FLUFORECASTBENCH: INFLUENZA FORECASTING WITH PUBLICATION-TIME INFORMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable influenza forecasting requires accurate predictions and useful uncertainty estimates from surveillance reports that are delayed, incomplete and revised. Retrospective evaluation using later reports can conceal the information constraints faced by a forecaster. Although prior studies establish the importance of reporting revisions, comparisons must also account for missing observations and changes in the data used to calibrate predictive uncertainty. To make these constraints explicit in evaluation, we introduce **FluForecastBench**, covering four countries, eight tasks and 90 surveillance series. It pairs histories selected under explicit publication rules with specified later editions and evaluates both forecasts against the same targets. The US condition uses archived early reports. Other tasks follow their recorded publication or first-seen rules. We evaluate twelve baseline procedures across seven per-task leaderboards. Relative performance varies across tasks: seven fitted procedures have higher error than a seasonal naive forecast on the German state panel. A separate five-baseline experiment in the United States and Japan separates model selection from interval calibration over two seasons. All four fitted baselines improve mean point accuracy under publication-constrained inputs, but nominal 80% intervals cover only 39–54% of US targets and 15–40% of Japanese targets under a shared preceding-season residual calibration scheme. Further analyses distinguish revised values from differences in available history and calibration data. These results identify incomplete histories and the transfer of uncertainty calibration across seasons as concrete challenges for influenza forecasting.

1 INTRODUCTION

Influenza forecasts inform public-health planning, which requires accurate predictions and reliable uncertainty estimates (Reich et al., 2019; McGowan et al., 2019). Surveillance reports arrive late, contain gaps and may subsequently be revised. The latest observation can lag the forecast target by several weeks, while reporting practices vary across surveillance systems. Models must address these information constraints alongside epidemic dynamics.

Retrospective evaluation can obscure these constraints. A chronological train–test split separates training and evaluation targets but leaves past observation versions unspecified. Records obtained after the evaluation period may contain subsequently reported observations or revised values, changing both model inputs and interval-calibration data. Reporting revisions matter for influenza forecasting (Brooks et al., 2018). A shared benchmark should test whether gains persist under restricted information and separate value revisions from changes in available history.

We introduce **FluForecastBench**, covering the United States, Japan, Canada and Germany, with eight tasks and 90 surveillance series (Figure 1). **Condition A (publication-constrained history)** selects observations under a declared task-specific rule. US A uses archived early-report values, a stricter condition than the complete history available at issuance. Other tasks use their recorded publication or first-seen rules. **Condition B (later edition)** uses a specified later history, including subsequently reported or revised observations. The same procedure forecasts under both conditions, and a fixed evaluator scores both against identical targets. A is the primary condition. B measures sensitivity to a specified later edition, which may still be subject to further revision.

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

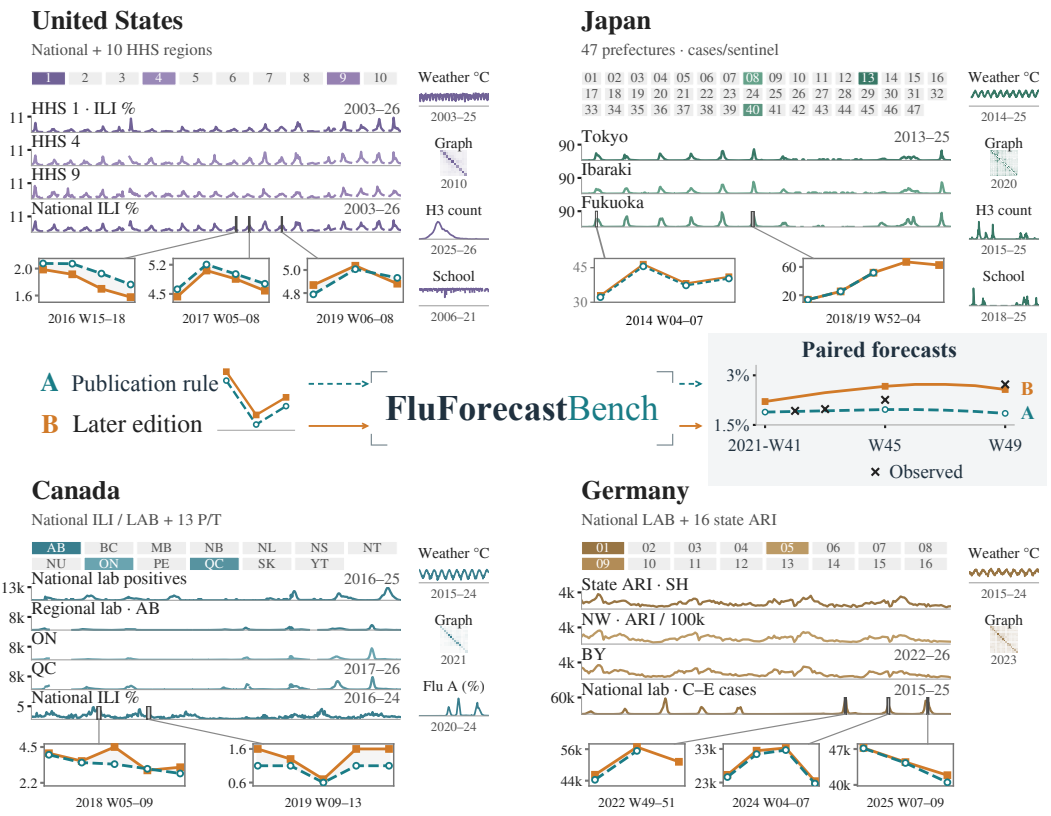


Figure 1: FluForecastBench pairs publication-constrained histories (A, teal dashed) with later editions (B, orange solid). Country panels show surveillance and auxiliary resources in native units. The centre shows frozen US forecasts against shared targets. The forecast example is a post-hoc illustration (Appendix B.1).

The main comparison reports per-task leaderboards across all four countries, making both gains over seasonal naive forecasting and unresolved comparisons visible. A separate controlled experiment compares seasonal naive, ridge regression (Hoerl & Kennard, 1970), COLA-GNN (Deng et al., 2020), the regional model of Murayama et al. (2021), and CAMul (Kamarthi et al., 2022a) on shared US and Japanese targets. It separates model selection from interval calibration in time. Fixed-state controls and regional peak analyses reveal failures and information constraints obscured by aggregate scores.

Our contributions are:

- We introduce **FluForecastBench**, with paired histories, fixed evaluators and a common submission interface for eight tasks across four countries.
- We evaluate twelve baseline procedures across seven per-task leaderboards and provide a controlled five-baseline comparison with separate model-selection and calibration periods, three random seeds for each neural model and two seasons per task.
- We examine how observation availability, revised values and residual calibration affect forecast scores, using shared targets and fixed model or calibration states.

2 Related Work

Forecast evaluation and revision-aware models. Collaborative influenza studies established shared targets and multi-season evaluation (Reich et al., 2019; McGowan et al., 2019). Hubs preserve submitted forecasts (Cramer et al., 2022). Real-time forecasting systems use information

Table 1: Task coverage. Counts describe source series, not the number of predictions available for every baseline. National and regional targets retain their native units and calendars.

Country	Task	Series	Weekly target
United States	US-NAT	1	Weighted ILI, % of outpatient visits
	US-HHS	10	Regional weighted ILI, % of visits
Japan	JP-PREW15	47	Influenza cases per reporting sentinel
Canada	CA-ILI	1	ILI, % of sentinel consultations
	CA-LAB	1	Influenza-positive specimen count
	CANADA-PT	13	Provincial/territorial positive detections
Germany	DE-LAB	1	Laboratory-confirmed C–E notifications
	DE-LAENDER-ARI	16	ARI consultations per 100,000

available when forecasts are issued (Shaman et al., 2013). ARGO examines vintage sensitivity (Yang et al., 2015), iterative forecasting models revisions (Brooks et al., 2018), and Back2Future learns backfill dynamics (Kamarthi et al., 2022b). FluForecastBench evaluates these approaches with paired inputs and controls that hold fitted states fixed.

Vintage archives and forecasting benchmarks. Dated vintages are established in macroeconomics (Croushore & Stark, 2001). COVIDcast records event and issue dates (Reinhart et al., 2021). Monash and GIFT-Eval compare forecasting methods across domains, frequencies and horizons (Godahewa et al., 2021; Aksu et al., 2024). Our focus is the information state supplied to the learner: we combine task-specific history rules, paired forecasts, shared evaluation targets and calibration controls. An archive supplies editions. This evaluation measures how changing editions affects both forecast quality and procedure comparisons (Appendix D.4).

3 FluForecastBench

3.1 Tasks and data construction

The benchmark contains eight tasks with distinct native targets (Table 1). US tasks measure influenza-like illness (ILI). Japanese and German regional tasks measure sentinel or consultation rates, while Canadian and German laboratory tasks measure surveillance counts. These are observed surveillance quantities rather than interchangeable measures of infection incidence (Japan Institute for Health Security, n.d.b; Centers for Disease Control and Prevention, n.d.; Public Health Agency of Canada, n.d.; Robert Koch Institute, n.d.b). Scores in different measurement units are reported separately. Japanese targets precede the 2025-W15 sentinel-system transition (Japan Institute for Health Security, n.d.a).

Each task provides historical observations, edition identifiers and missingness masks. A selects from the recorded publication states. B selects a fixed later edition. A procedure returns point forecasts and predictive quantiles, and the evaluator attaches target observations only after predictions are frozen. Archived data, forecast dates and predictions have different coverage. Some models lack enough predictions for evaluation in a given season. US A selects only recorded lag-0/lag-1 editions, omitting intermediate revisions already public by later origins (Appendix B).

3.2 What makes the benchmark challenging?

Prediction lead depends on observation availability. A forecast coordinate fixes task, geography, origin, target week and nominal horizon. At h2, the target is two native-calendar positions beyond the origin, not necessarily two weeks after the publication cutoff. The *effective lead* is its distance from the most recent observation used by the model. At recent Japanese origins, median effective lead is five weeks for A and three for B: A requires bridging a longer unobserved interval. Missing weeks retain their calendar positions rather than becoming zeros or being removed.

Figure 2 contrasts two information constraints: the US example contains revised values at a shared endpoint, while Tokyo and NRW have endpoint gaps. Ontario's A history ends at W23 while B extends to W50. Across the recent Canadian provincial cohort, median effective lead is thirteen weeks

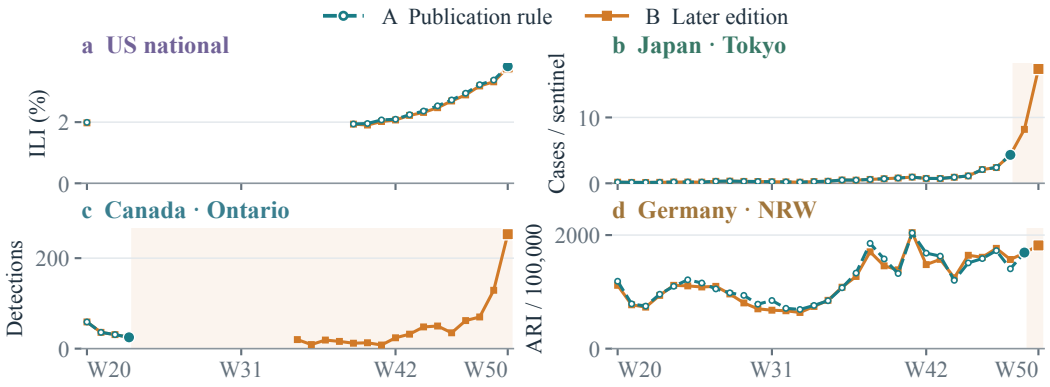


Figure 2: Actual input histories at origin 2024-W51. Dashed A and solid B lines retain resolved observations and gaps. US national histories share an endpoint but contain revisions. Tokyo, Ontario and NRW have different endpoints. Shading marks endpoint gaps, including archive gaps in Ontario.

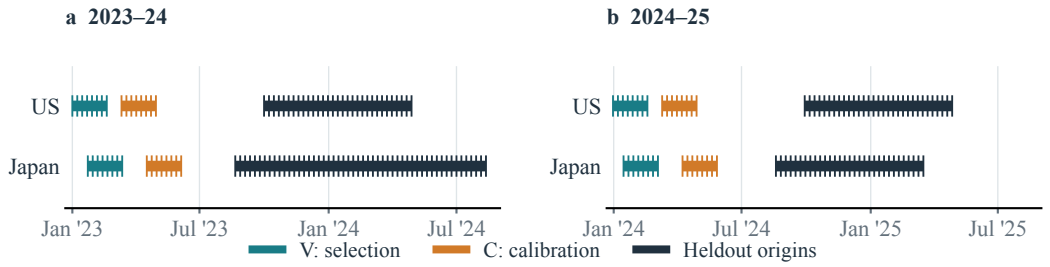


Figure 3: Actual origin dates for the two heldout folds. Models used for selection are fitted before V. Final A/B models are fitted before C and forecast both C and the held-out period. Heldout marks show common evaluation origins. C labels only calibrate intervals around frozen point forecasts. Exact cutoffs appear in Appendix D.3.

under A, ranging from four to thirty-three. Archive incompleteness contributes to these measured ages. They are not estimates of official publication delay.

Seasonal and regional variation poses challenges for transfer. Historical and recent seasons span different epidemic patterns. These comparisons examine transfer, not causal pandemic effects. Different regional observation systems require native-unit metrics and a seasonal-naive reference.

Accurate point forecasts can have poorly calibrated prediction intervals. Quantiles come from native samples, dropout, residual calibration or mechanistic models. Calibration data can change with the input vintage even when point forecasts agree. The benchmark requests points and nine predictive deciles, with the target and calibration policy declared.

4 Evaluation Protocol

4.1 Chronological seasonal evaluation

Figure 3 shows the controlled US/Japan experiment. Fitting and development precede heldout origins. A/B histories, fitted states and residual states remain separate. Recipes record cutoffs, permissible inputs, update and seed policies (Appendix C.1).

Disjoint selection and calibration. The controlled experiment covers US-HHS and Japan in 2023–24/2024–25. Before computing errors, it reserves the last eight preceding-season origins with A/B-common labels available at the first heldout origin as calibration block C. The preceding eight

eligible origins with labels available at C's start form selection block V. Models fit before V. Epochs or penalties are selected on V's condition-A data, then models are refitted for A/B before C. The final state predicts C and the heldout season without updates. Scalers, imputations and graph inputs follow the declared cutoff rules. C residuals never enter selection.

The adapter pools signed residuals across regions within task, season, seed and condition. Type-7 deciles are centered by their median, retaining the point as $q_{0.5}$ without clipping negative lower quantiles. All procedures and seeds share calibration coordinates, with at least 30 coordinates and five origins required. Each neural seed is scored separately. Appendix D.3 reports exact dates and earlier-recipe comparisons.

4.2 Metrics, support and uncertainty

For each coordinate, point error is $|\hat{y} - y|$. The probabilistic score is

$$\ell_{\mathrm{WIS}}(y, q) = \frac{2}{9} \sum_{\tau \in \mathcal{Q}} \big(\tau - \mathbf{1}\{y < q_\tau\}\big)(y - q_\tau), \tag{1}$$

where y is the evaluation observation, q_τ the predicted quantile and $\mathcal{Q} = \{0.1, \ldots, 0.9\}$ the exact decile grid. This equally weighted doubled pinball score is called WIS in the recent results (Bracher et al., 2021). Lower MAE and WIS are better. Historical 23-quantile grids remain separate. Missing quantiles are not interpolated.

We intersect A/B predictions by their full coordinate key and average hierarchically:

$$\overline{L}^k = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \frac{1}{|\mathcal{O}_s|} \sum_{o \in \mathcal{O}_s} \frac{1}{|\mathcal{G}_{so}|} \sum_{g \in \mathcal{G}_{so}} \ell^k_{gos}, \qquad k \in \{A, B\}, \tag{2}$$

where $\mathcal{S}$ contains the eligible seasons, $\mathcal{O}_s$ their origins, $\mathcal{G}_{so}$ their geographies and ℓ^k_{gos} the coordinate loss. Losses are averaged equally over regions within origins, origins within seasons, and seasons. Task units stay separate. Table 3 fixes these sets across procedures, seeds, A/B and point/probability metrics. Relative MAE, $\overline{L}^k_m / \overline{L}^k_{\mathrm{SN52}}$, divides mean loss by SN52's, with values below one indicating improvement. Earlier WIS analyses use different sets of evaluation coordinates (Appendix D).

For paired changes, $\Delta = L^A - L^B$ is positive when B has lower loss. The controlled comparison uses 2,000 circular origin-block replicates with shared resamples (Appendix D.3). Leaderboards and reference contrasts use 10,000 season-stratified stationary-block replicates with Type-7 endpoints (Politis & Romano, 1994; Hyndman & Fan, 1996). Retained-origin and calendar-gap rules and block lengths 4/8/12 assess sensitivity. The intervals are conditional on the observed seasons.

4.3 Baselines and submission interface

Appendix Table 8 lists twelve procedures. Repeat Last predicts the latest valid observation. SN52 uses the observation exactly 52 calendar positions before the target week. Ridge Direct and Recursive (Hoerl & Kennard, 1970) differ in prediction strategy. Chronos (Ansari et al., 2024) and FlowState (Graf et al., 2025) use fixed retrospective checkpoints. COLA-GNN (Deng et al., 2020), Murayama's regional model (Murayama et al., 2021), CAMul (Kamarthi et al., 2022a), the two EiNNs variants (Rodríguez et al., 2023) and SIR-INN (Rama et al., 2026) use declared task adaptations. Latent count-target SIR and observation-predictive distributions are kept separate.

New procedures freeze point or quantile submissions keyed by task, geography, origin, horizon and target week. The evaluator validates keys, preserves missing outputs and scores shared targets. The anonymous supplement supports controlled-experiment replay, input reconstruction and baseline refitting through a locked `uv` environment (Appendix G).

5 Results and diagnostic analysis

5.1 Which procedures work across surveillance tasks?

Table 2 compares seven tasks across four countries. Each panel shares coordinates, targets and weighting across procedures and A/B, but fitting policies remain procedure-specific. Appendix D.1

gives B and retrospective pretrained results, excluded from the primary advantage-change analysis. CA-LAB is a post-hoc 2024–26 archive extension. Pointwise intervals condition on observed seasons. Overlapping rank intervals indicate uncertainty in the ordering of methods.

Table 2: Publication-constrained accuracy on common task coordinates. A rMAE is MAE relative to SN52 (lower is better), with conditional pointwise 95% intervals. Models have a fixed display order. Rank intervals describe the nonpretrained member ranking. Teal/rose indicates a supported decrease/increase against SN52 under both resampling rules.

US regions
2023–25; 12, 12 origins

Baseline	A rMAE ↓ [95% CI]	Rank [95% CI]
SN52	1.00	3–5
Repeat	1.74 [1.36,2.18]	8–9
Ridge-D	0.75 [0.58,0.97]	2
Ridge-R	0.80 [0.63,1.02]	3–4
COLA-GNN	0.60 [0.52,0.70]	1
Murayama	1.31 [0.99,1.67]	5–8
CAMul	1.34 [1.17,1.60]	6–9
EiNNs-T	2.67 [2.45,2.96]	10
EiNNs-X	0.97 [0.82,1.23]	4–6
SIR-INN	1.26 [1.07,1.54]	6–8
Median	0.55 [0.47,0.65]	—

US national
2023–25; 27, 27 origins

Baseline	A rMAE ↓ [95% CI]	Rank [95% CI]
SN52	1.00	3–5
Repeat	0.94 [0.71,1.32]	3–5
Ridge-D	0.61 [0.41,0.78]	1
Ridge-R	0.64 [0.46,0.82]	2
CAMul	2.02 [1.44,2.55]	6–7
EiNNs-T	2.46 [2.08,2.95]	8
EiNNs-X	1.77 [1.36,2.14]	6–7
SIR-INN	0.97 [0.77,1.22]	3–5
Median	0.78 [0.57,0.98]	—

Japan
2023–25; 52, 31 origins

Baseline	A rMAE ↓ [95% CI]	Rank [95% CI]
SN52	1.00	7–10
Repeat	0.79 [0.43,1.18]	1–10
Ridge-D	0.80 [0.67,0.96]	3–9
Ridge-R	0.82 [0.71,0.95]	3–9
COLA-GNN	0.66 [0.42,0.89]	1–4
Murayama	0.63 [0.37,0.87]	1–4
CAMul	0.78 [0.47,1.06]	2–9
EiNNs-T	0.78 [0.60,0.99]	2–8
EiNNs-X	0.79 [0.62,0.99]	4–9
SIR-INN	0.75 [0.41,1.16]	1–10
Median	0.69 [0.49,0.88]	—

Germany lab.
2023–25; 26, 32 origins

Baseline	A rMAE ↓ [95% CI]	Rank [95% CI]
SN52	1.00	1–7
Repeat	0.87 [0.38,2.70]	1–5
Ridge-D	1.05 [0.61,3.40]	2–5
Ridge-R	1.19 [0.73,4.11]	4–6
CAMul	0.77 [0.33,1.94]	1–3
EiNNs-T	1.40 [0.74,4.84]	4–7
EiNNs-X	1.44 [0.78,5.09]	5–8
SIR-INN	1.93 [0.74,4.36]	2–8
Median	0.88 [0.45,2.63]	—

Germany states
2024–25; 50 origins

Baseline	A rMAE ↓ [95% CI]	Rank [95% CI]
SN52	1.00	1–2
Repeat	1.08 [0.86,1.32]	1–3
Ridge-D	1.45 [1.23,1.74]	3–8
Ridge-R	1.45 [1.22,1.75]	3–9
COLA-GNN	1.54 [1.29,1.79]	3–8
Murayama	1.55 [1.19,1.84]	3–8
CAMul	1.48 [1.30,1.68]	4–7
EiNNs-T	2.14 [1.46,2.59]	6–9
SIR-INN	1.40 [1.14,1.65]	2–7
Median	1.21 [1.01,1.41]	—

Canada provinces
2024–25; 16 origins

Baseline	A rMAE ↓ [95% CI]	Rank [95% CI]
SN52	1.00	1–4
Repeat	1.36 [0.85,3.10]	2–7
Ridge-D	0.85 [0.71,1.22]	1–2
Ridge-R	1.22 [0.93,2.17]	3–6
CAMul	1.13 [1.01,1.48]	3–5
EiNNs-T	1.38 [1.33,1.51]	4–6
SIR-INN	2.75 [1.88,3.34]	5–7
Median	0.80 [0.69,0.91]	—

Canada lab.
2024–26; 12, 26 origins

Baseline	A rMAE ↓ [95% CI]	Rank [95% CI]
SN52	1.00	4–8
Repeat	0.68 [0.43,0.98]	2–5
Ridge-D	0.49 [0.27,0.68]	1
Ridge-R	0.62 [0.46,0.79]	2–5
CAMul	0.84 [0.52,1.22]	5–7
EiNNs-T	0.74 [0.42,1.11]	3–6
EiNNs-X	0.73 [0.41,1.09]	2–5
SIR-INN	0.91 [0.59,1.32]	7–8
Median	0.66 [0.36,0.99]	—

Reading the panels. Origin counts follow season order. Seasons have equal weight. Printed score and rank intervals use retained-origin blocks of expected length eight. Colour also requires the calendar-gap rule. Intervals condition on observed seasons.

Rows. References and linear procedures precede neural procedures. Median is the post-hoc ensemble on the same coordinates, outside the member ranks. D/R: direct/recursive; T/X: target/external inputs.

Scope. A controls surveillance histories. Auxiliary inputs follow Section 3. CA-ILI has no qualified common panel. Raw errors, both interval rules, recipe variants and the CA-LAB extension are documented in Appendix D.1.

Accuracy gains depend on the task. Ridge Direct's A error is below SN52 under both interval rules in four of seven tasks. COLA-GNN improves on SN52 in Japan and US regions, while Murayama improves in Japan. All seven fitted procedures on the German state panel have higher A error than SN52. On the German laboratory panel, no individual procedure's relative-error interval excludes one under both resampling rules. The exploratory median ensemble is reported separately from the rankings of its members. These differences expose regional and target-specific challenges,

Table 3: Joint point/probability comparison over two heldout seasons with disjoint selection and calibration, h2. All methods, seeds and A/B share coordinates within each task. Small ± denotes seed SD for neural MAE/WIS. Seasonal SD and other seed SDs are in Appendix D.3. Bold rMAE below one indicates improvement over SN52, not significance. Coverage targets 80%. Width is interpreted jointly with coverage.

	MAE ↓		rMAE ↓		WIS9 ↓		Coverage (%)		Width	
Baseline	A	B	A	B	A	B	A	B	A	B
US HHS regions	620 shared coordinates; ILI percentage points									
SN52	1.277	1.287	1.000	1.000	1.141	1.150	39.0	36.8	1.21	1.22
Ridge Direct (Hoerl & Kennard, 1970)	0.831	0.838	**0.651**	**0.651**	0.722	0.724	53.9	50.2	1.28	1.17
COLA-GNN (Deng et al., 2020)	0.837±0.061	0.809±0.044	**0.655**	**0.628**	0.745±0.055	0.716±0.026	44.5	49.0	0.82	0.87
Murayama (Murayama et al., 2021)	0.789±0.013	0.843±0.040	**0.618**	**0.655**	0.700±0.006	0.756±0.039	44.9	36.6	0.76	0.69
CAMul Bounded (Kamarthi et al., 2022a)	0.987±0.086	1.048±0.138	**0.773**	**0.814**	0.857±0.073	0.915±0.113	45.9	37.8	1.16	0.98
Japan prefectures	3,618 shared coordinates; cases per sentinel									
SN52	11.86	11.86	1.000	1.000	11.42	11.42	14.9	14.9	2.46	2.46
Ridge Direct (Hoerl & Kennard, 1970)	11.08	7.35	**0.934**	**0.620**	10.29	6.78	39.8	23.2	7.23	3.95
COLA-GNN (Deng et al., 2020)	11.66±0.90	5.40±0.30	**0.983**	**0.455**	11.23±0.88	5.16±0.35	27.8	40.5	3.84	2.18
Murayama (Murayama et al., 2021)	9.56±0.10	4.68±0.13	**0.806**	**0.394**	9.36±0.09	4.56±0.12	32.2	30.9	1.43	1.19
CAMul Bounded (Kamarthi et al., 2022a)	8.53±0.64	5.50±0.48	**0.719**	**0.464**	8.11±0.55	5.29±0.49	30.3	42.0	3.46	2.12

with no consistent evidence of a universally superior architecture. US-HHS covers twelve late-season origins per season. The Canadian provincial and German state panels each cover one season.

5.2 Can one recipe deliver accurate points and useful uncertainty?

Table 3 uses shared US/Japan coordinates and a common residual adapter. MAE and WIS are computed from forecasts produced by the same fitted model. Cutoffs, variants and support differ from Table 2, so cross-table score differences cannot estimate improvement. Neural contexts and adaptations differ by model. Appendix D.3 gives seasonal scores, seed SD and coverage losses, with paired 4/8/12-origin block intervals.

Point accuracy improves while coverage remains low. All 72 neural final fits succeed. On 620 common US coordinates, A-condition Murayama has MAE 0.789 and WIS 0.700, versus SN52's 1.277 and 1.141. Its 80% intervals cover only 44.9% overall and 12.2% near peaks. Across procedures, US A coverage is 39.0–53.9% overall and 9.3–29.7% near peaks. Coverage thus remains poor despite the lower WIS relative to SN52.

The difficulty persists across input conditions. On 3,618 common Japanese coordinates, B lowers mean point error for all four fitted procedures, but overall coverage remains 14.9–42.0%. A coverage is 14.9–39.8% overall and 0–5.7% near peaks (Figure 4). CAMul Bounded has the lowest A mean MAE, 8.534, but only 0.7% peak coverage. These results challenge the fixed, pooled residual adapter, not the possibility of calibrated forecasting. The short calibration window and its first-fold support of only 38 coordinates remain limitations.

5.3 Does a later history change model comparisons?

On 1,783 common Japanese coordinates in 2023–24, Ridge Recursive has 61.6% higher WIS than SN52 under A, but 37.8% lower WIS under B (Appendix Figure 8). Statistical evidence of a reversal requires uncertainty estimates for both advantages and their change.

We test $\Gamma = (L_i^A - L_j^A) - (L_i^B - L_j^B)$ using simultaneous bootstrap bands over pairwise A, B and change contrasts within each panel. Excluding pretrained procedures while retaining the original

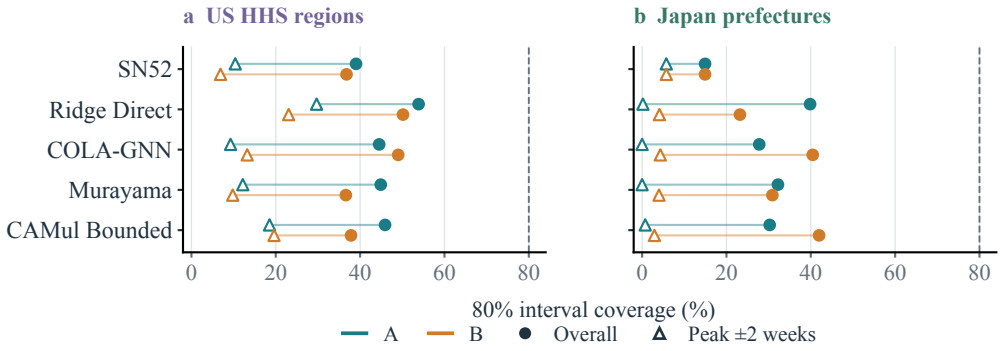


Figure 4: Coverage of nominal 80% intervals on the controlled common panel, overall and within two weeks of each region's peak. Symbols show mean coverage across seeds, with coverage computed separately for each seed. Peaks are located after forecast freeze using the maximum observed value over registered target weeks. The dashed line is nominal coverage. Interval widths and WIS must be read alongside coverage in Table 3.

coordinates and bands, seven of 42 recent WIS contrasts support a change under both resampling rules. Six involve Ridge Recursive in Japan or Canada. No comparison statistically supports opposite rankings under A and B. The multi-season MAE reference supports changes in four of eighteen comparisons. The comparisons share seasons and US surveillance data and therefore cannot be treated as independent replications. Results and block-length sensitivities remain in Appendix E.6. These diagnostics are exploratory and were designed after earlier results were inspected.

5.4 Which information constraint changes the forecasts?

Availability explains the Japanese point response. Truncating B at A's latest observation makes every Japanese Direct and Recursive Ridge prediction equal its A-input counterpart, under either fixed fitted state. Shared values agree, but effective leads are five versus three weeks. Recursive MAE differences of 2.43–2.52 and 1.72–1.85 vanish (Appendix Figure 7). This control isolates availability effects on point forecasts. WIS requires a separate decomposition.

A smaller value-revision effect remains in Canada. On 196 CANADA-PT coordinates in 2024–25, endpoint matching leaves revised values in 107 coordinates. Recursive Ridge's fixed-state MAE differences fall from 149.8/79.1 to 6.86/7.62 for the two fitted states. After endpoint matching, intervals exclude zero under both resampling rules. Availability explains much of the response, with a smaller value effect.

Identical points can conceal a calibration effect. All 3,242 paired Japanese SN52 points agree, but exchanging frozen residual states changes mean WIS by 0.381 [0.109, 0.649], or 3.3% of the B-state score. Restricting B residuals to shared development coordinates reproduces A's WIS. Adding B-only support recovers the lower B score. The single-season 2024–25 effect remains uncertain.

Effects are setting-dependent. The A–B difference changes sign across seasons in eleven of 53 comparable probabilistic series. Input-age restrictions also change eligibility and supported contrasts. Gains from later histories and suitable age thresholds depend on the setting (Appendix E.6).

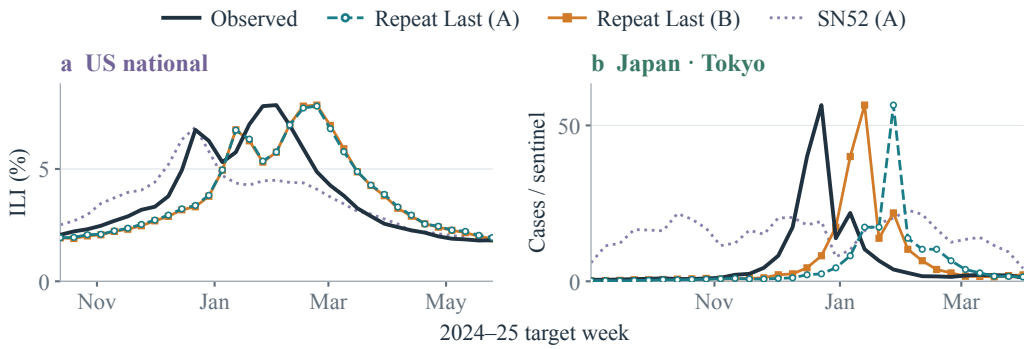


Figure 5: Qualitative forecasts over all scored 2024–25 weeks in two locations. Curves connect rolling h2 predictions at their target weeks. Repeat Last uses A/B inputs. SN52 uses A. Prediction gaps are retained.

Qualitative forecast behavior. Figure 5 shows failures hidden by aggregate losses. US A/B persistence forecasts nearly coincide but peak three weeks late. Tokyo's A and B peaks lag by five and three weeks, while SN52 misses the peak amplitude. These replays illustrate timing and seasonal-transfer errors for the selected deterministic baselines. Rankings across the full roster require broader comparisons.

6 Limitations and Reproducibility

The controlled experiment covers two tasks over two seasons, with eight calibration origins per fold. Missing SN52 lags leave Japan's first fold with 38 common calibration coordinates, versus 376 in the second. Dependent pooled residuals do not guarantee nominal coverage. Freezing states before C sacrifices recent training information. Comparisons between earlier and current configurations cannot isolate the training-cutoff effect. Peaks are identified after forecasts are frozen, and experiments were designed with earlier results known. The leaderboard retains task-specific recipes and short or single-season panels.

For the US, condition A uses early reports rather than complete issuance-time history. A–B differences can include revisions already available to a real-time forecaster. Archive gaps and auxiliary availability also limit operational interpretation. The comparison also lacks revision-aware models.

The anonymous supplement reproduces 5,700 controlled-experiment numerical checks and all four CPU Ridge/SN52 folds. The full set of 72 GPU fits has not been rerun since packaging, and external replication remains unverified. Leaderboard aggregates and restricted-input access routes are documented separately (Appendix G).

7 Conclusion

FluForecastBench makes surveillance information constraints part of the forecasting problem. Per-task leaderboards, controlled point/interval comparisons and peak diagnostics give new methods a concrete objective: accurate operational forecasts with useful uncertainty under delayed observations and changing epidemics.

Future work. Future evaluations should compare delay-aware and revision-aware models under matched publication rules, forecast origins and tuning budgets. Adaptive calibration could address low coverage near epidemic peaks, but requires evaluation across additional heldout seasons and regions. Extending the US archive to complete issuance-time histories and collecting prospective submissions would test whether retrospective findings transfer to operational forecasts. Independent reconstruction of permitted source data and replication of the packaged GPU pipeline are also needed to establish portability beyond the current local checks.

AI USE STATEMENT

Generative AI tools assisted conceptual framing and hypothesis refinement, the design and critique of evaluation methods and bounded diagnostic experiments, interpretation of results and uncertainty, literature search and related-work discovery, source-document review and data-formatting code, method implementation, analysis execution, figure and artifact preparation, and drafting of manuscript sections, translation and prose revision. This included feedback on common-support comparisons, endpoint controls and the sequential residual-state diagnostic. AI-assisted implementation also produced the mechanistic family's higher-precision posterior approximation, the Canadian EiNNs condition-B development fit and calibration layer, Canadian Ridge Direct horizon-8 selection and predictions, and German Murayama horizon-8 recursive inference. Reported surveillance observations and evaluation targets come from the bound source records, not AI-generated synthetic observations.

Verification used source and artifact identity checks, exact coordinate/support joins, regression tests, recomputation of scores from frozen predictions, original-state and controlled-replay identity checks, clean-extraction analysis regeneration, and rendered-manuscript inspection, as documented in Appendix G. These checks also used AI-assisted code and review. They are not independent external replication. The authors take responsibility for the final methods, claims, text, figures, code and artifacts, including AI-assisted contributions. AI tools are not authors.

ETHICS STATEMENT

This benchmark uses population-level public-health surveillance rather than individual clinical data, retaining source limitations and explicit unavailability. The public release policy excludes content without redistribution authorization, including the AGI/NRZ virological extracts, derivatives and source metadata covered by the provider clarification. A local reconstruction procedure documents their use for researchers who obtain the sources under applicable terms. Processing instructions do not confer access or redistribution rights. Local validation artifacts are distinct from a publicly cleared distribution (Appendix G). Prediction generation and evaluation remain separate: forecasts are frozen before the evaluator is attached. The authors are responsible for the scientific claims and final submission.

REPRODUCIBILITY STATEMENT

Seven-task leaderboard aggregates are supplied with scores, intervals, rank ranges and colour criteria. Six-task means are independently recomputed from frozen coordinate losses. The CA-LAB extension and bootstrap records are checked against delivered aggregates. The anonymous code supplement directly supports the controlled five-procedure comparison (Table 3). It includes resolved US/Japan histories, numerical model inputs, separate calibration and evaluation observations, frozen predictions, and scoring and training code with a locked `uv` environment. A fresh extraction reproduces 99,696 score rows and passes 5,700 numerical checks of aggregate, seasonal, seed, peak and paired-interval results. Four-fold input reconstruction is exact. CPU Ridge/SN52 refits reproduce 22,900 predictions within 7.1×10^{-13}. The original experiment executed twelve A-only selections and 72 final neural fits. Refactored neural code has CPU numerical and partial-pipeline checks, but the entire GPU matrix has not been repeated after packaging. Independently implemented checks reconstruct selection/calibration eligibility, state bindings and residual offsets. B-forbidden resolution reproduces all A query inputs across 7,043 calls (Appendix D.3). These checks use resolved histories and do not verify independent source acquisition or replication by an external researcher.

Separate reference and diagnostic artifacts reproduce 128 reference-table mean errors, fixed-state Chronos outputs and selected history-to-baseline routes (Appendix G). The supplement's executable scope is US-HHS and Japan. The four-country overview does not imply that all tasks and auxiliaries are included. Restricted AGI/NRZ virological extracts, derivatives and source metadata are excluded from the public release policy. User-side processing steps are documented, while independent raw-source reconstruction, full-roster retraining and complete public archival hosting remain unverified.

REFERENCES

Taha Aksu, Gerald Woo, Juncheng Liu, Xu Liu, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. GIFT-Eval: A benchmark for general time series forecasting model evaluation. *arXiv preprint arXiv:2410.10393*, 2024. URL https://arxiv.org/abs/2410.10393.

Abdul Fatir Ansari, Lorenzo Stella, Caner Türkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=gerNCVqqtR.

Johannes Bracher, Evan L. Ray, Tilmann Gneiting, and Nicholas G. Reich. Evaluating epidemic forecasts in an interval format. *PLOS Computational Biology*, 17(2):e1008618, 2021. doi: 10.1371/journal.pcbi.1008618.

Logan C. Brooks, David C. Farrow, Sangwon Hyun, Ryan J. Tibshirani, and Roni Rosenfeld. Nonmechanistic forecasts of seasonal influenza with iterative one-week-ahead distributions. *PLOS Computational Biology*, 14(6):e1006134, 2018. doi: 10.1371/journal.pcbi.1006134.

Centers for Disease Control and Prevention. U.s. influenza surveillance: Purpose and methods. https://www.cdc.gov/fluview/overview/index.html, n.d. Official surveillance overview; accessed 2026-09-24.

Estee Y. Cramer, Yuxin Huang, Yijin Wang, et al. The united states COVID-19 forecast hub dataset. *Scientific Data*, 9:462, 2022. doi: 10.1038/s41597-022-01517-w.

Dean Croushore and Tom Stark. A real-time data set for macroeconomists. *Journal of Econometrics*, 105(1):111–130, 2001. doi: 10.1016/S0304-4076(01)00072-0.

Songgaojun Deng, Shusen Wang, Huzefa Rangwala, Lijing Wang, and Yue Ning. Cola-GNN: Cross-location attention based graph neural networks for long-term ILI prediction. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, 2020. doi: 10.1145/3340531.3411975.

Rakshitha Godahewa, Christoph Bergmeir, Geoffrey I. Webb, Rob J. Hyndman, and Pablo Montero-Manso. Monash Time Series Forecasting Archive. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/eddea82ad2755b24c4e168c5fc2ebd40-Abstract-round2.html.

Lars Graf, Thomas Bohnstingl, Stanisław Woźniak, and Angeliki Pantazi. FlowState: Sampling-rate invariant time series foundation model with dynamic forecasting horizons. In *NeurIPS 2025 Workshop on Recent Advances in Time Series Foundation Models (BERT2S)*, 2025. URL https://openreview.net/forum?id=R50AT6nAsM.

Heikki Haario, Marko Laine, Antonietta Mira, and Eero Saksman. DRAM: Efficient adaptive MCMC. *Statistics and Computing*, 16(4):339–354, 2006. doi: 10.1007/s11222-006-9438-0.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. doi: 10.1080/00401706.1970.10488634.

Rob J. Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, 3 edition, 2021. URL https://otexts.com/fpp3/simple-methods.html.

Rob J. Hyndman and Yanan Fan. Sample quantiles in statistical packages. *The American Statistician*, 50(4):361–365, 1996. doi: 10.1080/00031305.1996.10473566.

Japan Institute for Health Security. Influenza (excluding avian influenza and novel influenza). https://id-info.jihs.go.jp/surveillance/idwr/article/influenza/article.html, n.d.a. Translated title; includes notice of the sentinel reporting change (in Japanese); accessed 2026-09-24.

Japan Institute for Health Security. Infectious diseases weekly report surveillance. https://id-info.jihs.go.jp/en/surveillance/idwr/index.html, n.d.b. Official IDWR surveillance documentation; accessed 2026-08-27.

Japan Institute for Health Security. Terms of use for the infectious disease information website, n.d.c. URL https://id-info.jihs.go.jp/usage-contract.html. Accessed 24 September 2026.

Harshavardhan Kamarthi, Lingkai Kong, Alexander Rodríguez, Chao Zhang, and B. Aditya Prakash. CAMul: Calibrated and accurate multi-view time-series forecasting. In *Proceedings of the ACM Web Conference 2022*, 2022a. doi: 10.1145/3485447.3512037.

Harshavardhan Kamarthi, Alexander Rodríguez, and B. Aditya Prakash. Back2Future: Leveraging backfill dynamics for improving real-time predictions in future. In *International Conference on Learning Representations*, 2022b. URL https://openreview.net/forum?id=L01Nn_VJ9i. OpenReview forum L01Nn_VJ9i; arXiv:2106.04420.

Craig J. McGowan, Matthew Biggerstaff, Michael Johansson, Karyn M. Apfeldorf, Michal Ben-Nun, Logan Brooks, Matteo Convertino, Madhav Erraguntla, David C. Farrow, John Freeze, Saurav Ghosh, Sangwon Hyun, Sasikiran Kandula, Joceline Lega, Yang Liu, Nicholas Michaud, Haruka Morita, Jarad Niemi, Naren Ramakrishnan, Evan L. Ray, Nicholas G. Reich, Pete Riley, Jeffrey Shaman, Ryan Tibshirani, Alessandro Vespignani, Qian Zhang, Carrie Reed, and The Influenza Forecasting Working Group. Collaborative efforts to forecast seasonal influenza in the united states, 2015–2016. *Scientific Reports*, 9:683, 2019. doi: 10.1038/s41598-018-36361-9.

Taichi Murayama, Nobuyuki Shimizu, Sumio Fujita, Shoko Wakamiya, and Eiji Aramaki. Predicting regional influenza epidemics with uncertainty estimation using commuting data in japan. *PLOS ONE*, 16(4):e0250417, 2021. doi: 10.1371/journal.pone.0250417.

Art B. Owen. Randomly permuted (t, m, s)-nets and (t, s)-sequences. In *Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, pp. 299–317. Springer, 1995. doi: 10.1007/978-1-4612-2552-2_19.

Dimitris N. Politis and Joseph P. Romano. The stationary bootstrap. *Journal of the American Statistical Association*, 89(428):1303–1313, 1994. doi: 10.1080/01621459.1994.10476870.

Public Health Agency of Canada. Respiratory virus monitoring in canada. https://www.canada.ca/en/public-health/services/diseases/flu-influenza/influenza-surveillance/about-fluwatch.html, n.d. Official surveillance and target documentation; accessed 2026-09-24.

Martina Rama, Gabriele Santin, Giulia Cencetti, Michele Tizzoni, and Bruno Lepri. Forecasting seasonal influenza epidemics with physics-informed neural networks. *Epidemics*, 55:100919, 2026. doi: 10.1016/j.epidem.2026.100919.

Nicholas G. Reich, Logan C. Brooks, Spencer J. Fox, Sasikiran Kandula, Craig J. McGowan, Evan Moore, Dave Osthus, Evan L. Ray, Abhinav Tushar, Teresa K. Yamana, Matthew Biggerstaff, Michael A. Johansson, Roni Rosenfeld, and Jeffrey Shaman. A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the united states. *Proceedings of the National Academy of Sciences*, 116(8):3146–3154, 2019. doi: 10.1073/pnas.1812594116.

Alex Reinhart, Logan Brooks, Maria Jahja, Aaron Rumack, Jingjing Tang, et al. An open repository of real-time COVID-19 indicators. *Proceedings of the National Academy of Sciences*, 118(51): e2111452118, 2021. doi: 10.1073/pnas.2111452118.

Robert Koch Institute. Arbeitsgemeinschaft influenza: Weekly report archive, n.d.a. URL https://influenza.rki.de/Wochenberichte.aspx. Accessed 25 September 2026; source access does not imply redistribution permission.

Robert Koch Institute. Arbeitsgemeinschaft influenza: Fachausdrücke. https://influenza.rki.de/Glossar.aspx, n.d.b. Official acute respiratory illness and consultation-incidence definitions; accessed 2026-09-24.

Robert Koch Institute. SurvStat@RKI 2.0: Data usage, n.d.c. URL https://survstat.rki.de/Content/Instruction/DataUsage.aspx. Accessed 24 September 2026.

Alexander Rodríguez, Jiaming Cui, Naren Ramakrishnan, Bijaya Adhikari, and B. Aditya Prakash. EINNs: Epidemiologically-informed neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12), 2023. doi: 10.1609/aaai.v37i12.26690.

Jeffrey Shaman, Alicia Karspeck, Wan Yang, James Tamerius, and Marc Lipsitch. Real-time influenza forecasts during the 2012–2013 season. *Nature Communications*, 4:2837, 2013. doi: 10.1038/ncomms3837. Article 2837.

I. M. Sobol'. On the distribution of points in a cube and the approximate evaluation of integrals. *USSR Computational Mathematics and Mathematical Physics*, 7(4):86–112, 1967. doi: 10.1016/0041-5553(67)90144-9.

Shihao Yang, Mauricio Santillana, and S. C. Kou. Accurate estimation of influenza epidemics using google search data via ARGO. *Proceedings of the National Academy of Sciences*, 112(47): 14473–14478, 2015. doi: 10.1073/pnas.1515373112.

A GUIDE TO THE SUPPLEMENT

The appendix provides data and publication rules (Appendix B), procedure definitions (Appendix C), the seven-task leaderboard and controlled US/Japan experiment (Appendix D), and information-state diagnostics (Appendix E). Reproducibility and access are documented in Appendix G. Throughout this appendix, the *code supplement* means the submitted `flu_forecast_bench_anonymous.zip`. The *internal research archive* means additional model-local tables, secondary experiments and execution logs retained by the authors but not included in that ZIP.

Distinct evaluation populations. The seven-task leaderboard compares procedure-specific recipes on shared coordinates. The controlled experiment instead shares selection/calibration cutoffs across five procedures in US/Japan. Older probability comparisons match each procedure to SN52, while model-comparison diagnostics intersect a declared group. These populations have different fitted states and support and must not be pooled. Retrospective pretrained checkpoints are separate from primary advantage-change counts. Seasonal and seed variation are reported separately.

Executable code supplement. Extract `flu_forecast_bench_anonymous.zip`, enter its directory and run:

```
uv sync --locked
uv run --locked python -m flu_forecast_bench verify
uv run --locked python -m flu_forecast_bench replay \
  --compare-reference
```

This route recalibrates frozen points and regenerates the controlled analyses. After dependency installation it needs no network, private account or original repository. Separate `prepare`, `baselines` and `train` commands reconstruct inputs and refit models. Their verified scope and the distinct seven-task aggregate records are described in Appendix G and the package README.

Files in the code supplement. Paths below are relative to the extracted `flu_forecast_bench/` directory. `FILE_GUIDE.md` maps individual source files and result tables to their roles.

File or directory	Purpose
`config/protocol.json`	Controlled US/Japan protocol and model settings
`flu_forecast_bench/`	Input reconstruction, model training, calibration and scoring source code
`data/`	Resolved histories, inputs, labels and frozen predictions
`reference/US-HHS/`, `reference/JP-PREW15/`	Expected scores, seasonal/seed values and paired intervals
`reference/leaderboard/`	Seven-task aggregate scores for inspection
`outputs/replay/`	Results generated by the replay command, including coordinate losses and missingness ledgers
`VALIDATION.json`	Completed checks and verification limits

`outputs/replay/` is created at runtime. It is not prepopulated in the ZIP.

B DATA AND TEMPORAL DEFINITIONS

B.1 DATA AND INFORMATION STATES

Every result retains its source edition, reader and task-specific recipe. The supplied histories define available data; the recorded model inputs determine what each procedure consumed. Additional numerical results and source bindings remain in the internal research archive.

The archived source-profile tables distinguish common and condition-only positions, revisions on common positions and effective leads. ISO Monday weeks are used outside the US; US tasks use MMWR Sunday weeks. Complete scheduled-origin inventories are retained in the internal research archive. Detailed source profiles are retained in the internal research archive.

Japanese 2020–21 and 2021–22 B histories contain source-native dash entries. Their numerical interpretation is withheld because the bound source audit found no explicit provider rule authorizing a numerical value; a dash is not converted to zero. These histories also contain a smaller set of positions with no available issue comparator. Thus non-numeric source entries and unavailable source rows are distinct; the provider's reason for each dash is not established by this audit.

B.1.1 CORE AND REGIONAL TARGET ROLES

JP-PREW15 retains 47 prefectures and reported influenza cases per sentinel before the 2025-W15 regime change. National contextual IASR virology is one task-global signal, not separately measured prefecture virology. US national and ten HHS regional weighted-ILI series retain percent outpatient visits and the accepted historical surveillance source role. Canadian national ILI percentages and influenza-positive laboratory counts remain different tasks. DE-LAB trains on the accepted C–E notification definition. Its 336-coordinate attachment is scoring-only and cannot enter training. The CAMul German training-population attachment is the accepted one. The COLA-GNN JP evaluator attachment is likewise separate from model history.

CANADA-PT has 13 province/territory count series. Earlier regional results and the common-history release retain separate source identities; reuse requires exact input equality. For static 2018–19 through 2023–24 reports, B selects the latest verified issue containing the reporting week, at or after the Monday following ISO week 35 of the closing year and before 2025-09-11 00:00 UTC. This study-defined maturity window is not an official finalization date. Dashboard editions and the 2025–26 fixed snapshot have their own recorded rules. The evaluator selects the target-week value after predictions are fixed, whereas model histories stop before the origin.

DE-LAENDER-ARI has 16 state consultation-rate series. First-seen evidence starts at 2023-W36; earlier publication dates are not reconstructed. Complete 2024/25 and trailing partial 2025/26 results remain separate.

B.1.2 GRAPHS AND AUXILIARY INPUTS

Graphs are procedure-specific: COLA-GNN uses geographic contact, while Murayama uses directed mobility or commuting. Canada's contact graph derives from 2021 administrative geometry and is a retrospective static input. For Germany, the two relations each have 16 nodes: boundary contact for COLA-GNN and 256 reported residence-to-workplace counts for Murayama. The latter retains within-state diagonal counts, has no suppressed or imputed cells, and reproduces published residence-column totals. The matrices are separate inputs. Availability evidence also differs: boundary contact uses a metadata-based reconstruction bound of 2024-07-17; commuting has an archive-capture public-by bound of 2024-05-22. Neither authenticates all earlier training origins. Historical graph availability requires separate evidence beyond node alignment and surveillance vintage. Each recipe retains its own temporal-use rule.

Canadian observed weather follows the existing ECCC weekly national reader and its permitted retrospective role. A constructed week-end-plus-one-day availability value is not an actual publication timestamp. US retrospective virology similarly is not origin-admissible merely because the observation week is old. Canadian laboratory subtype sums, where used, remain same-target derived history features rather than independent exogenous observations. School availability is recorded without automatically adding it to the primary models. The completed separate additive school-view experiment evaluates CAMul and EiNNs External on six fixed Japanese seasons (the separately retained school-input table); its condition-specific gains and losses do not change 2023–24 and 2024–25 membership. Metadata-only school resources remain distinct from consumed weekly channels: annual selected-attendance measures are not weekly school closures, and JP designated-city measurements are not prefecture observations. In Figure 1, Canadian virology shows influenza-A positivity, and US and Japanese virology show H3 counts. The German virology plot is omitted because its redistribution permission is unresolved. Brooklyn attendance and Tokyo school closures are different measures. The gallery documents resource coverage, not universal model use.

Overview forecast illustration. Figure 1 retains a frozen US-NAT Chronos example from origin 2021-W41. The example was selected post hoc by A/B forecast difference, without reference to forecast errors. Its native vector curves are preserved; the diagram's name and A/B colours are

adapted for consistency. It illustrates paired forecasts and shared observations, not a typical effect size or the primary split-CV result. Chronos is a retrospective pretrained comparator whose historical admissibility is not established (Appendix C.1). The auxiliary miniatures describe resource coverage; they do not imply that all primary procedures consume those channels.

Japanese overview insets. Both insets in Figure 1 show Fukuoka: 2014-W04–W07 and 2018-W52–2019-W04. Their A records precede 2014-03-03 and 2019-02-11, respectively; B uses the specified later histories. The two windows are illustrative peak selections from the same regional series, not forecasts or representative effect estimates. The committed numerical records retain source identities and publication dates.

B.1.3 JAPANESE WEATHER SOURCE AND MISSINGNESS ENCODING

Japanese weather retains numerical, missing and unavailable positions separately. The numerical values and observed masks are identical across the verified source representations. CAMul additionally distinguishes missing from unavailable in its five-channel encoding; the reported states use the corrected encoding. EiNNs uses a binary observed mask and therefore has unchanged tensors. These are retrospective pre-origin weather observations: their measurement intervals end before the origin, but historical release timestamps are not established. Results using them are conditional on that auxiliary information. The post-2025-W15 Japanese ARI regime remains outside the influenza task.

B.2 TEMPORAL SEMANTICS OF THE TASKS

This appendix states the temporal contract of every task from the frozen membership manifests and procedure recipes. Origin membership is the executable h2 coordinate manifest (rule: existing frozen paired main membership, no score-dependent selection). The origins listed are scheduled task origins; the evaluated support of each procedure is its own A/B intersection and is reported per row in Appendix C.3. Required-season completeness therefore does not mean that every scheduled origin was evaluated.

B.2.1 CORE TASK CALENDARS AND REFERENCE SEASONS

Japan uses 2023–24/2024–25 reference seasons; US national/HHS use 2016–17 through 2018–19. Canadian national ILI uses 2019–20 through 2022–23, laboratory counts use 2020–21, 2021–22 and 2023–24, and German national counts use 2019–20 through 2023–24. These original reference recipes are separate from the recent seasonal table. Origin schedules preserve real week-53 positions and typed source gaps. Exact first/last origins and schedule counts are retained in the archived temporal inventory.

Scheduled-window overlap. Seven JP origins (2024–07–08 to 2024–08–19; 329 geography–target coordinates) occur in two seasonal schedules. No procedure among the 202 accepted result rows scores the same origin–geography coordinate under two season labels. Thus scheduled overlap creates no double-counting in those reported means; evaluation support, rather than schedule alone, determines inclusion.

Shared cutoff semantics. Three quantities are stated separately for every task: the observation cutoff, the publication cutoff and the indexing that links an origin to its admissible source state. Observation weeks must end before the origin week start in all tasks. For JP-PREW15, CA-ILI, CA-LAB and DE-LAB, a source row is admissible under A only when its publication or effective-state timestamp is strictly before the origin week start; rows at or after that instant are excluded, the latest admissible publication of a week is selected, and equal-date conflicts remain unresolved (never picked). Under B the accepted later-completed edition is used without an origin publication cutoff. The origin start is a calendar date (Monday for ISO weeks, Sunday for MMWR weeks); publication evidence is recorded at day precision and no wall-clock time zone is recorded for the core tasks, so none is used as a cutoff. JP-PREW15 origins are Monday dates of the pre-2025-W15 regime; the JP membership has more scheduled origins than the 52/31 origins that most procedures evaluate, because the evaluated support is the procedure-specific A/B intersection with the evaluator mask.

US-NAT and US-HHS. US A is an *early-report state*. The source projection contains historical values only from issues at lag 0 or 1 relative to the observation week. The reader selects the latest admissible issue *within that sparse projection*; it does not reconstruct all intermediate revisions available at a later origin. Specifically, the origin epiweek indexes its own FluView issue. Reference weeks must start strictly before the origin Sunday, and selected issue dates must be at or before that origin issue date; equal-date conflicts remain unresolved. This publication cutoff can follow the nominal origin week. All 401,900 A values in the packaged primary US-HHS histories match an available lag-0/lag-1 source value, with no unmatched value at tolerance 10^{-8}. Thus US A–B comparisons can include revisions already public by the forecast cutoff, and are not estimates of the effect of using the complete real-time history. Under B the frozen later values are used, preferring current target-position rows with issue epiweek equal to the reference week. US-NAT and US-HHS share a calendar but have separate procedures, fitted states and residual banks; their original three-season schedules contain 99 and 990 h2 coordinates respectively.

B.2.2 What each condition and the evaluator select

Table 4 specifies input-edition and evaluation-target selection for all eight tasks, including the meaning of the forecast origin, the observation cutoff, the A publication selection with its comparison operator, the B edition selection with its maturity boundary, and what the separately attached evaluator selects. Two of these are version-selection conventions of this study rather than official finalization events, and are marked as such in the table; unrecorded boundaries are reported as unspecified. All tasks exclude observations from the origin week onward. ISO tasks use Monday origins, US tasks MMWR Sunday origins. Core publication evidence has day precision without an authenticated time zone; DE regional uses the stated UTC timestamp. Equal-date conflicts remain unresolved. The evaluator is a condition-neutral target-week observation attached after prediction freeze.

Table 4: Task-specific A and B selection. The executable reader and source-edition bindings retain exact operators. B denotes a fixed study edition, not official finality.

Task	A publication selection	B edition and evaluation target
JP-PREW15	Latest component effective-state timestamp strictly before origin.	Fixed later component edition; pre-2025-W15 influenza cases per reporting sentinel.
US-NAT / HHS	Early-report projection (lag-0/lag-1 editions only). Latest recorded issue dated at or before the origin-epiweek issue date; reference week starts before origin.	Frozen retrospective retrieval, preferring current target-position rows with issue epiweek equal to reference week. No runtime requery; not official FluSight truth. Target: weighted ILI percent.
CA-ILI / LAB	Latest source publication strictly before origin.	Fixed later national edition; respectively ILI consultation percentage and influenza-positive specimen count.
DE-LAB	Latest report edition strictly before origin.	Fixed later C–E notification history; evaluator uses the same target definition.
CANADA-PT	Latest known-public-by edition admissible at the origin, retaining missing positions.	Static-report maturity window defined above for 2018–24; recorded dashboard-era edition for 2024–25 and fixed snapshot for 2025–26. Target: province/territory positive detections.
DE-ARI	First-seen timestamp no later than origin Sunday 00:00 UTC; evidence begins 2023-W36.	Fixed retrospective state snapshot; state ARI consultations per 100,000. Targets beyond the same-season evaluator remain unavailable.

One coordinate in dates. US-NAT origin 2016-W40 spans October 2–8, with target 2016-W42 (October 16–22). Observation weeks must start before October 2, so the last eligible week is W39; the associated FluView issue is dated October 14. A therefore uses W39's value 1.19803 from that issue. The last-observation-to-target distance is three weeks. October 14 is the information

cutoff; the forecast is not claimed to have been issued on October 2. Thus nominal h2 is a calendar-coordinate distance, not a two-week lead measured from operational issuance.

B.2.3 PROCEDURE LIFECYCLES AND EFFECTIVE LEAD

Nominal h2 means target week = origin week $+2$ in the calendar of the task: it is an index difference between two coordinates, not a promise that the target lies two weeks after the last observation or after the publication cutoff, and those distances differ by country (the dated US-NAT coordinate above has a three-week endpoint-derived lead behind nominal h2). The effective lead is the calendar distance, in whole weeks, from the week of the last observation the procedure actually consumed to the target week. In the regular case the consumed context ends at origin-1, so the effective lead is three weeks; a delayed endpoint gives a lead above three. This calendar lead is distinct from the zero-based decoder index of a multi-step decoder (index 2 addresses the origin$+2$ step of a decoder started at the origin week) and from the SN52 reference age of 52 weeks, which is the position of the selected seasonal value rather than an autoregressive step. Appendix C.1 explains each procedure; the procedure records in the internal research archive retains exact fitted-state, development, selection and calibration settings. The same model name can designate different task-specific recipes.

B.2.4 REGIONAL TASK SCHEDULES

The original Canadian regional heldout season is 2025–26, with 41 scheduled origins from 2025-W34 to 2026-W22. Earlier Canadian evaluations have their own cohorts. The German native recipe uses 2022–23 for initial training, 2023–24 for development and complete 2024–25 for holdout; trailing 2025–26 is separate. h2 schedules contain 50 and 44 origins, respectively; h8 contains 44 and 38. German A selection uses the recorded origin-Sunday 00:00 UTC first-seen cutoff, whereas the nominal origin is the ISO Monday coordinate. Exact schedules and horizon-specific choices remain in the archived temporal inventory; no additional output is implied by source availability.

B.2.5 TEMPORAL SCOPE: SOURCE HISTORY, ADJUDICATED TASK-SEASONS, FROZEN EVALUATION AND THE PRIMARY TEMPORAL PANEL

Four temporal extents differ: source-history coverage, seasons with admissible paired inputs and evaluator, seasons actually evaluated by a procedure, and seasons assigned to the reference/anchor/latest analyses. Model execution is recorded separately from input availability. The balanced anchors are 2018–19, 2020–21 and 2021–22; high-revision and latest roles are task-specific. They can overlap reference seasons, but the recipe and support still identify separate experiments. The archived temporal-scope inventory records source coverage, execution and exact task-season availability separately.

Data-side authorities and exclusions. The reported rows retain the source version and reader under which they were generated, and the Canadian regional temporal scores use a separately validated task-specific scoring authority that changes the post-freeze evaluator and scoring mask, not the model inputs; exact version identities are retained in the internal research archive. For US-HHS 2003-04 to 2008-09, no region-level condition-A histories are archived (the B state exists; A is not reconstructed). JP-PREW15 2015-16 and 2016-17 have both information states but no accepted evaluator or scoring mask. CA-ILI 2015-16 and CA-LAB 2015-16 have neither state; CA-LAB 2016-17 has no Operational state; CA-LAB 2022-23 has inputs whose evaluation mask contains no eligible coordinates. CA-ILI 2023-24 has 32 evaluable h2 coordinates; each procedure is assessed on its available paired subset. CA-ILI 2024-25 has no supplied evaluation batch. CA-ILI, CA-LAB and DE-LAB have no validated evaluation coordinates for nominal horizon h8 in most seasons, so h8 is never implied from h1–h4. DE-LAENDER-ARI 2022-23 has no first-seen publication authority before 2023-W36 and is excluded by authority as retrospective-only; no publication timestamp is invented. The post-2025-W15 Japanese sentinel regime (JP-POSTW15) is a separate target epoch outside the tasks included in this release and is never merged with JP-PREW15.

Canadian regional evaluator. The task uses 2018–19 for development; 2019–20 through 2024–25 are evaluation candidates, with 2020–21, 2021–22 and 2024–25 forming its task-native panel. The original regional heldout season is 2025–26. Evaluators use the fixed target-week B edition,

with unresolved or unavailable values excluded by the scoring mask. The 2024–25 dashboard-era edition has 624 nonmissing target cells; the 2025–26 reference retains its original 624-row snapshot. Evaluator attachment changes scored support, not frozen input histories or forecasts. Exact edition-selection bounds are specified above and in the supplied reader.

Secondary temporal analyses. The historical anchor and metadata-selected high-revision analyses retain separate recipes and eligibility rules in the internal research archive. They are not pooled with the recent leaderboards or controlled experiment.

C PROCEDURES, PROBABILITY AND UNCERTAINTY

C.1 FORECASTING PROCEDURES

Each reported numerical row is linked to a task-specific procedure recipe that specifies its fit, development and selection rule, seed policy, probability route, exact quantile grid and required seasons; exact row-level mappings are retained in the internal research archive. Two examples of what those recipes fix: the target-only EiNNs controls on US-NAT, US-HHS and DE-LAB use one fixed 60-step schedule (20 warm-up and 40 joint-continuation steps) in every condition–season–seed cell with no development sweep, and Ridge Direct and Ridge Recursive on JP-PREW15 and US-HHS use one task-wide penalty per procedure selected on Operational development origins only (the procedure records in the internal research archive). Task calendars, cutoffs and origin schedules are in Appendix B.2; residual probability routes in Appendix C.2. The following summarizes the executable mechanisms. No shared fitted state is implied by a shared model name.

Baselines. Repeat Last selects the latest admissible finite observation, unless a later unresolved required position blocks that choice. SN52 requires the exact target calendar position minus 52 and has no nearby-week or persistence fallback. Neither model has fitted weights. Their point support is independent of residual-distribution qualification. R outputs use the declared development residual bank without changing frozen points.

Ridge. Direct and recursive versions use their declared recent-lag and target-aligned seasonal features over an uncompressed 104-position window. A and B have separately fitted normalization and preseason coefficients; the intercept is unpenalized. The candidate regularization set is $\{0.01, 0.1, 1, 10, 100\}$ and hyperparameters are selected using condition-A development data, with ties resolved in favor of the larger penalty. Direct models use separate training targets and coefficients for each forecast horizon; recursive inference updates its autoregressive recent-lag state while keeping observed seasonal references. Source-specific history/label eligibility and minimum data requirements for fitting are retained, rather than imposing a new cross-task minimum. The $\alpha = 1$ fallback belongs to the routes whose own freeze records it and is not a rule of every Ridge procedure; the procedure records in the internal research archive list the recorded selection of every Ridge cell. The regional German transfer uses existing national-DE selections ($\alpha = 10$ for h2, 100 for h8), explicitly not local tuning; that transfer statement is about the German regional cells and is not the rule of the Canadian ones.

Canada Ridge Direct, horizon 8. No complete A development cases exist at h8 in the registered 2018–19 development season. The separate h8 recipe therefore selects on the latest four eligible pre-heldout origins, 2025-W09/W11/W13/W15, chosen by feature/label availability. There are 51 complete cases; the declared five-value penalty grid selects $\alpha = 100$. All development targets precede the first heldout origin (2025-W34); their labels come from its B snapshot. The h2 recipe retains $\alpha = 0.01$. Exact development scores and bindings remain in the internal research archive.

Pretrained inference. Chronos uses the fixed T5-small checkpoint, a trailing window of up to 156 time steps on the task's calendar, nine decoded steps and 2,000 samples. FlowState uses its fixed research checkpoint, native missing masks, weekly scaling and nine-step mapping; only the deciles returned by the model are evaluated. Decoder selection respects the actual endpoint, with no heldout tuning. Chronos was published in March 2024 and FlowState r1.1 in March 2026. These retrospective checkpoints are excluded from the primary selection analysis. Record-level overlap between

their public/synthetic training collections and surveillance series remains unverified, regardless of checkpoint publication dates.

CAMul. Each view retains its native encoder and latent uncertainty; query/reference relations and trainable input-dependent fusion feed the predictive decoder. Calendar, weather and virology are added only through existing architectural view mechanisms. Configurations with fewer and more input views differ in their model-selection procedures and training budgets. The maximal recipe uses the transferred P07 configuration with batch size 10 and five training epochs where declared, with condition-owned normalization, fitted weights and reference banks. Native uncertainty integrates 2,000 replicate-indexed latent/reference and observation draws; it does not draw 2,000 observations conditional on just one latent realization. Type-7 quantiles are computed from the frozen native samples. German BASE retains target/calendar/weather; virology is additional. Weather observations are missing from the early portions of some condition-A training sets; the original training cutoffs are retained.

Graph models. COLA-GNN retains its recurrent temporal encoder and graph attention using the geographical graph specified for the task. Its declared core settings include hidden width 10, dropout 0.2, Adam learning rate 0.001, weight decay 0.0005, batch size 32 and seed 2026. Training durations selected on development data are distinguished from durations transferred to regional tasks and held fixed. For Canadian h8 forecasts, a separate model is fitted for each required effective lead for 1,500 epochs, with separate weights and scalers for A and B; it is not labeled task-tuned. Murayama retains directed mobility and the original temporal/spatial implementation. For the German states the two configurations use different graph datasets and preprocessing steps: COLA-GNN consumes the 2023-01-01 boundary-contact relation, where the explicit diagonal of the source is distinct from the symmetric degree scaling the model applies, and Murayama consumes the 2023-06 commuting relation under global-maximum scaling followed by the existing entering/leaving transformation, with the source diagonal preserved. The recorded graph temporal-use mode belongs to the relation and not to the surveillance history: both relations are recorded retrospective at the 2023-W40 origin and operational at the 2024-W40 origin. An operational mode at evaluation origins is an admissibility record for that origin; for the boundary-contact asset it does not amount to archive-verified byte evidence, and historical publication dates during development and training have not been verified. The recorded permission to use the relation is unchanged. The recovered point predictions are retained without modification. Native uncertainty combines dropout and development residual terms at the same decoder lead and calendar phase; JP's fixed window is five, HHS's selected window three.

German Murayama, horizon 8. The unchanged dual-GRU checkpoint has context 26 and training horizon nine. Delayed A endpoints require inference to steps 10–12: 800 of 832 rows use lead ten. Recursive inference extends the decoder without refitting or teacher forcing, retaining 100 dropout draws and the existing seed rule. The first nine steps reproduce 5,200 original comparisons; h2 differences across devices are at most 6.1×10^{-5}. Extended outputs are point-only. Complete 2024/25 and partial 2025/26 are separate, as are this implementation and Canadian Murayama v1. Japanese and HHS window sensitivities retain their own recipes; the HHS selection has only one shared development origin.

EiNNs. Target-only controls and External variants remain distinct. The target-only variant receives target-series values and indicators of whether they are observed. The External variant adds weather and virological measurements, their observation indicators, and calendar features as inputs to the feature module. A causal suffix of at most 52 positions is preserved. Training stages fit time-data, time/dynamics constraints, feature alignment, then shared output/parameter gradient transfer with the prescribed frozen encoders. Parameters are not replaced by metadata-only columns. Models are fitted separately for each condition, season and seed (1729 and 2027), and residuals are computed for each fitted model; losses are computed separately for each seed and then averaged at each evaluation coordinate, without averaging quantiles across seeds. Development evaluation compares training budgets of 5, 10 and 20 full passes through the training data; JP's selected five-pass budget transfers to its richer recipe without claiming weather-based early tuning. The residual adapter for each fitted model uses that model's out-of-sample residuals.

Canadian EiNNs probability. Each condition fits once at the earliest development origin with a valid scale, fit anchors and an observed label at every declared horizon (1, 2, 4, 8). A qualifies at

2018-W34; B at W44 (cutoff 2018-10-28), after three origins without B sources and subsequent missing h8 labels. The unchanged two-seed recipe (1729/2027, five epochs per stage, calendar only) gives B queries W44–W50. Signed residual states have 91 B coordinates (seven origins, 13 regions) and 104 A coordinates (eight origins), separately by seed at leads 14/21/35/63 days. Minimum support remains 30. Reference values come from each condition's first heldout-season input snapshot, restricted to development target weeks; no heldout label enters calibration. Median quantiles equal frozen points, whose weights, scalers and input identities are unchanged.

SIR-INN. The scientific procedure and the numerical method that approximates it are stated separately, because only the second changed.

Scientific procedure (unchanged). The following defines the estimand under the benchmark configuration, which differs from that of the source paper. Each unit is one (task, geography, origin, condition) conditioning problem with three unknowns $\theta = (\beta, \gamma, \tau_0)$: transmission rate, recovery rate and the time offset of the epidemic in days. A susceptible-depletion surrogate network, trained on synthetic dynamics before the benchmark and frozen, maps (t, β, γ) to a susceptible fraction S; no surrogate training, fine-tuning or checkpoint selection happens inside the benchmark. Weekly incidence on the span of m observed weeks plus the lead is the negative difference of that surrogate on consecutive weekly nodes, indexed from zero,

$$z_j(\theta) = S(\tau_0 + 7j) - S(\tau_0 + 7(j+1)), \qquad j = 0, \ldots, m + \text{lead} - 1,$$

where S is the frozen susceptible-fraction surrogate, m is the number of observed weeks and lead is the forecast offset in weeks. Observations $y_0, \ldots, y_{m-1}$ are paired with $z_0, \ldots, z_{m-1}$; the forecast position is $z_{m+\text{lead}-1}$. The prior is uniform on the rectangle $\beta \in [0.12, 0.45]$, $\gamma \in [1/12, 0.4]$, $\tau_0 \in [0, 400]$, restricted by $\beta - \gamma \le 0.135$ and by the time-domain condition $\tau_0 + 7(m + \text{lead}) \le 595$ days, which keeps the whole observed-plus-forecast span inside the surrogate's trained domain. The support restriction requires every z_j to be finite and nonnegative over the entire observed and forecast period. Parameter values that violate it receive log density $-\infty$. Both adaptations scale the latent incidence by a fixed plug-in nuisance estimated on Operational development data and reused unchanged in both conditions, so differences between A and B cannot arise from condition-specific refitting of the nuisance parameter. They differ in the observation model. The count route sets $\mu_j = e\, z_j$ with a frozen exposure e and uses a Poisson likelihood on nonnegative integer counts,

$$\log p(y \mid \theta) = \sum_{j=0}^{m-1} \left(y_j \log \mu_j - \mu_j - \log \Gamma(y_j + 1) \right),$$

where y_j is the weekly count, μ_j its conditional mean and Γ the gamma function. Units with noninteger observations are rejected. The rate route sets $\mu_j = s\, z_j$ with a frozen scale s and uses a Gaussian of frozen standard deviation σ truncated to the native admissible interval $[0, u]$, so each term carries its truncation normalizer $\log(\Phi((u - \mu_j)/\sigma) - \Phi(-\mu_j/\sigma))$ and no pseudo-counts are introduced. The forecast functional is the same scaling applied to the last latent position of the span, $f(\theta) = s\, z_{m+\text{lead}-1}$ (or $e\, z_{m+\text{lead}-1}$). The input windows are consecutive $M5$ — the count route permits $M4$ only at its declared first scheduled round, not as a general missingness fallback — and the leads and the quantile grid are those of the task. The target definitions imply different uncertainty quantities. Count-task quantiles describe posterior uncertainty in expected incidence $f(\theta)$, rather than realised counts. Rate-task quantiles integrate truncated-Gaussian observation noise over the posterior. These two uncertainty targets are never pooled into one interval.

Numerical approximation used for the reference results. The reported rows approximate that identical target by the self-normalised importance approximation with replicate-precision assessment and a frozen refinement policy (Appendix F); its per-unit and propagated precision is reported in the archived same-target numerical-comparison table. The earlier fixed-budget analysis instead ran one DRAM chain (Haario et al., 2006) per unit with 10,000 iterations, 9,000 burn-in and 1,000 retained draws, its condition-specific seed derived from the procedure/unit identity without retries; the earlier estimates are reported separately from the higher-precision estimates. The Poisson-mixture predictive sensitivity and the independent-chain validation belong to that earlier analysis and replace neither procedure.

C.1.1 PROBABILITY GRIDS AND RESULT ROLES

The historical score uses one of two declared grids:

$$\mathcal{Q}_{\text{conv}} = \{.01, .025, .05, .10, .15, .20, .25, .30, .35, .40, .45, .50, .55, .60, .65, .70, .75, .80, .85, .90, .95, .975, .99\},$$

where $\mathcal{Q}_{\text{conv}}$ is the conventional 23-level grid, or

$$\mathcal{Q}_{\text{JP}} = \{.01, .025, .05, .075, .10, .15, .20, .25, .30, .35, .40, .50, .60, .65, .70, .75, .80, .85, .90, .925, .95, .975, .99\},$$

where $\mathcal{Q}_{\text{JP}}$ is the native grid for JP COLA-GNN and the JP baseline R routes. Each historical score averages doubled pinball loss at its exact 23 probabilities, equivalently using its matching central-interval WIS weights. WIS uses only .1 through .9, each with equal doubled-pinball weight $1/9$. Native (N), residual-adapter (R), posterior-mean and observation-predictive targets are explicitly separate. In particular, FlowState provides nine quantiles, which are not expanded to the 23-quantile grid.

C.2 CONSTRUCTION OF PROBABILISTIC FORECASTS

Probability output may be native or adapted from frozen points. Each result record specifies the method used to obtain its predictive quantiles. Probability outputs require separate qualification from point forecasts. Residual states use development observations only.

Centered signed-residual adapter. Within a procedure/task/condition/recorded-lead stratum and one seed, at least 30 unique development coordinates are required. The centered signed-residual adapter defines

$$q_\tau(x) = c(x) + Q_\tau^{(7)}(R) - Q_{0.5}^{(7)}(R), \qquad q_{0.5}(x) = c(x),$$

where $c(x)$ is the frozen point forecast for input x, $R = y - c$ contains signed out-of-sample development residuals and $Q_\tau^{(7)}$ is their Type-7 empirical quantile. There is no clipping, borrowing across conditions, or fallback for missing support. Japanese baselines and COLA-GNN use the second native grid specified above; US-national Ridge uses this signed adapter, rather than the other historical Ridge routes below. Table 5 records membership.

Historical Ridge alternatives. Ridge Recursive's specified routes share A absolute residuals across conditions. Each central coverage level uses the Type-7 residual-magnitude quantile as a symmetric radius around the frozen point, followed by native bounds (nonnegative; at most 100 for US-HHS). Minimum support is 30. Direct Ridge instead adds 1,000 sampled signed, condition-specific residuals to the frozen point (seed 20260813), taking Type-7 quantiles without recentering or clipping. These are predictive-error draws, not a bootstrap sampling distribution of an estimated quantile. CA-LAB Direct pools by condition; Japanese and US-HHS Direct pool by condition/horizon, and US-HHS Recursive by A/horizon.

Residual keys and input age. EiNNs External's recorded 21-day stratum is nominal, not conditioned on consumed input age. In its reference input audit it agrees with US-NAT/HHS's three-week endpoint distance, but differs in CA-ILI (4–48 weeks), CA-LAB (5–47), DE-LAB (4–36) and Japan (4–6). The eight recent JP/US/DE probability units also use this declared key. Their scores evaluate the nominal-lead recipe. Calibration conditional on consumed input age remains untested. Canadian regional EiNNs instead has actual 14/21/35/63-day strata, separately by condition and seed. These implementations remain distinct.

Native predictive routes. Japanese Murayama combines the population variance of 100 dropout draws with the uncentered mean squared development residual. Calibration is stratified by condition, native lead, geography and the input endpoint's 52-week phase, retaining the latest origin per phase and excluding selection targets. The normal distribution is centered on the point forecast. Its interval width incorporates the uncentered mean squared development residual, which includes a squared-bias component. SIR quantiles come from its posterior approximation: latent expected incidence for native count output, observation-inclusive prediction for rate output. Poisson observation-predictive count extensions remain separately labeled. CAMul and pretrained procedures use their frozen native sample distributions.

Identical points need not imply identical distributions. All 3,618 recovered Japanese SN52 input coordinates select identical values and calendar positions; 3,242 belong to the paired probability cohort. The state crossing and development-support diagnostic quantify this case in Tables 22 and 23. Elsewhere identical-value counts are CA-ILI 23/86, CA-LAB 22/56, DE-LAB 7/147, US-HHS 0/930 and US-NAT 0/93, despite identical selected positions. Their contrasts can include value and state changes. On CANADA-PT 2024–25's 196-coordinate common panel, only SN52's 0.1 quantile differs (A is three units higher); the coordinate WIS difference is constantly -0.0667, explaining its degenerate bootstrap interval.

Table 5: Residual-route membership in the reference recipes. Core = JP, US-NAT, US-HHS, CA-ILI, CA-LAB, DE-LAB. Each rule is defined immediately above; task/condition/seed strata remain separate.

Route	Members
Signed, condition-specific	EiNNs Target and External: all six core tasks. COLA-GNN: US-HHS. Repeat Last: DE-LAB, US-NAT/HHS. Ridge Direct/Recursive: US-NAT. SN52: CA-ILI/LAB, DE-LAB, US-NAT/HHS.
Shared A absolute residuals, bounded	Ridge Recursive: CA-ILI/LAB, JP, US-HHS. US-HHS pools A by horizon; other recipes retain their recorded pool.
Signed residual draws	Ridge Direct: CA-LAB (condition pool), JP and US-HHS (condition/horizon pool). CA-LAB uses its frozen rolling pseudo-origin development.
Japanese centered 23-level route	JP COLA-GNN, Repeat Last and SN52. The second native grid includes .075/.925 instead of .45/.55; exact-distribution grid checks preserve this identity.

C.3 Common-support probability inference

All common-panel losses use the same coordinate intersection for every group member and both conditions. The bootstrap resamples whole origins with their geographies, retains season strata and shares the resampling indices across methods. Retained-origin resampling uses the sequence of available origins; calendar-gap resampling respects gaps in the native schedule. Bootstrap ratios are recorded as undefined when the reference loss is zero. The complete grid, reference matching and zero-denominator counts are retained with the frozen numerical records. The primary recent score is the exact nine-decile score in Equation 1.

D Seasonal results and complete coverage

D.1 Seven task-local leaderboards

Table 2 uses frozen point forecasts on seven task-local panels. The first six panels retain the earlier reference recipes; the seventh adds CA-LAB. For the first six, candidate seasons are 2023–24 and 2024–25, and the roster is the union of procedures with an authenticated paired result in either season. A season qualifies when every roster member is present and the all-member A/B intersection contains at least eight origins. Eligibility does not depend on losses or ranks. The roster includes twelve procedures in total, with 74 procedure–task pairs across the seven panels: 60 nonpretrained and fourteen retrospective-checkpoint pairs. Main-table rows follow a fixed procedure order, not their observed rank.

Scores, intervals and ranks. Within a task–season, all members and both conditions share targets and coordinates. Where seeds exist, errors are averaged on common seed support; forecasts are not first averaged. Regional losses are averaged within origin, origins within season, and seasons equally. Relative MAE divides this mean by SN52's mean on the same coordinates. The stationary origin-block bootstrap synchronizes resamples across models and conditions, stratifies by season, and uses 10,000 replicates and block lengths 4/8/12 under retained-origin and calendar-gap rules. Main-table intervals use retained-origin blocks of length eight; colouring requires both rules to lie on the same side of one. Intervals are pointwise and conditional on observed seasons, not simultaneous superiority claims. Rank ranges exclude Chronos and FlowState; the median ensemble pools

Table 6: Task-local leaderboard support. Counts within a cell follow the listed season order. These panels use their original fitting recipes and differ from the controlled experiment in Table 3.

Task	Seasons	Procedures	Origins	Coordinates
US-HHS	2023–24, 2024–25	12	12, 12	120, 120
US-NAT	2023–24, 2024–25	10	27, 27	27, 27
JP-PREW15	2023–24, 2024–25	12	52, 31	1,876, 1,457
DE-LAB	2023–24, 2024–25	10	26, 32	26, 32
DE-LAENDER-ARI	2024–25	11	50	800
CANADA-PT	2024–25	9	16	196
CA-LAB	2024–25, 2025–26	10	12, 26	12, 26

nonpretrained point forecasts post hoc and is outside the member ranks. All reported figures are checked against bound score and uncertainty records; six-panel mean losses are additionally recomputed from coordinate losses. The CA-LAB bootstrap and full-roster training are not rerun in this integration.

CA-LAB extension. This post-hoc panel uses archived RVDSS weekly editions. Source construction retains 15 origins from 18 editions in 2024–25 and 29 origins from 36 editions in 2025–26; six undated July–August 2024 captures are excluded. The supplied audit reports exact agreement with the four earlier 2024–25 A/B histories and evaluation observations. Twelve and 26 origins survive the ten-procedure intersection. Truth is the latest visible archived national row. The batch scorer is checked against the earlier accepted scorer because its fixed coordinate registry omits the new coordinates. Graph procedures do not apply to this single series. EiNNs External uses national weather extended through September 2026 as a retrospective covariate. These source and execution records are inherited from the coauthor's extension; independent reacquisition is not established by the present display audit.

Other coverage qualifications. US-HHS contains twelve February–April origins per season after intersecting the graph and baseline outputs, not complete seasons. CANADA-PT and German states retain one season because their full rosters are absent in the other candidate season. Canadian EiNNs uses the corrected A-only eligibility variant described below; earlier probability outputs retain separate identities. CA-ILI has no qualified recent common panel. A new submission must cover the declared coordinates without shrinking other models' support. Procedure-specific fitting, auxiliary availability and retrospective checkpoints remain part of each recipe; these panels are separate from the disjoint-selection/calibration experiment.

Table 7: Retrospective pretrained checkpoints on the same coordinates as Table 2. Each score is MAE relative to SN52 in its own condition. These checkpoints remain outside the primary ranking; record-level pretraining independence is unverified.

Procedure	Input	JP	US-HHS	US-NAT	DE-LAB	DE-ARI	CA-PT	CA-LAB
Chronos	A	0.776	1.052	0.733	0.760	0.977	1.032	0.474
	B	0.572	0.930	0.719	0.535	1.016	0.691	0.397
FlowState	A	0.682	0.655	0.667	0.484	0.878	0.863	0.449
	B	0.540	0.595	0.663	0.547	0.846	0.568	0.334

Packaged leaderboard files. The code supplement's `reference/leaderboard/` directory contains `scores.csv` (unrounded A/B aggregates) and `display.csv` (main-table scores, intervals, ranks and colours). These files support aggregate inspection, without seven-task retraining or bootstrap regeneration. Earlier six-task native-unit results follow, and `scores.csv` includes the CA-LAB extension.

D.2 Common-coordinate point leaderboard

Table 8 uses the R058 frozen point panels. We retain the declared task-specific roster and admit a season only when every member has A/B predictions on at least eight common origins. Eligibility

Table 8: Common-coordinate h2 point forecasting. MAE: equal-season mean, with ± between-season SD below; rMAE: the same mean divided by SN52. A: publication rule; B: later edition. **Bold**: rMAE below one.

	US: national				US: HHS regions				Japan: prefectures			
Origins	23-24: 27; 24-25: 27				23-24: 12; 24-25: 12				23-24: 52; 24-25: 31			
Baseline	MAE ↓		rMAE ↓		MAE ↓		rMAE ↓		MAE ↓		rMAE ↓	
	A	B	A	B	A	B	A	B	A	B	A	B
Repeat Last (Hyndman & Athanasopoulos, 2021)	1.06 ±0.28	1.06 ±0.30	**0.94**	**0.93**	0.95 ±0.46	0.93 ±0.48	1.74	1.63	9.47 ±2.71	7.32 ±2.57	**0.79**	**0.61**
SN52 (Hyndman & Athanasopoulos, 2021)	1.13 ±0.26	1.14 ±0.27	1.00	1.00	0.55 ±0.13	0.57 ±0.11	1.00	1.00	12.06 ±4.08	12.06 ±4.08	1.00	1.00
Ridge Direct (Hoerl & Kennard, 1970)	0.69 ±0.14	0.69 ±0.15	**0.61**	**0.61**	0.41 ±0.14	0.34 ±0.11	**0.75**	**0.59**	9.67 ±2.20	8.31 ±1.95	**0.80**	**0.69**
Ridge Recursive (Hoerl & Kennard, 1970)	0.72 ±0.20	0.74 ±0.21	**0.64**	**0.65**	0.44 ±0.16	0.35 ±0.14	**0.80**	**0.62**	9.88 ±0.61	7.76 ±1.57	**0.82**	**0.64**
Chronos^R (Ansari et al., 2024)	0.82 ±0.08	0.82 ±0.10	**0.73**	**0.72**	0.58 ±0.13	0.53 ±0.13	1.05	**0.93**	9.36 ±1.29	6.90 ±2.24	**0.78**	**0.57**
FlowState^R (Graf et al., 2025)	0.75 ±0.02	0.76 ±0.03	**0.67**	**0.66**	0.36 ±0.02	0.34 ±0.01	**0.65**	**0.60**	8.22 ±1.38	6.52 ±1.60	**0.68**	**0.54**
COLA-GNN (Deng et al., 2020)	n/a	n/a	n/a	n/a	0.33 ±0.01	0.33 ±0.02	**0.60**	**0.58**	7.92 ±0.46	5.90 ±0.43	**0.66**	**0.49**
Murayama (Murayama et al., 2021)	n/a	n/a	n/a	n/a	0.72 ±0.05	0.57 ±0.04	1.31	**0.99**	7.61 ±0.66	4.69 ±0.23	**0.63**	**0.39**
CAMul (Kamarthi et al., 2022a)	2.27 ±0.62	2.62 ±0.14	2.02	2.30	0.73 ±0.08	2.30 ±0.23	1.34	4.01	9.38 ±0.54	9.44 ±1.03	**0.78**	**0.78**
EiNNs Target (Rodríguez et al., 2023)	2.77 ±0.26	2.79 ±0.26	2.46	2.45	1.46 ±0.12	1.52 ±0.12	2.67	2.65	9.39 ±0.07	9.42 ±0.02	**0.78**	**0.78**
EiNNs External (Rodríguez et al., 2023)	1.99 ±0.29	2.02 ±0.28	1.77	1.77	0.53 ±0.12	0.56 ±0.11	**0.97**	**0.97**	9.55 ±0.06	9.31 ±0.22	**0.79**	**0.77**
SIR-INN (Rama et al., 2026)	1.09 ±0.13	1.11 ±0.14	**0.97**	**0.98**	0.69 ±0.20	0.70 ±0.20	1.26	1.21	9.08 ±2.74	8.08 ±2.86	**0.75**	**0.67**

	Canada: provinces				Germany: laboratory				Germany: states			
Origins	24-25: 16				23-24: 26; 24-25: 32				24-25: 50			
Baseline	MAE ↓		rMAE ↓		MAE ↓		rMAE ↓		MAE ↓		rMAE ↓	
	A	B	A	B	A	B	A	B	A	B	A	B
Repeat Last (Hyndman & Athanasopoulos, 2021)	458	297	1.36	**0.88**	4.81k ±1.19k	4.53k ±1.70k	**0.87**	**0.82**	413	401	1.08	1.12
SN52 (Hyndman & Athanasopoulos, 2021)	336	336	1.00	1.00	5.52k ±1.10k	5.54k ±1.29k	1.00	1.00	384	357	1.00	1.00
Ridge Direct (Hoerl & Kennard, 1970)	284	198	**0.85**	**0.59**	5.82k ±1.91k	4.24k ±2.11k	1.05	**0.77**	558	427	1.45	1.19
Ridge Recursive (Hoerl & Kennard, 1970)	408	212	1.22	**0.63**	6.58k ±1.81k	4.12k ±2.49k	1.19	**0.74**	558	435	1.45	1.22
Chronos^R (Ansari et al., 2024)	347	232	1.03	**0.69**	4.20k ±244	2.96k ±1.07k	**0.76**	**0.53**	375	363	**0.98**	1.02
FlowState^R (Graf et al., 2025)	290	191	**0.86**	**0.57**	2.68k ±2.36k	3.03k ±1.59k	**0.48**	**0.55**	337	302	**0.88**	**0.85**
COLA-GNN (Deng et al., 2020)	†	†	†	†	n/a	n/a	n/a	n/a	590	572	1.54	1.60
Murayama (Murayama et al., 2021)	†	†	†	†	n/a	n/a	n/a	n/a	594	581	1.55	1.63
CAMul (Kamarthi et al., 2022a)	379	320	1.13	**0.95**	4.27k ±3.92k	5.80k ±5.30k	**0.77**	1.05	569	674	1.48	1.89
EiNNs Target (Rodríguez et al., 2023)	462	462	1.38	1.38	7.74k ±4.82k	7.42k ±4.94k	1.40	1.34	821	556	2.14	1.56
EiNNs External (Rodríguez et al., 2023)	–	–	–	–	7.98k ±4.76k	7.63k ±4.90k	1.44	1.38	–	–	–	–
SIR-INN (Rama et al., 2026)	924	455	2.75	1.36	10.66k ±814	14.28k ±4.93k	1.93	2.58	537	465	1.40	1.30

Every reported procedure and A/B share coordinates within each task. Canada provinces and Germany states use only 2024–25 (no SD); US-HHS covers February–April. MAE retains native units; $k = 10^3$. rMAE normalizes MAE; bold is descriptive, not significance. R: retrospective checkpoint; n/a: outside the single-series roster; †: separate partial-observation recipe; –: unavailable output. Coverage, seasonal values and missing-cell reasons: Appendix D.2. WIS uses separate probability cohorts (Figure 8; Appendix D).

depends on output coverage, not the observed score. We do not remove a poorly performing member to improve the table. National Canadian tasks do not qualify for this all-procedure comparison; their available results remain in the internal research archive. Retrospective checkpoints stay in the common intersection but are identified separately from the primary non-pretrained roster.

The coordinate key comprises task, native calendar, season, forecast origin, geography, h2 and target week. Losses are averaged equally across geographies within an origin, origins within a season, and included seasons. The two-season SD is the sample SD of the two seasonal MAEs; single-season panels have no SD. Relative MAE divides the mean model loss by the mean SN52 loss under the same condition and coordinates. It is an interpretable rescaling of MAE, not additional independent evidence. Seed handling belongs to the declared recipe; these SDs do not measure training-seed variation. Table 9 reports dates and retained coverage. The US-HHS intersection is a February–April window in each season, so its improvements cannot establish performance over the full epidemic cycle.

Corrected Canadian EiNNs input admission. The provincial EiNNs Target row uses the calendar-transfer5 A-only variant. An earlier variant used B availability when admitting A input windows; the current variant bases A admission only on A. Its two seeds (1729 and 2027) are scored separately and their absolute errors averaged at each coordinate before hierarchical aggregation. The verification script binds every scored point to its pre-score prediction freeze and checks the absolute errors. The model-local paired support contains 447 and 507 coordinates in 2023–24 and 2024–25, respectively; the all-procedure panel retains only 196 coordinates in 2024–25. On those common coordinates, MAE is 462.126 for A and 462.323 for B. A later 2024–25 probability adapter for the corrected recipe is reported separately in Appendix D.6; the earlier Canadian EiNNs WIS values in the internal research archive describe the earlier recipe and are not its scores.

Comparison scope. Common coordinates, horizon, evaluator and aggregation are held fixed, while architectures, auxiliary inputs, fit windows, update rules and selection budgets remain recipe-specific. The table compares complete forecasting procedures, rather than isolating architectural effects under identical tuning. WIS needs complete predictive quantiles and its own declared common panel; point predictions alone cannot supply that comparison. The separate probability analyses retain their original support and uncertainty qualifications.

Missing cells are not a single failure mode. The declared single-series rosters omit COLA-GNN and Murayama (n/a); this is a task-design choice, not a claim that a one-node implementation is mathematically impossible. Both methods have main-panel results for Japan, US-HHS and German states. Their Canadian provincial results use a separately named partial-observation recipe (†), with condition-owned missing-input handling and observed-target training losses, rather than the primary complete-case recipe. Those runs already exist and must be reconciled with the current common panel before any insertion. The Canadian provincial and German state EiNNs External outputs are unavailable in this intake (–), rather than structurally excluded. Appendix D.5 separately audits missing origins within a reported cell.

Table 9: Common point-panel coverage. All listed procedures, including retrospective checkpoints, share these coordinates in A and B. An origin range need not be contiguous.

Task	Season	Models	Origins	Coordinates	First–last origin
US-NAT	2023-24	10	27	27	2023-10-29–2024-04-28
US-NAT	2024-25	10	27	27	2024-10-27–2025-04-27
US-HHS	2023-24	12	12	120	2024-02-11–2024-04-28
US-HHS	2024-25	12	12	120	2025-02-09–2025-04-27
JP-PREW15	2023-24	12	52	1876	2023-08-21–2024-08-12
JP-PREW15	2024-25	12	31	1457	2024-08-19–2025-03-17
CA-PT	2024-25	9	16	196	2025-01-13–2025-04-28
DE-LAB	2023-24	10	26	26	2023-11-06–2024-09-09
DE-LAB	2024-25	10	32	32	2024-11-11–2025-09-08
DE states	2024-25	11	50	800	2024-09-30–2025-09-08

Table 10: Seasonal native MAE on the exact common point panels used in Table 8. –: no eligible common panel or procedure outside the task roster.

		US: national		US: HHS regions		Japan: prefectures	
Procedure	Season	A	B	A	B	A	B
Repeat Last	2023-24	0.86	0.85	0.62	0.59	7.56	5.51
	2024-25	1.25	1.28	1.28	1.28	11.39	9.14
SN52	2023-24	1.31	1.33	0.64	0.65	9.18	9.18
	2024-25	0.94	0.95	0.45	0.50	14.95	14.95
Ridge Direct	2023-24	0.59	0.59	0.31	0.26	8.11	6.93
	2024-25	0.79	0.80	0.51	0.42	11.22	9.69
Ridge Recursive	2023-24	0.58	0.59	0.32	0.26	9.45	6.65
	2024-25	0.87	0.89	0.55	0.45	10.31	8.86
Chronos^R	2023-24	0.77	0.75	0.48	0.44	8.45	5.32
	2024-25	0.88	0.89	0.67	0.63	10.28	8.49
FlowState^R	2023-24	0.74	0.73	0.38	0.35	7.25	5.39
	2024-25	0.77	0.78	0.34	0.33	9.20	7.65
COLA-GNN	2023-24	–	–	0.32	0.35	7.60	6.21
	2024-25	–	–	0.34	0.32	8.24	5.59
Murayama	2023-24	–	–	0.75	0.54	8.08	4.85
	2024-25	–	–	0.68	0.60	7.14	4.53
CAMul	2023-24	1.83	2.52	0.79	2.14	9.76	10.17
	2024-25	2.71	2.72	0.68	2.46	8.99	8.71
EiNNs Target	2023-24	2.58	2.61	1.38	1.43	9.43	9.41
	2024-25	2.96	2.98	1.55	1.61	9.34	9.44
EiNNs External	2023-24	1.79	1.82	0.44	0.48	9.51	9.15
	2024-25	2.19	2.22	0.62	0.64	9.60	9.46
SIR-INN	2023-24	1.00	1.02	0.54	0.55	7.14	6.06
	2024-25	1.18	1.21	0.83	0.84	11.02	10.11

Table 11: Seasonal native MAE on the exact common point panels used in Table 8. –: no eligible common panel or procedure outside the task roster.

		Canada: provinces		Germany: laboratory		Germany: states	
Procedure	Season	A	B	A	B	A	B
Repeat Last	2023-24	–	–	3.97k	3.33k	–	–
	2024-25	458	297	5.65k	5.73k	413	401
SN52	2023-24	–	–	6.31k	6.45k	–	–
	2024-25	336	336	4.74k	4.63k	384	357
Ridge Direct	2023-24	–	–	4.47k	2.75k	–	–
	2024-25	284	198	7.16k	5.74k	558	427
Ridge Recursive	2023-24	–	–	5.30k	2.36k	–	–
	2024-25	408	212	7.86k	5.89k	558	435
Chronos^R	2023-24	–	–	4.37k	2.21k	–	–
	2024-25	347	232	4.03k	3.72k	375	363
FlowState^R	2023-24	–	–	1.01k	1.91k	–	–
	2024-25	290	191	4.35k	4.16k	337	302
COLA-GNN	2023-24	–	–	–	–	–	–
	2024-25	–	–	–	–	590	572
Murayama	2023-24	–	–	–	–	–	–
	2024-25	–	–	–	–	594	581
CAMul	2023-24	–	–	1.49k	2.05k	–	–
	2024-25	379	320	7.04k	9.55k	569	674
EiNNs Target	2023-24	–	–	4.34k	3.93k	–	–
	2024-25	462	462	11.15k	10.91k	821	556
EiNNs External	2023-24	–	–	4.61k	4.17k	–	–
	2024-25	–	–	11.35k	11.10k	–	–
SIR-INN	2023-24	–	–	11.24k	17.76k	–	–
	2024-25	924	455	10.09k	10.79k	537	465

D.3 JOINT POINT AND PROBABILITY COMPARISON WITH DISJOINT SELECTION AND CALIBRATION

The controlled evaluation reserves disjoint selection and calibration periods. Its recorded recipe identifier is *SelectionCalibrationSplit CV*. It is distinct from the procedure-specific leaderboard and earlier common-calendar control. The registered neural matrix is three models $\times$ two tasks $\times$ two seasons $\times$ three seeds $\times$ A/B, with twelve A-selection fits in addition to the 72 final fits. No seed is replaced based on failure or heldout error.

Availability determines the split. For each task/fold, C is the final eight registered preceding-season h2 origins with at least one coordinate whose target label is available in both A and B histories at the first heldout origin. V comprises the eight latest earlier origins with target labels available in both histories at C's first origin. The input freeze lists candidate origins, full-geography label availability and model-common prediction support separately. The registered forecast-coordinate identifiers are preserved. V/C origin and target sets are disjoint, with actual history resolution enforcing publication delays. Selection training excludes V, C and heldout targets; final training excludes C and heldout targets. Condition-A inputs are reconstructed with access to condition-B histories disabled. Calibration labels are stored separately from the fitting/prediction input tensors.

Table 12: Availability-selected origin blocks and cutoffs. Final fitting stops before the first C origin. Every block contains eight origins; dates use native task calendars.

Task / heldout	V origins	C origins	Final cutoff	Heldout start
US / 2023-24	2023-W01–2023-W08	2023-W11–2023-W18	2023-W11	2023-W40
US / 2024-25	2024-W01–2024-W08	2024-W11–2024-W18	2024-W11	2024-W40
Japan / 2023-24	2023-01-23–2023-03-13	2023-04-17–2023-06-05	2023-04-17	2023-07-10
Japan / 2024-25	2024-01-15–2024-03-04	2024-04-08–2024-05-27	2024-04-08	2024-07-08

Training and numerical policy. All neural models use selection seed 1729 and candidates 25/50/100 epochs. Selection minimizes A's V MAE, with ties resolved in favor of fewer epochs. The selected duration is shared across final seeds 1729/2027/314159 and A/B. Ridge uses penalties 0.1/1/10 with smallest-penalty tie breaking. Histories for each condition at its fit cutoff determine fills and min/max scalers; original static geographical graphs are reused, but no old scaler or final state is reused. COLA-GNN uses a 20-week context and learned mask projection; Murayama uses 26 weeks and the same static geographical graph, without commuting-data expansion. CAMul uses ten weeks, target/mask inputs, calendar month sine/cosine features, 64 reference windows and the mean of 16 latent draws within each state. Ridge uses 20 weeks with masks. Neural losses mask unavailable targets and gradients are clipped at norm five. These are declared adaptations, not identical input or computation budgets. Final states are held fixed from C through heldout; SN52 continues to resolve each origin's condition-owned history.

For *Bounded* CAMul, every `OfficialNormal` construction replaces log-scale s by $\min(5, \max(-8, s))$ for both likelihood and sampling, including latent, correlation and output distributions. Gaussian density and gradient-norm intermediates retain FP64 evaluation; weights and optimizer remain FP32. This constraint changes the distributions and the effective objective. The bound is fixed for all CAMul cells before evaluation. It addresses Gaussian-loss overflow observed in a preceding development fit; the replayed failure batch has finite loss and gradients. This does not prove general numerical stability.

Strict deterministic CUDA is used from the initial validation run onward: `CUBLAS_WORKSPACE_CONFIG=:4096:8`, deterministic algorithms with errors on unsupported operations, deterministic cuDNN, one CPU thread, and no AMP, TF32 or cuDNN benchmarking. The environment is B200, Python 3.12.3, PyTorch 2.8.0+cu128, CUDA 12.8 and cuDNN 9.10.2. The training records include hashes of the source code, inputs, configuration and trained model, together with elapsed run time and peak allocated GPU memory.

One state, two prediction blocks. Each final state produces C and heldout points from origin-specific A/B inputs. After both point files are frozen, C residuals use only labels admitted before the first heldout origin. The shared C panel intersects all models, seeds and A/B. Within each

task/fold/condition/seed, signed errors $e = y_C - \hat{y}_C$ yield offsets $d_\tau = Q_\tau(e) - Q_{0.5}(e)$ using Type-7 quantiles; heldout quantiles are $q_\tau = \hat{y} + d_\tau$. Thus $q_{0.5} = \hat{y}$ exactly. Quantiles are computed without clipping or mixing with native CAMul distributions. Each seed is evaluated separately, and calibration loss is excluded from epoch selection. At least 30 coordinates and five origins are required. US C retains 80 coordinates at eight origins in each fold. Japan has 376 common label candidates in both folds, but SN52 lacks 338 exact seasonal lags under both A and B in the first fold. The all-procedure C intersection therefore contains only 38 coordinates across 22 prefectures and eight origins for 2023–24, versus 376 across all 47 prefectures and eight origins for 2024–25. Both exceed the predeclared minimum, but the first is sparse and geographically uneven. Pooling adds regional residuals within the same observed seasons, without adding independent seasons.

Evaluation and uncertainty. Every primary metric uses the same all-model/seed/A/B heldout point/probability panel. Losses average geography, origin and season within seed, then average seeds. Relative MAE is the ratio of these mean losses to the same-condition SN52 loss. Seed sample SD and within-seed season sample SD are reported separately; deterministic baselines have no seed SD. Peaks are the earliest maximum evaluator value over registered target weeks within each season/region, identified only after forecast freeze. Paired intervals use 2,000 shared circular blocks of retained origins within season, with lengths 4/8/12. They condition on the observed seasons and do not account for model selection, calibration estimation or familywise multiple comparisons. The code supplement stores expected values in `reference/US-HHS/` and `reference/JP-PREW15/`. The replay command writes seasonal/seed tables, coordinate losses, missingness ledgers and intervals to `outputs/replay/` (Appendix A).

Table 13: Primary split-CV means ± seed SD. Deterministic references have no seed SD. A and B are separate rows; all metrics use the primary common panel. The code supplement stores seasonal values in `season_seed_metrics.csv` and seasonal SDs in `summary.csv`, under each task's `reference/` directory (Appendix A).

Procedure	Input	MAE	rMAE	WIS9	Coverage (%)	Width
US HHS regions						
SN52	A	1.28	1.00	1.14	39.0	1.21
SN52	B	1.29	1.00	1.15	36.8	1.22
Ridge Direct	A	0.83	0.65	0.72	53.9	1.28
Ridge Direct	B	0.84	0.65	0.72	50.2	1.17
COLA-GNN	A	0.84±0.06	0.66±0.05	0.74±0.05	44.5±1.3	0.82±0.15
COLA-GNN	B	0.81±0.04	0.63±0.03	0.72±0.03	49.0±5.3	0.87±0.20
Murayama	A	0.79±0.01	0.62±0.01	0.70±0.01	44.9±1.3	0.76±0.02
Murayama	B	0.84±0.04	0.65±0.03	0.76±0.04	36.6±2.2	0.69±0.00
CAMul Bounded	A	0.99±0.09	0.77±0.07	0.86±0.07	45.9±2.7	1.16±0.16
CAMul Bounded	B	1.05±0.14	0.81±0.11	0.91±0.11	37.8±6.7	0.98±0.13
Japan prefectures						
SN52	A	11.86	1.00	11.42	14.9	2.46
SN52	B	11.86	1.00	11.42	14.9	2.46
Ridge Direct	A	11.08	0.93	10.29	39.8	7.23
Ridge Direct	B	7.35	0.62	6.78	23.2	3.95
COLA-GNN	A	11.66±0.90	0.98±0.08	11.23±0.88	27.8±1.9	3.84±1.11
COLA-GNN	B	5.40±0.30	0.45±0.03	5.16±0.35	40.5±8.7	2.18±0.76
Murayama	A	9.56±0.10	0.81±0.01	9.36±0.09	32.2±0.5	1.43±0.03
Murayama	B	4.68±0.13	0.39±0.01	4.56±0.12	30.9±1.9	1.19±0.04
CAMul Bounded	A	8.53±0.64	0.72±0.05	8.11±0.55	30.3±10.3	3.46±1.09
CAMul Bounded	B	5.50±0.48	0.46±0.04	5.29±0.49	42.0±4.6	2.12±0.39

Earlier cutoffs are a material change. The previous common-CV final fit occurs immediately before heldout; this recipe's final fit occurs before C in the preceding season. The exact old/new coordinate comparison in the artifacts preserves this distinction. CAMul also changes its numerical constraint, while all new neural models use strict CUDA. Consequently, an old/new score difference

is a comparison of complete recipes, not an isolated causal effect of calibration or the earlier cutoff. Under the earlier fitting schedule, no eligible calibration observations remained after selection; it does not imply that probabilistic forecasting is impossible on the benchmark.

Local delivery and numerical verification. A separate local review verifies the integrity of all 21,376 files in the delivered archive and independently recalculates 99,696 coordinate-level point and probability scores, 1,320 season/seed metrics and 200 aggregate means. Maximum numerical discrepancy is below 3×10^{-14}. It recomputes 12,628 calibration residuals, 792 quantile offsets and all 1,024,830 attached quantiles; the original point fields are unchanged in all 114,818 registered rows. The 144 saved neural prediction files for calibration and evaluation are linked to their corresponding 72 training records and input hashes. Accepted histories independently reconstruct the four V/C splits, training arrays and scalers, label eligibility and all 227 A query origins. This is input and score verification, without local GPU retraining, checkpoint inference or independent-researcher replication; bootstrap intervals are authenticated but not regenerated in this review.

D.4 Resource comparison

The benchmark builds on established vintage-aware forecasting. Table 14 compares the evaluation functions supported by these resources. FluForecastBench's contribution is their integration into a declared multi-system influenza comparison, with the scope and remaining execution limits reported explicitly.

Table 14: Closest resource types and the evaluation function added here. The comparison describes resource scope and evaluation functionality.

Resource	Established capability	Function in FluForecastBench
Influenza revision studies (Brooks et al., 2018)	Model reporting revisions and compare real-time with later histories	A declared roster across four national surveillance systems, with common-coordinate A/B evaluation
COVIDcast (Reinhart et al., 2021)	Versioned indicators with event and issue dates	Bind resolved histories to task-specific origins, frozen submissions and fixed targets
FluSight collaborative evaluation (Reich et al., 2019); COVID-19 Forecast Hub (Cramer et al., 2022)	Shared targets and archived real-time forecast submissions	Rerun paired input conditions under declared procedures and separate input, fitted-state and calibration pathways

D.5 Point uncertainty, support and epidemic phase

The diagnostic protocol was specified after the original results were inspected and before the additional analyses were run. These analyses retain the reference-table roster and scores. All 768 combinations of procedure, condition, resampling rule and block length reproduce the archived MAE and relative-MAE interval endpoints exactly (3,072 checked bounds). Tables below display calendar-gap, block-eight pointwise intervals; the internal research archive retains both rules and block lengths four, eight and twelve. Intervals resample origins within observed seasons and do not quantify uncertainty over future seasons or provide simultaneous winner claims.

Coverage bottleneck. US-HHS COLA-GNN produces 14 of 33 registered origins in each recent season; its intersection with the other procedures retains twelve. For each season and condition, its frozen prediction file reports 190 unavailable regional coordinates at the first nineteen origins with reason `NO_COMPLETE_C20_OR_EFFECTIVE_LEAD_BEYOND_20`. Their recorded consumed-history endpoints imply effective leads of 22–40 weeks. Later produced coordinates have lead three. The completed source audit in Appendix D.6 confirms these gaps in the accepted archive. That appendix evaluates a separately named mask-aware recipe that extends coverage; it does not retrospectively change the original procedure's admissibility.

Removing each member in turn identifies COLA-GNN as the US-HHS coverage bottleneck. The eleven-member sensitivity panel retains 27 origins per season (540 geography–origin coordinates across two seasons), compared with 240 reference coordinates. Table 15 reports every remaining procedure in both conditions. These results describe changed seasonal support, not the missing full-season performance of COLA-GNN. The separate procedure-versus-SN52 sensitivity retains each task's original included seasons and exact A/B intersection; those differently sized pairs must not be ranked against one another.

Table 15: US-HHS support sensitivity. Primary: twelve procedures, twelve origins per season. Wider: eleven procedures excluding COLA-GNN, 27 origins per season. Each column uses its own condition-matched SN52 denominator; every remaining procedure is retained.

	Primary rMAE		Wider rMAE	
Procedure	A	B	A	B
Repeat Last	1.738	1.628	0.865	0.850
SN52	1.000	1.000	1.000	1.000
Ridge Direct	0.752	0.592	0.619	0.595
Ridge Recursive	0.797	0.619	0.638	0.613
Chronos	1.052	0.930	0.760	0.736
FlowState	0.655	0.595	0.671	0.657
Murayama	1.307	0.993	1.086	0.974
CAMul	1.343	4.006	1.074	2.122
EiNNs Target	2.671	2.648	1.976	1.997
EiNNs External	0.974	0.974	0.651	0.658
SIR-INN	1.257	1.212	0.886	0.882

Errors around observed peaks. For each geography and season, we locate the earliest maximum among all supplied h2 evaluator targets in the frozen baseline score file. Targets within fourteen days of that maximum form the peak-near group; remaining targets fall before or after it. This is a retrospective diagnostic based on observed evaluation coverage, not an input available to the forecaster or a claim that every rising/falling week is identified. Within each stratum we score the original all-procedure common coordinates, retaining geography–origin–season weighting and reporting missing strata. Japanese evaluation ends before the 2025-W15 transition. Coverage is 811/342/2,180 Japanese coordinates before/near/after the peak. Some US-HHS strata are absent or represented in only one season because its primary window is short; all counts appear in the released analysis CSVs.

In Japan, COLA-GNN and Murayama reduce mean seasonal error yet exceed SN52 near the observed peak: relative MAE is 1.104 and 1.149, respectively. Their before-peak ratios are 0.489 and 0.514. These results identify peak-period prediction as a remaining accuracy challenge. The phase analysis is descriptive; no hypothesis test or causal interpretation is attached to its subgroup differences.

Table 16: Relative MAE and pointwise 95% intervals on the main-table panels. Calendar-gap stationary bootstrap: block eight, 10,000 replicates. Intervals condition on observed seasons and do not provide simultaneous ranking guarantees. R: retrospective checkpoint; –: no primary result.

Procedure	Input	US national	US HHS	Japan
Repeat Last	A	0.94 [0.71, 1.32]	1.74 [1.36, 2.18]	0.79 [0.43, 1.18]
	B	0.93 [0.71, 1.30]	1.63 [1.24, 2.09]	0.61 [0.30, 0.94]
SN52	A	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]
	B	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]
Ridge Direct	A	0.61 [0.41, 0.78]	0.75 [0.58, 0.97]	0.80 [0.67, 0.96]
	B	0.61 [0.42, 0.77]	0.59 [0.45, 0.78]	0.69 [0.58, 0.84]
Ridge Recursive	A	0.64 [0.46, 0.82]	0.80 [0.63, 1.02]	0.82 [0.71, 0.95]
	B	0.65 [0.46, 0.82]	0.62 [0.48, 0.81]	0.64 [0.50, 0.82]
Chronos^R	A	0.73 [0.54, 0.94]	1.05 [0.84, 1.31]	0.78 [0.43, 1.13]
	B	0.72 [0.52, 0.91]	0.93 [0.71, 1.19]	0.57 [0.28, 0.89]
FlowState^R	A	0.67 [0.49, 0.84]	0.65 [0.53, 0.83]	0.68 [0.40, 0.96]
	B	0.66 [0.48, 0.83]	0.60 [0.47, 0.78]	0.54 [0.29, 0.81]
COLA-GNN	A	–	0.60 [0.52, 0.70]	0.66 [0.42, 0.89]
	B	–	0.58 [0.51, 0.69]	0.49 [0.31, 0.66]
Murayama	A	–	1.31 [0.99, 1.67]	0.63 [0.37, 0.87]
	B	–	0.99 [0.81, 1.23]	0.39 [0.26, 0.52]
CAMul	A	2.02 [1.44, 2.55]	1.34 [1.17, 1.60]	0.78 [0.47, 1.06]
	B	2.30 [1.80, 2.80]	4.01 [3.05, 4.98]	0.78 [0.47, 1.06]
EiNNs Target	A	2.46 [2.08, 2.95]	2.67 [2.45, 2.96]	0.78 [0.60, 0.99]
	B	2.45 [2.09, 2.90]	2.65 [2.40, 3.00]	0.78 [0.61, 1.00]
EiNNs External	A	1.77 [1.36, 2.14]	0.97 [0.82, 1.23]	0.79 [0.62, 0.99]
	B	1.77 [1.38, 2.12]	0.97 [0.81, 1.25]	0.77 [0.61, 0.97]
SIR-INN	A	0.97 [0.77, 1.22]	1.26 [1.07, 1.54]	0.75 [0.41, 1.16]
	B	0.98 [0.79, 1.21]	1.21 [1.02, 1.51]	0.67 [0.34, 1.08]

Procedure	Input	Canada provinces	Germany lab	Germany states
Repeat Last	A	1.36 [0.85, 3.10]	0.87 [0.11, 5.38]	1.08 [0.86, 1.32]
	B	0.88 [0.65, 1.58]	0.82 [0.10, 5.20]	1.12 [0.87, 1.38]
SN52	A	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]
	B	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]
Ridge Direct	A	0.85 [0.71, 1.22]	1.05 [0.35, 17.02]	1.45 [1.23, 1.74]
	B	0.59 [0.46, 0.89]	0.77 [0.33, 8.70]	1.19 [0.94, 1.52]
Ridge Recursive	A	1.22 [0.93, 2.17]	1.19 [0.48, 24.86]	1.45 [1.22, 1.75]
	B	0.63 [0.50, 0.91]	0.74 [0.26, 9.04]	1.22 [0.96, 1.57]
Chronos^R	A	1.03 [0.86, 1.56]	0.76 [0.06, 3.59]	0.98 [0.78, 1.15]
	B	0.69 [0.61, 0.88]	0.53 [0.05, 1.86]	1.02 [0.79, 1.24]
FlowState^R	A	0.86 [0.69, 1.00]	0.48 [0.05, 1.43]	0.88 [0.71, 1.03]
	B	0.57 [0.37, 0.79]	0.55 [0.05, 1.62]	0.85 [0.64, 1.04]
COLA-GNN	A	–	–	1.54 [1.29, 1.79]
	B	–	–	1.60 [1.23, 2.14]
Murayama	A	–	–	1.55 [1.19, 1.84]
	B	–	–	1.63 [1.37, 1.87]
CAMul	A	1.13 [1.01, 1.48]	0.77 [0.11, 7.63]	1.48 [1.30, 1.68]
	B	0.95 [0.80, 1.36]	1.05 [0.08, 2.78]	1.89 [1.25, 3.00]
EiNNs Target	A	1.38 [1.33, 1.51]	1.40 [0.35, 30.63]	2.14 [1.46, 2.59]
	B	1.38 [1.33, 1.51]	1.34 [0.30, 26.73]	1.56 [1.32, 1.90]
EiNNs External	A	–	1.44 [0.38, 33.10]	–
	B	–	1.38 [0.33, 29.15]	–
SIR-INN	A	2.75 [1.88, 3.34]	1.93 [0.06, 9.32]	1.40 [1.14, 1.65]
	B	1.36 [0.60, 2.41]	2.58 [0.05, 15.85]	1.30 [1.00, 1.63]

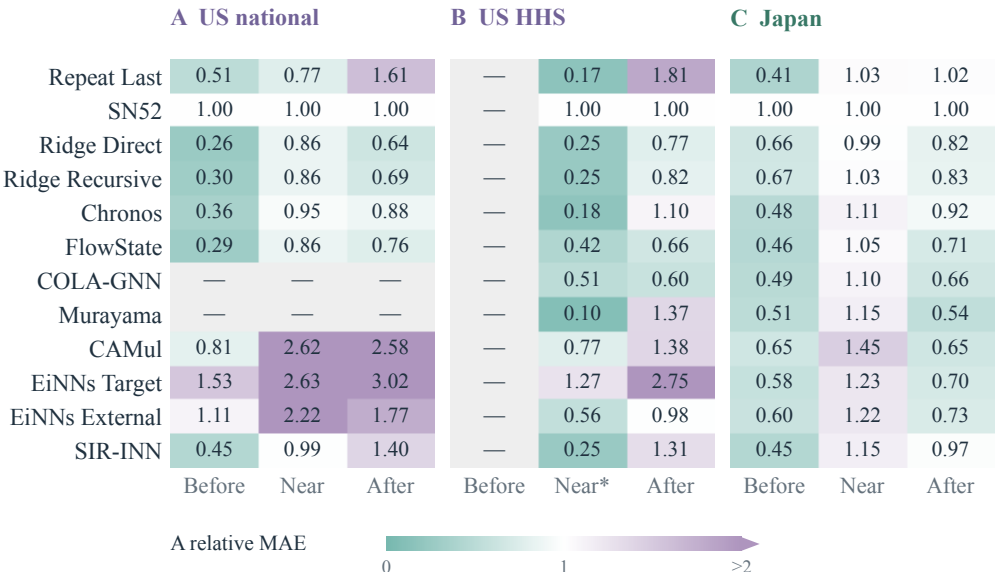


Figure 6: Condition-A relative MAE around the observed peak, on original reference-table coordinates within each stratum. Values below one improve on SN52. Color saturates at two; gray cells are unavailable. *US-HHS peak-near covers only four coordinates from two regions in 2023–24; before-peak support is absent. Retrospective checkpoints are descriptive.

D.6 Recipe variants and execution qualifications

Table 3 uses the disjoint selection/calibration recipe in Appendix D.3. Earlier GPU experiments used different cutoffs, input handling or residual adapters. They are summarized here to distinguish their results from the primary comparison; complete outputs and execution records remain in the internal research archive.

Earlier common-calendar control. This control fits final states immediately before heldout, with settings chosen on preceding-season A forecasts. It shares 620 US and 3,618 Japanese evaluation coordinates across all models, seeds and A/B. Stable CAMul replaces Gaussian-density and gradient-norm intermediates by FP64 calculations after a gradient-overflow failure; it does not bound the distribution. Its 24 final fits succeed, but a separate development fit still fails at epoch 44. Stable CAMul raises US A/B MAE from 0.974/1.125 to 0.994/1.174 on unchanged support. Japanese neural variants exceed SN52 under A in 2023–24 despite improving in the two-season average. After excluding origins already used for selection, none of the four task/folds has A/B-common calibration support. The 114,818 registered point rows are retained, but no strictly disjoint probability scores can be attached to those fixed states. The primary experiment reserves calibration origins before fitting and uses Bounded CAMul; its score differences do not isolate a numerical correction or cutoff effect.

Mask ablation and an alternative US adapter. The mask-aware COLA recipe expands the reference intersection from 240 to 540 US coordinates. On the original panel, seed-2026 A/B MAEs change from 0.328/0.334 to 0.329/0.338. A three-seed mask/no-mask ablation has lower masked error on the narrow panel but higher error on the expanded panel: A/B differences are 0.037/0.062, with eight-origin block intervals $[-0.074, 0.163]$ and $[-0.032, 0.171]$. Thus the experiment supports expanded coverage, not a general accuracy gain from masks. A separate forward-season residual adapter achieves 81.6%/82.7% overall 80% coverage, falling to 53.2%/56.4% near regional peaks. Its selection rules and 23 calibration origins per season differ from the primary eight-origin adapter.

Canadian adaptations. Corrected EiNNs uses A-only input admission and an adapter calibrated on authenticated 2023–24 forward predictions. On the 196-coordinate 2024–25 panel, A/B MAE remains 462.126/462.323, WIS is 422.191/416.938 and 80% coverage is 63.8%/64.0%. Calibration

for 2023–24 is unavailable because required earlier histories and fit inputs are missing. Separate partial-observation COLA-GNN and Murayama recipes have A rMAE 1.550/1.199 and coverage 45.3%/36.7%; A-only input eligibility has not been verified for these variants. They cannot replace primary missing cells or be pooled with the corrected EiNNs comparison.

Repeatability. Earlier Japanese CPU and CUDA training outputs disagree; fixed-input execution alone does not reproduce the primary row. In the examined 2023–24 lead-five cell, strict deterministic CUDA removes the observed convolution-backward discrepancy. Two fixed-47-epoch repeats and two A-selected repeats produce byte-identical predictions and checkpoints within each variant; selection chooses epoch 59. Each checks 5,076 produced predictions. Repeatability was verified for this configuration. Agreement with the original CPU weights, reproduction of all model states and independent replication have not been verified. The primary split-CV campaign uses strict deterministic CUDA throughout.

D.7 Earlier seasonal probability comparisons

The seasonal matrix in the internal research archive retains all earlier probability recipes, including unqualified or unavailable cells. These use procedure-specific reference matching, not the all-member intersection of Table 2. Canadian EiNNs probability values belong to the earlier input-admission recipe and must not be assigned to the corrected point variant. The internal research archive retains the complete model-local point table; the PDF reports common-coordinate point results above.

E Common-panel comparisons and controlled replays

The following comparisons hold evaluation coordinates fixed within each declared panel. They test changes in model gaps and distinguish input availability, revised values and residual calibration; full-season or cross-task causal claims do not follow from these controls.

Table 17: Supported changes in procedure gaps. Counts require both resampling rules at block length eight. The two cohorts use different recipes and metrics; their denominators must not be combined.

Evaluation	Primary pairs	Supported change	Opposite advantage
2023–25 common panels, WIS	42	7	0
Multi-season reference, MAE	18	4	0

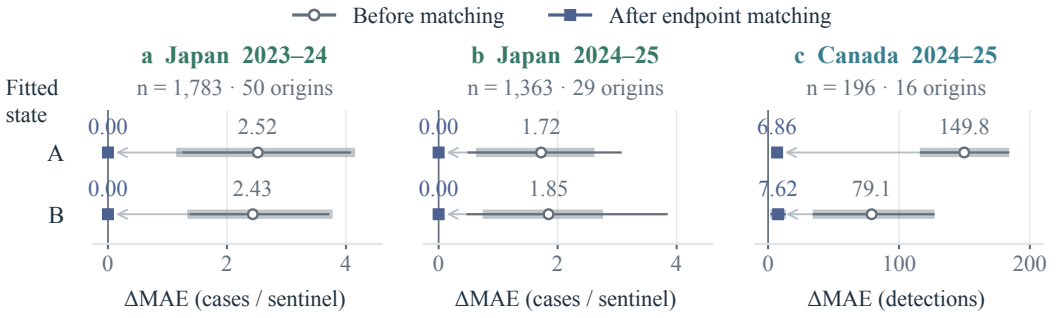


Figure 7: Endpoint controls for Ridge Recursive. Each row holds fitted state A or B fixed; n counts paired coordinates. Markers show MAE(A)−MAE(B) before and after truncating B at A's endpoint. Wide pale and thin dark segments show retained-origin and calendar-gap 95% intervals. Japanese responses vanish; Canadian responses remain smaller but nonzero. Native units stay separate.

E.1 All primary selection contrasts

The analysis retains all 42 nonpretrained pairs on seven original common panels. Excluding Chronos and FlowState does not shrink the original simultaneous uncertainty family. Let $G_A = L_i^A - L_j^A$, $G_B = L_i^B - L_j^B$ and $\Gamma = G_A - G_B$; positive gaps favor the second procedure. Seven changes exclude zero under both resampling rules, but no pair establishes opposite-direction advantages. Intervals are simultaneous within each original panel, not across panels. RL = Repeat Last, RD/RR = direct/recursive Ridge. The complete A-gap and B-gap bands, and the expanded retrospective-checkpoint analysis, remain in the internal research archive.

Table 20: All 42 primary WIS contrasts. A/B gaps are point estimates; R/G are retained-origin/calendar-gap simultaneous bands for the direct change, at expected block length eight. A star marks support under both rules. Unsupported contrasts remain in the same table.

Pair	G_A	G_B	Γ	R band	G band
JP-PREW15 2023–24; 50 origins, 1783 coordinates					
RL–SN52	-2.332	-3.243	0.911	[-0.1497, 1.972]	[-0.1401, 1.962]
RL–RD	-0.4617	-0.3293	-0.1324	[-2.589, 2.325]	[-2.656, 2.392]
RL–RR	-7.81	-0.1495	-7.661*	[-10.53, -4.792]	[-10.49, -4.829]
SN52–RD	1.871	2.914	-1.043	[-2.79, 0.7036]	[-2.894, 0.8067]
SN52–RR	-5.478	3.094	-8.572*	[-10.79, -6.358]	[-10.72, -6.424]
RD–RR	-7.348	0.1798	-7.528*	[-8.367, -6.69]	[-8.23, -6.827]
JP-PREW15 2024–25; 29 origins, 1363 coordinates					
RL–SN52	-3.025	-5.153	2.128	[-0.92, 5.176]	[-6.464, 10.72]
RL–RD	1.012	-0.469	1.481	[-0.5712, 3.534]	[-2.544, 5.507]
RL–RR	-3.207	0.6427	-3.85	[-7.405, -0.295]	[-12.67, 4.973]
SN52–RD	4.037	4.684	-0.6468	[-2.909, 1.615]	[-6.393, 5.099]
SN52–RR	-0.1824	5.796	-5.978*	[-7.501, -4.455]	[-7.498, -4.458]
RD–RR	-4.219	1.112	-5.331	[-8.22, -2.443]	[-11.56, 0.8981]
US-NAT 2023–24; 28 origins, 28 coordinates					
RL–SN52	-0.5052	-0.5055	0.0002315	[-0.02101, 0.02147]	[-0.02101, 0.02147]

Pair	G_A	G_B	Γ	R band	G band
RL–RD	0.1905	0.2054	-0.0149	[-0.035, 0.005201]	[-0.035, 0.005201]
RL–RR	0.1944	0.2001	-0.005746	[-0.03458, 0.02309]	[-0.03458, 0.02309]
SN52–RD	0.6957	0.7108	-0.01513	[-0.04497, 0.0147]	[-0.04497, 0.0147]
SN52–RR	0.6996	0.7056	-0.005978	[-0.04167, 0.02972]	[-0.04167, 0.02972]
RD–RR	0.003908	-0.005247	0.009155	[-0.002002, 0.02031]	[-0.002002, 0.02031]
US-NAT 2024–25; 28 origins, 28 coordinates					
RL–SN52	0.1851	0.2217	-0.03659	[-0.09927, 0.0261]	[-0.09927, 0.0261]
RL–RD	0.3344	0.3648	-0.03035	[-0.08289, 0.0222]	[-0.08289, 0.0222]
RL–RR	0.2791	0.3077	-0.02858	[-0.09404, 0.03689]	[-0.09404, 0.03689]
SN52–RD	0.1493	0.1431	0.006239	[-0.01763, 0.0301]	[-0.01763, 0.0301]
SN52–RR	0.09401	0.086	0.008007	[-0.01298, 0.02899]	[-0.01298, 0.02899]
RD–RR	-0.05529	-0.05706	0.001768	[-0.01404, 0.01758]	[-0.01404, 0.01758]
US-HHS 2023–24; 28 origins, 280 coordinates					
RL–SN52	-0.5713	-0.5816	0.01032	[-0.05137, 0.072]	[-0.05137, 0.072]
RL–RD	0.1827	0.1998	-0.01707	[-0.039, 0.004863]	[-0.039, 0.004863]
RL–RR	0.1796	0.1958	-0.01616	[-0.04267, 0.01035]	[-0.04267, 0.01035]
SN52–RD	0.754	0.7814	-0.02739	[-0.09965, 0.04488]	[-0.09965, 0.04488]
SN52–RR	0.7509	0.7774	-0.02647	[-0.09488, 0.04193]	[-0.09488, 0.04193]
RD–RR	-0.003073	-0.003987	0.0009138	[-0.01173, 0.01356]	[-0.01173, 0.01356]
US-HHS 2024–25; 28 origins, 280 coordinates					
RL–SN52	0.1272	0.123	0.004116	[-0.02067, 0.0289]	[-0.02067, 0.0289]
RL–RD	0.3069	0.3335	-0.02659	[-0.1081, 0.05489]	[-0.1081, 0.05489]
RL–RR	0.2672	0.3011	-0.03387	[-0.1125, 0.04479]	[-0.1125, 0.04479]
SN52–RD	0.1798	0.2105	-0.03071	[-0.09632, 0.0349]	[-0.09632, 0.0349]
SN52–RR	0.1401	0.1781	-0.03798	[-0.1034, 0.02747]	[-0.1034, 0.02747]
RD–RR	-0.03968	-0.0324	-0.007277	[-0.02543, 0.01087]	[-0.02543, 0.01087]
CANADA-PT 2024–25; 16 origins, 196 coordinates					
RL–SN52	118.3	-42.51	160.8*	[9.762, 311.9]	[9.762, 311.9]
RL–RD	177.3	105.2	72.12	[-161.4, 305.6]	[-161.4, 305.6]
RL–RR	45.29	90.91	-45.61	[-206.1, 114.9]	[-206.1, 114.9]
SN52–RD	58.97	147.7	-88.7	[-181.4, 4.031]	[-181.4, 4.031]
SN52–RR	-73.03	133.4	-206.4*	[-255.9, -156.9]	[-255.9, -156.9]
RD–RR	-132	-14.26	-117.7*	[-207.3, -28.18]	[-207.3, -28.18]

Retrospective-checkpoint sensitivity. The expanded analysis contains 100 qualified pairs in 17 panels, with 18 supported direct changes under both rules. This larger dependent family is separate from the 42 primary comparisons. Its full supported and unsupported bands remain in the numerical source; it does not add independent discoveries to the primary count.

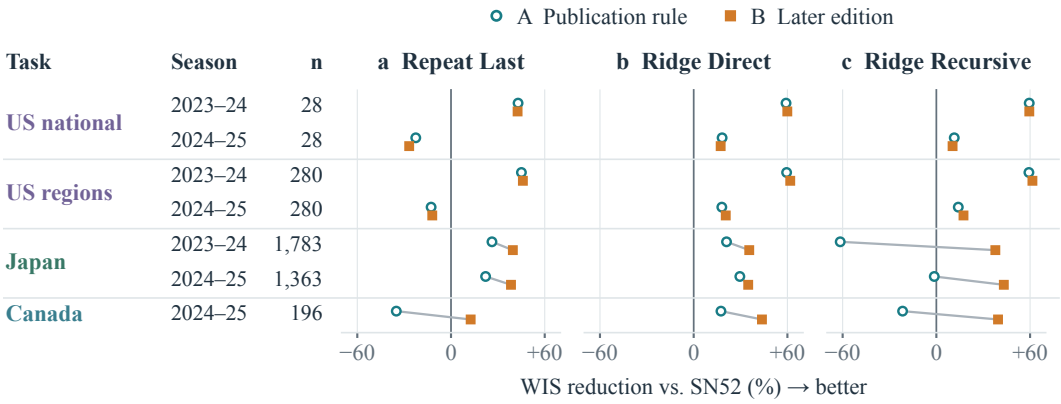


Figure 8: WIS reduction against SN52: $100(1-L^k_m/L^k_{\mathrm{SN52}})$ for $k \in \{A, B\}$. Connected A/B points share n coordinates and use condition-specific references. Positive values indicate lower loss; zero matches SN52. Small vertical offsets separate overlapping markers. These are descriptive estimates; simultaneous bands appear in Appendix E.

Table 18: Common-support 2023–24 and 2024–25 comparison: SN52-relative WIS A/B for persistence, linear and pretrained procedures. Every row uses one identical panel for all five procedures and both conditions; SN52 is 1/1. O/G/N: origins/geographies/coordinates. Ratios across different rows are not a leaderboard. Direct ordering bands, full raw losses, ratio intervals and support losses are in Appendix E.

Task	Season	O/G/N	Repeat Last	Ridge direct	Ridge rec.	Chronos^R
JP	2023–24	50/47/1783	0.738/0.604	0.79/0.644	1.62/0.622	0.858/0.601
JP	2024–25	29/47/1363	0.779/0.616	0.704/0.651	1.01/0.568	0.683/0.582
US-NAT	2023–24	28/1/28	0.571/0.574	0.409/0.401	0.405/0.405	0.526/0.517
US-NAT	2024–25	28/1/28	1.23/1.27	0.818/0.827	0.885/0.896	0.882/0.884
US-HHS	2023–24	28/10/280	0.549/0.54	0.405/0.382	0.407/0.385	0.569/0.547
US-HHS	2024–25	28/10/280	1.13/1.12	0.82/0.795	0.86/0.826	0.919/0.891
CA-PT	2024–25	16/13/196	1.35/0.874	0.825/0.563	1.22/0.605	0.96/0.637

Table 19: For 2023–24 and 2024–25, primary comparisons exclude Chronos and FlowState (42 pairs); the expanded analysis retains retrospective checkpoints (100 pairs). Both use the original exact panel coordinates and simultaneous bands. N counts condition-specific evidence; Γ counts direct changes supported under both resampling rules. Bands cover contrasts within each panel; these counts are dependent, not independent discoveries.

	Primary		Expanded	
Supported advantage	N	Γ	N	Γ
Same direction in A and B	8	0	16	0
A only	5	4	9	7
B only	4	1	11	2
Neither under both rules	25	2	64	9
Opposite directions	0	0	0	0

E.2 Multi-season reference recipes: common-panel point MAE

These calculations reuse the original common coordinates and losses. The original A, B and difference aggregates reproduce before resampling. All pairwise A, B and direct-change contrasts form one simultaneous family per panel, with the existing retained-origin and calendar-gap rules and expected blocks 4/8/12. Primary counts omit pairs involving Chronos without shrinking the simultaneous family. MAE and the separate recent nine-decile WIS contrasts have different estimands and denominators. Expected-block sensitivity supports 3/18 primary and 5/30 expanded changes at block 4, 4/18 primary and 6/30 expanded changes at block 8, 6/18 primary and 10/30 expanded

changes at block 12. The reported block-8 count is conditional on that primary resampling scale; support is not invariant to block length.

Table 21: All reference-panel pairs: point-MAE G_A, G_B and direct change Γ. Japan covers 2023–24 and 2024–25; both US panels cover 2016–17, 2017–18 and 2018–19. Each cell gives the estimate and block-8 simultaneous bands. Identical retained-origin and calendar-gap bands are printed once; otherwise they appear in that order. ^{R} involves retrospective Chronos.

Task	Pair	G_A	G_B	Γ
JP-PREW15	RL–SN52	-2.3 [-9.854, 5.255]	-4.037 [-11.14, 3.064]	1.737 [-0.05771, 3.532]
JP-PREW15	RL–RD	-0.00687 [-2.912, 2.898]	-0.3307 [-3.139, 2.477]	0.3239 [-0.6, 1.248]
JP-PREW15	RL–RR	-0.5048 [-5.215, 4.206]	-0.2508 [-3.228, 2.726]	-0.254 [-2.159, 1.651]
JP-PREW15	RL–Ch^R	-0.6069 [-1.828, 0.6145]	0.3127 [-0.5992, 1.225]	-0.9196 [-2.29, 0.4507]
JP-PREW15	SN52–RD	2.293 [-3.374, 7.959]	3.706 [-1.557, 8.97]	-1.414 [-2.733, -0.09436]
JP-PREW15	SN52–RR	1.795 [-2.68, 6.269]	3.786 [-1.418, 8.99]	-1.991 [-3.578, -0.4048]
JP-PREW15	SN52–Ch^R	1.693 [-5.874, 9.259]	4.35 [-2.617, 11.32]	-2.657 [-5.365, 0.05142]
JP-PREW15	RD–RR	-0.4979 [-2.414, 1.419]	0.07992 [-0.2617, 0.4215]	-0.5778 [-2.323, 1.168]
JP-PREW15	RD–Ch^R	-0.6 [-3.991, 2.791]	0.6434 [-2.218, 3.504]	-1.243 [-2.947, 0.4597]
JP-PREW15	RR–Ch^R	-0.1021 [-5.226, 5.022]	0.5635 [-2.447, 3.574]	-0.6657 [-3.253, 1.921]
US-NAT	RL–SN52	0.02475 [-0.4665, 0.516]	-0.006181 [-0.494, 0.4816]	0.03093 [-0.007325, 0.06919]
US-NAT	RL–RD	-0.3555 [-1.138, 0.4271]	-0.5093 [-1.426, 0.4074]	0.1538 [0.004752, 0.3028]
US-NAT	RL–RR	0.0519 [-0.408, 0.5118]	-0.06684 [-0.6215, 0.4879]	0.1187 [-0.006187, 0.2437]
US-NAT	RL–Ch^R	0.2668 [0.05605, 0.4775]	0.2683 [0.05411, 0.4825]	-0.00156 [-0.03188, 0.02876]
US-NAT	SN52–RD	-0.3803 [-0.86, 0.09954]	-0.5031 [-1.131, 0.1249]	0.1229 [-0.03959, 0.2853]
US-NAT	SN52–RR	0.02715 [-0.2131, 0.2674]	-0.06066 [-0.3697, 0.2484]	0.08781 [-0.04535, 0.221]
US-NAT	SN52–Ch^R	0.242 [-0.09379, 0.5778]	0.2745 [-0.05769, 0.6067]	-0.03249 [-0.0841, 0.01912]
US-NAT	RD–RR	0.4074 [0.02192, 0.7929]	0.4425 [0.01371, 0.8712]	-0.03505 [-0.1122, 0.04213]
US-NAT	RD–Ch^R	0.6223 [-0.02272, 1.267]	0.7776 [0.006733, 1.548]	-0.1553 [-0.2977, -0.01299]
US-NAT	RR–Ch^R	0.2149 [-0.08329, 0.513]	0.3352 [-0.04724, 0.7175]	-0.1203 [-0.2311, -0.009491]
US-HHS	RL–SN52	-0.2564 [-0.6681, 0.1552]	-0.2738 [-0.679, 0.1313]	0.01739 [-0.0127, 0.04747]
US-HHS	RL–RD	-0.001983 [-0.3427, 0.3388]	-0.02827 [-0.3696, 0.3131]	0.02628 [0.00458, 0.04799]
US-HHS	RL–RR	-0.1321 [-0.5613, 0.2972]	-0.1435 [-0.5641, 0.2771]	0.01143 [-0.03201, 0.05488]
US-HHS	RL–Ch^R	0.09429 [-0.1187, 0.3073]	0.08896 [-0.1289, 0.3069]	0.005331 [-0.02007, 0.03073]
US-HHS	SN52–RD	0.2545 [0.05759, 0.4513]	0.2456 [0.05753, 0.4336]	0.008894 [-0.01574, 0.03353]

Table 22: Historical Japanese SN52: crossing frozen point forecasts and residual states on identical coordinates. Since $p_A = p_B$ at every coordinate, each state column is unchanged when the point condition is crossed. Losses are nine-decile WIS; the interval is conditional on the displayed seasons. This 3,242-pair probability cohort differs from the 3,618-coordinate forward-input diagnostic.

Season	Pairs	State c_A	State c_B	Δ	95% interval
2023–24	2161	8.55	7.99	0.556	[0.0891, 1]
2024–25	1081	15.1	14.9	0.205	[-0.0514, 0.509]
Equal-season mean	3242	11.8	11.4	0.381	[0.109, 0.649]

Task	Pair	G_A	G_B	Γ
US-HHS	SN52–RR	0.1244 [-0.02818, 0.2769]	0.1303 [-0.01981, 0.2805]	-0.005955 [-0.03857, 0.02666]
US-HHS	SN52–Ch^R	0.3507 [0.07183, 0.6297]	0.3628 [0.1042, 0.6214]	-0.01206 [-0.04377, 0.01966]
US-HHS	RD–RR	-0.1301 [-0.3809, 0.1207]	-0.1152 [-0.3565, 0.1261]	-0.01485 [-0.04684, 0.01714]
US-HHS	RD–Ch^R	0.09628 [-0.1459, 0.3385]	0.1172 [-0.124, 0.3584]	-0.02095 [-0.04301, 0.001111]
US-HHS	RR–Ch^R	0.2264 [-0.06923, 0.522]	0.2325 [-0.03914, 0.5041]	-0.006102 [-0.0511, 0.03889]

RL: Repeat Last; RD/RR: direct/recursive Ridge; Ch: Chronos. Expected-block 4/12 results and exact support identities accompany the numerical analysis.

At block length four, the Japanese SN52–Ridge Direct contrast loses support; the Japanese SN52–Recursive and both US Repeat Last–Direct contrasts remain supported. At block twelve, Japanese Repeat Last–SN52 and US-NAT Repeat Last–Recursive are additionally supported. These are sensitivity results for the same eighteen primary contrasts.

E.2.1 Sequential development-state diagnostic

The development cohort and contrasts were fixed before counterfactual calculation. All 12,262 A coordinates are shared with B; B adds 5,004. The original Type-7 fitting function reproduces both full residual states exactly. The original Japanese quantile mapping then reproduces all frozen evaluation distributions and scores. A third state uses B residuals restricted to the shared development coordinates, preserving the 358-day stratum, minimum support 30, median centering and quantile grid. All three states are applied to the same 3,242 identical evaluation points. Evaluation truths enter scoring only. The two sequential contrasts sum to the original effect; they are order-dependent and do not constitute independent causal contributions. On this ordered diagnostic, common-support value

Table 23: Japanese SN52: development values on common support, then B support extension. Evaluation averages 2023–24 and 2024–25 equally. Intervals condition on frozen development states and resample evaluation origins, using block eight.

State	Development N	Residual median	WIS
A, shared (= full)	12262	0.00263158	11.816414
B, shared	12262	0.00223242	11.816455
B, full	17266	0.02	11.435651

Contrast	WIS difference	Retained interval	Gap-aware interval
A shared − B shared	-4.05045e-05	[-0.000150951, 7.026e-05]	[-0.000150951, 7.026e-05]
B shared − B full	0.380804	[0.109532, 0.648454]	[0.109532, 0.648454]
A full − B full	0.380763	[0.109372, 0.648534]	[0.109372, 0.648534]

revisions change the evaluation score negligibly, while adding B-only development coordinates accounts for nearly all of the observed state-dependent difference. This is a development-support effect, not the five-versus-three-week Ridge endpoint mechanism. The full states, seasonal results and both block-length sensitivities accompany the executable diagnostic.

E.3 Matching the usable endpoint

Table 24 compares the three inputs under each fixed Japanese state. Table 25 describes the associated recent input ages. Canadian full contrasts and endpoint-matched residuals are separated in Tables 26 and 27; Table 30 quantifies recent US input revisions independently of losses. We extend the fitted-state crossing with a later-vintage history truncated after the last numeric position in A, denoted $B_{\leq A}$. The cohort includes every coordinate of both Japanese common panels: 1,783 in 2023–24 and 1,363 in 2024–25. Both Ridge states and variants, Repeat Last and SN52 are fixed before calculating the new forecasts. Calendar positions and missingness are preserved; no value is imputed, no coefficient or residual state is fitted, and the evaluator is unchanged. Native forecast functions reproduce 25,168 original predictions exactly. This is a retrospective point-MAE diagnostic, separate from the primary probability comparison.

For every coordinate, $B_{\leq A}$ produces exactly the A-input forecast under either fixed Ridge state and for both baselines. The usable target distance is five weeks under A and three under B. Thus, on these panels, removing the additional later observations eliminates the fixed-state point response. The control isolates the availability/endpoint pathway from common-position value changes for these forecasts; it does not separate the information in the added observations from recursive propagation, or attribute the full WIS difference, which also includes fitted and residual states.

Table 24: Endpoint control on exact Japanese common support. Point MAE; state is held fixed across the three input columns. 2023–24: 50 origins/1,783 coordinates; 2024–25: 29/1,363. Comparing A/B input columns holds state fixed; comparing state rows holds input fixed. A minus truncated B is exactly zero on every coordinate.

Season	Procedure	State	A	B	$B_{\leq A}$
2023–24	Direct	A	8.189	7.682	8.189
2023–24	Direct	B	7.742	6.848	7.742
2023–24	Recursive	A	9.412	6.894	9.412
2023–24	Recursive	B	8.981	6.547	8.981
2023–24	Repeat Last	—	7.569	5.592	7.569
2023–24	SN52	—	9.332	9.332	9.332
2024–25	Direct	A	11.47	10.33	11.47
2024–25	Direct	B	11.07	9.784	11.07
2024–25	Recursive	A	10.68	8.961	10.68
2024–25	Recursive	B	10.71	8.861	10.71
2024–25	Repeat Last	—	11.26	8.63	11.26
2024–25	SN52	—	15.2	15.2	15.2

For each fixed state, the $B_{\leq A} - B$ MAE difference and its two origin-block intervals equal the A-input minus B-input result in the frozen Ridge input-crossing inventory. Every A-input minus $B_{\leq A}$ contrast is exactly zero, including both interval endpoints. These equalities are coordinate-level controls, rather than evidence of population equivalence.

E.4 Actual usable target distance across tasks

The following profiles use the exact numeric observation selected by frozen Repeat Last forecasts, joined to their current coordinate evaluator. Effective lead is target week minus that observation week in the task calendar. It is distinct from nominal h2 and from the 52-position seasonal lookup used by SN52. It characterizes this procedure, rather than assuming all procedures consume an identical endpoint. The primary analysis retains full paired support. The input-age sensitivity below applies fixed four-, eight- and twelve-week caps before recomputing shared-support losses and

intervals; these restricted estimands describe fresher-input subsets. Medians and ranges below are descriptive across coordinates.

Table 25: Effective lead in weeks on recent Repeat Last paired support. A/B entries are median [minimum, maximum]; O/N denotes origins/coordinates.

Task	Season	O/N	A	B	Median gap
CA-ILI	2023–24	32/32	22.5 [5, 45]	3 [3, 8]	19.5
CA-LAB	2023–24	25/25	17 [7, 36]	3 [3, 4]	14
CA-LAB	2024–25	4/4	4 [4, 4]	3 [3, 3]	1
CANADA-PT	2023–24	38/485	8 [4, 16]	3 [3, 7]	5
CANADA-PT	2024–25	41/533	13 [4, 33]	3 [3, 14]	9
DE-LAB	2023–24	42/42	4 [4, 5]	3 [3, 3]	1
DE-LAB	2024–25	48/48	3 [3, 6]	3 [3, 3]	0
DE-LAENDER-ARI	2023–24	52/832	4 [4, 5]	3 [3, 3]	1
DE-LAENDER-ARI	2024–25	52/832	4 [4, 6]	3 [3, 3]	1
JP-PREW15	2023–24	52/2444	5 [5, 6]	3 [3, 3]	2
JP-PREW15	2024–25	31/1457	5 [4, 6]	3 [3, 3]	2
US-HHS	2023–24	33/330	3 [3, 3]	3 [3, 3]	0
US-HHS	2024–25	33/330	3 [3, 3]	3 [3, 3]	0
US-NAT	2023–24	33/33	3 [3, 3]	3 [3, 3]	0
US-NAT	2024–25	33/33	3 [3, 3]	3 [3, 3]	0

E.5 Canadian endpoint control and input-age sensitivity

The same endpoint-truncation rule is applied to all 196 coordinates and sixteen origins of the CANADA-PT 2024–25 five-procedure common panel. This is the available complete Canadian panel, selected by input and procedure availability before computing the new controls. Each original fitted state, calendar position, feature rule and evaluator is retained. The original prediction function reproduces all 1568 original controls exactly. Crossed and endpoint-truncated outputs are supplementary point-MAE diagnostics; their losses do not replace primary WIS cells. A and truncated B have the same consumed feature positions on all 196 coordinates. Their numerical features agree exactly on 89; 107 coordinates contain 192 changed feature values. Endpoint matching leaves a value-revision pathway in this Canadian control. This is conditional on the fixed state and endpoint; the effects need not be additive under a different intervention order.

Table 26: Canadian endpoint control: point MAE on identical support. Each row fixes the fitted state.

Procedure	State	A	B	$B_{\leq A}$
Ridge Direct	A	284.3	220.8	284.9
Ridge Direct	B	292.1	197.6	294.1
Ridge Recursive	A	408.4	258.6	401.5
Ridge Recursive	B	291.5	212.4	283.9
Repeat Last	—	457.8	296.6	461.8
SN52	—	335.7	335.7	335.7

Table 27: Residual input-value response after endpoint matching: A minus truncated-B MAE, with pointwise 95% origin-block intervals. Retained-origin and calendar-gap intervals coincide on this contiguous-origin panel.

Procedure	State	Difference	Both intervals
Ridge Direct	A	-0.6095	[-7.717, 6.802]
Ridge Direct	B	-2.017	[-12.59, 8.595]
Ridge Recursive	A	6.857	[2.918, 10.68]
Ridge Recursive	B	7.621	[1.773, 13.48]
Repeat Last	—	-3.978	[-11.42, 3.754]
SN52	—	0	[0, 0]

Eight-week input-age sensitivity. A review-motivated maximum target-to-last-observation distance of eight weeks was fixed before recomputation. The filter uses the actual Repeat Last input endpoint as a shared coordinate policy; it does not assert that every model consumes that endpoint. All original group members and coordinate intersections are retained before applying this one filter. Original prediction distributions, evaluator observations, the minimum eight-origin interval rule, both resampling methods and within-panel simultaneous-band construction are unchanged. Panels with identical support reuse their original numerical results. This threshold is a sensitivity choice, not a retrospectively optimized eligibility rule.

Table 28: Panels changed by the eight-week A-input-age cap. All other common panels retain their original support.

Task	Season	Group	Original N	Capped N	Origins
CANADA-PT	2024-25	Linear	196	184	15
CANADA-PT	2023-24	Pretrained	475	259	21
CANADA-PT	2024-25	Pretrained	485	192	15

After the cap, 7/42 primary and 18/100 expanded direct changes remain supported. The equal counts do not imply identical discoveries: sixteen of the original eighteen expanded comparisons remain supported; two Canadian 2023–24 pretrained contrasts lose support and two 2024–25 contrasts gain it. The original seven primary supported pairs remain supported. Thus very stale baseline inputs do not account for all primary findings. Support changes alter the estimand, and no population-level robustness follows.

Stricter and looser input-age policies. Four- and twelve-week thresholds were requested and fixed together after the eight-week result, before their computation. All are reported; none changes primary eligibility. The four-week policy excludes the Japanese panels, whose A distance is five

Table 29: Input-age cap sensitivity. Denominators include only contrasts retaining at least eight origins; changing the cap changes the evaluated population.

Maximum A age	Primary supported/qualified	Expanded supported/qualified
4 weeks	0/24	3/58
8 weeks	7/42	18/100
12 weeks	7/42	19/100

weeks, and leaves no supported primary change. It cannot establish equality on the discarded coordinates. The eight-week policy retains the original seven primary supported pairs; twelve weeks does likewise. Thus the inference depends on the declared input-age policy and its resulting support.

Other requested Canadian controls. CANADA-PT 2023–24 has no scored A Ridge Recursive output, so the four-procedure paired control does not exist. CA-LAB 2023–24 has A endpoint

positions absent from B: simple truncation cannot impose identical endpoints on its full point intersection. These failures concern this control design, not absence of all Canadian forecasts. We retain their original point/probability results and do not impute positions or silently narrow the control cohort.

Procedure-specific endpoints in the existing consumed-input cohorts. These are previously validated CAMul-Reduced and EiNNs-External inputs, with their own cohorts and recipes. They are not substituted for current recent-panel tensors. Medians use recovered finite target-history endpoints separately by condition; the coverage columns expose missing endpoint evidence rather than filling it with Repeat Last values.

Value revisions in recent US histories. We measure the mean absolute A/B revision over common finite positions in the four native weeks immediately before each nominal origin, then summarize across origin–geography coordinates. Missing common positions are excluded explicitly, never assigned zero. Units are wILI percentage points. HHS quantiles describe coordinates rather than treating regions as independent replications. These input summaries cannot by themselves distinguish model insensitivity from a small input perturbation.

Table 30: Recent US four-position common-value revision magnitude. Each season has 33 origins; the regional task has ten regions.

Task	Season	Coordinates	Median	90th percentile
US-HHS	2023-24	330	0.03385	0.2723
US-HHS	2024-25	330	0.02139	0.1784
US-NAT	2023-24	33	0.03048	0.05163
US-NAT	2024-25	33	0.03822	0.07067

E.6 Additional diagnostic qualifications

Fixed-state value replay. At the earliest US-NAT origin in each anchor season, six original Chronos sample arrays reproduce exactly. Context positions, masks and endpoints agree, while 102 finite input values differ. Crossing both inputs with both RNG lineages gives WIS effects 0.00770/0.00767, 0.0264/0.0288 and $-0.00577/-0.00363$. The last original contrast was -0.0147, so sampling variation materially changes its magnitude. These are controls evaluated at the selected forecast origins, not a conclusion that whole-season US effects are negligible. In the separate 2025–26 observations, Chronos US-NAT's MAE difference is 0.102 [0.009,0.195] and CAMul US-HHS's is -0.642 $[-0.928, -0.385]$. The original full results remain in the internal research archive.

Unchanged values with different availability. For CA-ILI 2020–21, 26 origins have equal values at shared observed positions but different available positions and endpoints. Fixed-RNG WIS effects are $-0.0320/-0.0281$, or 23.0%/20.4% of B loss. For the primary lineage the retained-origin interval is $[-0.0583, -0.00575]$, while the calendar-gap interval is $[-0.0896, 0.0255]$. Availability, endpoint and scaling move together; this control does not identify their separate effects. Other anchor seasons fail the unchanged-values criterion. Japanese Chronos replay did not reproduce across the recorded hardware/software environments and was excluded from this analysis.

Broader seasons and horizons. The point-MAE contrast has the same direction of effect across all three reference seasons for 9/10 Japanese procedures, 2/9 US-NAT, 4/11 US-HHS, 1/7 CA-ILI, 0/6 CA-LAB and 4/7 DE-LAB. Exact zero is a separate descriptive state; neither sign persistence nor sign change establishes significance. Low incidence in 2020–21 makes relative effects sensitive to small denominators. Shared-support h1/h2/h4 comparisons show attenuation in Japan but increasing absolute effects in US-HHS 2024–25. Separate Canadian 2025–26 graph experiments favor B for COLA-GNN h1/h2 and A for Murayama h1/h2/h4. There is no universal horizon trend. School and partial-observation experiments contain gains and losses and remain separately reported, without changing primary cohort membership.

Recent probability coverage and short windows. Japanese Murayama probability output is qualified for eleven origins per recent season, despite wider point coverage. Reference matching cannot turn that short interval into a full-season experiment. CA-LAB 2024–25 has only four recent origins for several probability routes and therefore no qualified origin-block interval. Canadian national and German national SN52 residual policies differ between recent seasons; unsupported cross-season pooling is not performed. Missing reference output, missing predictive output, archive-limited inputs and missing evaluation observations are distinct statuses in the seasonal matrix in the internal research archive. Count-target latent SIR uncertainty is not included as observation-predictive WIS.

Separate Canadian recipes. Appendix D.6 summarizes the partial-observation graph adaptations and their admission qualifications. They are distinct from primary complete-context procedures; the internal research archive retains all four calibrated results.

F MECHANISTIC POSTERIOR APPROXIMATION

The higher-precision SIR approximation holds the posterior target, inputs, prior bounds, growth constraint, nuisance values, output target and evaluator fixed. Older estimates remain a numerical sensitivity in the internal research archive.

Diagnosis. The validation panel contains 50 units, each with a production chain and two validation chains. Production classifications are 2 trapped, 3 under-explored, 33 imprecise and 12 adequate; all 150 chains give 7/6/91/46. Trapped chains have best log density more than ten below the reference; under-explored chains cover less than half its mass; imprecise quantiles differ by over five reference standard errors despite posterior coverage. The batched and frozen scalar log targets agree within 10^{-9}. These classifications describe the examined chains, not every production fit.

Reference estimator. Eight independently scrambled Sobol replicates (Sobol', 1967; Owen, 1995) use self-normalized importance weights. The defensive proposal mixes 30% uniform mass over the parameter box with Gaussians from an adaptive grid, valid local-optimum Hessians and previous weighted samples. Each stage holds its proposal fixed; points outside the prior, growth, time or nonnegative-incidence constraints have zero weight. Expected-incidence quantiles use weighted forecast atoms; observation-predictive quantiles invert the weighted Poisson or truncated-Gaussian mixture CDF. Shifted-grid, uniform and multistart probes check for missed modes; sampled log maxima more than one below the best optimum are flagged. Boundary-biased midpoint quadrature is used only to construct proposals and cross-check, not as the final estimator. Proposal covariances, fallback rules and unit-level records remain in the numerical source.

Precision rule and budget. For quantile k, let SE_k be its standard error across the eight replicates. The recorded criterion is

$$2\sum_k w_k \mathrm{SE}_k \le \epsilon_{\mathrm{abs}} + 10^{-2}\sum_k w_k |q_k|,$$

where $w_k = 2\max(\tau_k, 1-\tau_k)/23$, with effective sample size at least 2,000. Absolute tolerances are 0.1 for counts and German regional rates, 0.001 for Japan, 10^{-6} for US percentages and 0.01 for CA-ILI. The primary target is expected incidence for counts and observation-inclusive prediction for rates. Tail, integer-CDF and secondary-target precision are recorded separately, without gating acceptance.

Budgets double from 2^{16} to 2^{20} points per replicate. Units above tolerance with adequate effective sample size retain that status; unresolved units retain their historical forecast and are excluded from higher-precision support. A fixed refinement uses weighted-cluster proposals, optimum restarts, independent seeds and at most 2^{21} points per replicate, without pooling initial samples. The policy is fixed before scoring and does not tune tolerance to an A/B gap. Of 15,906 units, final statuses are 15,814 meeting the criterion, 91 above tolerance and one unresolved. On the 50-unit validation panel, 44 meet the criterion and six remain above tolerance.

Interpretation. Replicate error does not bound missed modes, self-normalization bias or deterministic error. The 1% weighted quantile criterion is neither a guarantee on true WIS nor a per-tail

or A/B-error bound. Numerical and origin-bootstrap uncertainty are distinct; plug-in nuisance uncertainty is excluded. Meeting the criterion does not prove convergence or parameter identification. Historical numerical changes do not quantify error for every recent SIR recipe, whose targets and support can differ.

G REPRODUCIBILITY AND ACCESS

Scoring frozen forecasts, replaying fitted states and training from scratch are separate verification levels. Local inclusion also differs from permission to redistribute provider content. The code supplement's `README.md` lists execution commands, and `FILE_GUIDE.md` identifies source and result files. The internal research archive is not included in the submitted ZIP (Appendix A).

Primary executable supplement. The `uv`-managed code supplement covers the five-procedure US-HHS/JP-PREW15 comparison in Table 3. It includes resolved histories, fitting tensors, separate calibration/evaluation observations, frozen points, reference outputs and scoring/training code. Python 3.12 and `uv.lock` fix dependencies; CPU/CUDA extras select the backend. A fresh extraction reproduces 99,696 scored rows and passes 5,700 numerical comparisons for the table, peak slices, season/seed summaries and paired intervals (maximum difference 1.5×10^{-14}). Input reconstruction reproduces all four folds. CPU Ridge/SN52 refits match 22,900 predictions within 7.1×10^{-13}, including unavailable outputs. These checks start from resolved histories, not original reports.

The package supports twelve A-only neural selection trajectories and 72 final A/B fits. CPU boundary checks cover initialization, loss, updates and predictions; a partial training run checks execution and resumption. A full post-refactoring GPU rerun and replication by an independent research team have not been verified. Other countries, restricted virological auxiliaries and raw-report acquisition are outside this executable supplement. Seven-task leaderboard aggregates are included for inspection, without a claim of full-roster training or bootstrap regeneration.

Reference-score and diagnostic verification. The internal verification bundle `common-score-replay.zip`, which is not part of the code supplement, recomputes the exact six-task all-member intersections and hierarchical losses. An isolated standard-library extraction reproduces 128 mean MAEs, 216 seasonal values and ten coverage rows; four exported files match byte-for-byte. Differences from upstream floating-point means are below 1.9×10^{-12}. Corrected Canadian EiNNs losses also bind to pre-score prediction freezes. This verifies score arithmetic and supplied artifacts, not independent model regeneration.

Diagnostic records in the internal research archive support regeneration of 18 common-support panels, both bootstrap rules and block sensitivities, and the 3,242-pair Japanese residual-state crossing. The sequential SN52 diagnostic reconstructs residual quantiles from 29,528 development rows before applying the counterfactual to fixed evaluation points. The bounded Chronos replay reproduces six US-NAT and 52 CA-ILI original sample arrays with recorded checkpoints and CUDA dependencies, then crosses inputs and RNG lineages. Japanese controls fail cross-environment reproduction and are excluded. The authors performed these replay checks; independent replication and the historical admissibility of the pretrained models have not been verified.

History and baseline access. Table 31 describes the locally checked eight-task history artifact. Its isolated standard-library route reproduces 62 h2 baseline cells within 10^{-9} on identical support, generating forecasts before evaluator attachment. A representative historical US-NAT Ridge refit creates 30 states and reproduces point/probability scoring within 5.4×10^{-15}; development-selected settings and residual states are supplied fixed inputs. It does not refit recent leaderboard procedures. Figure 5 replays all registered 2024–25 US-national and Tokyo weeks, preserves prediction gaps and uses no fitted model. Exact historical editions remain necessary: a current API response can return a later edition than the required cutoff.

Provider terms. US histories retain federal public-domain attribution; Canadian histories retain non-commercial reproduction terms; German regional data retain CC-BY-4.0 attribution. JIHS permits attributed reuse and adaptation subject to third-party exceptions. Permission to publish SurvStat analyses is distinct from permission to redistribute the underlying data (Japan Institute for Health

Table 31: Locally verified history artifact and task-level terms. All eight tasks contain A/B histories, evaluators and masks. Counts record reproduced h2 baseline cells matching to 10^{-9}. Redistribution rights require a separate source-specific assessment. AGI/NRZ auxiliary access is discussed separately.

Task	Provider terms	Cells
JP-PREW15	JIHS attributed reuse of derived numerical tables; provider PDFs and third-party material excluded	10
US-NAT	US federal public domain; Source: CDC, non-endorsement notice, marked derived table	9
US-HHS	as US-NAT	9
CA-ILI	PHAC, non-commercial reproduction with attribution and source URL	8
CA-LAB	as CA-ILI	8
DE-LAB	RKI/IfSG notification targets; publication and redistribution assessed separately from AGI/NRZ virology	8
CANADA-PT	PHAC, non-commercial reproduction with attribution	6
DE-LAENDER-ARI	RKI, CC-BY-4.0 with attribution; derived vintage selection declared	4

Security, n.d.c; Robert Koch Institute, n.d.c). DE-LAB targets, regional ARI targets and AGI/NRZ virological auxiliaries require separate assessments. Original provider PDFs, third-party works and source content without redistribution authorization are excluded from public packaging.

Restricted virology and local reconstruction. The provider does not permit the proposed public-download redistribution of AGI/NRZ sentinel virology extracts, derived tables or source metadata. These auxiliaries use a user-side reconstruction route. Users obtain the scoped 2015-W40–2024-W45 weekly reports under terms applicable to their use (Robert Koch Institute, n.d.a); retain the exact edition and publication evidence locally; extract subtype, submitted-specimen and influenza-positive counts separately; and calculate proportions only with an available, explicitly identified denominator. Missing counts remain missing. Issue-specific records remain separate until the origin resolver selects A or the specified later edition B. Apply the recorded normalization and recipe before attaching evaluation targets. PDF metadata alone does not authenticate publication dates. Attribution or conversion to proportions does not substitute for permission. This restriction concerns the named auxiliaries, not a newly inferred restriction for other German targets. Independent end-to-end reconstruction and a complete public archival release have not been verified.

Execution and resources. The primary GPU campaign uses an NVIDIA B200, Python 3.12.3, PyTorch 2.8.0 and CUDA 12.8; fit receipts record wall time and peak allocated memory. Appendix D.3 gives deterministic settings and the delivered prediction/state checks. The successful fits in the earlier common-calendar experiment required a total of 4,746 seconds of recorded run time with at most 0.50 GB allocated GPU memory; the later Stable CAMul campaign adds 6,305 final-fit seconds. These sums can include concurrent work and do not estimate end-to-end or full-benchmark cost. The internal research archive retains failed attempts, numerical corrections, CPU/CUDA disagreements and scoped replay timings. A complete per-procedure compute ledger and external replication remain outstanding.