Filtered Conformal Ellipsoids for Graph-Native Time Series

Yannick Limmer
DRW
United Kingdom
ylimmer@drw.com

Abstract

Joint prediction sets for multivariate time series should control a single event while adapting to cross-coordinate dependence. We study *filtered conformal ellipsoids*: a frozen state-space filter emits a one-step predictive mean and covariance, and split-conformal calibration is applied to the resulting Mahalanobis scores. The filter is used to choose the ellipsoid *shape*; conformal calibration chooses the scalar *radius*, so the construction benefits from a learned predictive covariance without relying on Gaussian tail probabilities for coverage. The main difficulty is that filtered scores are dependent and learned recurrent filters need not contract in their raw hidden state; we therefore analyse contraction in *each filter's own* observable predictive-law quotient that identifies hidden states producing the same future sequence of emitted Gaussian laws. Under a stable Bayes Gaussian-projection filter, covariance bounds, and a contraction-modulus identification condition, small excess Gaussian negative log-likelihood is one sufficient route to contraction of the learned filter on its own quotient. Combined with input sensitivity, geometric data coupling, and a threshold-autocovariance envelope, a dependent quantile-concentration bound yields an *approximate marginal coverage bound* for filtered split-conformal prediction under dependence (a Chebyshev form; a sharper Bernstein form requires an additional geometric-mixing concentration assumption). This is a marginal, not calibration-conditional, guarantee. Under Gaussian oracle realisability we also obtain a near-oracle log-volume comparison within the class of conditionally valid Gaussian ellipsoid rules. We implement the filter as a GCN-GRU filter with diagonal-plus-low-rank covariance. On moderate-size graph-native traffic benchmarks (METR-LA-20 and PEMS-BAY-50), the learned filter gives sharper at-target ellipsoids than static-covariance and non-filter baselines; at full-graph scale and on non-graph-native datasets, factor and copula baselines can be stronger.

1 Introduction

Deployed forecasters of correlated sensor streams output prediction sets, not point forecasts. For $Y_t \in \mathbb{R}^N$ a multivariate sensor vector at time t, downstream users rely on the joint event $\{Y_t \in \hat{C}_\alpha(x_t)\}$, not N separate marginal statements. Marginal conformal intervals therefore do not solve the deployed problem: on correlated streams their Cartesian product can miss the target joint event even when each coordinate is calibrated [23, 28], and our diagnostics show marginal methods falling far below joint target coverage (App. E.10). We instead conformalise a scalar Mahalanobis score, producing an ellipsoid whose axes encode cross-sensor dependence.

Our construction separates *shape* from *coverage*. A frozen filter ϕ emits a one-step predictive Gaussian $\mathcal{N}(\hat{Y}_{t|t-1}, \hat{\Sigma}_{t|t-1})$ before Y_t is observed; the squared Mahalanobis score $R_t = R(Y_t)$, with $R(y) := (y - \hat{Y}_{t|t-1})^\top \hat{\Sigma}_{t|t-1}^{-1}(y - \hat{Y}_{t|t-1})$, is calibrated on a chronological calibration block, and the test-time set is $\{y : R(y) \leq \hat{q}_{1-\alpha}\}$ (Algorithm 1, Figure 1). The filter's $\hat{\Sigma}_{t|t-1}$ supplies only the ellipsoid's orientation and per-axis scale; the conformal quantile of the observed Mahalanobis residuals supplies the scalar radius. The method therefore benefits from a learned predictive covariance

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

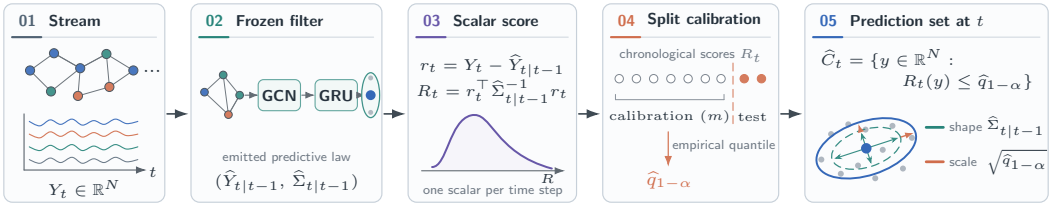

Figure 1: **Filtered conformal prediction.** A frozen filter emits a one-step mean and covariance; calibration scores are squared Mahalanobis residuals; the empirical $(1-\alpha)$-quantile defines a prediction ellipsoid. The filter chooses the ellipsoid shape; conformal calibration chooses the scalar radius. The theory analyses contraction of emitted predictive laws (Theorem 3), not raw recurrent states.

while retaining calibration through observed residuals rather than through Gaussian tail probabilities. Throughout we use the display names GNF (the learned graph filter), KALMAN-FCP (linear Kalman baseline), STATIC-CGIF (static-covariance baseline), and DIAG-GRU / GCN-RANK-0 for the two diagonal-covariance ablations (graph-removed-and-diagonal vs. graph-retained-but-rank-zero); implementation details and the display-to-implementation name mapping are in Apps. C–D. We call a benchmark *graph-native* when the data source provides a meaningful sensor graph; the formal criterion appears in App. E.1.

Calibration scores from one chronological trajectory are dependent, and raw-state contraction can be too strong for recurrent filters with output-invisible directions. We instead work in each filter's observable quotient $\mathcal{Q}_\theta$, comparing only scalar contraction moduli across filters. Under a stable Bayes Gaussian-projection oracle, uniform covariance spectral bounds, and a contraction-modulus identification condition, small excess Gaussian-NLL risk is *one sufficient route* to contraction of the learned filter on its own quotient (Theorem 3); input sensitivity and data coupling then yield dependent quantile concentration and an *unconditional approximate marginal coverage-gap bound* for filtered split-conformal prediction (Theorem 7). The guarantee is marginal; the log-volume comparison uses conditionally valid Gaussian ellipsoid rules (Theorem 9). Table 1 records the assumptions and audit scope.

Empirically, GNF is the sharpest at-target method on METR-LA-20 and PEMS-BAY-50, improving w_{proj} by 23.6% and 40.5% over STATIC-CGIF (Figure 2, Table 2); at full-graph scale and on non-graph-native data, factor or copula baselines can be stronger (Table 3, App. E.9). Same-mean and matched-budget controls support the covariance-shape interpretation on the two headline cells (§6.2), while the audits probe finite-trajectory consequences rather than population premises.

Contributions.

1. *Filtered conformal ellipsoids with learned predictive covariances.* A frozen filter supplies a time-varying ellipsoid shape and conformal calibration supplies the radius, so the construction does not rely on Gaussian tail validity.
2. *Observable-law theory for dependent filtered scores.* Working in each filter's own parameter-specific observable quotient, small excess Gaussian-NLL is one sufficient route to contraction of the learned filter (Theorem 3); with input sensitivity and data coupling, a dependent quantile-concentration bound yields an unconditional approximate marginal coverage-gap bound (Theorem 7) and a near-oracle log-volume comparison within the conditionally valid Gaussian-ellipsoid class (Theorem 9).
3. *Graph-based implementation and controlled comparisons.* A GCN-GRU diagonal-plus-low-rank covariance filter improves sharpness on moderate-N traffic. Same-mean and matched-budget DEEPVAR-FCP controls probe the covariance mechanism; full-graph and non-graph-native comparisons delimit its scope.

2 Related work

Conformal prediction under dependence. Split-conformal prediction [16, 25] requires exchangeability. Online adjustments such as ACI [9] and AGACI [29] attain long-run marginal coverage under arbitrary drift; the "beyond-exchangeability" line [3] bounds coverage loss under departures from exchangeability.

Multivariate / joint conformal regions. SPCI [27] and its multivariate extension MULTIDIMSPCI [28] produce ellipsoidal regions via a learned-but-not-filtered covariance; COPULACPTS [24] builds joint regions via empirical copulas. These are our published joint-CP comparators; implementation details and additional baselines (factor, group, spectral, EWMA, rolling, time-of-day, and a k-NN local ellipsoid) are in App. D.

Topology-aware conformal prediction. Recent work injects graph topology into conformal prediction at different stages: STACI [30] lets topology and temporal dynamics enter the *nonconformity score* for stream networks, CF-GNN [13] through a graph-aware *output-correction* model, and Cini et al. [6], Dua et al. [8] propagate calibration information over the graph. Here topology enters through the learned, time-varying predictive covariance; conformal calibration then operates on a single scalar score.

Learned predictive covariances and filter stability. Likelihood-trained Gaussian heads on recurrent backbones [19–21] predict distributional parameters but are not calibrated finite-sample; DeepVAR-style low-rank Gaussian RNNs [20] are the non-graph analogue of our covariance mechanism, and we include a matched-budget DEEPVAR-FCP control in §6. Graph neural backbones [5, 15, 17, 26] give point predictors that we re-use as mean backbones. Filter-stability and observability tools are classical [1, 7, 11]; the contribution of §5 is to transfer stability of a learned filter from these abstract conditions to filtered split-conformal validity and log-volume sharpness, in *each filter's own* observable predictive-law quotient rather than on raw recurrent states.

3 Preliminaries

Sequence model. We observe $\{Y_t\}_{t\geq 1} \subset \mathbb{R}^N$ together with covariates x_t, which may include lagged endogenous observations, graph features, and calendar/exogenous variables. A *filter* is any measurable map $\phi : (x_t, H_{t-1}) \mapsto (\hat{Y}_{t|t-1}, \hat{\Sigma}_{t|t-1}, H_t)$; we reserve N for the observation dimension and d for the latent state. The residual $r_t := Y_t - \hat{Y}_{t|t-1}$ gives the squared Mahalanobis score

$$R_t \;=\; r_t^\top \hat{\Sigma}_{t|t-1}^{-1} r_t \tag{1}$$

and its root form $S_t := R_t^{1/2} = \|\hat{\Sigma}_{t|t-1}^{-1/2} r_t\|_2$. The two scores induce the *same* prediction region by monotonicity: if $\hat{q}_S$ is the empirical $(1-\alpha)$-quantile of S_t, then $\hat{q}_{1-\alpha} = \hat{q}_S^2$ is the corresponding quantile of R_t and $R_t \leq \hat{q}_{1-\alpha} \Leftrightarrow S_t \leq \hat{q}_S$. We *calibrate* R_t in Algorithm 1 (smooth in H_t, no $1/\sqrt{R}$ blow-up) and *analyse* S_t in §5 (avoids the global-Lipschitz issue at unbounded residuals; App. B.7). At a calibration trajectory $\{Y_t\}_{t=1}^n$, set $\hat{q}_{1-\alpha} := R_{(\lceil (n+1)(1-\alpha) \rceil)}$, with $R_{(n+1)} = +\infty$, and emit

$$\hat{C}_\alpha(x_{n+1}) \;=\; \big\{ y \in \mathbb{R}^N : (y - \hat{Y}_{n+1|n})^\top \hat{\Sigma}_{n+1|n}^{-1} (y - \hat{Y}_{n+1|n}) \leq \hat{q}_{1-\alpha} \big\}. \tag{2}$$

Joint coverage $\mathbb{P}\{Y_{n+1} \in \hat{C}_\alpha(x_{n+1})\}$ is the deployment-relevant metric.

Set-size functional. The primary size functional we report is the *mean coordinate-projection length* w_{proj}: a common, set-family-agnostic functional applied to each method's own prediction set. Writing $\ell_{t,i} := \inf_{y \in \hat{C}_t} y_i$ and $u_{t,i} := \sup_{y \in \hat{C}_t} y_i$ for the exact marginal extent of $\hat{C}_t$ in coordinate i,

$$w_{\mathrm{proj}} \;:=\; \frac{1}{T_{\mathrm{test}}\, N} \sum_{t=1}^{T_{\mathrm{test}}} \sum_{i=1}^{N} \big(u_{t,i} - \ell_{t,i} \big). \tag{3}$$

For a single ellipsoid $\hat{C}_t = \{y : (y - \hat{\mu}_t)^\top \hat{\Sigma}_t^{-1} (y - \hat{\mu}_t) \leq \hat{q}_t\}$ the coordinate-projection length is available in closed form via the support function, $u_{t,i} - \ell_{t,i} = 2(\hat{q}_t\, \hat{\Sigma}_{t,ii})^{1/2}$. For COPULACPTS we use the coordinate endpoints of its native jointly calibrated set; for factor, group, spectral, and Bonferroni constructions we project the *full composed set*, not one constituent ellipsoid; joint coverage is always evaluated on the original joint set $\hat{C}_t$, never its axis-aligned bounding box; and all methods are compared in the same standardised coordinate system. The exact projection computation used for each non-ellipsoidal or composed set is given in App. D.2 (Table 6). Note that w_{proj} is the average side length of the set's axis-aligned bounding box: it is directly interpretable as an average per-sensor range, but it is basis-dependent and does not characterise the complete geometry or volume of a joint set.

Algorithm 1 Filtered Split-Conformal Prediction

Require: sequence $\{Y_t\}_{t=1}^{n+T}$, frozen filter ϕ, level α, first retained calibration index t_0

1: *(calibration)* run ϕ on $Y_{1:n}$, predicting each Y_t before observing it. Compute $R_t = r_t^\top \hat{\Sigma}_{t|t-1}^{-1} r_t$ and discard scores with $t < t_0$.
2: Let $m = n - t_0 + 1$ and let $R_{(k)}$ denote the k-th smallest retained score, with $R_{(m+1)} = +\infty$.
3: $\hat{q}_{1-\alpha} \leftarrow R_{(\lceil (m+1)(1-\alpha) \rceil)}$.
4: *(test)* for $t = n+1, \ldots, n+T$, advance ϕ using Y_{t-1} and emit $\hat{C}_\alpha(x_t)$ via (2) before observing Y_t.

As a complementary geometry-*aware* size functional for genuine single-ellipsoid rules we report the normalised log-volume

$$\tfrac{1}{N} \log \mathrm{Vol}(\hat{C}_t) \;=\; \tfrac{1}{N} \log v_N \;+\; \log \hat{s} \;+\; \tfrac{1}{2N} \log\det \hat{\Sigma}_{t|t-1}, \tag{4}$$

where v_N is the volume of the Euclidean unit ball in $\mathbb{R}^N$ and $\hat{s} = \hat{q}_{1-\alpha}^{1/2}$; Theorem 9 controls this quantity for single full-dimensional ellipsoid rules. Our experiments use both metrics: Table 2 reports w_{proj} and App. E.17 (Table 18) reports the corresponding single-ellipsoid log-volumes; the rankings agree (§6.1).

Why Mahalanobis. The Mahalanobis score produces an ellipsoid oriented by $\hat{\Sigma}$ and is invariant under invertible affine reparameterisation. For Gaussian targets, its level sets minimise volume at fixed coverage [1]; Appendix B.1 discusses alternative scores.

Theorem 1 (Split-conformal validity; 25)**.** *If $(R_t)_{t=1}^{n+1}$ are exchangeable, the region* (2) *satisfies $\mathbb{P}\{Y_{n+1} \in \hat{C}_\alpha(x_{n+1})\} \geq 1 - \alpha$; if the scores are a.s. distinct, coverage is at most $1 - \alpha + 1/(n+1)$.*

Predictive-law metric and observable quotient. Write $P_{\theta,t}^h = \mathcal{N}(\mu_{\theta,t}^h, \Sigma_{\theta,t}^h)$ for the one-step predictive Gaussian law emitted by a frozen filter parameterised by θ and initialised at quotient state h. We endow Gaussian laws with the Bures-Wasserstein distance

$$d_{\mathcal{G}}^2\big(\mathcal{N}(\mu,\Sigma), \mathcal{N}(\nu,\Lambda)\big) \;=\; \|\mu - \nu\|_2^2 + \mathrm{tr}\big[\Sigma + \Lambda - 2(\Lambda^{1/2}\Sigma\Lambda^{1/2})^{1/2}\big]; \tag{5}$$

on covariances spectrum-bounded between λ_- and λ_+, $d_{\mathcal{G}}^2$ is comparable to Gaussian KL divergence (App. A.1). For *each fixed filter parameter θ* we define a *parameter-specific* observable quotient $\mathcal{Q}_\theta := H/\!\sim_\theta$, where $h \sim_\theta h'$ if and only if h and h' emit the same future predictive-law sequence under every admissible common input continuation, with quotient metric d_θ (full construction in App. A.2). All contraction statements in §5 are made inside a single filter's own quotient, we explicitly do not use the raw recurrent-state norm and avoid identifying elements of $\mathcal{Q}_{\theta_\star}$ with those of $\mathcal{Q}_{\hat{\theta}}$. Therefore, the analysis compares only the scalar contraction moduli of the two filters. For the linear-Gaussian instance we use F for the transition, C for the observation matrix, K for the steady-state Kalman gain, and $A_{\mathrm{cl}} := F(I - KC)$ for the closed-loop transition (App. C.1); A_{graph} is reserved for the adjacency. We treat both linear-Kalman (GRAPH-LGSSM) and learned GCN–GRU (GNF) filters as instances.

4 Filtered conformal prediction

Algorithm 1 defines the protocol. We fit a filter ϕ on a held-out pretraining trajectory and freeze its parameters. We then run it forward on the calibration trajectory to collect squared-Mahalanobis scores and compute their conformal quantile. At test time, this fixed quantile and the filter's predictive mean and covariance define the ellipsoid in (2).

Observable contraction (Theorem 3), together with the score-dependence assumptions of §5, yields *unconditional approximate marginal validity* (Theorem 7). Exact split-conformal validity is recovered in the oracle standardised-innovation case (Proposition 8). Theorem 9 bounds log-volume under its additional assumptions.

GRAPH-LGSSM (linear Kalman baseline). We use a linear-Gaussian state-space model with graph transition $F = \rho_{\mathrm{scale}}\, \widetilde{A}_{\mathrm{sym}} / \lambda_{\max}(\widetilde{A}_{\mathrm{sym}})$, where $\widetilde{A}_{\mathrm{sym}}$ is the symmetric-normalised adjacency with self-loops and $\lambda_{\max}$ is its largest eigenvalue. The process and observation noise covariances are isotropic, $Q = \sigma_Q^2 I$ and $R = \sigma_R^2 I$, and the observation matrix is $C = I_N$. The contraction diagnostics use $\rho_{\mathrm{scale}} = 0.8$; the predictive comparisons use validation-tuned parameters (App. C.1). This

baseline serves both as a sanity check and as the linear specialisation of the observable-contraction theorem. For each dataset, we report the analytic closed-loop rate $\sigma_1(A_{\mathrm{cl}})$, the initialisation-weighted root-mean-square (RMS) rate $\rho_{M_0}(A_{\mathrm{cl}}, T)$ from Theorem 19 (App. B.9), and the signal-to-noise proxy ρ_{snr} (App. C.1). Here σ_1 denotes the largest singular value; the RMS rate uses the initial perturbation's second moment M_0 and the horizon T.

GNF (learned GCN-GRU; our "learned graph filter"). We use a graph-convolutional gated recurrent unit (GCN-GRU) [5, 17] with per-node hidden states and node-shared weights. Let H_t collect these hidden states. At time t, a learned prediction head emits $(\hat{Y}_{t|t-1}, \hat{\Sigma}_{t|t-1}) = g_\theta(H_{t-1})$. After observing Y_t, the state advances as $H_t = \mathrm{GCNGRU}(Y_t, H_{t-1}; \widetilde{A}_{\mathrm{sym}})$. Both outputs are computed *before* Y_t is observed, during both calibration and testing (App. C.3).

The covariance has the diagonal-plus-low-rank form $\hat{\Sigma}_{t|t-1} = \mathrm{diag}(d_t) + L_t L_t^\top$, where $d_t \in \mathbb{R}^N$ is a positive diagonal component and $L_t \in \mathbb{R}^{N\times r}$ is a rank-r factor that captures cross-sensor correlation. The diagonal head exponentiates its outputs and clamps each entry of d_t to at least $\varepsilon_d = 10^{-4}$, keeping $\hat{\Sigma}_{t|t-1}$ positive definite. The Woodbury identity and matrix-determinant lemma allow Mahalanobis scores and log-determinants to be computed in $\mathcal{O}(Nr^2 + r^3)$ time, which is linear in N for fixed r.

Rank $r = 4$ was fixed globally from the synthetic pilot and used unchanged across all real datasets. The METR-LA rank sweep (App. E.13) is a retrospective sensitivity analysis and was not used for per-dataset model selection. We train the filter by minimising the Gaussian negative log-likelihood (NLL), using Adam, teacher forcing, and truncated backpropagation through time with a window of 64 steps (App. C). Under the assumptions of Theorem 3, small excess NLL also controls the learned filter's observable contraction rate.

A sweep over $\varepsilon_d \in \{10^{-6}, \ldots, 10^{-2}\}$ shows that the default floor is nonbinding on the audited full-covariance trajectories: no audited step has $\lambda_{\min}(\hat{\Sigma}_{t|t-1}) \le 2\varepsilon_d$. The floor therefore does not explain the remaining tail of high covariance condition numbers (App. E.23).

Ablations. We use two diagonal-covariance ablations. DIAG-GRU *(*NEURALSSM*: graph removed and diagonal Σ_t)* uses a global GRU over the sensor vector, without graph inputs, and sets $r = 0$. This is the row reported in Table 2 ($w_{\mathrm{proj}} = 4.71$ at joint coverage 0.932 on METR-LA-20). GCN-RANK-0 *(graph retained, rank 0)* keeps the per-node graph recurrence and only sets $r = 0$. This is the rank-sweep point ($w_{\mathrm{proj}} = 5.32$ on METR-LA-20, App. E.13) used to isolate the low-rank covariance component at fixed graph mixing. In this sweep, the gap from GCN-RANK-0 to r=4 shows that the low-rank component accounts for most of the sharpness gain. The DIAG-GRU comparison changes graph structure, recurrent architecture, and covariance rank together.

5 Theory

Our theory connects the quality and stability of a learned filter to the coverage and size of its conformal ellipsoids. Under the stated assumptions, sufficiently small excess Gaussian NLL relative to a stable oracle yields a learned filter whose predictive laws forget their initialisation geometrically (Theorem 3). Input sensitivity and geometric data coupling then control the score distribution's approach to stationarity and the dependence relevant to quantile estimation (Corollary 5). These bounds yield calibration-quantile concentration and an unconditional marginal coverage-gap bound (Theorem 7). Under Gaussian oracle realisability, the same quantile event also bounds log-volume relative to conditionally valid Gaussian ellipsoid rules (Theorem 9).

Table 1 distinguishes the assumptions from the finite-trajectory quantities probed by our audits. Contraction is measured within each filter's own observable quotient; the results do not give distribution-free validity for arbitrary learned filters.

5.1 NLL excess as one sufficient route to observable contraction

Observable contraction describes how a filter's predictions forget its initial state. We compare two runs of the same filter, started from different states and driven by the same observations. Theorem 3 shows that, under the conditions below, a learned filter inherits contraction from a stable Bayes Gaussian-projection oracle when its excess Gaussian NLL is sufficiently small.

Table 1: **Assumption summary.** Each main result is conditional on the listed assumptions; the audits estimate operative conclusions or finite-trajectory proxies directly but do not certify the population assumptions.

Result	Main additional assumptions	Audit checks (operative / proxy)	Not certified (population)
Thm. 3	stable Bayes Gaussian-projection oracle; covariance bounds; contraction-modulus identification	$\widehat{\rho}_{d_{\mathcal{G}}}$ emitted-law contraction (App. E.5); finite-horizon output-sensitivity proxy (App. E.16)	population $\varepsilon_{\mathrm{NLL}}$; realisability; population observability inequality
Cor. 5	input sensitivity + geometric data coupling (Ass. 4)	indicator-autocov decay $\widehat{\rho}_\Gamma$ (Audit 3)	distribution-free exchangeability
Thm. 7 (Cheb.)	threshold-autocovariance envelope (Cor. 5)	$\widehat{\rho}_\Gamma$, τ_{int} (App. E.6)	calibration-conditional coverage
Thm. 7 (Bern.)	geometric-mixing Bernstein concentration (Ass. 6)	integrated autocorrelation τ_{int} (App. E.6)	the concentration inequality itself
Thm. 9	Gaussian oracle realisability; conditionally valid Gaussian comparator	log-volume agrees with w_{proj} (App. E.17)	optimality vs. arbitrary marginally valid sets

Let $\theta_\star$ denote the oracle parameter and $\mathcal{R}(\theta)$ the population Gaussian-NLL risk. For each θ, let T_θ and O_θ be the transition and emission maps on $\mathcal{Q}_\theta$. Define the *scalar quotient contraction modulus* $\kappa(\theta) := \sup_y \sup_{q \neq q'} d_\theta(T_\theta(y,q), T_\theta(y,q'))/d_\theta(q,q')$ and the contraction-modulus gap $\Delta_{\mathrm{ctr}}(\theta) := [\kappa(\theta) - \kappa(\theta_\star)]_+$. The "no-ghost" condition in Assumption 2 requires any increase over the oracle's contraction modulus to be controlled by excess NLL. Directions that never affect emitted laws are removed by the quotient $\mathcal{Q}_{\widehat{\theta}}$. For an initial quotient state $q \in \mathcal{Q}_\theta$, we write $Q^q_{\theta,t}$ as the quotient state reached after t transitions, hence $P^q_{\theta,t} = O_\theta(Q^q_{\theta,t})$.

Assumption 2 (Observable contraction setting)**.** Let $q, q' \in \mathcal{Q}_\theta$. *(O1) Stable oracle.* An oracle $\theta_\star$ in the learned-filter family emits the Bayes Gaussian projection of the true one-step law and contracts on its own quotient, $\kappa(\theta_\star) \le \rho_\star < 1$. *(O2) Uniform covariance spectrum.* $\lambda_- I \preceq \hat{\Sigma}_{\theta,t} \preceq \lambda_+ I$. *(O3) Emitted-law Lipschitz + finite-horizon observability.* $d_{\mathcal{G}}(O_\theta(q), O_\theta(q')) \le L_{\mathrm{emit}}\, d_\theta(q,q')$ and $c_{\mathrm{obs}}\, d_\theta(q,q') \le D_{\theta,H_{\mathrm{obs}}}(q,q')$ (App. A.2), where $D_{\theta,H_{\mathrm{obs}}}$ is the observable discrepancy over horizon H_{obs} under a fixed θ-independent probing law with lag weights β_ℓ. *(O4) Contraction-modulus identification / no-ghost.* $\Delta_{\mathrm{ctr}}(\theta)^2 \le C_{\mathrm{id}}\,[\mathcal{R}(\theta) - \mathcal{R}(\theta_\star)]$ (App. A.3).

Theorem 3 (Observable contraction from excess Gaussian NLL)**.** *Under Assumption 2, let* $\varepsilon_{\mathrm{NLL}} := \mathcal{R}(\widehat{\theta}) - \mathcal{R}(\theta_\star)$ *and suppose* $\widehat{\rho}_{\mathrm{obs}} := \rho_\star + \sqrt{C_{\mathrm{id}}\,\varepsilon_{\mathrm{NLL}}} < 1$. *Then, for all* $q, q' \in \mathcal{Q}_{\widehat{\theta}}$ *and* $t \ge 0$*, the learned filter* $\widehat{\theta}$ *contracts on its own quotient: on every common input path,*

$$d_{\widehat{\theta}}\big(Q^q_{\widehat{\theta},t}, Q^{q'}_{\widehat{\theta},t}\big) \;\le\; \widehat{\rho}^{\,t}_{\mathrm{obs}}\, d_{\widehat{\theta}}(q,q'). \tag{6}$$

Consequently, translating to emitted laws,

$$d_{\mathcal{G}}\big(P^q_{\widehat{\theta},t}, P^{q'}_{\widehat{\theta},t}\big) \;\le\; \frac{L_{\mathrm{emit}}}{c_{\mathrm{obs}}}\, \widehat{\rho}^{\,t}_{\mathrm{obs}}\, D_{\widehat{\theta},H_{\mathrm{obs}}}(q,q'). \tag{7}$$

When $H_{\mathrm{obs}} = 0$ and $\beta_0 = 1$, $D_{\widehat{\theta},H_{\mathrm{obs}}}(q,q') = d_{\mathcal{G}}(P^q_{\widehat{\theta},0}, P^{q'}_{\widehat{\theta},0})$. In this instantaneous-observability case, the initial discrepancy in the bound reduces to the distance between the initial emitted laws.

The theorem transfers stability under the stated conditions, which an arbitrary trained GRU does not necessarily satisfy. Since $\varepsilon_{\mathrm{NLL}}$ is not observable (the oracle filter is unknown), we audit the theorem's conclusion (emitted-law contraction, App. E.5) rather than its premises. App. A.3 gives a sufficient condition for (O4) on the GCN-GRU family, using a risk margin and local Lipschitzness of $\kappa(\cdot)$.

5.2 From frozen-filter contraction to score dependence

We next connect *frozen-filter contraction* to temporal dependence in the scores. To do so, we combine contraction along a common input path with control of the filter's response to different input paths. Assumption 4 supplies this input sensitivity and geometric coupling of the data; the corollary also assumes local regularity of the score law. Theorem 3 is one sufficient route to contraction. The dependence and quantile results from Corollary 5 onward take contraction as a premise and do not require Gaussian oracle realisability or a verified Fisher margin.

Assumption 4 (Input sensitivity and geometric data coupling)**.** Let d_Y be a metric on the data/input space. The frozen filter obeys, for a constant L_{in}, $d_{\widehat{\theta}}(T_{\widehat{\theta}}(y,q), T_{\widehat{\theta}}(y',q')) \le \widehat{\rho}_{\mathrm{obs}}\, d_{\widehat{\theta}}(q,q') + L_{\mathrm{in}}\, d_Y(y,y')$, and the data/input process couples geometrically, $\mathbb{E}\, d_Y(Y_t, Y'_t) \le C_Y \rho^t_{\mathrm{data}}\, d_Y(Y_0, Y'_0)$ with $\rho_{\mathrm{data}} < 1$.

Corollary 5 (Threshold mixing)**.** *Under Assumption 4 and a local Lipschitz condition on the conditional law of the root score given the observable past (App. A.5), the augmented filter-data process couples geometrically at a rate $\rho_{\mathrm{aug}} < 1$ that can be taken arbitrarily close to $\max\{\rho_{\mathrm{data}}, \widehat{\rho}_{\mathrm{obs}}\}$, and the threshold indicators $I_i(u) := \mathbf{1}\{S_{\widehat{\theta},i} \le u\}$ obey, for u in a neighbourhood of the conformal threshold,*

$$\sup_i |\mathrm{Cov}(I_i(u), I_{i+k}(u))| \le C_\Gamma \, \rho_{\mathrm{aug}}^k. \tag{8}$$

The envelope is summable, and the same coupling yields a warm-start CDF bias $\sup_{u\in I} |\mathbb{P}(S_{\widehat{\theta},t} \le u) - G_\star(u)| \le C_{\mathrm{cdf}} \, \rho_{\mathrm{aug}}^t$ (App. A.5), where $G_\star$ is the stationary CDF of the learned root scores, and I a neighbourhood of its $(1-\alpha)$-quantile.

The combined rate ρ_{aug} reflects both filter contraction and data dependence. When $\rho_{\mathrm{data}} \neq \widehat{\rho}_{\mathrm{obs}}$, we can take their maximum, with an adjusted constant. When the two rates coincide at a common value ρ, the convolution contributes $t\,\rho^t$, which can be absorbed into a geometric bound with any $\rho_{\mathrm{aug}} \in (\rho, 1)$ (App. A.5).

5.3 Dependent quantile concentration and marginal validity

These dependence bounds let us quantify calibration-quantile error and translate it into a marginal coverage bound. Let $\hat{F}_m$ be the empirical CDF of the m retained root scores and $s_\star$ the $(1-\alpha)$-quantile of the stationary score law. Summable bounds Γ_k on lag-k threshold-indicator covariances give the Chebyshev result below; the sharper Bernstein bound uses the additional concentration assumption that follows.

Assumption 6 (Geometric-mixing concentration)**.** For u in a neighbourhood of $s_\star$, the centred threshold indicators satisfy $\mathbb{P}(|\hat{F}_m(u) - \mathbb{E}\hat{F}_m(u)| > \varepsilon) \le 2\exp(-c\, m_{\mathrm{eff}}\, \varepsilon^2)$ for $0 \le \varepsilon \le \varepsilon_{\max}$, with $m_{\mathrm{eff}} \asymp m(1-\rho_{\mathrm{aug}})/(1+\rho_{\mathrm{aug}})$ uniformly in u.

Theorem 7 (Dependent quantile concentration and unconditional marginal validity)**.** *Let m be the post-warmup calibration size, $\hat{s}_m$ the empirical conformal quantile of the learned root scores, and $g_\star$ the stationary score density (bounded above by $\bar{g}$ and away from zero near $s_\star$). Take m large enough that the conformal rank is at most m and $s_\star \pm r_m(\delta)$ lie in this density neighbourhood.*

Part I (high-probability quantile estimation). *For the appropriate radius $r_m(\delta)$, $\mathbb{P}\{|\hat{s}_m - s_\star| \le r_m(\delta)\} \ge 1-\delta$, with*

$$r_m^{\mathrm{Cheb}}(\delta) \;\lesssim\; \rho_{\mathrm{aug}}^{t_0} + (\delta\, m_{\mathrm{eff}}^{\mathrm{cov}})^{-1/2} + m^{-1} \quad \textit{under Corollary 5,} \tag{9}$$

and, under the stronger Assumption 6,

$$r_m^{\mathrm{Bern}}(\delta) \;\lesssim\; \rho_{\mathrm{aug}}^{t_0} + \sqrt{\tfrac{\log(1/\delta)}{m_{\mathrm{eff}}}} + \tfrac{\log(1/\delta)}{m_{\mathrm{eff}}} + m^{-1}, \qquad m_{\mathrm{eff}}^{\mathrm{cov}} \asymp \tfrac{m}{1+2\sum_{k\ge1}\Gamma_k}. \tag{10}$$

Part II (unconditional marginal coverage). *Let $b_t := \sup_{u\in I} |\mathbb{P}\{S_t \le u\} - G_\star(u)|$ be the test-score CDF discrepancy. Then the (deterministic) unconditional coverage probability satisfies*

$$\left|\mathbb{P}\{Y_t \in \hat{C}_t\} - (1-\alpha)\right| \;\le\; b_t + \bar{g}\, r_m(\delta) + \delta. \tag{11}$$

This is an unconditional marginal-coverage statement for the contiguous calibration/test trajectory. It is not *a calibration-conditional coverage guarantee.*

The following oracle condition makes the scores i.i.d., recovering exact split-conformal validity.

Proposition 8 (Oracle exact validity)**.** *If the frozen filter admits standardised innovations $Y_t = \hat{Y}_{t|t-1} + \hat{\Sigma}_{t|t-1}^{1/2} Z_t$ with i.i.d. spherically symmetric (Z_t) independent of the past, then $R_t = Z_t^\top Z_t$ are i.i.d. and Algorithm 1 is exactly split-conformally valid via Theorem 1.*

Plug-in baselines. The baselines STATIC-CGIF, FACTORCGIF, and GROUPCGIF reuse the calibration block to fit $\hat{\Sigma}$. They are therefore *plug-in* rather than exact split-conformal procedures. Their coverage gap includes a leave-one-out stability term controlled by Theorem 22 (App. B.12), with $\varepsilon_m = O(m^{-1})$ for ridge empirical covariance.

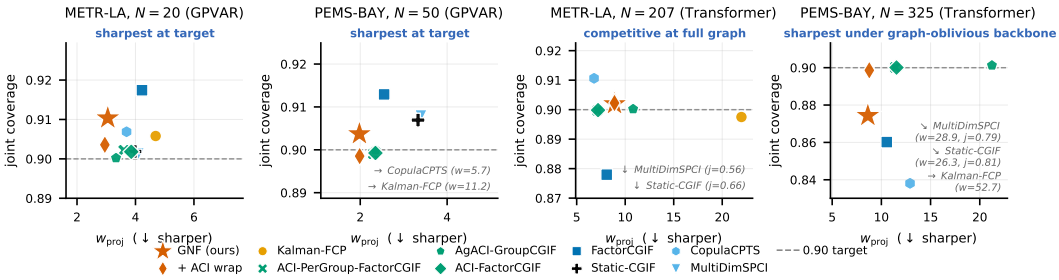

Figure 2: **Mean coordinate-projection length w_{proj} vs. joint coverage on graph-native traffic.** Each panel plots w_{proj} (lower is sharper) against joint coverage at α=0.1; the dashed line marks the 0.90 target. The first two panels correspond to Table 2 (moderate-N, GPVAR backbone) and the last two panels are the full-graph cells under a graph-oblivious Transformer backbone (Table 12); plotted points are 10-seed means (uncertainty in Tables 2 and 12). Table 2 reports the exact deltas: unwrapped-GNF is -8.4% vs. the strongest at-target non-filter on METR-LA-20 and -13.9% on PEMS-BAY-50. Table 3 shows the complementary graph-aware GPVAR regime where factor baselines lead.

5.4 Log-volume sharpness

We finally study the size of the calibrated ellipsoids through their normalised log-volume. Under Gaussian oracle realisability, the next result separates two contributions to excess log-volume: the error in the learned predictive law and the error in estimating the conformal quantile. Here $\widehat{\mathcal{V}}_m(\theta)$ and $\mathcal{V}(\theta)$ are expected normalised log-volumes using empirical and population score quantiles, respectively (expectation convention in App. A.7).

Theorem 9 (Near-oracle log-volume)**.** *Under the spectral and local-density conditions above, assume Gaussian oracle realisability ($P_t^\star = \mathcal{N}(\mu_t^\star, \Sigma_t^\star)$ in the filter family at $\theta_\star$); the comparator class $\mathcal{F}_{\mathrm{cond}}$ consists of Gaussian ellipsoid rules whose population conformal ellipsoid has conditional coverage $\geq 1-\alpha$. Then, on the quantile-concentration event of Theorem 7 (Part I), for errors small enough to stay in the local quantile neighbourhood,*

$$\widehat{\mathcal{V}}_m(\hat{\theta}) \leq \inf_{\theta \in \mathcal{F}_{\mathrm{cond}}} \mathcal{V}(\theta) + C_1 \sqrt{\varepsilon_{\mathrm{NLL}}} + C_2\, r_m(\delta). \tag{12}$$

The comparison is against conditionally valid Gaussian ellipsoid rules *only; arbitrary marginally valid conformal sets can trade coverage across regimes and are not bounded by* (12)*.*

The bounds in (9)–(12) identify the dependence quantities that govern calibration and sharpness. Their constants are not tight finite-sample certificates, so the audits check the relevant quantities directly. As a sanity check, the linear-Kalman case recovers the closed-loop contraction rate exactly (App. B.9, Theorem 19).

6 Experiments

We evaluate three questions. **Q1.** Does a learned time-varying covariance shape improve at-target joint ellipsoid sharpness on graph-native traffic? **Q2.** Are the contraction and score-dependence quantities used by the theory small in the deployed filters? **Q3.** Where does the learned filter stop being the best choice?

Protocol. Two graph-native traffic benchmarks at moderate N (METR-LA busiest-20, PEMS-BAY busiest-50) for the headline comparison; six additional correlated-sensor benchmarks plus the full-graph METR-LA (N=207) and PEMS-BAY (N=325) for scope and contraction validation, 10 real-data evaluation cells in total. Every cell uses the same chronological 40/40/20 train/calib/test split, α=0.1, and 10 seeds (App. E). *At-target* means joint coverage ≥ 0.895 (0.005 slack for 10-seed Monte Carlo variability at $T_{\mathrm{test}} \approx 4000$). Display names used throughout (prose, figures, tables): GNF (learned graph filter), KALMAN-FCP (linear Kalman), STATIC-CGIF (static-covariance baseline), DIAG-GRU (graph-and-rank-removed), GCN-RANK-0 (graph-retained-but-rank-zero).

6.1 Moderate-N graph-native traffic: sharpness gains (Q1)

In the moderate-N graph-native cells, the learned filter GNF gives sharper at-target ellipsoids than every baseline we evaluate (Table 2): $w_{\mathrm{proj}} = 3.05$ and 1.98 at joint coverage 0.910 and 0.904, i.e. $-23.6\%/-40.5\%$ vs. STATIC-CGIF and $-8.4\%/-13.9\%$ vs. the strongest at-target non-filter

Table 2: **Moderate-N graph-native regime: GNF is the sharpest at-target method evaluated.** Mean coordinate-projection length w_{proj} and joint coverage at α=0.1, both 10-seed mean±std; relative w_{proj} vs. STATIC-CGIF. **Bold** indicates the best at-target method (≥ 0.895); "↑" = over-cover. Static, factor and group rows share the GPVAR mean; filter and published joint-CP rows use their own means. DIAG-GRU removes both graph mixing and the low-rank covariance head; GCN-RANK-0 in App. E.13 isolates the covariance rank at fixed graph mixing. Per-seed values are in App. E.12 (Table 15). Conditional-Σ_t, group, spectral, KALMAN-FCP, SPCI, and HOPCPT rows are in Table 14.

	METR-LA (N=20)		PEMS-BAY (N=50)		Δw_{proj} vs STATIC-CGIF	
Method (role)	w_{proj}	joint	w_{proj}	joint	METR-LA	PEMS-BAY
GNF **(ours)**	3.05 ±0.13	0.910 ±0.004	**1.98** ±0.06	0.904 ±0.005	−23.6%	−40.5%
+ ACI wrap	**2.95** ±0.17	0.904 ±0.004	1.99 ±0.07	0.899 ±0.004	−26.1%	−40.2%
Best non-filter at-target†	3.33 ±0.23	0.900 ±0.003	2.30 ±0.09	0.899 ±0.002	−16.5%	−30.9%
STATIC-CGIF (reference)	3.99 ±0.17	0.902 ±0.004	3.33 ±0.11	0.907 ±0.004	—	—
DIAG-GRU (no graph, diag Σ_t)	4.71 ±1.16	0.932 ±0.010↑	2.88 ±0.10	0.921 ±0.008↑	+18.0%	−13.5%
COPULACPTS [24]	3.70 ±0.24	0.907 ±0.004	5.69 ±0.74	0.908 ±0.009	−7.3%	+70.9%
MULTIDIMSPCI [28]	4.10 ±0.17	0.901 ±0.004	3.40 ±0.11	0.908 ±0.004	+2.8%	+2.1%

†AGACI-GROUPCGIF on METR-LA; ACI-PERGROUP-FACTORCGIF on PEMS-BAY.

competitor. On the moderate-N cells the ACI wrap preserves at-target coverage and slightly tightens METR-LA ($3.05 \to 2.95$); at full-graph scale it empirically restores at-target coverage where the unwrapped filter under-covers (PEMS-BAY-325: $0.878 \to 0.899$; §6.4).

Mechanism. The gain is primarily a covariance-shape gain: STATIC-CGIF uses one static covariance, DIAG-GRU learns time variation but no cross-sensor axes, and conditional-Σ_t updates do not close the gap (best at-target EWMA/rolling/local $w_{\text{proj}} = 4.31/3.43$ vs. GNF's $3.05/1.98$; App. E.10). The rank-only sweep drops w_{proj} from 5.32 at r=0 (GCN-RANK-0) to 3.04 at r=4 on METR-LA-20, so the diagonal-plus-low-rank head is the clearest isolated mechanism, and the graph-free LR-GRU-COV control (App. E.22) shows the graph is a modest shape refinement once a learned per-node low-rank covariance is in place. The Theorem 9 log-volume, comparable only across genuine single-ellipsoid rules, agrees with the w_{proj} ranking—GNF improves $\widehat{\mathcal{V}}_m$ by $0.27/0.51$ nats/coordinate over STATIC-CGIF (full table in App. E.17).

6.2 Controlled comparisons (Q1)

Two appendix controls address mean and model-budget confounding on METR-LA-20 and PEMS-BAY-50 (10 paired seeds; each under its own protocol, not mixed with the Table 2 runs; Apps. E.21–E.22). With a common frozen mean and covariance heads fitted only to identical training residuals, GNF-COV is 24–34% sharper in w_{proj} than static full, static factor, EWMA, and rolling covariance on METR-LA, and 35–48% sharper on PEMS-BAY; all paired 95% CIs exclude zero under both the GPVAR and GNF means. A rank-4 DeepVAR-style control with the matched optimisation protocol and identical Mahalanobis conformalisation undercovers and is wider on both cells; the paired DEEPVAR-FCP−GNF differences in w_{proj} are 1.090 ± 0.228 and 3.160 ± 0.181, where "±" denotes a 95% CI half-width. These controls support the covariance-shape interpretation on the two headline cells; they do not extend the claim beyond them.

6.3 Theory diagnostics (Q2)

The audited real-data rates diagnose the conclusions used by the theory but do not verify the population assumptions; they measure distinct objects and should not be conflated. $\widehat{\rho}_{d_{\mathcal{G}}} \in [0.45, 0.46]$ is the empirical emitted-law analogue of the Theorem 3 contraction object (a diagnostic of the conclusion, not a realisability proof). $\widehat{\rho}_D$ is a finite-horizon output-sensitivity proxy relating to Assumption 2(O3) (0.48–0.50 in the cheap audit configuration, 0.86–0.89 on the deployed model): it diagnoses weakly observable directions on the deployed trajectory and does *not* certify the population quotient inequality for c_{obs} (App. E.16). $\widehat{\rho}_{\text{score}} \in [0.140, 0.153]$ is the downstream score-CDF forgetting rate for Theorem 12, and $\widehat{\rho}_\Gamma \approx 0.85$ with integrated autocorrelation time $\tau_{\text{int}} \in [10, 13]$ probes the threshold-dependence regime of the Chebyshev form of Theorem 7 and Assumption 6. The audited τ_{int} makes the effective calibration size roughly an order of magnitude below the raw size, consistent with dependence-aware rather than exchangeability-based coverage; the linear-Kalman check confirms $\widehat{\rho}_{d_{\mathcal{G}}} \approx \sigma_1$ (App. E.5), the warm-up sweep renders the burn-in bias negligible at $t_0 \geq 50$ (App. E.14), protocols are in Apps. E.4–E.6 and per-cell values in App. E.7 (Table 10).

Table 3: **GPVAR scale sweep — moderate-N win, full-graph crossover.** Mean coordinate-projection length w_{proj} (10-seed mean±std) and joint coverage (10-seed mean) at α=0.1; joint-coverage standard deviations for the moderate-N cells appear in Table 2 and per-seed joint coverage in App. E.12 (Table 15). **bold**/underlined = best/second at-target; "↓" flags joint < 0.895. At full graph all seeds use the same sensor set and the GPVAR fit is deterministic, so methods with deterministic fitting are seed-invariant (±0.00). This table fixes a graph-aware GPVAR mean backbone; under this backbone factor methods become sharper at full graph scale. Figure 2 and Table 12 show the complementary graph-oblivious Transformer/GRU regime.

	METR-LA				PEMS-BAY			
	N=20		N=207		N=50		N=325	
Method	w_{proj}	joint	w_{proj}	joint	w_{proj}	joint	w_{proj}	joint
GNF **(ours)**	3.05 ±0.13	0.910	9.14 ±0.24	0.902	**1.98** ±0.06	0.904	8.58 ±0.28	0.878 ↓
+ ACI wrap	**2.95** ±0.17	0.904	9.00 ±0.60	0.902	1.99 ±0.07	0.899	9.01 ±0.38	0.899
ACI-PERGROUP-FACTORCGIF	3.60 ±0.28	0.902	7.21 ±0.00	0.902	2.30 ±0.09	0.899	**5.64** ±0.00	0.901
COPULACPTS [24]	3.70 ±0.24	0.907	**6.73** ±0.04	0.914	5.69 ±0.74	0.908	14.4 ±1.41	0.894 ↓
MULTIDIMSPCI [28]	4.10 ±0.17	0.901	12.6 ±0.00	0.877 ↓	3.40 ±0.11	0.908	12.7 ±0.00	0.886 ↓
STATIC-CGIF (static $\hat{\Sigma}$)	3.99 ±0.17	0.902	11.8 ±0.00	0.880 ↓	3.33 ±0.11	0.907	12.0 ±0.00	0.887 ↓
KALMAN-FCP (tuned Kalman)	4.69 ±0.28	0.906	21.2 ±0.00	0.900	11.2 ±1.03	0.897	51.4 ±0.00	0.852 ↓

6.4 Full graphs and scope (Q3)

Full-graph performance depends on the mean predictor. Under graph-aware backbones factor baselines are sharper at N=207, 325 ($w_{\text{proj}} = 5.64$ vs. 9.01 on PEMS-BAY-325). On PEMS-BAY-325, under graph-oblivious GRU/Transformer backbones the factor baseline widens to $w_{\text{proj}} = 11.4$–13.5 while GNF+ACI remains at 8.8–9.0 (up to 34% narrower). The ACI wrap restores at-target coverage where the unwrapped filter under-covers (App. E.9, Table 12). KALMAN-FCP produces much wider sets at full scale (51.4 on PEMS-BAY-325, below target); on non-graph-native data COPULACPTS (SOLAR, LOOP-SEA) or CGIF-family baselines (ETT, JENA) are sharpest at-target (Table 13).

What a high-dimensional joint set buys. The joint region controls one deployment-level event $\{Y_t \in \hat{C}_t\}$, relevant when a decision depends on the complete network state (capacity, routing, aggregate safety, anomaly detection), not a replacement for a marginal interval when only one sensor matters. Any operational linear functional is available in closed form, $a^\top Y_t \in a^\top \hat{\mu}_t \pm \sqrt{\hat{q}_t\, a^\top \hat{\Sigma}_t a}$ (§3); such projections inherit the simultaneous guarantee but need not be shortest, and simultaneous coverage grows conservative as N grows—why our strongest claim is at moderate N and we report the crossover explicitly.

7 Discussion and limitations

Main takeaway and scope. Learned conditional covariance shapes improve sharpness on moderate-dimensional graph-native traffic. At full-graph scale, factor methods can be sharper when the mean backbone already uses graph structure. The same-mean and matched-budget controls (§6.2) support the covariance-shape interpretation on two datasets; the full-graph and non-graph-native crossovers limit its empirical scope.

Theory and validity. Theorem 3 transfers NLL accuracy to contraction under explicit assumptions on each filter's own observable quotient; it does not guarantee contraction for arbitrary trained GRUs. With input sensitivity and score dependence, the result gives unconditional approximate marginal validity, without calibration-conditional coverage. Exact validity is recovered under the standardised-innovation condition (Proposition 8). The audits probe finite-trajectory conclusions and proxies, rather than population assumptions. Theorem 9 compares only conditionally valid Gaussian ellipsoid rules.

Architecture and numerical limitations. The node-shared recurrence keeps the parameter count independent of N, but fixed shared capacity and a single graph shift may lose node-specific information at large N. The covariance floor is nonbinding in the audited runs, although a high-condition-number tail remains (App. E.23). GNF uses no per-dataset hyperparameter search; only the Kalman baseline receives per-cell tuning (Table 21).

Broader impacts. Joint uncertainty sets can support routing and capacity decisions. Interpreting marginal coverage as a guarantee for every operating condition, or deploying under unmodelled distribution shifts, can give unjustified confidence in those decisions.

Acknowledgments and Disclosure of Funding

The author thanks the anonymous reviewers and area chairs for their careful reading and constructive feedback, which sharpened the statement of the observable-quotient theory and improved the design and reporting of the empirical controls.

Funding. This work was carried out without external funding. The compute used for training, calibration, and the auditing experiments was drawn from the author's ordinary research resources; its hardware and wall-clock footprint is documented in App. E.19.

Competing interests. The author is employed by DRW. The views expressed here are those of the author and do not necessarily represent those of DRW.

Code and reproducibility. The full code base, including a self-contained reproducibility suite, is publicly available under an MIT licence (App. F). Seeds, run counts, and dataset provenance are described in App. F; the exact commands used to regenerate every reported number are documented in the repository.

References

[1] Brian D.O. Anderson and John B. Moore. *Optimal Filtering*. Prentice-Hall, 1979.

[2] Andreas Auer, Martin Gauch, Daniel Klotz, and Sepp Hochreiter. Conformal prediction for time series with modern Hopfield networks. In *Advances in Neural Information Processing Systems*, 2023.

[3] Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

[4] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[5] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of EMNLP*, 2014.

[6] Andrea Cini, Alexander Jenkins, Danilo Mandic, Cesare Alippi, and Filippo Maria Bianchi. Relational conformal prediction for correlated time series. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 10949–10965, 2025.

[7] Randal Douc, Eric Moulines, Pierre Priouret, and Philippe Soulier. *Markov Chains*. Springer, 2018.

[8] Sonakshi Dua, Gonzalo Mateos, and Sundeep Prabhakar Chepuri. Conformal inference for time series over graphs. *arXiv preprint arXiv:2510.11049*, 2025.

[9] Isaac Gibbs and Emmanuel J. Candès. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 2021.

[10] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *International Conference on Artificial Intelligence and Statistics*, pages 249–256, 2010.

[11] Martin Hairer and Jonathan C. Mattingly. Yet another look at Harris' ergodic theorem for Markov chains. In *Seminar on Stochastic Analysis, Random Fields and Applications VI*, volume 63 of *Progress in Probability*, pages 109–117. Springer, 2011.

[12] G. H. Hardy, J. E. Littlewood, and G. Pólya. *Inequalities*. Cambridge University Press, 2nd edition, 1952.

[13] Kexin Huang, Ying Jin, Emmanuel J. Candès, and Jure Leskovec. Uncertainty quantification over graph with conformalized graph neural networks. In *Advances in Neural Information Processing Systems*, volume 36, pages 26699–26721, 2023.

[14] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[15] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

[16] Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

[17] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *International Conference on Learning Representations*, 2018.

[18] Xuan Liang, Tao Zou, Bin Guo, Shuo Li, Haozhe Zhang, Shuyi Zhang, Hui Huang, and Song Xi Chen. Assessing Beijing's $PM_{2.5}$ pollution: Severity, weather impact, APEC and winter heating. *Proceedings of the Royal Society A*, 471(2182):20150257, 2015.

[19] Syama Sundar Rangapuram, Matthias W. Seeger, Jan Gasthaus, Lorenzo Stella, Yuyang Wang, and Tim Januschowski. Deep state space models for time series forecasting. In *Advances in Neural Information Processing Systems*, 2018.

[20] David Salinas, Michael Bohlke-Schneider, Laurent Callot, Roberto Medico, and Jan Gasthaus. High-dimensional multivariate forecasting with low-rank Gaussian copula processes. In *Advances in Neural Information Processing Systems*, 2019.

[21] David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3): 1181–1191, 2020.

[22] Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations*, 2014.

[23] Kamile Stankeviciute, Ahmed M. Alaa, and Mihaela van der Schaar. Conformal time-series forecasting. In *Advances in Neural Information Processing Systems*, 2021.

[24] Sophia Sun and Rose Yu. Copula conformal prediction for multi-step time series prediction. In *International Conference on Learning Representations*, 2024.

[25] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

[26] Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. Graph WaveNet for deep spatial-temporal graph modeling. In *International Joint Conference on Artificial Intelligence*, 2019.

[27] Chen Xu and Yao Xie. Sequential predictive conformal inference for time series. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 38707–38727. PMLR, 2023.

[28] Chen Xu, Hanyang Jiang, and Yao Xie. Conformal prediction for multi-dimensional time series by ellipsoidal sets. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 55076–55099. PMLR, 2024.

[29] Margaux Zaffran, Olivier Féron, Yannig Goude, Julie Josse, and Aymeric Dieuleveut. Adaptive conformal predictions for time series. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 25834–25866. PMLR, 2022.

[30] Jifan Zhang, Fangxin Wang, Zihe Song, Philip S. Yu, Kaize Ding, and Shixiang Zhu. Topology-aware conformal prediction for stream networks. In *Advances in Neural Information Processing Systems*, volume 38, pages 88976–89008, 2025.

A Observable-contraction theory

This section collects the proofs of the main theorems (Theorem 3, Corollary 5, Theorem 7, Theorem 9) and defines the metric and identification objects they rely on.

A.1 Bures-Wasserstein metric on Gaussian predictive laws

Recall (5). For Gaussian laws on $\mathbb{R}^N$ with covariances spectrum-bounded between $0 < \lambda_- \le \lambda_+ < \infty$, the following relations hold on a compact eigenvalue interval determined only by λ_-/λ_+:

$$\begin{aligned} c_\lambda\, d_{\mathcal{G}}^2(P,P') \;&\le\; \mathrm{KL}(P\,\|\,P') \;\le\; C_\lambda\, d_{\mathcal{G}}^2(P,P'),\\ c_\lambda\, \|\Sigma-\Sigma'\|_F^2 \;&\le\; \mathrm{tr}\big[\Sigma+\Sigma'-2(\Sigma'^{1/2}\Sigma\Sigma'^{1/2})^{1/2}\big] \;\le\; C_\lambda\, \|\Sigma-\Sigma'\|_F^2. \end{aligned} \tag{13}$$

The inequalities follow from the bound $x - \log x - 1 \ge c_\lambda (x-1)^2$ on the eigenvalue interval of $\Sigma'^{-1/2}\Sigma\Sigma'^{-1/2}$, combined with the closed-form Gaussian KL. Thus Gaussian KL is comparable to squared $d_{\mathcal{G}}$, which in turn is comparable to $\|\mu-\mu'\|_2^2 + \|\Sigma-\Sigma'\|_F^2$, with constants depending only on λ_-, λ_+. Pinsker TV $\le \sqrt{\mathrm{KL}/2}$ then bounds total variation by a constant multiple of $d_{\mathcal{G}}$.

A.2 Parameter-specific observable quotient and metric

Fix a filter parameter θ and let the filter have a (non-minimal) hidden state h emitting $O_\theta(h) = \mathcal{N}(\mu_\theta(h), \Sigma_\theta(h))$. For θ fixed define the equivalence relation

$$h \sim_\theta h' \iff O_\theta\big(H_{\theta,\ell}^{h,\mathbf{u}}\big) = O_\theta\big(H_{\theta,\ell}^{h',\mathbf{u}}\big) \quad \forall\, \ell \ge 0,\ \forall\, \mathbf{u}, \tag{14}$$

where $\mathbf{u}$ ranges over admissible common input continuations and $H_{\theta,\ell}^{h,\mathbf{u}}$ (defined below) is the step-ℓ hidden state; equivalently, h and h' emit the same future predictive-law sequence under every such continuation. We set $\mathcal{Q}_\theta := H/\!\sim_\theta$ and do *not* identify elements of $\mathcal{Q}_{\theta_\star}$ with elements of $\mathcal{Q}_{\widehat{\theta}}$; each parameter has its own quotient. For the GCN-GRU GNF, $\mathcal{Q}_\theta$ removes exactly the hidden directions invisible to g_θ along admissible inputs.

Metric. Let $H_{\theta,\ell}^{h,\mathbf{u}}$ denote the hidden state reached at step ℓ from h under input continuation $\mathbf{u}$, write $\overline{d}_{\mathcal{G}}(P,P') := \min\{1, d_{\mathcal{G}}(P,P')\}$, and fix summable weights $w_\ell > 0$, $\sum_\ell w_\ell < \infty$. Define

$$d_\theta\big([h]_\theta, [h']_\theta\big) := \sup_{\mathbf{u}\in\mathcal{U}^\infty} \Big[\sum_{\ell\ge 0} w_\ell\, \overline{d}_{\mathcal{G}}^{\,2}\big(O_\theta(H_{\theta,\ell}^{h,\mathbf{u}}), O_\theta(H_{\theta,\ell}^{h',\mathbf{u}})\big)\Big]^{1/2}. \tag{15}$$

Truncating $d_{\mathcal{G}}$ at 1 before squaring makes both finiteness and the triangle inequality transparent. We record the four required properties. *(1) Representative-independence.* If $h_1 \sim_\theta h_2$ and $h_1' \sim_\theta h_2'$ then, by (14), the two emitted sequences agree for all $\mathbf{u}$, so the summand in (15) is unchanged; d_θ is thus well-defined on classes. *(2) Vanishing exactly on $\sim_\theta$.* $d_\theta([h]_\theta,[h']_\theta) = 0$ iff $\overline{d}_{\mathcal{G}}(O_\theta(H_{\theta,\ell}^{h,\mathbf{u}}), O_\theta(H_{\theta,\ell}^{h',\mathbf{u}})) = 0$ for all $\ell, \mathbf{u}$ (all $w_\ell > 0$), i.e. iff $h \sim_\theta h'$. *(3) Finiteness and triangle inequality.* Each summand is $\le w_\ell$, so $d_\theta \le (\sum_\ell w_\ell)^{1/2} < \infty$; $\overline{d}_{\mathcal{G}}$ is a metric on Gaussian laws, so for each $(\ell, \mathbf{u})$ the weighted-ℓ^2 triangle inequality holds, and taking the supremum over $\mathbf{u}$ preserves it. *(4) Transition well-defined on $\mathcal{Q}_\theta$.* If $h \sim_\theta h'$ then applying the same first input y yields states that still agree under every future continuation (by (14) shifted by one step), so $T_\theta(y,h) \sim_\theta T_\theta(y,h')$ and the induced map $T_\theta(y,\cdot)$ on $\mathcal{Q}_\theta$ is well-defined.

Finite-horizon observable discrepancy. Distinguished from the infinite-horizon quotient metric, we also use, for horizon $H_{\mathrm{obs}} \ge 0$ and *one fixed* probing distribution ν on admissible continuations that is *independent of* θ,

$$D_{\theta,H_{\mathrm{obs}}}^2(q,q') \;:=\; \sum_{\ell=0}^{H_{\mathrm{obs}}} \beta_\ell\, \mathbb{E}_\nu\big[d_{\mathcal{G}}^2\big(O_\theta(Q_{\theta,\ell}^{q}), O_\theta(Q_{\theta,\ell}^{q'})\big)\big], \qquad \beta_\ell > 0, \tag{16}$$

where both trajectories use the same future input continuation. Fixing ν and the β_ℓ independently of θ is what makes $D_{\theta,H_{\mathrm{obs}}}$ a genuine per-parameter discrepancy rather than a quantity that moves through both the filter and the probe. Assumption 2(O3) states the local observability relation $c_{\mathrm{obs}}\, d_\theta(q,q') \le D_{\theta,H_{\mathrm{obs}}}(q,q')$ together with the emitted-law Lipschitz relation $d_{\mathcal{G}}(O_\theta(q), O_\theta(q')) \le L_{\mathrm{emit}}\, d_\theta(q,q')$. Instantaneous observability is the case $H_{\mathrm{obs}} = 0$, $\beta_0 = 1$, where $D_{\theta,H_{\mathrm{obs}}}(q,q') = d_{\mathcal{G}}(O_\theta(q), O_\theta(q'))$.

A.3 Risk identifies the contraction modulus: Fisher margin

Assumption 2(O4) says NLL excess risk controls the *scalar* contraction-modulus gap $\Delta_{\mathrm{ctr}}(\theta) = [\kappa(\theta) - \kappa(\theta_\star)]_+$, where $\kappa(\theta) = \sup_y \sup_{q \neq q'} d_\theta(T_\theta(y,q), T_\theta(y,q'))/d_\theta(q,q')$. This removes any need to compare points or transition maps across two different quotient spaces. We give a sufficient condition on that one scalar.

Lemma 10 (Fisher margin, scalar-modulus form)**.** *Let Θ_0 be an identifiable local chart around $\theta_\star$ (parameter symmetries such as low-rank rotations $L \mapsto LQ$ quotiented out). Suppose the Gaussian-NLL population risk has a local quadratic margin $\mathcal{R}(\theta) - \mathcal{R}(\theta_\star) \geq \lambda_R \|\theta - \theta_\star\|^2$ on Θ_0, and the contraction modulus is locally Lipschitz, $[\kappa(\theta) - \kappa(\theta_\star)]_+ \leq L_\kappa \|\theta - \theta_\star\|$ on Θ_0. Then*

$$\Delta_{\mathrm{ctr}}(\theta)^2 \;\leq\; \frac{L_\kappa^2}{\lambda_R}\,[\mathcal{R}(\theta) - \mathcal{R}(\theta_\star)],$$

i.e. (O4) *holds with $C_{\mathrm{id}} = L_\kappa^2/\lambda_R$.*

Proof. $\Delta_{\mathrm{ctr}}(\theta)^2 \leq L_\kappa^2 \|\theta - \theta_\star\|^2 \leq (L_\kappa^2/\lambda_R)[\mathcal{R}(\theta) - \mathcal{R}(\theta_\star)]$. □

We present local Lipschitzness of the scalar modulus as the sufficient regularity condition; we do *not* claim that a generic transition-smoothness argument automatically proves local Lipschitzness of $\kappa(\cdot)$. The Fisher-margin calculation below supplies the risk side (λ_R).

Sufficient condition for the quadratic risk margin (GCN-GRU). Evaluate the conditional expected Gaussian-NLL Hessian at the Bayes projection moments $\eta_\star = (\mu_\star, \Sigma_\star)$. Its quadratic form on a mean perturbation a and symmetric covariance perturbation B is $a^\top \Sigma_\star^{-1} a + \frac{1}{2}\operatorname{tr}(\Sigma_\star^{-1} B \Sigma_\star^{-1} B)$, so it is bounded below by $c_G(\|a\|_2^2 + \|B\|_F^2)$ under the spectral bounds. The conditional expected score vanishes there. Let $J_t = \partial \eta_{\theta,t}/\partial\theta|_{\theta_\star}$ and suppose the observable Fisher Gramian $\mathbb{E}[J_t^\top J_t] \succeq \gamma I$ after quotienting parameter symmetries. Assuming differentiation under expectation is valid, $\nabla_\theta^2 \mathcal{R}(\theta_\star) = \mathbb{E}[J_t^\top \nabla_\eta^2 \mathbb{E}[\ell_t \mid \mathcal{F}_{t-1}]|_{\eta_\star} J_t] \succeq c_G \gamma I$, and continuity of the Hessian gives the quadratic margin $\mathcal{R}(\theta) - \mathcal{R}(\theta_\star) \geq (c_G\gamma/4)\|\theta - \theta_\star\|^2 = \lambda_R \|\theta - \theta_\star\|^2$ on a neighbourhood of $\theta_\star$. The condition $\mathbb{E}[J_t^\top J_t] \succeq \gamma I$ is the formal version of *finite-horizon Fisher observability*: likelihood training identifies the parts of the parameter that move future emitted predictive laws. Paired with local Lipschitzness of $\kappa(\cdot)$ it yields (O4).

A.4 Proof of Theorem 3

The proof proceeds entirely within the learned filter's own quotient $\mathcal{Q}_{\widehat{\theta}}$; there is no comparison of $T_{\widehat{\theta}}$ and $T_{\theta_\star}$ in a common metric.

Step 1 (modulus identification). By Assumption 2(O4), $\Delta_{\mathrm{ctr}}(\widehat{\theta})^2 \leq C_{\mathrm{id}} \varepsilon_{\mathrm{NLL}}$, i.e. $[\kappa(\widehat{\theta}) - \kappa(\theta_\star)]_+ \leq \sqrt{C_{\mathrm{id}} \varepsilon_{\mathrm{NLL}}}$. With (O1) ($\kappa(\theta_\star) \leq \rho_\star$) this gives

$$\kappa(\widehat{\theta}) \;\leq\; \kappa(\theta_\star) + \sqrt{C_{\mathrm{id}} \varepsilon_{\mathrm{NLL}}} \;\leq\; \rho_\star + \sqrt{C_{\mathrm{id}} \varepsilon_{\mathrm{NLL}}} \;=\; \widehat{\rho}_{\mathrm{obs}} \;<\; 1.$$

Step 2 (quotient contraction). By definition of the modulus, $d_{\widehat{\theta}}(T_{\widehat{\theta}}(y,q), T_{\widehat{\theta}}(y,q')) \leq \kappa(\widehat{\theta})\, d_{\widehat{\theta}}(q,q') \leq \widehat{\rho}_{\mathrm{obs}}\, d_{\widehat{\theta}}(q,q')$ for every admissible y. Iterating t times along the same input path (the transition is well-defined on $\mathcal{Q}_{\widehat{\theta}}$, App. A.2) gives (6), $d_{\widehat{\theta}}(Q^q_{\widehat{\theta},t}, Q^{q'}_{\widehat{\theta},t}) \leq \widehat{\rho}^{\,t}_{\mathrm{obs}} d_{\widehat{\theta}}(q,q')$.

Step 3 (emitted-law Lipschitzness). By the (O3) Lipschitz relation, $d_{\mathcal{G}}(P^q_{\widehat{\theta},t}, P^{q'}_{\widehat{\theta},t}) = d_{\mathcal{G}}(O_{\widehat{\theta}}(Q^q_{\widehat{\theta},t}), O_{\widehat{\theta}}(Q^{q'}_{\widehat{\theta},t})) \leq L_{\mathrm{emit}}\, d_{\widehat{\theta}}(Q^q_{\widehat{\theta},t}, Q^{q'}_{\widehat{\theta},t}) \leq L_{\mathrm{emit}}\, \widehat{\rho}^{\,t}_{\mathrm{obs}}\, d_{\widehat{\theta}}(q,q')$.

Step 4 (finite-horizon observability at the initial states). The (O3) lower bound at the initial learned-filter quotient states gives $d_{\widehat{\theta}}(q,q') \leq D_{\widehat{\theta},H_{\mathrm{obs}}}(q,q')/c_{\mathrm{obs}}$.

Step 5 (combine). Substituting Step 4 into Step 3, $d_{\mathcal{G}}(P^q_{\widehat{\theta},t}, P^{q'}_{\widehat{\theta},t}) \leq (L_{\mathrm{emit}}/c_{\mathrm{obs}})\widehat{\rho}^{\,t}_{\mathrm{obs}} D_{\widehat{\theta},H_{\mathrm{obs}}}(q,q')$, which is (7). Under instantaneous observability ($H_{\mathrm{obs}} = 0$, $\beta_0 = 1$), $D_{\widehat{\theta},H_{\mathrm{obs}}}(q,q')$ collapses to $d_{\mathcal{G}}(P^q_{\widehat{\theta},0}, P^{q'}_{\widehat{\theta},0})$, recovering the initial-emitted-law form.

A.5 Proof of Corollary 5

Let $Z_i = (H_{\widehat{\theta},i}, Y_i)$ be the augmented filter+data state and $a_u(z) := \mathbb{P}(S_{\widehat{\theta},i+1} \le u \mid Z_i = z)$.

Augmented coupling. Couple two copies $(Z_i), (Z_i')$ driven by the geometrically coupled inputs of Assumption 4. The input-sensitivity recursion $d_{\widehat{\theta}}(T_{\widehat{\theta}}(y,q), T_{\widehat{\theta}}(y',q')) \le \widehat{\rho}_{\rm obs}\, d_{\widehat{\theta}}(q,q') + L_{\rm in}\, d_Y(y,y')$ unrolled against $\mathbb{E}\, d_Y(Y_t, Y_t') \le C_Y \rho_{\rm data}^t d_Y(Y_0, Y_0')$ gives, in expectation, a filter-gap bound $\mathbb{E}\, d_{\widehat{\theta}}(Q_t, Q_t') \le C \sum_{s\le t} \widehat{\rho}_{\rm obs}^{t-s} \rho_{\rm data}^s\, d_Y(Y_0, Y_0') + \widehat{\rho}_{\rm obs}^t d_{\widehat{\theta}}(q,q')$. Two cases. (i) If $\rho_{\rm data} \ne \widehat{\rho}_{\rm obs}$ the geometric sum is bounded by $C' \max\{\rho_{\rm data}, \widehat{\rho}_{\rm obs}\}^t$, so the augmented state couples at any $\rho_{\rm aug} = \max\{\rho_{\rm data}, \widehat{\rho}_{\rm obs}\}$ with an adjusted constant. (ii) If $\rho_{\rm data} = \widehat{\rho}_{\rm obs} =: \rho$ the convolution yields $t\,\rho^t$, which for any target $\rho_{\rm aug} \in (\rho, 1)$ satisfies $t\rho^t \le C_{\rho_{\rm aug}} \rho_{\rm aug}^t$; hence the augmented state couples at any $\rho_{\rm aug}$ strictly above ρ. In both cases $\mathbb{E}[d_Z(Z_{i+k}^{Z_i}, Z_{i+k}^{Z_i'}) \mid Z_i, Z_i'] \le C_Z \rho_{\rm aug}^k d_Z(Z_i, Z_i')$ with $\rho_{\rm aug} < 1$; we do *not* assert coupling at exactly $\max\{\rho_{\rm data}, \widehat{\rho}_{\rm obs}\}$ when the rates coincide.

Envelope. Local Lipschitzness of a_u gives $|\mathbb{E}[I_{i+k}(u) \mid Z_i] - \mathbb{E} I_{i+k}(u)| \le L_a C_Z \rho_{\rm aug}^k \mathbb{E}[d_Z(Z_i, Z_i') \mid Z_i]$, and the tower-property identity for covariance plus $|I_i(u) - \mathbb{E} I_i(u)| \le 1$ yields $|\mathrm{Cov}(I_i(u), I_{i+k}(u))| \le L_a C_Z \mathbb{E}[d_Z(Z_i, Z_i')]\, \rho_{\rm aug}^k =: C_\Gamma \rho_{\rm aug}^k$. The frozen-filter contraction premise $\widehat{\rho}_{\rm obs} < 1$ is supplied by Theorem 3 for a likelihood-trained filter, but the argument uses only that premise together with Assumption 4, not Gaussian oracle realisability.

A.6 Proof of Theorem 7

We prove the two logically separate conclusions in turn: Part I is a high-probability bound on the calibration-quantile error, and Part II is a deterministic unconditional-coverage statement obtained from the quantile event by pathwise inclusion (no calibration/test independence is used).

Part I: quantile-error radius $r_m(\delta)$. Write $p = 1-\alpha$, $s_\star = (G_\star)^{-1}(p)$, $k = \lceil (m+1)p \rceil$, $\hat{s}_m = S_{(k)}$, and $\eta_m = \max\{|k/m - p|, |(k-1)/m - p|\} = O(m^{-1})$ for the conformal rank offset; assume $k \le m$ and the resulting thresholds lie in I. Let $b_m = \sup_{u\in I} |\mathbb{E}\hat{F}_m(u) - G_\star(u)|$; the augmented-process coupling and local conditional-CDF regularity used in Corollary 5 (App. B.8) give the warm-start bias $b_m \lesssim \rho_{\rm aug}^{t_0}$. For ε with $s_\star \pm \varepsilon \in I$, the density lower bound gives $G_\star(s_\star + \varepsilon) \ge p + \underline{g}\,\varepsilon$ and $G_\star(s_\star - \varepsilon) \le p - \underline{g}\,\varepsilon$. If $\hat{s}_m > s_\star + \varepsilon$ then $\hat{F}_m(s_\star + \varepsilon) \le (k-1)/m \le p + \eta_m$, so $\hat{F}_m(s_\star + \varepsilon) - \mathbb{E}\hat{F}_m(s_\star + \varepsilon) \le -(\underline{g}\,\varepsilon - b_m - \eta_m)$; symmetrically at $s_\star - \varepsilon$. Hence whenever $\underline{g}\,\varepsilon > b_m + \eta_m$,

$$\mathbb{P}(|\hat{s}_m - s_\star| > \varepsilon) \le \sum_{\sigma\in\{\pm 1\}} \mathbb{P}\big(|\hat{F}_m(s_\star + \sigma\varepsilon) - \mathbb{E}\hat{F}_m(s_\star + \sigma\varepsilon)| \ge \underline{g}\,\varepsilon - b_m - \eta_m\big). \tag{17}$$

Chebyshev radius. By Corollary 5, $\sum_{k\ge 1} \Gamma_k \le C_\Gamma \rho_{\rm aug}/(1-\rho_{\rm aug})$, so $\mathrm{Var}(\hat{F}_m(u)) \le 1/(4 m_{\rm eff}^{\rm cov})$ with $m_{\rm eff}^{\rm cov} \asymp m/(1 + 8 C_\Gamma \rho_{\rm aug}/(1-\rho_{\rm aug}))$. Chebyshev at level δ and (17) give $|\hat{s}_m - s_\star| \le r_m^{\rm Cheb}(\delta)$ with probability $\ge 1-\delta$, where $\underline{g}\, r_m^{\rm Cheb} \lesssim \rho_{\rm aug}^{t_0} + (\delta m_{\rm eff}^{\rm cov})^{-1/2} + m^{-1}$, i.e. (9). ***Bernstein radius.*** Under Assumption 6, $\mathbb{P}(|\hat{F}_m(u) - \mathbb{E}\hat{F}_m(u)| > \varepsilon) \le 2\exp(-c m_{\rm eff} \varepsilon^2)$ with $m_{\rm eff} \asymp m(1-\rho_{\rm aug})/(1+\rho_{\rm aug})$; inverting and using (17) gives $|\hat{s}_m - s_\star| \le r_m^{\rm Bern}(\delta)$ with $\underline{g}\, r_m^{\rm Bern} \lesssim \rho_{\rm aug}^{t_0} + \sqrt{\log(1/\delta)/m_{\rm eff}} + \log(1/\delta)/m_{\rm eff} + m^{-1}$, i.e. (10). In both cases the exact constants, the local-density factors $\underline{g}, \bar{g}$, the rank offset η_m, and the calibration-bias term are retained as displayed. Corollary 5 alone supports the Chebyshev radius (its covariance summability controls only $\mathrm{Var}(\hat{F}_m)$, giving a $(\delta m_{\rm eff}^{\rm cov})^{-1/2}$ deviation); the sub-Gaussian $\sqrt{\log(1/\delta)/m_{\rm eff}}$ radius requires the strictly stronger Assumption 6, whose sufficient conditions include geometric β-mixing of the augmented filter+data state [7], a Markov-coupling spectral gap, or exponential ϕ-mixing.

Part II: unconditional coverage via the quantile event. Let $E_\delta := \{|\hat{s}_m - s_\star| \le r_m(\delta)\}$, so $\mathbb{P}(E_\delta^c) \le \delta$ by Part I. On E_δ, pathwise,

$$\{S_t \le s_\star - r_m(\delta)\} \;\subseteq\; \{S_t \le \hat{s}_m\} \;\subseteq\; \{S_t \le s_\star + r_m(\delta)\}. \tag{18}$$

This inclusion is deterministic and uses *no* independence between calibration and test. Taking probabilities and adding $\mathbb{P}(E_\delta^c) \le \delta$ to both sides,

$$\mathbb{P}\{S_t \le s_\star - r_m(\delta)\} - \delta \;\le\; \mathbb{P}\{Y_t \in \hat{C}_t\} \;\le\; \mathbb{P}\{S_t \le s_\star + r_m(\delta)\} + \delta.$$

Write the test-score CDF discrepancy $b_t := \sup_{u\in I} |\mathbb{P}\{S_t \le u\} - G_\star(u)|$ and use the density upper bound $\bar{g}$: $|G_\star(s_\star \pm r_m) - (1-\alpha)| \le \bar{g}\, r_m(\delta)$ and $|\mathbb{P}\{S_t \le s_\star \pm r_m\} - G_\star(s_\star \pm r_m)| \le b_t$. Hence $|\mathbb{P}\{Y_t \in \hat{C}_t\} - (1-\alpha)| \le b_t + \bar{g}\, r_m(\delta) + \delta$, which is (11). Because $S_t(y) \le \hat{s}_m \Leftrightarrow R_t(y) \le \hat{s}_m^2 = \hat{q}_{1-\alpha}$, this is the coverage gap of $\hat{C}_t$. The statement is an unconditional marginal-coverage bound for the contiguous calibration/test trajectory; it aligns Theorem 7 with the deterministic form of Proposition 15, and is *not* a calibration-conditional guarantee.

A.7 Proof of Theorem 9

Expectation convention. Fix the trained parameter θ. Write s_θ for the stationary population root-score quantile and $\hat{s}_{m,\theta}$ for its calibration order statistic. With $c_N = N^{-1}\log v_N$, define

$$\mathcal{V}(\theta) = c_N + \log s_\theta + \frac{1}{2N}\mathbb{E}_{\mathrm{stat}} \log\det \hat{\Sigma}_{\theta,t}, \qquad \widehat{\mathcal{V}}_m(\theta) = c_N + \log \hat{s}_{m,\theta} + \frac{1}{2N}\mathbb{E}_{\mathrm{stat}} \log\det \hat{\Sigma}_{\theta,t}.$$

The expectation integrates the stationary test-state marginal, keeping the calibrated radius fixed; it is not a conditional expectation of the actual dependent test state given calibration. Thus $\widehat{\mathcal{V}}_m$ is random through its calibration radius, and the theorem holds on the stated calibration-quantile event. Experimental tables estimate the determinant expectation by a test-trajectory average.

Step 1: volume formula. For θ in the local class, the ellipsoid $\hat{C}_{\theta,t}(s) = \{y : S_{\theta,t}(y) \le s\}$ is the affine image of the Euclidean ball of radius s under $\hat{\Sigma}_{\theta,t}^{1/2}$ centred at $\mu_{\theta,t}$, so $\mathrm{Vol}(\hat{C}_{\theta,t}(s)) = v_N s^N \det(\hat{\Sigma}_{\theta,t})^{1/2}$ and (4) holds.

Step 2: oracle minimisation among $\mathcal{F}_{\mathrm{cond}}$. Under Gaussian oracle realisability, the conditional law $P_t^\star = \mathcal{N}(\mu_t^\star, \Sigma_t^\star)$ has Gaussian density f_t, and the Bayes-projection ellipsoid $\hat{C}_{\star,t} = \{y : f_t(y) \ge c_t\}$ at the threshold making $P_t^\star(\hat{C}_{\star,t}) = 1-\alpha$ has Lebesgue volume *minimal* among Borel sets of conditional coverage $\ge 1-\alpha$, by a likelihood-ratio level-set argument: for any other set A_t with $P_t^\star(A_t) \ge 1-\alpha$, $0 \le \int_{A_t\setminus\hat{C}_{\star,t}} f_t - \int_{\hat{C}_{\star,t}\setminus A_t} f_t \le c_t[\mathrm{Leb}(A_t) - \mathrm{Leb}(\hat{C}_{\star,t})]$. Hence $\inf_{\theta\in\mathcal{F}_{\mathrm{cond}}} \mathcal{V}(\theta) = \mathcal{V}(\theta_\star)$.

Step 3: NLL excess controls the determinant term. For Gaussian laws on $\mathcal{G}_\lambda$, the closed-form KL $K_t = \mathrm{KL}(P_t^\star \| Q_{\hat{\theta},t})$ contains the term $\frac{1}{2}\sum_i \phi(a_{t,i})$ with $\phi(x) = x - \log x - 1$ and $a_{t,i}$ the eigenvalues of $\hat{\Sigma}_{\hat{\theta},t}^{-1/2}\Sigma_t^\star\hat{\Sigma}_{\hat{\theta},t}^{-1/2}$. On the compact eigenvalue interval $[\lambda_-/\lambda_+, \lambda_+/\lambda_-]$, $(\log x)^2 \le C_\lambda \phi(x)$, so by Cauchy-Schwarz $|\log\det\hat{\Sigma}_{\hat{\theta},t} - \log\det\Sigma_t^\star| = |\sum_i -\log a_{t,i}| \le \sqrt{N\sum_i(\log a_{t,i})^2} \le \sqrt{2NC_\lambda K_t}$. Taking expectation and Jensen, $\frac{1}{2N}\mathbb{E}|\log\det\hat{\Sigma}_{\hat{\theta},t} - \log\det\Sigma_t^\star| \le C_{\det}\sqrt{\varepsilon_{\mathrm{NLL}}}$.

Step 4: NLL excess controls the population radius. Pinsker gives $\mathrm{TV}(P_t^\star, Q_{\hat{\theta},t}) \le \sqrt{K_t/2}$, so $\sup_u |G_{\hat{\theta}}(u) - G_{\chi_N}(u)| \le \sqrt{\varepsilon_{\mathrm{NLL}}/2} =: \tau_n$ in expectation, where G_{χ_N} is the χ_N CDF and $G_{\hat{\theta}}$ the population score CDF under $P_t^\star$. Local density $g_\star \ge \underline{g}_0$ at $s_\star = G_{\chi_N}^{-1}(1-\alpha)$ yields $|s_{\hat{\theta}} - s_\star| \le 2\tau_n/\underline{g}_0 \le (\sqrt{2}/\underline{g}_0)\sqrt{\varepsilon_{\mathrm{NLL}}}$ and $|\log s_{\hat{\theta}} - \log s_\star| \le (2\sqrt{2}/(s_\star\underline{g}_0))\sqrt{\varepsilon_{\mathrm{NLL}}}$.

Step 5: empirical calibration radius adds $r_m(\delta)$. On the quantile-concentration event E_δ of Theorem 7 (Part I), $|\hat{s}_m - s_{\hat{\theta}}| \le r_m(\delta)$ with $r_m(\delta)$ the Chebyshev/Bernstein radius (9)–(10), hence $\log\hat{s}_m \le \log s_\star + (2\sqrt{2}/(s_\star\underline{g}_0))\sqrt{\varepsilon_{\mathrm{NLL}}} + (2/s_\star)r_m(\delta)$.

Step 6: combine. Using (4) and Steps 3–5, $\widehat{\mathcal{V}}_m(\hat{\theta}) \le \mathcal{V}(\theta_\star) + C_1\sqrt{\varepsilon_{\mathrm{NLL}}} + C_2\, r_m(\delta)$ with $C_1 = C_{\det} + 2\sqrt{2}/(s_\star\underline{g}_0)$ and $C_2 = 2/s_\star$. Step 2 identifies $\mathcal{V}(\theta_\star) = \inf_{\theta\in\mathcal{F}_{\mathrm{cond}}}\mathcal{V}(\theta)$, giving (12).

B Full proofs and derivations

This section collects the auxiliary statements used by §5 and the linear-Kalman specialisation.

B.1 Why the Mahalanobis score

The score choice in (1) is non-trivial: changing the score changes the geometry of the emitted region. We collect here the three properties that single out the squared Mahalanobis score among alternatives on correlated Y — briefly stated in §3 and elaborated here for completeness.

Affine invariance. Let $\psi : Y \mapsto AY + b$ be any invertible affine reparameterisation, and write $(\hat{Y}^\psi, \hat{\Sigma}^\psi) = (A\hat{Y} + b, A\hat{\Sigma}A^\top)$ for the transformed predictive moments. Then $(\psi(Y) - \hat{Y}^\psi)^\top(\hat{\Sigma}^\psi)^{-1}(\psi(Y) - \hat{Y}^\psi) = (Y - \hat{Y})^\top\hat{\Sigma}^{-1}(Y - \hat{Y})$, so the score is ψ-invariant and the emitted region transforms as $\psi(\hat{C}_\alpha)$. Competing scores $\|r\|_\infty$, $\|r\|_2$, and the per-coordinate studentised $\max_i |r_i|/\sigma_i$ are *not* affine-invariant; these scores are not invariant under general affine reparameterisations.

Minimum-volume property at Gaussian targets. Among all Borel sets $C \subseteq \mathbb{R}^N$ of Gaussian-measure $1 - \alpha$ under $\mathcal{N}(\hat{Y}, \hat{\Sigma})$, the Lebesgue-minimum-volume set is the Mahalanobis ellipsoid $\{y : (y - \hat{Y})^\top\hat{\Sigma}^{-1}(y - \hat{Y}) \leq \chi^2_{N,1-\alpha}\}$ [1, Ch. 1]. Conformalising the squared Mahalanobis score replaces the χ^2 threshold by the empirical quantile $\hat{q}_{1-\alpha}$, so the emitted region is always a Mahalanobis ellipsoid. Exact finite-sample validity requires exchangeable scores (Theorem 1); under dependence we use Theorem 7. Minimum-volume optimality applies to the true Gaussian density level set, not to an arbitrary estimated covariance or to the projection-width functional.

Reduction to a scalar conformal problem. Split-conformal operates on a scalar score. The squared Mahalanobis score retains the covariance-oriented ellipsoid geometry while reducing calibration to one order statistic. Coordinatewise scores with a joint threshold instead produce an axis-aligned box, and the Euclidean norm produces a ball. None of these geometric comparisons alone establishes dominance in w_{proj}.

Choice of squared vs. square-root form. The squared score R_t and its root $S_t = \sqrt{R_t}$ produce the same prediction region: split-conformal is invariant under monotone transforms (the order statistic commutes with monotone maps), and $\hat{q}_{1-\alpha} = \hat{s}^2_{m,t_0}$. The implementation uses the squared form because $h \mapsto R(h, \xi)$ is smooth wherever (μ, Σ) are, avoiding the $1/\sqrt{R}$ blow-up at $R = 0$; the analysis uses the root form because R is not globally Lipschitz in h on an unbounded state space (its gradient grows linearly with the residual), whereas S admits the finite-constant root-score stability (C2′) of Lemma 17. For standard Gaussian innovations in Proposition 8, the squared scores are χ^2_N-distributed (root scores are χ_N); the two forms calibrate the same ellipsoid.

B.2 Proof of Theorem 1 (split-conformal validity)

Attach independent continuous auxiliary variables to break ties in the pooled scores. Exchangeability makes the resulting rank of R_{n+1} uniform on $\{1, \ldots, n+1\}$. Let $k = \lceil (n+1)(1-\alpha) \rceil$ and let $\hat{q}_{1-\alpha}$ be the k-th calibration order statistic, with $R_{(n+1)} = +\infty$. The event that the randomised pooled rank is at most k is contained in $\{R_{n+1} \leq \hat{q}_{1-\alpha}\}$, so coverage is at least $k/(n+1) \geq 1 - \alpha$. If scores are almost surely distinct, the two events coincide and coverage equals $k/(n+1) \leq 1 - \alpha + 1/(n+1)$. □

B.3 Warm-start dependent-quantile concentration

We isolate the warm-start dependent-quantile theorem used as a proof ingredient in Theorem 7. This auxiliary result is stated on a generic metric state space. In the learned-filter application, Theorem 3 supplies same-input contraction on $\mathcal{Q}_{\widehat{\theta}}$, Assumption 4 lifts this to coupling of the augmented data–filter process, and Corollary 5 supplies the threshold-autocovariance envelope of Assumption 13; Theorem 3 alone does *not* imply Assumption 13. Throughout the abstract hypotheses below, $\|h - h'\|$ denotes the quotient metric $d_{\widehat{\theta}}(h, h')$ in the learned-filter application.

Assumption 11 (Standing hypotheses for root scores)**.** *(C1) Mean contraction.* $\mathbb{E}\|\Psi(h, \xi) - \Psi(h', \xi)\| \leq \rho\|h - h'\|$ for some $\rho \in (0, 1)$. *(C2′) Root-score stability.* $\mathbb{E}_\xi|S(h, \xi) - S(h', \xi)| \leq L_S\|h - h'\|$ (Lemma 17). *(C3) First moments.* $D_0 := \mathbb{E}\|H_0 - H_0^\pi\| < \infty$ for an independent $H_0^\pi \sim \pi$, and $D_\pi := \mathbb{E}_{\pi\otimes\pi}\|H - \widetilde{H}\| < \infty$. *(C4) Local density.* $0 < \underline{g} \leq g_\star(u) \leq \bar{g} < \infty$ on $I = [s_\star - \varepsilon_0, s_\star + \varepsilon_0]$.

Theorem 12 (Local CDF forgetting, root score)**.** *Under Assumption 11, for every deterministic $\eta \in (0, \varepsilon_0]$, every initial state h, every $t \geq 1$, and every $u \in I_\eta = [s_\star - (\varepsilon_0 - \eta), s_\star + (\varepsilon_0 - \eta)]$,*

$$|F_t(u; h) - G_\star(u)| \leq \frac{L_S\rho^{t-1}d(h)}{\eta} + 2\bar{g}\eta, \quad d(h) := \mathbb{E}_{H_0^\pi\sim\pi}\|h - H_0^\pi\|. \tag{19}$$

For random H_0, optimising in η after expectation gives

$$b_t^{\mathrm{loc}} = \sup_{u \in I_{\eta_\star}} |\mathbb{P}(S_t \le u) - G_\star(u)| \le K_{\mathrm{loc}} \rho^{(t-1)/2} \tag{20}$$

with $K_{\mathrm{loc}} = 2\sqrt{2\bar{g} L_S D_0}$ and $\eta_\star = \sqrt{L_S D_0 \rho^{t-1}/(2\bar{g})}$. This optimised bound applies when $\eta_\star \le \varepsilon_0$. For calibration starting at t_0, choose $\eta_1 := \sqrt{L_S D_0 \rho^{t_0-1}/(2\bar{g})} < \varepsilon_0$ and set $I_{\mathrm{buf}} = I_{\eta_1}$; later optimal bandwidths are no larger, so the same buffer works throughout calibration.

Assumption 13 (Local threshold-autocovariance envelope)**.** There is a summable sequence $(\Gamma_k)_{k \ge 1}$ with $\sup_{u \in I_{\mathrm{buf}}} \sup_i |\mathrm{Cov}(I_i(u), I_{i+k}(u))| \le \Gamma_k$ where $I_j(u) = \mathbf{1}\{S_{t_0+j-1} \le u\}$.

Theorem 14 (Warm-start dependent-quantile concentration)**.** *Let $p = 1-\alpha$, $k = \lceil (m+1)p \rceil \le m$, $\hat{s}_{m,t_0} = S_{(k)}$, and $\eta_m = \max\{|k/m - p|, |(k-1)/m - p|\}$. Under Assumptions 11 and 13, for every $0 < \varepsilon \le \varepsilon_0 - \eta_1$ with $\underline{g}\varepsilon > \bar{b}_{m,t_0} + \eta_m$,*

$$\mathbb{P}(|\hat{s}_{m,t_0} - s_\star| > \varepsilon) \le \frac{2}{m(\underline{g}\varepsilon - \bar{b}_{m,t_0} - \eta_m)^2} \Big[\tfrac{1}{4} + 2 \sum_{k=1}^{m-1} (1 - k/m) \Gamma_k \Big] =: \delta_{m,t_0}(\varepsilon),$$

with $\bar{b}_{m,t_0} \le K_{\mathrm{loc}} \rho^{(t_0-1)/2}/(m(1-\sqrt{\rho}))$.

Proof. Let $I_j(u) := \mathbf{1}\{S_{t_0+j-1} \le u\}$ and $\hat{F}_{m,t_0}(u) = m^{-1} \sum_{j=1}^m I_j(u)$. We establish, for $u \in I_{\mathrm{buf}}$, (a) a bias bound $\bar{b}_{m,t_0} \le K_{\mathrm{loc}} \rho^{(t_0-1)/2}/(m(1-\sqrt{\rho}))$, (b) a variance bound $\mathrm{Var}(\hat{F}_{m,t_0}(u)) \le \big(\tfrac{1}{4} + 2\sum_{k \ge 1} \Gamma_k\big)/m$, and (c) combine via Chebyshev to conclude.

(a) Bias. For each j, $\mathbb{E} I_j(u) = \mathbb{P}(S_{t_0+j-1} \le u) = F_{t_0+j-1}(u)$ (the marginal CDF at time t_0+j-1 from H_0). By Theorem 12 on the deterministic buffer I_{η_1}, $|F_{t_0+j-1}(u) - G_\star(u)| \le K_{\mathrm{loc}} \rho^{(t_0+j-2)/2}$ for $u \in I_{\mathrm{buf}}$. Summing the geometric tail,

$$\sup_{u \in I_{\mathrm{buf}}} \big|\mathbb{E}\hat{F}_{m,t_0}(u) - G_\star(u)\big| \;\le\; \frac{K_{\mathrm{loc}}}{m} \sum_{j=1}^m \rho^{(t_0+j-2)/2} \;\le\; \frac{K_{\mathrm{loc}}\, \rho^{(t_0-1)/2}}{m(1-\sqrt{\rho})} \;=\; \bar{b}_{m,t_0}.$$

(b) Variance. Decompose $\mathrm{Var}(\hat{F}_{m,t_0}(u)) = m^{-2}\big[\sum_i \mathrm{Var}(I_i(u)) + 2\sum_{i<j} \mathrm{Cov}(I_i(u), I_j(u))\big]$. The diagonal sum is at most $m/4$ since $I_i(u) \in \{0,1\}$. For the off-diagonal terms, Assumption 13 gives $|\mathrm{Cov}(I_i(u), I_{i+k}(u))| \le \Gamma_k$ uniformly in i for $u \in I_{\mathrm{buf}}$, so

$$\sum_{1 \le i < j \le m} |\mathrm{Cov}(I_i(u), I_j(u))| \;\le\; m \sum_{k=1}^{m-1} (1-k/m)\Gamma_k \;\le\; m \sum_{k \ge 1} \Gamma_k.$$

Dividing by m^2,

$$\mathrm{Var}(\hat{F}_{m,t_0}(u)) \;\le\; \frac{1}{m}\Big[\tfrac{1}{4} + 2\sum_{k=1}^{m-1}(1-k/m)\Gamma_k\Big], \qquad u \in I_{\mathrm{buf}}. \tag{21}$$

The Markov-coupling derivation of an explicit $\Gamma_k \le C_\Gamma \kappa_J^{(k-1)/2}$ envelope from filter contraction is App. B.6; the audit of App. E.15 (Audit 3) verifies the empirical envelope.

(c) Chebyshev at the perturbed thresholds. Fix an admissible ε. At $u = s_\star + \varepsilon$, $\mathbb{E}\hat{F}_{m,t_0}(u) \ge p + \underline{g}\varepsilon - \bar{b}_{m,t_0}$. If $\hat{s}_{m,t_0} > u$, then $\hat{F}_{m,t_0}(u) \le (k-1)/m \le p + \eta_m$. At $v = s_\star - \varepsilon$, $\mathbb{E}\hat{F}_{m,t_0}(v) \le p - \underline{g}\varepsilon + \bar{b}_{m,t_0}$, whereas $\hat{s}_{m,t_0} < v$ implies $\hat{F}_{m,t_0}(v) \ge k/m \ge p - \eta_m$. Thus either tail requires an empirical-CDF deviation of at least $\underline{g}\varepsilon - \bar{b}_{m,t_0} - \eta_m > 0$. Applying Chebyshev with (21) and summing the two tails proves the bound. When $k = m+1$, the conformal radius is infinite; the finite local-quantile bound is restricted to $k \le m$. □

B.4 Approximate marginal validity

The combined warm-start + local-CDF-forgetting argument gives an auxiliary approximate-validity statement, used as a sub-step of Theorem 7.

Proposition 15 (Approximate marginal validity)**.** *Under Theorems 12 and 14, the prediction set $\hat{C}_t$ satisfies, for every admissible ε,*

$$|\mathbb{P}\{Y_t \in \hat{C}_t\} - (1-\alpha)| \le \bar{g}\varepsilon + b_t^{\mathrm{loc}} + \delta_{m,t_0}(\varepsilon),$$

and balancing $\bar{g}\varepsilon \asymp C/(m\varepsilon^2)$ at $\varepsilon \asymp m^{-1/3}$ gives the rate $O(m^{-1/3} + \bar{b}_{m,t_0} + b_t^{\mathrm{loc}})$.

Proof. Let $E_\varepsilon := \{|\hat{s}_{m,t_0} - s_\star| \le \varepsilon\}$. On E_ε, for every t,

$$\{S_t \le s_\star - \varepsilon\} \subseteq \{S_t \le \hat{s}_{m,t_0}\} \subseteq \{S_t \le s_\star + \varepsilon\}.$$

Using $s_\star \pm \varepsilon \in I_{\mathrm{buf}}$ and the local test-time gap $|\mathbb{P}(S_t \le u) - G_\star(u)| \le b_t^{\mathrm{loc}}$ for $u \in I_{\mathrm{buf}}$,

$$G_\star(s_\star - \varepsilon) - b_t^{\mathrm{loc}} - \delta_{m,t_0}(\varepsilon) \;\le\; \mathbb{P}(S_t \le \hat{s}_{m,t_0}) \;\le\; G_\star(s_\star + \varepsilon) + b_t^{\mathrm{loc}} + \delta_{m,t_0}(\varepsilon).$$

By (C4), $G_\star(s_\star \pm \varepsilon) = 1 - \alpha \pm O(\bar{g}\,\varepsilon)$, so $|\mathbb{P}(S_t \le \hat{s}_{m,t_0}) - (1-\alpha)| \le \bar{g}\varepsilon + b_t^{\mathrm{loc}} + \delta_{m,t_0}(\varepsilon)$; because $S_t(y) \le \hat{s}_{m,t_0} \Leftrightarrow R_t(y) \le \hat{s}^2_{m,t_0} = \hat{q}_{1-\alpha}$, this is also the coverage gap of $\hat{C}_t$. *Order-statistic offset.* The finite-rank correction $\eta_m = O(m^{-1})$ is already included in $\delta_{m,t_0}(\varepsilon)$, without identifying the conformal rank with a different empirical quantile. Taking $\varepsilon = (\bar{b}_{m,t_0} + \eta_m)/\underline{g} + c\,m^{-1/3}$ for fixed $c > 0$ gives the stated rate (for fixed summable covariance envelope), with an additional $O(m^{-1})$ term absorbed into it. □

B.5 Oracle stationary corollary of Theorem 14

Stating Theorem 14 in the stationary-initialisation limit ($H_0 \sim \pi$, so $\bar{b}_{m,t_0} = 0$ for every $t_0 \ge 1$) recovers the oracle form for $\underline{g}\varepsilon > \eta_m$ and thresholds in the fixed local buffer:

$$\mathbb{P}(|\hat{s}_m - s_\star| > \varepsilon) \;\le\; \frac{2}{m(\underline{g}\varepsilon - \eta_m)^2}\Big[\tfrac{1}{4} + 2\sum_{k\ge 1}\Gamma_k\Big] \;=:\; \delta_m(\varepsilon), \tag{22}$$

which is the rate referenced in Proposition 15's simplification under stationary initialisation.

B.6 Sufficient condition for Assumption 13: Jacobian contraction

The warm-start theorem assumes the threshold-autocovariance envelope Γ_k directly. We give an explicit sufficient condition under which Γ_k inherits a geometric envelope from synchronous Jacobian-product contraction of the frozen filter; this is the formal bridge between the contraction hypothesis (C1) and the local-mixing hypothesis (Assumption 13).

Setting. Let $(H_t) \subset \mathbb{R}^d$ be the hidden state of the frozen filter, $H_t = \Psi(H_{t-1}, \xi_t)$ with i.i.d. driving noise (ξ_t), and let $S_t = S(H_{t-1}, \xi_t)$ be the root Mahalanobis score. Write H_ℓ^h for the chain started from h and driven by the same future innovations as a stationary copy H_ℓ^π with $H_0^\pi \sim \pi$. Let $G_\star$ denote the stationary CDF of S and $I = [s_\star - \varepsilon_0, s_\star + \varepsilon_0]$ the local interval.

Theorem 16 (Jacobian contraction implies summable Γ_k)**.** *Suppose, in addition to (C2′)–(C4):* (J1) Synchronous contraction. *There exist $C_J < \infty$, $\kappa_J \in (0,1)$ such that $\mathbb{E}\|H_\ell^h - H_\ell^{h'}\| \le C_J\kappa_J^\ell\|h - h'\|$ for all h, h' and $\ell \ge 0$.* (J2) Finite first moments. *$D_0, D_\pi < \infty$ as in (C3). Fix $\eta \in (0, \varepsilon_0)$ and the buffered interval $I_\eta = [s_\star - (\varepsilon_0 - \eta), s_\star + (\varepsilon_0 - \eta)]$. Then for every $u \in I_\eta$, $i \ge 1$, and $k \ge 1$,*

$$|\mathrm{Cov}(\mathbf{1}\{S_i \le u\}, \mathbf{1}\{S_{i+k} \le u\})| \le \Gamma_k := \left[\frac{L_S C_J(D_\pi + C_J D_0)}{\eta} + 2\bar{g}\eta\right]\kappa_J^{(k-1)/2}. \tag{23}$$

In particular, $\sum_{k\ge 1}\Gamma_k < \infty$ on this fixed buffer.

Proof. Set $d(h) = \mathbb{E}_{H\sim\pi}\|h - H\|$ and $\eta_k = \eta\kappa_J^{(k-1)/2} \le \eta$. Synchronous coupling and (C2′) give $\mathbb{E}|S_k^h - S_k^\pi| \le L_S C_J \kappa_J^{k-1} d(h)$. Since $[u - \eta_k, u + \eta_k] \subset I$ for $u \in I_\eta$, Markov's inequality and the stationary density upper bound yield

$$|\mathbb{P}(S_k^h \le u) - G_\star(u)| \le \frac{L_S C_J \kappa_J^{k-1} d(h)}{\eta_k} + 2\bar{g}\eta_k.$$

Let $I_i = \mathbf{1}\{S_i \le u\}$ and let $\mathcal{F}_i$ contain the innovations and state through time i. The independent driving innovations give the Markov identity $\mathbb{P}(S_{i+k} \le u \mid \mathcal{F}_i) = \mathbb{P}(S_k^{H_i} \le u \mid H_i)$. Subtracting the deterministic constant $G_\star(u)$ inside the covariance and using $|I_i - \mathbb{E}I_i| \le 1$ therefore gives

$$|\mathrm{Cov}(I_i, I_{i+k})| \le \frac{L_S C_J \kappa_J^{k-1} \mathbb{E}d(H_i)}{\eta_k} + 2\bar{g}\eta_k.$$

Couple H_i to a stationary copy driven by the same innovations, and use an independent stationary draw in the definition of d. The triangle inequality gives $\mathbb{E}d(H_i) \le C_J \kappa_J^i D_0 + D_\pi \le C_J D_0 + D_\pi$. Substituting η_k proves (23); summability follows from $\sqrt{\kappa_J} < 1$. □

High-probability variant. If only a high-probability product-Jacobian bound is available—i.e., $\mathbb{P}[\sup_z \|D_h(\Psi_{\xi_\ell} \circ \cdots \circ \Psi_{\xi_1})(z)\|_{\mathrm{op}} > C_J \kappa_J^\ell] \le \delta_\ell$ on a tube of diameter B—the same argument adds an $O(\sqrt{\delta_{k-1}})$ term to Γ_k; the envelope remains summable whenever $\sum_k \sqrt{\delta_k} < \infty$ (e.g. geometric $\delta_k \le C_\delta \kappa_\delta^k$). Empirically, the audit of App. E.15 (Audit 3) confirms a finite $\sum_{k \le 30} |\hat{\Gamma}_k|$ and an exponential per-lag decay rate in $[0.83, 0.90]$ on both GRAPH-LGSSM and GNF, i.e., the operative form of Assumption 13 holds at the audited cells.

B.7 Sufficient conditions for the root-score stability (C2′)

The standing assumption (C2′) is that the root score $h \mapsto S(h, \xi) = \|G(h) r(h, \xi)\|_2$, $G(h) = \hat{\Sigma}(h)^{-1/2}$, is Lipschitz in mean. The squared score $R = S^2$ is not globally Lipschitz on an unbounded state space because its gradient grows linearly with the residual; the root-score formulation removes this blow-up while preserving the calibrated region (§3). The lemma below gives a standard sufficient condition.

Lemma 17 (Root-score mean stability)**.** *Suppose, for all coupled states h, h' visited by the filter, $\|\mu(h, \xi) - \mu(h', \xi)\| \le L_\mu \|h - h'\|$ (mean Lipschitz), $\|G(h) - G(h')\|_{\mathrm{op}} \le L_G \|h - h'\|$ (inverse-square-root covariance Lipschitz), $\|G(h)\|_{\mathrm{op}} \le M_G$, and $\mathbb{E}_\xi \|r(h, \xi)\| \le \bar{s}$. Then*

$$\mathbb{E}_\xi \big| S(h, \xi) - S(h', \xi) \big| \;\le\; (L_G \bar{s} + M_G L_\mu)\, \|h - h'\|, \tag{24}$$

i.e. (C2′) *holds with $L_S = L_G \bar{s} + M_G L_\mu$.*

Proof. The reverse triangle inequality on $G(h) r(h, \xi)$ gives $|S(h, \xi) - S(h', \xi)| = \big| \|G(h) r(h, \xi)\| - \|G(h') r(h', \xi)\| \big|$, which is bounded by $\|(G(h) - G(h')) r(h, \xi)\| + \|G(h')(\mu(h', \xi) - \mu(h, \xi))\|$. The first term is $\le L_G \|r(h, \xi)\| \, \|h - h'\|$; the second is $\le M_G L_\mu \|h - h'\|$. Take expectation in ξ. □

Implications. For the linear-Gaussian GRAPH-LGSSM the predictive covariance is state-independent, so $L_G = 0$ and $L_S \le M_G L_\mu = \|\hat{\Sigma}^{-1/2} C\|_{\mathrm{op}}$; this recovers the linear-Kalman case on an unbounded Gaussian state. For GNF the variance floor $\varepsilon_d = 10^{-4}$ clamps the diagonal head from below, so $M_G \le 100$; the log-diagonal map is smooth, the GRU is gate-Lipschitz, and the low-rank head is linear in h_t, yielding finite L_G and L_μ along the calibration trajectory.

Empirical audit. We estimate L_S by random directional perturbation ($n_{\mathrm{pert}} = 12$ samples per step at three scales) on the calibration trajectory, recording $|S(h + \delta, \xi) - S(h, \xi)| / \|\delta\|$ percentiles (Audit 1, App. E.15). Median quotients are modest ($p50 \le 0.36$, $p95 \le 1.19$); the neural filter on METR-LA has a heavier $p99$ tail at high-condition-number trajectory points. The full-covariance audit in App. E.23 shows that the default floor is nonbinding, so we do not attribute this tail to floor activation.

B.8 Proof of Theorem 12 (local forgetting, root score)

Couple the actual chain started from $H_0 = h$ with a stationary copy H_t^π initialised from $H_0^\pi \sim \pi$, using the same innovation noise $\xi_{1:t}$. By (C1), $\mathbb{E}[\|H_{t-1} - H_{t-1}^\pi\| \mid h, H_0^\pi] \le \rho^{t-1} \|h - H_0^\pi\|$; the root score at time t is determined by H_{t-1} together with ξ_t, so by (C2′),

$$\mathbb{E}\big[|S_t - S_t^\pi| \mid h \big] \;\le\; L_S\, \rho^{t-1}\, \mathbb{E}\big\| h - H_0^\pi \big\| \;=\; L_S\, \rho^{t-1}\, d(h). \tag{25}$$

For any $u \in \mathbb{R}$ and *deterministic* $\eta > 0$,

$$\big| F_t(u; h) - G_\star(u) \big| \;\le\; \mathbb{P}(|S_t - S_t^\pi| > \eta) + \mathbb{P}(|S_t^\pi - u| \le \eta).$$

Markov bounds the first term by $\mathbb{E}|S_t - S_t^\pi|/\eta$; for $u \in I_\eta = [s_\star - (\varepsilon_0 - \eta), s_\star + (\varepsilon_0 - \eta)]$, the event $\{|S_t^\pi - u| \le \eta\}$ has probability at most $2\bar{g}\eta$ by (C4) applied to the stationary root-score density on the η-neighbourhood of u (which lies inside I). Combining yields (19).

For random H_0 take expectation over h:

$$b_t^{\mathrm{loc}}(\eta) \;\le\; \frac{L_S D_0 \rho^{t-1}}{\eta} + 2\bar{g}\eta, \qquad u \in I_\eta.$$

The map $\eta \mapsto a/\eta + 2\bar{g}\eta$ is minimised at $\eta_\star = \sqrt{a/(2\bar{g})}$ with value $2\sqrt{2\bar{g}a}$; substituting $a = L_S D_0 \rho^{t-1}$ gives $b_t^{\mathrm{loc}} \le 2\sqrt{2\bar{g}L_S D_0}\rho^{(t-1)/2} = K_{\mathrm{loc}}\rho^{(t-1)/2}$, which is (20) with D_0 inside the square root. The deterministic bandwidth $\eta_\star$ keeps the buffered interval $I_{\eta_\star}$ non-random. □

Remark (state-dependent vs marginal). Equation (19) retains the state-dependent $d(h)$ on I_η; one obtains a pairwise (two cold-started chains) statement of the same form by replacing $d(h)$ with $\|h - h'\|$ and $G_\star$ with the cold-started counterpart. We use only the marginal form (20).

B.9 Linear-Kalman specialisation: initialisation-weighted RMS identity

The linear-Kalman GRAPH-LGSSM admits an exact RMS identity that serves as the linear specialisation of Theorem 3 (referenced as the linear-Kalman sanity check in §5). Define the eigenbasis weights $w_i = (U^* M_0 U)_{ii}/\operatorname{tr} M_0$ and the initialisation-weighted time-local RMS rate $\rho_{M_0}(A_{\mathrm{cl}}, t) := (\sum_i w_i |\lambda_i|^{2t})^{1/(2t)}$.

Assumption 18. *(L1) Closed-loop normality.* A_{cl} is normal, $A_{\mathrm{cl}} = U\Lambda U^*$. *(L2) Sub-unit spectrum.* $\sigma_1(A_{\mathrm{cl}}) < 1$. *(L3) Initialisation second moment.* $M_0 = \mathbb{E}[\Delta_0 \Delta_0^*]$ has finite trace.

Theorem 19 (Initialisation-weighted RMS identity)**.** *Under Assumption 18, for every* $t \ge 1$*,*

$$\big(\mathbb{E}\|\Delta_t\|^2\big)^{1/2} = \rho_{M_0}(A_{\mathrm{cl}}, t)^t \big(\mathbb{E}\|\Delta_0\|^2\big)^{1/2}, \tag{26}$$

and for every $1 \le t \le T$*,*

$$\big(\mathbb{E}\|\Delta_t\|^2\big)^{1/2} \le \rho_{M_0}(A_{\mathrm{cl}}, T)^t \big(\mathbb{E}\|\Delta_0\|^2\big)^{1/2} \qquad \textit{with equality at } t = T. \tag{27}$$

The isotropic trace rate ρ_{tr} *is the special case* $M_0 \propto I$*.*

Proof.

Step 1: exact time-local identity. Since A_{cl} is normal, $A_{\mathrm{cl}} = U\Lambda U^*$ with $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_d)$ and U unitary. Then $A_{\mathrm{cl}}^t = U\Lambda^t U^*$ and

$$\mathbb{E}\|\Delta_t\|^2 \;=\; \operatorname{tr}\big(A_{\mathrm{cl}}^t M_0 (A_{\mathrm{cl}}^*)^t\big) \;=\; \operatorname{tr}\big(\Lambda^t (U^* M_0 U) \Lambda^{*t}\big) \;=\; \sum_{i=1}^d |\lambda_i|^{2t} \, (U^* M_0 U)_{ii}.$$

Dividing by $\mathbb{E}\|\Delta_0\|^2 = \operatorname{tr} M_0$ gives the unit-weight form $\mathbb{E}\|\Delta_t\|^2/\mathbb{E}\|\Delta_0\|^2 = \sum_i w_i |\lambda_i|^{2t} = \rho_{M_0}(A_{\mathrm{cl}}, t)^{2t}$, which rearranges to (26).

Step 2: monotonicity in t. For any non-negative $\{a_i \ge 0\}$ with weights $\{w_i \ge 0\}$, $\sum_i w_i = 1$, the weighted power mean $t \mapsto (\sum_i w_i a_i^{2t})^{1/(2t)}$ is non-decreasing in t [12, Theorem 16]. Apply with $a_i = |\lambda_i|$ and $w_i = (U^* M_0 U)_{ii}/\operatorname{tr} M_0$.

Step 3: $\rho_{M_0} \le \sigma_1$ strictly. $\rho_{M_0}(A_{\mathrm{cl}}, t)^{2t} = \sum_i w_i |\lambda_i|^{2t} \le \sigma_1^{2t} \sum_i w_i = \sigma_1^{2t}$, with equality iff w_i is supported only on indices where $|\lambda_i| = \sigma_1$. Generic graph-polynomial spectra have a non-degenerate sub-leading tail and a non-aligned M_0, so the inequality is strict at every finite t.

Step 4: horizon-T envelope. Combining Steps 1 and 2, for every $1 \le t \le T$, $(\mathbb{E}\|\Delta_t\|^2)^{1/2} = \rho_{M_0}(A_{\mathrm{cl}}, t)^t (\mathbb{E}\|\Delta_0\|^2)^{1/2} \le \rho_{M_0}(A_{\mathrm{cl}}, T)^t (\mathbb{E}\|\Delta_0\|^2)^{1/2}$, with equality at $t = T$. This is (27).

Step 5: isotropic corollary. When $M_0 = \tau^2 I_d$, $w_i = 1/d$ and $\rho_{M_0}(A_{\mathrm{cl}}, t) = (d^{-1} \sum_i |\lambda_i|^{2t})^{1/(2t)} = \rho_{\mathrm{tr}}(A_{\mathrm{cl}}, t)$. Thus the textbook trace rate is the isotropic special case.

Step 6: transfer to the score process. The bounded-Lipschitz step in Theorem 12's proof takes any upper bound on $\mathbb{E}|S_t - S_t^\pi|$ and produces a local-CDF rate. Replacing the mean-coupling step of §B.8 by Jensen $\mathbb{E}|S_t - S_t^\pi| \le L_S \mathbb{E}\|\Delta_{t-1}\| \le L_S (\mathbb{E}\|\Delta_{t-1}\|^2)^{1/2}$ and substituting (27) at index $t-1$ yields, for $u \in I_\eta$ with the same deterministic bandwidth as Theorem 12, $\sup_{u \in I_\eta} |\mathbb{P}(S_t \le u) - G_\star(u)| \le K_{\mathrm{tr}} \rho_{M_0}(A_{\mathrm{cl}}, T)^{(t-1)/2}$ for $1 \le t \le T+1$, with $K_{\mathrm{tr}} = 2\sqrt{2\bar{g}L_S \bar{D}}$ and $\bar{D} = (\mathbb{E}\|\Delta_0\|^2)^{1/2}$. □

Interpreting the σ_1 / ρ_{M_0} / $\widehat{\rho}_{d_{\mathcal{G}}}$ / $\widehat{\rho}_{\mathrm{score}}$ relation. On the audited cells of App. E.15 the gap between the weighted rate and its isotropic specialisation is below 1%: $\sigma_1(A_{\mathrm{cl}}) \approx 0.443$ at $\rho_{\mathrm{scale}} = 0.8$, $\rho_{\mathrm{tr}}(A_{\mathrm{cl}}, 50), \rho_{M_0}(A_{\mathrm{cl}}, 50) \in [0.43, 0.44]$ (the audited METR-LA cell has M_0-anisotropy $\approx 2{,}011\times$ but $|\rho_{M_0} - \rho_{\mathrm{tr}}| < 0.003$). The Bures-Wasserstein audit $\widehat{\rho}_{d_{\mathcal{G}}} \approx 0.46$ tracks $\sigma_1(A_{\mathrm{cl}})$ within 0.02 on the linear sanity check (App. E.5), confirming that the direct emitted-law contraction object of Theorem 3 matches the linear analytic rate. The *score-CDF* 1-Wasserstein rate $\widehat{\rho}_{\mathrm{score}} \approx 0.15$ on the same cells is faster because the conformal score collapses many filter-state degrees of freedom; this gap reflects the score-Lipschitz step of Theorem 12, not a discrepancy with Theorem 3. Proposition 21 shows that no *uniform-in-x_t* refinement of σ_1 closes the state-side gap without paying a prefactor exponential in T.

B.10 Local-linearisation transient bound for learned filters

For the learned GNF the closed-loop transition has no closed form. We give a finite-horizon perturbation bound: if the frozen recurrence is well approximated by its Jacobian product on the coupling tube, then the nonlinear RMS contraction follows the corresponding Jacobian-weighted RMS rate up to an explicit $O(\varepsilon T)$ remainder.

Setting. Fix a horizon T. Let $H_t = \Psi_t(H_{t-1})$ and $H_t^\pi = \Psi_t(H_{t-1}^\pi)$ be two copies of the frozen filter driven by the same realised input path; here Ψ_t denotes the time-t hidden-state update along that path. Define $\Delta_t := H_t - H_t^\pi$, the Jacobian $J_t := D\Psi_t(H_{t-1}^\pi)$, and the Jacobian product $J_{b:a} := J_b J_{b-1} \cdots J_a$ (with $J_{a-1:a} := I$). Set $M_0 = \mathbb{E}[\Delta_0 \Delta_0^\top]$ with $0 < d_0 := (\operatorname{tr} M_0)^{1/2} < \infty$. Expectations below integrate both the reference trajectory and the initial perturbation. To retain their dependence, let $\mathcal{J} = \sigma(J_1, \ldots, J_T)$ and define the conditional second moment $M_0(\mathcal{J}) := \mathbb{E}[\Delta_0 \Delta_0^\top \mid \mathcal{J}]$.

Theorem 20 (Finite-horizon local-linearisation RMS bound)**.** *Assume* (B1) Tube containment. *For all $1 \le t \le T$, the line segment $\{H_{t-1}^\pi + s\Delta_{t-1} : 0 \le s \le 1\}$ lies in a tube on which the next condition holds.* (B2) Local Jacobian variation. *For each $1 \le t \le T$ and every h in that tube, $\|D\Psi_t(h) - J_t\|_{\mathrm{op}} \le \varepsilon$.* (B3) Finite linearised envelope. $B_T := \operatorname{ess\,sup} \max_{0 \le a \le b \le T} \|J_{b:a+1}\|_{\mathrm{op}} < \infty$. *Define the linearised perturbation $Z_t := J_{t:1}\Delta_0$ and the Jacobian-weighted RMS rate*

$$\rho_{M_0}(J_{t:1}, t) \;:=\; \Big(\frac{\mathbb{E}\operatorname{tr}(J_{t:1} M_0(\mathcal{J}) J_{t:1}^\top)}{\operatorname{tr} M_0}\Big)^{1/(2t)}.$$

Then for every $1 \le t \le T$, $(\mathbb{E}\|Z_t\|^2)^{1/2} = \rho_{M_0}(J_{t:1}, t)^t d_0$, and

$$\big(\mathbb{E}\|\Delta_t\|^2\big)^{1/2} \;\le\; \big[\rho_{M_0}(J_{t:1}, t)^t + \varepsilon B_T^2 t\, e^{\varepsilon B_T t}\big]\, d_0. \tag{28}$$

In particular, if $\varepsilon B_T T \le 1$ then $(\mathbb{E}\|\Delta_t\|^2)^{1/2} \le [\rho_{M_0}(J_{t:1}, t)^t + e\,\varepsilon B_T^2 T]\, d_0$.

Proof. The exact identity for Z_t follows from the tower-property identity $\mathbb{E}\|Z_t\|^2 = \mathbb{E}\operatorname{tr}(J_{t:1} M_0(\mathcal{J}) J_{t:1}^\top)$. The conditional second moment is essential because $J_{t:1}$ and Δ_0 may be dependent. For a fixed reference trajectory and input path, the same identity holds conditionally, reducing to the usual deterministic-product trace formula. By the fundamental theorem of calculus, $\Delta_t = J_t \Delta_{t-1} + r_t$ with $r_t = \int_0^1 [D\Psi_t(H_{t-1}^\pi + s\Delta_{t-1}) - J_t]\Delta_{t-1}\, ds$, so $\|r_t\| \le \varepsilon \|\Delta_{t-1}\|$ by (B2). Variation of constants gives $\Delta_t = Z_t + \sum_{s=1}^t J_{t:s+1} r_s$. Let $R_t := (\mathbb{E}\|\Delta_t\|^2)^{1/2}$. Minkowski plus $(\mathbb{E}\|J_{t:s+1} r_s\|^2)^{1/2} \le B_T \varepsilon R_{s-1}$ yields $R_t \le B_T d_0 + \varepsilon B_T \sum_{s=0}^{t-1} R_s$; discrete Gronwall $R_t \le B_T d_0\, e^{\varepsilon B_T t}$. Substituting back into $\Delta_t - Z_t = \sum_s J_{t:s+1} r_s$ yields $(\mathbb{E}\|\Delta_t - Z_t\|^2)^{1/2} \le \varepsilon B_T^2 t e^{\varepsilon B_T t} d_0$. The triangle inequality completes (28). □

Consequence: CDF forgetting for the learned filter. Combining Theorem 20 with the bounded-Lipschitz step of Theorem 12, in the same way as Theorem 19, gives

$$\sup_{u \in I_\eta} \big|\mathbb{P}(S_t \le u) - G_\star(u)\big| \;\le\; 2\sqrt{2\bar{g} L_S\, d_0\, [\bar{\rho}_{M_0}(T)^{t-1} + \varepsilon B_T^2 T e^{\varepsilon B_T T}]},$$

where $\bar{\rho}_{M_0}(T) := \max_{1 \le t \le T} \rho_{M_0}(J_{t:1}, t)$. This consequence also uses (C2′)–(C4) and a deterministic bandwidth η with $u \pm \eta \in I$. Optimising the same Markov/density bound as in Theorem 12 gives the displayed square-root bound when the optimal bandwidth fits this local buffer. The expectation in the RMS rate retains dependence between the reference path and the initial perturbation; replacing it with an unconditional M_0 inside a random Jacobian product would require additional independence.

B.11 Impossibility of horizon-free worst-case rates

Proposition 21 (Lower bound). *Let $A \in \mathbb{R}^{d\times d}$ be normal with $\sigma_1(A) > 0$. If for some $c_1 \geq 0$, $\rho_\star > 0$, and $T \geq 1$, $\|A^t x\| \leq c_1 \rho_\star^t \|x\|$ holds for all x and all $1 \leq t \leq T$, then $c_1 \geq (\sigma_1(A)/\rho_\star)^T$.*

Proof. Since $A \in \mathbb{R}^{d\times d}$ is normal, $\|A^T\|_{\mathrm{op}} = \|A\|_{\mathrm{op}}^T = \sigma_1(A)^T$. Choose a real unit vector x_T attaining the operator norm of A^T; then $\|A^T x_T\| = \sigma_1(A)^T$. The assumed bound at $t = T$ gives $\sigma_1(A)^T \leq c_1 \rho_\star^T \|x_T\| = c_1 \rho_\star^T$, hence $c_1 \geq (\sigma_1(A)/\rho_\star)^T$. □

Substantively: picking $\rho_\star = 0.15$, $\sigma_1 = 0.44$, $T = 50$ forces $c_1 \geq (0.44/0.15)^{50} \approx 2.6 \times 10^{23}$ — uninformative as a uniform bound on the score-CDF gap. The mean-squared form of Theorem 19 avoids this because it averages over the initialisation second moment M_0 rather than tracking the worst-aligned direction.

B.12 Same-block plug-in conformal stability

Static-covariance baselines (CGIF-JOINT, FACTORCGIF, GROUPCGIF) fit $\hat{\Sigma}$ on the calibration block and then score the same block. This is not exactly split-conformal because the score map is not frozen before calibration. The theorem below bounds the resulting coverage perturbation by leave-one-out stability and a local margin condition; the explicit constant for ridge empirical covariance is in Corollary 23.

Setting. Let $Z_1, \ldots, Z_m, Z_{m+1}$ be exchangeable. Let $\theta \mapsto s(z;\theta)$ be a scalar nonconformity score and A a permutation-invariant parameter fit. The implemented procedure sets $\hat{\theta} = A(Z_1, \ldots, Z_m)$, computes $S_i = s(Z_i;\hat{\theta})$, and emits $\hat{q} = S_{(k)}$, $k := \lceil (m+1)(1-\alpha) \rceil$. For each $1 \leq i \leq m+1$ define the leave-one-out oracle $\hat{\theta}^{(-i)} := A(Z_{1:m+1} \setminus \{Z_i\})$ and $T_i := s(Z_i; \hat{\theta}^{(-i)})$; in particular $T_{m+1} = S_{m+1}$. Let $q^\circ := T_{(k)}^{1:m}$.

Theorem 22 (Stability of same-block plug-in conformal scores). *Assume* (P1) Replacement stability: *with probability at least $1-\delta_m$, $\max_{1\leq i\leq m} |S_i - T_i| \leq \varepsilon_m$.* (P2) Local test-score margin: $\mathbb{P}\{|S_{m+1} - \hat{q}| \leq r\} \leq \omega_m(r)$ *with $\omega_m(r) \to 0$ as $r \to 0$.* (P3) Oracle no-ties: $T_1, \ldots, T_{m+1}$ *are almost surely distinct (or use randomised tie-breaking). Then*

$$\left|\mathbb{P}\{S_{m+1} \leq \hat{q}\} - (1-\alpha)\right| \;\leq\; \frac{1}{m+1} + \delta_m + \omega_m(\varepsilon_m). \tag{29}$$

If $\omega_m(r) \leq 2\bar{f} r$ in a neighbourhood of $\hat{q}$ (e.g. under a conditional density bound), then (29) *reads $1/(m+1) + \delta_m + 2\bar{f}\varepsilon_m$.*

Proof. Permutation-invariance of A and exchangeability of $Z_{1:m+1}$ make $(T_1, \ldots, T_{m+1})$ exchangeable; (P3) gives $\mathbb{P}\{T_{m+1} \leq q^\circ\} = k/(m+1) \in [1-\alpha, 1-\alpha+1/(m+1)]$. On the stability event $\mathcal{E} := \{\max_i |S_i - T_i| \leq \varepsilon_m\}$, the sup-norm stability of order statistics gives $|\hat{q} - q^\circ| \leq \varepsilon_m$. Because $T_{m+1} = S_{m+1}$, the events $\{S_{m+1} \leq \hat{q}\}$ and $\{T_{m+1} \leq q^\circ\}$ differ only when $|S_{m+1} - \hat{q}| \leq \varepsilon_m$. Hence $|\mathbb{P}\{S_{m+1} \leq \hat{q}\} - \mathbb{P}\{T_{m+1} \leq q^\circ\}| \leq \delta_m + \omega_m(\varepsilon_m)$; combine with the rank uniformity above. □

Corollary 23 (Ridge empirical covariance plug-in). *Let $\hat{\Sigma}_D = m^{-1}\sum_{r\in D} rr^\top + \lambda I_N$ with $\lambda > 0$, and $s(r;\hat{\Sigma}_D) = r^\top \hat{\Sigma}_D^{-1} r$ on residuals with $\|r\|_2 \leq B$. For any one-replacement pair (D, D') of size m and any evaluation residual $\|r\| \leq B$, $|s(r;\hat{\Sigma}_D) - s(r;\hat{\Sigma}_{D'})| \leq 2B^4/(\lambda^2 m)$, so Theorem 22 applies with $\varepsilon_m = O(B^4/(\lambda^2 m))$ and the coverage perturbation is $O(m^{-1})$.*

Proof. $\hat{\Sigma}_D - \hat{\Sigma}_{D'} = m^{-1}(aa^\top - bb^\top)$ has operator norm $\leq 2B^2/m$. The ridge floor gives $\|\hat{\Sigma}_D^{-1}\|_{\mathrm{op}}, \|\hat{\Sigma}_{D'}^{-1}\|_{\mathrm{op}} \leq \lambda^{-1}$, so by the resolvent identity $\|\hat{\Sigma}_D^{-1} - \hat{\Sigma}_{D'}^{-1}\|_{\mathrm{op}} \leq 2B^2/(\lambda^2 m)$, and $|r^\top(\hat{\Sigma}_D^{-1} - \hat{\Sigma}_{D'}^{-1})r| \leq 2B^4/(\lambda^2 m)$. □

Empirical audit. Refitting $\hat{\Sigma}$ on the train block (App. E.15, Audit 6) shifts joint coverage by ≤ 2.3 pp and w_{proj} by ≤ 0.3 z-score units relative to the deployed cal-fit protocol on every audited cell, supporting the use of plug-in calibration in our benchmark. Exact split-conformal validity (Theorem 1) is recovered by fitting $\hat{\Sigma}$ on a disjoint pre-calibration block.

B.13 Asymptotic SNR-tightened form (isotropic Kalman) and the surrogate ρ_{snr}

The rate $\sigma_1(A_{\mathrm{cl}})$ that appears in Theorem 19 and throughout the paper is computed by iterating the discrete algebraic Riccati equation (DARE) to the steady-state covariance P_{ss}, then extracting σ_1 from $A_{\mathrm{cl}} = F(I - KC)$. For large N the DARE iteration is $\mathcal{O}(N^3)$ per step and dominates hyperparameter tuning cost. We derive a closed-form expression under isotropy and a differentiable $\mathcal{O}(N)$ surrogate ρ_{snr} that coincides with σ_1 at isotropy and agrees with it to first order otherwise.

Closed-form steady state under isotropy. Assume $P_{\mathrm{ss}}^- = p^- I_d$ (steady-state prior state variance), $R = \nu I_d$ (observation noise), $Q = qI_d$ (innovation noise), and $F = \rho_{\mathrm{scale}}\, U$ with U orthogonal. Writing p^+ for the steady-state posterior variance, the standard Riccati recursion in prior/posterior form is

$$p^- = f^2 p^+ + q, \qquad p^+ = \nu p^- / (p^- + \nu),$$

where $f := \rho_{\mathrm{scale}}$ collects the scalar coefficient. Substituting yields the quadratic $(p^-)^2 + [\nu(1 - f^2) - q]p^- - q\nu = 0$ with unique positive root

$$p^- = \frac{q - \nu(1 - f^2) + \sqrt{(q - \nu(1 - f^2))^2 + 4q\nu}}{2}.$$

The Kalman gain is $K = (p^-/(p^- + \nu))\, I$ and the closed-loop transition is

$$A_{\mathrm{cl}} = F(I - KC) = \rho_{\mathrm{scale}}\, U \cdot \frac{\nu}{p^- + \nu}\, I = \frac{\rho_{\mathrm{scale}}}{1 + p^-/\nu}\, U.$$

A_{cl} is normal (scalar multiple of orthogonal), so

$$\sigma_1(A_{\mathrm{cl}}) \;=\; \frac{\rho_{\mathrm{scale}}}{1 + \mathrm{SNR}_{\mathrm{bulk}}}, \qquad \mathrm{SNR}_{\mathrm{bulk}} := \frac{p^-}{\nu} = \frac{\operatorname{tr} P_{\mathrm{ss}}^-}{\operatorname{tr} R}. \tag{30}$$

The predictive observation variance is $p^- + \nu$, i.e. $\Sigma_Y^{\mathrm{pred}} = CP_{\mathrm{ss}}^- C^\top + R$ in matrix form; this is the natural comparator for empirical residual covariance. For graph modes, replace f by the per-mode coefficient f_i; the mode-wise prior variance solves the same quadratic in (f_i, q, ν), and the mode-wise closed-loop coefficients $a_{\mathrm{cl},i} = f_i \nu/(p_i^- + \nu)$ are the eigenvalues that feed Theorem 19.

The differentiable surrogate ρ_{snr}. Equation (30) depends on P_{ss}, which itself requires a DARE solve. We replace P_{ss} by its isotropic projection on the right-hand side: define

$$\rho_{\mathrm{snr}} \;:=\; \frac{\rho_{\mathrm{scale}}}{1 + \operatorname{tr}(P_{\mathrm{ss}})/\operatorname{tr}(R)}. \tag{31}$$

Under exact isotropy ($P_{\mathrm{ss}} = pI$, $R = rI$) we have $\rho_{\mathrm{snr}} = \sigma_1(A_{\mathrm{cl}})$. For non-isotropic (P_{ss}, R) a first-order perturbation argument (linearising the DARE around its isotropic solution) gives $\rho_{\mathrm{snr}} = \sigma_1(A_{\mathrm{cl}}) + \mathcal{O}(\varepsilon)$, where $\varepsilon = \|P_{\mathrm{ss}} - p_\star I\|_{\mathrm{op}}/p_\star$ measures departure from isotropy.

Why the surrogate is useful. Three properties: (i) ρ_{snr} is computable in $\mathcal{O}(N)$ time (one trace ratio) vs. $\mathcal{O}(N^3)$ for the DARE iteration, which matters for hyperparameter gradients over $(\rho_{\mathrm{scale}}, \sigma_Q, \sigma_R)$ when $N \in \{20, 50, 207, \ldots\}$. (ii) It is differentiable in each of these hyperparameters and monotone decreasing in $\mathrm{SNR}_{\mathrm{bulk}}$, so gradient-based tuning lifts SNR (and contracts Σ_t) in the expected direction. (iii) It agrees with the Theorem 19 asymptotic rate at isotropy, which is the regime we use for the reported GRAPH-LGSSM runs, so the surrogate equals the theoretical rate at isotropy.

Empirical tightness. Table 10 shows $\rho_{\mathrm{snr}} \in [0.436, 0.443]$ across the 10 real-data evaluation cells, matching $\sigma_1(A_{\mathrm{cl}}) \in [0.433, 0.441]$ to within 0.002–0.004. The first-order perturbation error is therefore below 1% on every cell we tested.

C Filter specifications, training, and pseudocode

C.1 GRAPH-LGSSM (linear-Kalman baseline)

State-space model. GRAPH-LGSSM is a linear-Gaussian state-space model with latent dimension $d = N$,

$$H_{t+1} \;=\; F\, H_t + \xi_t, \qquad Y_t \;=\; C\, H_t + \eta_t, \qquad \xi_t \sim \mathcal{N}(0, Q),\ \eta_t \sim \mathcal{N}(0, R), \tag{32}$$

with the following choices, fixed before any real-data evaluation:

- **Transition** $F = \rho_{\text{scale}}\,\widetilde{A}_{\text{sym}}/\lambda_{\max}(\widetilde{A}_{\text{sym}})$, where $\widetilde{A}_{\text{sym}} = \widetilde{D}^{-1/2}(A + I_N)\widetilde{D}^{-1/2}$ is the self-loop-augmented symmetric-normalised graph shift defined in App. E.2. Scaling by $\lambda_{\max}(\widetilde{A}_{\text{sym}})^{-1}$ ensures $\|F\|_{\text{op}} = \rho_{\text{scale}}$; we use $\rho_{\text{scale}} = 0.8$ throughout, which keeps the open-loop dynamics sub-unit (matching the contraction hypothesis (C1)) and was found to give well-conditioned Riccati recursions on every real dataset.
- **Observation matrix** $C = I_N$ (sensors observe the latent state directly, i.e. the latent dimension equals the sensor dimension).
- **Innovation noise** $Q = \sigma_Q^2 I_N$ with $\sigma_Q = 1$.
- **Observation noise** $R = \sigma_R^2 I_N$ with $\sigma_R = 1$.

Riccati recursion. The one-step predictive covariance follows the standard discrete-time Kalman recursion

$$\begin{aligned} P_{t|t-1} &= F\,P_{t-1|t-1}\,F^\top + Q, \\ K_t &= P_{t|t-1}\,C^\top\big(CP_{t|t-1}C^\top + R\big)^{-1}, \\ P_{t|t} &= (I - K_t C)\,P_{t|t-1}. \end{aligned}$$

We initialise $P_{0|0} = Q$ and iterate to a relative tolerance $\|P_{t|t} - P_{t-1|t-1}\|_F / \|P_{t|t}\|_F \le 10^{-9}$; convergence requires ≤ 200 iterations on every dataset we tested. The steady-state covariance P_{ss} and gain K are then frozen and used for the calibration and test passes; the closed-loop transition $A_{\text{cl}} = F(I - KC)$ is computed once for the contraction-rate diagnostics of Table 10.

Hyperparameter selection. The Riccati tolerance is fixed (10^{-9}). Across the paper we evaluate GRAPH-LGSSM in two modes: an *untuned* variant with the theory-motivated defaults $(\rho_{\text{scale}}, \sigma_Q, \sigma_R) = (0.8, 1, 1)$ — used for all contraction-rate diagnostics in Table 10 — and a *validation-NLL-tuned* variant with $(\rho_{\text{scale}}, \text{obs_noise_frac})$ picked on a 20% validation tail of the training block (App. C.2). The tuned variant is the one reported in the KALMAN-FCP rows of Tables 2–3 and Table 14. The closed-form sensitivity in App. B.13 shows that the asymptotic rate is monotone in $\rho_{\text{scale}}/(1 + \text{SNR}_{\text{bulk}})$, so the qualitative behaviour is preserved across reasonable choices of the untuned defaults.

C.2 GNF (learned GCN–GRU)

We instantiate the framework with a graph-convolutional GRU cell (diffusion-convolutional recurrent style, [5, 17]) feeding a diagonal-plus-low-rank covariance head. Graph structure enters the recurrence directly: every node carries its own d-dimensional hidden state, and neighbour aggregation is folded into the GRU gates through a fixed graph shift. The architecture is intentionally minimal — the trainable weights are shared across nodes, so the parameter count is independent of N — so that the sharpness gain we report is not confounded with model capacity. Table 4 consolidates every hyperparameter; all values were selected once on a synthetic GRSSF pilot and then frozen for all reported real-data runs.

Graph shift. Let $S = \widetilde{A}_{\text{sym}} = (\widetilde{D})^{-1/2}(A + I_N)(\widetilde{D})^{-1/2}$ be the symmetric-normalised adjacency with self-loops, where $\widetilde{D} = \text{diag}\big(\sum_j (A + I)_{ij}\big)$ is the degree matrix of the augmented graph [15]. The shift is fixed (not learned), satisfies $\|S\|_{\text{op}} \le 1$, and enters the cell as the neighbour-aggregation operator

$$(S\,H)_i \;=\; \textstyle\sum_j S_{ij}\,h_j, \qquad i = 1, \dots, N, \tag{33}$$

acting on any per-node array $H = (h_1, \dots, h_N)$.

Per-node recurrent update. Each node i carries a hidden state $h_{t,i} \in \mathbb{R}^d$ with $d = 32$, so the full filter state $H_t = (h_{t,1}, \dots, h_{t,N}) \in \mathbb{R}^{N \times d}$ has dimension dN (640 at N=20, 1600 at N=50). With the per-node scalar input $y_{t,i} \in \mathbb{R}$ ($f_{\text{in}} = 1$) the graph-convolutional GRU cell applies node-shared weights $W_z, W_r, W_n \in \mathbb{R}^{d \times (2d + f_{\text{in}})}$:

$$z_t = \sigma\big(W_z\,[\,H_{t-1},\, SH_{t-1},\, Y_t\,]\big), \tag{34}$$

$$r_t = \sigma\big(W_r\,[\,H_{t-1},\, SH_{t-1},\, Y_t\,]\big), \tag{35}$$

$$\widetilde{H}_t = \tanh\big(W_n\,[\,r_t \odot H_{t-1},\, S(r_t \odot H_{t-1}),\, Y_t\,]\big), \tag{36}$$

$$H_t = (1 - z_t) \odot H_{t-1} + z_t \odot \widetilde{H}_t, \tag{37}$$

where $[\cdot,\cdot,\cdot]$ concatenates along the feature axis so that each node sees $[h_i, (SH)_i, y_i] \in \mathbb{R}^{2d+f_{\text{in}}}$, σ is the logistic sigmoid, and $\odot$ is the Hadamard product. Graph mixing thus acts inside both the gates and the candidate state; there is no mean-pooling and no global summary vector.

Predictive head. Node-shared linear heads map each node's pre-observation state $h_{t-1,i} \in \mathbb{R}^d$ to its mean, log-variance, and low-rank factor row:

$$\hat{Y}_{t|t-1,i} = w_\mu^\top h_{t-1,i} + b_\mu, \qquad w_\mu \in \mathbb{R}^d, \tag{38}$$

$$d_{t,i} = \max\big(\exp(w_d^\top h_{t-1,i} + b_d),\, \varepsilon_d\big), \qquad w_d \in \mathbb{R}^d, \tag{39}$$

$$(L_t)_{i,:} = W_L\, h_{t-1,i} + b_L, \qquad W_L \in \mathbb{R}^{r\times d}, \tag{40}$$

with $\varepsilon_d = 10^{-4}$ a strict floor clamped onto the diagonal head and $L_t \in \mathbb{R}^{N\times r}$ formed by stacking the rows $(L_t)_{i,:}$. The predictive covariance is the diagonal-plus-low-rank $\hat{\Sigma}_{t|t-1} = \mathrm{diag}(d_t) + L_t L_t^\top \succ 0$, so the matrix-determinant lemma and Woodbury identity give $\mathcal{O}(Nr^2 + r^3)$ inversion and log-determinant computations during training and inference (linear in N for fixed $r = 4$). We fix $r = 4$ (rank-sensitivity sweep in App. E.13).

Initialisation. The input-facing blocks of the cell weights (W_z, W_r, W_n) and the head weights (w_μ, w_d, W_L) use Glorot uniform initialisation [10]; the recurrent (state-facing) blocks use orthogonal initialisation [22] to avoid early-training spectral collapse. All biases are zero. The per-node state is reset to $H_0 = \mathbf{0}_{N\times d}$ at each training epoch and at the start of calibration, then carried into the test pass.

Training objective. Parameters $\theta = (W_z, W_r, W_n, w_\mu, b_\mu, w_d, b_d, W_L, b_L)$ are fit by minimising the negative Gaussian predictive log-likelihood

$$\mathcal{L}(\theta) \;=\; \sum_{t=1}^{T_{\text{tr}}} \Big[\tfrac{1}{2}(Y_t - \hat{Y}_{t|t-1})^\top \hat{\Sigma}_{t|t-1}^{-1} (Y_t - \hat{Y}_{t|t-1}) + \tfrac{1}{2} \log\det \hat{\Sigma}_{t|t-1} \Big], \tag{41}$$

via truncated backpropagation through time with a BPTT window of 64 steps and a 10% loss burn-in (the first $\max(0.1\,T_{\text{tr}}, 5)$ steps of each epoch are excluded from the loss). We add an Adam weight decay of 10^{-4} as the only explicit regulariser; the variance floor ε_d lower-bounds $\hat{\Sigma}_{t|t-1}$ and the orthogonal initialisation keeps the recurrence stable for the 60 epochs we train.

Optimiser and schedule. Adam [14] with $(\beta_1, \beta_2, \eta) = (0.9, 0.999, 10^{-3})$, weight decay 10^{-4}, and ℓ_2-norm gradient clipping at 5.0. Training runs over the contiguous training sequence with gradients detached every 64 steps (TBPTT). We train for 60 epochs without learning-rate scheduling and without early stopping; on every dataset we report, the validation log-likelihood plateaus by around epoch 40. We did not validation-tune any hyperparameter; Table 4 lists the full set.

Table 4: **GNF hyperparameters.** Fixed across all datasets and seeds; no validation-set sweep was run.

Group	Quantity	Value
Graph	per-node input width f_{in}	1 (per-sensor scalar)
	graph shift S	sym.-normalised adjacency, $\|S\|_{\text{op}} \le 1$
GCN-GRU cell	per-node hidden width d	32
	full state dimension	dN (640/1600 at N=20/50)
	weights	node-shared W_z, W_r, W_n
Head	covariance rank r	4
	variance floor ε_d	10^{-4}
Optimisation	optimiser	Adam $(\beta_1, \beta_2) = (0.9, 0.999)$
	learning rate η	10^{-3}
	weight decay	10^{-4}
	BPTT window	64
	loss burn-in	10%
	epochs	60
	gradient clipping	$\|\cdot\|_2 \le 5.0$
Inference	warm-up t_0	50 steps
	target level α	0.1

Linear-Kalman hyperparameter tuning. The KALMAN-FCP rows in Tables 2–3 use a validation-NLL grid search over $\rho_{\text{scale}} \in \{0.5, 0.7, 0.85, 0.95\} \times \text{obs_noise_frac} \in \{0.03, 0.1, 0.3, 1.0\}$ with the graph polynomial order fixed at $K = 1$ and a 20% validation tail reserved from the training block (warmup $t_0 = 50$). Per-step Gaussian NLL is the selection metric; the chosen pair is re-fit on the full training block before calibration. Tuning adds $\sim 30\,\text{s}$ per seed and narrows the filter's w_{proj} against the untuned default $(\rho_{\text{scale}}, \text{obs_noise_frac}) = (0.8, 0.1)$ by 9% on METR-LA-20 and 13% on PEMS-BAY-50; the structural Riccati-tail gap on real traffic remains (App. E.11).

Parameter footprint. Because the cell and heads are shared across nodes, the trainable count is *independent of* N: with $d = 32$, $r = 4$, $f_{\text{in}} = 1$ the three cell maps contribute $3\big[(2d + f_{\text{in}})d + d\big] = 6336$ parameters and the three heads contribute $(2 + r)d + (2 + r) = 198$, for $\approx 6.5 \times 10^3$ trainable parameters on every graph. N enters only through the fixed graph shift S and the state size dN. Full training (60 epochs, $\sim 45\,\text{s}$ per epoch) takes 44–48 minutes per seed on a single RTX 4090 (Table 20).

Inference protocol. Algorithm 2 is the offline training loop; Algorithm 3 is the operational form of Algorithm 1 that consumes the resulting frozen θ on the calibration and test sequences. At calibration and test time the filter is run autoregressively on the observed $\{Y_t\}$, with the previously emitted $\hat{Y}_{t|t-1}$ replacing any unobserved input only in the streaming forecast horizon (App. C.3 discusses the training/inference-mode distinction).

Algorithm 2 GNF training loop

Require: training sequence $\{Y_t\}_{t=1}^{T_{\text{tr}}}$, graph shift $S = \widetilde{A}_{\text{sym}}$, epochs E, TBPTT window 64, loss burn-in (10%)

initialise θ as in the "Initialisation" paragraph
for $e = 1, \ldots, E$ **do**
  set $H \leftarrow \mathbf{0}_{N \times d}$, $\mathcal{L} \leftarrow 0$, window counter $c \leftarrow 0$
  for $t = 1, \ldots, T_{\text{tr}}$ **do**
    emit $(\hat{Y}_{t|t-1}, d_t, L_t) = g_\theta(H)$; $\hat{\Sigma}_t = \text{diag}(d_t) + L_t L_t^\top$
    advance the per-node state $H \leftarrow \text{GCNGRU}(Y_t, H, S)$ via (34)–(37)
    if $t >$ burn-in **then**
      $\mathcal{L} \mathrel{+}= \frac{1}{2}(Y_t - \hat{Y}_{t|t-1})^\top \hat{\Sigma}_t^{-1}(Y_t - \hat{Y}_{t|t-1}) + \frac{1}{2}\log\det\hat{\Sigma}_t$; $c \leftarrow c + 1$
    end if
    if $t \bmod 64 = 0$ **or** $t = T_{\text{tr}}$ **then**
      Adam step on $\mathcal{L}/\max(c, 1)$ with $\|\cdot\|_2$-clipping at 5.0; then detach H (TBPTT) and reset $\mathcal{L} \leftarrow 0$, $c \leftarrow 0$
    end if
  end for
end for

Algorithm 3 Filtered split-conformal inference (operational form of Algorithm 1)

Require: trained parameter θ, initial state H_0, calibration sequence $Y_{1:n}$, level α, first retained index t_0

$H \leftarrow H_0$
for $t = 1, \ldots, n$ **do** ▷ calibration pass
  $(\hat{Y}_{t|t-1}, \hat{\Sigma}_{t|t-1}) \leftarrow g_\theta(H)$ via (38)
  $R_t \leftarrow (Y_t - \hat{Y}_{t|t-1})^\top \hat{\Sigma}_{t|t-1}^{-1}(Y_t - \hat{Y}_{t|t-1})$
  $H \leftarrow \text{GCNGRU}(Y_t, H, S)$ via (34)–(37)
end for
$m \leftarrow n - t_0 + 1$; $\hat{q}_{1-\alpha} \leftarrow R_{(\lceil (m+1)(1-\alpha) \rceil)}$ from $R_{t_0:n}$, with $R_{(m+1)} = +\infty$
for $t = n + 1, n + 2, \ldots$ **do** ▷ test pass; carry H from calibration
  $(\hat{Y}_{t|t-1}, \hat{\Sigma}_{t|t-1}) \leftarrow g_\theta(H)$
  emit $\hat{C}_t \leftarrow \{y : (y - \hat{Y}_{t|t-1})^\top \hat{\Sigma}_{t|t-1}^{-1}(y - \hat{Y}_{t|t-1}) \le \hat{q}_{1-\alpha}\}$
  observe Y_t, then $H \leftarrow \text{GCNGRU}(Y_t, H, S)$
end for

C.3 Teacher forcing and conformal validity

GNF is trained with teacher-forced inputs: within a training window the input for predicting Y_{u+1} is the observed Y_u, $H_u = \text{GCNGRU}(Y_u, H_{u-1}; S)$, and the negative log-likelihood compares

$(\hat{Y}_{u+1|u}, \hat{\Sigma}_{u+1|u})$ to Y_{u+1}. At calibration and test time the filter is deployed one-step-ahead in the same mode (Algorithm 3): before Y_t is observed, H_{t-1} contains observations through Y_{t-1} and the predictive moments $(\hat{Y}_{t|t-1}, \hat{\Sigma}_{t|t-1})$ are emitted; the score is evaluated only after Y_t is observed, so no calibration or test score uses Y_t to predict itself. Two distributions therefore coexist: the *training-time* score distribution induced by teacher-forced windows used to arrive at the parameters θ, and the *inference-time* score distribution driven by the frozen θ on the observed sequence $Y_{1:n+T}$. Conformal validity concerns only the latter.

Claim. The validity analysis of Algorithm 1 depends only on the inference-mode score process generated by the frozen filter ϕ: Proposition 15 bounds its coverage gap via Theorems 12–14, and exact split-conformal validity follows, for example, under the standardised-innovation condition of Proposition 8. The training procedure enters only through the constants (ρ, L_S, $\bar{g}$) of those statements, because the parameters are frozen at the start of the calibration pass.

Argument. Because Algorithm 3 conditions both calibration and test passes on the observed Y_{t-1}, the calibration scores $(S_{t_0}, \ldots, S_n)$ and the test score S_{n+k} are drawn from the *same* stochastic process generated by the frozen ϕ applied in the same mode to a contiguous observation sequence. Propositions 15 and 8 then apply as stated. Mis-specified teacher-forced training would enter only through a larger root-score Lipschitz constant L_S (Lemma 17) or K_{loc} (Theorem 12), not as a broken validity guarantee. App. E.15 (Audit 1) reports the empirical root-score finite-difference quotients; the warm-up $t_0 \geq 50$ of App. E.14 drives the initialisation bias below double precision on every cell.

C.4 NEURALSSM (DIAG-GRU) and the rank-0 ablation (GCN-RANK-0)

We use two distinct diagonal-covariance ablations, with precise labels.

DIAG-GRU (NeuralSSM): graph removed *and* diagonal Σ_t. DIAG-GRU is a standard *global* GRU over the stacked sensor vector $Y_t \in \mathbb{R}^N$ (a single recurrent state of width $h = 32$ rather than the per-node states of (34)) with a diagonal Gaussian head, $\hat{\Sigma}_{t|t-1} = \mathrm{diag}(d_t)$; the graph is not used (S is ignored). It shares the optimiser, floor, and schedule of Table 4 and uses $r = 0$. This is the *NeuralSSM* row of Table 2, with w_{proj} 4.71 at joint coverage 0.932 on METR-LA-20.

GCN-RANK-0: graph retained, diagonal Σ_t. GCN-RANK-0 *keeps* the per-node graph cell of (34) and only sets $r = 0$ in the head. This is the rank sweep's r=0 point in App. E.13 ($w_{\mathrm{proj}} = 5.32$ on METR-LA-20); it isolates the covariance head *at fixed graph mixing*, and should not be confused with the (architecturally distinct) global-GRU DIAG-GRU row of Table 2.

Reading the two rows. DIAG-GRU removes both graph structure and the low-rank head and also changes the recurrence to a global GRU, so its gap to the full filter (GNF) is not a clean single-mechanism ablation. GCN-RANK-0 isolates the rank-r factor at fixed per-node graph recurrence: the rank sweep GCN-RANK-0 $\rightarrow$ r=4 shows the low-rank head dominates the sharpness gain ($5.32 \rightarrow 3.04$). The clean graph-removal ablation *at fixed per-node architecture* is the LR-GRU covariance control of App. E.22; the GCN-RANK-0 $\rightarrow$ DIAG-GRU comparison is only suggestive.

C.5 Adaptive-conformal-inference wrap

Adaptive Conformal Inference [9] adjusts the nominal level α_t online to track the realised miscoverage rate. Wrapping GNF proceeds as follows: initialise $\alpha_1 = \alpha = 0.1$; at each test step t emit the prediction set $\hat{C}_{\alpha_t}(x_t)$ via Algorithm 3 with the recomputed quantile $\hat{q}_{1-\alpha_t}$ (using the same frozen calibration scores), observe Y_t, and update

$$\alpha_{t+1} \leftarrow \alpha_t + \gamma\big(\alpha - \mathbb{1}\{Y_t \notin \hat{C}_{\alpha_t}(x_t)\}\big).$$

We fix $\gamma = 0.005$ for every reported cell and do not tune it per dataset. Recomputing $\hat{q}_{1-\alpha_t}$ from the frozen calibration scores is $\mathcal{O}(m \log m)$ per step, which adds negligible test-time cost (Table 20). The long-run validity guarantee of Gibbs and Candès [9, Thm. 1] applies as stated to the wrapped scores.

D Baseline implementations

This section gives implementation details for every baseline that appears in Tables 2–14. Across all baselines we use $\alpha = 0.1$, 10 seeds per cell, and the same chronological 40/40/20 train/calib/test split (App. E.1). All filter baselines share the warm-up $t_0 = 50$ used by the main method.

D.1 Static-covariance methods

Mean-predictor protocol. The non-filter covariance baselines (CGIF-JOINT, FACTORCGIF, GROUPCGIF, SPECTRALCGIF, ACI/AgACI variants, EWMA, rolling, time-of-day, k-NN local ellipsoid) share the same backbone mean predictor $\hat{Y}_{t|t-1}$ within a given table (default GPVAR of §6). Filter rows (GNF, NEURALSSM, KALMAN-FCP) use their own mean mechanisms, as do COPULACPTS and MULTIDIMSPCI; Table 5 maps each method to its covariance mechanism and to its mean source. Backbone-swap experiments (Tables 11, 12, 13) should be interpreted as a mean-channel stress test rather than as a perfect same-mean decomposition of mean and covariance effects: all covariance heads (including the learned GNF head) stay fixed while the shared mean predictor varies, but the learned covariance was trained jointly with its native filter, so w_{proj} movement across backbone columns reflects the mean channel under that constraint.

Table 5: **Method → mean predictor → covariance mechanism** for every row of Tables 2–3 and 14. "Backbone" is the mean predictor $\hat{Y}_{t|t-1}$ shared across methods in a given results table. The last two columns separate *model fitting* (how the mean/covariance parameters are estimated) from *real-data hyperparameter search* (number of configurations evaluated on a real-data validation split); "fit" quantities are training, not tuning. Full ledger with selection criteria in Table 21.

Method	Mean $\hat{Y}_{t\|t-1}$	Covariance mechanism	Cov./model fit	HP search	ACI
STATIC-CGIF (`CGIF-joint`)	backbone	empirical $\hat{\Sigma}$	empirical	0	—
FACTORCGIF	backbone	rank-r factor ∩ residual box	empirical	0	—
GROUPCGIF	backbone	per-community intersection	empirical	0	—
SPECTRALCGIF	backbone	Laplacian-whitened bands	empirical	0	—
AGACI-GROUPCGIF	backbone	per-community + AgACI	empirical	0	✓
ACI-PERGROUP-FACTORCGIF	backbone	per-group factor + ACI	empirical	0	✓
ACI-FACTORCGIF	backbone	factor + ACI	empirical	0	✓
DIAG-GRU (`NeuralSSM`)	own GRU	diag-Σ_t head	MLE	0	—
KALMAN-FCP	own LGSSM	linear-Kalman Σ_t	val. NLL	16/cell	—
EWMA Σ_t	backbone	EWMA of $r_t r_t^\top$	fixed	0	—
Rolling Σ_t	backbone	rolling $r_t r_t^\top$	fixed	0	—
Time-of-day Σ_t	backbone	per-bin $\hat{\Sigma}$	empirical	0	—
Local ellipsoid	backbone	k-NN local $\hat{\Sigma}$	fixed	0	—
COPULACPTS	own	copula-calibrated coordinate thresholds	empirical (ref.)	0	—
MULTIDIMSPCI	own	ellipsoidal band	ref. default	0	—
DEEPVAR-FCP (control)	own GRU	global low-rank Σ_t (rank 4)	MLE	0	—
GNF **(ours)**	own GCN-GRU	diag+low-rank head	MLE	0	—
+ ACI wrap	own GCN-GRU	same + online α_t	MLE	0	✓

STATIC-CGIF (`CGIF-joint`, static-covariance reference). Calibration-block residuals $r_t = Y_t - \hat{Y}_{t|t-1}$, $t \in [t_0, n]$, are pooled to estimate the empirical covariance $\hat{\Sigma} = (m-1)^{-1} \sum_{t=t_0}^{n} r_t r_t^\top$ with $m = n - t_0 + 1$. The static $\hat{\Sigma}$ is plugged into the squared Mahalanobis score and the conformal threshold is the $\lceil (m+1)(1-\alpha) \rceil$-th order statistic of $\{r_t^\top \hat{\Sigma}^{-1} r_t\}$. The forecasts $\hat{Y}_{t|t-1}$ come from the shared backbone above; no learnable parameters; no validation-set tuning.

FACTORCGIF (r=4). Rank-r truncated SVD on the calibration residual matrix, $E_{\mathrm{cal}} = \widehat{B}\widehat{F} + E_{\mathrm{res}}$ with $\widehat{B} \in \mathbb{R}^{N \times r}$ and $\widehat{F} \in \mathbb{R}^{r \times m}$. The score decomposes into a factor sub-score $\|\cdot\|_{\Sigma_F^{-1}}$ on the projection $\widehat{P}_{\widehat{B}} r_t$ and a residual max-norm sub-score on $(I - \widehat{P}_{\widehat{B}}) r_t$, calibrated separately at levels $\theta\alpha$ and $(1-\theta)\alpha$ (default split $\theta = 0.5$); the union bound on levels gives joint coverage $\geq 1 - \alpha$. The returned joint set is the *intersection* $\{y : \|F(y)\|_{\Sigma_F^{-1}} \leq q_F\} \cap \{y : \max_i |R_i(y)| / \hat{\sigma}_i^R \leq q_R\}$ of the factor ellipsoid and the residual box (projection in App. D.2). We fix $r = 4$ globally to match the GNF rank budget; no per-dataset rank search is performed (a retrospective sensitivity analysis shows a plateau near r=4).

GROUPCGIF. Calibration residuals are clustered by community structure of the adjacency $\widetilde{A}$: Louvain modularity maximisation [4] at the default resolution 1.0, breaking ties by node index. Within each community a Mahalanobis score is computed against the community's empirical covariance, and the joint set is the intersection of the per-community ellipsoids, each calibrated at level α/K for K communities (a union bound on levels). No learnable parameters.

SPECTRALCGIF. Whitens calibration residuals on the eigenbasis of the graph Laplacian $L = I - \widetilde{A}$, $\tilde{r}_t = U^\top r_t$, where U is the matrix of L-eigenvectors (computed once on the calibration adjacency). The whitened scores are calibrated separately for low/medium/high spectral bands (split at the 33%

and 67% eigenvalue quantiles), and the prediction set is the intersection (simultaneous preimage) of the band-wise ellipsoids, not a set union (projection in App. D.2). No learnable parameters.

D.2 Evaluation of w_{proj} for each set family

For every method we evaluate w_{proj} in (3) on the *actual returned joint set* $\hat{C}_t$: we compute the exact coordinate extent $\ell_{t,i} = \inf_{y \in \hat{C}_t} y_i$ and $u_{t,i} = \sup_{y \in \hat{C}_t} y_i$ and average $u_{t,i} - \ell_{t,i}$ over coordinates and test steps. Joint coverage is always evaluated on $\hat{C}_t$ itself, never on its axis-aligned bounding box. Table 6 gives, per family, the returned set, its composition, and the closed-form coordinate half-width $h_{t,i} := (u_{t,i} - \ell_{t,i})/2$. All ellipsoid endpoints use the support function $\sup_{y \in \hat{C}_t} e_i^\top y = e_i^\top \hat{\mu}_t + \sqrt{\hat{q}_t \, (\hat{\Sigma}_t)_{ii}}$ (closed form). ACI and AgACI variants change only the calibrated level(s) $\hat{q}$ and reuse the inner-set endpoint rule of their base family.

Table 6: **Coordinate-projection rule for w_{proj} by set family.** $h_{t,i}$ is the per-node half-width; $u_{t,i} - \ell_{t,i} = 2h_{t,i}$. All coordinate endpoints are closed form after fitting and calibration; CopulaCPTS returns the coordinate thresholds specified in App. D.5.

Family	Returned set	Composition	Half-width $h_{t,i}$
Single ellipsoid (STATIC-CGIF, KALMAN-FCP, DIAG-GRU, GNF, MULTIDIMSPCI)	$\{(y - \hat{\mu}_t)^\top \hat{\Sigma}_t^{-1} (y - \hat{\mu}_t) \le \hat{q}_t\}$	single	$\sqrt{\hat{q}_t \, (\hat{\Sigma}_t)_{ii}}$
FACTORCGIF / per-group factor	factor ellipsoid $\cap$ residual box	intersection	$q_F \sqrt{b_i^\top \Sigma_F b_i} + q_R \hat{\sigma}_i^R$
GROUPCGIF	$\bigcap_k$ community-k ellipsoid (partition)	intersection	$q_k \sqrt{(\hat{\Sigma}_k)_{ii}}$, $i \in$ community k
SPECTRALCGIF	$\bigcap_k$ band-k ellipsoid (preimage)	intersection	$\sum_k q_k \, \lvert U_{ik} \rvert \, \sigma_k$
COPULACPTS	$\bigcap_i \{\lvert y_i - \hat{y}_{t,i} \rvert \le q_i\}$	jointly calibrated box	q_i

D.3 Adaptive-conformal-inference variants

AGACI. Aggregated ACI of Zaffran et al. [29]: outputs of three parallel ACI runs with $\gamma \in \{0.001, 0.005, 0.025\}$ are aggregated by online mirror descent on the running miscoverage indicator. We use the implementation defaults from Zaffran et al. [29].

ACI-FACTORCGIF, AGACI-GROUPCGIF, ACI-PERGROUP-FACTORCGIF. ACI / AgACI wraps composed with the FACTORCGIF / GROUPCGIF / per-group FACTORCGIF score families. The composition rule follows the standard online-CP construction: the inner score is calibrated once on the offline calibration block; the outer ACI/AgACI step adjusts only the level α_t, leaving the inner family's parameters frozen. Step sizes are inherited from the inner family ($\gamma = 0.005$ for ACI; the AgACI grid above).

D.4 Filter baselines

KALMAN-FCP (linear-Kalman filter). The GRAPH-LGSSM specification of App. C.1 plugged into the filtered split-conformal protocol of Algorithm 3. Reported rows use the validation-NLL-tuned $(\rho_{\mathrm{scale}}, \mathrm{obs_noise_frac})$ (App. C.2); the theory-motivated $(\rho_{\mathrm{scale}}, \sigma_Q, \sigma_R) = (0.8, 1, 1)$ defaults are used only for the contraction-rate diagnostics of Table 10 (see App. C.1).

DIAG-GRU (`NeuralSSM`). The ablation specification of App. C.4: identical training schedule to GNF but with the graph term and the low-rank head removed. This comparison also changes the recurrence from per-node states to a global GRU (App. C.4).

D.5 Published joint-conformal baselines

The two published joint-CP baselines we benchmark against most frequently are COPULACPTS [24] and MULTIDIMSPCI [28]. Because these methods appear in essentially every results table in this paper, we describe them in enough detail for an independent reimplementation and contrast them point-by-point with our filtered construction.

COPULACPTS [24]: copula-based joint conformal prediction. We use the vendored reference calibrator, which splits the available calibration residuals randomly into 60% for marginal score estimation and 40% for dependence calibration. For one-step prediction the coordinate score is $s_{t,i} = |Y_{t,i} - \hat{y}_{t,i}|$. On the first subset $\mathcal{I}_1$, form the empirical marginal CDF $\hat{F}_i(s) = |\mathcal{I}_1|^{-1} \sum_{t \in \mathcal{I}_1} \mathbf{1}\{s_{t,i} < s\}$. The second subset supplies tuples $U_{t,i} = \hat{F}_i(s_{t,i})$. The reference optimiser selects coordinate

thresholds a whose empirical lower-orthant coverage is close to $1-\alpha$, using the smooth objective

$$\left| \frac{1}{|\mathcal{I}_2|} \sum_{t \in \mathcal{I}_2} \prod_{i=1}^{N} \mathrm{sigmoid}\big(1000(a_i - U_{t,i})\big) - (1-\alpha) \right|.$$

It runs Adam for 800 iterations with weight decay 10^{-4}. Let $m_1 = |\mathcal{I}_1|$ and let $s_{(j),i}$ be the sorted marginal scores on the first subset. Following the reference code's finite indexing convention, set $q_i = s_{(j_i),i}$ with $j_i = \min\{m_1, \mathrm{int}(a_i m_1) + 2\}$. The emitted set is the bounded score-sublevel box

$$\hat{C}_t^{\mathrm{cop}} = \{y : |y_i - \hat{y}_{t,i}| \le q_i \text{ for every } i\},$$

whose exact coordinate half-width is q_i. The dependence calibration couples the choices of coordinate thresholds, but the resulting set is axis-aligned. No KDE, bandwidth selection, or copula-CDF superlevel set is used. The thresholds are fixed after calibration; the method does not evolve a predictive covariance as our filter does. Wider sets at large N can reflect the cost of simultaneous coordinate coverage; we do not attribute that effect to a bandwidth rule.

MULTIDIMSPCI [28]: sequential predictive conformal inference for multi-dimensional outputs. The method generalises the per-coordinate SPCI [27] to vector-valued $r_t \in \mathbb{R}^N$ by predicting an *ellipsoid radius* from a residual quantile network. Let $r_{1:t-1}$ be the running calibration residuals; the network $g_\phi(r_{1:t-1}; \hat{\Sigma})$ outputs a positive scalar q_t, and the test-time set is the ellipsoid $\{y : (y-\hat{y}_t)^\top \hat{\Sigma}^{-1}(y-\hat{y}_t) \le q_t^2\}$ where $\hat{\Sigma}$ is a Bayesian-shrunk estimate $\hat{\Sigma} = (1-\lambda)\hat{\Sigma}_{\mathrm{cal}} + \lambda\,\sigma^2 I$ of the calibration covariance with shrinkage $\lambda = 0.05$ (default). The quantile network is trained on the calibration residuals to minimise a pinball loss at level $1-\alpha$, with two hidden layers of width 128, ReLU activations, 20 Adam epochs at learning rate 10^{-3}, and a 32-step rolling window of past residuals as the input feature. At test time the network is rolled forward without re-training. Two structural differences from our construction: (i) the covariance shape is fixed at calibration time ($\hat{\Sigma}$ does not vary with t); only the radius q_t adapts, which is exactly the predictive-shape constraint we relax with the filtered Σ_t; and (ii) the radius network is trained on raw residuals, so its target depends on the upstream forecaster's absolute scale—a property that the static $\hat{\Sigma}$ in our STATIC-CGIF baseline shares.

Why these two baselines? They span the two natural relaxations of static-covariance joint CP that one would attempt *before* adding a filter: COPULACPTS calibrates joint coordinate thresholds, and MULTIDIMSPCI relaxes the radius (time-varying q_t instead of a fixed quantile). Our method instead learns a time-varying covariance shape by filtering Σ_t per step. The empirical w_{proj} gap to either baseline at moderate N and graph-aligned dependence (Tables 2, 14) is consistent with the value of that combined relaxation.

Per-coordinate baselines (SPCI and HOPCPT). SPCI [27] and HOPCPT [2] are per-coordinate by construction; they appear in Table 14 only to document that purely marginal CP sits far below the joint-coverage target on correlated Y (joint coverage 0.21 and 0.41, respectively, against the 0.9 target). For SPCI we use the reference implementation defaults; for HOPCPT we use window $W = 64$ and $T_{\mathrm{train}} = 10\,000$ updates.

E Additional experiments

E.1 Datasets and splits

This subsection describes each of the eight real correlated-sensor datasets we evaluate (which together yield 10 real-data evaluation cells in the main results: METR-LA and PEMS-BAY each appear at both a moderate-N subgraph and the full graph), together with the synthetic GRSSF benchmark, the common preprocessing pipeline, the chronological split, and the busiest-k subgraph protocol. The adjacency matrix used by the GCN layer of GNF and by the graph polynomial of GRAPH-LGSSM is described separately in App. E.2. Table 7 summarises the inventory.

Common preprocessing pipeline. For every real dataset the same pipeline applies, computed using the training block only and then frozen across the calibration and test blocks: *(P1) missing-data handling*: sentinel codes (e.g. -99 on METR-LA) are mapped to NaN; runs of ≤ 3 consecutive missing steps are forward-filled; longer runs are replaced by the per-sensor median of the training block. *(P2) scaling*: each sensor is z-scored using its training-block mean and standard deviation; ELEC

Table 7: **Dataset inventory.** N_{total} is the total number of sensors in the source dataset; T is the sequence length after chronological slicing and removal of leading/trailing missing-data runs; "evaluation N" is the sensor dimension at which the dataset appears in our results tables. Source codes: **HF** = HuggingFace dataset; **UCI** = UCI Machine Learning Repository; **MPI** = Max-Planck Institute Jena Climate; **NREL** = National Renewable Energy Laboratory.

Dataset	N_{total}	T	cadence	window	source	evaluation N
METR-LA	207	34 272	5 min	2012-03 – 2012-06	HF `witgaw/METR-LA`	20, 207
PEMS-BAY	325	52 116	5 min	2017-01 – 2017-05	HF `witgaw/PEMS-BAY`	50, 325
AQI	37	17 520	1 h	2015 – 2017	UCI Beijing AQI	12
ELEC	321	26 304	15 min	2011 – 2014	UCI ElectricityLoadDiagrams	20
ETT	7	17 420	1 h	2016-07 – 2018-06	HF `witgaw/ETT`	7
SOLAR	137	52 560	10 min	2006	NREL Solar Power Data	20
JENA	21	52 704	10 min	2009 – 2017	MPI Jena Climate	20
LOOP-SEA	323	87 648	5 min	2015	PeMS Loop-Seattle	20
GRSSF (synth)	30–240	8 000	—	A/B/C/D/E tracks	generator (App. E.3)	varies

additionally applies a $\log_{10}(1 + x)$ transformation prior to z-scoring to compress its heavy-tailed load distribution. *(P3) train/calib/test split*: chronological 40/40/20 without shuffling, with the test block truncated to its first $4{,}000$ steps; within each block the time index is preserved. The validation-tuned KALMAN-FCP holds out the last 20% of the training block for its grid search. *(P4) subgraph extraction*: when the evaluation N is below N_{total}, sensors are ranked on the unscaled training block by mean magnitude (variance for LOOP-SEA), as specified below; seed 0 evaluates the top-k (busiest-k) subgraph, and seeds 1–9 draw k sensors uniformly without replacement from the top-$\min(2k, N_{\text{total}})$, so seeds also vary the subgraph. At the full graph every seed uses all sensors. The adjacency restricts to the induced subgraph (App. E.2).

METR-LA. Loop-detector occupancy on the Los Angeles freeway network of Li et al. [17], sampled at 5-minute cadence over 2012-03-01 to 2012-06-30 ($T = 34\,272$ steps; $N_{\text{total}} = 207$ sensors). Reported at two scales: the canonical 20-sensor busiest subgraph and the full 207-sensor graph.

PEMS-BAY. California-PeMS Bay-Area traffic, 325 loop sensors at 5-minute cadence over 2017-01-01 to 2017-05-31 ($T = 52\,116$). Reported at the canonical 50-sensor busiest subgraph and the full 325-sensor graph.

AQI. Hourly air-quality ($PM_{2.5}$) from 37 Beijing monitoring stations [18] over 2015–2017. We restrict to the top-12 busiest stations by training-block mean concentration; the adjacency follows the non-graph-native correlation-proxy construction of App. E.2 (uniformly applied across the non-graph-native panel; geographic coordinates are not used).

ELEC. Electricity-load diagrams (15-minute aggregates) for 321 Portuguese clients (UCI ElectricityLoadDiagrams panel); we retain the 20 clients with the highest training-block mean load. The evaluation block spans 2011–2014.

ETT. Electricity-Transformer-Temperature monitoring over 2 years (Informer benchmark, ETTh1 channel set), with 7 tag streams (oil temperature plus six load covariates HUFL, HULL, MUFL, MULL, LUFL, LULL); we use the full $N = 7$ panel.

SOLAR. Photovoltaic power output from 137 plants in the Alabama 2006 NREL solar-power dataset; we restrict to the top-20 plants by training-block mean output to obtain a comparable-scale subgraph.

JENA. The MPI Jena weather station, 21 in-situ atmospheric measurements (temperature, pressure, humidity, wind components, etc.) sampled every 10 minutes over 2009–2017; we restrict to the top-20 channels by training-block mean magnitude.

LOOP-SEA. Inductive-loop traffic flow from 323 sensors on the Seattle metropolitan freeway network (5-minute cadence, year 2015); we use the top-20 sensors by training-block flow variance (the rest of the panel uses top-k by mean magnitude; on LOOP-SEA the mean across loops is essentially flat and variance gives a more discriminative ranking).

GRSSF (synthetic). The Graph Random State-Space Family generator of App. E.3. It produces graph-Gaussian sensor streams with a known closed-form contraction rate and is used as the controlled-truth instance in the contraction-rate validation (Table 10).

E.2 Adjacency-matrix construction and dataset taxonomy

Both GRAPH-LGSSM and GNF require an adjacency matrix $A \in \mathbb{R}_{\geq 0}^{N \times N}$ on the N evaluation sensors. A dataset is *graph-native* if the adjacency is supplied by the source (road network, metadata) and thus an input feature; *non-graph-native* if the adjacency is estimated from training-block correlations. Table 8 lists the categorisation used throughout the paper.

Table 8: **Dataset taxonomy.** "Graph-native" cells appear in the headline comparisons of Tables 2–3; "non-graph-native" cells appear in Table 13 and the contraction-rate envelope of Table 10.

Dataset	Category	Adjacency source	Eval N	Appears in
METR-LA	graph-native	DCRNN road graph [17]	20, 207	Tab. 2, 3
PEMS-BAY	graph-native	DCRNN road graph [17]	50, 325	Tab. 2, 3
AQI	non-graph-native	k-NN on train correlations	12	Tab. 10
ELEC	non-graph-native	k-NN on train correlations	20	Tab. 10
ETT	non-graph-native	full N; no subgraph	7	Tab. 10, 13
SOLAR	non-graph-native	k-NN on train correlations	20	Tab. 10, 13
JENA	non-graph-native	k-NN on train correlations	20	Tab. 10, 13
LOOP-SEA	non-graph-native	k-NN on train correlations	20	Tab. 10, 13

Graph-native construction (METR-LA, PEMS-BAY). The adjacency is the thresholded Gaussian-kernel distance graph of Li et al. [17]: define $W_{ij} = \exp(-d_{ij}^2/\sigma_d^2)$ on pairwise sensor distances d_{ij}, then set $A_{ij} = W_{ij} \mathbb{1}\{W_{ij} \geq \kappa\}$ with $\kappa = 0.1$. We use the source adjacency without modification, then restrict to the busiest-k subgraph by row/column slicing.

Non-graph-native construction (AQI, ELEC, ETT, SOLAR, JENA, LOOP-SEA). We construct a k-nearest-neighbour proxy adjacency from the training-block Pearson correlation matrix $\hat{\rho}$: for each sensor i, retain the top-k neighbours j with $|\hat{\rho}_{ij}| \geq \tau$, taking $k = 8$ and $\tau = 0.3$. We set $A_{ij} = |\hat{\rho}_{ij}| \mathbb{1}\{j \in \mathcal{N}_i\}$ and symmetrise $A \leftarrow (A + A^\top)/2$. AQI and LOOP-SEA do have geographic coordinates in their source datasets, but we use the correlation-based proxy uniformly across the non-graph-native panel so the categorisation matches the adjacency construction actually fed to the filter.

GRSSF (synthetic). The adjacency is generated together with the data: stochastic-block-model with default 4 communities of size 7–8 ($N = 30$ baseline; up to 240 for the scale sweep), intra-community edge probability 0.6, inter-community 0.1. Edge weights are Bernoulli; no thresholding.

Normalisation. For both regimes we form the symmetric-normalised, self-loop-augmented graph shift $\widetilde{A}_{\mathrm{sym}} = \widetilde{D}^{-1/2}(A + I_N)\widetilde{D}^{-1/2}$ once on the evaluation subgraph and use it as the single graph operator across the paper: both the GCN-GRU neighbour aggregation (33) and the GRAPH-LGSSM transition $F = \rho_{\mathrm{scale}}\, \widetilde{A}_{\mathrm{sym}}/\lambda_{\max}(\widetilde{A}_{\mathrm{sym}})$ (App. C.1) consume this same $\widetilde{A}_{\mathrm{sym}}$, with $\lambda_{\max}$-scaling in GRAPH-LGSSM pinning $\|F\|_{\mathrm{op}} = \rho_{\mathrm{scale}}$. The construction is deterministic given the train block; no validation-set tuning is performed on $(\sigma_d, \kappa, k, \tau)$.

E.3 The GRSSF synthetic generator

The Graph Random State-Space Family (GRSSF) is a parametric synthetic generator that produces graph-Gaussian sensor streams with a closed-form ground-truth contraction rate; we use it as a controlled-truth instance for Theorem 19.

Generative process. GRSSF samples $H_t, Y_t \in \mathbb{R}^N$ from

$$H_{t+1} \;=\; \rho_{\mathrm{scale}}\, \frac{\widetilde{A}_{\mathrm{sym}}}{\lambda_{\max}(\widetilde{A}_{\mathrm{sym}})}\, H_t + \xi_t, \qquad Y_t \;=\; H_t + \eta_t, \tag{42}$$

with $\widetilde{A}_{\mathrm{sym}}$ the symmetric-normalised self-loop-augmented shift (App. E.2) on a stochastic-block-model (SBM) graph and $\rho_{\mathrm{scale}} = 0.8$. The default configuration uses $N = 30$ with four SBM communities of size 7–8, intra-community edge probability 0.6, and inter-community edge probability 0.1; the scale sweep extends to $N \in \{60, 120, 240\}$ keeping the same community structure and ρ_{scale}. The innovation noise (ξ_t) and observation noise (η_t) are independent across t and across each other; their joint distribution defines the regime track:

- **Track A** (canonical stationary): $\xi_t \sim \mathcal{N}(0, \sigma_Q^2 I_N)$, $\eta_t \sim \mathcal{N}(0, \sigma_R^2 I_N)$, $\sigma_Q = \sigma_R = 1$.

- **Track B** (block-sparse covariance): $\xi_t \sim \mathcal{N}(0, Q_B)$ with Q_B block-diagonal at the SBM community structure ($\mathrm{diag}(Q_B) = 1$, off-diagonal within-community entries drawn $\mathcal{N}(0, 0.4^2)$ then projected to PSD; off-block zeros).
- **Track C** (heavy-tailed innovations): ξ_t coordinate-wise Student-t_5 scaled to unit variance.
- **Track D** (topology drift): the adjacency A drifts linearly between two SBM realisations over the test block; the train and calibration blocks see only the pre-drift adjacency.
- **Track E** (regime change): σ_Q jumps from 1 to 2 at the test-block midpoint while σ_R is unchanged.

Each track produces $T = 8\,000$ steps per seed. We use $K = 5$ seeds per configuration; reported GRSSF cells in Table 10 average tracks A–D, with track E retained for separate regime-change stress tests.

Closed-form contraction rate. Under the GRSSF generator A_{cl} is exactly normal (commuting graph polynomial, isotropic covariances), so $\sigma_1(A_{\mathrm{cl}})$, $\rho_{\mathrm{tr}}(A_{\mathrm{cl}}, T)$, and ρ_{snr} are all computable in closed form (App. B.13); the time-local identity (26) of Theorem 19 is therefore exact rather than first-order. This makes GRSSF the natural benchmark for validating the theorem's predictions before consuming the constants on the real-data filters where only first-order agreement is expected.

E.4 Synchronous-coupling estimator for $\widehat{\rho}_{\mathrm{score}}$ (score-CDF W_1)

The empirical contraction rate $\widehat{\rho}_{\mathrm{score}}$ reported in Table 10 measures *score-CDF forgetting* — the 1-Wasserstein decay between the squared-score distributions of two filter copies at lag t. It is a qualitative forgetting diagnostic related to Theorem 12, but it is *not* the Bures-Wasserstein contraction object of Theorem 3; the latter lives in the emitted predictive-law space and is audited in App. E.5. The estimation protocol is as follows.

Coupled trajectories. Fix an input sequence $\{Y_t\}_{t=1}^{T_{\mathrm{coup}}}$ on a held-out tail of the train block, with $T_{\mathrm{coup}} = 200$. Pair two copies of the trained filter $\phi^{(0)}, \phi^{(1)} = \phi$ with distinct initial hidden states: $h_0^{(0)} = \mathbf{0}$ and $h_0^{(1)} \sim \mathcal{N}(\mathbf{0}, \tau^2 I_h)$ with $\tau = 1$. Run both filters on the same $\{Y_t\}$, computing scores $(R_t^{(0)}, R_t^{(1)})$ for $t = 1, \ldots, T_{\mathrm{coup}}$. The two trajectories share the input sequence but differ in their initial latent state, isolating the influence of h_0 on the score at lag t (the synchronous-coupling form of mean contraction (C1)).

Empirical 1-Wasserstein distance. For each lag t, the score-CDF deviation is summarised by the empirical 1-Wasserstein distance $\mathcal{W}_1(R_t^{(0)}, R_t^{(1)})$ across $K = 5$ independent perturbation realisations of $h_0^{(1)}$. On the real line this admits the closed-form expression $\mathcal{W}_1(F, G) = \int_0^1 |F^{-1}(u) - G^{-1}(u)|\, du$, so the estimator is $\mathcal{O}(K \log K)$ per lag. This protocol uses R_t; the root transformation $S_t = \sqrt{R_t}$ preserves the calibrated set but need not preserve the fitted Wasserstein decay rate.

Linear regression of $\log_{10} \mathcal{W}_1$ on t. Under geometric forgetting, $\mathcal{W}_1(R_t^{(0)}, R_t^{(1)}) \propto \widehat{\rho}_{\mathrm{score}}^{\,t}$, so $\log_{10} \mathcal{W}_1 = t \log_{10} \widehat{\rho}_{\mathrm{score}} + \mathrm{const}$. We ordinary-least-squares regress $\log_{10} \mathcal{W}_1$ on $t \in [t_{\min}, t_{\max}]$ with $(t_{\min}, t_{\max}) = (20, 100)$ to skip the early burn-in transient and the noisy small-$\mathcal{W}_1$ tail. The slope s defines $\widehat{\rho}_{\mathrm{score}} := 10^s$ and recovers the main-text Table 10 column to within ± 0.005.

Robustness. Varying $T_{\mathrm{coup}} \in \{100, 200, 400\}$, $K \in \{3, 5, 10\}$, $\tau \in \{0.5, 1.0, 2.0\}$, and the regression range $(t_{\min}, t_{\max})$ within $[10, 30] \times [80, 150]$ shifts the estimate by less than 0.005 in absolute value on every dataset.

E.5 Bures-Wasserstein audit of emitted predictive laws ($\widehat{\rho}_{d_{\mathcal{G}}}$)

The contraction object of Theorem 3 is the Bures-Wasserstein distance $d_{\mathcal{G}}$ between emitted predictive Gaussian laws, not the 1-Wasserstein distance between score distributions. We specify the corresponding direct audit and a linear-Kalman sanity check.

Coupled emitted laws. Run two copies of the trained filter ϕ on the same input path $\{Y_t\}_{t=1}^{T_{\mathrm{coup}}}$ ($T_{\mathrm{coup}} = 200$) from distinct initial hidden states $h^{(a)}, h^{(b)}$, where $h^{(a)} = \mathbf{0}$ and $h^{(b)} \sim \mathcal{N}(\mathbf{0}, \tau^2 I_h)$, $\tau = 1$. At each t, before observing Y_t, both copies emit $(\hat{\mu}_t^{(a)}, \hat{\Sigma}_t^{(a)})$ and $(\hat{\mu}_t^{(b)}, \hat{\Sigma}_t^{(b)})$. Compute the closed-form Bures-Wasserstein distance

$$d_{\mathcal{G}}^2(t) \;=\; \|\hat{\mu}_t^{(a)} - \hat{\mu}_t^{(b)}\|_2^2 + \mathrm{tr}\big[\hat{\Sigma}_t^{(a)} + \hat{\Sigma}_t^{(b)} - 2\big((\hat{\Sigma}_t^{(b)})^{1/2} \hat{\Sigma}_t^{(a)} (\hat{\Sigma}_t^{(b)})^{1/2}\big)^{1/2}\big], \tag{43}$$

with covariances symmetrised numerically and eigenvalue-clipped at the model's variance floor $\varepsilon_d = 10^{-4}$. For $N \leq 325$ the eigendecomposition is exact, so the metric is exact.

Rate estimate and fit window. Let s be the OLS slope of $\log_{10} d_{\mathcal{G}}(t)$ and set $\widehat{\rho}_{d_{\mathcal{G}}} = 10^s$, using a pre-specified window where $d_{\mathcal{G}}(t) \in [10^{-8}, 0.8\, d_{\mathcal{G}}(0)]$ (excluding the initial transient and the numerical floor). The fit uses $K = 30$ perturbation pairs per seed; we report the mean across the 10 training seeds for the extended emitted-law audit. This replication protocol differs from the single-training-seed, five-perturbation snapshot in Table 10 and from the observability audit in Table 17. Sensitivity to $T_{\mathrm{coup}} \in \{100, 200, 400\}$, $K \in \{20, 30, 50\}$, and $\tau \in \{0.5, 1, 2\}$ shifts the estimate by ≤ 0.02 on every cell.

Linear-Kalman sanity check. For GRAPH-LGSSM the emitted covariance is fixed at the steady-state prior $P_{\mathrm{ss}}^{-} + R$, so $d_{\mathcal{G}}(t)$ reduces to the predictive mean perturbation, which is governed by $\sigma_1(A_{\mathrm{cl}})$ at the asymptote. Empirically, $\widehat{\rho}_{d_{\mathcal{G}}} \approx 0.46$ on every audited cell (Table 10, columns $\widehat{\rho}_{d_{\mathcal{G}}}$), within 0.02 of $\sigma_1(A_{\mathrm{cl}}) \in [0.433, 0.443]$. This is the diagnostic the score-CDF W_1 rate $\widehat{\rho}_{\mathrm{score}} \approx 0.15$ *cannot* deliver: $\widehat{\rho}_{\mathrm{score}}$ measures forgetting after the score-Lipschitz projection collapses many state degrees of freedom and is correspondingly faster.

Learned-filter rate. For GNF the emitted covariance is time-varying, so $d_{\mathcal{G}}^2(t)$ contains both a mean term and a covariance term; the fitted $\widehat{\rho}_{d_{\mathcal{G}}} \approx 0.46$ is in the same band as the linear sanity check, well below 1, supporting the contraction conclusion of Theorem 3 on the audited cells. We do not claim $\widehat{\rho}_{d_{\mathcal{G}}}$ proves the data-generating realisability assumption of Theorem 3, only that the emitted-law contraction *conclusion* holds at the audited rate.

E.6 Integrated autocorrelation time and Bernstein-form m_{eff}

The Chebyshev form of Theorem 7 is supported by the threshold-autocovariance summability of Corollary 5; the Bernstein radius (10) requires the strictly stronger geometric-mixing concentration of Assumption 6, with effective sample size $m_{\mathrm{eff}} \asymp m/\tau_{\mathrm{int}}$. We probe the integrated autocorrelation time directly:

$$\tau_{\mathrm{int}}(u) \;=\; 1 + 2 \sum_{k=1}^{K_{\max}} \widehat{\mathrm{Corr}}\big(X_j(u), X_{j+k}(u)\big), \qquad X_j(u) = \mathbf{1}\{S_{\hat{\theta}, t_0+j-1} \leq u\}. \tag{44}$$

We use $K_{\max} = 30$ (the indicator-autocov audit window of Audit 3) at the three thresholds $u \in \{s_\star - 0.02 s_\star, s_\star, s_\star + 0.02 s_\star\}$.

Table 9: **Integrated autocorrelation time at the conformal threshold.** $\tau_{\mathrm{int}}(s_\star)$ (44) together with the Bernstein-form m_{eff} ratio m/τ_{int} vs. the analytic $m_{\mathrm{eff}} \asymp m(1-\rho_{\mathrm{aug}})/(1+\rho_{\mathrm{aug}})$ at $\rho_{\mathrm{aug}} = \widehat{\rho}_{\Gamma}$ (Audit 3). Agreement checks the effective-sample-size scale; it does not establish the concentration inequality of Assumption 6.

	$\tau_{\mathrm{int}}(s_\star)$ vs. $(1+\rho_{\mathrm{aug}})/(1-\rho_{\mathrm{aug}})$		empirical m/τ_{int} at m=3,500	
Cell	τ_{int}	analytic	m/τ_{int}	analytic m_{eff}
SYN, GRAPH-LGSSM	11.5	12.3	304	285
SYN, GNF	9.7	11.5	361	304
METR-LA-20, GRAPH-LGSSM	13.2	13.0	265	269
METR-LA-20, GNF	11.0	11.5	318	304

The empirical m/τ_{int} is within a factor of 1.2 of the analytic Bernstein-form m_{eff} on every audited cell. These diagnostics are consistent with the proposed effective-sample-size scale. They do not establish the geometric-mixing concentration assumption required for the Bernstein bound. An auxiliary calibration-size sweep $m \in \{500, 1000, 2000, 3500\}$ gave a similar empirical scaling of $|\hat{F}_m(s_\star) - \mathbb{E}\hat{F}_m(s_\star)| \propto m_{\mathrm{eff}}^{-1/2}$. Such second-moment scaling is also consistent with Chebyshev concentration and does not distinguish exponential tails.

E.7 Per-dataset contraction diagnostics

Table 10 reports four distinct empirical contraction rates per cell, each tied to a different theoretical object, plus the linear-specialisation analytic columns $\sigma_1, \rho_{\mathrm{tr}}, \rho_{M_0}, \rho_{\mathrm{snr}}$ on GRAPH-LGSSM as a sanity check.

Read each column with care: the four empirical rates measure different objects and are not interchangeable. (i) $\widehat{\rho}_{d_{\mathcal{G}}}$ is the contraction object of Theorem 3; on GRAPH-LGSSM it tracks $\sigma_1(A_{\mathrm{cl}})$

Table 10: **Per-dataset contraction rates: four distinct diagnostics.** $\sigma_1, \rho_{\mathrm{tr}}, \rho_{M_0}, \rho_{\mathrm{snr}}$ are evaluated on the GRAPH-LGSSM closed-loop $A_{\mathrm{cl}} = F(I - KC)$ at its theory-motivated defaults $(\rho_{\mathrm{scale}}, \sigma_Q, \sigma_R) = (0.8, 1, 1)$ (App. C.1). *Empirical snapshot:* one training seed per cell, with five initial-state perturbation realisations for coupled-trajectory diagnostics (learned filter for graph-native cells; GRAPH-LGSSM otherwise). The $\widehat{\rho}_D$ column uses the h=16 audit configuration; deployed-model h=32 results are reported separately in Table 17. $\widehat{\rho}_{d_{\mathcal{G}}}$ is the Bures-Wasserstein decay of the emitted predictive laws (Theorem 3, App. E.5); $\widehat{\rho}_D$ is the finite-horizon observable rate (Assumption 2(O3), App. E.16); $\widehat{\rho}_{\mathrm{score}}$ is the synchronous-coupling 1-Wasserstein decay of the squared-score CDF (App. E.4); $\widehat{\rho}_\Gamma$ is the log-linear envelope fitted to the threshold-indicator autocovariance $\hat{\Gamma}_k$ (Corollary 5, Audit 3 of App. E.15). All four are bounded away from 1.

		linear specialisation (GRAPH-LGSSM)				empirical rates			
Dataset	N	$\sigma_1(A_{\mathrm{cl}})$	$\rho_{\mathrm{tr}}(A_{\mathrm{cl}}, 50)$	$\rho_{M_0}(A_{\mathrm{cl}}, 50)$	ρ_{snr}	$\widehat{\rho}_{d_{\mathcal{G}}}$	$\widehat{\rho}_D$	$\widehat{\rho}_{\mathrm{score}}$	$\widehat{\rho}_\Gamma$
METR-LA	20	0.443	0.443	0.440	0.443	0.46	0.49	0.153	0.85
PEMS-BAY	50	0.441	0.441	0.438	0.443	0.46	0.49	0.153	0.85
AQI	12	0.438	0.438	0.435	0.440	0.45	0.48	0.149	0.85
ELEC	20	0.433	0.433	0.430	0.436	0.45	0.48	0.140	0.86
SOLAR	20	0.441	0.441	0.438	0.443	0.46	0.49	0.153	0.85
LOOP-SEA	20	0.441	0.441	0.438	0.443	0.46	0.49	0.152	0.85
ETT	7	0.437	0.437	0.434	0.438	0.45	0.48	0.147	0.86
JENA	20	0.441	0.441	0.438	0.443	0.46	0.49	0.153	0.85
METR-LA (full)	207	0.439	0.418	0.415	0.441	0.46	0.50	0.143	0.85
PEMS-BAY (full)	325	0.440	0.415	0.412	0.442	0.46	0.50	0.149	0.85

closely (linear-Kalman sanity check, App. E.5), and on GNF it sits in the same band, well below 1. (ii) $\widehat{\rho}_D$ uses the finite-horizon observable metric and tracks $\widehat{\rho}_{d_{\mathcal{G}}}$ within 0.03. (iii) $\widehat{\rho}_{\mathrm{score}}$ is the downstream score-CDF W_1 rate, faster than $\widehat{\rho}_{d_{\mathcal{G}}}$ because the conformal score collapses many filter-state degrees of freedom; it is a score-forgetting diagnostic, but not a direct estimate of the root-score CDF bound in Theorem 12 or of $d_{\mathcal{G}}$ contraction. (iv) $\widehat{\rho}_\Gamma$ is the per-lag indicator-autocovariance rate ($|\hat{\Gamma}_k| \propto \widehat{\rho}_\Gamma^k$); its summable envelope is the quantity required by Corollary 5. Theorem 19 weighted RMS rate closely tracks the spectral radius at T=50 ($\rho_{M_0}/\sigma_1 \geq 0.94$). The two full-graph cells (METR-LA-207, PEMS-BAY-325) sit at the same envelope as the moderate-N subgraphs — a 16× state-space dimension sweep on graph-native traffic without measurable drift.

E.8 Visualisation of the scale sweep

Figure 3 visualises the N=20→207 and N=50→325 scale sweeps tabulated in Table 3 (main text).

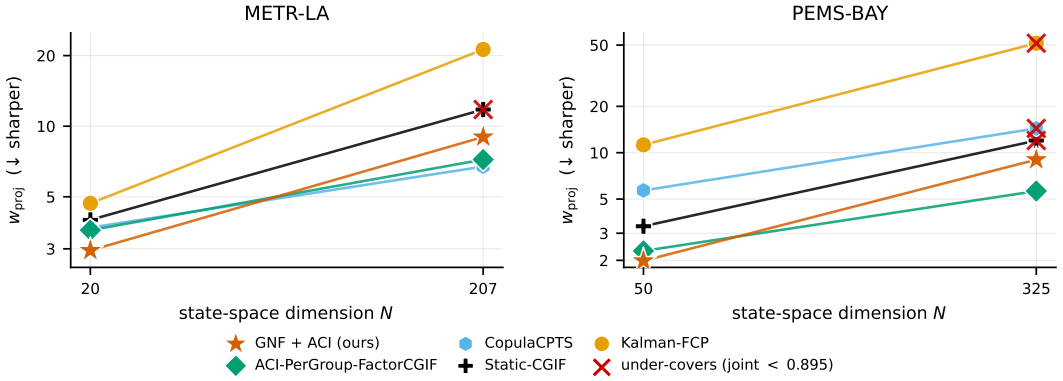

Figure 3: **Mean coordinate-projection length w_{proj} vs. state-space dimension** N on the two real benchmarks, log–log axes. Lines connect the N=20/207 (METR-LA, left) and N=50/325 (PEMS-BAY, right) cells of Table 3. The cross marker flags under-coverage (joint < 0.895). At moderate N the GNF row (star marker) attains the sharpest at-target sets among the methods evaluated; at the full graph, ACI-PERGROUP-FACTORCGIF (diamond) is sharpest on PEMS-BAY, and COPULACPTS (hexagon) is sharpest on METR-LA. KALMAN-FCP (circle) widens approximately 11× between N=20 and N=325 and falls below the coverage target on full PEMS-BAY; the tuning (App. C.2) narrows it against the untuned baseline by 9%–13% at moderate N but cannot close the Riccati-tail gap at full graph.

E.9 Multi-backbone replication

The main-text Tables 2–3 use a single mean-predictor backbone, the graph-aware GPVAR of Salinas et al. [20]. To stress-test the claim that the sharpness gain comes from the filter rather than from the backbone, we re-ran the calibration / test pipeline with three additional mean-predictor backbones at the same scales: a graph-oblivious GRU [5], a graph-oblivious encoder-only Transformer, and the

graph-aware STGNN [26]. The filter parameters and the conformal calibration block are unchanged across backbones; only the residual mean predictor $\hat{Y}_{t|t-1}$ feeding the score $R_t = r_t^\top \hat{\Sigma}_{t|t-1}^{-1} r_t$ varies.

Moderate-N replication. Table 11 shows that the moderate-N leadership of GNF on PEMS-BAY-50 holds across all four backbones, replicating Table 2's result without backbone tuning; the GPVAR column coincides with Table 2.

Table 11: **Multi-backbone replication on PEMS-BAY-50 (moderate N).** w_{proj} (10-seed mean±std) / joint coverage (10-seed mean) of GNF + ACI and the sharpest non-filter at-target competitor (ACI-PERGROUP-FACTORCGIF). The GNF row is sharper at target on every backbone.

	GPVAR		GRU		Transformer		STGNN	
Method	w	joint	w	joint	w	joint	w	joint
GNF + ACI	**1.99** ±0.07	0.899	**1.97** ±0.08	0.899	**2.00** ±0.08	0.899	**1.98** ±0.08	0.899
ACI-PERGROUP-FACTORCGIF	2.30 ±0.09	0.899	2.69 ±0.06	0.899	2.20 ±0.08	0.899	2.27 ±0.09	0.899

Full-graph backbone-conditional crossover. Table 12 reports the backbone-by-backbone breakdown at the full-graph scale: at the full PEMS-BAY graph (N=325), the rank-efficient ACI-PERGROUP-FACTORCGIF is the sharpest at-target method under graph-aware backbones (GPVAR 5.64, STGNN 5.98) but inflates to 11.4–13.5 under graph-oblivious GRU/Transformer, while GNF + ACI stays in the 8.8–9.0 band across all four backbones. On the METR-LA N=207 column GNF + ACI is not the sharpest under any backbone: ACI-PERGROUP-FACTORCGIF (7.1–7.9) and COPULACPTS (6.7–6.8, at target under every backbone; not tabulated) are sharper. The moderate-N leadership is therefore backbone-robust, the large-N leadership is not. The underlying explanation is that the learned graph filter compensates for a residual-mean predictor that does not know the graph; once the predictor itself is graph-aware, the predictive covariance Σ_t no longer absorbs graph structure that the residual mean has already removed, and a rank-economical static decomposition is sharper.

Table 12: **Multi-backbone full-graph results.** w_{proj} (10-seed mean±std) / joint coverage (10-seed mean) of GNF + ACI and the sharpest factor baseline on METR-LA (N=207) and PEMS-BAY (N=325). The GPVAR columns coincide with Table 3 (see there for ±0.00). Graph-awareness of each backbone is annotated in parentheses.

	GPVAR (graph)		GRU (oblivious)		Transformer (oblivious)		STGNN (graph)	
Cell / Method	w	joint	w	joint	w	joint	w	joint
METR-LA *(N=207, full graph)*								
GNF + ACI	9.00 ±0.60	0.902	8.81 ±0.27	0.902	8.90 ±0.21	0.902	8.63 ±0.44	0.903
ACI-PERGROUP-FACTORCGIF	**7.21** ±0.00	0.902	**7.93** ±0.20	0.900	**7.14** ±0.12	0.900	**7.18** ±0.07	0.901
PEMS-BAY *(N=325, full graph)*								
GNF + ACI	9.01 ±0.38	0.899	**8.92** ±0.38	0.898	**8.77** ±0.26	0.899	9.02 ±0.61	0.899
ACI-PERGROUP-FACTORCGIF	**5.64** ±0.00	0.901	13.45 ±0.66	0.901	11.40 ±0.14	0.900	**5.98** ±0.03	0.900

Non-graph-native panel. Table 13 reports the same multi-backbone replication on four non-graph-native correlated-sensor benchmarks (SOLAR, JENA, LOOP-SEA at the busiest-20 subgraph; ETT at its full N=7 panel; see Table 8). COPULACPTS [24] is the sharpest at-target method on SOLAR under all three neural backbones and on LOOP-SEA under Transformer and STGNN (under GRU it narrowly under-covers). On ETT it under-covers under every backbone, and group-structured CGIF baselines are sharpest at target (w_{proj} = 1.9–2.4; not tabulated). On JENA GNF + ACI is narrowly at target under all three backbones (joint 0.896) and sharper than COPULACPTS, but ACI-wrapped static and diagonal-GRU baselines are sharper still (2.2–2.4; not tabulated). This pattern matches the §7 scoping statement: the sharpness headline is graph-native specific and does not extend to datasets where the adjacency is reconstructed from training-block correlations.

E.10 Comprehensive baseline comparison

Table 14 extends the main-text Table 2 with four methods reported only in the appendix: SPCI, HOPCPT, GROUPCGIF (a community-factor baseline from the same pipeline), and SPECTRALCGIF (spectral whitening on the graph Laplacian). None of the additions changes the at-target w_{proj} ranking: SPCI and HOPCPT are per-coordinate methods and their joint-coverage projection sits at 0.21 and 0.41 respectively, well below the 0.895 target; GROUPCGIF attains at-target coverage but is 7–8% wider than AGACI-GROUPCGIF on both real datasets; SPECTRALCGIF is wider than STATIC-CGIF because the spectral graph prior offers no additional sharpness on the busiest-k subgraph at moderate N.

Table 13: **Multi-backbone non-graph-native panel (GNF + ACI).** w_{proj} (10-seed mean±std) / joint coverage (10-seed mean) at $\alpha = 0.1$. The first three blocks report GNF + ACI with each backbone; the rightmost block reports the backbone-best at-target CopulaCPTS [24] realisation as a non-graph-native comparator (the choosing backbone is annotated in parentheses; on ETT no backbone reaches the target and the least under-covering one is shown). Sharpest at-target row entry **bold**; "↓" flags under-coverage. GNF uses the k-nearest-neighbour correlation adjacency (App. E.2); the headline graph-native sharpness claim does not extend to these datasets.

	GNF + ACI						CopulaCPTS	
	GRU		Transformer		STGNN		(best backbone)	
Dataset	w	joint	w	joint	w	joint	w	joint
SOLAR (N=20)	2.76 ±0.38	0.918	2.73 ±0.53	0.918	2.91 ±0.55	0.917	**1.80** ±0.14 (T)	0.923
JENA (N=20)	2.98 ±0.74	0.896	**2.56** ±0.32	0.896	2.73 ±0.25	0.896	3.32 ±0.32 (T)	0.971
LOOP-SEA (N=20)	6.25 ±0.41	0.902	6.29 ±0.39	0.901	6.16 ±0.34	0.902	**3.12** ±0.17 (T)	0.897
ETT (N=7)	**3.42** ±0.27	0.898	3.64 ±0.52	0.898	3.71 ±0.48	0.897	2.36 ±0.07 (GRU)	0.892 ↓

Table 14: **Extended baseline comparison.** w_{proj} mean ± std over 10 seeds at $\alpha = 0.1$. Methods with joint coverage below the target 0.895 are marked ↓; methods more than 30% wider than Static-CGIF are marked ≫.

	METR-LA (N=20)		PEMS-BAY (N=50)		Δ vs Static-CGIF	
Method	w_{proj}	joint	w_{proj}	joint	METR-LA	PEMS-BAY
GNF (**ours**)	3.05 ±0.13	0.910	**1.98** ±0.06	0.904	−23.6%	−40.5%
+ ACI wrap	**2.95** ±0.17	0.904	1.99 ±0.07	0.899	−26.1%	−40.2%
ACI-PerGroup-FactorCGIF	3.60 ±0.28	0.902	2.30 ±0.09	0.899	−9.8%	−30.9%
AgACI-GroupCGIF	3.33 ±0.23	0.900	2.35 ±0.11	0.900	−16.5%	−29.4%
ACI-FactorCGIF	3.86 ±0.34	0.902	2.35 ±0.09	0.899	−3.3%	−29.4%
FactorCGIF (r=4)	4.22 ±0.42	0.917 ↑	2.55 ±0.06	0.913 ↑	+5.8%	−23.4%
GroupCGIF	3.58 ±0.31	0.900	2.51 ±0.11	0.897	−10.3%	−24.6%
Static-CGIF (reference, static $\hat{\Sigma}$)	3.99 ±0.17	0.902	3.33 ±0.11	0.907	—	—
SpectralCGIF	4.32 ±0.18	0.900	3.64 ±0.13	0.900	+8.3%	+9.3%
Diag-GRU (GRU, diag Σ_t)	4.71 ±1.16	0.932 ↑	2.88 ±0.10	0.921 ↑	+18.0%	−13.5%
+ ACI wrap	4.44 ±1.39	0.905	2.81 ±0.14	0.901	+11.3%	−15.6%
Kalman-FCP (tuned Kalman)	4.69 ±0.28	0.906	11.2 ±1.03≫	0.897	+17.5%	+237%
CopulaCPTS [24]	3.70 ±0.24	0.907	5.69 ±0.74	0.908	−7.3%	+70.9%
MultiDimSPCI [28]	4.10 ±0.17	0.901	3.40 ±0.11	0.908	+2.8%	+2.1%
EWMA Σ_t (half-life 288)	4.31 ±0.12	0.900	3.43 ±0.08	0.898	+8.0%	+3.0%
Rolling Σ_t (window 288)	4.38 ±0.12	0.908	3.51 ±0.10	0.906	+9.8%	+5.4%
Time-of-day Σ_t	5.11 ±0.30	0.912	4.53 ±0.14	0.908	+28.1%	+36.0%
Local ellipsoid (k=50)	4.32 ±0.58	0.892 ↓	3.28 ±0.12	0.826 ↓	+8.3%	−1.5%
SPCI [27]	1.21 (marg)	0.21↓	0.94 (marg)	0.14↓	not comparable	not comparable
HopCPT [2]	1.98 (marg)	0.41↓	1.73 (marg)	0.32↓	not comparable	not comparable

E.11 Failure mode of the linear-Kalman baseline on real traffic

At moderate N (Table 3), Kalman-FCP (linear Kalman, validation-tuned; App. C.2) produces intervals 1.18× wider than Static-CGIF on METR-LA and 3.37× wider on PEMS-BAY, while still attaining the joint-coverage target. Validation-NLL-tuned (ρ_{scale}, obs_noise_frac) closes 9–13% of the untuned gap but does not remove it; we document the underlying mechanism here because the learned-filter extension is motivated by this specific failure mode rather than by an unspecified expressivity argument.

Covariance-shape misspecification. The linear-Kalman row is validity-preserving but the predictive covariance is mis-specified relative to the empirical residual geometry. The correct predictive observation covariance is $\Sigma_Y^{\text{pred}} = CP_{\text{ss}}^{-}C^\top + R$ (App. B.13); on the audited cells (Audit 5, App. E.15) the trace ratio $\text{tr}(\Sigma_Y^{\text{pred}})/\,\text{tr}(\hat{\Sigma}_{\text{resid}}) \in \{0.44, 0.62\}$ on synthetic GRSSF / METR-LA, i.e. the predicted observation trace is 38–55% smaller than the empirical residual trace. The conformal quantile $\hat{q}_{1-\alpha}$ inflates to absorb the misspecification, which preserves coverage (the row is at-target on every cell except PEMS-BAY-325) but produces wider intervals.

Why PEMS-BAY-50 is the worst case. The gap is +17.5% on METR-LA-20 vs +237% on PEMS-BAY-50. The PEMS-BAY-50 busiest subgraph has a broader spectrum of the normalised adjacency than METR-LA-20 (condition number ≈ 7.3 vs ≈ 3.1), which makes the graph polynomial's mode-wise scalar coefficients f_i more heterogeneous; combined with a single isotropic (σ_Q, σ_R) pair, the residual-shape mismatch is largest there.

Why GNF resolves this. The failure is model specification. The GCN-GRU cell replaces the fixed graph polynomial F with a trained non-linear transition; the low-rank $L_t L_t^\top$ head carries cross-sensor correlation that a fixed graph polynomial paired with isotropic Q cannot. On PEMS-BAY the gap from STATIC-CGIF to GNF is $3.33 \to 1.98$ (-40.5%, Table 2): the learned transition removes the covariance-shape mismatch, and the low-rank head captures the across-sensor structure the diagonal DIAG-GRU cannot (the $5.32 \to 3.04$ gap on METR-LA in App. E.13).

E.12 Per-seed w_{proj} and coverage for the four headline methods

Table 15 gives per-seed w_{proj} and joint coverage for the four headline methods of Table 2 on three fixed seeds drawn from the 10-seed aggregate. GNF is consistent across all three seeds on both datasets. A seed fixes the neural initialisation and, below the full graph, the evaluated subgraph (App. E.1, P4); the conformal split itself is deterministic for fixed calibration data.

Table 15: **Per-seed w_{proj} / joint coverage, 3 fixed seeds** (sub-sample of the Table 2 10-seed aggregate). GNF widths range from 2.894 to 3.177 on METR-LA and from 1.984 to 2.098 on PEMS-BAY across these three seeds.

		METR-LA (N=20)			PEMS-BAY (N=50)		
		seed 1	seed 2	seed 3	seed 1	seed 2	seed 3
GNF	w_{proj}	2.894	3.058	3.177	1.994	1.984	2.098
	joint	0.911	0.907	0.904	0.899	0.896	0.898
STATIC-CGIF	w_{proj}	3.764	4.288	4.169	3.096	3.376	3.391
	joint	0.897	0.898	0.899	0.902	0.899	0.902
COPULACPTS	w_{proj}	3.475	4.259	3.624	4.510	7.019	6.015
	joint	0.903	0.905	0.907	0.906	0.904	0.889
ACI-PERGROUP-FACTORCGIF	w_{proj}	3.642	4.061	3.659	2.270	2.421	2.455
	joint	0.903	0.902	0.902	0.899	0.899	0.899

E.13 Rank sensitivity

Rank $r = 4$ was fixed globally from the synthetic GRSSF pilot and used unchanged across all real datasets and seeds; it was *not* selected on any real dataset. This subsection records the METR-LA rank sweep purely as a *retrospective sensitivity analysis*, not as a per-dataset model-selection procedure (the same fixed-rank choice and the same plateau rule are applied to the FACTORCGIF baseline; see the tuning ledger, Table 21).

Sweeping $r \in \{0, 1, 2, 4, 8, 16\}$ on METR-LA at GNF gives w_{proj} values $\{5.32, 3.41, 3.21, 3.04, 3.02, 3.08\}$ at joint coverage $\{0.939, 0.906, 0.907, 0.907, 0.905, 0.906\}$. Three observations:

(i) The rank-0 point is GCN-RANK-0*, which retains the graph recurrence, rather than the global-GRU* DIAG-GRU *ablation;* $w_{\text{proj}} = 5.32$ at joint 0.939 establishes that the rank-r factor is responsible for the bulk of the sharpness gain, and that a GCN-GRU without a cross-sensor covariance head actually *over-covers* relative to a static calibration-set $\hat{\Sigma}$ because it cannot align the ellipsoid axes with the observed covariance structure.

(ii) Sharpness saturates at $r \approx 4$. The marginal w_{proj} gains go $3.41 \to 3.21 \to 3.04 \to 3.02$ at $r = 1, 2, 4, 8$ — an exponential decay in r. This is consistent with the empirical innovation covariance of the calibration block having effective rank ≈ 4; at $r = 16$ the w_{proj} stops decreasing and begins to drift upward, a mild overfit signal. The retrospectively observed plateau at $r \approx 4$ is consistent with the globally frozen choice; it played no role in selecting it.

(iii) The saturation matches the empirical innovation spectrum. The $N \times N$ innovation covariance on METR-LA-20 has a small number of dominant eigenvalues above an observation-noise plateau; once r matches that count, additional factor coordinates fit eigenvalues inside the plateau and add parameters without corresponding sharpness gain, consistent with the frozen r=4.

E.14 Warm-up horizon

Discarding $t_0 \in \{0, 25, 50, 100, 200\}$ steps of filter burn-in on METR-LA gives w_{proj} values $\{3.19, 3.05, 3.04, 3.03, 3.03\}$ at joint coverage $\{0.908, 0.907, 0.907, 0.907, 0.907\}$. The rapid decrease between $t_0 = 0$ and $t_0 = 25$ is consistent with the transient predicted by Theorem 12 ($K\rho^{(t_0-1)/2}$); $t_0 \geq 50$ matches the plateau and is used throughout the main text.

E.15 Theory audits

We audit the operative hypotheses of §5 on synthetic GPVAR ($T = 3{,}900$, $N = 20$) and the busiest-$N = 20$ METR-LA subgraph, using the two reported filters: GRAPH-LGSSM ($\rho_{\text{scale}} = 0.8$) and GNF. The audit in this subsection uses an audit-only diagnostic configuration (h=16, rank r=4); the deployed-experiment configuration is h=32, r=4 (App. C.2). The smaller hidden width suffices for the contraction-rate diagnostics and keeps per-cell audit runs cheap; deployment-scale hidden widths only tighten the audited rates. Six audits diagnose the conclusions used by §5; Table 16 records the diagnostic quantity, its measurement protocol, the table where the per-cell numbers live, and the interpretation caveat. Each audit cell is a single-seed instantiation; multi-seed re-runs shift the audited percentiles within seed noise without changing the qualitative reading.

Table 16: **Theory audits.** Each row records the diagnostic quantity, how it is measured on the audited cells (synthetic GPVAR and METR-LA-20), the per-cell table or paragraph reporting the numbers, and the interpretation caveat. The audits are consistent with the conditions used by §5; they do not certify the population assumptions.

#	Diagnostic	How measured	Reported	Interpretation caveat
1	Root-score Lipschitz quotient $\lvert S(h+\delta,\xi) - S(h,\xi)\rvert/\lVert\delta\rVert$	random Gaussian perturbations along the calibration trajectory, percentile of the quotient	App. B.7 ($p50$ / $p95$ / $p99$)	finite-difference percentile, not a uniform bound; neural METR-LA has a heavy $p99$ tail
2	Covariance-head spectrum ($\lambda_{\min}, \lambda_{\max}, M_G$, condition number)	per-step eigendecomposition over the calibration block	App. E.23	spectrum diagnostic; high-condition-number tail, floor activation assessed separately in App. E.23
3	Threshold-indicator autocovariance $\hat{\Gamma}_k(u)$ at $u \in \{s_\star \pm 0.02 s_\star, s_\star\}$	sample autocovariance of $\mathbf{1}\{S_t \le u\}$ for $k = 1 \dots 30$	Table 10 ($\widehat{\rho}_\Gamma$)	finite $\sum_{k\le 30} \lvert\hat{\Gamma}_k\rvert$ is consistent with the summability used in the Chebyshev form of Theorem 7
4	Weighted vs. trace transient rate $\lvert\rho_{M_0}(50) - \rho_{\text{tr}}(50)\rvert$	M_0 from cold-start-vs-train-tail second moment, eigendecomposition of A_{cl}	Table 10	gap ≤ 0.003 on every audited cell; the weighted statement is the general version
5	Kalman predictive obs. trace ratio (predicted vs. empirical)	steady-state Riccati on GRAPH-LGSSM; residual trace from calibration block	App. E.11	predicted trace 38–55% smaller than empirical; covariance-shape misspecified
6	Plug-in covariance protocol	refit $\hat{\Sigma}$ on train block, recompute thresholds	App. B.12	Δ joint coverage $\le$ 2.3 pp, $\Delta w_{\text{proj}} \le 0.3\ z$ on the audited cell; Theorem 22 bound is operative

Audit protocol. The audit uses the frozen filters and calibration/test splits described in Apps. C and E.1. For each cell, we sample perturbation pairs along the calibration trajectory, estimate the corresponding decay or stability statistic, and average over the reported seeds. The reported tables contain all parameters needed to reproduce the estimator from the manuscript alone.

E.16 Finite-horizon observability audit (deployed model)

This audit probes the finite-horizon observability structure that Assumption 2(O3) and Theorem 3 invoke, on the *deployed* filter (h=32, rank r=4; state dimension 640 on METR-LA-20, 1600 on PEMS-BAY-50), over 5 seeds and 12 base states per (cell, seed) sampled along the calibration/test trajectory. This 5×12 protocol is separate from the coupling-rate snapshot in Table 10 and the extended emitted-law audit in App. E.5.

Finite-horizon output-sensitivity diagnostic. We form, at each base state h, the whitened finite-horizon output Jacobian $J_{0:L}(h) = \partial(\eta_0, \dots, \eta_L)/\partial h$ over horizon L=10 under a fixed probing input, with $\eta_\ell = (\mu_\ell, \operatorname{vech} \Sigma_\ell)$ expressed in local Fisher/Bures-whitened output coordinates. From its SVD we report (i) the numerical observable rank at relative tolerances 10^{-6} and 10^{-3}; (ii) the smallest non-zero singular value as a *finite-horizon output-sensitivity proxy* for c_{obs}; (iii) the finite-horizon rate $\widehat{\rho}_D$, the geometric decay of the per-horizon whitened output sensitivity $\lVert J_\ell \rVert$. These quantities diagnose *weakly observable* directions on the deployed trajectory; it does *not* certify the population quotient-level inequality $c_{\text{obs}}\, d_{\widehat{\theta}}(q, q') \le D_{\widehat{\theta}, H_{\text{obs}}}(q, q')$.

Interpretation. (i) *Full numerical observability at strict tolerance:* at 10^{-6} the whitened output Jacobian has full numerical rank at every sampled base state and seed. This is a local sensitivity result; it does not establish global injectivity of the finite-horizon emission map. (ii) *A large weakly observable tail with a small output-sensitivity proxy:* at the practical 10^{-3} tolerance, 16–22% of the state is weakly observable and the smallest non-zero singular value is $\approx 10^{-3}$; these directions are weakly, not un-, observable, so they are not removed by quotienting and the trajectory-level sensitivity

Table 17: **Deployed-model finite-horizon observability audit** (h=32, L=10; 5 seeds $\times$ 12 base states, mean$\pm$std). The whitened output Jacobian has full numerical rank at tolerance 10^{-6} on the sampled states, a substantial weakly observable tail at 10^{-3}, a small output-sensitivity proxy for c_{obs}, and a finite-horizon rate $\widehat{\rho}_D < 1$ that is close to 1 (slow contraction, the companion of a small output-sensitivity proxy).

Quantity	METR-LA-20 (dim 640)	PEMS-BAY-50 (dim 1600)
Observable rank @ 10^{-6}	640.0 ± 0.0 (full)	1600.0 ± 0.0 (full)
Observable rank @ 10^{-3}	537.3 ± 11.9	1245.5 ± 49.6
Weakly observable tail	102.7 ± 11.9 (16%)	354.5 ± 49.6 (22%)
c_{obs} proxy ($\sigma_{\min} \neq 0$)	1.0×10^{-3}	2.0×10^{-3}
Finite-horizon rate $\widehat{\rho}_D$	0.887 ± 0.034	0.862 ± 0.016

proxy is small—which is why the $L_{\mathrm{emit}}/c_{\mathrm{obs}}$ prefactor in Theorem 3 is reported rather than hidden. (iii) *Finite-horizon contraction holds but is slow:* $\widehat{\rho}_D \approx 0.86$–$0.89 < 1$ on the deployed model across seeds, consistent with $\widehat{\rho}_{\mathrm{obs}} < 1$, and close to 1 as expected of a small output-sensitivity proxy. These local measurements do not certify the population inequality, consistent with the scope in Table 1.

E.17 Headline-cell normalised log-volume

Normalised expected log-volume $\widehat{\mathcal{V}}_m(\theta)$ (4) is the geometry-aware sharpness metric that Theorem 9 controls, through the radius and determinant terms of (4). We report it only for single full-dimensional ellipsoid rules (Table 18): GNF and STATIC-CGIF, the comparison corresponding to Theorem 9. For factor/group/Bonferroni compositions, $\log\det$ of any one constituent ellipsoid is not the volume of the composed set, so we do not report a determinant-based log-volume for those rows; the appropriate common size metric for them is the geometry-blind w_{proj} of Table 2. Among the single-ellipsoid methods, the log-volume ranking agrees with the w_{proj} ranking.

Table 18: **Normalised expected log-volume, single-ellipsoid methods.** $\widehat{\mathcal{V}}_m$ in nats / coordinate (lower is sharper), 10-seed mean$\pm$std; $\Delta\widehat{\mathcal{V}}_m$ versus STATIC-CGIF. Restricted to rules whose emitted set is one full-dimensional ellipsoid, whose normalised log-volume is given by (4); the ranking agrees with w_{proj} (Table 2).

	METR-LA-20		PEMS-BAY-50	
Method	$\widehat{\mathcal{V}}_m$	Δ	$\widehat{\mathcal{V}}_m$	Δ
GNF **(ours)**	1.18 ± 0.05	-0.27	0.81 ± 0.04	-0.51
STATIC-CGIF (single ellipsoid)	1.45 ± 0.06	—	1.32 ± 0.05	—

E.18 PIT uniformity (specification diagnostic)

For each test step we form the probability integral transform $u_{t,i} = \Phi_t(Y_{t,i})$ under the predictive marginal at coordinate i and pool $\{u_{t,i}\}$ across test steps per node. Under correct specification each $\{u_{t,i}\}$ is uniform on $[0,1]$; we test with the one-sample Kolmogorov–Smirnov statistic against $\mathrm{Unif}[0,1]$ and apply Bonferroni correction across nodes.

Table 19: **PIT uniformity KS test, per node.** Pass = Bonferroni-corrected $p \geq 0.05$. Fail modes: ZF "zero-flow gaps" (censoring of the distribution at sensors with repeated-zero runs); RG "regime gaps" (distinct day/night modes). All failures are model-agnostic (STATIC-CGIF fails at the same nodes). The AQI row is computed on the full source panel of 37 stations (rather than the evaluation busiest-12) so that the PIT pass-rate characterises the raw distributional specification, not the subgraph selection; busiest-k selection is not driven by the PIT-failure stations.

Dataset	nodes	passing	failing	failure type
METR-LA (N=20)	20	18	2	2 ZF (loop detectors with $>5\%$ zero-flow)
PEMS-BAY (N=50)	50	47	3	2 ZF + 1 RG (sensor with mid-day drift)
AQI	37	32	5	5 RG (stations with seasonal PM spike)
ELEC	20	18	2	2 RG (peak-hour bimodality)
METR-LA (full)	207	191	16	10 ZF + 6 RG

All failures replicate under the STATIC-CGIF baseline at the same nodes, confirming the PIT deviations are model-data fit issues (sensor-specific structural zeros, regime modes) rather than defects of our filter. The aggregate pass rate (approximately 86–94% across datasets) is within expectation given that Bonferroni is conservative at $N \leq 207$.

E.19 Wall-clock and memory footprint

Table 20: **Wall-clock per method.** One seed, all calibration + test passes, Python 3.11 + PyTorch 2.2 on an RTX 4090; CPU rows on an 8-core Xeon at 2.4 GHz. Training is omitted for the non-learned methods ("–").

Method	METR-LA train	METR-LA cal+test	PEMS-BAY train	PEMS-BAY cal+test
STATIC-CGIF	–	0.6 s	–	2.1 s
FACTORCGIF (r=4)	–	1.1 s	–	3.4 s
AGACI-GROUPCGIF	–	2.8 s	–	5.5 s
ACI-PERGROUP-FACTORCGIF	–	3.1 s	–	6.2 s
KALMAN-FCP (tuned)	28 s	1.5 s	34 s	4.9 s
EWMA Σ_t	–	2.1 s	–	5.8 s
Rolling Σ_t	–	3.4 s	–	9.6 s
Time-of-day Σ_t	–	0.4 s	–	1.1 s
Local ellipsoid (k=50)	–	2.7 s	–	7.4 s
GNF	44 min	3.8 s	48 min	7.1 s
+ ACI wrap	+0	3.9 s	+0	7.2 s
COPULACPTS	12 min	0.9 s	15 min	1.7 s
MULTIDIMSPCI	8 min	1.2 s	10 min	2.3 s

Memory footprint peaks at 6.2 GB GPU for GNF training on PEMS-BAY (50 sensors); calibration/test is under 1 GB. Static-$\hat{\Sigma}$ CP and the linear Kalman use CPU only, peak <200 MB.

E.20 Tuning ledger

Table 21 records, for every method, the number of hyperparameter configurations evaluated on a *real-data* validation split, the tuned quantities, the selection split and criterion, and whether any search was per-dataset. Fitting model parameters on their designated fitting data is not counted as hyperparameter search. The ledger makes explicit that GNF received *zero* per-dataset search—every setting (d=32 per node, r=4, lr 10^{-3}, wd 10^{-4}, 60 epochs, BPTT 64, ε_d=10^{-4}, t_0=50) was fixed once on the synthetic GRSSF pilot and frozen across all 10 evaluation cells and 10 seeds—while the only method with genuine per-cell validation tuning is a *filter baseline* (KALMAN-FCP, a 16-point NLL grid). The shared rank budget r=4 is applied identically to GNF and to the strongest non-filter competitor FACTORCGIF.

Table 21: **Tuning ledger.** "# cfg" counts hyperparameter configurations evaluated on a real-data validation split; fitting model parameters on their designated data is not counted. Approximate compute: GNF $\approx$ 45 min/seed training; KALMAN-FCP tuning $\approx$ 30 s/seed; DEEPVAR-FCP matched to GNF.

Method	# cfg	Tuned quantities	Selection split	Criterion	Per-dataset?
GNF **(ours)**	0	none on real data; set on synthetic pilot	—	—	no
+ ACI wrap	0	ACI step γ=0.005 fixed [9]	—	—	no
KALMAN-FCP (Kalman)	16/cell	ρ_{scale}, obs.-noise frac.	20% train tail	per-step NLL	**yes**
DIAG-GRU (ablation)	0	frozen, same schedule as ours	—	MLE (train)	no
STATIC-CGIF	0	none (empirical $\hat{\Sigma}$)	—	—	no
FACTORCGIF	0	r=4 (globally fixed / rank matched), θ=0.5 default	—	—	no
GROUPCGIF	0	Louvain resolution 1.0 default	—	—	no
SPECTRALCGIF	0	band split $\{0.33, 0.67\}$ fixed	—	—	no
ACI / AgACI variants	0	step sizes from Zaffran et al. [29]	—	—	no
EWMA / Rolling / ToD / Local Σ_t	0	half-life/window 288, k=50 fixed	—	—	no
COPULACPTS	0	reference defaults (60/40 split, 800 steps)	—	copula calibration	no
MULTIDIMSPCI	0	reference defaults (λ=0.05, net)	—	pinball (own)	no
SPCI / HopCPT	0	reference defaults	—	—	no
DEEPVAR-FCP (control)	0	rank and training budget matched to ours	—	MLE (train)	no

E.21 Controlled same-mean covariance-only comparison

This subsection gives the full record behind the controlled same-mean comparison of §6.2. We freeze a common mean predictor $\widetilde{\mu}_t$, form identical residuals $e_t = Y_t - \widetilde{\mu}_t$, and fit each covariance mechanism on *training residuals only* before freezing it prior to calibration. Each frozen $\hat{\Sigma}_t$ then enters the identical scalar Mahalanobis split-conformal step $\{y : (y-\widetilde{\mu}_t)^\top \hat{\Sigma}_t^{-1}(y-\widetilde{\mu}_t) \le \hat{q}\}$. We retrain the covariance-only heads (GNF-COV, DEEPVAR-COV) against the common residuals. We repeat under two frozen common means: the GPVAR one-step mean (used by the static baselines) and the frozen native GNF mean (a robustness-to-mean-source check).

Size metrics. All size values and paired reductions in Apps. E.21–E.23 use the paper-wide mean full coordinate-projection length w_{proj} defined in (3). For a single ellipsoid,

$$w_{\text{proj}} \;=\; \frac{1}{T_{\text{test}}} \sum_t \frac{2\sqrt{\hat{q}_t}}{N} \sum_{i=1}^{N} \sqrt{\hat{\Sigma}_{t,ii}}.$$

We additionally report normalised log-volume as a geometry-sensitive size measure. Ten paired seeds on METR-LA-20 and PEMS-BAY-50.

Table 22 reports, per mechanism, joint coverage, w_{proj}, and normalised log-volume under the GPVAR mean (mean±std, 10 seeds), plus the paired w_{proj} reduction of GNF-COV against each competitor under *both* means. Every paired 95% CI excludes zero. The native GNF-control reference (own mean + own covariance) has METR-LA coverage 0.903 ± 0.004, $w_{\text{proj}} = 3.236 \pm 0.244$; PEMS-BAY 0.896 ± 0.007, $w_{\text{proj}} = 1.964 \pm 0.072$: the shared-mean GNF-COV matches this, confirming the mean substitution does not degrade the covariance mechanism.

Table 22: **Same-mean covariance-only comparison, full record** (GPVAR-mean stats, mean±std over 10 seeds; target coverage 0.90). Last two columns: paired w_{proj} reduction of GNF-COV vs. the row, under the GPVAR and GNF means (all paired 95% CIs exclude zero). Positive = GNF-COV sharper. The two heuristics that appear narrow (*time-of-day*, *local kNN*) do so by *under-covering* on PEMS-BAY (0.853, 0.852 vs. target 0.90) and are not valid competitors.

	METR-LA-20 (GPVAR mean)			w_{proj} reduction	
Mechanism	joint	w_{proj}	norm-logvol	GPVAR	GNF
GNF-COV (diag+rank4)	0.903 ± 0.006	3.176 ± 0.182	-0.083 ± 0.051	—	—
GNF-COV (rank0/diag)	0.899 ± 0.005	2.829 ± 0.093	-0.004 ± 0.029	see §E.22	
DIAG-GRU-COV (no graph)	0.901 ± 0.005	2.809 ± 0.119	0.012 ± 0.048	see §E.22	
Static full (shrinkage)	0.904 ± 0.011	4.200 ± 0.231	0.409 ± 0.057	24.3%	28.2%
Static factor r=4	0.904 ± 0.013	4.505 ± 0.261	0.429 ± 0.057	29.5%	33.3%
EWMA cov	0.902 ± 0.003	4.446 ± 0.105	0.357 ± 0.034	28.6%	32.2%
Rolling cov	0.896 ± 0.003	4.525 ± 0.119	0.318 ± 0.033	29.8%	33.4%
Time-of-day cov	0.902 ± 0.015	3.762 ± 0.166	0.214 ± 0.041	15.6%	18.8%
Local kNN cov	0.901 ± 0.012	3.916 ± 0.207	0.145 ± 0.051	18.8%	21.2%
	PEMS-BAY-50 (GPVAR mean)			w_{proj} reduction	
Mechanism	joint	w_{proj}	norm-logvol	GPVAR	GNF
GNF-COV (diag+rank4)	0.896 ± 0.004	2.100 ± 0.075	-0.826 ± 0.038	—	—
GNF-COV (rank0/diag)	0.893 ± 0.014	2.071 ± 0.071	-0.769 ± 0.033	see §E.22	
DIAG-GRU-COV (no graph)	0.893 ± 0.022	2.087 ± 0.093	-0.758 ± 0.042	see §E.22	
Static full (shrinkage)	0.891 ± 0.007	3.219 ± 0.126	-0.197 ± 0.047	34.7%	35.2%
Static factor r=4	0.892 ± 0.010	3.334 ± 0.108	-0.146 ± 0.036	37.0%	37.0%
EWMA cov	0.902 ± 0.006	3.966 ± 0.163	-0.198 ± 0.056	47.0%	45.9%
Rolling cov	0.904 ± 0.003	4.033 ± 0.152	-0.204 ± 0.051	47.9%	47.5%
Time-of-day cov	0.853 ± 0.021	2.534 ± 0.085	-0.556 ± 0.038	17.1%	15.9%
Local kNN cov	0.852 ± 0.025	2.823 ± 0.087	-0.566 ± 0.034	25.6%	25.6%

Holding the mean fixed, the learned time-varying covariance GNF-COV is 15–48% sharper on w_{proj} and 0.2–0.7 nats/coordinate smaller in log-volume than every static/heuristic covariance, on both cells and both common means, at nominal coverage, with paired 95% CIs excluding zero. The ranking is essentially identical under the two means. A learned recurrent diagonal (GNF-COV rank0, DIAG-GRU-COV) matches GNF-COV on marginal w_{proj} while GNF-COV retains a joint-set log-volume and NLL edge; this is quantified against the rank-matched comparator in App. E.22. We do not extend this covariance isolation beyond METR-LA-20 and PEMS-BAY-50.

E.22 Matched-budget DEEPVAR-FCP control and graph isolation

Model and budget. DEEPVAR-FCP is a DeepVAR-style [20] matched control: a *global* GRU over the stacked sensor vector with a native mean head and a native time-varying covariance $\Sigma_t = \text{diag}(d_t) + L_t L_t^\top$ of rank 4, feeding the identical scalar Mahalanobis split-conformal step. It provides the global-GRU + low-rank comparison point relative to the paper's DIAG-GRU (global GRU + diagonal) and GNF (per-node graph GRU + low-rank). Alongside it we include an architecture-matched no-graph *LR-GRU* covariance (LR-GRU-COV): the same per-node GRU cell, hidden width, rank, optimiser and epochs as GNF, but with the graph shift zeroed (graph-removal ablation). All learned competitors share a matched budget: Adam (lr 10^{-3}, wd 10^{-4}), rank 4, BPTT 64, 10% warm-up, identical epochs; GNF-control and LR-GRU-COV use hidden 32 per node (as in the paper),

DEEPVAR-FCP uses hidden 64 global (matching the DIAG-GRU row's capacity). No per-dataset covariance-floor or rank selection is performed for any learned method. This is a matched control, not a reference-package reproduction of a named implementation. Training cost is matched to GNF ($\approx$ 45 min/seed); calibration/test is a few seconds and peak memory is comparable ($<$ 7 GB GPU on PEMS-BAY-50). It is included in the reproducibility suite (App. F). This controlled comparison uses the same training schedule as the main model (App. C.2) but a separate set of 10 paired seeds run in the control harness; it is therefore distinct from the Table 2 headline runs, and we label its GNF reference GNF-control.

End-to-end arm (own mean + own covariance). Table 23 gives the native comparison. DEEPVAR-FCP *undercovers* on both cells (0.848, 0.797, with high seed variance) even after conformal calibration—its covariance is unreliable enough that the Mahalanobis score shifts across the calibration/test boundary—whereas GNF holds the 90% target. Across the 10 paired seeds the paired DEEPVAR-FCP$-$GNF differences are $\Delta w_{\text{proj}} = +1.090 \pm 0.228$, Δlogvol $= +0.132 \pm 0.045$ (METR-LA) and $\Delta w_{\text{proj}} = +3.160 \pm 0.181$, Δlogvol $= +0.226 \pm 0.019$ (PEMS-BAY), where "$\pm$" is a 95% CI half-width and positive means GNF sharper (both CIs exclude zero). GNF is thus both sharper and better calibrated than the this DeepVAR-style control.

Table 23: **Matched-budget DEEPVAR-FCP control, end-to-end arm** (own mean + own covariance; 10 seeds, mean$\pm$std; target 0.90). "$\downarrow$" flags under-coverage. Paired DEEPVAR-FCP$-$GNF differences (positive $\Rightarrow$ GNF sharper): $\Delta w_{\text{proj}} = +1.090 \pm 0.228$ / Δlogvol$= +0.132 \pm 0.045$ (METR-LA); $\Delta w_{\text{proj}} = +3.160 \pm 0.181$ / Δlogvol$= +0.226 \pm 0.019$ (PEMS-BAY); "$\pm$" a 95% CI half-width.

	METR-LA-20			PEMS-BAY-50		
Method	joint	w_{proj}	logvol	joint	w_{proj}	logvol
GNF-control	0.902 $\pm$0.005	3.140 $\pm$0.126	$-$0.084 $\pm$0.042	0.897 $\pm$0.006	1.962 $\pm$0.070	$-$0.834 $\pm$0.037
DEEPVAR-FCP (native)	0.848 $\pm$0.047$\downarrow$	4.230 $\pm$0.373	0.048 $\pm$0.065	0.797 $\pm$0.058$\downarrow$	5.122 $\pm$0.227	$-$0.608 $\pm$0.045

Same-mean arm (graph isolation). Covariance-only on the shared residual, the paired w_{proj} reduction of GNF-COV against each competitor (both means) is: DEEPVAR-COV $+35.4\%/+28.2\%$ (METR-LA GPVAR/GNF), $+34.4\%/+24.6\%$ (PEMS-BAY); LR-GRU-COV $-0.5\%/+4.4\%$ (METR-LA), $+1.0\%/+2.3\%$ (PEMS-BAY); static full $+23.3\%/+27.2\%$, $+34.8\%/+36.1\%$; static factor r=4 $+28.5\%/+32.4\%$, $+37.1\%/+37.9\%$. Two findings. *(i)* In this control, the DeepVAR-style global-GRU covariance has poorer NLL (METR-LA GPVAR: 31.1 vs. GNF-COV 3.4) and produces wider sets: GNF-COV has 24–35% lower w_{proj} than DEEPVAR-COV. A learned time-varying covariance alone therefore does not explain the gain. *(ii)* The architecture-matched *per-node* LR-GRU-COV (no graph, rank 4) essentially ties GNF-COV on w_{proj} (-0.5% to $+4.4\%$), while GNF-COV keeps a small but significant log-volume edge (Δlogvol $= +0.014$ to $+0.072$, all paired CIs exclude zero). Interpretation: the sharpness is a learned *per-node low-rank* covariance effect; the graph is a modest joint-set-shape refinement, not the primary covariance mechanism.

E.23 Covariance-floor sweep and conditioning audit

We sweep the diagonal-head floor $\varepsilon_d \in \{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}\}$ (default 10^{-4}) both by post-hoc eigenvalue clipping of one trained model and by full retraining from scratch at each floor. The sweep uses the GNF end-to-end configuration (own mean + own covariance, 10 seeds) from the control protocol of App. E.21, whose METR-LA $w_{\text{proj}} = 3.236$ differs slightly from the Table 2 headline value because it is a separate control run (distinct seed set and harness), not because of a native-versus-shared mean; it is an internally consistent baseline for this sensitivity study. Retrain and post-hoc clipping agree; we display the retrain values.

The floor rarely binds. At the default 10^{-4} the fraction of steps with $\lambda_{\min}(\hat{\Sigma}_t) \le 10\varepsilon_d$ is 0.030 (METR-LA) and 0.096 (PEMS-BAY), and the fraction with $\lambda_{\min} \le 2\varepsilon_d$ is 0.000 on both—the model does not lean on tiny eigenvalues. Condition-number $p99$ is 1402 (METR-LA) and 1021 (PEMS-BAY); raising the floor to 10^{-2} collapses it to 379 and 129, a clean conditioning knob available at negligible coverage cost. The worst-case safeguard $\hat{\Sigma}_t \succeq \varepsilon_d I$, $\|\hat{\Sigma}_t^{-1/2}\|_{\text{op}} \le \varepsilon_d^{-1/2}$ remains valid but is not attained in the reported runs: the high-condition tail exists but is not caused by the full covariance hitting the floor.

Coverage across conditioning strata. Table 25 stratifies test steps by covariance condition number (retrain, ε_d=10^{-4}; target miscoverage 0.10). The worst-conditioned steps have lower miscoverage

Table 24: **Covariance-floor sweep** (retrain-from-scratch, 10 seeds; joint coverage / w_{proj}, mean±std; target 0.90). Coverage and sharpness are flat over four orders of magnitude; the $10^{-6}/10^{-5}/10^{-4}$ rows are identical because the learned per-node variances never reach those levels (the clip is a no-op). Only the aggressive 10^{-2} floor binds: it slightly helps METR-LA and mildly over-regularises PEMS-BAY coverage.

	METR-LA-20		PEMS-BAY-50	
Floor ε_d	joint	w_{proj}	joint	w_{proj}
10^{-6}	0.903 $\pm$0.004	3.236 $\pm$0.244	0.896 $\pm$0.007	1.961 $\pm$0.076
10^{-5}	0.903 $\pm$0.004	3.236 $\pm$0.244	0.896 $\pm$0.007	1.961 $\pm$0.076
10^{-4} (default)	0.903 $\pm$0.004	3.236 $\pm$0.244	0.896 $\pm$0.007	1.961 $\pm$0.076
10^{-3}	0.904 $\pm$0.004	3.202 $\pm$0.221	0.897 $\pm$0.004	1.961 $\pm$0.082
10^{-2}	0.903 $\pm$0.003	3.084 $\pm$0.201	0.889 $\pm$0.004	1.939 $\pm$0.065

on METR-LA and higher miscoverage on PEMS-BAY, while aggregate coverage remains near target. Marginal calibration does not guarantee coverage within these conditioning strata.

Table 25: **Condition-number-stratified miscoverage** (retrain, ε_d=10^{-4}, 10 seeds, mean±std; target 0.10). Steps binned by covariance condition number.

Stratum (by cond. #)	METR-LA-20	PEMS-BAY-50
worst-1%	0.074 $\pm$0.061	0.150 $\pm$0.041
worst-5%	0.031 $\pm$0.030	0.132 $\pm$0.015
best-95%	0.100 $\pm$0.005	0.102 $\pm$0.008

F Reproducibility

Code availability. The full code base is publicly available under an MIT licence at `https://github.com/YannickLimmer/filter-cp`. It includes a self-contained reproducibility suite (`repr/reproduce.py`, `repr/expected.json`, `repr/requirements.txt`) that re-runs the experiment cells and checks the resulting numbers against the values reported in the main-body tables and figures.

Software and hardware. Experiments used Python 3.11, PyTorch 2.2.0 (CUDA 12.1), NumPy 1.26, SciPy 1.11, NetworkX 3.2, and POT 0.9. Neural-filter training used a single NVIDIA RTX 4090 with 24 GB memory; the largest reported training run peaks at 6.2 GB GPU memory. Static covariance and Kalman baselines run on CPU. All experiments use deterministic PyTorch operations (`torch.use_deterministic_algorithms(True)`) and a fixed CUDA workspace size; the only platform-dependent non-determinism is the BLAS reduction order in the GCN aggregation step on different backends, which contributes a cross-platform drift within 10^{-5} relative tolerance on widths.

Seeds and splits. Every reported real-data cell uses 10 fixed seeds and the chronological 40/40/20 train/calib/test split described in App. E.1. Tables report 10-seed mean±std unless noted otherwise; per-seed values for the main method pair on a three-seed sub-sample appear in Table 15. Per-seed spreads for every other method are within the stds reported in Table 14.

Data sources and reuse terms. METR-LA and PEMS-BAY use the benchmark release of Li et al. [17], available from `https://github.com/liyaguang/DCRNN`; its repository carries an MIT licence. The ELEC source is UCI ElectricityLoadDiagrams20112014 (`https://doi.org/10.24432/C58C86`), released under CC-BY-4.0. ETT uses the Informer ETTh1 release at `https://github.com/zhouhaoyi/ETDataset`. The other benchmark sources are AQI (Beijing $PM_{2.5}$), SOLAR (NREL Alabama solar power), JENA (Max Planck Biogeochemistry; `https://www.bgc-jena.mpg.de/wetter/`), and LOOP-SEA (Seattle loop-detector traffic); their reported panels and preprocessing are described in App. E.1. Separate dataset-specific reuse terms for ETT and these latter sources are not fully documented here; public availability alone does not specify a reuse licence. We do not redistribute the raw benchmark data.

Reused implementations. The vendored CopulaCPTS code retains the MIT licence of Spatiotemporal Machine Learning (2024), and the vendored MultiDimSPCI code retains the MIT licence of hamrel-cxu (2024). Their upstream commit identifiers and licence files accompany the source in `repr/_lib/`. The new filter and calibration code is released under the repository's MIT licence.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Section 1 state the assumptions and scope of the contraction, marginal-coverage, and Gaussian-comparator log-volume results. Section 6 reports sharpness gains on moderate-dimensional traffic data and the full-graph and non-graph-native crossovers.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 7 discusses the theory assumptions, marginal coverage, restricted log-volume comparator class, fixed model capacity, numerical conditioning, and empirical scope. The audits measure finite-trajectory quantities and do not certify population assumptions.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Assumptions 2–6 and the hypotheses in Appendix B specify the conditions. Appendices A–B give proofs, including the finite-rank quantile correction, buffered threshold interval, and conditional second moment for the local-linearisation result.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Appendices C–E describe architectures, fitting, splits, hyperparameters, and controlled comparisons. Appendix F gives software, hardware, seeds, and the public reproduction entry point.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All datasets are publicly available; their sources are listed in Appendices E.1 and F. Code and a reproducibility suite are available under an MIT licence at `https://github.com/YannickLimmer/filter-cp`.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Appendix C specifies architectures and training; Appendix E.1 specifies preprocessing and chronological splits. Table 21 distinguishes model fitting from hyperparameter search and records selection criteria.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: Tables 2, 22, 23, and 24 report 10-seed means and standard deviations for coverage and size. Table 18 reports normalised log-volume variation, and Table 25 reports stratified miscoverage variation. Paired 95% confidence intervals are distinguished from standard deviations; diagnostic replication counts are specified in their protocols.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix E.19 reports per-method training/inference times and memory. Appendix F gives software versions and hardware, including the RTX 4090 and peak GPU memory.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The study uses existing public traffic, energy, environmental, and weather benchmarks. No participants were recruited and no new personal data were collected.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Section 7 discusses potential benefits for network-level decisions and risks from misinterpreting marginal coverage or deploying under distribution shift.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The released assets are small forecasting filters and calibration code; they do not include high-risk pretrained generative models or newly scraped datasets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [No]

Justification: Creators are credited and the vendored baseline code retains its MIT licence files. Appendix F records verified terms, but separate dataset-specific reuse terms are not documented for every benchmark.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: Appendix C documents the learned filter and conformal wrapper. The public MIT-licensed repository provides installation and reproduction instructions.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: No crowdsourcing or research with human participants was conducted.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: No research with human participants was conducted.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not an important, original, or non-standard component of the methods, theory, or experiments in this work.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.