HISTORICAL INFORMATION IN FORECASTER SELECTION: REPLAY, RETENTION, AND CANDIDATE EXCLUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Historical archives contain reports unavailable when forecasts would have been issued. These reports can change both candidate rankings and which checkpoints survive for final selection. We vary the information available at retention and final selection while holding fitted trajectories fixed and testing selected forecasters on common deployment inputs. Our main study predicts vehicle positions 60 seconds ahead across eight refitted episodes at two transit agencies, SEPTA and VTA. Retrospective selection increases position error in seven episodes. At equal additional realistic evaluation counts, retaining runner-up checkpoints and replaying fewer dates improves mean error over complete replay of one winner per group in six episodes, with agency-average gains of 4.48 m at SEPTA and 0.10 m at VTA. Restricting the library to origin-available fits remains competitive. Finite-candidate analysis separates exclusion, replay-score error, and changes in candidate performance between validation and testing. Full-replay controls distinguish the benefit of wider retention from the cost of reduced date coverage. Accurate rescoring cannot recover a discarded alternative, while retaining more candidates need not improve deployment loss. Retention and final selection must therefore be evaluated together.

1 INTRODUCTION

A forecaster deployed at noon can use only information received by noon. A historical archive may later contain delayed reports or revised measurements, as studied in surveillance backfill (Kamarthi et al., 2022; Tang et al., 2025). Forecast evaluation must respect the prediction setting (Hewamalage et al., 2023); chronological splits alone do not reconstruct the information available at issuance. Data vintages affect estimates and model rankings (Koenig et al., 2003; Feng, 2009; Pfarrhofer, 2024). In a multistage selection procedure, a further question arises: **does realistic evaluation still have access to the candidates needed for a good deployment decision?**

Forecasting systems select among heterogeneous alternatives (Shchur et al., 2023), while resource-limited search can eliminate candidates before full evaluation (Jamieson & Talwalkar, 2016; Li et al., 2018). Replaying validation with origin-available inputs can correct the ranking of survivors, but cannot recover a discarded checkpoint. Restricting the library to origin-available fits is another response, although it may remove useful models trained on more complete historical records. We compare these choices by the selected forecaster's deployment loss.

We compare origin-available (A) and retrospective (F) information at two stages: inner validation retains candidates, and outer validation selects one for deployment. In AA, AF, FA, and FF, the **first letter specifies retention; the second specifies final selection**. All four procedures use fixed fitted trajectories and common A test inputs (Figure 1). Matched libraries help distinguish which candidates are available from how many survive and how well they score on inner validation.

Three questions guide the study. **RQ1** asks whether historical information changes selected forecasters and their test loss across refitted episodes. **RQ2** asks how candidate composition affects these changes beyond candidate count and inner-validation performance. **RQ3** compares restriction, realistic replay, and wider retention under a fixed evaluation allowance.

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.

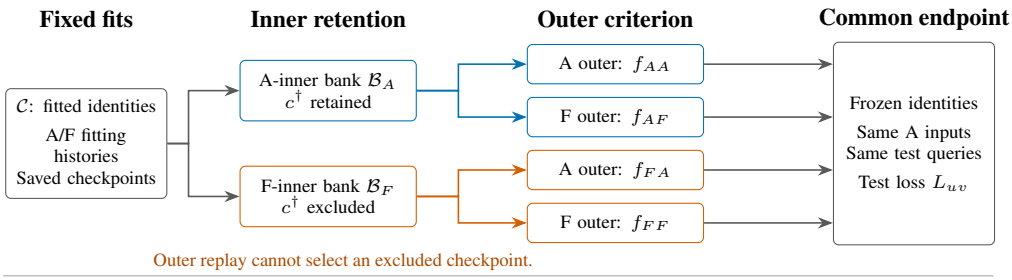

Figure 1: **Ranking and candidate access.** A uses origin-available information; F uses retrospective information. The first letter specifies retention, the second final selection. Candidate $c^\dagger$ illustrates exclusion before replay. Below: q is the number of queries per date, so both allocations score $35q$ candidate–query pairs.

Retrospective selection raises mean test loss in seven of eight refitted episodes; realistic replay removes most of the increase. Retaining runner-up checkpoints at equal additional A evaluation counts improves on complete Top-1 replay by 4.48 m at SEPTA and 0.10 m at VTA. A-only restriction wins in repeated SEPTA, while mixed-history replay wins on other panels. Closing an AA–FF gap therefore need not identify the best procedure.

Our contribution is an empirical account of candidate survival, supported by finite-candidate analysis. The crossed design distinguishes retention from final ranking; matched libraries separate composition from candidate count and inner-validation quality. Full-replay controls then separate the benefit of retaining more candidates from the cost of evaluating fewer dates: at SEPTA, widening gains 6.20 m, while reduced coverage gives back 1.73 m. The analysis explains these outcomes through exclusion, replay-score error, and validation-to-test transfer. These are diagnostics computed after testing; the evaluated policies select using only pre-test information.

2 Related Work

Real-time information and selection. Macroeconomic studies distinguish estimation from evaluation information (Stark & Croushore, 2001; Koenig et al., 2003; Croushore, 2006) and show that revisions alter forecasts and model rankings (Feng, 2009; Pfarrhofer, 2024). Back2Future uses backfill dynamics for prediction and evaluation (Kamarthi et al., 2022); Delphi-RF forecasts mature surveillance values (Tang et al., 2025). We examine how information also determines which fitted candidates survive retention.

Candidate elimination and evaluation resources. Successive Halving eliminates configurations (Jamieson & Talwalkar, 2016); Hyperband and multi-fidelity methods balance candidate breadth against evaluation resources (Li et al., 2018; Li & Li, 2026). We hold fitted trajectories fixed and trade saved-checkpoint retention against realistic replay coverage. Paired replay uses difference estimation (Särndal et al., 1992), related to proxy-plus-correction inference (Angelopoulos et al., 2023). Whereas proxy-assisted LLM ranking studies preference uncertainty (Chatzi et al., 2024), our endpoint is selected forecasting loss; matched subset-only controls isolate the value of auxiliary scores.

Forecasting libraries and evaluation. Irregular and regular-grid models (Yalavarthi et al., 2024; Zhang et al., 2024; Nie et al., 2023), heterogeneous systems, and benchmarks (Shchur et al., 2023; Qiu et al., 2024; Aksu et al., 2024; Godahewa et al., 2021) motivate selection across fitted alternatives. fev-bench emphasizes covariates and aggregation (Shchur et al., 2025); TIME emphasizes fresh data and operational tasks (Qiao et al., 2026); TimesX adds aligned textual context and contamination controls (Liu et al., 2026). Receipt timing is complementary: uncontaminated series can still contain versions unavailable at issuance. Dataset-selection bias limits generalization (Roque et al., 2025), while selection overfitting, leakage, and vintage-aware inference motivate evaluating the complete

procedure (Cawley & Talbot, 2010; Kapoor & Narayanan, 2023; Gonçalves et al., 2025; Hewamalage et al., 2023).

3 INFORMATION-AWARE RETENTION AND SELECTION

3.1 HISTORIES, CANDIDATES, AND THE DEPLOYMENT ENDPOINT

A record version $r = (s_r, a_r, z_r)$ contains event time, availability time, and value. At origin d and role cutoff $c \geq d$, define

$$H_{\mathrm{A}}(d) = \mathrm{resolve}\{r : s_r \leq d,\ a_r \leq d\}, \qquad H_{\mathrm{F}}(d; c) = \mathrm{resolve}\{r : s_r \leq d,\ a_r \leq c\}. \tag{1}$$

Resolution selects the latest eligible version of each event. Both histories exclude future events; F admits later reports about past events. A position recorded at 12:00:10 and received at 12:00:40 is absent from A at 12:00:30, but enters F if its cutoff is at least 12:00:40. Each chronological role has a cutoff and target-maturity rule; compared procedures share supported queries and targets.

A fitted identity records family, fitting history, representation, preprocessing, configuration, seed, and checkpoint. Rescoring changes validation inputs, not weights. An F-fitted candidate remains deployable when its fitting information was available by the fitting cutoff: mature inputs at historical training origins do not imply future inputs at deployment.

Let $\mathcal{C}_g$ contain eligible candidates in family–fitting-history group g and $\mathcal{C} = \bigcup_g \mathcal{C}_g$. Fitting, inner validation, outer validation, and testing use separate chronological roles. Inner criterion $u \in \{\mathrm{A}, \mathrm{F}\}$ retains $c_g^{(u)} = \arg\min_{c \in \mathcal{C}_g} V_u^{\mathrm{in}}(c)$, forming $\mathcal{B}_u = \{c_g^{(u)}\}_g$. Outer criterion v selects $f_{uv} = \arg\min_{f \in \mathcal{B}_u} V_v^{\mathrm{out}}(f)$, with fixed tie order. The endpoint $L_{uv} = R_{\mathrm{A}}^{\mathrm{test}}(f_{uv})$ is empirical loss on common A test inputs. AA uses A at both selection stages; it is not a test oracle.

3.2 STAGE EFFECTS AND CANDIDATE EXCLUSION

The retrospective-selection cost $C = L_{FF} - L_{AA}$ decomposes as

$$\begin{aligned} \Delta_{\mathrm{in}} &= L_{FA} - L_{AA}, & \Delta_{\mathrm{out}} &= L_{AF} - L_{AA}, \\ I &= L_{FF} - L_{FA} - L_{AF} + L_{AA}, & C &= \Delta_{\mathrm{in}} + \Delta_{\mathrm{out}} + I. \end{aligned} \tag{2}$$

Interaction I measures how retention changes outer-criterion sensitivity. Replaying the F-retained bank recovers $G_{\mathrm{out}} = L_{FF} - L_{FA} = \Delta_{\mathrm{out}} + I$, leaving residual Δ_{in}. Thus AA-to-AF sensitivity and FF-to-FA recovery concern different banks. Their signed residuals can reflect different validation-to-test transfer.

For $\varnothing \neq \mathcal{B} \subseteq \mathcal{C}$, let $R^*(\mathcal{B}) = \min_{f \in \mathcal{B}} R_{\mathrm{A}}^{\mathrm{test}}(f)$. Define exclusion E and within-bank selection loss S:

$$E(\mathcal{B}) = R^*(\mathcal{B}) - R^*(\mathcal{C}), \qquad S(\widehat{f}; \mathcal{B}) = R_{\mathrm{A}}^{\mathrm{test}}(\widehat{f}) - R^*(\mathcal{B}), \tag{3}$$

$$R_{\mathrm{A}}^{\mathrm{test}}(\widehat{f}) - R^*(\mathcal{C}) = E(\mathcal{B}) + S(\widehat{f}; \mathcal{B}). \tag{4}$$

Both are nonnegative diagnostics computed from test losses after policies are fixed. Exclusion E measures the cost of candidates missing from the retained set; S measures the cost of not selecting its best candidate. Unlike E, the FA–AA residual compares selected procedures and can be negative.

Proposition 1 (Exclusion, replay error, and transfer)**.** *Fix $\mathcal{B}$, its full A outer scores V_{A}, and its test losses. Set $g_f = R_{\mathrm{A}}^{\mathrm{test}}(f) - V_{\mathrm{A}}(f)$ and $\omega_{\mathcal{B}} = \max_{f \in \mathcal{B}} g_f - \min_{f \in \mathcal{B}} g_f$. If $\widehat{f}$ minimizes estimated scores with $\max_{f \in \mathcal{B}} |\widehat{V}(f) - V_{\mathrm{A}}(f)| \leq \epsilon$, then*

$$E(\mathcal{B}) \leq R_{\mathrm{A}}^{\mathrm{test}}(\widehat{f}) - R^*(\mathcal{C}) \leq E(\mathcal{B}) + 2\epsilon + \omega_{\mathcal{B}}. \tag{5}$$

The lower bound follows because the selector cannot choose outside its bank; the upper bound compares it with the bank's best test candidate (Appendix B). Retention addresses exclusion, replay addresses score error, and $\omega_{\mathcal{B}}$ measures how differently candidates' losses change from validation to testing. Exact replay removes score error, but not exclusion. These test-based quantities explain outcomes rather than guide selection.

3.3 RESTRICTION AND RETENTION UNDER A REPLAY ALLOWANCE

F-inner Top-1 keeps one winner per group. **Top-2** also keeps the lowest-F-inner-loss checkpoint at least ten epochs from each graph trajectory's winner. Singleton trees and the history-independent reference are unchanged. **Spread** adds the farthest eligible checkpoint, preferring the earlier epoch on a distance tie; it controls count and separation without using runner-up quality. The **A/F union** keeps both criterion-specific inner winners, deduplicating identities.

A-only F/F restriction removes F-fitted groups but retains retrospective validation. Because group searches are separable, it requires no additional A scoring from the completed mixed library. Removing models fitted on more complete histories may reduce sensitivity but also discard useful candidates; Top-2 preserves both fitting histories.

For J candidates, K equally sized blocks of q queries, and additional A evaluation allowance B, feasible replay coverage is

$$m(J;B) = \min\{K, \lfloor B/(Jq) \rfloor\}, \qquad B \geq Jq. \tag{6}$$

Top-1 on seven dates and Top-2 on five dates each use $35q$ A candidate–query evaluations: 5×7 versus 7×5. Top-2 and Spread share dates and use subset A means; full Top-2 uses $49q$. Fits are fixed. Counts do not equate heterogeneous runtimes; reconstruction is charged per date/history, and AA/union additionally score A-inner checkpoints. Top-2 and Spread use only F-inner scores and saved identities before permitted A outer queries select the deployed model.

3.4 WIDENING AND REDUCED TEMPORAL COVERAGE

For nested banks $\mathcal{B}_1 \subseteq \mathcal{B}_2$, put $D = R^*(\mathcal{B}_1) - R^*(\mathcal{B}_2) \geq 0$. The gain between selected identities is

$$R_{\mathrm{A}}^{\mathrm{test}}(\widehat{f}_1) - R_{\mathrm{A}}^{\mathrm{test}}(\widehat{f}_2) = D - \big[S(\widehat{f}_2;\mathcal{B}_2) - S(\widehat{f}_1;\mathcal{B}_1)\big]. \tag{7}$$

The change in S includes both bank and coverage changes. To separate them, let L_r^{full} be bank r's full-A winner's test loss and L_2^{part} the wider bank's partial-replay loss. Then

$$\begin{aligned} W &= L_1^{\mathrm{full}} - L_2^{\mathrm{full}}, & H &= L_2^{\mathrm{part}} - L_2^{\mathrm{full}}, \\ G &= L_1^{\mathrm{full}} - L_2^{\mathrm{part}} = W - H, & W &= D - (S_2^{\mathrm{full}} - S_1^{\mathrm{full}}). \end{aligned} \tag{8}$$

W measures widening under complete replay; H changes coverage within that same bank. Negative H means a partial-panel choice transfers better. These identities hold per seed and replay realization and under common aggregation weights.

Proposition 2 (A sufficient condition for useful widening)**.** *Let j_1, j_2 be full-A winners in nested banks and $\gamma = V_{\mathrm{A}}(j_1) - V_{\mathrm{A}}(j_2) \geq 0$. If wider-bank replay has uniform score error at most ϵ_2, then*

$$R_{\mathrm{A}}^{\mathrm{test}}(\widehat{j}_2) - R_{\mathrm{A}}^{\mathrm{test}}(j_1) \leq -\gamma + 2\epsilon_2 + \omega_{\mathcal{B}_2}. \tag{9}$$

Thus $\gamma > 2\epsilon_2 + \omega_{\mathcal{B}_2}$ suffices for improvement over the narrow bank's full-A winner.

The condition explains the balance of added opportunity, precision, and transfer. Since $\omega_{\mathcal{B}_2}$ uses test losses, it diagnoses outcomes rather than determining a replay allowance. Top-2 chooses candidates directly from F-inner scores.

3.5 REALISTIC SUBSETS AND RETROSPECTIVE SCORE CORRECTION

For a fixed bank with full F outer scores, sample m of K blocks uniformly without replacement. With N queries, block query sets I_k, and sampled blocks $\mathcal{U}$, paired replay uses

$$\widehat{V}_{\mathrm{A}}^{\mathrm{pair}}(j) = V_{\mathrm{F}}(j) + \frac{K}{mN} \sum_{k \in \mathcal{U}} \sum_{q \in I_k} [\ell_{\mathrm{A}}(j,q) - \ell_{\mathrm{F}}(j,q)]. \tag{10}$$

Subset-only replay replaces the bracket by $\ell_{\mathrm{A}}(j,q)$ and omits the F mean. Both are unbiased for full query-weighted A scores over block sampling, including unequal block sizes (Särndal et al., 1992). Pairing reduces pairwise ranking variance when candidate correction differences vary less than A-loss differences. Appendices B.5–B.6 give exact variances and a margin-weighted regret bound. Hourly comparisons use identical banks and subsets; transit retention uses subset-only A scores.

4 EXPERIMENTAL DESIGN

Study structure. Eight refitted transit episodes provide the main evidence for RQ1–RQ3. Separate June/July diagnostic panels identify changes in selected architecture and fitting history. Hourly benchmarks test sensitivity to missing recent context and compare paired with subset-only replay. Their fitting schedules differ, so these studies are reported separately.

Tasks and chronological roles. Transit predicts vehicle position at $+60$ seconds using receipt-time versus event-time histories (MobilityData, 2026; Jarvus Innovations, 2026). Its library contains persistence, ExtraTrees, and GraFITi (Geurts et al., 2006; Yalavarthi et al., 2024). Shared anchoring uses A's last position for both histories; original anchoring uses each history's last position. The eight episodes comprise four per agency, with 14/7/7/14 fitting/inner/outer/test days and three seeds. Test windows are disjoint, but development windows overlap. The June/July diagnostic panels use two fitting days and one day per selection stage; their SEPTA fits pool both agencies. Appendix A gives daily windows, dates, and maturity rules.

Electricity and Traffic use 32-series Monash panels, 60/10/10/20% chronological roles, and a 48-hour context (Godahewa et al., 2021). Their library contains classical references, Ridge, ExtraTrees, and PatchTST-based models (Nie et al., 2023). A imposes a three-hour trailing-context restriction; this is not an observed release delay. Origin-hour nowcasting and a separately fitted $+23$-hour task distinguish near-term from longer-horizon effects.

Candidates and matched controls. Each transit seed saves persistence, two trees, and 51 checkpoints from each of two graph trajectories: 105 identities. Three-candidate controls preserve persistence, one tree, and one graph across all four fitting-history assignments. Graph swaps hold tree-A fixed. Quality matching admits corresponding fits within 5% A-inner loss, using the same admitted pairs for both endpoints; this diagnostic screen is separate from F-score retention.

Evaluation and uncertainty. Each repeated validation date uses 2,048 common supported keys sampled without consulting target values or losses. Support-deficient episodes are excluded before scoring. Top-2 and Spread share all 21 five-date subsets. Losses average queries within date, dates, replay subsets where applicable, and seeds; oracle minima are computed within seed. Replay subsets do not count as development replications. Diagnostic-panel calendar intervals condition on selected fits; repeated results report all episodes and leave-one-out sensitivity. Errors are meters for transit and fitting-standardized MAE for hourly tasks. Window, weighting, fitting-duration, native-label, separation, and horizon checks are specified in Appendices A and C.3.

5 EMPIRICAL FINDINGS

5.1 RQ1: HISTORICAL INFORMATION CHANGES CONSEQUENTIAL SELECTIONS

Across eight refitted agency–time episodes, retrospective selection has positive cost in seven (Table 1). SEPTA's mean FF loss is 465.65 m versus AA at 431.23 m, a 34.43 m cost. VTA's cost is 4.55 m, with a 0.50 m reversal in V3. Historical information changes consequential selections under longer fitting and validation windows, but the magnitude varies substantially by agency and episode.

Table 1: **Episode outcomes and retention opportunity.** All values are meters; Gain = FA − Top-2, and D is the reduction in oracle exclusion.

	SEPTA (S1–S4)						VTA (V1–V4)					
Ep.	AA	FF	FA	Top-2	Gain	D	AA	FF	FA	Top-2	Gain	D
1	438.6	469.8	449.4	443.0	**6.4**	7.0	208.3	217.2	209.1	208.8	**0.3**	0.9
2	412.8	455.4	422.1	416.8	**5.3**	5.7	192.4	198.5	193.0	192.8	**0.2**	0.6
3	475.1	521.7	489.8	482.1	**7.7**	8.6	224.6	224.1	224.4	225.0	**-0.6**	0.1
4	398.4	415.7	402.0	403.5	**-1.5**	1.0	201.7	205.4	202.5	202.0	**0.5**	0.9

Shaded bold cells show gains; negative values mean Top-2 is worse. Four episodes per agency; means over three fitted seeds. Equal-work Top-2 averages 21 five-date subsets. AF, Spread, union, and oracle outcomes: Appendix C.

Table 2: **Diagnostic-panel crossed selections on common A inputs.** Letters index inner/outer criteria; $C = \mathrm{FF} - \mathrm{AA}$.

Condition	AA	AF	FA	FF	C	95% interval for C
Transit: meters; O/S denote original/shared anchor						
SEPTA Jul./O	350.420	392.504	352.739	392.504	**42.084**	[37.929,45.060]
SEPTA Jul./S	350.420	375.465	352.739	429.274	**78.853**	[75.773,81.847]
SEPTA Jun./S	590.212	613.790	587.084	633.838	**43.626**	[40.956,47.443]
VTA Jul./O	196.696	196.318	196.594	196.875	**0.178**	[-0.968,0.970]
VTA Jul./S	197.581	200.300	196.275	199.580	**1.998**	[1.696,2.263]†
Hourly: standardized MAE; M/G denote median/fill+age						
Electricity/M	0.2023	0.5995	0.2090	0.6717	**0.4694**	[0.4458,0.4937]
Electricity/G	0.2017	0.4236	0.2019	0.4490	**0.2473**	[0.2362,0.2574]
Traffic/M	0.2177	0.5211	0.2177	0.5367	**0.3190**	[0.3008,0.3379]
Traffic/G	0.2156	0.4979	0.2156	0.5392	**0.3236**	[0.2857,0.3543]

Shaded bold C values emphasize selection cost, not significance. Calendar intervals condition on fitted choices; † seed/calendar interval: [-1.750,5.878] m. Contrasts use unrounded losses. Hourly: origin-hour targets, three-hour imposed restriction, 32 series. Restricted banks: Appendix D.

The June/July diagnostic panels show that historical information changes both architecture and fitting history (Table 2). July SEPTA A validation selects GraFITi-A; F validation selects ExtraTrees-F with original anchoring and GraFITi-F with shared anchoring. The resulting costs are 42.08 and 78.85 m. June shows the same direction, whereas VTA's original-anchor effect is negligible and its shared-anchor interval widens with seed variation.

Electricity's A-selected neural candidate improves over its best A-fitted classical comparator by 8.90% with median inputs and 8.32% with fill+age, yet F selection increases loss by 232.07% and 122.57%. Traffic selects trees, so the effect does not require neural dominance. These large hourly effects concern origin-hour targets with three-hour context restrictions. All twelve zero-delay conditions preserve identical forecast vectors.

5.2 RQ2: Retention and composition shape sensitivity

The effect of outer replay depends on which candidates survived. In shared-anchor July SEPTA, changing the outer criterion matters much more after F retention: the interaction is 51.49 m, compared with an 8.10 m repeated-episode mean. Complete outer replay removes most of the repeated selection cost (Table 4). Yet its remaining FA–AA gaps, 9.60 m at SEPTA and 0.50 m at VTA, differ from oracle exclusion, 7.75 and 0.80 m. The gaps compare selected forecasters; exclusion measures the cost of missing candidates. One cannot substitute for the other.

Matched controls show that candidate count alone does not explain sensitivity (Table 3A). With three candidates throughout, the SEPTA contrast is 2.80 m for A-only, 22.70 m averaged over the four history assignments, and 20.40 m when only the graph history changes. Among fits matched within 5% A-inner loss, swapping histories increases the contrast by 8.10 m at SEPTA and 1.10 m at VTA. Similar inner-validation scores therefore do not eliminate the effect either.

5.3 RQ3: Restriction, replay, and wider retention

Complete procedures. No policy dominates across panels. A-only F/F restriction needs no added A scoring and reaches 434.80 m at SEPTA, 1.55 m below equal-count Top-2. At VTA, Top-2 is better: 207.15 versus 207.40 m (Table 3B). Mixed-bank outer replay also beats restriction in original-anchor July SEPTA and Electricity/M (Table S7). Restriction's value therefore depends on which candidates it removes. Compare complete procedures by their test loss, not by the size of their internal AA–FF gap.

Reallocating realistic evaluations. At 71,680 additional A candidate–query evaluations, Top-2 on five dates improves over complete Top-1 replay in six of eight episode means. The benefit is concentrated at SEPTA: 4.48 m, or 46.6% of the remaining FA–AA gap. VTA gains only 0.10 m (Table 4). The overall mean gain is 2.29 m; leave-one-out means remain positive at 1.51–2.83 m. S4

Table 3: **Candidate composition and retention policies.** Deployment losses and $C = \mathrm{FF} - \mathrm{AA}$ are meters, averaged over four episodes per agency.

A. Family-matched composition controls

	SEPTA			VTA		
Control (bank size)	AA	FF	C	AA	FF	C
Mixed (5)	431.23	465.65	34.43	206.75	211.30	4.55
A-only (3)	432.00	434.80	2.80	207.00	207.40	0.40
Uniform histories (3)	437.40	460.10	22.70	207.80	210.10	2.30
Graph swap (3)	434.60	455.00	20.40	207.60	210.30	2.70
Quality-matched A (3)	432.70	435.20	2.50	207.20	207.80	0.60
Quality-matched F (3)	433.40	444.00	10.60	207.30	209.00	1.70

B. Restriction, retention, and replay

Policy / criterion	Outer dates	A-count	SEPTA	VTA
F-top-1 / F (FF)	7 F	0	465.65	211.30
A-only restriction / F	7 F	0	434.80	207.40
F-top-1 / A (FA)	7/7	1.00	440.83	207.25
Spread / subset A	5/7	1.00	439.68	207.45
F-top-2 / subset A	5/7	1.00	**436.35**	**207.15**
F-top-2 / full A	7/7	1.40	434.63	206.88
A-top-1 / full A (AA)	7/7	21.40	431.23	206.75
A/F union / full A	7/7	$\leq$ 21.80	432.80	206.85

A: Matched rows share 18/24 SEPTA and 14/24 VTA pairs within 5% A-inner error. *B:* A-count normalizes added A candidate–query evaluations by 71,680; blue rows share this allowance; bold marks their lowest loss. Gray marks restriction without added A scoring. Fits and F selection are common sunk work; AA/union include A-inner scoring. Counts are not runtimes (Table S2).

and V3 nevertheless worsen by 1.50 and 0.60 m. Averaging over seeds and replay subsets does not create independent deployment replications.

Spread matches size, dates, and A counts but has higher SEPTA loss than Top-2 (Table 3B). Thus, preserving a good F-inner runner-up matters beyond checkpoint separation alone. Full Top-2 uses 40% more outer evaluations; AA and the A/F union also search A-inner checkpoints. Their lower SEPTA losses come with different evaluation requirements. Counts do not equate runtime or total compute.

Widening versus coverage. Complete Top-2 replay separates access to additional candidates from the reduced temporal coverage imposed by equal evaluation counts (Figure 2; Table 4). SEPTA's widening benefit is 6.200 m; reducing coverage within that bank costs 1.725 m, leaving 4.475 m. At VTA, 0.375 m of widening benefit and 0.275 m of coverage loss nearly cancel.

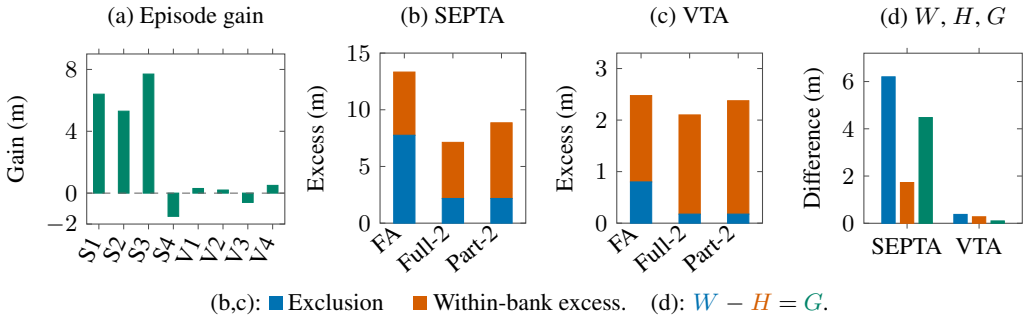

Figure 2: **Opportunity, coverage, and selected loss.** (a) Equal-count gains of partial Top-2 over FA. (b,c) Excess above seed-specific complete-bank test minima: FA is full Top-1; Full-2/Part-2 are full/partial Top-2. (d) Widening W, coverage H, and gain G. Agency panels average four episodes.

Table 4: **Retention-gain decomposition.** Four-episode agency means; absolute differences in meters.

	Outer replay	Exclusion	Widening	Coverage	Equal-count	Gain relative to	
Agency	FF−FA	D	W	H	$G = W - H$	FA loss	FA−AA
SEPTA	24.825	5.575	6.200	1.725	**4.475**	1.015%	46.6%
VTA	4.050	0.625	0.375	0.275	**0.100**	0.048%	20.0%

Shaded bold G is the net equal-count gain. D: oracle exclusion reduction; W: full-replay widening; H: reduced-coverage effect; $G = W - H$. Percentages use unrounded agency means.

The oracle decomposition explains where the SEPTA gain comes from. Wider retention lowers exclusion from 7.75 to 2.18 m. It also improves selection within the bank under full replay, but evaluating fewer dates offsets that improvement:

$$\underbrace{4.475}_{G} = \underbrace{5.575}_{D} - \underbrace{(4.950 - 5.575)}_{S_2^{\text{full}} - S_1^{\text{full}}} - \underbrace{(6.675 - 4.950)}_{H}.$$

Thus the change in within-bank excess is not solely a sampling-precision cost. Full Top-2 already worsens S4 and V3 relative to FA, while partial replay slightly improves on full replay in V2. Added opportunity, validation ranking, and test transfer jointly determine the gain.

5.4 Validation support and the incremental value of pairing

More validation dates complement wider retention. With seven outer days, extending SEPTA inner validation from one to seven days lowers AA loss by 10.38 m; extending outer validation with seven inner days lowers it by 7.08 m. Extending both reduces selection cost from 51.80 to 34.43 m (Figure 3a,b). VTA changes are smaller (Table S5).

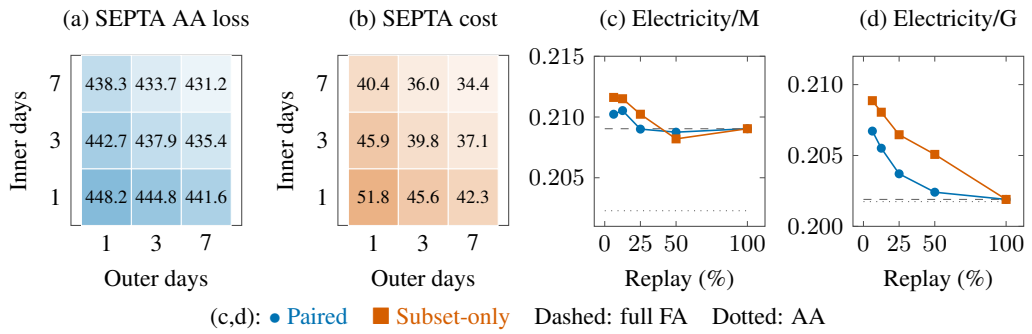

Figure 3: **Validation support and replay.** (a,b) SEPTA AA loss and selection cost (m); lighter cells indicate lower values. (c,d) Electricity standardized MAE, median (M) and fill+age (G); 200 matched subset draws, three seeds. FF: Table 2; agreement: Table S8.

Realistic subset evaluation supplies most of the hourly improvement; paired correction adds a smaller benefit. At four of sixteen Electricity blocks, pairing lowers loss by 0.001224 for median inputs and 0.002757 for fill+age relative to subset-only replay (Table S8). Traffic reaches full-FA loss with either rule at three of eleven blocks. At half-panel Electricity/M, subset-only achieves lower test loss despite matching the full-FA winner less often. Winner recovery therefore does not substitute for deployment loss.

The pairwise diagnostics explain why pairing helps unevenly (Table S5B). Electricity has lower correction variance and better winner agreement under pairing. Traffic's larger score margins allow both rules to recover the winner, whereas VTA has slightly higher correction variance and lower paired agreement. These summaries compare the full-panel winner with its nearest competitor; they diagnose difficulty without tuning the partial-replay selector. Appendices B.5–B.6 give the variance and regret analysis.

Table 5: **Protocol robustness and a longer prediction horizon.** Positive gain favors Top-2 at equal additional A candidate–query counts.

A. Transit robustness (m)

	SEPTA		VTA	
Protocol	C	Gain	C	Gain
Date-balanced	34.43	**4.48**	4.55	**0.10**
Query-weighted	33.10	**4.20**	4.30	**0.20**
28 fitting days	29.40	**3.60**	3.90	**0.10**
Native labels	32.80	**4.60**	4.60	**0.10**

B. Hourly $+23$-hour horizon

Standardized MAE

Task	AA	FF	C	C/AA
E/M	0.344	0.417	0.073	21.22%
E/G	0.340	0.392	0.052	15.29%
T/M	0.391	0.407	0.016	4.09%
T/G	0.389	0.402	0.013	3.34%

Shaded bold cells track retention gains across protocols. E/T: Electricity/Traffic; M/G: median/fill+age. Refitting, support reconstruction, and query-weighted reselection: Appendix C.3.

5.5 Sensitivity and transfer

SEPTA's equal-count gain persists under query-weighted reselection, 28-day refitting, and native-label support reconstruction; VTA gains remain small (Table 5). Query-weighted reselection changes both the supported panel and its weights (Appendix C.3). Ten epochs gives the lowest SEPTA mean loss among the four reported separations (Table S6); this sensitivity comparison did not determine the fixed policy.

Independently refitted $+23$-hour tasks have smaller selection contrasts (Table 5): controlled-delay effects are largest near the origin, where recent context matters most. This extends the evidence for information-sensitive selection across horizons; the equal-count retention comparison remains a transit result.

6 Discussion and conclusion

Information fidelity changes access as well as ranking. Our crossed design extends vintage-sensitive ranking studies (Pfarrhofer, 2024; Tang et al., 2025) by distinguishing exclusion from failures to select useful survivors within the same fitted collection. Improving outer evaluation can leave a deployment-loss gap because accurate rescoring cannot recover an excluded candidate.

Interventions address different selection bottlenecks. Replay evaluates survivors, retention preserves alternatives, and restriction removes retrospective fits. Their relative value depends on candidate composition and the trade-off between breadth and realistic date coverage (Li & Li, 2026). SEPTA's 6.20 m widening benefit exceeds its 1.73 m coverage cost; VTA's effects nearly cancel. Restriction nevertheless outperforms equal-count Top-2 on repeated SEPTA.

Implications for forecasting-library design. Save checkpoints and inner scores, reconstruct origin-available inputs, and compare common supported queries (Hewamalage et al., 2023; Kapoor & Narayanan, 2023). Fix retention rules and allowances using earlier periods (Cawley & Talbot, 2010). Compare complete policies by deployment loss, and proxy correction against subset-only replay. Test-based exclusion and winner-agreement diagnostics can explain outcomes and inform subsequent development; they do not replace deployment loss.

Scope and transfer. Observed receipt times cover two agencies and specified daily windows; repeated fitting adds periods, not agencies (Roque et al., 2025). Hourly restrictions are imposed. Finite-bank results condition on evaluated candidates, and calendar intervals exclude full-development uncertainty (Gonçalves et al., 2025). Evaluation counts do not equate inference, storage, or reconstruction costs. Transfer to Chronos-2 (Ansari et al., 2025) and TimesX (Liu et al., 2026) remains untested and requires recording release times, model versions, and selection decisions.

Evaluate retention and final selection together: replay can rank only surviving candidates, and wider retention helps only when the selected forecaster's deployment loss improves.

AI USE STATEMENT

Generative AI assisted manuscript proofreading, figure preparation, and LaTeX formatting and compilation checks. The authors are responsible for experiment, interpretation, references, and the final submission.

REPRODUCIBILITY STATEMENT

Appendix A specifies temporal roles, admission and target-maturity rules, fitted identities, selection policies, replay sampling, and resource accounting. Appendix B gives the finite-bank and sampling proofs. Appendices C and D provide complete transit episode endpoints, oracle decompositions, restricted-library comparisons, and hourly replay curves. Code for anonymous review is available at `https://anonymous.4open.science/r/historical-information-forecaster-selection-E247`.

ETHICS STATEMENT

The analysis uses public transport and aggregated forecasting archives without inferring individual passenger trajectories. Position-error reductions are forecasting outcomes, not estimates of passenger, service, or health benefits. Data access and redistribution are governed by source licenses and archive restrictions.

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, 2024. URL https://arxiv.org/abs/2410.10393.

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023. doi: 10.1126/science.adi6000.

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, Mononito Goswami, Shubham Kapoor, Danielle C. Maddix, Pablo Guerron, Tony Hu, Junming Yin, Nick Erickson, Prateek Mutalik Desai, Hao Wang, Huzefa Rangwala, George Karypis, Yuyang Wang, and Michael Bohlke-Schneider. Chronos-2: From univariate to universal forecasting, 2025. URL https://arxiv.org/abs/2510.15821. arXiv preprint.

Gavin C. Cawley and Nicola L. C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11(70):2079–2107, 2010. URL https://www.jmlr.org/papers/v11/cawley10a.html.

Ivi Chatzi, Eleni Straitouri, Suhas Thejaswi, and Manuel Gomez Rodriguez. Prediction-powered ranking of large language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 113096–113133. Curran Associates, Inc., 2024. doi: 10.52202/079017-3594. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/cd47cd67caa87f5b1944e00f6781598f-Abstract-Conference.html.

Dean Croushore. Forecasting with real-time macroeconomic data. In *Handbook of Economic Forecasting*, volume 1, pp. 961–982. Elsevier, 2006. doi: 10.1016/S1574-0706(05)01017-7.

Hui Feng. Real-time or current vintage: Does the type of data matter for forecasting and model selection? *Journal of Forecasting*, 28(3):183–193, 2009. doi: 10.1002/for.1089.

Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63:3–42, 2006. doi: 10.1007/s10994-006-6226-1.

Rakshitha Godahewa, Christoph Bergmeir, Geoffrey I. Webb, Rob J. Hyndman, and Pablo Montero-Manso. Monash time series forecasting archive. In *Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. URL https://arxiv.org/abs/2105.06643.

Sílvia Gonçalves, Michael W. McCracken, and Yongxu Yao. Bootstrapping out-of-sample predictability tests with real-time data. *Journal of Econometrics*, 247:105916, 2025. doi: 10.1016/j.jeconom.2024.105916.

Hansika Hewamalage, Klaus Ackermann, and Christoph Bergmeir. Forecast evaluation for data scientists: common pitfalls and best practices. *Data Mining and Knowledge Discovery*, 37(2): 788–832, 2023. doi: 10.1007/s10618-022-00894-5.

Kevin Jamieson and Ameet Talwalkar. Non-stochastic best arm identification and hyperparameter optimization. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pp. 240–248, 2016. URL https://proceedings.mlr.press/v51/jamieson16.html.

Jarvus Innovations. gtfsrt.io: Open GTFS-Realtime data archive, 2026. URL https://gtfsrt.io/. Archive and access documentation.

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*, 2022. URL https://arxiv.org/abs/2106.04420.

Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804, 2023. doi: 10.1016/j.patter.2023.100804.

Evan F. Koenig, Sheila Dolmas, and Jeremy Piger. The use and abuse of real-time data in economic forecasting. *The Review of Economics and Statistics*, 85(3):618–628, 2003. URL https://jeremypiger.com/assets/files/KDP_2003_RESTAT.pdf.

Ke Li and Fan Li. Multi-fidelity methods for optimization: A survey. *ACM Computing Surveys*, 58 (12):1–38, 2026. doi: 10.1145/3801959.

Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research*, 18(185):1–52, 2018. URL https://jmlr.org/papers/v18/16-558.html.

Haoxin Liu, Yichen Zhou, Rajat Sen, B. Aditya Prakash, and Abhimanyu Das. Rethinking multimodal time-series forecasting evaluation. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2607.06973.

MobilityData. GTFS Realtime reference, 2026. URL https://gtfs.org/documentation/realtime/reference/. Specification.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Jbdc0vTOcol.

Michael Pfarrhofer. Forecasts with Bayesian vector autoregressions under real time conditions. *Journal of Forecasting*, 43(3):771–801, 2024. doi: 10.1002/for.3055.

Zhongzheng Qiao, Sheng Pan, Anni Wang, Viktoriya Zhukova, Yong Liu, Xudong Jiang, Qingsong Wen, Mingsheng Long, Ming Jin, and Chenghao Liu. It's TIME: Towards the next generation of time series forecasting benchmarks. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.12147.

Xiangfei Qiu, Jilin Hu, Lekui Zhou, Xingjian Wu, Junyang Du, Buang Zhang, Chenjuan Guo, Aoying Zhou, Christian S. Jensen, Zhenli Sheng, and Bin Yang. TFB: Towards comprehensive and fair benchmarking of time series forecasting methods. *Proceedings of the VLDB Endowment*, 17(9): 2363–2377, 2024. doi: 10.14778/3665844.3665863.

Luis Roque, Vítor Cerqueira, Carlos Soares, and Luís Torgo. Cherry-picking in time series forecasting: How to select datasets to make your model shine. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 20192–20199, 2025. doi: 10.1609/aaai.v39i19.34224. URL https://ojs.aaai.org/index.php/AAAI/article/view/34224.

Carl-Erik Särndal, Bengt Swensson, and Jan Wretman. *Model Assisted Survey Sampling*. Springer, 1992. URL https://link.springer.com/book/9780387406206.

Oleksandr Shchur, Ali Caner Turkmen, Nick Erickson, Huibin Shen, Alexander Shirkov, Tony Hu, and Bernie Wang. AutoGluon-TimeSeries: AutoML for probabilistic time series forecasting. In *Proceedings of the Second International Conference on Automated Machine Learning*, volume 224 of *Proceedings of Machine Learning Research*, pp. 9/1–21, 2023. URL https://proceedings.mlr.press/v224/shchur23a.html.

Oleksandr Shchur, Abdul Fatir Ansari, Caner Turkmen, Lorenzo Stella, Nick Erickson, Pablo Guerron, Michael Bohlke-Schneider, and Yuyang Wang. fev-bench: A realistic benchmark for time series forecasting, 2025. URL https://arxiv.org/abs/2509.26468. arXiv preprint; version 4, June 2026.

Tom Stark and Dean Croushore. Forecasting with a real-time data set for macroeconomists. Technical Report 01-10, Federal Reserve Bank of Philadelphia, 2001. URL https://www.philadelphiafed.org/-/media/frbp/assets/working-papers/2001/wp01-10.pdf.

Jingjing Tang, Aaron Rumack, Bryan Wilder, and Roni Rosenfeld. Real-time forecasting of data revisions in epidemic surveillance streams. *PLOS Computational Biology*, 21(11):e1013709, 2025. doi: 10.1371/journal.pcbi.1013709.

Vijaya Krishna Yalavarthi, Kiran Madhusudhanan, Randolf Scholz, Nourhan Ahmed, Johannes Burchert, Shayan Jawed, Stefan Born, and Lars Schmidt-Thieme. GraFITi: Graphs for forecasting irregularly sampled time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 16255–16263, 2024. doi: 10.1609/aaai.v38i15.29560.

Weijia Zhang, Chenlong Yin, Hao Liu, Xiaofang Zhou, and Hui Xiong. Irregular multivariate time series forecasting: A transformable patching graph neural networks approach. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 60179–60196, 2024. URL https://proceedings.mlr.press/v235/zhang24bw.html.

A EXPERIMENTAL PROTOCOL

A.1 TEMPORAL ROLES, SUPPORTED QUERIES, AND MATURITY

For every origin, both histories exclude events after that origin. Availability determines which version of an earlier event may be used. Fitting, inner selection, outer selection, and testing occupy separate chronological roles. A target may enter a role only after its maturity rule is satisfied by the role cutoff. Preprocessing moments and imputation values are fitted on fitting rows alone. Each query key records source, entity, origin, horizon, and target definition. Comparisons take the intersection of supported keys before scoring; unsupported targets retain exclusion reasons rather than substituted outcomes.

Transit uses the consumer receipt timestamp rather than treating the producer event timestamp as a receipt (MobilityData, 2026; Jarvus Innovations, 2026). Timestamps are stored in UTC and dates use agency-local time. Thirty-second origins span ten-minute windows beginning at 07:00, 12:00, and 17:00. Context is 300 seconds, with a 420-second target buffer. Primary labels interpolate across brackets no more than 60 seconds apart; native labels use a five-second tolerance. Shared anchoring uses the last A-available position for both views and residual responses. Identical fallbacks apply when the required admitted positions are absent.

July fitting is July 6–7, 2026, with inner/outer selection July 8/9 and testing July 10–23. June SEPTA fits June 1–2, selects June 3/4, and tests June 5–18. June/July SEPTA fits pool SEPTA and VTA examples; VTA uses VTA-only examples. The repeated study uses only the named agency and is not numerically pooled with the diagnostic panels.

Table S1: **Repeated-development chronology.** Inclusive agency-local dates in 2026; all fits and both selection stages are rebuilt within each episode.

Episode	Fitting	Inner	Outer	Test
S1	Jan. 4–17	Jan. 18–24	Jan. 25–31	Feb. 1–14
V1	Jan. 18–31	Feb. 1–7	Feb. 8–14	Feb. 15–28
S2	Feb. 1–14	Feb. 15–21	Feb. 22–28	Mar. 1–14
V2	Feb. 15–28	Mar. 1–7	Mar. 8–14	Mar. 15–28
S3	Mar. 4–17	Mar. 18–24	Mar. 25–31	Apr. 1–14
V3	Mar. 18–31	Apr. 1–7	Apr. 8–14	Apr. 15–28
S4	Apr. 3–16	Apr. 17–23	Apr. 24–30	May 1–14
V4	Apr. 17–30	May 1–7	May 8–14	May 15–28

S/V denote SEPTA/VTA. Test periods are disjoint across episodes; development windows may overlap in calendar time. These are evaluated agency–time units, not independent draws from a population of agencies.

Every repeated validation date requires 2,048 supported keys. They are sampled without replacement using a fixed pseudorandom order determined only by source, episode, date, and key, with seed 20260922. Neither candidate score nor target value enters the sampling order. All histories and models use identical keys. A support-deficient episode is inadmissible before losses are inspected. All supported test keys are retained, with within-date averaging followed by equal date weight. Window interventions take the last one, three, or seven dates of each role without changing its endpoint.

A.2 FITTING RECIPES AND COMPLETE CANDIDATE IDENTITIES

Transit ExtraTrees has 64 estimators and minimum leaf size 20, with count, age, span, velocity, acceleration/turning, and fallback features (Geurts et al., 2006). GraFITi uses three channels, latent width 64, two layers, and four attention heads; its adapter predicts coordinates relative to the admitted anchor (Yalavarthi et al., 2024). AdamW uses learning rate and weight decay 10^{-3}, batch size 128, masked coordinate MSE, and 50 epochs. Epoch zero is eligible. Seeds are 20260912, 20260913, and 20260914. A/F branches share optimization budgets. Per seed, persistence, two trees, and 51 checkpoints from each graph trajectory give 105 identities.

Ties prefer the history-independent reference, tree-A, tree-F, graph-A, then graph-F, with earlier epoch within a trajectory. Top-1 retains one graph checkpoint per trajectory. Top-2 adds the lowest-F-inner-loss checkpoint at least ten epochs away; such an alternative exists for any winning epoch from zero through 50. Spread chooses the farthest eligible epoch and prefers the earlier epoch on

a distance tie. The A/F union deduplicates the two criterion-specific winners. Different seeds are repeated fits, not candidates competing in a common bank.

Hourly F context at origin d is $d-48,\dots,d-1$; A omits its last three rows. Targets are d or, in the separate task, $d+23$. Origins occur every six row-hours. Thirty-two series exclude the sixteen-series development panel and use 60/10/10/20% chronological roles. A matched eight-series panel varies delay over zero, one, three, and six hours. These restrictions are imposed, not measured release delays.

PatchTST has patch length eight, stride four, two encoder layers, four heads, latent width 64, and feed-forward width 128; a linear adapter adds missingness, age, calendar, and series metadata (Nie et al., 2023). Training uses 40 epochs, fitting-only standardization, and seeds 20260915–20260917. References are fitting mean, available persistence, and seasonal persistence. Ridge penalties are $0.1, 1, 10, 100$; ExtraTrees has 64 trees and minimum leaf size five or twenty. Median+mask and causal fill+age are distinct representations. Each horizon is independently fitted and selected, with target maturity determining role-end purges.

A.3 Composition controls and quality matching

A restricted three-candidate bank contains persistence, one tree history, and one graph history. All four A/F assignments preserve family counts and size. Uniform histories averages those assignments; the graph swap changes graph-A to graph-F while keeping tree-A fixed. Each surviving group retains its own criterion-specific winner, rather than exposing a union of checkpoints to both outer criteria.

Quality matching compares corresponding A/F fits after A-inner retention, admitting a pair when

$$\frac{|V_{\mathrm{A}}^{\mathrm{in}}(c_A) - V_{\mathrm{A}}^{\mathrm{in}}(c_F)|}{\min\{V_{\mathrm{A}}^{\mathrm{in}}(c_A), V_{\mathrm{A}}^{\mathrm{in}}(c_F)\}} \le 0.05.$$

The companion family stays A-fitted and persistence remains eligible. Tree and graph swaps across four episodes and three seeds supply 24 possible pairs per agency. Admitted pairs are equally weighted and both endpoints use that same subset. The screen controls the declared quality measure, not every candidate attribute. Its A-inner score computation is diagnostic work outside the inexpensive F-based retention policy.

A.4 Replay sampling, costs, and aggregation

Transit retention uses subset-only A scores; newly retained candidates need no F-outer scoring. Top-2 and Spread enumerate all $\binom{7}{5} = 21$ subsets, paired within seed. A deployment uses one subset; enumeration measures the rule's mean and variability as separate offline work. Table S2 gives per-deployment counts: Top-1 and partial Top-2 each use 71,680 A candidate–query evaluations, while full Top-2 uses 100,352.

Table S2: **Incremental evaluation requirements.** Each candidate–date evaluation scores 2,048 queries. The starting point is the completed mixed-library fits and F selection.

Policy	A-inner	A-outer	Ratio	Additional A dates
FF	0	0	0	None
A-only restriction / F	0	0	0	None
Top-1 / A	0	$5 \cdot 7$	1.00	7 outer
Top-2 or Spread / subset A	0	$7 \cdot 5$	1.00	5 outer
Top-2 / full A	0	$7 \cdot 7$	1.40	7 outer
AA	$102 \cdot 7$	$5 \cdot 7$	21.40	7 inner + 7 outer
A/F union	$102 \cdot 7$	$\le 7 \cdot 7$	≤ 21.80	7 inner + 7 outer

Only 102 graph checkpoints require A-inner search; singleton groups need no search. A-only F/F restricts the existing F-retained groups. Counts exclude fitting, storage, history reconstruction, and test-oracle evaluation, and do not equate runtime. Reconstruction and feature preparation occur once per date/history.

For unequal blocks, added A work is $J\sum_{k\in\mathcal{U}}|I_k|$, or Jmq on balanced blocks. The block policy is infeasible for $B < Jq$. Zero replay preserves FF; full replay returns the retained bank's A winner (FA for F-Top-1).

Hourly replay fixes nine F-inner alternatives: three references plus A/F Ridge, trees, and PatchTST. Seven-date blocks retain all series, with sixteen Electricity blocks and eleven Traffic blocks and a shorter final block where needed. Nominal fractions $.05, .10, .25, .50, 1$ round up to full blocks. The 200 matched draws share subsets across methods and seeds. Block-total weighting preserves the query-weighted criterion; full replay recovers FA under the same tie rule, not necessarily AA.

Transit pairing diagnostics use five F-inner candidates and all two-of-seven date subsets. Their per-agency agreement denominator is $4 \times 3 \times \binom{7}{2} = 252$. Ratios use the full-A winner and nearest distinct-score competitor per seed/episode, then average. Tied winners or zero denominators would be excluded with counts; displayed diagnostics have no such exclusions. Diagnostic-panel calendar intervals use 10,000 paired circular block draws, with three-day transit and seven-day hourly blocks retaining related series together. Repeated studies report all episodes and leave-one-out sensitivity rather than relabeling replay draws as independent replications.

Policies freeze an identity using their permitted selection scores before common-input testing and oracle diagnosis. Losses average queries within date, dates, replay subsets where applicable, and seeds.

B Analysis and proofs

The analysis conditions on the fitted candidate collection, supported queries, observed histories, and declared weights. R_A^{test} is empirical common-input test loss. Randomness in the replay statements comes from uniform sampling of calendar blocks, allowing arbitrary dependence among neighboring observations within the fixed panel.

B.1 Crossed stages and candidate exclusion

Substitution in Equation 2 gives $C = \Delta_{\text{in}} + \Delta_{\text{out}} + I$ and $L_{FF} - L_{FA} = \Delta_{\text{out}} + I$. The FA–AA residual can have either sign. Adding and subtracting $R^*(\mathcal{B})$ proves Equation 4; nonnegativity follows from $R_{\text{A}}^{\text{test}}(\widehat{f}) \geq R^*(\mathcal{B}) \geq R^*(\mathcal{C})$.

Nested banks have nonincreasing oracle exclusion but need not have decreasing validation-selected test loss. A candidate with validation/test losses $(0.2, 0.2)$ can be displaced by an added candidate with $(0.1, 0.3)$, although the bank's test minimum is unchanged.

B.2 Proof of Proposition 1

Let $f_B^* \in \arg\min_{f \in \mathcal{B}} R_{\text{A}}^{\text{test}}(f)$. Estimated-score optimality gives

$$V_{\text{A}}(\widehat{f}) \leq \widehat{V}(\widehat{f}) + \epsilon \leq \widehat{V}(f_B^*) + \epsilon \leq V_{\text{A}}(f_B^*) + 2\epsilon.$$

Adding $g_{\widehat{f}} - g_{f_B^*} \leq \omega_{\mathcal{B}}$ gives $R_{\text{A}}^{\text{test}}(\widehat{f}) - R^*(\mathcal{B}) \leq 2\epsilon + \omega_{\mathcal{B}}$. Equation 4 yields the upper bound, and membership yields the lower bound. No unique minimizer is required. With complete replay, $\epsilon = 0$; if g_f is also constant within the bank, its A-validation winner attains the bank's test minimum.

B.3 Proof of Proposition 2

Let j_2 minimize full A score in $\mathcal{B}_2$. Estimated-score optimality implies $V_{\text{A}}(\widehat{j}_2) - V_{\text{A}}(j_2) \leq 2\epsilon_2$. Since $V_{\text{A}}(j_2) = V_{\text{A}}(j_1) - \gamma$ and $j_1 \in \mathcal{B}_2$,

$$R_{\text{A}}^{\text{test}}(\widehat{j}_2) - R_{\text{A}}^{\text{test}}(j_1) \leq 2\epsilon_2 - \gamma + g_{\widehat{j}_2} - g_{j_1} \leq 2\epsilon_2 - \gamma + \omega_{\mathcal{B}_2}.$$

The reference is the narrow bank's full-A winner.

B.4 Widening, coverage, and realized gains

Writing $R_{\text{A}}^{\text{test}}(\widehat{f}_r) = R^*(\mathcal{B}_r) + S(\widehat{f}_r; \mathcal{B}_r)$ and subtracting proves Equation 7. With $S_r^{\text{full}} = L_r^{\text{full}} - R^*(\mathcal{B}_r)$ and $S_2^{\text{part}} = L_2^{\text{part}} - R^*(\mathcal{B}_2)$,

$$\begin{aligned} G &= \left(L_1^{\text{full}} - L_2^{\text{full}}\right) - \left(L_2^{\text{part}} - L_2^{\text{full}}\right) \\ &= D - \left(S_2^{\text{full}} - S_1^{\text{full}}\right) - \left(S_2^{\text{part}} - S_2^{\text{full}}\right) = W - H. \end{aligned}$$

Only H changes coverage within an unchanged bank. Unlike $D \geq 0$, W and H are signed. The identities survive averaging with common weights; oracle minima must be computed within seed first. A positive D need not imply positive G.

B.5 Finite-panel score correction

Define

$$X_{kj} = \frac{K}{N} \sum_{q \in I_k} [\ell_{\mathrm{A}}(j,q) - \ell_{\mathrm{F}}(j,q)], \qquad Y_{kj} = \frac{K}{N} \sum_{q \in I_k} \ell_{\mathrm{A}}(j,q).$$

Let $Z_k = \mathbf{1}\{k \in \mathcal{U}\}$ for a uniformly sampled m-subset $\mathcal{U}$ of the K blocks. Then $\mathbb{E}Z_k = m/K$ and $\mathbb{E}Z_k Z_l = m(m-1)/(K(K-1))$ for $k \neq l$. Hence

$$\mathbb{E}_{\mathcal{U}}\left[V_{\mathrm{F}}(j) + \frac{1}{m}\sum_k Z_k X_{kj}\right] = V_{\mathrm{F}}(j) + \frac{1}{K}\sum_k X_{kj} = V_{\mathrm{A}}(j).$$

The same calculation with Y proves subset-only unbiasedness. Unequal block sizes are incorporated through block totals and the K/N factor.

For fixed values $D_1, \ldots, D_K$ with $K > 1$, the inclusion covariances imply

$$\mathrm{Var}\left(\frac{1}{m}\sum_k Z_k D_k\right) = \left(\frac{1}{m} - \frac{1}{K}\right)\frac{1}{K-1}\sum_k (D_k - \overline{D})^2.$$

Substituting candidate differences gives

$$\mathrm{Var}(\widehat{V}_j - \widehat{V}_i) = \left(\frac{1}{m} - \frac{1}{K}\right)\begin{cases} S^2_{X_j - X_i}, & \text{paired,} \\ S^2_{Y_j - Y_i}, & \text{subset-only,} \end{cases} \tag{11}$$

where S^2 has denominator $K-1$. For $m < K$ and a positive denominator, pairing reduces this pair's variance exactly when $\rho_{ij} = S^2_{X_j - X_i}/S^2_{Y_j - Y_i} < 1$. If $S^2_{Y_j - Y_i} = 0$, the subset-only pair difference is already exact. Candidate-common noise cancels in differences, so marginal candidate variances do not determine ranking precision. Complete replay has zero sampling variance; when $K = 1$, exactness follows directly without the variance formula.

B.6 Margin-weighted replay regret

Proposition 3 (Margin-weighted replay regret)**.** *Let j_* uniquely minimize full A score in a fixed bank and let $\delta_j = V_{\mathrm{A}}(j) - V_{\mathrm{A}}(j_*) > 0$. For paired replay put $\tau_j^2 = (1/m - 1/K)S^2_{X_j - X_{j_*}}$. Then*

$$\mathbb{E}_{\mathcal{U}}[V_{\mathrm{A}}(\widehat{j}) - V_{\mathrm{A}}(j_*)] \leq \sum_{j \neq j_*} \delta_j \frac{\tau_j^2}{\tau_j^2 + \delta_j^2}. \tag{12}$$

Expected test excess above $R^*(\mathcal{C})$ is at most the right-hand side plus $E(\mathcal{B}) + \omega_{\mathcal{B}}$. Subset-only replay has the same result with X replaced by Y.

Proof. The estimated difference $\widehat{D}_j = \widehat{V}(j) - \widehat{V}(j_*)$ has mean δ_j and variance τ_j^2. A one-sided variance inequality gives

$$\Pr(\widehat{D}_j \leq 0) \leq \frac{\tau_j^2}{\tau_j^2 + \delta_j^2}.$$

Selecting j requires $\widehat{D}_j \leq 0$, including a fixed tie rule. Therefore

$$\mathbb{E}[V_{\mathrm{A}}(\widehat{j}) - V_{\mathrm{A}}(j_*)] = \sum_{j \neq j_*} \delta_j \Pr(\widehat{j} = j) \leq \sum_{j \neq j_*} \delta_j \frac{\tau_j^2}{\tau_j^2 + \delta_j^2}.$$

For the test statement, let f_B^* minimize test loss in the bank. Since $V_{\mathrm{A}}(j_*) \leq V_{\mathrm{A}}(f_B^*)$,

$$\begin{aligned} R_{\mathrm{A}}^{\mathrm{test}}(\widehat{j}) - R_{\mathrm{A}}^{\mathrm{test}}(f_B^*) &= [V_{\mathrm{A}}(\widehat{j}) - V_{\mathrm{A}}(j_*)] + [V_{\mathrm{A}}(j_*) - V_{\mathrm{A}}(f_B^*)] + g_{\widehat{j}} - g_{f_B^*} \\ &\leq V_{\mathrm{A}}(\widehat{j}) - V_{\mathrm{A}}(j_*) + \omega_{\mathcal{B}}. \end{aligned}$$

Take expectations and add $E(\mathcal{B})$. □

No independence across candidate comparisons is required. A union bound also yields

$$\Pr(\widehat{j} \neq j_*) \leq \min\left\{1, \sum_{j \neq j_*} \frac{\tau_j^2}{\tau_j^2 + \delta_j^2}\right\}.$$

The margin-weighted bound concerns validation regret; the latter bound concerns winner identity. At full replay the sampling regret vanishes. Tied full-panel minima are resolved by the declared tie order, while the positive-margin result assumes a unique minimum. Margins and variances refer to the full fixed panel and are used for interpretation rather than assumed known from a small replay subset.

B.7 General weights

For nonnegative weights w_q summing to one, replace the correction in Equation 10 by $(K/m)\sum_{k\in\mathcal{U}}\sum_{q\in I_k} w_q[\ell_{\mathrm{A}}(j,q) - \ell_{\mathrm{F}}(j,q)]$. Inclusion probabilities prove unbiasedness. Date-balanced weights $w_q = 1/(K|I_k|)$ coincide with query weights on balanced dates, but not on an hourly panel with an incomplete final block.

C Additional transit results

C.1 Episode endpoints, exclusion, and replay coverage

Table S3 reports crossed endpoints and retention policies on common A test keys; Table S4 separates oracle exclusion from within-bank selection loss. S4 and V3 worsen under full as well as partial Top-2 replay. V2's negative coverage effect means partial-panel choices transfer slightly better. The eight-episode mean gain is 2.2875 m, with leave-one-out means of 1.5143–2.8286 m.

Table S3: **Episode endpoints and replay-coverage effects.** Meters; three-seed means. Partial Top-2 and Spread average the same 21 five-date subsets.

	Crossed stages				Retention policies				Coverage effects		
					Top-2						
Ep.	AA	AF	FA	FF	Partial	Full	Spread	Union	W	H	G
S1	438.6	451.9	449.4	469.8	443.0	441.2	447.3	439.9	8.2	1.8	6.4
S2	412.8	434.4	422.1	455.4	416.8	415.0	420.7	413.7	7.1	1.8	5.3
S3	475.1	501.4	489.8	521.7	482.1	479.6	487.6	477.8	10.2	2.5	7.7
S4	398.4	404.1	402.0	415.7	403.5	402.7	403.1	399.8	-0.7	0.8	-1.5
V1	208.3	213.7	209.1	217.2	208.8	208.2	209.3	208.6	0.9	0.6	0.3
V2	192.4	198.1	193.0	198.5	192.8	192.9	193.4	192.7	0.1	-0.1	0.2
V3	224.6	223.9	224.4	224.1	225.0	224.8	224.7	224.3	-0.4	0.2	-0.6
V4	201.7	203.0	202.5	205.4	202.0	201.6	202.4	201.8	0.9	0.4	0.5

S/V: SEPTA/VTA. Partial Top-2 and Spread match FA's A-evaluation count; full Top-2 and union use all dates. $W = \mathrm{FA} - \mathrm{Full}$, $H = \mathrm{Partial} - \mathrm{Full}$, and $G = W - H$. Stage contrasts follow Equation 2; exclusion and within-bank loss appear in Table S4.

C.2 Composition, validation windows, and ranking diagnostics

Quality matching admits 18/24 SEPTA and 14/24 VTA family–episode–seed pairs; both endpoints use that same subset (Table 3A). Table S5A varies inner and outer windows independently; the seven-day setting matches the repeated-episode means. Panel B summarizes the full-A winner's nearest distinct-score competitor, one pair per fit rather than the entire bank. Full hourly loss and agreement curves are in Table S8.

C.3 Weighting, fitting duration, separation, and horizon

The query-weighted sensitivity in Table 5 uses all supported validation keys and reruns selection. Within selected dates, queries are sampled under an equal-query evaluation allowance and the actual evaluated count is charged. This changes the supported validation panel as well as its weighting; on

Table S4: **Oracle minima and within-bank selection loss.** Meters; minima are computed within seed before averaging.

	Oracle minima			Exclusion		Selection loss	
Episode	$R^*(\mathcal{C})$	$R^*(\mathcal{B}_1)$	$R^*(\mathcal{B}_2)$	E_1	E_2	S_1	S_2
S1	434.0	443.2	436.2	9.2	2.2	6.20	6.80
S2	409.5	416.7	411.0	7.2	1.5	5.40	5.80
S3	470.8	482.6	474.0	11.8	3.2	7.20	8.10
S4	395.7	398.5	397.5	2.8	1.8	3.50	6.00
V1	206.6	207.7	206.8	1.1	0.2	1.40	2.00
V2	190.9	191.7	191.1	0.8	0.2	1.30	1.70
V3	221.8	222.0	221.9	0.2	0.1	2.40	3.10
V4	199.8	200.9	200.0	1.1	0.2	1.60	2.00

$\mathcal{C}$ contains 105 identities; $\mathcal{B}_1/\mathcal{B}_2$ retain Top-1/Top-2. $E_k = R^*(\mathcal{B}_k) - R^*(\mathcal{C})$; S_1 uses FA and S_2 uses partial Top-2. The gain satisfies $G = (E_1 - E_2) - (S_2 - S_1)$; its episode values appear in Table S3.

Table S5: **Validation-window and candidate-ranking diagnostics.** A: Inner/outer windows vary independently with fixed role endpoints, trajectories, and test support.

A. Validation windows (m)

Days		SEPTA		VTA	
Inner	Outer	AA	C	AA	C
1	1	448.200	51.800	208.200	6.100
1	3	444.800	45.600	207.700	5.700
1	7	441.600	42.300	207.500	5.400
3	1	442.700	45.900	207.500	5.500
3	3	437.900	39.800	207.100	5.000
3	7	435.400	37.100	206.900	4.800
7	1	438.300	40.400	207.100	5.000
7	3	433.700	36.000	206.900	4.700
7	7	431.225	34.425	206.750	4.550

B. Ranking diagnostics

			Agreement (%)	
Task	Ratio	Margin	Paired	Subset
E/M	0.32	0.55	88.50	72.00
E/G	0.41	0.48	90.00	74.83
T/M	0.82	2.40	100.00	100.00
T/G	0.88	2.80	100.00	100.00
SEPTA	0.54	0.61	77.38	68.65
VTA	1.12	0.88	83.33	84.52

B: Ratio is correction-difference variance divided by A-loss-difference variance; Margin is $\delta/\sqrt{S^2_{Y_j - Y_{j*}}}$. Both average seed/episode diagnostics. Hourly agreement uses quarter-panel replay; transit uses two of seven dates. Transit counts are 195/252 versus 173/252 (SEPTA) and 210/252 versus 213/252 (VTA).

an identical balanced panel, date- and query-weighted means would coincide. The 28-day comparison extends fitting by 14 days while keeping later boundaries fixed and retraining both histories. Native-label comparisons reconstruct matched support and rerun both selection stages.

Table S6: **Checkpoint separation at equal A evaluation counts.** Position error in meters; ten epochs is the fixed primary policy.

Minimum separation (epochs)	1	5	10	20
SEPTA	438.90	437.40	436.35	438.10
VTA	207.35	207.30	207.15	207.60

Separation changes the eligible runner-up while leaving the candidate count and evaluation allowance fixed. Spread controls temporal distance separately from F-inner quality. The $+23$-hour comparisons independently fit and select for the longer-horizon target, retaining architecture, chronological role fractions, and fitting-only standardization.

D ADDITIONAL FORECASTING RESULTS

D.1 COMPLETE NESTED ENDPOINTS

Restricted banks retain their own criterion-specific group winners. Removing a group leaves surviving fits and winners unchanged, allowing reconstruction from saved groupwise scores. Table S7 reports the crossed endpoints on common deployment inputs.

Table S7: **Complete nested endpoints on the diagnostic panels.** Each restricted bank uses its own criterion-specific inner winners.

A. July transit: A-input position error (m)						
Task	Bank	AA	AF	FA	FF	C
SEPTA/O	A-only	350.420	352.739	352.739	355.615	5.195
	F-only	370.016	392.504	386.677	392.504	22.488
	Mixed	350.420	392.504	352.739	392.504	42.084
SEPTA/S	A-only	350.420	352.739	352.739	352.739	2.319
	F-only	371.545	375.465	371.545	429.274	57.729
	Mixed	350.420	375.465	352.739	429.274	78.853
VTA/O	A-only	197.581	196.275	197.195	196.275	-1.306
	F-only	196.590	196.318	196.875	196.875	0.284
	Mixed	196.696	196.318	196.594	196.875	0.178
VTA/S	A-only	197.581	196.275	196.275	196.275	-1.306
	F-only	197.184	200.300	198.307	199.580	2.396
	Mixed	197.581	200.300	196.275	199.580	1.998

B. Hourly: standardized MAE						
Task	Bank	AA	AF	FA	FF	C
E/M	A-only	0.202287	0.222052	0.209037	0.222052	0.019765
	F-only	0.283586	0.599467	0.283586	0.671726	0.388140
	Mixed	0.202287	0.599467	0.209037	0.671726	0.469439
E/G	A-only	0.201747	0.207591	0.201900	0.201900	0.000153
	F-only	0.283586	0.423619	0.283586	0.449019	0.165433
	Mixed	0.201747	0.423619	0.201900	0.449019	0.247272
T/M	A-only	0.217663	0.217663	0.217663	0.217663	$\equiv 0$
	F-only	0.284250	0.521125	0.284250	0.536678	0.252428
	Mixed	0.217663	0.521125	0.217663	0.536678	0.319015
T/G	A-only	0.215632	0.215632	0.215632	0.215632	$\equiv 0$
	F-only	0.284250	0.497926	0.284250	0.539211	0.254961
	Mixed	0.215632	0.497926	0.215632	0.539211	0.323580

O/S: original/shared anchoring; E/T: Electricity/Traffic; M/G: median/fill+age. Hourly rows use the origin-hour target, three-hour restriction, and 32 series. Costs retain unrounded source results; displayed rounding may cause a last-digit difference.

SEPTA A-only cost has conditional calendar intervals [-0.704,4.047] m under shared anchoring and [2.171,6.931] m under original anchoring. Electricity/G's A-only interval is [-0.000061,0.000374]. Traffic's A-only forecasts coincide under all four criteria; this exact null is distinct from an interval containing zero. Diagnostic-panel and VTA seed/calendar intervals appear in Table 2.

D.2 FULL HOURLY REPLAY CURVES

Table S8 reports the actual replay-block counts; an incomplete final block makes block and query fractions differ. With zero delay, all twelve source/representation/library comparisons preserve selected forecasts exactly.

Table S8: **Complete hourly replay curves.** Paired and subset-only losses, with agreement against the full-FA identity.

		Median (M)				Fill+age (G)			
		A-input MAE		Agreement (%)		A-input MAE		Agreement (%)	
Task	Blocks	Paired	Subset	Paired	Subset	Paired	Subset	Paired	Subset
E	1/16	0.210228	0.211612	73.17	62.17	0.206721	0.208867	73.33	61.50
	2/16	0.210528	0.211504	80.33	64.33	0.205513	0.208052	80.00	66.17
	4/16	0.209004	0.210228	88.50	72.00	0.203707	0.206464	90.00	74.83
	8/16	0.208752	0.208200	98.17	84.83	0.202409	0.205076	97.17	82.50
	16/16	0.209037	0.209037	100.00	100.00	0.201900	0.201900	100.00	100.00
T	1/11	0.219393	0.222990	95.00	92.00	0.215632	0.221464	100.00	91.50
	2/11	0.217663	0.219661	100.00	97.00	0.215632	0.218033	100.00	96.50
	3/11	0.217663	0.217663	100.00	100.00	0.215632	0.215632	100.00	100.00
	6/11	0.217663	0.217663	100.00	100.00	0.215632	0.215632	100.00	100.00
	11/11	0.217663	0.217663	100.00	100.00	0.215632	0.215632	100.00	100.00

E/T: Electricity/Traffic. Means use 200 matched block draws and three fitted seeds. A shorter final block makes query and block fractions differ.