Online Resource 1

Budgeted Day-Ahead Reserve Allocation from Conformal Forecasts under Delayed Feedback

International Journal of Machine Learning and Cybernetics

Ran Wang
School of Computer Science and Technology, Hangzhou Dianzi University
Hangzhou 310018, Zhejiang, China
Corresponding author: wangran@hdu.edu.cn

This supplement accompanies *Budgeted Day-Ahead Reserve Allocation from Conformal Forecasts under Delayed Feedback*. The current study is presented first: Section A gives the event contract and exact policies, Section B the complete annual and reserve-response results, and Section C the analytic proofs. Sections D–H preserve the earlier pipeline evaluations as secondary evidence with their original cohorts and statistical contracts. Their observation-index feedback adapter differs from the current declared-calendar adapter; each pipeline is reported with its own numerical rows and evaluation support. Section H.8 additionally carries the synthetic range checks and active-premise counterexamples for the current study's propositions.

A Current study: event contract, policies, and grids

This section states the executable semantics of the end-to-end design of the main article at the level needed to reproduce it from the description, with the policy definitions matched to the released reproduction code.

Daily availability. For every declared delivery date d, the expected hours are that date's $H_d \in \{23, 24, 25\}$ local clock hours. A date is *issuable* (`scorable`) when it has at least one expected hour and every expected hour has a finite issued point forecast; a date is *observed* when every expected hour has a finite actual. Dates with partial actuals are recorded as partial: they never enter modeled service or the physical window endpoints, while their finite issued zone-hour rows still contribute to the annual uncovered energy and their committed hours still book capacity-time. Missing rows are never imputed.

Shared event list. Two events, both derived from the declared calendar and the availability maps, drive the region adapter and every policy:

- `issuance_absent`(d): date d is not issuable. It is known at the decision for d, so the reset takes effect at d itself; d has no decision and can never be scored.
- `missing_outcome`(u): date u is issuable but not observed. It is discovered at the reveal date $u + 2$, so the reset takes effect there.

A reset starts a new segment, drops every pending release record whose source segment is the old one, clears the shortfall history, the bank, and the tension rank history, and never restores spent budget. Commitments made on a date that later stops being observed remain booked.

Policy state and commit rule. Each policy g walks the declared calendar once with its own history. At the decision for an issuable date d, the eligible records are those with source date at most $d - \ell$ *and* the same segment, where $\ell = 2$; the state $\sigma_g(d)$ is computed from them and the commit rule is applied. Table S1 lists the six action policies exactly; the banked variant is stated in full because its transition is the only one with memory beyond the state.

Table S1 Exact policy definitions. Each policy shares the declared calendar, the event list, the eligibility rule ($\ell = 2$, same segment), and the cap B; only the state and the commit rule differ.

Policy	State at the decision for d	Commit rule
none	—	never commits; the walk records zero commitments, zero blocked days, and the same event-derived segments as every other policy
P	$h = \sum$ of the last $m = 6$ eligible shortfalls $1 - s$	commit iff $h \geq \theta = 1$ and used $< B$; otherwise record a blocked trigger iff $h \geq \theta$
F1	scheduled flag at within-segment accessible index i (first issuable day of the segment has $i = 0$)	commit iff $i \geq 6$ and $(i - 6) \bmod 7 = 0$ and used $< B$; otherwise record a cap skip iff scheduled and the cap is exhausted
F2	credit bank $b \in \{0, 1, 2, \dots\}$, incremented by 15 at every issuable date before the rule is applied	if $b \geq 100$: commit and $b \leftarrow b - 100$ when used $< B$, otherwise record a cap skip; if $b < 100$: no commitment
F3	$k = \#\{$of the last six eligible days with $s < 0.9\}$	commit iff $k \geq 2$ and used $< B$; otherwise record a blocked trigger iff $k \geq 2$
F4	$s_{\text{last}} =$ service of the most recent eligible day	commit iff $s_{\text{last}} \leq 0.5$ and used $< B$
F5	bank $c \in \{0, 1\}$; tension rank history of the current segment	see the transition below

The cap is shared and nonreplenishable; a commitment is recorded whether or not the day is later observed. "Cap skip" and "blocked" are the same event under different policy labels; both mean that the policy's own rule called for a commitment and the budget refused it.

F5 transition. F5 uses the production rule P as its shadow path and the pre-action tension τ_d of the date, which is computed at the decision and never depends on the current or future actual. For an issuable date d with remaining days R_d through December 31,

$$\text{flush}_d = \mathbf{1}\{R_d \leq B + 1\}, \qquad \text{high}_d = \mathbf{1}\{\rho_d \leq 0.5\},$$

with the tension rank of the day,

$$\rho_d = \frac{1 + \#\{v \geq \tau_d\}}{1 + \#\{\text{segment history}\}},$$

so high_d marks high tension under this regularized rank rule: few prior days had an equal-or-higher tension. The bank and commitment update is

$$(a_d,\ c_{d+1}) = \begin{cases} (1,\ \text{mint}_d), & c_d = 1, \\ (0,\ 0), & c_d = 0,\ \text{mint}_d = 0, \\ (1,\ 0), & c_d = 0,\ \text{mint}_d = 1,\ (\text{high}_d \text{ or } \text{flush}_d), \\ (0,\ 1), & \text{otherwise}, \end{cases}$$

where mint_d is the shadow action of P on date d. A banked credit is therefore executed on the next issuable day whatever that day's own state is, and if that day also mints, the new mint is retained in the bank in place of the spent credit. The tension is then inserted into the rank history. A reset discards a banked credit, which is why the executed commitment count of F5 can fall below B.

Tension. For each hour slot the frozen threshold of the decision is ranked against that slot's own released threshold history, $p = (1 + \#\{v \leq \text{thr}\})/(1 + \#\{\text{history}\})$, and $\tau_d = \text{mean}_{\text{slots}}(-\log p)$, computed on every issuable date from pre-decision information only, whether or not the date's actual exists.

Shared walk. Algorithm 1 states the skeleton that all seven policies execute; the policy-specific state and commit rule are the ones tabulated above.

Algorithm 1 Declared-calendar walk shared by all policies (production code semantics).

init: used $\leftarrow 0$; empty release queue; empty method-owned history; segment $\leftarrow$ none
build `scorable`$[d]$, `observed`$[d]$ for every declared date d; build the event list
for declared date d in chronological order **do**
    apply `issuance_absent`(d): start a new segment, drop old-segment pending records, clear history, bank, rank history
    release queued records with source date $\leq d - \ell$ in chronological order into the history (same segment only)
    apply `missing_outcome` resets whose reveal date is d
    if `scorable`$[d]$ **then**
        freeze the day's slot thresholds from the released scores
        $i \leftarrow$ within-segment accessible index; $\sigma \leftarrow$ policy state from the eligible history
        apply the policy's commit rule, respecting used $< B$; record blocked or cap-skipped when the rule fires and the budget is exhausted
        if committed **then** used $\leftarrow$ used $+ 1$
        end if
    end if
After delivery of day d, and only if `scorable`$[d]$ *and* `observed`$[d]$*:*
    compute q_d and $s_d \leftarrow a_d + (1 - a_d)q_d$
    queue the day's scores and s_d, tagged with their source segment, for the decision indexed $d + \ell$
end for

Availability maps for the offline simulation are built ahead of the walk from the data support, but a missing outcome is a property of the future and becomes visible to every policy only at its reveal date, when the `missing_outcome` event fires; no policy sees a future support mask.

Reserve grids. The response multipliers $\eta \in \{0, 0.5, 1, 2\}$ and the nine declared price ratios from $\kappa = 0$ to $\kappa = 0.5$ are reported in full for every policy and origin; H_z is the reference-year 95th percentile (linear interpolation) of the positive corrected reference residual; deficits, adequacy, and bookings follow the reserve-requirement, deficit, and cost equations of the main article, and the finite-hour support of a given base is identical across policies by construction.

Table S2 Declared quantities of the end-to-end design: symbols, units, and pre-declared grids. Every value is fixed before evaluation support is read.

Symbol	Meaning	Unit or value
T_{planned}	declared calendar length, Jan 2–Dec 31	364 or 365 days; $H_t \in \{23, 24, 25\}$ hours per day
B	horizon commitment cap	committed days, $\text{round}(0.15\, T_{\text{planned}}) = 55$; nonreplenishable
ℓ	feedback lag of a completed daily outcome	2 delivery days
$m,\ \theta$	memory and threshold of the production rule P	6 records; $\theta = 1.0$ normalized shortfall
policies	commitment menu	none, P, F1 pace-7, F2 bank-pace, F3 two-low, F4 persistence, F5 declared-end bank
$e_{t,h,z}$	residual, actual minus issued point	MW, main-text residual equation
$r^0_{t,h,z}$	no-commitment zonal requirement	MW, main-text base-reserve equation
H_z	reference headroom scale	MW, reference-year 95th percentile of the positive corrected residual, main-text headroom equation
η	capacity multiplier on committed days	dimensionless, $\{0, 0.5, 1, 2\}$
Δ_t	uncovered deficit of a day	MWh, positive part integrated over real hours, main-text deficit equation
adequacy	one-sided all-zone adequacy	indicator per hour; daily mean over complete observed hours
κ	capacity-to-deficit price ratio	dimensionless, nine declared values (table note), main-text cost equation
λ	modeled penalty per committed day	$\{0.25, 0.5, 1, 2, 4\}$ on the modeled-shortfall scale, main-text modeled-loss equation
inference	paired per-origin intervals	10,000 IID-over-origin bootstrap draws, seed 20261011, plus leave-one-origin-out means

The nine declared κ values are 0, 0.001, 0.005, 0.01, 0.025, 0.05, 0.1, 0.25 and 0.5. Capacity quantities are capacity-time (MW·h) booked on committed days. η scales the reference headroom, while $\kappa = c_R/c_D$ converts booked capacity into an MWh-equivalent normalization using declared price ratios.

Algorithm 2 The production policy P on the declared calendar.

input: $\theta = 1.0$; $m = 6$; $\ell = 2$; $B = \mathrm{round}(0.15\,T_{\mathrm{planned}})$; used $\leftarrow 0$; empty release queue and shortfall history
for declared delivery date d, in chronological order **do**
  if an input gap for d is known at the cutoff **then**
    reset the segment, discard old-segment pending records
  end if
  At the cutoff for d:
  release queued same-segment records with source date $\leq d - \ell$
  apply resets for missing outcomes whose reveal date is d
  freeze the day's region thresholds from the released scores
  if d is issuable **then**
    $h \leftarrow$ sum of the m most recent eligible same-segment shortfalls
    if $h \geq \theta$ **and** used $< B$ **then**
      $a_d \leftarrow 1$; used $\leftarrow$ used $+ 1$
    else
      $a_d \leftarrow 0$; record a blocked trigger if $h \geq \theta$
    end if
  end if
  After delivery:
  if d is issuable **and** completely observed **then**
    compute q_d; $s_d \leftarrow a_d + (1 - a_d)q_d$
    enqueue the day's scores and s_d, tagged with their source segment, for release at $d+\ell$
  end if
end for

B End-to-end study: complete results

Every number in this section is emitted by the frozen pipeline on the seven exposed annual origins and is reproduced by the asset generator named in Section B.5; the tables here are the complete versions of the main-text summaries, and the machine-readable aggregate carries every remaining cell.

B.1 Modeled endpoints for all policies

Table S3 Per-origin modeled endpoints for all seven declared policies.

Origin	Policy	Mean	W_3	W_7	Actions	Blocked
2019	none	0.9114	0.4306	0.565476	0	0
2019	P	0.9206	0.4306	0.732143	43	0
2019	F1	0.9280	0.4861	0.613095	52	0
2019	F2	0.9236	0.4861	0.636905	54	0
2019	F3	0.9228	0.4306	0.565476	55	83
2019	F4	0.9143	0.4306	0.690476	14	0
2019	F5	0.9184	0.4306	0.726190	43	0
2020	none	0.9089	0.3194	0.642857	0	0
2020	P	0.9143	0.3194	0.648810	32	0
2020	F1	0.9235	0.5556	0.750000	52	0
2020	F2	0.9190	0.3194	0.672619	54	0
2020	F3	0.9192	0.3194	0.642857	55	147
2020	F4	0.9098	0.3194	0.648810	8	0
2020	F5	0.9152	0.3194	0.642857	32	0
2021	none	0.9109	0.4861	0.696429	0	0
2021	P	0.9189	0.4861	0.755952	40	0

Continued on next page

Table S3 (continued)

Origin	Policy	Mean	W_3	W_7	Actions	Blocked
2021	F1	0.9234	0.4861	0.732143	49	0
2021	F2	0.9261	0.5972	0.755952	52	0
2021	F3	0.9225	0.4861	0.696429	55	111
2021	F4	0.9123	0.4861	0.732143	7	0
2021	F5	0.9190	0.4861	0.755952	40	0
2022	none	0.8930	0.1250	0.404762	0	0
2022	P	0.9106	0.3194	0.708333	54	0
2022	F1	0.9091	0.3194	0.547619	52	0
2022	F2	0.9079	0.1250	0.488095	54	0
2022	F3	0.9069	0.1250	0.404762	55	117
2022	F4	0.8978	0.3611	0.553571	11	0
2022	F5	0.9109	0.3194	0.708333	54	0
2023	none	0.9085	0.4583	0.696429	0	0
2023	P	0.9215	0.4583	0.738095	53	0
2023	F1	0.9172	0.4583	0.696429	52	0
2023	F2	0.9239	0.4583	0.720238	54	0
2023	F3	0.9222	0.4583	0.696429	55	100
2023	F4	0.9120	0.4583	0.696429	14	0
2023	F5	0.9207	0.4583	0.714286	53	0
2024	none	0.8983	0.3056	0.583333	0	0
2024	P	0.9139	0.3056	0.678571	55	6
2024	F1	0.9178	0.4722	0.607143	52	0
2024	F2	0.9171	0.3056	0.636905	54	0
2024	F3	0.9134	0.3056	0.583333	55	99
2024	F4	0.9030	0.3056	0.630952	16	0
2024	F5	0.9132	0.3056	0.678571	55	0
2025	none	0.9003	0.2500	0.482143	0	0
2025	P	0.9115	0.5417	0.732143	33	0
2025	F1	0.9173	0.2500	0.577381	52	0
2025	F2	0.9158	0.4722	0.577381	54	0
2025	F3	0.9163	0.2500	0.482143	55	138
2025	F4	0.9045	0.4722	0.619048	12	0
2025	F5	0.9133	0.5417	0.732143	33	0

Higher mean, W_3 and W_7 are better. Actions and blocked triggers are counts over that origin's declared calendar; the cap is $B = 55$ committed days.

B.2 Complete modeled contrast set

Table S4 Complete paired annual contrast set for modeled endpoints: candidate minus comparator, 10,000-draw IID-over-origin bootstrap (seed 20261011), leave-one-origin-out range of the mean, and the count of strictly positive per-origin annual effects.

Contrast	Endpoint	Mean effect	95% interval	LOO range	Annual > 0
P − none	mean	0.0114	[0.0085, 0.0143]	[0.0104, 0.0124]	7/7
P − none	W_3	0.0694	[0.0000, 0.1528]	[0.0324, 0.0810]	2/7 (5 ties)
P − none	W_7	0.1318	[0.0595, 0.2117]	[0.1032, 0.1528]	7/7
P − none	W_{14}	0.0591	[0.0259, 0.0969]	[0.0466, 0.0670]	7/7
P − none	L_1	0.1108	[0.0953, 0.1259]	[0.1067, 0.1160]	7/7
P − F1	mean	-0.0036	[-0.0067, -0.0002]	[-0.0049, -0.0027]	2/7
P − F1	W_3	-0.0238	[-0.1310, 0.0933]	[-0.0764, 0.0116]	1/7 (3 ties)
P − F1	W_7	0.0672	[0.0017, 0.1241]	[0.0516, 0.0952]	6/7
P − F1	W_{14}	0.0264	[0.0055, 0.0468]	[0.0203, 0.0337]	5/7 (1 tie)
P − F1	L_1	-0.0165	[-0.0321, -0.0012]	[-0.0213, -0.0116]	2/7
P − F2	mean	-0.0031	[-0.0050, -0.0008]	[-0.0041, -0.0025]	1/7
P − F2	W_3	0.0139	[-0.0476, 0.0853]	[-0.0162, 0.0347]	2/7 (3 ties)
P − F2	W_7	0.0723	[0.0153, 0.1378]	[0.0476, 0.0883]	5/7 (1 tie)
P − F2	W_{14}	0.0268	[0.0017, 0.0553]	[0.0164, 0.0332]	4/7 (1 tie)
P − F2	L_1	-0.0229	[-0.0395, -0.0064]	[-0.0278, -0.0175]	1/7
P − F3	mean	-0.0017	[-0.0037, 0.0005]	[-0.0026, -0.0012]	2/7
P − F3	W_3	0.0694	[0.0000, 0.1528]	[0.0324, 0.0810]	2/7 (5 ties)
P − F3	W_7	0.1318	[0.0595, 0.2117]	[0.1032, 0.1528]	7/7
P − F3	W_{14}	0.0591	[0.0259, 0.0969]	[0.0466, 0.0670]	7/7
P − F3	L_1	-0.0279	[-0.0446, -0.0117]	[-0.0325, -0.0229]	0/7
P − F4	mean	0.0082	[0.0063, 0.0103]	[0.0075, 0.0089]	7/7
P − F4	W_3	0.0040	[-0.0179, 0.0298]	[-0.0069, 0.0116]	1/7 (5 ties)

Contrast	Endpoint	Mean effect	95% interval	LOO range	Annual > 0
P − F4	W_7	0.0604	[0.0272, 0.0995]	[0.0446, 0.0704]	6/7 (1 tie)
P − F4	W_{14}	0.0196	[0.0047, 0.0391]	[0.0109, 0.0228]	5/7 (2 ties)
P − F4	L_1	0.0817	[0.0672, 0.0950]	[0.0778, 0.0869]	7/7
P − F5	mean	0.0001	[-0.0008, 0.0010]	[-0.0003, 0.0004]	3/7
P − F5	W_3	0.0000	[0.0000, 0.0000]	[0.0000, 0.0000]	0/7 (7 ties)
P − F5	W_7	0.0051	[0.0009, 0.0119]	[0.0020, 0.0060]	3/7 (4 ties)
P − F5	W_{14}	0.0009	[-0.0034, 0.0051]	[-0.0010, 0.0020]	2/7 (3 ties)
P − F5	L_1	-0.0001	[-0.0010, 0.0008]	[-0.0004, 0.0003]	4/7

Positive effects favor the production rule on mean, W_3, W_7 and W_{14}; for L_1, which adds the abstract penalty per committed day, a negative effect favors the production rule. The annual column counts origins with a strictly positive paired annual difference on that endpoint, so a positive mean, interval or leave-one-origin-out range must not be read as every origin being positive.

B.3 Complete finite zonal reserve response grid

Table S5 Complete finite zonal reserve response grid by policy and declared capacity multiplier η: annual means over seven exposed origins.

Policy	η	Uncovered (MWh/yr)	W_7 deficit (MWh)	W_7 adequacy (fraction)	Extra booked (10^6 MW·h/yr)
none	0	1305.4	606.3	0.9549	0.0000
none	0.5	1305.4	606.3	0.9549	0.0000
none	1	1305.4	606.3	0.9549	0.0000
none	2	1305.4	606.3	0.9549	0.0000
P	0	1305.4	606.3	0.9549	0.0000
P	0.5	1066.9	516.8	0.9575	0.5618
P	1	967.0	516.7	0.9592	1.1236
P	2	937.6	516.7	0.9592	2.2472
F1	0	1305.4	606.3	0.9549	0.0000
F1	0.5	1219.7	590.5	0.9592	0.6545
F1	1	1197.2	590.5	0.9592	1.3090
F1	2	1185.8	590.5	0.9592	2.6179
F2	0	1305.4	606.3	0.9549	0.0000
F2	0.5	1010.2	387.9	0.9592	0.6816
F2	1	861.6	323.2	0.9626	1.3631
F2	2	818.7	316.3	0.9626	2.7262
F3	0	1305.4	606.3	0.9549	0.0000
F3	0.5	1262.7	603.1	0.9549	0.6973
F3	1	1254.2	603.1	0.9549	1.3947
F3	2	1254.0	603.1	0.9549	2.7894
F4	0	1305.4	606.3	0.9549	0.0000
F4	0.5	1209.3	528.0	0.9549	0.1497
F4	1	1152.1	528.0	0.9549	0.2994
F4	2	1149.2	528.0	0.9549	0.5987
F5	0	1305.4	606.3	0.9549	0.0000
F5	0.5	1128.3	516.8	0.9575	0.5618
F5	1	1047.3	516.7	0.9592	1.1236
F5	2	1033.4	516.7	0.9592	2.2472

Lower uncovered energy and lower W_7 deficit are better; higher W_7 adequacy is better. The $\eta = 0$ row of each policy is identical by construction, because a policy books no extra capacity without a commitment, so those rows restate the shared no-commitment physical baseline. The table covers all 28 declared policy and multiplier combinations; together with the nine declared price ratios and the seven origins, these are the 1,764 cells of the admitted machine-readable grid.

Concentration of the reduction. The rule protects several high-deficit events that drive the aggregate annual energy reduction. At the declared multiplier $\eta = 1$, twelve of the 310 committed days across the seven origins reduce the physical deficit of their own origin, and three of them lie in 2025 (2025-05-06, 2025-06-23 and 2025-06-24, saving 573.1, 255.4 and 1063.2 MWh respectively), which is where the 1891.7 MWh 2025 total comes from. The per-origin annual deficit improves in five of seven origins and is unchanged in two; the physical worst-seven-day deficit improves in two origins, 2023 and 2025, and is unchanged in the other five.

Table S6 Paired annual physical contrasts at $\eta = 1$. Effects are candidate minus comparator in the unit shown; intervals and the LOO range are computed as in Table S4.

Contrast	Endpoint	Unit	Effect	95% interval	LOO range
P − none	Uncovered	MWh/yr	-338.43	[-868.05, -23.29]	[-394.83, -79.54]
P − none	W_7 deficit	MWh	-89.60	[-246.07, 0.00]	[-104.53, -13.25]
P − none	W_7 adequacy	fraction	0.0043	[0.0000, 0.0094]	[0.0020, 0.0050]
P − none	Extra booked	10^6 MW·h/yr	1.1236	[0.9446, 1.2924]	[1.0745, 1.1855]
P − F2	W_7 deficit	MWh	193.51	[12.03, 430.36]	[92.67, 225.76]
P − F2	Uncovered	MWh/yr	105.35	[-253.56, 444.58]	[-10.19, 260.99]

A negative uncovered or deficit effect is better; a positive adequacy effect is better. The W_7 deficit and adequacy intervals touch zero on their favorable side, while the annual uncovered reduction excludes zero. The $(P - F2)$ rows quantify the physical-protection/capacity trade-off between these policies. Inference follows the main-text study design.

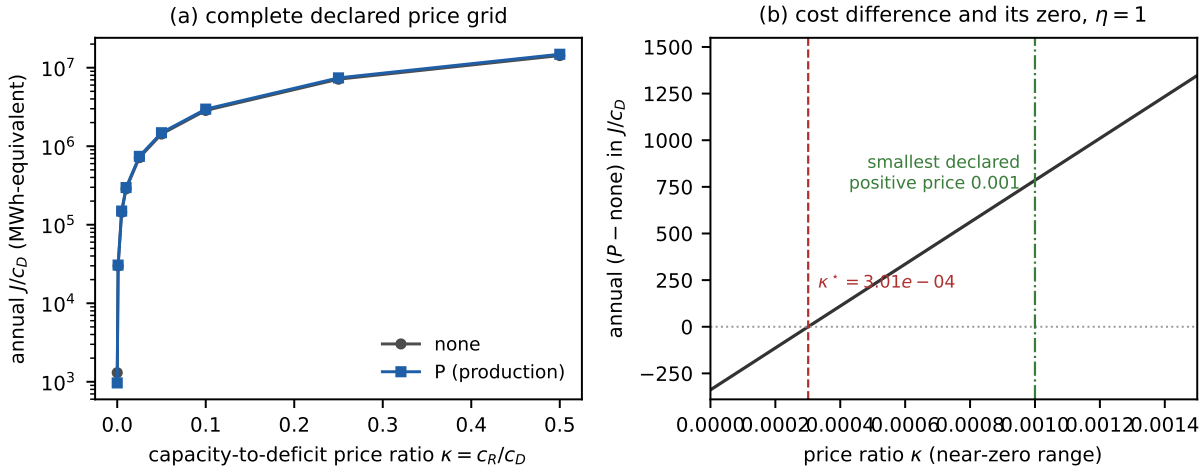


Figure S1 Normalized annual cost J/c_D in MWh-equivalent units at $\eta = 1$. (a) The complete declared nine-point price grid for the production rule and the no-extra-commitment comparator. (b) The production-minus-comparator difference over the near-zero range, with the crossing $\kappa^\star$ where the two normalized costs are equal and the smallest declared positive price marked; the difference is a post-run algebraic diagnostic of the fixed cost model. Higher κ favors the no-extra-commitment comparator in this normalized objective. Prepared with AI-assisted code

B.4 Normalized cost grid

	none		P	
κ	$\eta = 1$	$\eta = 2$	$\eta = 1$	$\eta = 2$
0	1,305	1,305	967	938
0.001	29,856	29,856	30,641	31,736
0.005	144,060	144,060	149,339	154,928
0.01	286,814	286,814	297,712	308,918
0.025	715,077	715,077	742,829	770,890
0.05	1,428,848	1,428,848	1,484,690	1,540,842
0.1	2,856,391	2,856,391	2,968,414	3,080,746
0.25	7,139,020	7,139,020	7,419,584	7,700,458
0.5	14,276,734	14,276,734	14,838,201	15,399,978

Normalized annual cost J/c_D in MWh-equivalent units at the declared price ratios, for the two policies and the two largest declared response multipliers; the $\eta = 0.5$ and $\eta = 0$ columns and the remaining policies are in the machine-readable aggregate. Higher κ makes booked capacity relatively more expensive.

B.5 Machine-readable results and asset provenance

The end-to-end artifacts comprise the aggregate summary, the seven per-origin payloads with their manifests and sidecars, the frozen protocol with its dated amendments, the guard results, the complete declared grid of policy, response and price cells, and the generator that produces every table and figure of the main article and this supplement together with the SHA-256 manifest of the files it reads. Table and figure cells are computed from those artifacts rather than transcribed, and the generator fails closed if the admitted fingerprint does not match.

C Analytic proofs of the main-text local deficit results

The statements and operating implications remain in Section 5 of the main article. This section preserves their complete analytic proofs; the synthetic range checks and active-premise counterexamples remain in Section H.8. Proof numbering below follows the main-text propositions. Main-text propositions and the corollary share one counter, so the coupling bound is Proposition 4.

C.1 Proof of main-text Proposition 1

Proof. Cauchy–Schwarz in the Σ^{-1} inner product gives $(e_z^{\mathsf{T}}(e-\mu))^2 \le \Sigma_{zz}\,(e-\mu)^{\mathsf{T}}\Sigma^{-1}(e-\mu) \le \Sigma_{zz}R$ for $R \ge 0$. The bound is attained at $e = \mu + c\,\Sigma e_z$ with $c = \sqrt{R/\Sigma_{zz}}$, which lies on the boundary and satisfies the inequality with equality. □

C.2 Proof of main-text Proposition 2

Proof. Let j be the last position in the window with $d_{w+j} > 0$ whose trigger was *not* blocked, and write $j = -\infty$ if no such position exists. Every positive-shortfall day after position j is blocked by maximality of j, so the tail contributes at most $\sum_{i\in T} d_{w+i}$ and it remains to bound the head, the positive-shortfall days with index at most j.

If $j = -\infty$ the head is empty. Otherwise, day $w + j$ has positive shortfall and its trigger was not blocked; under the full-service model that is only possible if its state was below threshold, $h_{w+j} < \theta$. The eligible in-window records of day $w+j$ are exactly the indices $w, \dots, w+j-\ell$, which are at most $n-\ell \le m$ in number and therefore all among the m records the state sums; any eligible record from before the window only adds a nonnegative term, so $\sum_{i=0}^{j-\ell} d_{w+i} \le h_{w+j} < \theta$. If $j < \ell$ there are no such indices. The remaining head indices lie in $(j-\ell,\, j]$, at most ℓ of them, each with shortfall at most one. Hence the head is strictly below $\theta + \ell$, and adding the tail gives the main-text blocked-trigger inequality. □

C.3 Proof of main-text Proposition 4

Proof. On an uncommitted day, main-text Proposition 1 makes every covered hour adequate, so $s_t^{\mathrm{phys}} \ge s_t$ and $1 - s_t^{\mathrm{phys}} \le 1 - s_t$. On a committed day $s_t = 1$, so $1 - s_t^{\mathrm{phys}} = (1 - s_t) + a_t(1 - s_t^{\mathrm{phys}})$ holds. Summing the per-day inequality over the window gives the claim. □

D Earlier pipeline evaluations

The evaluations in this section were produced by an earlier, observation-index replay adapter on the same public sources. They are reported here in full, under their original supports, information contracts and multiplicity

Table S7 Region-only coverage, worst-window service, and volume across exposed cohorts and retrospective matching views.

		Raw common shell			Coverage matched		Static-volume matched	
Cohort	Method	Cov.	W_7	Mean vol.	W_7	Mean vol.	Cov.	W_7
NYISO	Static	0.873	0.287	1.14	0.325	2.75	0.873	0.287
	Rolling	0.875	0.293	7.94	0.350	15.6	0.766	0.156
	ACI	0.882	0.554	6.43e+03	0.583	9.9e+03	0.261	0.064
	PID-P	0.792	0.158	0.145	0.344	1.46	0.899	0.344
	PID-PI	0.783	0.151	0.141	0.341	1.55	0.896	0.329
	SAOCP	0.909	0.636	880	0.620	666	0.511	0.165
	UP-OCP	0.875	0.549	186	0.592	323	0.471	0.142
	Areces	0.795	0.161	0.149	0.343	1.45	0.899	0.343
GEFCom	Static	0.897	0.429	1	0.371	4.03	0.897	0.429
	Rolling	0.873	0.314	2.68	0.400	8.04	0.884	0.371
	ACI	0.882	0.514	2.24×10^4	0.571	2.65×10^4	0.494	0.057
	PID-P	0.770	0.171	0.0374	0.343	4.21	0.895	0.429
	PID-PI	0.759	0.143	0.0342	0.371	4.21	0.892	0.429
	SAOCP	0.897	0.571	2.91×10^4	0.600	7.36e+03	0.604	0.086
	UP-OCP	0.866	0.543	2.85×10^4	0.543	1.43×10^5	0.519	0.086
	Areces	0.772	0.171	0.0392	0.343	4.19	0.892	0.429
CAISO dev	Static	0.857	0.429	1	0.464	5.17×10^4	0.857	0.429
	Rolling	0.874	0.452	317	0.464	5.52×10^5	0.815	0.369
	ACI	0.872	0.548	1.46×10^{21}	0.607	5.62×10^{21}	0.207	0.095
	PID-P	0.648	0.155	0.0123	0.452	4.54×10^4	0.854	0.429
	PID-PI	0.658	0.167	0.0142	0.464	4.79×10^4	0.855	0.429
	SAOCP	0.909	0.655	5.73×10^{11}	0.643	2.27×10^{11}	0.327	0.036
	UP-OCP	0.872	0.595	1.65×10^9	0.631	5.54×10^9	0.418	0.060
	Areces	0.658	0.167	0.0134	0.452	4.54×10^4	0.854	0.429
CAISO post	Static	0.735	0.286	1	0.571	883	0.735	0.286
	Rolling	0.835	0.429	134	0.571	1.68×10^4	0.675	0.310
	ACI	0.873	0.571	5.06×10^6	0.643	2.90×10^7	0.114	0.000
	PID-P	0.514	0.119	0.0739	0.571	861	0.727	0.310
	PID-PI	0.525	0.119	0.0892	0.571	928	0.724	0.310
	SAOCP	0.892	0.571	2.12×10^7	0.619	1.57×10^8	0.300	0.000
	UP-OCP	0.870	0.667	1.07×10^7	0.667	2.18×10^7	0.281	0.024
	Areces	0.523	0.119	0.0773	0.571	817	0.727	0.310

The two matched views are retrospective diagnostics. Each row uses one cohort and one common OAS-region estimand; volume comparisons are made within that cohort and geometry.

families. The production rule appears below under its historical label DRB. These results complement the current study with region-volume, forecasting-base and allocation diagnostics from that earlier pipeline.

D.1 Coverage, worst-window service, and region size

The first question is whether region adaptation alone improves short-window service at a comparable region size. The unified matrix compares eight methods across four exposed cohorts in raw, retrospective coverage-matched, and retrospective static-volume-matched views (Xu et al., 2024; Gibbs and Candès, 2021; Angelopoulos et al., 2023; Areces et al., 2025; Liu et al., 2026; Bhatnagar et al., 2023). These views characterize how the evaluated methods trade coverage, short-window service and region volume. Small raw PID volumes on NYISO coincide with undercoverage. SAOCP and ACI can provide higher worst-window values in the coverage-matched view, but at large volume costs; constraining them to static mean volume reduces their coverage and service within this evaluated method set. The stratified table below preserves the three views and cohort strata; Supplementary Figure S9 shows their service–volume movements. Retrospective matching is an outcome-conditioned diagnostic; the region-only rows evaluate the forecast base without fallback.

D.2 Fallback gains, action cost, and matched-count allocation

On exposed NYISO, adding the production DRB rule to the same SAOCP base increases origin-averaged worst-seven-day modeled service from 0.636054 to 0.730442. This gain uses 280 actions, an action rate of approximately 11.01%; $L_{\lambda=1}$ rises from about 0.091 to 0.192. The improvement therefore trades modeled short-window protection for action use. The primary-base descriptive mean-service break-even coefficient is $\lambda^* = 0.0843$ (Supplementary Table S23). This coefficient is expressed in the units of the mean-shortfall loss.

To distinguish action placement from action quantity, the exposed NYISO COCD diagnostic reallocates the shadow-DRB actions at the same count in every origin and contiguous segment, totaling 280 for each policy. Its daily ranking uses strict-lag-2 pre-action information, conditional on the completed segment's shadow-action budget and end. This makes COCD a budget/end-conditioned retrospective allocation diagnostic. At the matched count, DRB attains 0.730442 worst-seven-day service versus COCD's 0.702381; all three descriptive paired origin-bootstrap intervals exclude zero in the DRB-minus-COCD direction. Rescue precision is 0.494240 versus 0.461925,

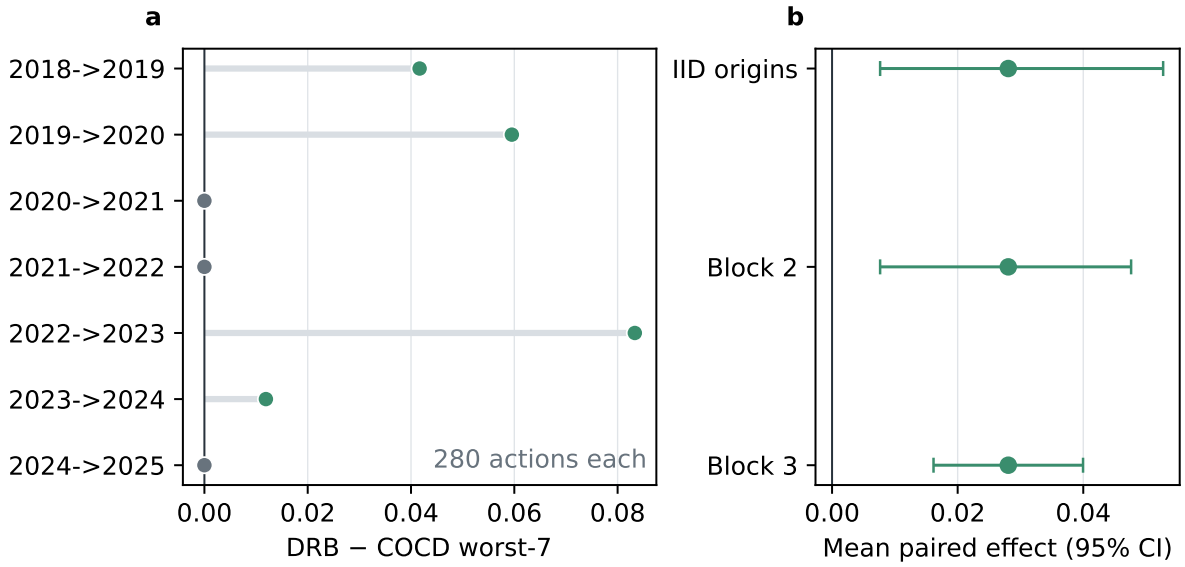


Figure S2 Modeled service and allocation diagnostics on exposed NYISO. DRB and COCD match action counts within each origin and segment, with 280 actions each. COCD conditions on the completed segment's budget and end, defining a retrospective allocation diagnostic. Prepared with AI-assisted code

Table S8 Modeled NYISO service, action use, and loss for exposed allocation diagnostics.

Policy	Mean	W_3	W_7	W_{14}	Actions	Rate (%)	Rescue	L_1
No fallback	0.909	0.383	0.636	0.769	0	0.0	–	0.091
DRB	0.918	0.462	0.730	0.809	280	11.0	0.494	0.192
COCD	0.917	0.435	0.702	0.803	280	11.0	0.462	0.193
Random matched	0.919	0.417	0.662	0.786	280/draw	11.0	0.549	0.191
Exact offline oracle	0.952	0.825	0.912	0.918	280	11.0	1.000	0.158

Modeled service and the abstract action loss L_1; all rows share the DRB base and horizon cap. COCD is budget/end-conditioned and the exact oracle uses the full outcome path, defining two retrospective allocation diagnostics. Random matched averages outcome-independent positions at the realized count; DRB and COCD use exactly 280 actions with per-origin, per-segment count matching.

while $L_{\lambda=1}$ is approximately 0.192 versus 0.193. The paired comparison separates allocation quality from action quantity. Figure S2 and Table S8 show these diagnostics; random matched placement remains a descriptive sensitivity.

An exact offline allocation provides a separate full-outcome benchmark. For each exposed NYISO origin, a zero-gap mixed-integer program observes the entire base-service path and maximizes minimum within-segment seven-day service, subject to exactly that origin's DRB action count. It reaches 0.911565 versus DRB's 0.730442 with the same 280 total actions; its paired origin-averaged difference is +0.181122, positive in 7/7 origins, with descriptive block-2 interval [0.149660, 0.215136]. Its $L_{\lambda=1}$ is approximately 0.158. This full-outcome optimization quantifies allocation headroom for the fixed-count replay problem. Its retrospective information set distinguishes it from the causal policies (Supplementary Section H.11).

D.3 Registered recovery: base gain and stronger comparator

The recovery comparison asks whether the frozen rule improves the registered endpoint beyond the base and a causal allocation comparator. The $A^\star$ specialization uses a pre-registered horizon and causal one-credit bank, without querying future missingness or a future segment end. It shares the SAOCP region trace and cap with D1, the production DRB policy. On the released recovery path both use 27 actions out of 181 days (14.92%); the general action contract is at most B commitments (Supplementary Section H.5).

The registered average of the 12 blockwise worst-seven-day effects for $D1 - \text{BASE}$ is +0.0143849, or 1.44 percentage points, with primary block-2 95% CI [0.0059524, 0.0243056] and Holm-adjusted $p = 0.00704965$. The stronger $D1 - (SAOCP{+}A^\star)$ contrast has effect +0.0009921 (95% CI [0, 0.0019841]) and Holm-adjusted $p = 0.0686966$. The guarded $C^\star$ contrast is non-evaluable and remains in the registered family. The recovery figure and table below report all three outcomes.

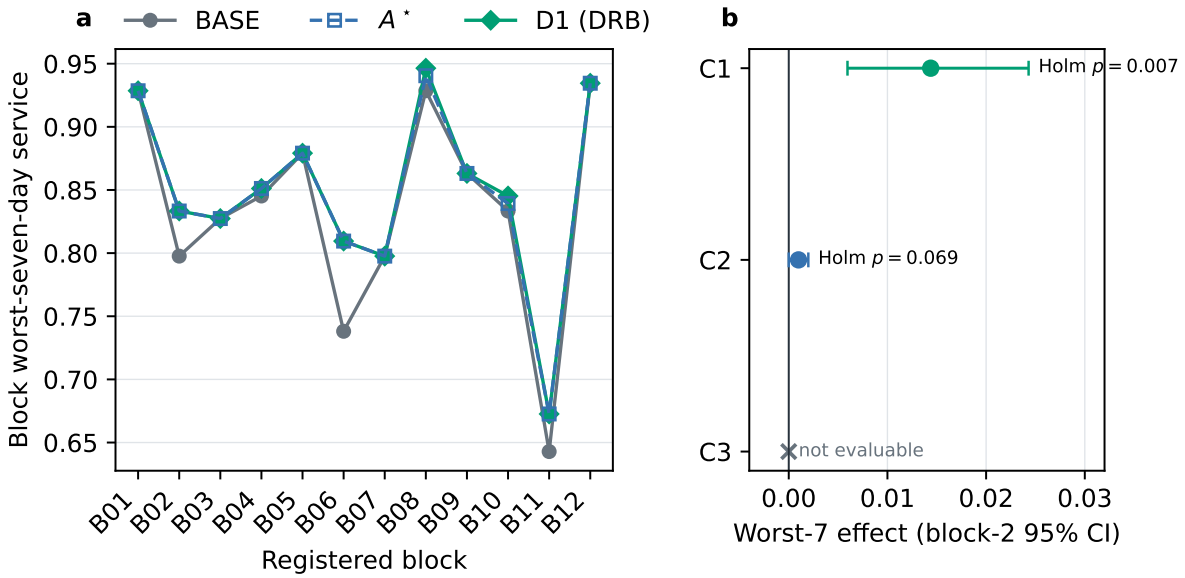


Figure S3 Registered recovery endpoints and contrasts for modeled service. The analysis was retained after technical recovery with prior outcome access. Full-period summaries are non-primary; the registered endpoint averages 12 blockwise worst-seven-day values. The base contrast is positive under the registered Holm test; the $SAOCP{+}A^\star$ contrast has Holm-adjusted $p = 0.0687$. Prepared with AI-assisted code

Table S9 Registered recovery comparisons with prior outcome access: modeled service, action use, and loss.

Panel A: policy summaries

Policy	Block mean W_7	Full mean	Full W_7	Actions	L_1
BASE	0.835	0.894	0.643	0	0.106
$A^\star$	0.848	0.906	0.673	27	0.243
D1 (DRB)	0.849	0.904	0.673	27	0.245

Panel B: registered family

Contrast	Effect	Block-2 95% CI	Raw p	Holm p	Decision
C1: D1 − BASE	0.0144	[0.0060, 0.0243]	0.002	0.007	Reject
C2: D1 $-A^\star$	0.0010	[0.0000, 0.0020]	0.034	0.069	Do not reject
C3: D1 $-C^\star$	–	–	–	1.000	Not evaluable

Panel A full-period values are non-primary sensitivity summaries; Panel B reports the registered three-member family on 12 chronological blocks. Here $A^\star$ denotes $SAOCP{+}A^\star$, sharing the base region trace with D1. The frozen analysis was retained after technical recovery with prior outcome access. The base contrast passes the registered Holm test; the $SAOCP{+}A^\star$ contrast remains statistically unresolved.

The benefit is accompanied by action cost: the non-primary full-period $L_{\lambda=1}$ summaries are approximately 0.106 for BASE, 0.243 for $SAOCP{+}A^\star$, and 0.245 for D1. These full-period values are separate from the registered blockwise endpoint. The recovery retained its frozen analysis after technical failures with prior outcome access. The base contrast remains significant under the existing block-3 sensitivity (Holm $p = 0.00105$), and an assumption-changing, post-release finite-support sign-flip stress test gives Holm $p = 0.09375$: the registered result is block-length stable and sensitive to the null/resampling construction (Supplementary Section H.6 and Table S25).

D.4 Forecasting bases: accuracy, fallback benefit, and cost

We next ask whether the fallback result depends on the forecasting base. The exposed evaluations include development-selected P7-residual adaptations of DeepAR (Salinas et al., 2020), N-HiTS (Challu et al., 2023), PatchTST (Nie et al., 2023), TFT (Lim et al., 2021), and ProLoaF (Gürses-Tran et al., 2023); LightGBM (Ke et al., 2017) and CatBoost (Prokhorenkova et al., 2018) also enter the point-MAE comparison. Selection uses the common 2018 development calendar and normalized weighted quantile loss, freezing the selected configuration

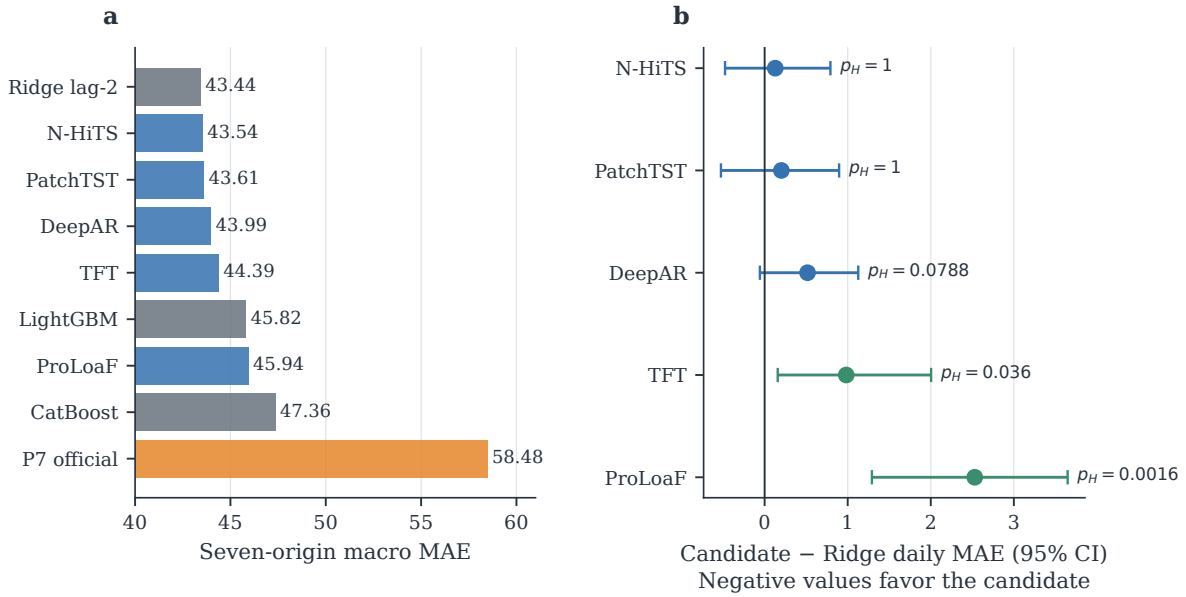


Figure S4 Point-forecast accuracy on exposed NYISO under the frozen evaluation protocol. Deep models are development-selected P7-residual adaptations; point accuracy is distinct from common-region service. Prepared with AI-assisted code

before the seven annual evaluations. Common-OAS evaluation uses six origins, with the first prediction origin supplying reference history. Data splits, feedback eligibility, selection loss, and evaluation rules are aligned; model interfaces and source-paper configurations need not be identical. Supplementary Table S27 records search sizes, training caps, selected configurations, environments, and retained failures.

DeepAR significantly improves point accuracy over LightGBM, CatBoost and P7, with macro MAE 43.9913. Its paired daily-MAE difference from lag-2 Ridge is +0.5166 (block-7 CI [−0.0571, 1.1270], Holm $p = 0.0788$). All four further deep adaptations significantly improve on P7; N-HiTS and PatchTST have statistically unresolved MAE contrasts with Ridge, while TFT and ProLoaF have higher MAE than Ridge under the Holm test. The base-only common-OAS worst-seven-day contrasts with P7 remain statistically unresolved, highlighting the distinct roles of point accuracy and short-window allocation.

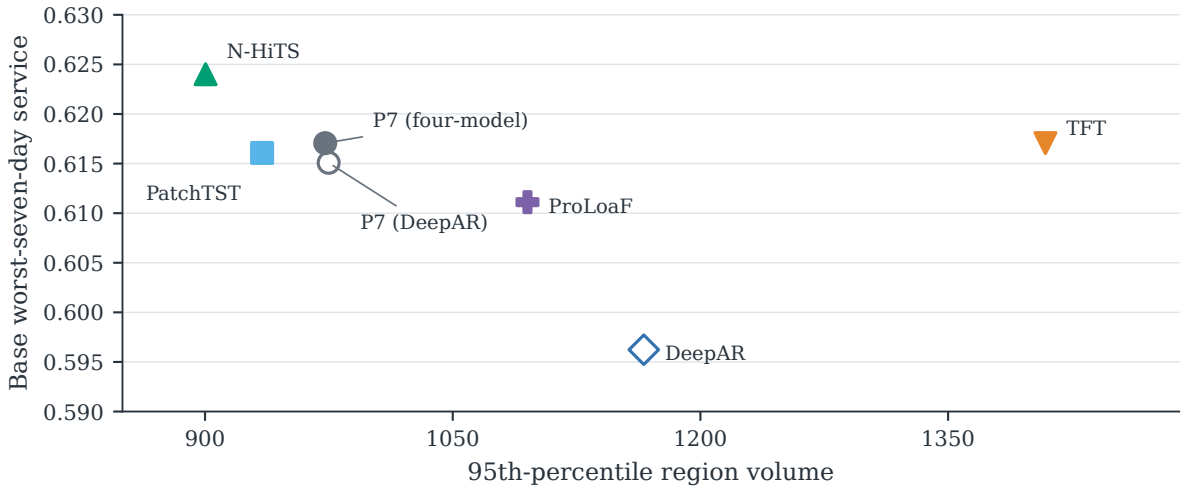
Adding production DRB to each deep base significantly improves mean modeled service for five of five bases and worst-seven-day service for three of five: DeepAR, N-HiTS, and PatchTST. These counts summarize decisions from the separate frozen DeepAR and four-model families; they are not a pooled five-model test. Worst-seven-day effects for TFT and ProLoaF do not pass Holm correction ($p = 0.0594$ and $p = 0.0540$). The fallback gains use origin-averaged action rates of approximately 12.1%–12.6%, while $L_{\lambda=1}$ rises from about 0.095–0.098 without fallback to 0.210–0.213. These loss values quantify the action penalty accompanying the service gains. Figure S4 reports point accuracy; Figures S5–S6 and Table S10 separate region size, service, action, rescue and modeled loss. Paired point and native-geometry metrics remain in Supplementary Tables S28–S29.

The existing block-14 and leave-one-origin-out checks retain positive directions for all ten deep-base fallback service effects. Formal decisions remain those from the frozen block-7 Holm families; these checks characterize directional robustness. The final four-model common-UQ partition (p3) ran on Windows Python 3.11 after 30 primary-server native failures, with identical inputs, code, and SAOCP pin. The platform change is retained in the reproducibility record for these common-UQ results.

The fallback response is also a modeling assumption. In a closed-loop sensitivity across the primary SAOCP base and the five deep bases, an action removes a deterministic fraction η of current shortfall and the resulting service feeds back into subsequent decisions. At $\eta = .75$, origin-averaged worst-seven-day gains span 0.0660–0.0967, with a corresponding $L_{\lambda=1}$ increase of 0.1058–0.1214. Each base first exceeds a 0.01 gain at the tested grid point $\eta = .25$. Here η is a deterministic modeled shortfall reduction, and the grid characterizes response sensitivity (Supplementary Table S22).

D.5 Cross-system response, budget pressure, and variant sensitivity

Finally, the production rule is replayed without retuning on the admitted binary daily GEFCom/CAISO service traces. It exhausts $B = \text{round}(0.15T)$ in all 23 origins and records 498 blocked triggers: 107 on GEFCom, 253 on CAISO development, and 138 on CAISO exposed post-split. Origin-averaged worst-seven-day gain over the base is zero on GEFCom and CAISO post, with service unchanged at 0.5714. CAISO development rises from



Open markers: DeepAR evaluation; filled markers: four-model evaluation (inference remains separate).

Figure S5 Coverage, worst-window service, and region size in the common OAS shell without fallback. DeepAR and the four-model evaluation retain separate supports and inference families; native rectangular metrics are not included in this comparison. Prepared with AI-assisted code

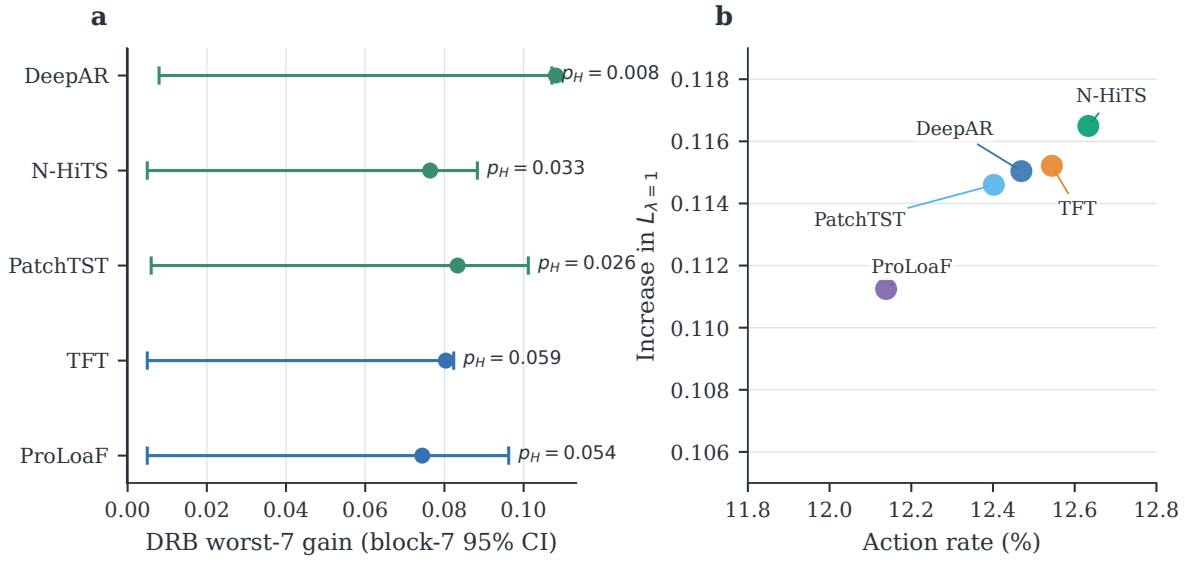


Figure S6 Modeled DRB service gains, action use, and abstract action penalty on each deep base. Gains are relative to that base without fallback; $L_{\lambda=1}$ is a modeled loss, not a measured operating cost. DeepAR and four-model Holm decisions remain separate. Prepared with AI-assisted code

Table S10 Common-OAS region results and modeled production-DRB benefits, actions, and loss across forecasting bases.

Panel A: common OAS shell without fallback

Evaluation	Base	Joint cov.	Base W_7	Median vol.	P95 vol.
DeepAR study	DeepAR	0.905	0.596	2.14	1165.7
Four-model study	N-HiTS	0.903	0.624	1.70	900.3
Four-model study	PatchTST	0.904	0.616	1.96	934.5
Four-model study	TFT	0.903	0.617	2.42	1408.9
Four-model study	ProLoaF	0.902	0.611	2.08	1095.3

Panel B: production DRB service effect

Evaluation	Base	DRB W_7	Effect	Holm p
DeepAR study	DeepAR	0.704	+0.108	0.008
Four-model study	N-HiTS	0.700	+0.076	0.033
Four-model study	PatchTST	0.699	+0.083	0.026
Four-model study	TFT	0.697	+0.080	0.059
Four-model study	ProLoaF	0.686	+0.074	0.054

Panel C: production DRB action and cost

Evaluation	Base	Action (%)	Rescue	L_1 base→DRB
DeepAR study	DeepAR	12.5	0.437	0.095→0.210
Four-model study	N-HiTS	12.6	0.429	0.097→0.213
Four-model study	PatchTST	12.4	0.423	0.096→0.210
Four-model study	TFT	12.5	0.413	0.097→0.212
Four-model study	ProLoaF	12.1	0.482	0.098→0.210

DeepAR and the four-model evaluation retain separate supports and frozen Holm families. Panel B compares production DRB with its corresponding forecast base before fallback. Service and L_1 use the modeled fallback and abstract action penalty. The final four-model common partition ran on Windows Python 3.11 after 30 server-native failures, with identical inputs, code, and SAOCP pin.

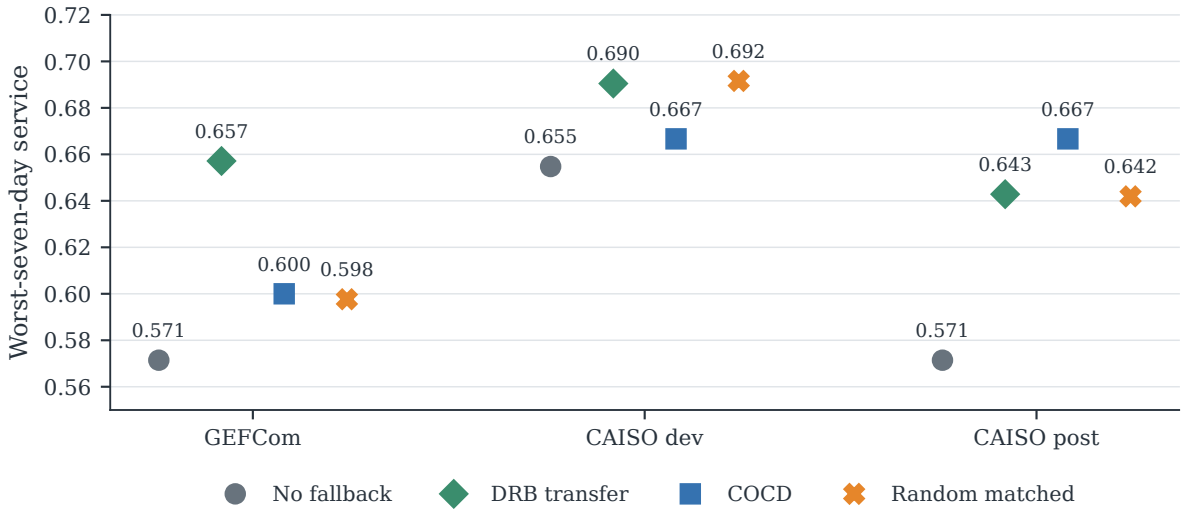


Figure S7 Exposed cross-dataset allocation sensitivity of the thresholded-count transfer variant, distinct from production DRB. COCD and random placement are conditioned allocation diagnostics; service and action cost are modeled. Production-rule transfer results are reported separately in Supplementary Table S21. Prepared with AI-assisted code

0.6548 to 0.6667, an effect of +0.011905 with descriptive block-2 interval [0, 0.0357]; this does not establish a significant improvement. Origin-averaged action rates are approximately 15.3%, 15.2%, and 15.1%, reflecting integer rounding of the horizon caps. $L_{\lambda=1}$ rises from approximately 0.103 to 0.247, 0.091 to 0.240, and 0.108 to 0.254, respectively. This retrospective replay identifies the role of budget pressure in cross-system performance. Its blocked paths are described by the blocked-trigger bound, while the no-block floor applies to eligible unblocked windows (Supplementary Table S21).

The separate thresholded-count transfer variant follows a distinct trigger and action path from the continuous shortfall-sum rule. In its allocation sensitivities, GEFCom favors the variant over budget-conditioned COCD; CAISO development provides no clear advantage over matched random placement, while CAISO exposed post-split favors COCD. Action rates are approximately 12.3%, 9.7%, and 11.2%, and $L_{\lambda=1}$ for the variant increases over the no-fallback base in all three cohorts. Figure S7 and Table S11 report this variant separately from the production replay. Its COCD and random rows provide conditioned allocation sensitivities under the variant's own action path.

The outcome-exposed, post-hoc mechanism trace also concerns only that thresholded-count variant. Trigger-lift point estimates are below one in all three cohorts, but only the CAISO-development exploratory interval excludes one. Direct rescues occur in restrictive worst-seven-day windows in 5/5, 4/5, and 3/4 evaluated cases for GEFCom, CAISO development, and CAISO post. These denominators refer to the diagnostic key-window cases, not all origins or an intervention-success probability. With only 5/12/6 origin clusters, the intervals are exploratory. The observations are compatible with localized window rescue without broad same-day trigger enrichment; they do not identify a causal mechanism or establish the mechanism of the production shortfall-sum rule.

Table S11 Exposed allocation sensitivity of the thresholded-count transfer variant, distinct from production DRB.

Cohort	Policy	Mean	W_3	W_7	W_{14}	Act. (%)	Rescue	L_1
GEFCom	No fallback	0.897	0.200	0.571	0.729	0.0	–	0.103
	DRB K2 variant	0.908	0.267	0.657	0.786	12.3	0.094	0.215
	COCD	0.904	0.200	0.600	0.757	12.3	0.043	0.219
	Random matched	0.910	0.238	0.598	0.752	12.3	0.104	0.212
CAISO dev	No fallback	0.909	0.306	0.655	0.756	0.0	–	0.091
	DRB K2 variant	0.914	0.306	0.690	0.780	9.7	0.040	0.183
	COCD	0.913	0.306	0.667	0.774	9.7	0.031	0.184
	Random matched	0.921	0.355	0.692	0.787	9.7	0.118	0.176
CAISO post	No fallback	0.892	0.333	0.571	0.714	0.0	–	0.108
	DRB K2 variant	0.899	0.333	0.643	0.750	11.2	0.064	0.213
	COCD	0.901	0.333	0.667	0.762	11.2	0.078	0.211
	Random matched	0.908	0.393	0.642	0.773	11.2	0.151	0.204

DRB K2 variant uses thresholded-count feedback and is distinct from production shortfall-sum DRB. COCD is a budget/end-conditioned allocation sensitivity; random matched averages outcome-independent position draws conditional on the realized count. Service and L_1 use modeled fallback and an abstract action penalty. Production replay and its blocked triggers are reported in Table S21.

E Fair deep-baseline stratified metrics (per origin)

The per-origin rows in this section are read from the frozen DeepAR and four-model common-UQ policy objects. The two evaluations retain their own supports and statistical families; their P7 rows are not interchangeable. Here w3/w7/w14 denote minimum same-segment 3/7/14-day mean modeled service; vol p90/p95 are forecast-region volume percentiles; rescued is $\sum_{t:a_t=1}(1-q_t)$; and L1 is the modeled loss $L_{\lambda=1} = \overline{1-s} + \overline{a}$. Action rate and cap utilization are fractions, not percentages. Cov denotes base joint-region coverage, which is unchanged by the action layer. Base denotes no fallback; DRB denotes the production continuous shortfall-sum controller with strict-lag-2 feedback and modeled sure-cover actions. Service and rescue are evaluated on the modeled joint-coverage scale.

Table S12 Per-origin common-OAS results for the P7-residual DeepAR evaluation: no fallback versus production DRB.

model	origin	policy	w3	w7	w14	vol p90	vol p95	rescued	L1	rate	cap util	cov
DeepAR	2019→2020	base	0.2778	0.5298	0.7500	194.9323	1217.0877	0.0000	0.0930	0.0000	0.0000	0.9070
DeepAR	2019→2020	DRB	0.4444	0.6607	0.7976	194.9323	1217.0877	2.5000	0.1709	0.0847	0.5636	0.9070
DeepAR	2020→2021	base	0.3333	0.6726	0.7708	1067.4906	4376.0317	0.0000	0.1073	0.0000	0.0000	0.8927
DeepAR	2020→2021	DRB	0.3333	0.7143	0.7917	1067.4906	4376.0317	2.8317	0.2477	0.1488	1.0000	0.8927
DeepAR	2021→2022	base	0.1806	0.5714	0.7351	55.7020	214.5603	0.0000	0.0991	0.0000	0.0000	0.9009
DeepAR	2021→2022	DRB	0.3889	0.7083	0.8006	55.7020	214.5603	4.2083	0.2136	0.1260	0.8364	0.9009
DeepAR	2022→2023	base	0.4167	0.7024	0.7798	97.9673	299.0789	0.0000	0.0829	0.0000	0.0000	0.9171
DeepAR	2022→2023	DRB	0.4583	0.7024	0.8036	97.9673	299.0789	2.8333	0.2066	0.1315	0.8727	0.9171
DeepAR	2023→2024	base	0.2917	0.6131	0.7738	126.0822	502.8105	0.0000	0.0999	0.0000	0.0000	0.9002
DeepAR	2023→2024	DRB	0.2917	0.6845	0.7887	126.0822	502.8105	4.8750	0.2368	0.1503	1.0000	0.9002
DeepAR	2024→2025	base	0.2639	0.4881	0.6845	122.4767	384.7505	0.0000	0.0896	0.0000	0.0000	0.9104
DeepAR	2024→2025	DRB	0.5833	0.7560	0.8155	122.4767	384.7505	3.6703	0.1864	0.1068	0.7091	0.9104
P7	2019→2020	base	0.3333	0.6131	0.7708	149.0599	886.2089	0.0000	0.0912	0.0000	0.0000	0.9088
P7	2020→2021	base	0.3333	0.6786	0.7738	914.8587	3303.7834	0.0000	0.1091	0.0000	0.0000	0.8909
P7	2021→2022	base	0.1528	0.5476	0.7351	64.7593	283.3866	0.0000	0.1094	0.0000	0.0000	0.8905
P7	2022→2023	base	0.4583	0.7143	0.7708	118.8724	377.7918	0.0000	0.0847	0.0000	0.0000	0.9153
P7	2023→2024	base	0.2639	0.6310	0.7619	125.7482	546.8364	0.0000	0.1036	0.0000	0.0000	0.8964
P7	2024→2025	base	0.2917	0.5060	0.6756	104.9753	451.1806	0.0000	0.0976	0.0000	0.0000	0.9024

Modeled service uses $s_t = a_t + (1 - a_t)q_t$ under sure cover; L_1 is mean shortfall plus action rate. Action rate and cap utilization are fractions. Cov and volume describe the forecast region, which DRB leaves unchanged. The P7 reference belongs to this evaluation only.

Table S13 Per-origin common-OAS results for four P7-residual deep adaptations: no fallback versus production DRB.

model	origin	policy	w3	w7	w14	vol p90	vol p95	rescued	L1	rate	cap util	cov
N-HiTS	2019→2020	base	0.3333	0.6429	0.7560	120.3519	760.1560	0.0000	0.0932	0.0000	0.0000	0.9068
N-HiTS	2019→2020	DRB	0.3333	0.6667	0.7887	120.3519	760.1560	2.0833	0.1668	0.0792	0.5273	0.9068
N-HiTS	2020→2021	base	0.3333	0.7083	0.7500	759.6385	3202.1204	0.0000	0.1030	0.0000	0.0000	0.8970
N-HiTS	2020→2021	DRB	0.3333	0.7143	0.7530	759.6385	3202.1204	3.5000	0.2217	0.1285	0.8519	0.8970
N-HiTS	2021→2022	base	0.2222	0.5952	0.7589	53.8607	217.9661	0.0000	0.1046	0.0000	0.0000	0.8954
N-HiTS	2021→2022	DRB	0.3889	0.7024	0.8095	53.8607	217.9661	3.4583	0.2430	0.1479	0.9818	0.8954
N-HiTS	2022→2023	base	0.4167	0.6667	0.7827	100.7975	339.0891	0.0000	0.0837	0.0000	0.0000	0.9163
N-HiTS	2022→2023	DRB	0.4306	0.6964	0.7827	100.7975	339.0891	2.8333	0.2184	0.1425	0.9455	0.9163
N-HiTS	2023→2024	base	0.3611	0.5774	0.7470	114.0341	481.0481	0.0000	0.1050	0.0000	0.0000	0.8950
N-HiTS	2023→2024	DRB	0.3611	0.6964	0.8065	114.0341	481.0481	6.4167	0.2377	0.1503	1.0000	0.8950
N-HiTS	2024→2025	base	0.3333	0.5536	0.7113	101.0284	401.6350	0.0000	0.0904	0.0000	0.0000	0.9096
N-HiTS	2024→2025	DRB	0.5694	0.7262	0.8363	101.0284	401.6350	3.2083	0.1912	0.1096	0.7273	0.9096
PatchTST	2019→2020	base	0.3194	0.6429	0.7887	142.7990	992.5281	0.0000	0.0929	0.0000	0.0000	0.9071
PatchTST	2019→2020	DRB	0.3194	0.6488	0.7917	142.7990	992.5281	1.4583	0.1927	0.1038	0.6909	0.9071
PatchTST	2020→2021	base	0.3333	0.7083	0.7708	902.4998	3335.4094	0.0000	0.1031	0.0000	0.0000	0.8969
PatchTST	2020→2021	DRB	0.3333	0.7143	0.7738	902.4998	3335.4094	3.5000	0.2163	0.1229	0.8148	0.8969
PatchTST	2021→2022	base	0.1667	0.5714	0.7470	49.4473	207.4441	0.0000	0.1001	0.0000	0.0000	0.8999
PatchTST	2021→2022	DRB	0.3611	0.6845	0.8006	49.4473	207.4441	3.7083	0.2406	0.1507	1.0000	0.8999
PatchTST	2022→2023	base	0.3750	0.6786	0.7976	83.8511	269.8847	0.0000	0.0871	0.0000	0.0000	0.9129
PatchTST	2022→2023	DRB	0.4444	0.7143	0.8006	83.8511	269.8847	2.7500	0.2056	0.1260	0.8364	0.9129
PatchTST	2023→2024	base	0.3611	0.5774	0.7500	118.5942	436.0552	0.0000	0.1005	0.0000	0.0000	0.8995
PatchTST	2023→2024	DRB	0.3611	0.7024	0.8185	118.5942	436.0552	5.4583	0.2359	0.1503	1.0000	0.8995
PatchTST	2024→2025	base	0.3194	0.5179	0.6756	100.0537	365.9772	0.0000	0.0896	0.0000	0.0000	0.9104
PatchTST	2024→2025	DRB	0.5694	0.7321	0.8482	100.0537	365.9772	3.7083	0.1699	0.0904	0.6000	0.9104
ProLoaF	2019→2020	base	0.3194	0.6369	0.7917	180.7521	1114.5477	0.0000	0.0887	0.0000	0.0000	0.9113
ProLoaF	2019→2020	DRB	0.3194	0.6548	0.7917	180.7521	1114.5477	1.9167	0.1572	0.0738	0.4909	0.9113
ProLoaF	2020→2021	base	0.3333	0.6786	0.7143	1002.2907	3871.1113	0.0000	0.1098	0.0000	0.0000	0.8902
ProLoaF	2020→2021	DRB	0.3333	0.7143	0.7173	1002.2907	3871.1113	3.0000	0.2522	0.1508	1.0000	0.8902
ProLoaF	2021→2022	base	0.1250	0.4881	0.7083	70.6190	299.3704	0.0000	0.1053	0.0000	0.0000	0.8946
ProLoaF	2021→2022	DRB	0.3750	0.6667	0.7768	70.6190	299.3704	4.4167	0.2357	0.1425	0.9455	0.8946

Modeled service uses $s_t = a_t + (1 - a_t)q_t$ under sure cover; L_1 is mean shortfall plus action rate. Action rate and cap utilization are fractions. This P7 reference and inference family remain separate from the DeepAR evaluation. The final common partition used Windows Python 3.11 after 30 retained remote native failures, with the same sealed inputs, code, and SAOCP pin.

Table S13 Per-origin common-OAS results for four P7-residual deep adaptations: no fallback versus production DRB (continued)

model	origin	policy	w3	w7	w14	vol p90	vol p95	rescued	L1	rate	cap util	cov
ProLoaF	2022→2023	base	0.4028	0.6726	0.7619	111.9401	352.0423	0.0000	0.0865	0.0000	0.0000	0.9135
ProLoaF	2022→2023	DRB	0.4028	0.7202	0.7827	111.9401	352.0423	3.8333	0.1966	0.1205	0.8000	0.9135
ProLoaF	2023→2024	base	0.2639	0.6429	0.7619	118.4523	457.0805	0.0000	0.1019	0.0000	0.0000	0.8981
ProLoaF	2023→2024	DRB	0.2639	0.6488	0.8185	118.4523	457.0805	5.6667	0.2367	0.1503	1.0000	0.8981
ProLoaF	2024→2025	base	0.3056	0.5476	0.7054	102.1056	477.4091	0.0000	0.0977	0.0000	0.0000	0.9023
ProLoaF	2024→2025	DRB	0.4444	0.7083	0.7946	102.1056	477.4091	3.3333	0.1790	0.0904	0.6000	0.9023
TFT	2019→2020	base	0.3194	0.6369	0.7589	244.1890	1361.9859	0.0000	0.0925	0.0000	0.0000	0.9074
TFT	2019→2020	DRB	0.3194	0.6488	0.7946	244.1890	1361.9859	2.1250	0.1578	0.0710	0.4727	0.9074
TFT	2020→2021	base	0.3333	0.6905	0.7470	1242.0624	5074.6574	0.0000	0.1063	0.0000	0.0000	0.8937
TFT	2020→2021	DRB	0.3333	0.7143	0.7589	1242.0624	5074.6574	2.5833	0.2331	0.1341	0.8889	0.8937
TFT	2021→2022	base	0.1806	0.5179	0.7232	81.7399	380.9757	0.0000	0.1067	0.0000	0.0000	0.8933
TFT	2021→2022	DRB	0.3889	0.6548	0.7887	81.7399	380.9757	4.9167	0.2440	0.1507	1.0000	0.8933
TFT	2022→2023	base	0.4444	0.6488	0.7530	140.6594	475.5764	0.0000	0.0816	0.0000	0.0000	0.9184
TFT	2022→2023	DRB	0.4444	0.7262	0.7619	140.6594	475.5764	3.5000	0.2063	0.1342	0.8909	0.9184
TFT	2023→2024	base	0.3472	0.6250	0.7976	115.5516	427.0775	0.0000	0.1001	0.0000	0.0000	0.8999
TFT	2023→2024	DRB	0.3472	0.6905	0.8095	115.5516	427.0775	5.7917	0.2345	0.1503	1.0000	0.8999
TFT	2024→2025	base	0.4306	0.5833	0.7054	138.6575	733.3351	0.0000	0.0930	0.0000	0.0000	0.9070
TFT	2024→2025	DRB	0.4583	0.7500	0.8333	138.6575	733.3351	3.4583	0.1959	0.1123	0.7455	0.9070
P7	2019→2020	base	0.3333	0.6190	0.7708	156.8962	899.7453	0.0000	0.0892	0.0000	0.0000	0.9107
P7	2020→2021	base	0.3333	0.6964	0.7143	860.5258	3252.1887	0.0000	0.1103	0.0000	0.0000	0.8897
P7	2021→2022	base	0.1389	0.5298	0.7232	69.5129	300.1641	0.0000	0.1066	0.0000	0.0000	0.8934
P7	2022→2023	base	0.4722	0.7202	0.7708	120.2292	386.5019	0.0000	0.0831	0.0000	0.0000	0.9169
P7	2023→2024	base	0.2639	0.6310	0.7619	125.7482	546.8364	0.0000	0.1036	0.0000	0.0000	0.8964
P7	2024→2025	base	0.2917	0.5060	0.6756	104.9753	451.1806	0.0000	0.0976	0.0000	0.0000	0.9024

Modeled service uses $s_t = a_t + (1 - a_t)q_t$ under sure cover; L_1 is mean shortfall plus action rate. Action rate and cap utilization are fractions. This P7 reference and inference family remain separate from the DeepAR evaluation. The final common partition used Windows Python 3.11 after 30 retained remote native failures, with the same sealed inputs, code, and SAOCP pin.

Table S14 Paired inference for common-OAS modeled-service contrasts.

contrast / endpoint	effect	b7/b2 lo	b7/b2 hi	raw p	Holm p	inference
DeepAR:base_minus_P7:base [meanSvc]	0.0040	-0.0008	0.0091	0.0996	0.1992	B7+H
DeepAR:base_minus_P7:base [worst7]	-0.0188	-0.0298	0.0367	0.8413	0.8413	B7+H
DeepAR:DRB_minus_DeepAR:base [meanSvc]	0.0097	0.0061	0.0138	0.0002	0.0008	B7+H
DeepAR:DRB_minus_DeepAR:base [worst7]	0.1081	0.0079	0.1071	0.0026	0.0078	B7+H
TFT:base_minus_P7:base [meanSvc]	0.0017	-0.0027	0.0061	0.4454	1.0000	B7+H
TFT:base_minus_P7:base [worst7]	0.0000	-0.0288	0.0446	0.6617	1.0000	B7+H
NHiTS:base_minus_P7:base [meanSvc]	0.0018	-0.0032	0.0068	0.4890	1.0000	B7+H
NHiTS:base_minus_P7:base [worst7]	0.0069	-0.0308	0.0446	0.6389	1.0000	B7+H
PatchTST:base_minus_P7:base [meanSvc]	0.0029	-0.0028	0.0085	0.3230	1.0000	B7+H
PatchTST:base_minus_P7:base [worst7]	-0.0010	-0.0299	0.0446	0.6893	1.0000	B7+H
ProLoaF:base_minus_P7:base [meanSvc]	0.0001	-0.0026	0.0027	0.9417	1.0000	B7+H
ProLoaF:base_minus_P7:base [worst7]	-0.0060	-0.0248	0.0238	0.9471	1.0000	B7+H
TFT:DRB_minus_TFT:base [meanSvc]	0.0102	0.0062	0.0149	0.0002	0.0032	B7+H
TFT:DRB_minus_TFT:base [worst7]	0.0804	0.0050	0.0823	0.0066	0.0594	B7+H
NHiTS:DRB_minus_NHiTS:base [meanSvc]	0.0098	0.0059	0.0146	0.0002	0.0032	B7+H
NHiTS:DRB_minus_NHiTS:base [worst7]	0.0764	0.0050	0.0883	0.0030	0.0330	B7+H
PatchTST:DRB_minus_PatchTST:base [meanSvc]	0.0094	0.0057	0.0139	0.0002	0.0032	B7+H
PatchTST:DRB_minus_PatchTST:base [worst7]	0.0833	0.0060	0.1012	0.0022	0.0264	B7+H
ProLoaF:DRB_minus_ProLoaF:base [meanSvc]	0.0101	0.0063	0.0146	0.0002	0.0032	B7+H
ProLoaF:DRB_minus_ProLoaF:base [worst7]	0.0744	0.0050	0.0962	0.0054	0.0540	B7+H

Here `B7+H` is the single key for every row: paired circular day-block-7 resampling within segments, origins resampled as clusters, and the family-specific Holm correction, with separate families for DeepAR ($m = 4$) and the four-model evaluation ($m = 16$). MeanSvc and worst7 concern modeled service and positive effects favor the candidate.

Table S15 Paired daily-MAE inference for P7-residual deep adaptations.

contrast / endpoint	effect	b7/b2 lo	b7/b2 hi	raw p	Holm p	inference
DeepAR_minus_Ridge [MAE]	0.5166	-0.0571	1.1270	0.0788	0.0788	B7+H
DeepAR_minus_LightGBM_energy [MAE]	-1.8628	-3.2441	-0.6131	0.0014	0.0028	B7+H
DeepAR_minus_CatBoost_energy_quantile [MAE]	-3.3985	-5.1553	-1.8391	0.0002	0.0008	B7+H
DeepAR_minus_P7 [MAE]	-14.6159	-16.9763	-12.2729	0.0002	0.0008	B7+H
NHiTS_minus_Ridge [MAE]	0.1274	-0.4764	0.7921	0.7115	1.0000	B7+H
NHiTS_minus_P7 [MAE]	-14.9394	-17.1574	-12.6393	0.0002	0.0016	B7+H
PatchTST_minus_Ridge [MAE]	0.2015	-0.5258	0.8966	0.5773	1.0000	B7+H
PatchTST_minus_P7 [MAE]	-14.8704	-17.5464	-12.3944	0.0002	0.0016	B7+H
ProLoaF_minus_Ridge [MAE]	2.5293	1.2913	3.6476	0.0002	0.0016	B7+H
ProLoaF_minus_P7 [MAE]	-12.5393	-15.5897	-9.5078	0.0002	0.0016	B7+H
TFT_minus_Ridge [MAE]	0.9822	0.1577	2.0042	0.0120	0.0360	B7+H
TFT_minus_P7 [MAE]	-14.0905	-16.0609	-12.1149	0.0002	0.0016	B7+H

Here `B7+H` is the single key for every row: paired circular day-block-7 resampling within segments, origins resampled as clusters, and the family-specific Holm correction, with separate families for DeepAR ($m = 4$) and the four-model evaluation ($m = 8$). Effects are candidate-minus-comparator, so negative values favor the candidate.

F Per-contrast statistical contract (paired inference)

Effects are candidate-minus-comparator on the named endpoint. Positive mean-service and worst-7 effects favor modeled service; negative MAE effects favor point accuracy. Positive action-rate or L_λ effects indicate greater action use or modeled loss. The b7/b2 columns display the primary circular day-block-7 intervals for the deep-model contrasts and the primary ordered origin-block-2 intervals for the exposed NYISO contrasts, respectively. All displayed intervals are 95% bootstrap percentile intervals. The deep-model analysis resamples origins as clusters and paired days within segments; raw and Holm-adjusted values belong to separate frozen DeepAR and four-model families. Non-rejection establishes neither equivalence nor non-inferiority. The NYISO base-shield and COCD analyses are descriptive paired-origin bootstraps without a pre-declared Holm family; their raw and Holm entries are not evaluable (n.e.). Block-length sensitivity and other robustness checks are described in Section H.6.

Table S16 Production DRB minus the SAOCP base on exposed NYISO: descriptive paired-origin inference.

contrast / endpoint	effect	b7/b2 lo	b7/b2 hi	raw p	Holm p	inference
DRB_minus_base [worst7]	0.0944	0.0459	0.1565	n.e.	n.e.	descriptive
DRB_minus_base [mean_volume]	0.0000	0.0000	0.0000	n.e.	n.e.	descriptive
DRB_minus_base [action_rate]	0.1101	0.0865	0.1368	n.e.	n.e.	descriptive
DRB_minus_base [L_lambda_1]	0.1008	0.0791	0.1244	n.e.	n.e.	descriptive

The displayed interval is the ordered origin-block-2 95% percentile interval; block-3 robustness is described in Section H.6. Worst7 measures modeled service. $L_{\lambda=1}$ is mean shortfall plus action rate, not a monetary operating cost. No raw or Holm-adjusted test is defined for this descriptive family.

Table S17 COCD minus production DRB on exposed NYISO: budget-conditioned, action-matched allocation sensitivity.

contrast / endpoint	effect	b7/b2 lo	b7/b2 hi	raw p	Holm p	inference
worst7 [worst7]	-0.0281	-0.0476	-0.0077	n.e.	n.e.	descriptive
rescued_shortfall_per_action [rescued_shortfall]	-0.0124	-0.0206	-0.0045	n.e.	n.e.	descriptive
service_coverage [service_coverage]	-0.0013	-0.0020	-0.0005	n.e.	n.e.	descriptive
L_lambda_1 [L_lambda_1]	0.0013	0.0005	0.0020	n.e.	n.e.	descriptive

Both policies use 280 actions, matched within every origin and segment. COCD conditions on the completed segment's shadow-action total and end; it is a retrospective allocation diagnostic. The displayed intervals are descriptive ordered origin-block-2 95% percentiles, without a raw-p or Holm family. Service and rescued shortfall are modeled; L_1 is an abstract shortfall-plus-action loss.

G Recovery inference and supporting sensitivities

G.1 Registered recovery comparisons

The registered endpoint is the average of twelve pre-defined blockwise worst-seven-day modeled-service values. Circular block-2 inference is primary; block-3 inference is a sensitivity analysis. Holm correction uses the three-member registered family. The frozen $D1-$BASE contrast has effect 0.0144 and Holm-adjusted $p = 0.0070$. Recovery followed prior payload access and pre-endpoint technical repair, with the frozen endpoint and analysis retained. The $D1 - (SAOCP{+}A^{\star})$ contrast is not significant, and $C^{\star}$ is non-evaluable.

Table S18 Registered recovery comparisons of the average twelve-block worst-seven-day modeled-service endpoint.

contrast	effect	Holm p	rejected	family
C1_D1_minus_BASE	0.0144	0.0070	yes	3
C2_D1_minus_SAOCP_A_STAR	0.0010	0.0687	no	3
C3_D1_minus_C_STAR	n.e.	1.0000	no	3

Holm values use the three-member registered family with circular block-2 inference. Prior outcome access and technical recovery are disclosed in Sections G.3 and H.13. C3 is non-evaluable; its $p = 1$ entry retains the registered family placeholder.

G.2 Modeled loss and exposed-replay sensitivities

Tables S19–S23 provide loss, runtime, transfer, response, and break-even diagnostics. They do not extend the registered recovery testing family. The recovery ratio in Table S23 is post-release descriptive.

Table S19 Abstract action-penalty sensitivity of the production DRB modeled loss.

Base	Policy	$L_{0.25}$	$L_{0.5}$	L_1	L_2	L_4
DeepAR	base	0.0953	0.0953	0.0953	0.0953	0.0953
DeepAR	DRB	0.1168	0.1480	0.2103	0.3350	0.5844
N-HiTS	base	0.0966	0.0966	0.0966	0.0966	0.0966
N-HiTS	DRB	0.1184	0.1500	0.2131	0.3395	0.5921
PatchTST	base	0.0956	0.0956	0.0956	0.0956	0.0956
PatchTST	DRB	0.1171	0.1482	0.2102	0.3342	0.5822
TFT	base	0.0967	0.0967	0.0967	0.0967	0.0967
TFT	DRB	0.1178	0.1492	0.2119	0.3374	0.5883
ProLoaF	base	0.0983	0.0983	0.0983	0.0983	0.0983
ProLoaF	DRB	0.1185	0.1489	0.2096	0.3310	0.5737

Values are six-origin macro averages from the admitted DeepAR and four-model common-UQ objects. Here L_λ = mean shortfall + λ action rate; hence the no-fallback rows do not vary with λ.
The action penalty λ is expressed in service-shortfall units per action. The two evaluations are summarized under their respective supports and statistical families.

Table S20 Stage-specific runtime ledger.

Evaluation	Timed stage	Units	Total time (s)
Verified SAOCP	full method replay	7 origins	378.121
Unified stratification	table post-processing	1 run	0.528
COCD comparison	comparator post-processing	1 run	0.389
DeepAR	successful train+predict	7 origins	220.160
N-HiTS	successful train+predict	7 origins	462.715
PatchTST	successful train+predict	7 origins	467.468
TFT	successful train+predict	7 origins	1883.986
ProLoaF	successful train+predict	7 origins	246.185

The stages and execution environments differ, so these values are a reproducibility ledger rather than a speed ranking. The four-model totals include successful attempts only; retained failed server attempts are disclosed separately in the reproduction logs.

Table S21 Exposed no-retune production-DRB transfer and comparison with the distinct K2 variant.

Cohort	base W7	prod. W7	K2 W7	prod. act.	K2 act.	base L_1	prod. L_1	blocked
GEFCom replay	0.5714	0.5714	0.6571	0.153	0.123	0.103	0.247	107
CAISO development	0.6548	0.6667	0.6905	0.152	0.097	0.091	0.240	253
CAISO exposed post	0.5714	0.5714	0.6429	0.151	0.112	0.108	0.254	138

Origin macro averages on admitted exposed-replay SAOCP daily traces. "prod." is the continuous shortfall-sum controller and "K2" the distinct thresholded-count variant (eligible days with $s_u < 0.9$ in the recent six-record window, trigger at two); both use strict-lag-2 feedback, and their action counts are not matched. All 23 production origins exhaust the 15% cap. L_1 is mean shortfall plus action rate, and the blocked column counts production blocked-trigger days. Inference is descriptive and cohort-specific; the CAISO post row is an exposed sensitivity.

Table S22 Closed-loop sensitivity to a modeled partial shortfall response.

Base	W7 Δ, $\eta = .50$	W7 Δ, $\eta = .75$	W7 Δ, $\eta = 1$	first η	ΔL_1, $\eta = .75$
Verified SAOCP	0.0591	0.0846	0.0944	0.25	0.1058
DeepAR	0.0719	0.0967	0.1081	0.25	0.1188
N-HiTS	0.0561	0.0709	0.0764	0.25	0.1214
PatchTST	0.0585	0.0751	0.0833	0.25	0.1205
ProLoaF	0.0546	0.0660	0.0744	0.25	0.1164
TFT	0.0620	0.0754	0.0804	0.25	0.1196

Closed-loop macro averages over admitted exposed production traces. Partial service re-enters the strict-lag-2 state. "First η" is the smallest tested grid value with worst-seven-day gain at least 0.01. The response is $s_t = q_t + \eta a_t(1 - q_t)$; η is the deterministic fraction of base shortfall removed on action days. Gains are relative to the corresponding no-fallback base. $\Delta L_1 > 0$ denotes increased modeled shortfall-plus-action loss; these rows characterize modeled response sensitivity.

Table S23 Descriptive break-even coefficients for the modeled mean-shortfall-plus-action loss.

Stratum	gain	act.	λ^*	b2 lo	b2 hi	finite	$\Delta L_{.25}$	ΔL_1
Verified SAOCP	.0093	.1101	.0843	.0714	.0964	1.000	+.0183	+.1008
DeepAR	.0097	.1247	.0774	.0687	.0893	1.000	+.0215	+.1150
N-HiTS	.0098	.1263	.0779	.0633	.0947	1.000	+.0217	+.1165
PatchTST	.0094	.1240	.0760	.0630	.0952	1.000	+.0216	+.1146
ProLoaF	.0101	.1214	.0835	.0673	.0985	1.000	+.0202	+.1112
TFT	.0102	.1254	.0816	.0696	.0930	1.000	+.0211	+.1152
GEFCom replay	.0087	.1532	.0571	.0284	.0860	.978	+.0296	+.1444
CAISO development	.0027	.1519	.0179	.0059	.0299	.981	+.0352	+.1491
CAISO exposed post	.0055	.1507	.0366	.0121	.0616	.994	+.0322	+.1452
Recovery D1	.0099	.1492	.0664	.0336	.1210	1.000	+.0274	+.1393

λ^* = mean-service gain/action rate relative to the zero-action base. Intervals are descriptive paired circular block-2 bootstrap percentiles over origins, or the 12 frozen chronological blocks for Recovery D1. "finite" is the share of 10,000 draws with positive gain and action denominator. The recovery ratio is post-release descriptive. The coefficients express the mean-service/action trade-off in service units.

G.3 Recovery evidence chain

The recovery cohort is admitted only through a three-part evidence chain: the original frozen execution manifest, a metadata-correction record, and an independent execution verification. Full file paths and SHA-256 values are listed in the reproduction snapshot rather than repeated here. The chain preserves the original numerical objects while correcting a stale access classification; it does not alter the registered contrasts or multiplicity family.

H Reproducibility details

This section consolidates supporting protocol and verification details. The main article retains the complete method definition, theoretical statements, primary comparisons, and the statistical conclusions needed to interpret the claims.

H.1 Forecast horizon and feedback eligibility

The D-1 forecast horizon and strict-lag-2 feedback eligibility are distinct. For a delivery day t the four timestamps are: on $t-1$ the day-ahead decision is made and the day-t hourly joint region is committed; on t delivery occurs and hourly actuals accrue; at the end of t the actuals and the daily support q_t become observable; and under strict-lag-2 eligibility the post-action service of day t first becomes policy-eligible at the decision for delivery day $t+2$. The DRB trigger therefore uses only outcomes revealed through $t-2$ at decision time $t-1$. Same-day and next-day feedback never drive a trigger.

H.2 NYISO hourly region and DST aggregation

On NYISO the conformal region is an 11-dimensional hourly joint region $C_{t,h}$, with $h \in \{1, \ldots, H_t\}$. Its membership indicator is $r_{t,h} = \mathbf{1}\{y_{t,h} \in C_{t,h}\}$ and daily region support is $q_t = H_t^{-1} \sum_h r_{t,h}$. Native-support replay retains $H_t \in \{23, 24, 25\}$ on DST transition days without padding or resampling. The retrospective matched-region diagnostic instead uses the sealed, outcome-free exact-24-hour common support for every method and excludes the same NYISO transition dates. These support contracts are not interchangeable. Under modeled sure cover, daily service is $s_t = a_t + (1 - a_t)q_t$. On GEFCom/CAISO, q_t is the binary membership indicator of a daily 24-dimensional joint region. In all cases, service is a coverage-based evaluation quantity, not a measurement of electricity delivery.

H.3 Cohorts, origins, and partitions

NYISO point evaluation uses seven annual reference-to-evaluation origins $\{2018 \to 2019, \ldots, 2024 \to 2025\}$; reference-period training and calibration obey each frozen model contract. Deep-model selection uses only the frozen 2018 development calendar. The DeepAR and four-model common-UQ evaluations use six calendar origins $\{2019 \to 2020, \ldots, 2024 \to 2025\}$ and exclude the first point origin. Their separately frozen support and numeric-stack contracts remain distinct; matching calendar labels does not justify pooling their P7 rows or inference. Online feedback and pending queues reset at segment boundaries. GEFCom contributes five quarterly origins; CAISO contributes twelve development origins and six exposed post-split sensitivity origins. The recovery cohort spans 181 delivery days and twelve registered endpoint blocks, with prior outcome access disclosed. Cohorts and coverage-geometry strata are never pooled for inference.

H.4 Multivariate conformal forecasting base

The base combines a frozen point forecast, reference-period Oracle Approximating Shrinkage (OAS) covariance, and an online squared-Mahalanobis score threshold. Ellipsoidal score sublevel sets follow the multivariate construction of Xu et al. (Xu et al., 2024); Strongly Adaptive Online Conformal Prediction (SAOCP) supplies the threshold (Bhatnagar et al., 2023). The primitive unit is an hourly 11-dimensional joint region on NYISO and a daily 24-dimensional joint region on GEFCom/CAISO. The nominal common-region coverage target is 0.90; native rectangular intervals remain a separate estimand.

The adapter uses lifetime $g = 8$ and relevant algorithmic nodes extracted by abstract syntax tree (AST) from pinned public SAOCP source. The admitted result identifier is `saocp_verified_ast_lifetime8`. Manifests bind the upstream commit and tree, source-file and AST-node identities, configuration, and reference-only calibration. The delayed interface is checked for reveal order and segment resets: before the decision for delivery day t, only same-segment scores from $u \leq t - 2$ may update the threshold; completed day-t feedback is queued for $t + 2$.

Verification covers source identity and the delayed interface. The extracted kernel is evaluated in our data and adapter configuration; stepwise parity against a separately imported public SAOCP runtime is outside this verification. The manifest records `public_oracle_parity=false`. The 0.90 target is a nominal calibration setting; achieved coverage is reported empirically under the delayed adapter. The calendar-batched adapter of the end-to-end design (the region-construction subsection of the main article) is a further layer over the same pinned kernel and is stated in Section A; it is not the historical observation-index adapter used by the exposed replay cohorts.

H.5 Comparators and their applicability

COCD (conformal one-credit deferral) reallocates the current shadow-DRB action using pre-action tension. Let $m_t \in \{0,1\}$ be the current shadow action, z_t the tension, $b_t \in \{0,1\}$ the bank before the decision, and

$$r_t = \frac{1 + \#\{u < t : u \text{ is in the current observed segment, } z_u \geq z_t\}}{1 + \#\{u < t : u \text{ is in the current observed segment}\}}.$$

For either implementation, with flush indicator F_t, the transition is

$$(a_t, b_{t+1}) = \begin{cases} (1, m_t), & b_t = 1, \\ (0,0), & b_t = 0,\ m_t = 0, \\ (1,0), & b_t = 0,\ m_t = 1,\ (r_t \leq .5 \text{ or } F_t = 1), \\ (0,1), & \text{otherwise.} \end{cases}$$

The tension history is updated only after the decision. A newly observed segment resets the history and bank.

The exposed NYISO COCD diagnostic sets $F_t = \mathbf{1}\{R_{g,t} \leq K_g + 1\}$, where $R_{g,t}$ is the number of days remaining through the observed segment end and K_g is the full segment's shadow-action total. This flush closes the bank and preserves the exact shadow count in every segment (280 actions for each policy across 2,543 days and eight contiguous segments). Because K_g and the segment end contain information not available when the segment begins, this is a budget-conditioned exposed sensitivity, not a deployable online comparator.

The recovery $A^\star$ implementation instead sets $F_t = \mathbf{1}\{R_t^{\text{reg}} \leq B+1\}$ using only the pre-registered calendar end and shared global cap B. It queries neither future missingness nor a future segment end; an observed gap resets the bank and may discard one outstanding credit. It therefore guarantees at most B actions rather than exact equality in every possible path. On the released 181-day recovery path, the bank closes and $A^\star$ takes 27 actions, exactly matching D1. The $D1 - (SAOCP{+}A^\star)$ contrast remains statistically unresolved under its registered test. The allocator follows the registered-horizon information contract. Its reference implementation scans prior within-segment tensions, using $O(\sum_g n_g^2)$ time and $O(\max_g n_g)$ memory.

Table S24 Information and action-count contracts for exposed COCD and registered-horizon $A^\star$.

Variant	Flush information	Action contract	Evaluation role
Exposed COCD	Full observed-segment shadow budget and end	Exact within every segment; 280/280	Budget-conditioned exposed sensitivity
$A^\star$	Registered horizon end and global cap	At most B generally; 27/27 realized	Deployable recovery comparator

The exposed COCD flush uses information unavailable at segment start. The $A^\star$ allocator follows the registered-horizon information contract and the at-most-B action bound, with equality realized on the reported recovery path. Service uses the modeled sure-cover convention.

The production DRB rule reaches NYISO worst-seven-day service 0.730442 and matched-count rescue precision 0.494240, compared with COCD's 0.702381 and 0.461925, respectively. The $C^\star$ comparator is non-evaluable on the recovery cohort and is reported as such. Two cross-dataset replays are reported separately. The allocation sensitivity uses the distinct thresholded-count K2 variant and is not covered by the production-controller construction floor. Table S21 reports the no-retune production shortfall-sum replay; its blocked triggers enter the explicit blocked-deficit term. The two rules retain their own executed action counts.

H.6 Statistical contract

Statistical inference uses paired, cohort-specific families (Politis and Romano, 1992, 1994; Chernozhukov et al., 2018). The deep-model point and common-UQ contrasts use per-day paired circular-day-block inference within each (model, origin, segment), with origins resampled as clusters and raw and Holm-adjusted p-values reported within each declared family. The recovery-confirmatory analysis uses paired differences over twelve pre-defined blocks with block-2/block-3 bootstrap and Holm correction over a three-member registered family. The COCD contrast is a descriptive paired origin bootstrap without a pre-declared Holm family. Tables S14–S18 report the frozen per-contrast results.

Across all 32 DeepAR and four-model point/common-UQ contrasts, the block-7 and block-14 intervals have the same relation to zero. All ten DRB service contrasts (mean and worst-seven-day for five bases) remain positive at both block lengths, and their leave-one-origin-out means remain positive. The six direction-unstable LOO rows among 40 audited contrasts are five endpoints from four no-fallback common-base comparisons with P7 plus the N-HiTS–Ridge point comparison; none is used for a superiority claim.

For recovery C1, the registered block-2 Holm value is 0.00705 and the existing block-3 sensitivity gives Holm 0.00105. A separate post-release stress test enumerates the signs of the six nonzero effects among the twelve

registered blocks. Its two-sided raw value is 0.03125 and its three-member Holm value is 0.09375. This diagnostic changes the null/exchangeability assumption, is not a randomized assignment test, and does not replace the registered block-2 analysis. It instead makes the finite-support sensitivity explicit; C2 remains non-significant and C3 remains non-evaluable.

Table S25 Cohort-specific inference contracts.

Family	Paired unit / clusters	Block contract	Correction	Robustness role
NYISO exposed	Origin endpoint / 7	IID, 2, 3; 10,000 draws	None	Descriptive CI and LOO only
Recovery	Pre-defined endpoint block / 12	Circular 2 primary, 3 sensitivity; 20,000	Holm, $m = 3$	Post-release sign-flip stress
DeepAR point / common	Delivery day within segment; 7/6 origins	Circular 7 primary, 14 sensitivity; 10,000	Holm, $m = 4$ each	CI relation and LOO
Four-model point	Delivery day within segment; 7 origins	Circular 7 primary, 14 sensitivity; 10,000	Holm, $m = 8$	CI relation and LOO
Four-model common	Delivery day within segment; 6 origins	Circular 7 primary, 14 sensitivity; 10,000	Holm, $m = 16$	CI relation and LOO

H.7 Blocked-trigger applicability verification

For each cohort we replay the production continuous shortfall-sum controller on its per-day policy trace and count blocked-trigger days ($h_t \geq \theta$ with the horizon cap exhausted). The verified counts are 4 for the primary NYISO replay (all in $2023 \to 2024$), 6 for recovery, 7 for DeepAR, and 23 for the four-model evaluation (4 of 24 model-origin cells). Each replayed action vector is checked element-wise against its recorded trace, and the four-model trace SHA-256 is verified against the parsed external binding. The window-floor corollary of the main article is invoked only on windows whose triggers are served; on the exposed primary replay that is six of seven origins, while the remaining primary origin, the recovery path, and all production transfer origins contain blocked triggers, where the blocked-trigger bound of the main article keeps the realized blocked deficit as an explicit term instead of discarding the window. The check is encoded in a fail-closed, byte-deterministic machine object and script bound to the submission package.

H.8 Local deficit accounting: range checks and the chance extension

This section accompanies the support-inclusion, blocked-trigger, and physical-coupling propositions of the main article. The proofs are analytic; the material here records the synthetic range checks, the active premises, and the conditional chance extension that the main text states in one line.

Support inclusion. In the frozen guard suite, 60 random draws of (μ, Σ, R) in dimension eleven were used. Every generated interior point satisfied all eleven per-zone bounds (largest excess -0.2553, i.e. strictly inside), the bound was attained on the boundary in all 660 zone-direction trials, and the clipped requirement $r^0 = \max(0, \mu_z + \sqrt{R\Sigma_{zz}})$ was nonnegative in all draws. No evaluation outcome is read by these checks.

Blocked-trigger bound. The bound is checked by exhaustive enumeration rather than sampling: all binary shortfall paths of 11 days under the budget levels $\{0, 1, 2\}$ (6,144 paths, 30,720 window checks at width seven), with the controller state recomputed from the path exactly as the production rule computes it. The retained artifact records zero violations of the weighted form $\sum_P d < \theta + \ell + \sum_T d$ and zero violations of the count form $\sum_P d < \theta + \ell + |T|$; the maximum of $\sum_{\text{head}} d - (\theta + \ell)$ over all windows was -1.0, so the head inequality is satisfied strictly rather than on a boundary. Among the windows with positive shortfall, 2,621 had no blocked day and 10,320 had a last-non-trigger index below the lag, i.e. the two structural cases of the proof were both exercised.

Active premises. With $(\ell, m, n, \theta) = (2, 3, 8, 1)$ the sufficient-memory condition $m \geq n - \ell$ fails: from an empty history with ample budget and base coverage $q = 0$ on every day of the window, the state sees only the three most recent eligible deficits and the realized shortfall path is $d = (1, 1, 0, 0, 0, 0, 1, 1)$, whose total 4 exceeds $\theta + \ell = 3$. This is the verified counterexample. A window crossing a segment reset rebuilds the state and can exceed the bound; a partial-response commitment is a different method because the full-service step (an executed commitment makes the day's shortfall zero) no longer holds; and the blocked set T is realized ex post, which is why the bound conditions on it; the blocked term of the inequality is explicit rather than assumed away.

Physical coupling. The coupling was checked on 60 synthetic days per action branch: on uncommitted days the adequacy fraction was at least the coverage fraction in all 60 draws, and on committed days the identity $1 - s^{\text{phys}} = (1 - s) + (1 - s^{\text{phys}})$ held in all 60. Where a reported lower bound is clipped at zero the statement is the weak inequality, since clipping can only raise a negative bound to zero.

Conditional chance extension. The following extension replaces modeled full service by a probabilistic commitment. Fix the window width $n = 7$, the width of the service endpoints of the main article, and keep the general lag ℓ and memory $m \geq n - \ell$. Write d_t for the *realized* service deficit of day t: it is zero on a committed day whose commitment succeeds, it may be positive on a committed day whose commitment fails, and on an uncommitted day it equals the realized deficit $1 - q_t$ of the base service. This realized deficit replaces the modeled sure-cover deficit $d_t = (1 - a_t)(1 - q_t)$ of the main article, which sets a committed day to zero by assumption. Let V be the event that some same-segment seven-day window satisfying the deterministic premises below has total realized shortfall at least $\theta + \ell$, and let $N = \sum_t a_t$ count the executed commitments.

Proposition 1 (Conditional chance bound)**.** *Consider the shortfall-sum rule with lag ℓ, window width $n = 7$, memory $m \geq n - \ell$, and threshold $\theta > 0$, evaluated on windows that lie inside one segment, use complete observed support, and contain no blocked trigger. Suppose that every eligible trigger executes a commitment, that the budget is not exhausted on the analyzed horizon, and that each committed day satisfies*

$$P(d_t > 0 \mid \mathcal{G}_t, a_t = 1) \leq \varepsilon,$$

where $\mathcal{G}_t$ is the pre-decision information set. Then

$$\begin{aligned} P(V) \;&\leq\; P\big(\exists\, t:\; a_t = 1,\; d_t > 0\big) \\ &\leq\; \sum_t \mathbb{E}\big[a_t\, P(d_t > 0 \mid \mathcal{G}_t, a_t = 1)\big] \;\leq\; \varepsilon\, \mathbb{E}[N]. \end{aligned} \tag{1}$$

Proof. Condition on the event that every executed commitment succeeds, that is $d_t = 0$ on every day with $a_t = 1$. On that event the realized path satisfies the deterministic premises of the blocked-trigger bound term by term: no trigger in the window is blocked, the memory condition holds, and every committed day has zero shortfall, exactly the full-service case for which the deterministic bound gives total window shortfall strictly below $\theta + \ell$. A violating window therefore cannot occur on that event, so V implies the existence of at least one committed day with positive shortfall, which is the first inequality. The second inequality is a union bound over committed days; the third substitutes the conditional failure assumption, using that a_t is $\mathcal{G}_t$-measurable, so the executed count is $N = \sum_t a_t$. No step selects a particular day of the window, and no assumption is made about which committed day fails. □

The extension requires the displayed conditional per-action success model in addition to the deterministic premises. Establishing that model for an operational reserve mechanism would permit empirical use of the chance bound; the present load evaluation uses the modeled and finite-response endpoints.

H.9 Frozen controller-parameter sensitivity

On the exposed seven-origin primary NYISO trace, 168 configurations cross $m \in \{3, 4, 5, 6, 7, 9, 12\}$, $\theta \in \{.5, .75, 1, 1.25, 1.5, 2\}$, and budget fractions $\{.05, .10, .15, .20\}$. The window lengths 3, 7, and 14 change only the reported estimand. The 27-setting local neighborhood fixes $m \in \{5, 6, 7\}$, $\theta \in \{.75, 1, 1.25\}$, and budgets $\{.10, .15, .20\}$. No grid point replaces the production reference.

Table S26 Exposed, non-tuning sensitivity of the production DRB parameter rule.

Scope	ΔW_3 / ΔW_7 / ΔW_{14}	Action rate	ΔL_1	Inactive	Pressured
Full grid (168)	.0000–.1151 / .0000–.1148 / .0000–.0616	.0020–.1998	.0019–.1823	52	89
Local (27)	.0099–.0794 / .0034–.0978 / .0055–.0455	.0420–.1970	.0383–.1801	1	15
Production $(6, 1, .15)$	.0794 / .0944 / .0404	.1101	.1008	no	yes (1 origin)

Worst-window gains concern modeled sure-cover service. Inactive means macro worst-seven-day gain below 0.01; pressured means at least one origin has a blocked trigger. Positive ΔL_1 means greater modeled loss. The grid does not select a replacement production configuration.

"Inactive" means macro worst-seven-day gain below 0.01. "Cap-pressured" means at least one origin has a blocked trigger, so the global floor cannot be invoked. Every one of the 168 configurations has $\Delta L_1 > 0$. The production reference has four blocked triggers in one origin; its daily action, service, state, cap, and revealed-shortfall paths match the admitted primary trace in all seven origins. The full retrospective grid characterizes parameter sensitivity while retaining the frozen production reference.

H.10 Fair deep-baseline selection and compute contracts

All five deep baselines use the same 2018-only train/validation/selection calendar, three-quantile normalized weighted quantile loss, strict-lag-2 outcome boundary, and frozen 2019–2025 evaluation origins. The finite grids compare load-level and P7-residual targets; the winner is immutable before annual evaluation.

Table S27 Fair deep-baseline selection and compute contracts.

Model	Dev. runs	Frozen winner	Training cap	Implementation and retained failures
DeepAR	8/8	P7 residual; context 336; hidden 40; 2 layers	50 epochs with patience 5; refit 32 epochs	GluonTS 0.16.3; one evaluation serialization failure retained before the repaired complete run
TFT	2/4	P7 residual; encoder 336; hidden 64; 2 LSTM layers	20 epochs; 350 batches/epoch; patience 5	PyTorch Forecasting 1.8.0; two failed development candidates retained and excluded fail-closed
N-HiTS	4/4	P7 residual; input 336	1,000 steps; validation every 50; patience 5	NeuralForecast 1.7.7; no development-candidate failure
PatchTST	4/4	P7 residual; input 336	1,000 steps; validation every 50; patience 5	NeuralForecast 1.7.7; public interface uses no future exogenous variables
ProLoaF	4/4	P7 residual; history 168	40 epochs; patience 5	public ProLoaF core at commit `e4a85c9`; no development-candidate failure

Training and prediction ran on the primary Linux host with an RTX 4090. Table S20 reports successful train–predict time and is not a speed ranking. The DeepAR common-shell replay and the final four-model common partition used the same sealed inputs, code, and SAOCP pin on local Windows Python 3.11 after retained native-host failures; the latter required 30 failed remote attempts for that partition. TFT, N-HiTS, and ProLoaF use the known-future inputs supported by their adapters, whereas PatchTST does not. Thus fairness denotes aligned data, information cutoff, development loss, and evaluation—not identical model interfaces or vanilla-paper replication.

H.11 Exact same-count offline allocation ceiling

For each exposed NYISO origin, let c_t be no-fallback daily service, x_t the binary oracle action, and K the realized DRB action count. The first stage solves

$$\max_{x,z} \; z \quad \text{subject to} \quad |w|^{-1}\sum_{t\in w}\{c_t + (1-c_t)x_t\} \geq z \;\forall w, \qquad \sum_t x_t = K,$$

where w ranges only over seven-day windows contained in one sealed segment and $x_t \in \{0,1\}$. A second stage maximizes rescued shortfall while retaining the optimal z to numerical tolerance. SciPy/HiGHS returns zero MIP gap, and independent recomputation matches every origin objective within 10^{-8}.

The oracle and DRB each use 280 actions overall, with equal counts in every origin. Oracle-minus-DRB worst-seven-day headroom is 0.181122 on average and positive in all seven origins; the block-2 descriptive interval is [0.149660, 0.215136]. The oracle sees the full outcome path. It is therefore an exact fixed-base, fixed-count optimization ceiling for this within-segment retrospective replay.

H.12 Additional diagnostic displays

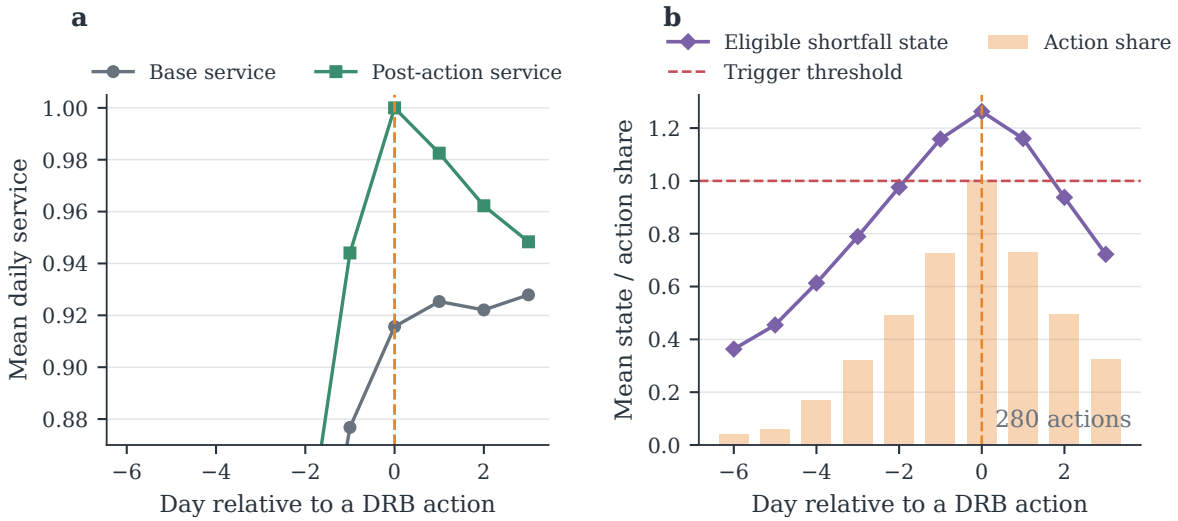


Figure S8 Event-aligned production-DRB diagnostics on exposed NYISO. Day zero is an executed action; 280 action-centered paths are pooled within their origins and segments, with available offsets from −6 to +3. (a) Base support q_t and post-action modeled service $s_t = a_t + (1 - a_t)q_t$ under sure cover. (b) Eligible continuous shortfall state h_t, threshold one, and the fraction of aligned observations with an action at each offset. Offset zero has action share one by construction; windows can overlap and have different available counts near segment boundaries. The display summarizes event-aligned associations for production DRB. Prepared with AI-assisted code

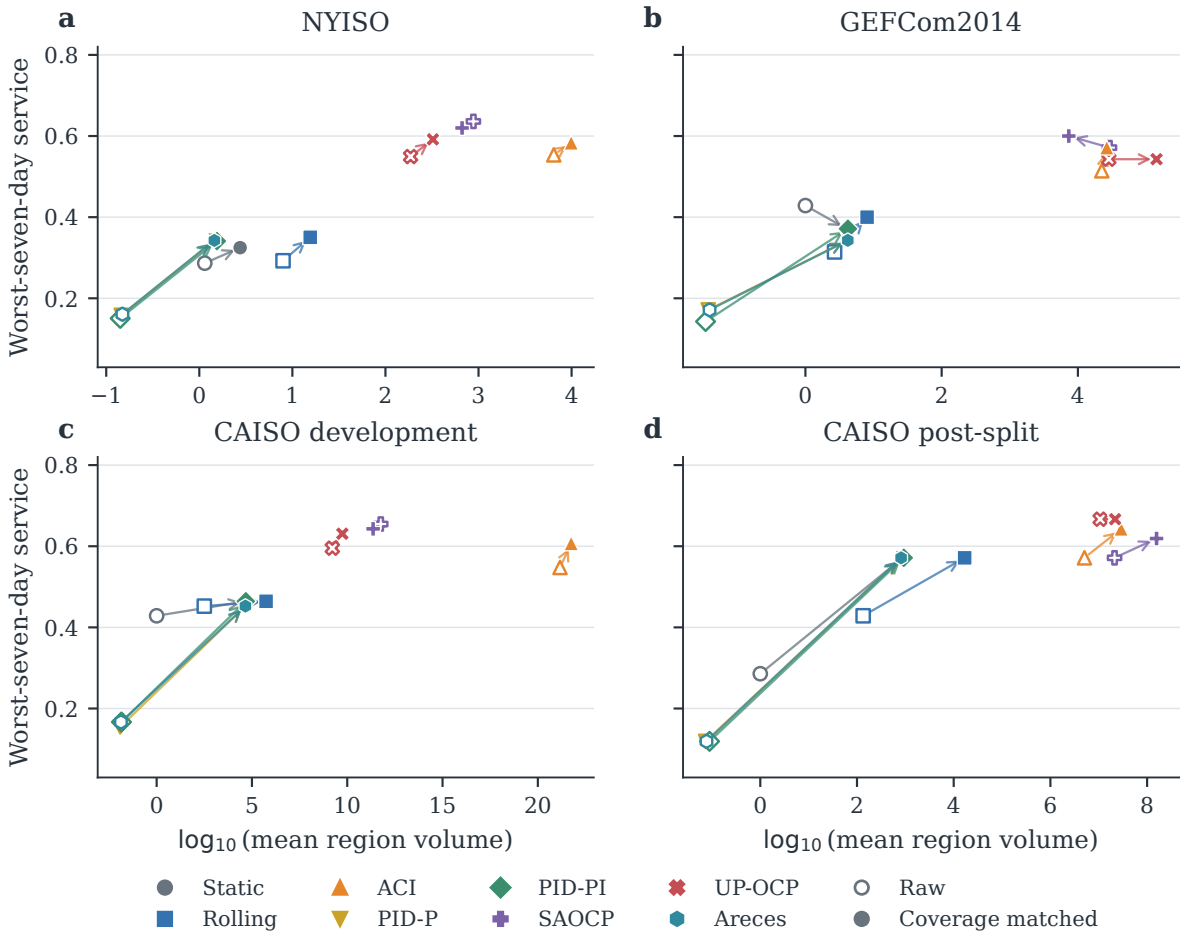


Figure S9 Region-only service–volume trade-offs in four exposed cohorts: (a) NYISO, (b) GEFCom2014, (c) CAISO development, and (d) CAISO exposed post-split sensitivity. Hollow markers are raw results; filled markers are retrospective 0.90 realized-coverage matches, with arrows joining the same method. Axes show the base's minimum seven-day mean support and $\log_{10}$ mean region volume; no fallback is used. Matching uses completed evaluation outcomes and is not an online controller. Volumes are interpreted within each cohort's geometry and support, not ranked across cohorts. The static-volume-matched view is tabulated in Section D.1. Prepared with AI-assisted code

Table S28 Point-forecast performance of energy baselines on exposed NYISO.

Model	Macro MAE	Effect vs Ridge	Block-7 95% CI	Holm p
Ridge lag-2	43.4423	–	–	–
N-HiTS	43.5444	+0.1274	[-0.4764, 0.7921]	1.000
PatchTST	43.6129	+0.2015	[-0.5258, 0.8966]	1.000
DeepAR	43.9913	+0.5166	[-0.0571, 1.1270]	0.079
TFT	44.3929	+0.9822	[0.1577, 2.0042]	0.036
LightGBM	45.8208	–	–	–
ProLoaF	45.9438	+2.5293	[1.2913, 3.6476]	0.002
CatBoost	47.3609	–	–	–
P7 official	58.4835	–	–	–

Effects are candidate-minus-Ridge paired daily MAE; negative values favor the candidate. Classical-model entries without a frozen direct Ridge contrast are left blank rather than assigned unpaired tests. Macro MAE summarizes seven point-evaluation origins. Deep models are development-selected P7-residual adaptations. DeepAR and four-model Holm families remain separate. Point accuracy and common-region service are evaluated as complementary endpoints.

Table S29 Native rectangular probabilistic forecast metrics on exposed NYISO.

Model	RMSE	WAPE (%)	Marginal cov.	11-zone cov.	Width	WQL
DeepAR	74.945	2.796	0.862	0.315	166.7	0.0148
N-HiTS	73.165	2.772	0.890	0.395	184.0	0.0146
PatchTST	73.704	2.776	0.892	0.412	181.7	0.0147
TFT	76.738	2.825	0.873	0.341	164.9	0.0149
ProLoaF	75.787	2.925	0.922	0.513	216.3	0.0155

These are native rectangular quantile outputs. Their marginal and simultaneous coverage characterize the native quantile geometry; the common multivariate OAS-region evaluation is reported separately in Section D.4. The five models are P7-residual adaptations. Coverage and interval width describe native forecast rectangles, without fallback service. WQL is weighted quantile loss; comparisons here use the native rectangular forecast geometry.

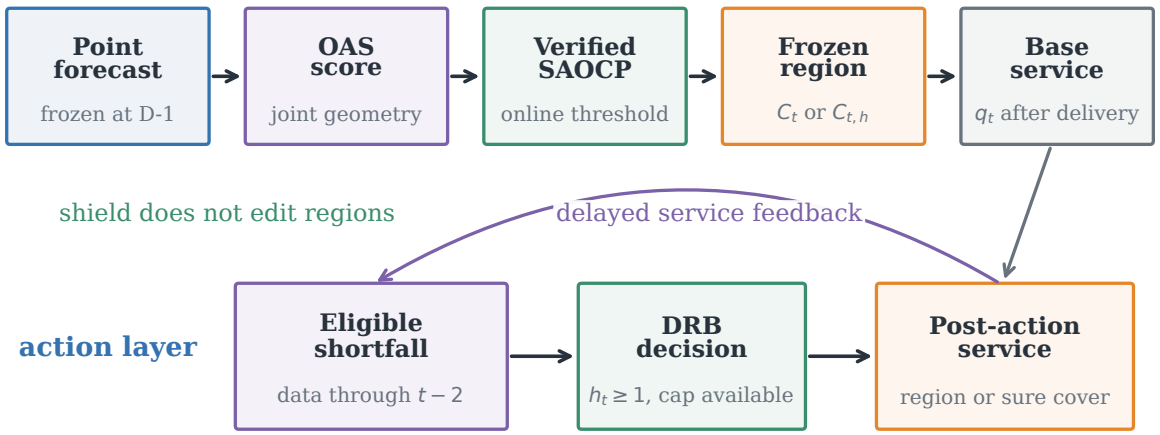


Figure S10 Architecture of the earlier observation-index replay pipeline used by the historical and recovery evaluations in this supplement. Region construction (frozen point forecast, OAS geometry, verified online threshold, frozen region) is separated from the action layer, which consumes eligible delayed service feedback through $t-2$ and leaves the committed region unchanged. "Verified" denotes the pinned source and AST-node identity checks of Section H.4, and full service on committed days is modeled. The current study's calendar-batched adapter is specified separately in Section A and the main article. Prepared with AI-assisted code

H.13 Result objects and access status

Reported numbers are read from frozen machine objects in the reproduction snapshot or from manifest-bound external artifacts. Manifests record input and code identities, execution metadata, and access flags. Frozen artifact replay, controller-trace verification, and model retraining with externally acquired data provide complementary reproducibility routes.

Recovery is admitted through the original execution package, its metadata-correction record, and independent execution verification. Prior outcome payload access occurred before registered endpoint computation; the repair retained the original scientific configuration and registered analysis. The metadata-correction record documents the prior-access classification of this technical-recovery evaluation.

Large four-model tensors remain externally bound by SHA-256; the compact policy trace required for verification is included in the reproduction snapshot. An external binding identifies an artifact but does not itself provide public access to its bytes. The Data availability statement in the main article distinguishes provider-hosted source data, released compact artifacts, and external artifacts. The legacy review snapshot remains available as an additional reproducibility route.

References

Angelopoulos A, Candès E, Tibshirani RJ (2023) Conformal PID control for time series prediction. In: Advances in Neural Information Processing Systems, vol 36. Curran Associates, Inc., pp 23047–23074, https://doi.org/10.52202/075280-1000

Areces F, Mohri C, Hashimoto T, et al (2025) Online conformal prediction via online optimization. In: Proceedings of the 42nd International Conference on Machine Learning, Proceedings of Machine Learning Research, vol 267. PMLR, pp 1604–1649, URL https://proceedings.mlr.press/v267/areces25a.html

Bhatnagar A, Wang H, Xiong C, et al (2023) Improved online conformal prediction via strongly adaptive online learning. In: Proceedings of the 40th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol 202. PMLR, pp 2337–2363, URL https://proceedings.mlr.press/v202/bhatnagar23a.html

Challu C, Olivares KG, Oreshkin BN, et al (2023) NHITS: Neural hierarchical interpolation for time series forecasting. Proceedings of the AAAI Conference on Artificial Intelligence 37(6):6989–6997. https://doi.org/10.1609/aaai.v37i6.25854

Chernozhukov V, Wüthrich K, Zhu Y (2018) Exact and robust conformal inference methods for predictive machine learning with dependent data. In: Bubeck S, Perchet V, Rigollet P (eds) Proceedings of the 31st Conference on Learning Theory, Proceedings of Machine Learning Research, vol 75. PMLR, pp 732–749, URL https://proceedings.mlr.press/v75/chernozhukov18a.html

Gibbs I, Candès E (2021) Adaptive conformal inference under distribution shift. In: Advances in Neural Information Processing Systems, vol 34. Curran Associates, Inc., pp 1660–1672, URL https://proceedings.neurips.cc/paper/2021/hash/0d441de75945e5acbc865406fc9a2559-Abstract.html

Gürses-Tran G, Oppermann F, Monti A (2023) ProLoaF: Probabilistic load forecasting for power systems. SoftwareX 23:101487. https://doi.org/10.1016/j.softx.2023.101487

Ke G, Meng Q, Finley T, et al (2017) LightGBM: A highly efficient gradient boosting decision tree. In: Advances in Neural Information Processing Systems (NeurIPS), URL https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

Lim B, Arık SÖ, Loeff N, et al (2021) Temporal fusion transformers for interpretable multi-horizon time series forecasting. International Journal of Forecasting 37(4):1748–1764. https://doi.org/10.1016/j.ijforecast.2021.03.012

Liu T, Dobriban E, Orabona F (2026) Online conformal prediction via universal portfolio algorithms. In: Proceedings of the 43rd International Conference on Machine Learning, Proceedings of Machine Learning Research, vol 306. PMLR, pp 75891–75952, URL https://proceedings.mlr.press/v306/liu26y.html

Nie Y, Nguyen NH, Sinthong P, et al (2023) A time series is worth 64 words: Long-term forecasting with transformers. In: International Conference on Learning Representations (ICLR), URL https://openreview.net/forum?id=bmaunQqYD

Politis DN, Romano JP (1992) A circular block-resampling procedure for stationary data. In: LePage R, Billard L (eds) Exploring the Limits of Bootstrap. John Wiley & Sons, New York, p 263–270, URL https://statistics.stanford.edu/technical-reports/circular-block-resampling-procedure-stationary-data

Politis DN, Romano JP (1994) The stationary bootstrap. Journal of the American Statistical Association 89(428):1303–1313. https://doi.org/10.1080/01621459.1994.10476870

Prokhorenkova L, Gusev G, Vorobev A, et al (2018) CatBoost: unbiased boosting with categorical features. In: Advances in Neural Information Processing Systems (NeurIPS), URL https://proceedings.neurips.cc/paper/2018/hash/14491b756b3a51daac41c24863285549-Abstract.html

Salinas D, Flunkert V, Gasthaus J, et al (2020) DeepAR: Probabilistic forecasting with autoregressive recurrent networks. International Journal of Forecasting 36(3):1181–1191. https://doi.org/10.1016/j.ijforecast.2019.07.001

Xu C, Jiang H, Xie Y (2024) Conformal prediction for multi-dimensional time series by ellipsoidal sets. In: Proceedings of the 41st International Conference on Machine Learning, Proceedings of Machine Learning Research, vol 235. PMLR, pp 55076–55099, URL https://proceedings.mlr.press/v235/xu24m.html