methods are not separated by the Nemenyi test at 0.05.

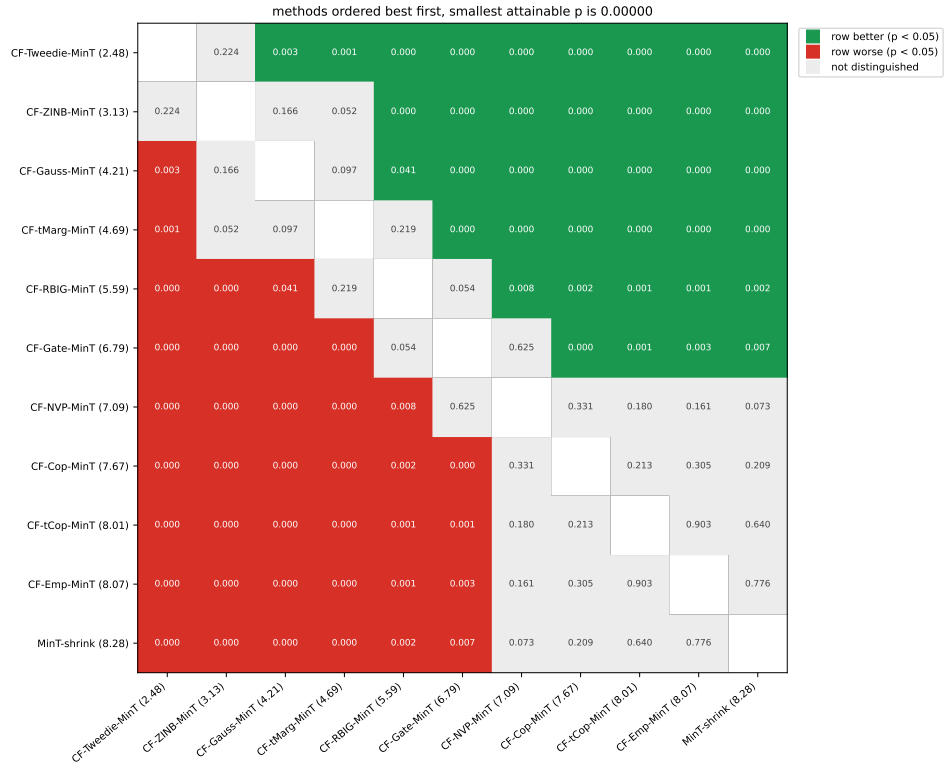


Figure 9: Pairwise dataset-clustered sign-flip decisions for marginal-family nCRPS over 87 origins. Green indicates that the row ranks better and red that it ranks worse.

A.9 Point-head ablations

The point head changes only the point forecast. Predictive samples and calibration scores remain fixed for a given leaf law, so the following studies compare point accuracy against MinT-shrink.

A.9.1 Attention support

Table 9 varies the information available to each correction, from independent per-leaf processing to dense attention and sparse routing along S. Cross-leaf attention achieves the best relMAE ratio, 0.967, and improves on the per-leaf network at 0.985 (p=0.027). However, no graph-structured attention head is separated from it. The closest reaches 0.974 (p=0.33). Cross-leaf context is therefore useful, but the exact attention support contributes little additional relMAE.

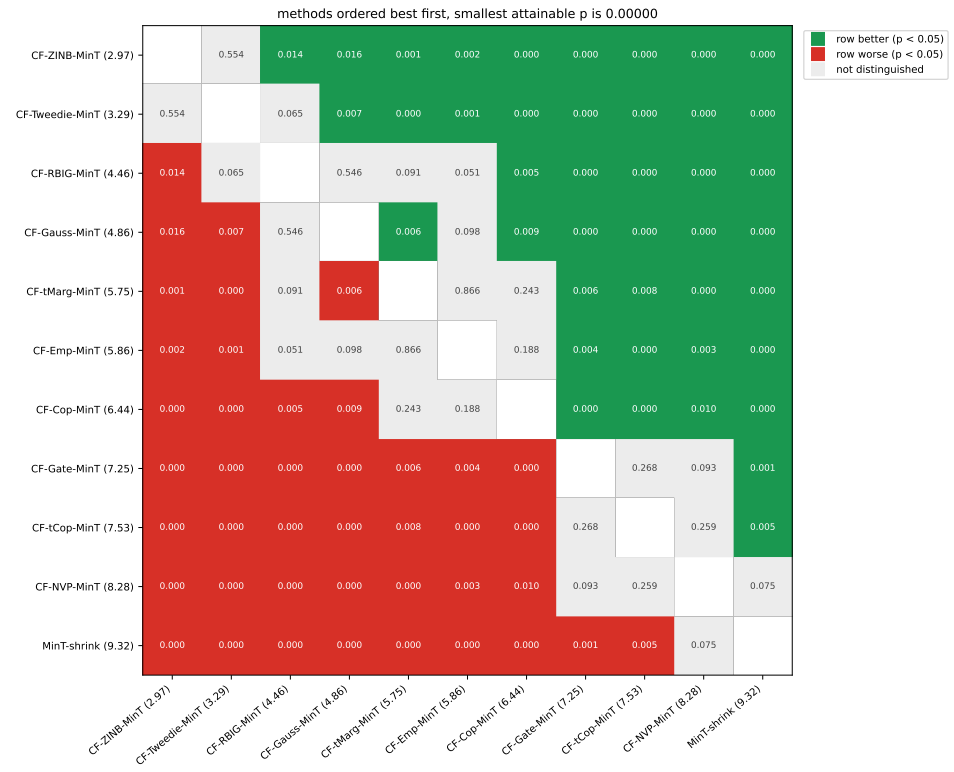
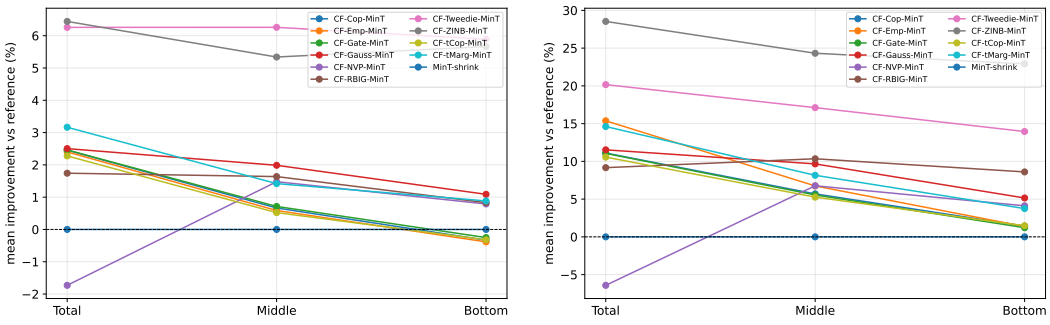


Figure 10: Pairwise dataset-clustered sign-flip decisions for marginal-family nMIS over 87 origins. Green indicates that the row ranks better and red that it ranks worse.



(a) nCRPS. (b) nMIS.

Figure 11: nCRPS and nMIS improvement over MinT-shrink by level class for the marginal-family ablation. Positive values favor the method.

WMAPE shows a different pattern. Constraint routing leads at 0.915, while the dense attention heads cluster between 0.994 and 0.996 and are all measurably worse ($p \leq 0.007$). Routing sends each aggregate discrepancy to its descendant leaves, directly targeting volume error. Dense attention instead redistributes information among leaves. Together with Section A.9.2, this shows that topology and objective alignment are distinct sources of improvement.

A.9.2 Metric-aligned loss weight

Table 10 varies the weight λ between per-series relMAE and pooled WMAPE for both attention architectures. The relMAE response is monotone and consistent across architectures. With cross-leaf attention, the ratio falls from 0.967 without metric alignment to 0.952 at λ=0.25 and 0.934 at

Table 7: Anchor and residual-source ablation over 29 datasets. Entries are geometric-mean ratios to MinT-shrink ± the standard deviation of per-dataset log ratios, with lower values better. Bold marks the column best, and brackets give the two-sided Wilcoxon p against it.

Family	Method	nCRPS / MinT	nCRPS rank	nMIS / MinT	nMIS rank
MinT anchor	CoFlow-Gauss-MinT	**0.983 ± 0.013**	**3.48 ± 2.11**	0.925 ± 0.048 (.080)	3.62 ± 1.80 (.194)
MinT anchor	CoFlow-RBIG-MinT	0.987 ± 0.025 (.495)	3.76 ± 2.32 (.663)	**0.914 ± 0.131**	**3.14 ± 2.22**
MinT anchor	CoFlow-NVP-MinT	0.993 ± 0.015 (<.001)	5.86 ± 2.33 (<.001)	0.966 ± 0.064 (<.001)	6.24 ± 2.29 (<.001)
Bottom-up anchor	CoFlow-Gauss-BU	0.985 ± 0.114 (.717)	4.59 ± 2.37 (.149)	0.924 ± 0.151 (.198)	4.21 ± 1.68 (.118)
Bottom-up anchor	CoFlow-RBIG-BU	0.992 ± 0.126 (.749)	4.79 ± 2.90 (.116)	0.920 ± 0.246 (.624)	3.38 ± 2.85 (.833)
Bottom-up anchor	CoFlow-NVP-BU	0.996 ± 0.113 (.132)	6.62 ± 2.27 (**.002**)	0.970 ± 0.148 (<.001)	6.72 ± 1.85 (<.001)
Reconciled residuals	CoFlow-Gauss.reconres-MinT	0.994 ± 0.011 (<.001)	5.59 ± 2.15 (<.001)	0.975 ± 0.052 (<.001)	6.48 ± 1.64 (<.001)
Reconciliation	MinT-shrink	1.000 ± 0.000 (<.001)	6.45 ± 2.63 (<.001)	1.000 ± 0.000 (<.001)	7.41 ± 2.21 (<.001)
Reconciliation	Bottom-up	0.984 ± 0.114 (.782)	3.86 ± 2.15 (.440)	0.922 ± 0.152 (.213)	3.79 ± 1.84 (.286)



(a) nCRPS. (b) nMIS.

Figure 12: Average ranks for the anchor and residual-source ablation over 29 datasets. Lower ranks are better, and connected methods are not separated by the Nemenyi test at 0.05.

λ=0.50. Over all node pairs, the corresponding sequence is 0.975, 0.965, and 0.955. The common ordering confirms that objective alignment is separate from attention topology.

The balanced cross-leaf head leads on relMAE, including against the balanced all-node head (p=0.006). WMAPE remains between 0.992 and 0.996, with no detectable difference among the rows. Metric alignment therefore improves per-series accuracy on both attention architectures but does not create the aggregate correction supplied by routing.

A.9.3 Routing loss frontier

Table 11 exposes a direct trade-off on the routing head. As λ increases from 0 to 1, relMAE improves from 0.973 to 0.899 at λ=0.50 and 0.75, then rises slightly to 0.906. Over the same sweep, WMAPE moves from 0.915 to 0.968. The relMAE difference between either endpoint and the minimum is well resolved ($p < 10^{-3}$), whereas the WMAPE changes are not distinguishable from the λ=0 endpoint.

The values used in the paper lie near the flatter parts of this curve. λ=0.25 preserves WMAPE at 0.920 while recovering most of the relMAE gain, and λ=0.50 reaches the relMAE minimum with a 1.2-point WMAPE cost. The weight is fixed across datasets, so the sweep characterizes a fixed metric trade-off rather than dataset-specific tuning.

A.10 Supporting comparisons

Table 12 carries the same five point heads on a Gaussian and on an RBIG leaf law, providing a direct test of the proposed separation. Within each law, predictive scores remain constant across the five heads. For example, the Gaussian MLP and Gaussian leaf-attention rows both attain 0.983 on nCRPS (p=1.000). Within a head, changing the law leaves point accuracy effectively unchanged: the two MLP rows both attain relMAE 0.985, while the two leaf-attention rows attain 0.967 and 0.968. Only one of the ten rows is separated from the best on any predictive metric. This is the empirical separability described in Section A.4.5.

Table 13 extends this comparison to every observed combination. Point columns follow the head and predictive columns follow the leaf law, with no visible interaction at the resolution of the corpus.

Table 14 compares the standard reconcilers under the same base forecaster. Their ordering agrees with prior work. MinT-sample degrades relMAE by nearly a factor of seven because the sample

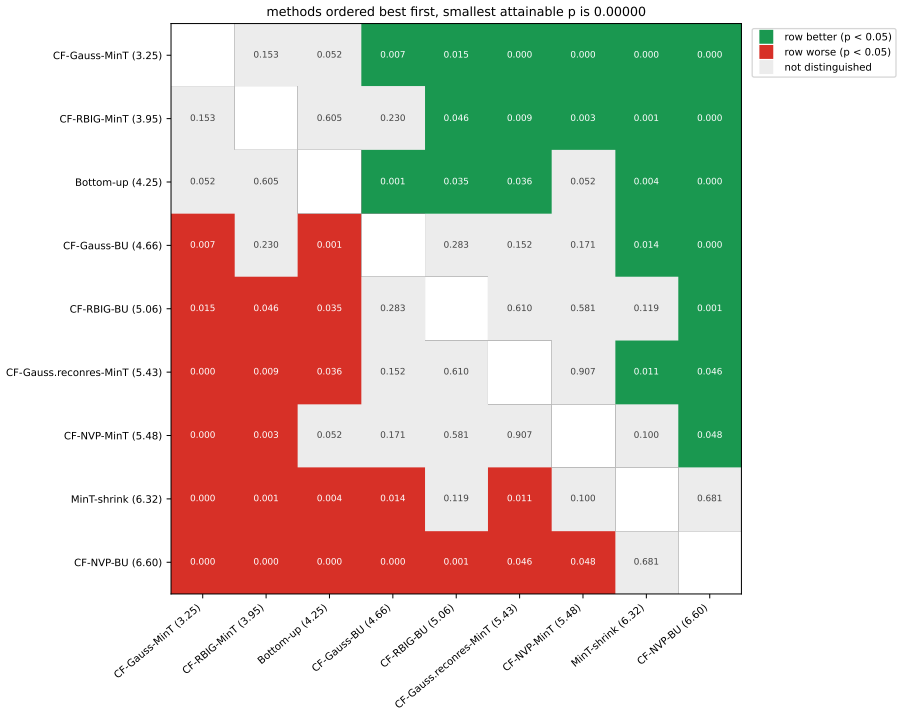


Figure 13: Pairwise dataset-clustered sign-flip decisions for anchor and residual-source nCRPS over 87 origins. Green indicates that the row ranks better and red that it ranks worse.

Table 8: Flow-objective and hedge ablation over 29 datasets. Entries are geometric-mean ratios to MinT-shrink ± the standard deviation of per-dataset log ratios, with lower values better. Bold marks the column best, and brackets give the two-sided Wilcoxon p against it.

Family	Method	nCRPS / MinT	nCRPS rank	nMIS / MinT	nMIS rank
Gaussian anchor	CoFlow-Gauss-MinT	0.983 ± 0.013 (.002)	2.93 ± 1.67 (.001)	0.925 ± 0.048 (.004)	3.21 ± 1.42 (<.001)
Flow alone	CoFlow-RBIG-MinT	0.987 ± 0.025 (<.001)	4.03 ± 2.29 (<.001)	0.914 ± 0.131 (.749)	2.45 ± 2.06 (.281)
Flow alone	CoFlow-NVP-MinT	0.993 ± 0.015 (<.001)	6.03 ± 1.05 (<.001)	0.966 ± 0.064 (<.001)	5.97 ± 1.21 (<.001)
Flow objective	CoFlow-NVP.energy-MinT	1.004 ± 0.023 (<.001)	7.10 ± 1.08 (<.001)	1.011 ± 0.103 (<.001)	7.24 ± 0.99 (<.001)
Hedged flows	CoFlow-RBIG+Gauss-MinT	**0.975 ± 0.016**	**1.66 ± 0.72**	**0.901 ± 0.079**	**2.03 ± 0.98**
Hedged flows	CoFlow-NVP+Gauss-MinT	0.987 ± 0.011 (<.001)	4.24 ± 1.09 (<.001)	0.939 ± 0.051 (<.001)	4.48 ± 1.15 (<.001)
Hedged flows	CoFlow-RBIG+NVP-MinT	0.981 ± 0.019 (<.001)	3.38 ± 1.32 (<.001)	0.924 ± 0.089 (<.001)	3.76 ± 1.12 (<.001)
Reconciliation	MinT-shrink	1.000 ± 0.000 (<.001)	6.62 ± 1.76 (<.001)	1.000 ± 0.000 (<.001)	6.86 ± 1.57 (<.001)

covariance is singular when the number of series exceeds the residual history. This is precisely the regime motivating the shrinkage estimator in equation 3.

A.11 External comparisons

Table 15 gives the full predictive comparison with Normality-MinT, Bootstrap-MinT, Conformal-MinT, and HINT. Ratios are relative to MinT-shrink, while coverage gap is a difference on the coverage scale. Lower values are better in both cases. HINT exposes marginal quantiles rather than joint samples, so its energy score is unavailable. Table 16 reports point accuracy and omits the three probabilistic MinT variants because they reuse the MinT mean exactly.

Coverage must be read together with a proper score. HINT-BU has the smallest coverage gap, but its nCRPS and nMIS are 1.737 and 1.584 times those of MinT-shrink. Wider intervals can therefore improve coverage without improving the predictive distribution. In contrast, the RBIG-Gaussian hedge reduces the coverage gap while also lowering nCRPS and nMIS.

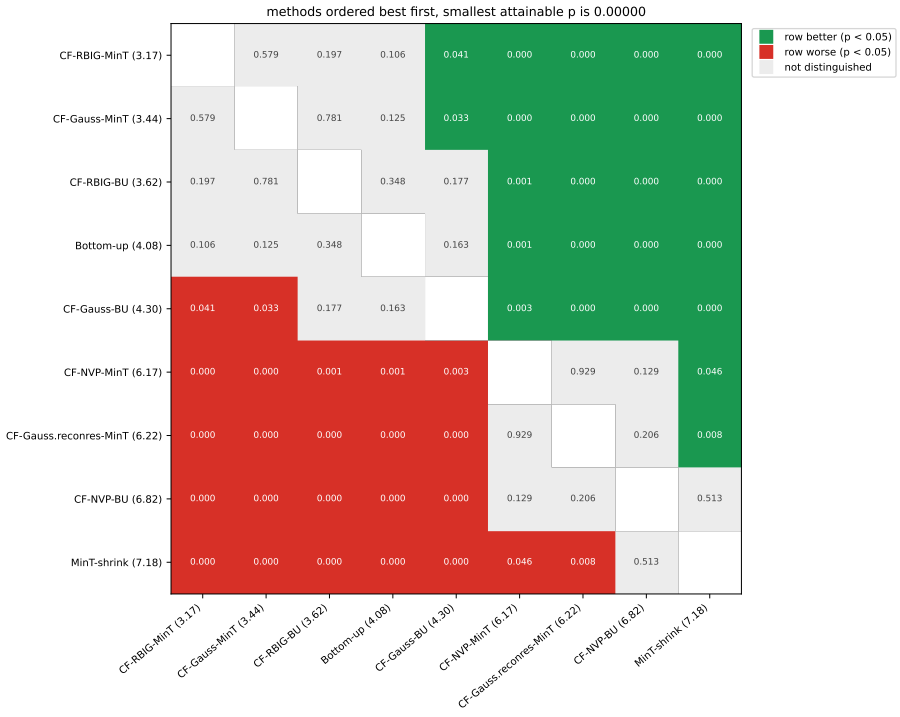
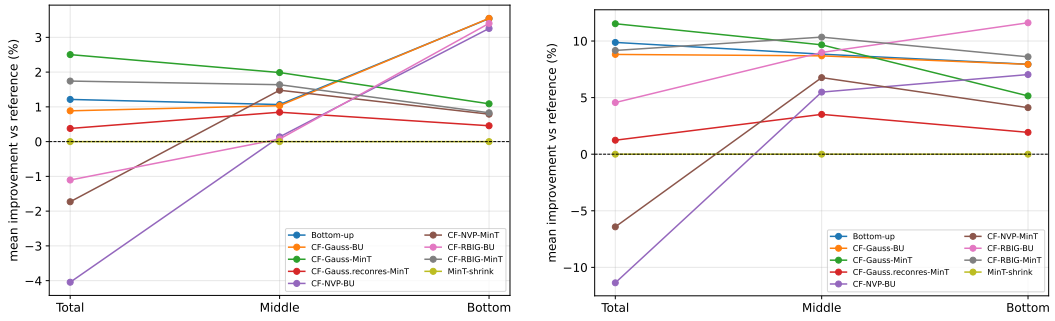


Figure 14: Pairwise dataset-clustered sign-flip decisions for anchor and residual-source nMIS over 87 origins. Green indicates that the row ranks better and red that it ranks worse.



(a) nCRPS. (b) nMIS.

Figure 15: nCRPS and nMIS improvement over MinT-shrink by level class for the anchor and residual-source ablation. Positive values favor the method.

A.12 Absolute per-dataset results

The main comparisons remove dataset difficulty through within-dataset ratios. Tables in this section instead report absolute metric values, allowing direct comparison with future results on the same benchmarks. The six columns contain the unreconciled forecast, three widely used reconcilers, and the two CoFlow configurations from Table 2. All six share the same base forecasts, so differences arise only from reconciliation and predictive modeling. HINT is excluded because it trains a different base forecaster. The probabilistic MinT variants are excluded because they repeat the MinT point forecast on relMAE and WMAPE. Each cell reports the mean and standard deviation over the same three origins and five seeds used in the main analysis.

Table 9: Point-head architecture ablation over 29 datasets. Entries are geometric-mean ratios to MinT-shrink ± the standard deviation of per-dataset log ratios, with lower values better. Bold marks the column best, and brackets give the two-sided Wilcoxon p against it.

Family	Method	relMAE / MinT	relMAE rank	WMAPE / MinT	WMAPE rank
No attention	CoFlow-Gauss-MLP	0.985 ± 0.044 (.027)	4.86 ± 2.26 (.027)	0.994 ± 0.021 (.005)	4.52 ± 2.13 (.224)
Cross-leaf attention	CoFlow-Gauss-LeafAttn	**0.967 ± 0.070**	4.03 ± 2.23 (.727)	0.996 ± 0.034 (.003)	5.14 ± 2.39 (.045)
Hierarchy graph	CoFlow-Gauss-TreeAttn	0.974 ± 0.057 (.325)	4.14 ± 1.30 (.321)	0.994 ± 0.033 (.003)	4.72 ± 1.79 (.068)
Hierarchy graph	CoFlow-Gauss-LeafTreeAttn	0.974 ± 0.052 (.831)	**3.79 ± 1.54**	0.994 ± 0.022 (.006)	4.48 ± 1.45 (.120)
Hierarchy graph	CoFlow-Gauss-FullAttn	0.975 ± 0.052 (.624)	4.14 ± 1.38 (.327)	0.995 ± 0.019 (.007)	4.34 ± 1.49 (.184)
Rolling supervision	CoFlow-Gauss-LeafAttn.ro	0.987 ± 0.107 (.733)	4.76 ± 3.28 (.200)	0.958 ± 0.129 (.275)	4.28 ± 3.25 (.173)
Constraint routing	CoFlow-Gauss-RouteAttn.L0	0.973 ± 0.102 (.831)	4.52 ± 2.98 (.245)	**0.915 ± 0.143**	**3.41 ± 2.77**
Reconciliation	MinT-shrink	1.000 ± 0.000 (.011)	5.76 ± 2.21 (<.001)	1.000 ± 0.000 (.004)	5.10 ± 2.30 (.032)



(a) relMAE. (b) WMAPE.

Figure 16: Average ranks for the point-head architecture ablation over 29 datasets. Lower ranks are better, and connected methods are not separated by the Nemenyi test at 0.05.

A.13 Per-dataset heatmaps

The heatmaps extend the absolute comparison to the full method set. Each cell averages three origins and five seeds and is printed to three significant figures. Colors are rescaled within each dataset between its fifth and ninety-fifth percentiles. This column-wise scale isolates method differences from dataset difficulty, which remains substantial even for normalized metrics. For example, nMIS spans 0.06 to more than 4000 across the matrix.

Rows are split into competing methods and CoFlow variants, with each block ordered from worst to best mean within-dataset rank and the unreconciled base placed first. Columns place the public benchmarks before the private corpus. Repeated row patterns make the separation in Section 3 visible: changing only the leaf law preserves the MinT point forecast, while changing only the point head preserves the predictive law.

A.14 Seed stability and computation

The selected variants are stable across the five seeds. Median seed coefficients of variation are 1.08% for relMAE and 1.43% for WMAPE with the balanced cross-leaf head, and 0.69% and 0.74% with CoFlow-Tweedie-RouteAttn.L50. Predictive variation is smaller still: median seed coefficients of variation for nCRPS are 0.13% for Tweedie, 0.20% for the RBIG hedge and 0.14% for the NVP hedge. HINT displays an order of magnitude more seed variation, 3.21% on relMAE and 7.34% on nCRPS, because each seed trains a complete neural base forecaster, while the CoFlow predictive variants reuse the shared base and vary only their post-hoc generative component.

Dense attention has quadratic complexity in the number of tokens, whereas sparse routing scales with the hierarchy edges. The predictive models remain post-hoc. Gaussian and Tweedie require no neural training, RBIG is gradient free, and NVP uses a leaf-space flow. These comparisons characterize the principal computational differences among the evaluated model classes.

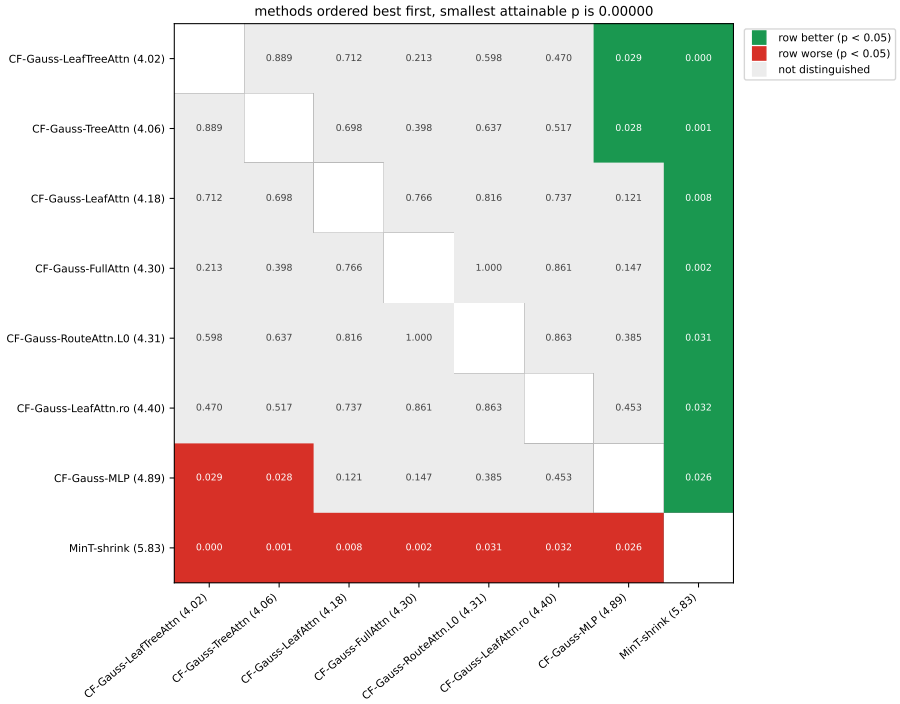


Figure 17: Pairwise dataset-clustered sign-flip decisions for architecture relMAE over 87 origins. Green indicates that the row ranks better and red that it ranks worse.

Table 10: Metric-aligned loss on the two attention heads over 29 datasets. Entries are geometric-mean ratios to MinT-shrink ± the standard deviation of per-dataset log ratios, with lower values better. Bold marks the column best, and brackets give the two-sided Wilcoxon p against it.

Family	Method	relMAE / MinT	relMAE rank	WMAPE / MinT	WMAPE rank
Cross-leaf attention	CoFlow-Gauss-LeafAttn	0.967 ± 0.070 (<.001)	4.86 ± 1.73 (<.001)	0.996 ± 0.034 (.284)	4.66 ± 2.09 (.025)
Cross-leaf attention	CoFlow-Gauss-LeafAttn.L25	0.952 ± 0.081 (<.001)	3.24 ± 1.55 (<.001)	0.996 ± 0.031 (.183)	4.55 ± 1.82 (.028)
Cross-leaf attention	CoFlow-Gauss-LeafAttn.L50	**0.934 ± 0.096**	**1.86 ± 1.33**	**0.992 ± 0.042**	4.17 ± 2.05 (.130)
Every node pair	CoFlow-Gauss-FullAttn	0.975 ± 0.052 (<.001)	5.07 ± 1.28 (<.001)	0.995 ± 0.019 (.393)	3.62 ± 1.50 (.377)
Every node pair	CoFlow-Gauss-FullAttn.L25	0.965 ± 0.066 (<.001)	4.28 ± 1.41 (<.001)	0.994 ± 0.018 (.314)	**3.28 ± 1.65**
Every node pair	CoFlow-Gauss-FullAttn.L50	0.955 ± 0.074 (.006)	3.00 ± 1.60 (.006)	0.995 ± 0.022 (.247)	3.48 ± 1.96 (.582)
Reconciliation	MinT-shrink	1.000 ± 0.000 (<.001)	5.69 ± 2.07 (<.001)	1.000 ± 0.000 (.654)	4.24 ± 2.54 (.064)

Table 11: Routing loss frontier over 29 datasets. Entries are geometric-mean ratios to MinT-shrink ± the standard deviation of per-dataset log ratios, with lower values better. Bold marks the column best, and brackets give the two-sided Wilcoxon p against it.

Family	Method	relMAE / MinT	relMAE rank	WMAPE / MinT	WMAPE rank
Constraint routing	CoFlow-Gauss-RouteAttn.L0	0.973 ± 0.102 (<.001)	5.10 ± 0.86 (<.001)	**0.915 ± 0.143**	3.52 ± 1.68 (.233)
Constraint routing	CoFlow-Gauss-RouteAttn.L25	0.922 ± 0.225 (<.001)	4.10 ± 1.08 (<.001)	0.920 ± 0.137 (.247)	3.03 ± 1.38 (.686)
Constraint routing	CoFlow-Gauss-RouteAttn.L50	0.899 ± 0.285 (<.001)	3.28 ± 1.28 (<.001)	0.932 ± 0.114 (.831)	**2.97 ± 1.57**
Constraint routing	CoFlow-Gauss-RouteAttn.L75	**0.899 ± 0.248**	**1.97 ± 0.73**	0.949 ± 0.088 (.670)	3.38 ± 1.24 (.380)
Constraint routing	CoFlow-Gauss-RouteAttn.L100	0.906 ± 0.192 (.247)	2.00 ± 1.10 (.972)	0.968 ± 0.085 (.230)	3.86 ± 1.83 (.093)
Reconciliation	MinT-shrink	1.000 ± 0.000 (<.001)	4.55 ± 1.96 (<.001)	1.000 ± 0.000 (.004)	4.24 ± 2.20 (.044)

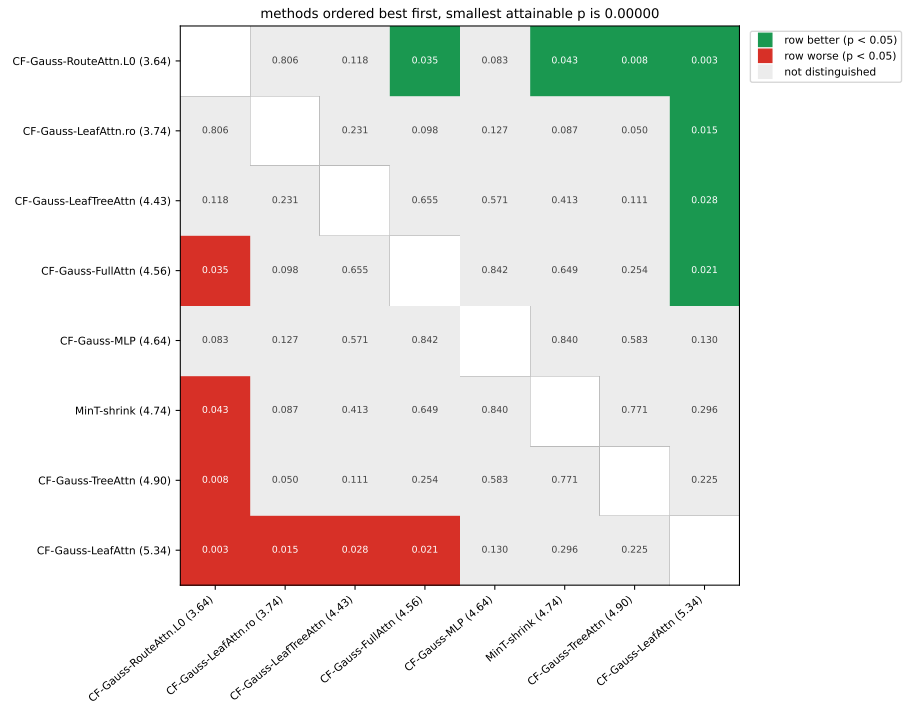
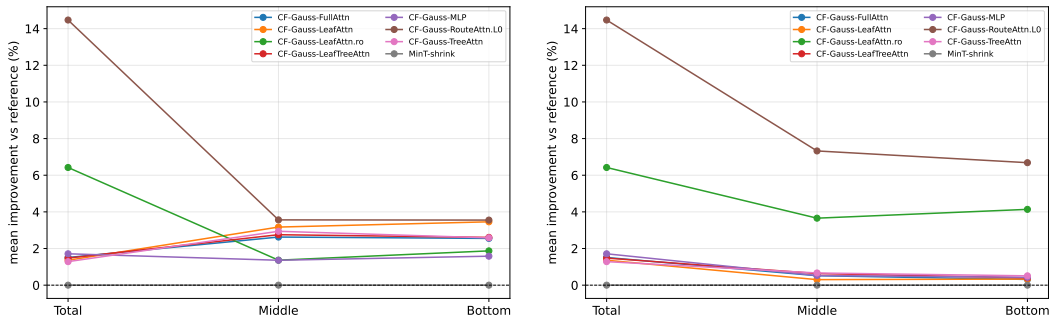


Figure 18: Pairwise dataset-clustered sign-flip decisions for architecture WMAPE over 87 origins. Green indicates that the row ranks better and red that it ranks worse.



(a) relMAE. (b) WMAPE.

Figure 19: relMAE and WMAPE improvement over MinT-shrink by level class for the point-head architecture ablation. Positive values favor the method.



(a) relMAE. (b) WMAPE.

Figure 20: Average ranks for the attention loss-weight ablation over 29 datasets. Lower ranks are better, and connected methods are not separated by the Nemenyi test at 0.05.

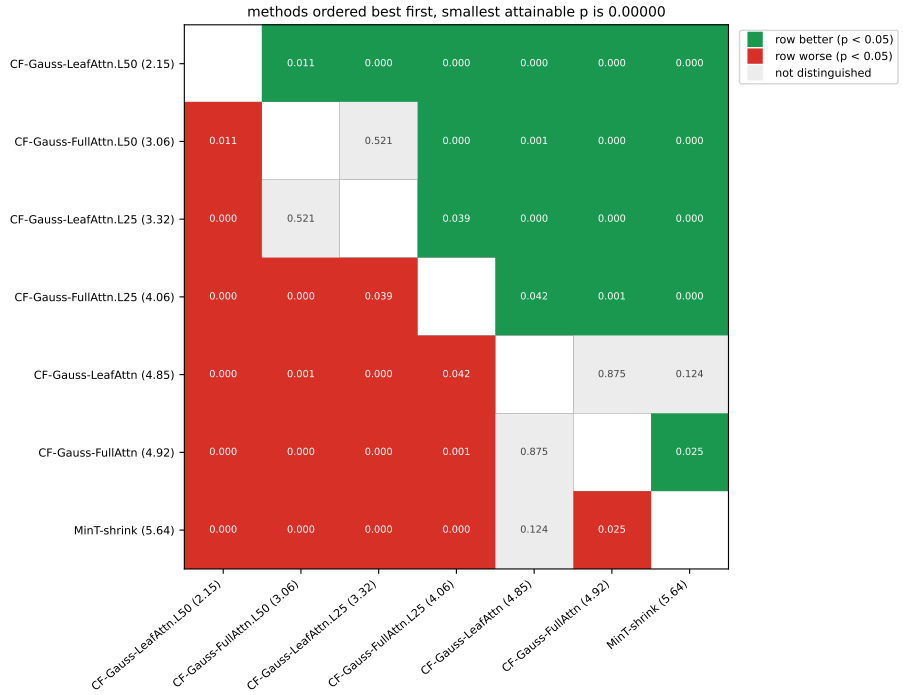


Figure 21: Pairwise dataset-clustered sign-flip decisions for attention-loss relMAE over 87 origins. Green indicates that the row ranks better and red that it ranks worse.

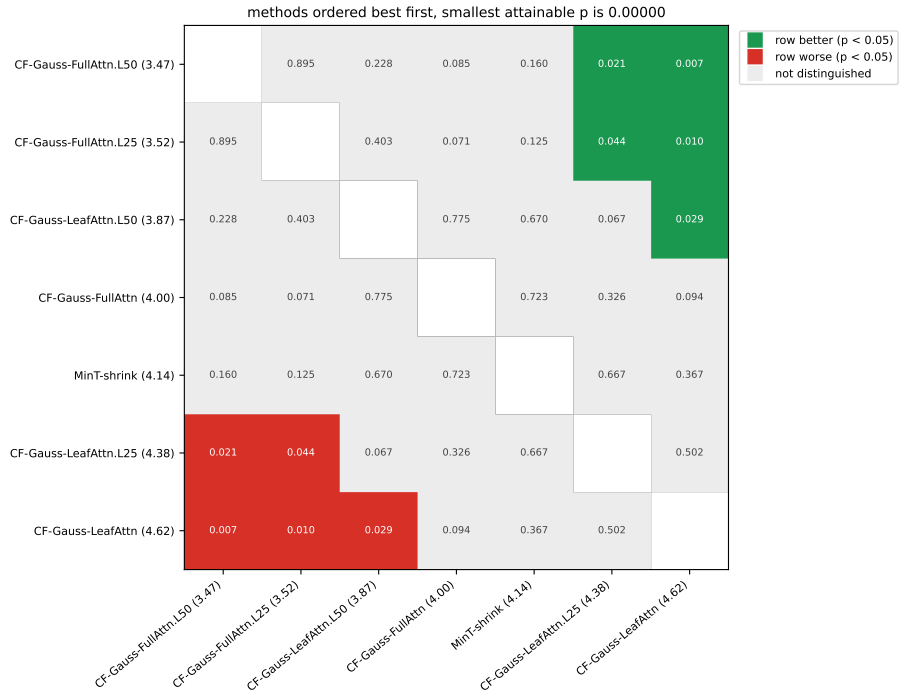
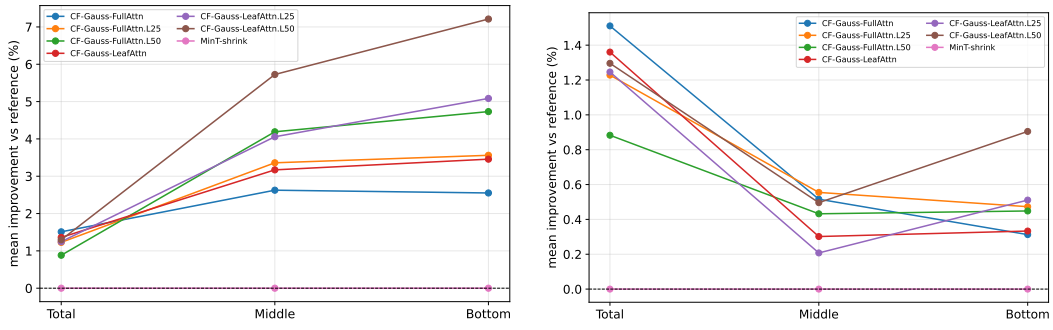


Figure 22: Pairwise dataset-clustered sign-flip decisions for attention-loss WMAPE over 87 origins. Green indicates that the row ranks better and red that it ranks worse.



(a) relMAE. (b) WMAPE.

Figure 23: relMAE and WMAPE improvement over MinT-shrink by level class for the attention loss-weight ablation. Positive values favor the method.



(a) relMAE. (b) WMAPE.

Figure 24: Average ranks along the routing loss frontier over 29 datasets. Lower ranks are better, and connected methods are not separated by the Nemenyi test at 0.05.

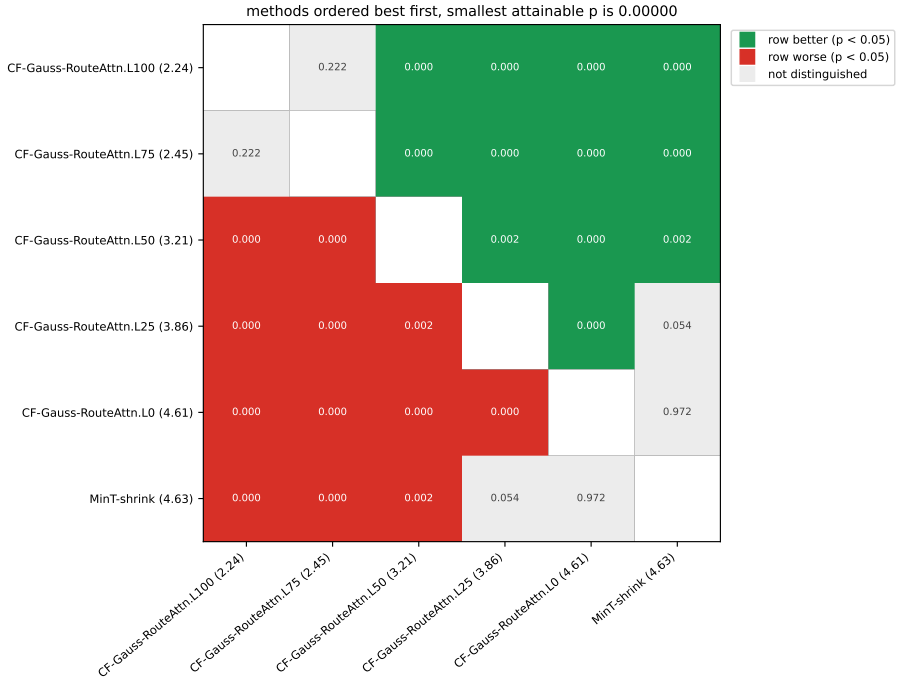


Figure 25: Pairwise dataset-clustered sign-flip decisions for routing-frontier relMAE over 87 origins. Green indicates that the row ranks better and red that it ranks worse.

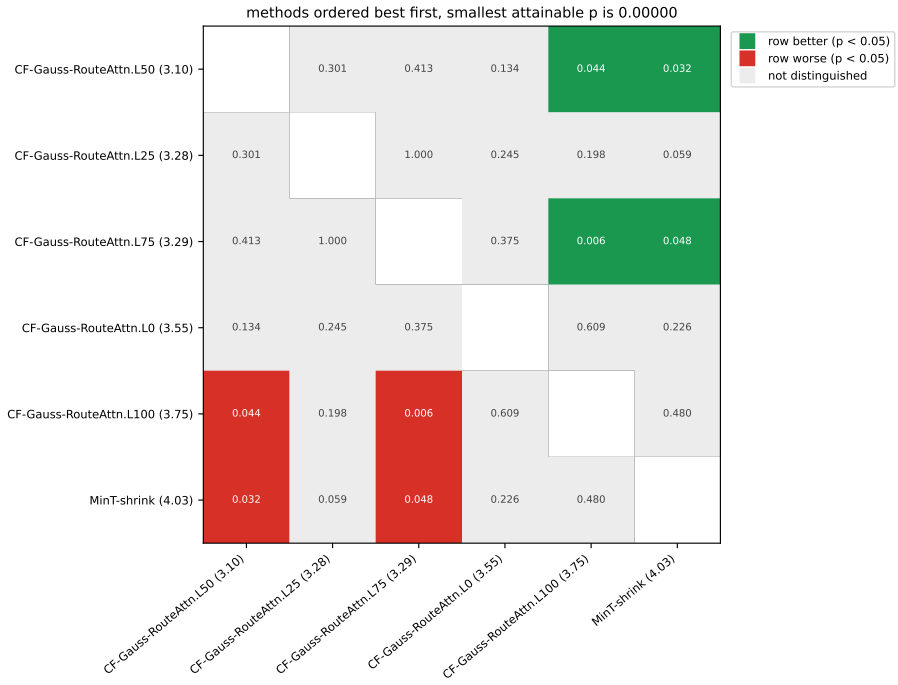
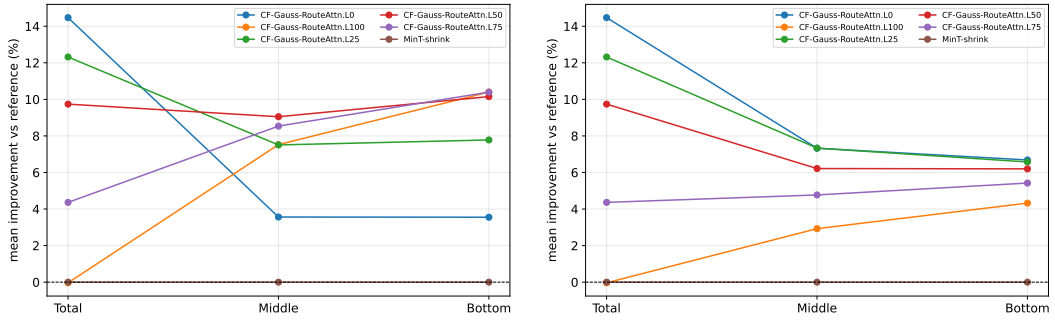


Figure 26: Pairwise dataset-clustered sign-flip decisions for routing-frontier WMAPE over 87 origins. Green indicates that the row ranks better and red that it ranks worse.



(a) relMAE. (b) WMAPE.

Figure 27: relMAE and WMAPE improvement over MinT-shrink by level class along the routing loss frontier. Positive values favor the method.

Table 12: Separation of point heads and leaf laws. The same five point heads are paired with Gaussian and RBIG leaf laws. Entries are geometric-mean ratios to MinT-shrink over 29 datasets ± the standard deviation of per-dataset log ratios. Lower is better. Bold marks the column best, and brackets give the two-sided Wilcoxon p against it.

Family	Method	relMAE / MinT	relMAE rank	nCRPS / MinT	nCRPS rank	nMIS / MinT	nMIS rank
Gaussian leaf law	CoFlow-Gauss-MLP	0.985 ± 0.044 (.027)	7.05 ± 3.22 (**.014**)	**0.983 ± 0.013**	**5.48 ± 2.63**	0.925 ± 0.048 (.080)	6.45 ± 2.35 (.450)
Gaussian leaf law	CoFlow-Gauss-LeafAttn	**0.967 ± 0.070**	5.31 ± 3.78 (.634)	**0.983 ± 0.013** (1.000)	**5.48 ± 2.63** (1.000)	0.925 ± 0.048 (.080)	6.45 ± 2.35 (.450)
Gaussian leaf law	CoFlow-Gauss-TreeAttn	0.974 ± 0.057 (.325)	5.59 ± 2.61 (.282)	**0.983 ± 0.013** (1.000)	**5.48 ± 2.63** (1.000)	0.925 ± 0.048 (.080)	6.45 ± 2.35 (.450)
Gaussian leaf law	CoFlow-Gauss-LeafTreeAttn	0.974 ± 0.052 (.831)	**4.78 ± 2.64**	**0.983 ± 0.013** (1.000)	**5.48 ± 2.63** (1.000)	0.925 ± 0.048 (.080)	6.45 ± 2.35 (.450)
Gaussian leaf law	CoFlow-Gauss-FullAttn	0.975 ± 0.052 (.624)	5.64 ± 2.56 (.168)	**0.983 ± 0.013** (1.000)	**5.48 ± 2.63** (1.000)	0.925 ± 0.048 (.080)	6.45 ± 2.35 (.450)
RBIG leaf law	CoFlow-RBIG-MLP	0.985 ± 0.044 (.027)	7.05 ± 3.22 (**.014**)	0.987 ± 0.025 (.495)	5.83 ± 2.74 (.337)	**0.914 ± 0.131**	**4.72 ± 2.63**
RBIG leaf law	CoFlow-RBIG-LeafAttn	0.968 ± 0.068 (.200)	5.38 ± 3.63 (.530)	0.987 ± 0.025 (.495)	5.83 ± 2.74 (.337)	**0.914 ± 0.131** (1.000)	**4.72 ± 2.63** (1.000)
RBIG leaf law	CoFlow-RBIG-TreeAttn	0.973 ± 0.056 (.393)	5.59 ± 2.20 (.287)	0.987 ± 0.025 (.495)	5.83 ± 2.74 (.337)	**0.914 ± 0.131** (1.000)	**4.72 ± 2.63** (1.000)
RBIG leaf law	CoFlow-RBIG-LeafTreeAttn	0.974 ± 0.051 (.579)	5.40 ± 2.74 (.068)	0.987 ± 0.025 (.495)	5.83 ± 2.74 (.337)	**0.914 ± 0.131** (1.000)	**4.72 ± 2.63** (1.000)
RBIG leaf law	CoFlow-RBIG-FullAttn	0.975 ± 0.053 (.508)	5.81 ± 2.68 (.192)	0.987 ± 0.025 (.495)	5.83 ± 2.74 (.337)	**0.914 ± 0.131** (1.000)	**4.72 ± 2.63** (1.000)
Reconciliation	MinT-shrink	1.000 ± 0.000 (**.011**)	8.41 ± 3.65 (<**.001**)	1.000 ± 0.000 (<**.001**)	9.45 ± 2.71 (<**.001**)	1.000 ± 0.000 (<**.001**)	10.14 ± 1.92 (<**.001**)

Table 13: Observed combinations of leaf laws and point heads over 29 datasets. Entries are geometric-mean ratios to MinT-shrink ± the standard deviation of per-dataset log ratios, with lower values better. Bold marks the column best, and brackets give the two-sided Wilcoxon p against it.

Family	Method	relMAE / MinT	relMAE rank	WMAPE / MinT	WMAPE rank	nCRPS / MinT	nCRPS rank	nMIS / MinT	nMIS rank
Leaf law alone	CoFlow-Gauss-MinT	1.000 ± 0.000 (<.001)	13.10 ± 4.25 (<.001)	1.000 ± 0.000 (.006)	10.97 ± 4.24 (.024)	0.983 ± 0.013 (<.001)	11.93 ± 3.95 (<.001)	0.925 ± 0.048 (<.001)	12.28 ± 3.64 (<.001)
Leaf law alone	CoFlow-Tweedie-MinT	1.000 ± 0.000 (<.001)	13.10 ± 4.25 (<.001)	1.000 ± 0.000 (.006)	10.97 ± 4.24 (.024)	0.938 ± 0.074 (.808)	**4.66 ± 4.13**	0.840 ± 0.143 (.316)	5.66 ± 5.25 (.175)
Leaf law alone	CoFlow-RBIG+Gauss-MinT	1.000 ± 0.000 (<.001)	13.10 ± 4.25 (<.001)	1.000 ± 0.000 (.006)	10.97 ± 4.24 (.024)	0.975 ± 0.016 (<.001)	9.07 ± 3.06 (.005)	0.901 ± 0.079 (.003)	8.17 ± 3.30 (.072)
Leaf law alone	CoFlow-RBIG+NVP-MinT	1.000 ± 0.000 (<.001)	13.10 ± 4.25 (<.001)	1.000 ± 0.000 (.006)	10.97 ± 4.24 (.024)	0.981 ± 0.019 (<.001)	12.90 ± 4.05 (<.001)	0.924 ± 0.089 (<.001)	12.67 ± 3.88 (<.001)
Cross-leaf head	CoFlow-Gauss-LeafAttn.L50	0.934 ± 0.096 (.882)	7.17 ± 4.49 (.583)	0.992 ± 0.042 (.002)	10.79 ± 5.33 (.046)	0.983 ± 0.013 (<.001)	11.93 ± 3.95 (<.001)	0.925 ± 0.048 (<.001)	12.28 ± 3.64 (<.001)
Cross-leaf head	CoFlow-Tweedie-LeafAttn.L50	0.930 ± 0.100 (.848)	7.48 ± 5.22 (.051)	0.994 ± 0.036 (.001)	11.03 ± 4.90 (.040)	**0.938 ± 0.074** (.317)	4.74 ± 4.09 (.827)	**0.840 ± 0.142**	**5.02 ± 3.95**
Cross-leaf head	CoFlow-RBIG+Gauss-LeafAttn.L50	0.929 ± 0.100 (.983)	**6.62 ± 4.80**	0.993 ± 0.040 (.001)	10.72 ± 5.23 (.089)	0.975 ± 0.016 (<.001)	8.93 ± 2.92 (.003)	0.901 ± 0.079 (.003)	8.21 ± 3.34 (.067)
Node-pair head	CoFlow-Gauss-FullAttn.L50	0.955 ± 0.074 (.149)	8.66 ± 3.28 (.022)	0.995 ± 0.022 (.003)	9.86 ± 4.85 (.190)	0.983 ± 0.013 (<.001)	11.93 ± 3.95 (<.001)	0.925 ± 0.048 (<.001)	12.28 ± 3.64 (<.001)
Node-pair head	CoFlow-Tweedie-FullAttn.L50	0.955 ± 0.073 (.132)	8.78 ± 3.82 (.019)	0.994 ± 0.024 (.002)	10.05 ± 4.69 (.144)	**0.938 ± 0.074** (.317)	4.79 ± 4.22 (.683)	0.840 ± 0.142 (.317)	5.10 ± 4.12 (.317)
Node-pair head	CoFlow-RBIG+Gauss-FullAttn.L50	0.955 ± 0.071 (.121)	8.33 ± 3.97 (.065)	0.995 ± 0.026 (.003)	10.02 ± 4.64 (.265)	0.975 ± 0.016 (<.001)	8.93 ± 2.92 (.003)	0.901 ± 0.079 (.003)	8.21 ± 3.34 (.067)
Routing head, L25	CoFlow-Gauss-RouteAttn.L25	0.922 ± 0.225 (<.001)	9.34 ± 4.48 (.076)	**0.920 ± 0.137**	**7.41 ± 4.79**	0.983 ± 0.013 (<.001)	11.93 ± 3.95 (<.001)	0.925 ± 0.048 (<.001)	12.28 ± 3.64 (<.001)
Routing head, L25	CoFlow-Tweedie-RouteAttn.L25	0.922 ± 0.228 (<.001)	9.67 ± 4.47 (.054)	0.920 ± 0.137 (.522)	7.55 ± 5.24 (.796)	**0.938 ± 0.074** (.317)	4.74 ± 4.09 (.827)	**0.840 ± 0.142** (1.000)	**5.02 ± 3.95** (1.000)
Routing head, L25	CoFlow-RBIG+Gauss-RouteAttn.L25	0.922 ± 0.228 (<.001)	9.28 ± 4.82 (.153)	0.920 ± 0.137 (.782)	7.64 ± 4.91 (.394)	0.975 ± 0.016 (<.001)	9.03 ± 2.88 (.002)	0.901 ± 0.079 (.003)	8.31 ± 3.17 (.056)
Routing head, L25	CoFlow-RBIG+NVP-RouteAttn.L25	0.921 ± 0.235 (<.001)	9.64 ± 4.80 (.089)	0.921 ± 0.137 (.417)	7.53 ± 5.29 (.764)	0.981 ± 0.019 (<.001)	12.86 ± 4.01 (<.001)	0.924 ± 0.089 (<.001)	12.60 ± 4.03 (<.001)
Routing head, L50	CoFlow-Gauss-RouteAttn.L50	**0.899 ± 0.285**	6.83 ± 5.51 (.820)	0.932 ± 0.114 (.639)	8.17 ± 5.86 (.288)	0.983 ± 0.013 (<.001)	11.93 ± 3.95 (<.001)	0.925 ± 0.048 (<.001)	12.28 ± 3.64 (<.001)
Routing head, L50	CoFlow-Tweedie-RouteAttn.L50	0.902 ± 0.277 (.898)	6.78 ± 5.56 (.991)	0.932 ± 0.115 (.481)	7.48 ± 6.14 (.664)	**0.938 ± 0.074**	4.69 ± 3.99 (.984)	0.840 ± 0.142 (.317)	5.07 ± 4.05 (.317)
Routing head, L50	CoFlow-RBIG+Gauss-RouteAttn.L50	0.900 ± 0.285 (.522)	6.91 ± 5.54 (.914)	0.932 ± 0.115 (.565)	7.90 ± 6.07 (.870)	0.975 ± 0.016 (<.001)	9.03 ± 2.88 (.002)	0.901 ± 0.079 (.003)	8.31 ± 3.17 (.056)
Reconciliation	MinT-shrink	1.000 ± 0.000 (<.001)	13.10 ± 4.25 (<.001)	1.000 ± 0.000 (.006)	10.97 ± 4.24 (.024)	1.000 ± 0.000 (<.001)	16.97 ± 2.53 (<.001)	1.000 ± 0.000 (<.001)	17.28 ± 2.00 (<.001)

Table 14: Standard reconcilers over 29 datasets. Entries are geometric-mean ratios to MinT-shrink ± the standard deviation of per-dataset log ratios, with lower values better. Bold marks the column best, and brackets give the two-sided Wilcoxon p against it.

Family	Method	relMAE / MinT	relMAE rank	WMAPE / MinT	WMAPE rank	nCRPS / MinT	nCRPS rank	nMIS / MinT	nMIS rank
Unreconciled	Base	0.945 ± 0.338 (<.001)	3.72 ± 1.07 (<.001)	1.045 ± 0.241 (.126)	4.00 ± 1.98 (.008)	1.013 ± 0.250 (.798)	3.38 ± 2.03 (.534)	0.922 ± 0.285 (.115)	**2.34 ± 1.70**
Single-level	Bottom-up	0.946 ± 0.338 (.002)	3.66 ± 1.76 (<.001)	1.005 ± 0.103 (.059)	3.55 ± 2.20 (.046)	0.984 ± 0.114 (.949)	3.31 ± 2.14 (.683)	**0.922 ± 0.152**	2.86 ± 1.83 (.434)
Single-level	Middle-out	0.996 ± 0.393 (.183)	3.97 ± 2.04 (.003)	1.103 ± 0.211 (<.001)	5.03 ± 1.70 (<.001)	1.118 ± 0.249 (<.001)	5.24 ± 1.84 (<.001)	1.172 ± 0.372 (<.001)	5.69 ± 1.85 (<.001)
Least squares	OLS	1.965 ± 0.644 (<.001)	6.66 ± 1.26 (<.001)	1.197 ± 0.347 (<.001)	6.00 ± 1.69 (<.001)	1.156 ± 0.355 (<.001)	5.62 ± 1.90 (<.001)	1.040 ± 0.389 (.110)	4.72 ± 2.10 (<.001)
Least squares	WLS-struct	1.313 ± 0.365 (<.001)	5.03 ± 1.68 (<.001)	1.042 ± 0.181 (.009)	3.79 ± 1.37 (.014)	1.026 ± 0.190 (.092)	3.79 ± 1.40 (.119)	0.981 ± 0.222 (.137)	3.45 ± 1.33 (.017)
Least squares	WLS-var	**0.935 ± 0.338**	**2.28 ± 1.03**	**0.983 ± 0.070**	**2.66 ± 1.29**	**0.976 ± 0.079**	**3.07 ± 1.39**	0.957 ± 0.108 (.002)	4.14 ± 1.46 (.003)
Minimum trace	MinT-sample	6.829 ± 1.697 (<.001)	7.79 ± 0.82 (<.001)	3.083 ± 1.427 (<.001)	7.55 ± 1.50 (<.001)	3.728 ± 1.581 (<.001)	7.55 ± 1.50 (<.001)	3.365 ± 1.641 (<.001)	7.48 ± 1.50 (<.001)
Minimum trace	MinT-shrink	1.000 ± 0.000 (.865)	2.90 ± 1.78 (.184)	1.000 ± 0.000 (.115)	3.41 ± 2.11 (.106)	1.000 ± 0.000 (.015)	4.03 ± 2.13 (.027)	1.000 ± 0.000 (<.001)	5.31 ± 1.73 (<.001)

Table 15: External predictive comparison over 29 datasets. nCRPS, nMIS, and Energy are ratios to MinT-shrink. CovGap is the mean signed difference from MinT-shrink, where negative values are closer to nominal coverage. Lower values and ranks are better, and bold marks the column best.

Method	nCRPS	nMIS	Energy	CovGap	Rank
CoFlow-RBIG+Gauss-RouteAttn.L25	0.975	0.898	0.972	-0.072	**3.937**
CoFlow-Tweedie-RouteAttn.L50	**0.940**	**0.842**	**0.939**	-0.017	4.187
Bottom-up	0.978	0.915	0.984	-0.063	5.535
HINT-BU	1.636	1.482	N/A	**-0.146**	5.536
Base	1.021	0.935	1.094	-0.062	5.717
WLS-struct	1.024	0.980	0.994	-0.018	6.200
HINT-WLS	1.621	1.512	N/A	-0.126	6.241
WLS-var	0.971	0.951	0.975	-0.026	6.641
MinT-shrink	1.000	1.000	1.000	0.000	7.458
Bootstrap-MinT	1.004	0.994	0.999	0.016	7.757
Normality-MinT	1.001	1.001	1.000	0.001	7.869
Conformal-MinT	1.005	0.994	1.001	0.018	8.116

Table 16: External point comparison over 29 datasets. Scores are geometric-mean ratios to MinT-shrink and ranks are averaged within dataset. Lower values and ranks are better, and bold marks the column best.

Method	relMAE	WMAPE	MASE	Rank
CoFlow-Tweedie-RouteAttn.L50	**0.903**	0.932	0.916	**3.111**
CoFlow-RBIG+Gauss-RouteAttn.L25	0.923	**0.922**	**0.911**	3.525
WLS-var	0.943	0.977	0.920	4.241
MinT-shrink	1.000	1.000	1.000	4.536
Bottom-up	0.953	0.999	0.966	5.293
Base	0.953	1.052	0.984	5.722
HINT-BU	1.223	1.870	1.497	5.870
HINT-WLS	1.981	1.791	2.670	6.226
WLS-struct	1.297	1.039	1.514	6.475

Table 17: Absolute relMAE by dataset. Cells report mean ± standard deviation over three origins and five seeds. Public benchmarks appear above the rule and anonymized private datasets below it. Lower is better and bold marks the row minimum. The Base column is one by definition. COFLOW is abbreviated CF-.

Dataset	Base	Bottom-up	WLS-var	MinT-shrink	CF-Tweedie-RouteAttn.L50	CF-RBIG+NVP-RouteAttn.L25
Labour	1.0000 ± 0.0000	0.8534 ± 0.0161	**0.8423 ± 0.0566**	0.8836 ± 0.0598	0.8957 ± 0.0577	0.8957 ± 0.0506
M5	1.0000 ± 0.0000	1.0015 ± 0.0001	0.9947 ± 0.0004	0.9727 ± 0.0178	**0.9375 ± 0.0105**	0.9441 ± 0.0120
TourismLarge	1.0000 ± 0.0000	1.0078 ± 0.0038	0.9767 ± 0.0029	0.9777 ± 0.0103	**0.9426 ± 0.0108**	0.9701 ± 0.0052
TourismSmall	1.0000 ± 0.0000	**0.9643 ± 0.0217**	0.9661 ± 0.0211	0.9869 ± 0.0412	0.9802 ± 0.0509	0.9843 ± 0.0483
Traffic	1.0000 ± 0.0000	1.0135 ± 0.0076	0.9585 ± 0.0144	**0.9560 ± 0.0473**	0.9608 ± 0.0573	0.9590 ± 0.0604
Wiki2	1.000 ± 0.000	1.118 ± 0.031	1.199 ± 0.036	1.211 ± 0.018	0.998 ± 0.044	**0.985 ± 0.074**
P01	1.000 ± 0.000	**0.995 ± 0.003**	1.006 ± 0.004	1.011 ± 0.006	0.998 ± 0.030	1.010 ± 0.020
P02	1.0000 ± 0.0000	1.0000 ± 0.0019	0.9967 ± 0.0004	0.9604 ± 0.0119	**0.8543 ± 0.0648**	0.8556 ± 0.0813
P03	1.0000 ± 0.0000	1.0212 ± 0.0104	0.9990 ± 0.0026	0.8878 ± 0.0065	**0.8215 ± 0.0299**	0.8293 ± 0.0394
P04	1.0000 ± 0.0000	1.0097 ± 0.0120	0.9762 ± 0.0077	1.0014 ± 0.0032	**0.9203 ± 0.0599**	0.9436 ± 0.0773
P05	1.0000 ± 0.0000	1.0047 ± 0.0007	0.9996 ± 0.0008	0.9890 ± 0.0051	**0.6836 ± 0.0165**	0.7001 ± 0.0197
P06	1.000 ± 0.000	0.998 ± 0.002	**0.997 ± 0.002**	1.010 ± 0.008	1.006 ± 0.009	1.006 ± 0.008
P07	1.000 ± 0.000	1.000 ± 0.004	**0.995 ± 0.003**	1.004 ± 0.010	1.033 ± 0.056	1.060 ± 0.054
P08	1.000 ± 0.000	0.999 ± 0.016	**0.997 ± 0.019**	1.522 ± 0.730	1.945 ± 1.070	1.946 ± 0.990
P09	1.000 ± 0.000	1.002 ± 0.003	0.998 ± 0.001	**0.996 ± 0.001**	1.012 ± 0.007	1.008 ± 0.009
P10	1.0000 ± 0.0000	1.0002 ± 0.0019	0.9972 ± 0.0010	0.9811 ± 0.0141	**0.8730 ± 0.0079**	0.8794 ± 0.0101
P11	1.000 ± 0.000	**0.976 ± 0.005**	1.007 ± 0.011	1.352 ± 0.274	1.006 ± 0.055	1.044 ± 0.056
P12	1.0000 ± 0.0000	1.0257 ± 0.0026	0.9698 ± 0.0030	0.8293 ± 0.0173	**0.8157 ± 0.0227**	0.8184 ± 0.0179
P13	1.000 ± 0.000	1.003 ± 0.019	**0.963 ± 0.013**	5.417 ± 2.214	1.300 ± 0.322	1.683 ± 0.564
P14	1.0000 ± 0.0000	0.9954 ± 0.0060	0.9915 ± 0.0055	0.9923 ± 0.0211	0.9370 ± 0.0287	**0.9362 ± 0.0368**
P15	1.0000 ± 0.0000	1.0124 ± 0.0030	0.9908 ± 0.0024	0.8892 ± 0.0116	**0.8606 ± 0.0486**	0.9076 ± 0.0262
P16	1.0000 ± 0.0000	0.9937 ± 0.0024	0.9969 ± 0.0025	0.9876 ± 0.0033	**0.9640 ± 0.0174**	0.9654 ± 0.0162
P17	1.0000 ± 0.0000	1.0148 ± 0.0284	0.9932 ± 0.0073	0.9957 ± 0.0082	**0.9322 ± 0.0249**	0.9361 ± 0.0284
P18	1.0000 ± 0.0000	**0.9973 ± 0.0022**	0.9974 ± 0.0036	0.9976 ± 0.0063	1.0270 ± 0.0319	1.0341 ± 0.0326
P19	1.0000 ± 0.0000	0.9996 ± 0.0006	0.9981 ± 0.0019	**0.9711 ± 0.0237**	0.9734 ± 0.0529	1.0178 ± 0.1124
P20	1.0000 ± 0.0000	0.9998 ± 0.0002	0.9995 ± 0.0002	1.0084 ± 0.0175	**0.9119 ± 0.0255**	0.9137 ± 0.0230
P21	1.0000 ± 0.0000	1.0006 ± 0.0037	0.9949 ± 0.0009	0.9939 ± 0.0079	**0.8504 ± 0.0331**	0.8528 ± 0.0305
P22	1.000 ± 0.000	0.995 ± 0.005	**0.982 ± 0.005**	1.011 ± 0.020	1.014 ± 0.023	1.007 ± 0.018
P23	1.0000 ± 0.0000	1.0338 ± 0.0121	0.9486 ± 0.0036	0.7994 ± 0.0121	**0.7134 ± 0.0277**	0.7690 ± 0.0490

Table 18: Absolute WMAPE by dataset. Cells report mean ± standard deviation over three origins and five seeds. Public benchmarks appear above the rule and anonymized private datasets below it. Lower is better and bold marks the row minimum. COFLOW is abbreviated CF-.

Dataset	Base	Bottom-up	WLS-var	MinT-shrink	CF-Tweedie-RouteAttn.L50	CF-RBIG+NVP-RouteAttn.L25
Labour	0.03442 ± 0.00234	**0.01026 ± 0.00169**	0.01053 ± 0.00244	0.01104 ± 0.00273	0.01124 ± 0.00266	0.01132 ± 0.00273
M5	0.1525 ± 0.0093	0.1785 ± 0.0069	0.1687 ± 0.0062	0.1450 ± 0.0117	**0.1376 ± 0.0026**	0.1377 ± 0.0032
TourismLarge	0.1939 ± 0.0131	0.1799 ± 0.0144	0.1761 ± 0.0157	**0.1752 ± 0.0160**	0.1838 ± 0.0200	0.1775 ± 0.0166
TourismSmall	0.1427 ± 0.0129	**0.1044 ± 0.0167**	0.1086 ± 0.0174	0.1111 ± 0.0169	0.1126 ± 0.0124	0.1140 ± 0.0118
Traffic	**0.05634 ± 0.01070**	0.06762 ± 0.01820	0.05811 ± 0.01688	0.05688 ± 0.01437	0.05914 ± 0.01581	0.05787 ± 0.01567
Wiki2	0.6154 ± 0.0982	0.8626 ± 0.2699	0.6545 ± 0.1300	0.6589 ± 0.1232	0.5678 ± 0.1456	**0.4502 ± 0.0474**
P01	0.5740 ± 0.1624	0.5781 ± 0.1687	0.5834 ± 0.1745	0.5849 ± 0.1747	**0.5533 ± 0.1368**	0.5588 ± 0.1421
P02	0.7244 ± 0.1269	0.7162 ± 0.1738	0.7113 ± 0.1783	0.7040 ± 0.1836	0.6759 ± 0.1376	**0.6725 ± 0.1230**
P03	0.3273 ± 0.0262	0.2999 ± 0.0234	0.2979 ± 0.0249	0.2986 ± 0.0263	**0.2933 ± 0.0276**	0.2935 ± 0.0275
P04	0.4342 ± 0.0338	0.4267 ± 0.0401	0.4193 ± 0.0336	0.4301 ± 0.0504	**0.4075 ± 0.0345**	0.4140 ± 0.0289
P05	0.9320 ± 0.0855	0.9432 ± 0.0792	0.9474 ± 0.0872	0.9562 ± 0.0928	0.7489 ± 0.0604	**0.7488 ± 0.0572**
P06	0.2909 ± 0.0049	**0.2809 ± 0.0102**	0.2812 ± 0.0096	0.2883 ± 0.0107	0.2863 ± 0.0066	0.2869 ± 0.0068
P07	**0.6572 ± 0.2060**	0.7529 ± 0.2936	0.7268 ± 0.2827	0.7224 ± 0.2843	0.6914 ± 0.2614	0.6890 ± 0.2502
P08	**2.243 ± 2.011**	2.969 ± 2.962	2.884 ± 2.892	3.102 ± 3.062	3.181 ± 3.199	3.175 ± 3.177
P09	0.5080 ± 0.0356	0.6129 ± 0.0656	0.5923 ± 0.0556	0.5928 ± 0.0556	0.4029 ± 0.0807	**0.3703 ± 0.0187**
P10	0.5367 ± 0.0276	0.5304 ± 0.0246	0.5310 ± 0.0275	0.5383 ± 0.0251	**0.4555 ± 0.0084**	0.4583 ± 0.0091
P11	0.3357 ± 0.0580	**0.2932 ± 0.0425**	0.3079 ± 0.0450	0.3907 ± 0.0858	0.3173 ± 0.0560	0.3235 ± 0.0525
P12	0.3122 ± 0.0105	0.3027 ± 0.0249	0.2953 ± 0.0152	**0.2936 ± 0.0203**	0.2961 ± 0.0358	0.2982 ± 0.0332
P13	1.509 ± 0.980	1.459 ± 0.914	1.429 ± 0.971	1.811 ± 0.089	1.294 ± 0.626	**1.234 ± 0.571**
P14	0.6728 ± 0.2563	0.5902 ± 0.3086	0.5953 ± 0.3144	0.5992 ± 0.3212	0.5341 ± 0.2443	**0.5260 ± 0.2216**
P15	0.5203 ± 0.0875	0.5064 ± 0.0818	**0.5023 ± 0.0818**	0.5110 ± 0.0832	0.5103 ± 0.0550	0.5072 ± 0.0772
P16	0.2638 ± 0.0066	0.2522 ± 0.0128	0.2533 ± 0.0126	**0.2514 ± 0.0112**	0.2538 ± 0.0052	0.2541 ± 0.0048
P17	0.5439 ± 0.0520	0.5792 ± 0.0815	0.5387 ± 0.0701	0.5430 ± 0.0706	**0.4716 ± 0.0234**	0.4823 ± 0.0308
P18	0.3517 ± 0.0359	0.3149 ± 0.0298	0.3142 ± 0.0283	**0.3131 ± 0.0261**	0.3299 ± 0.0326	0.3291 ± 0.0323
P19	**0.3834 ± 0.1384**	0.4078 ± 0.2087	0.4238 ± 0.1938	0.4227 ± 0.1950	0.4148 ± 0.1952	0.4193 ± 0.1977
P20	0.5735 ± 0.0893	0.5080 ± 0.0579	0.5095 ± 0.0576	0.4994 ± 0.0501	**0.4600 ± 0.0361**	0.4603 ± 0.0345
P21	0.8012 ± 0.1494	0.8324 ± 0.1724	0.8170 ± 0.1762	0.8316 ± 0.1938	**0.6883 ± 0.1235**	0.6894 ± 0.1249
P22	0.06249 ± 0.00238	0.05820 ± 0.00202	0.05726 ± 0.00180	**0.05585 ± 0.00201**	0.05756 ± 0.00157	0.05721 ± 0.00124
P23	0.3039 ± 0.0438	0.2646 ± 0.0367	**0.2592 ± 0.0353**	0.2613 ± 0.0351	0.2666 ± 0.0410	0.2634 ± 0.0355

Table 19: Absolute nCRPS by dataset. Cells report mean ± standard deviation over three origins and five seeds. Public benchmarks appear above the rule and anonymized private datasets below it. Lower is better and bold marks the row minimum. COFLOW is abbreviated CF-.

Dataset	Base	Bottom-up	WLS-var	MinT-shrink	CF-Tweedie-RouteAttn.L50	CF-RBIG+NVP-RouteAttn.L25
Labour	0.027953 ± 0.002196	**0.007787 ± 0.001359**	0.008196 ± 0.002134	0.008694 ± 0.002570	0.008160 ± 0.002322	0.008608 ± 0.002520
M5	0.1146 ± 0.0078	0.1333 ± 0.0062	0.1258 ± 0.0058	0.1082 ± 0.0095	**0.1037 ± 0.0066**	0.1090 ± 0.0089
TourismLarge	0.1486 ± 0.0123	0.1346 ± 0.0126	0.1360 ± 0.0146	0.1357 ± 0.0152	**0.1288 ± 0.0148**	0.1318 ± 0.0143
TourismSmall	0.11790 ± 0.01150	**0.09115 ± 0.01591**	0.09748 ± 0.01725	0.09998 ± 0.01686	0.09742 ± 0.01625	0.09324 ± 0.01717
Traffic	0.04398 ± 0.00906	0.05129 ± 0.01321	0.04463 ± 0.01216	0.04489 ± 0.01206	0.04724 ± 0.01015	**0.04373 ± 0.01211**
Wiki2	0.5239 ± 0.0866	0.7309 ± 0.2415	0.5507 ± 0.1190	0.5542 ± 0.1106	**0.3778 ± 0.0553**	0.5249 ± 0.1061
P01	**0.4752 ± 0.1324**	0.4979 ± 0.1482	0.5137 ± 0.1571	0.5156 ± 0.1574	0.4890 ± 0.1477	0.4978 ± 0.1508
P02	0.5818 ± 0.1059	0.5944 ± 0.1524	0.5927 ± 0.1589	0.5885 ± 0.1635	**0.5346 ± 0.1368**	0.5819 ± 0.1576
P03	0.2568 ± 0.0253	0.2392 ± 0.0213	0.2429 ± 0.0241	0.2443 ± 0.0245	**0.2315 ± 0.0235**	0.2405 ± 0.0242
P04	0.3310 ± 0.0328	0.3310 ± 0.0321	0.3273 ± 0.0301	0.3377 ± 0.0400	**0.3251 ± 0.0376**	0.3358 ± 0.0390
P05	0.7008 ± 0.0626	0.7303 ± 0.0580	0.7451 ± 0.0659	0.7538 ± 0.0706	**0.6893 ± 0.0653**	0.7275 ± 0.0681
P06	0.2171 ± 0.0058	**0.2075 ± 0.0072**	0.2095 ± 0.0073	0.2159 ± 0.0084	0.2107 ± 0.0068	0.2134 ± 0.0077
P07	**0.5844 ± 0.2029**	0.6809 ± 0.2885	0.6700 ± 0.2803	0.6655 ± 0.2816	0.6029 ± 0.2540	0.6462 ± 0.2772
P08	**2.170 ± 2.010**	2.899 ± 2.965	2.827 ± 2.893	3.046 ± 3.062	2.719 ± 2.758	3.016 ± 3.057
P09	**0.4265 ± 0.0337**	0.5314 ± 0.0636	0.5222 ± 0.0542	0.5229 ± 0.0542	0.4463 ± 0.0407	0.5106 ± 0.0532
P10	0.4176 ± 0.0232	0.4213 ± 0.0232	0.4250 ± 0.0260	0.4322 ± 0.0256	**0.4017 ± 0.0225**	0.4222 ± 0.0230
P11	0.2560 ± 0.0487	**0.2145 ± 0.0347**	0.2302 ± 0.0371	0.3082 ± 0.0774	0.3060 ± 0.0720	0.3088 ± 0.0773
P12	0.2432 ± 0.0161	0.2408 ± 0.0278	0.2341 ± 0.0213	0.2345 ± 0.0245	**0.2291 ± 0.0297**	0.2337 ± 0.0244
P13	1.317 ± 0.886	1.284 ± 0.835	**1.276 ± 0.903**	1.657 ± 0.148	1.591 ± 0.168	1.624 ± 0.158
P14	0.5135 ± 0.2210	0.4802 ± 0.2819	0.4918 ± 0.2929	0.4975 ± 0.2999	**0.4557 ± 0.2659**	0.4769 ± 0.2872
P15	**0.4216 ± 0.0735**	0.4242 ± 0.0703	0.4285 ± 0.0720	0.4347 ± 0.0740	0.4230 ± 0.0713	0.4228 ± 0.0696
P16	0.2021 ± 0.0066	0.1935 ± 0.0104	0.1934 ± 0.0104	0.1923 ± 0.0100	**0.1908 ± 0.0095**	0.1924 ± 0.0104
P17	0.4446 ± 0.0459	0.4787 ± 0.0738	0.4656 ± 0.0678	0.4697 ± 0.0692	**0.4325 ± 0.0550**	0.4500 ± 0.0655
P18	0.2742 ± 0.0300	0.2536 ± 0.0293	0.2580 ± 0.0280	0.2576 ± 0.0258	**0.2473 ± 0.0259**	0.2537 ± 0.0263
P19	**0.3269 ± 0.1328**	0.3460 ± 0.2027	0.3695 ± 0.1910	0.3689 ± 0.1921	0.3323 ± 0.1632	0.3623 ± 0.1871
P20	0.4244 ± 0.0637	0.3862 ± 0.0468	0.3881 ± 0.0471	**0.3822 ± 0.0383**	0.3904 ± 0.0397	0.3864 ± 0.0386
P21	0.5943 ± 0.1146	0.6241 ± 0.1328	0.6181 ± 0.1384	0.6317 ± 0.1519	**0.5940 ± 0.1397**	0.6265 ± 0.1503
P22	0.04867 ± 0.00182	0.04548 ± 0.00177	0.04494 ± 0.00154	0.04398 ± 0.00157	**0.04247 ± 0.00133**	0.04425 ± 0.00186
P23	0.2399 ± 0.0358	0.2105 ± 0.0299	0.2095 ± 0.0310	0.2136 ± 0.0307	**0.2040 ± 0.0277**	0.2094 ± 0.0292

Table 20: Absolute nMIS by dataset. Cells report mean ± standard deviation over three origins and five seeds. Public benchmarks appear above the rule and anonymized private datasets below it. Lower is better and bold marks the row minimum. COFLOW is abbreviated CF-.

Dataset	Base	Bottom-up	WLS-var	MinT-shrink	CF-Tweedie-RouteAttn.L50	CF-RBIG+NVP-RouteAttn.L25
Labour	0.35481 ± 0.03777	0.07879 ± 0.01755	0.09038 ± 0.03299	0.10012 ± 0.04552	**0.07345 ± 0.03232**	0.09434 ± 0.04393
M5	1.090 ± 0.106	1.243 ± 0.093	1.168 ± 0.088	1.015 ± 0.116	**0.853 ± 0.050**	1.009 ± 0.190
TourismLarge	1.510 ± 0.196	1.278 ± 0.187	1.465 ± 0.254	1.481 ± 0.279	**1.116 ± 0.215**	1.209 ± 0.186
TourismSmall	1.601 ± 0.209	1.407 ± 0.294	1.597 ± 0.339	1.650 ± 0.334	1.520 ± 0.303	**1.068 ± 0.241**
Traffic	0.4242 ± 0.1309	0.4622 ± 0.1342	**0.4150 ± 0.1415**	0.4594 ± 0.1850	0.4825 ± 0.1177	0.4374 ± 0.1980
Wiki2	6.526 ± 1.556	9.135 ± 4.041	7.094 ± 2.084	7.161 ± 1.919	**4.513 ± 0.900**	6.045 ± 1.716
P01	**6.413 ± 1.718**	7.452 ± 2.346	8.099 ± 2.601	8.143 ± 2.609	7.032 ± 2.312	7.308 ± 2.530
P02	7.312 ± 1.534	8.040 ± 2.485	8.150 ± 2.647	8.146 ± 2.715	**6.208 ± 1.838**	7.859 ± 2.714
P03	2.969 ± 0.454	2.917 ± 0.373	3.170 ± 0.451	3.196 ± 0.458	**2.615 ± 0.399**	2.941 ± 0.435
P04	**3.477 ± 0.721**	3.643 ± 0.689	3.683 ± 0.781	3.865 ± 0.778	3.507 ± 0.718	3.686 ± 0.777
P05	**6.948 ± 0.679**	8.133 ± 0.658	8.749 ± 0.801	8.934 ± 0.844	7.824 ± 0.825	8.687 ± 0.854
P06	2.002 ± 0.176	1.859 ± 0.043	1.953 ± 0.058	2.061 ± 0.104	**1.841 ± 0.068**	1.960 ± 0.118
P07	9.44 ± 3.97	11.36 ± 5.64	11.58 ± 5.52	11.49 ± 5.55	**8.40 ± 4.22**	10.28 ± 5.35
P08	**41.06 ± 40.19**	55.77 ± 59.37	54.73 ± 57.92	59.18 ± 61.25	45.02 ± 48.08	57.03 ± 60.80
P09	5.947 ± 0.660	8.060 ± 1.229	8.249 ± 1.039	8.267 ± 1.038	**4.891 ± 0.532**	7.527 ± 1.026
P10	**4.670 ± 0.376**	5.039 ± 0.440	5.234 ± 0.497	5.363 ± 0.535	4.752 ± 0.353	5.112 ± 0.443
P11	2.637 ± 0.668	**1.901 ± 0.464**	2.242 ± 0.505	3.595 ± 1.268	3.421 ± 1.151	3.504 ± 1.190
P12	2.611 ± 0.390	2.747 ± 0.565	2.727 ± 0.545	2.761 ± 0.575	**2.574 ± 0.670**	2.708 ± 0.545
P13	20.26 ± 14.58	**20.09 ± 14.05**	20.58 ± 15.86	28.16 ± 5.05	25.72 ± 6.28	25.35 ± 5.79
P14	5.475 ± 3.283	6.054 ± 4.895	6.522 ± 5.236	6.685 ± 5.387	**5.445 ± 4.152**	6.149 ± 5.012
P15	**5.249 ± 1.011**	5.788 ± 1.059	6.189 ± 1.152	6.305 ± 1.190	5.728 ± 1.176	5.570 ± 0.949
P16	1.983 ± 0.150	**1.915 ± 0.188**	1.929 ± 0.182	1.933 ± 0.183	1.926 ± 0.167	1.994 ± 0.195
P17	5.826 ± 0.723	6.434 ± 1.285	6.954 ± 1.250	7.032 ± 1.285	**5.486 ± 0.573**	6.065 ± 1.110
P18	3.049 ± 0.467	3.145 ± 0.568	3.395 ± 0.546	3.406 ± 0.508	**2.883 ± 0.492**	3.138 ± 0.518
P19	4.643 ± 2.556	4.971 ± 3.861	5.663 ± 3.725	5.669 ± 3.745	**3.959 ± 2.382**	5.121 ± 3.466
P20	**3.950 ± 0.674**	3.958 ± 0.648	4.009 ± 0.666	4.005 ± 0.502	4.661 ± 0.659	4.050 ± 0.572
P21	**5.504 ± 1.165**	6.081 ± 1.519	6.237 ± 1.664	6.492 ± 1.853	5.829 ± 1.719	6.784 ± 2.036
P22	0.5320 ± 0.0306	0.5049 ± 0.0252	0.5084 ± 0.0209	0.5039 ± 0.0183	**0.4278 ± 0.0156**	0.4992 ± 0.0397
P23	2.817 ± 0.464	2.518 ± 0.415	2.627 ± 0.479	2.764 ± 0.481	**2.302 ± 0.403**	2.470 ± 0.415



Figure 28: Absolute relMAE for every method and dataset, averaged over three origins and five seeds. Colors are scaled within each dataset, with blue denoting lower error and red higher error.

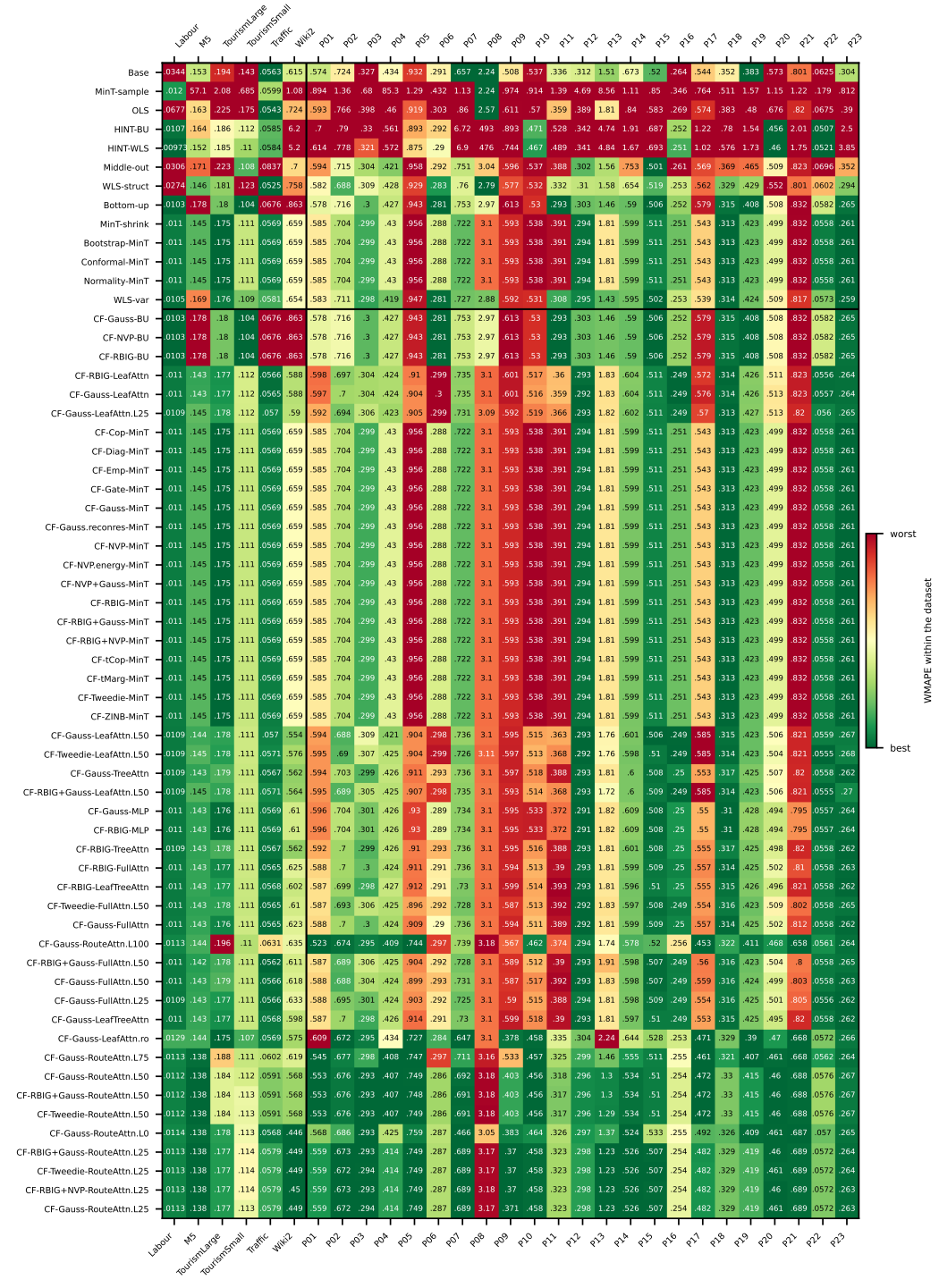


Figure 29: Absolute WMAPE for every method and dataset, averaged over three origins and five seeds. Colors are scaled within each dataset, with blue denoting lower error and red higher error.

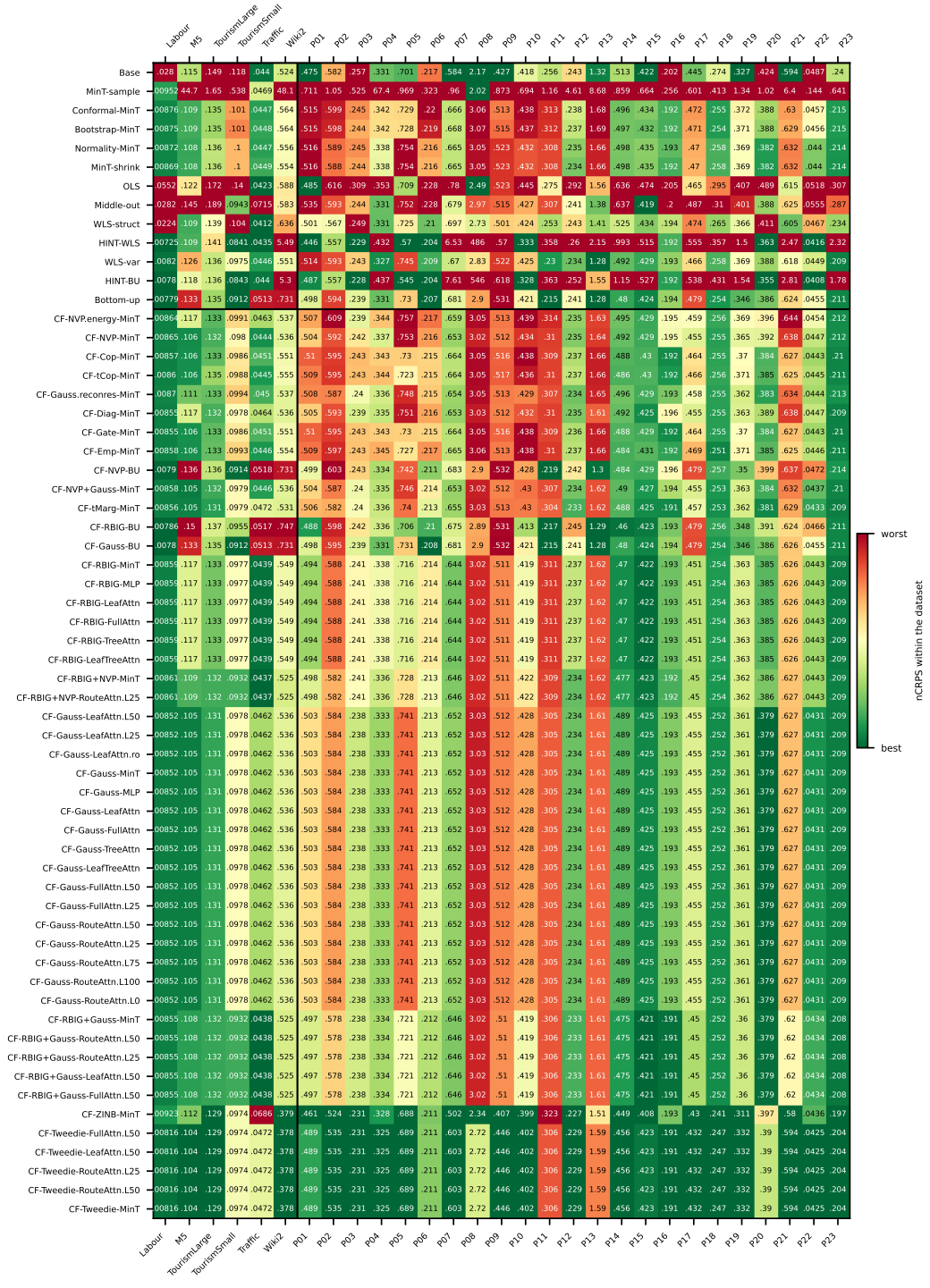


Figure 30: Absolute nCRPS for every method and dataset, averaged over three origins and five seeds. Colors are scaled within each dataset, with blue denoting lower error and red higher error.

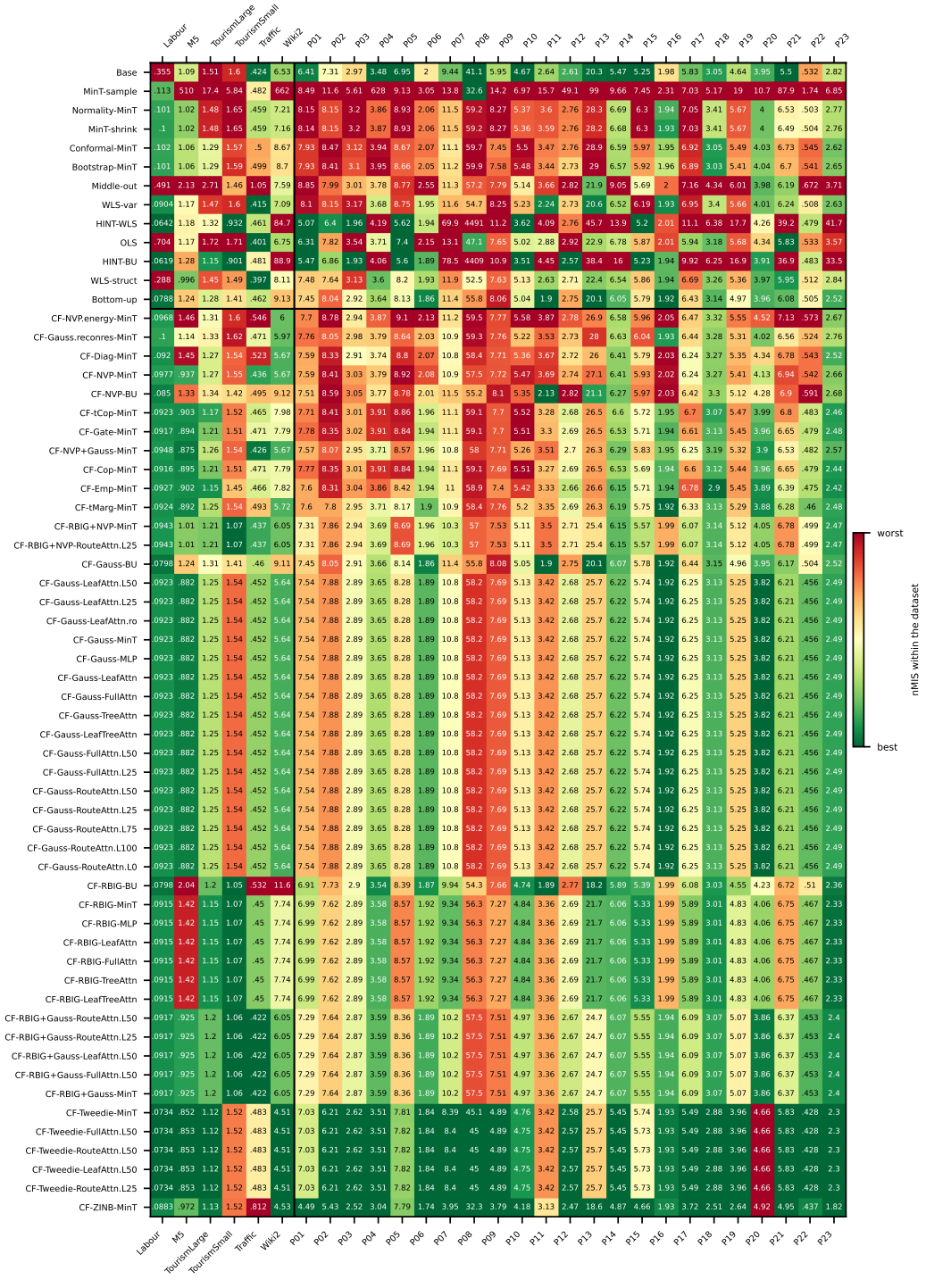


Figure 31: Absolute nMIS for every method and dataset, averaged over three origins and five seeds. Colors are scaled within each dataset, with blue denoting lower error and red higher error.