When the Counterfactual Is Adjudicated: A Grid-Curtailment Benchmark Separating Structure from One-Sided Estimation

Anonymous authors
Paper under double-blind review

Abstract

Methods that recover an outcome suppressed by an intervening policy are often evaluated against a simulator written by the evaluator, and their ranking changes with the simulator. On one power grid, the suppressed outcome is adjudicated instead. Whenever the Brazilian system operator restricts a wind or solar plant, it publishes a half-hourly estimate of what the plant would have generated. The estimate is computed from wind-speed or irradiance measurements at the plant, under the procedure used to settle compensation payments. It is not a second meter reading, but the process that produces it predates this work and uses those measurements, which the benchmark withholds from every method. This paper builds a benchmark from the published record. The benchmark covers 18 wind clusters over $40{,}896$ hours and 22 solar clusters over $18{,}984$ hours, and it enforces five invariants as build-time assertions. Its evaluation protocol declares what each method may read, splits the record by date, and scores recovery on the hours the operator restricts separately from false corrections on the hours it does not. At a configuration fixed before any test score is seen, a structured latent decomposition models the censoring in its likelihood and floors its predictions at realised output. On restricted hours it beats a least-squares ratio estimator on solar but loses on wind. A ratio estimator that instead fits the 0.75 quantile of realised output, with calendar inputs and no latent structure, beats the decomposition on censored hours on all 15 seeds of both technologies (on seeds 1 to 5, 0.1237 against 0.1327 on solar and 0.1388 against 0.1713 on wind). On solar it also cuts the error on unrestricted hours from 0.2483 to 0.0198. With or without calendar terms in its inputs, a least-squares fit loses to the decomposition on every solar seed, and a 0.75-quantile fit with the same inputs beats it on at least 14 of 15. For those inputs, the change of loss alone reverses the ordering. The quantile grid is fixed in advance and reported in full, and this arm is its best cell on solar. A validation block inside the training window picks cells that still beat the decomposition on 14 of 15 solar seeds and all 15 wind seeds. Over all clusters, the solar advantage of the decomposition over least squares therefore does not show a benefit from its latent structure. On the solar clusters live throughout the record, the decomposition gains more and keeps a lead of about 0.02 over the quantile fit on restricted hours, so one-sided estimation reproduces only part of that gain. Scoring the unrestricted hours separately is what exposes the false corrections of the decomposition.

1 Introduction

A potential outcome is what would happen under an intervention that does not happen. Estimating one is routine; checking the estimate is not, because the quantity being estimated is never realised. The usual answers are to assume the mechanism hiding it (Rubin, 1974; Ding, 2024) or to simulate the outcome and score against the simulation (Hill, 2011; Dorie et al., 2019), and neither is neutral. Treatment-effect estimators change rank across semi-synthetic datasets whose only difference is the generating process (Curth et al., 2021).

Electricity dispatch offers an unusual escape. When a grid cannot take all the output on offer, its operator instructs a wind or solar plant to generate less than the weather would allow. Because that

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

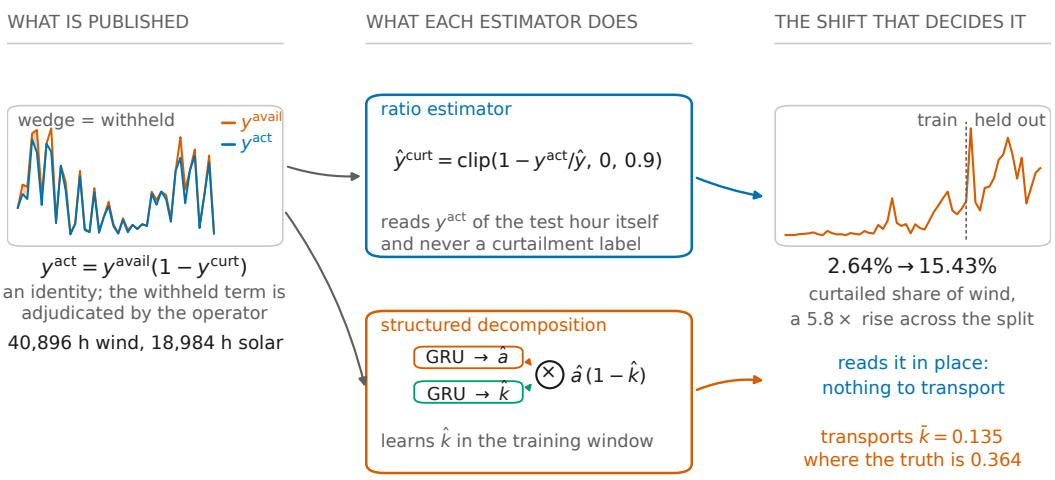


Figure 1: The benchmark, the two families of estimator it compares and the policy shift they face. **Left**, 14 days of the record, where realised output is metered and the withheld wedge is the adjudication by the operator of what the restriction costs. **Centre**, what each estimator computes, with $\hat{a}$ and $\hat{k}$ short for $\hat{y}^{\text{avail}}$ and $\hat{y}^{\text{curt}}$. **Right**, the policy shift the split imposes and its consequence for each. An estimator that learns a censoring magnitude must learn it in training and use it in the held-out window; one that reads the censored observation transfers nothing.

instruction costs the plant money, the operator has to say how much. For every restricted half hour, the Brazilian operator, ONS, publishes what the plant would have generated absent the restriction. It obtains this figure from a productivity curve fitted to the metered history of that plant and evaluated at the anemometry or irradiance the plant streams in real time. The figure is produced under the procedure used to settle compensation through the market operator (Operador Nacional do Sistema Elétrico (ONS), 2026; Operador Nacional do Sistema Elétrico, 2025). It is an estimate and this paper treats it as one. Nobody evaluating a method chooses it, so it is usable where a simulator is not. The procedure predates this benchmark and is contested by a counterparty with money at stake. The benchmark withholds the anemometry and irradiance of the plant from everything scored against it.

The problem curtailment poses is not one of missing data. Curtailment is informative censoring, applied when the grid cannot take the output rather than at random, so a model fitted to observed output learns the dispatch policy mixed with plant physics. The policy is moving, with the energy-weighted share of available wind energy withheld rising from 2.6% before 2025-01-05 to 15.4% after. The natural response is to exploit the factorisation, estimating availability and a curtailment ratio as separate factors under constraints that make them identifiable. At a configuration fixed in advance, such a decomposition beats a least-squares ratio estimator on solar once $y^{\text{avail}} \geq y^{\text{act}}$ is imposed on its output, and loses on wind by margins far outside seed noise. Over all clusters, the benchmark shows that the solar gain does not require the structure. A ratio estimator that fits the 0.75 quantile of realised output, with calendar inputs and no latent factors, beats the decomposition on both technologies and makes fewer false corrections where no restriction applies.

Contributions.

1. **A benchmark whose counterfactual the evaluator does not write** (Section 4). It covers $40{,}896$ hours over 18 wind clusters and $18{,}984$ hours over 22 solar clusters. The suppressed outcome is supplied per unit and per half hour by the operator that imposes the restriction, under a procedure published in advance and used for financial settlement (Table 1). The decomposition identity and four other invariants are enforced as executable build-time assertions. The fifth catches the defect reported here, 2,283 consecutive hours of forward-filled irradiance that the other four assertions and a manual inspection miss.
2. **An evaluation protocol, with the measurement that makes it necessary** (Section 5). It comprises a declared information set enforced by a permutation guard, a policy-shift split and paired metrics. Censored-hour MAE scores reconstruction where the operator adju-

Table 1: Where the counterfactual of a benchmark comes from. The last column shows the distinction that matters. A counterfactual specified by the benchmark author can match the assumptions of the estimators it scores, and rankings then move with that specification (Curth et al., 2021). The next three rows escape this dependence at the cost of moving to a different unit or to an expectation. The counterfactual here is produced per unit and per half hour by the party that imposes the intervention, under a procedure published in advance. It is an estimate, not a measurement. Section 3 gives the mechanism and Section 7 gives the consequences. Sources: IHDP and ACIC (Hill, 2011; Dorie et al., 2019); Twins (Louizos et al., 2017); interventional screens (Chevalley et al., 2025); off-policy evaluation over a logged policy (Dudík et al., 2011; Saito et al., 2021).

Benchmark	Counterfactual produced by	Chosen by whoever built the benchmark
IHDP, ACIC	a response surface	yes
Twins	the realised outcome of a matched sibling	no, but a different unit
Interventional screens	realising both potential outcomes	no, but different cells
Logged-policy OPE	reweighting by a known propensity	no, and only in expectation
This benchmark	**the intervening operator, from the instruments of the unit, under a published settlement procedure**	**no**

dicates a suppression, and uncensored-hour MAE scores false corrections where realised output already equals the adjudicated availability. Censored-hour error alone rewards correcting upward everywhere, and global error alone ranks the best-informed arm below doing nothing, so the protocol reports both.

3. **A test that separates latent structure from one-sided estimation** (Section 6). At a configuration fixed in advance, the structured decomposition with $y^{\text{avail}} \geq y^{\text{act}}$ imposed on its output beats a least-squares ratio estimator on every solar seed, even when that estimator reads the same inputs, and loses on wind. A ratio estimator that fits the 0.75 quantile of realised output instead, with calendar inputs and no latent structure, beats the decomposition on all 15 seeds of both technologies and makes fewer false corrections where no restriction applies. The quantile grid is fixed in advance and reported in full. Over all clusters, the solar gain of the decomposition therefore does not show a benefit from its structure.

The build-time assertions, the arms and the extraction rule of every figure will be released with the benchmark.

2 Related Work

Evaluating counterfactuals without a counterfactual. In practice, potential-outcome estimation is checked against manufactured ground truth. Standard treatment-effect benchmarks pair real covariates with simulated outcomes (Hill, 2011; Dorie et al., 2019), and the representation-learning methods built on them (Johansson et al., 2016; Shalit et al., 2017; Louizos et al., 2017) are compared on that basis. Curth et al. (2021) show the cost. Estimator rankings move with the simulation. Table 1 places this benchmark against the constructions that address the same difficulty. Off-policy evaluation is closest in spirit (Dudík et al., 2011; Saito et al., 2021) but differs in kind. It recovers the suppressed outcome in expectation through a known propensity, whereas here the outcome is adjudicated per unit and per half hour. The two constructions that realise both potential outcomes do so on *different* units: siblings (Louizos et al., 2017) or cells (Chevalley et al., 2025).

The counterfactual used here is an estimate, and this paper claims no more for it. What separates it from a simulated benchmark is exogeneity rather than exactness. The label-generating process predates this benchmark and pays out to a counterparty positioned to contest it. The wind speed or irradiance it reads at the plant is withheld from every method scored against it. **The search reported in Appendix A finds no other public benchmark in which the counterfactual is supplied per unit and per period by the party that imposes the intervention, under a procedure published in advance.** Two 2026 studies use the same records without releasing a benchmark, one estimating suppressed wind generation from SCADA anemometry across more than 1,000 plants (Nascimento et al., 2026) and one characterising the operational drivers of Brazilian curtailment (Vieira et al.,

2026). That an engineering literature exists to improve this quantity is itself evidence that it carries error, which Section 7 treats as such. Elsewhere, recent curtailment work scores a quantile model of available turbine power on synthetic censoring and on curtailment releases (Wang & Wang, 2027).

Censored data, structured latent models and the domain. Modelling an outcome seen only through a censoring mechanism runs from limited dependent variables (Tobin, 1958), censored regression quantiles (Powell, 1986) and proportional hazards (Cox, 1972) to neural survival models (Wiegrebe et al., 2024). In most of that literature, censoring falls at a known threshold, is assumed non-informative or is modelled through its *time* rather than the magnitude withheld. Here, censoring is applied precisely because output is high, and what is withheld is a continuous, zero-inflated fraction. The tested model applies the standard latent-world-model approach (Ha & Schmidhuber, 2018; Hafner et al., 2025) to that decomposition. Power forecasting is a mature applied field (Tawn & Browell, 2022) and curtailment is documented and rising (International Energy Agency, 2025; Vieira et al., 2026). Both are taken as given. The other arms use the gradient-boosted family a practitioner uses (Ke et al., 2017), and the domain is used only for what it uniquely supplies. This paper is not a contribution to power forecasting.

3 The Censoring Problem

For each resource cluster i and hour t, the contract records time-varying installed capacity $P_{i,t}$, realised generation $g_{i,t}$ and generation curtailed by a dispatch instruction $c_{i,t}$. These define

$$y^{\text{avail}}_{i,t} = \frac{g_{i,t} + c_{i,t}}{P_{i,t}}, \qquad y^{\text{curt}}_{i,t} = \frac{c_{i,t}}{g_{i,t} + c_{i,t}}, \qquad y^{\text{act}}_{i,t} = y^{\text{avail}}_{i,t}\left(1 - y^{\text{curt}}_{i,t}\right), \tag{1}$$

where $y^{\text{act}} = g/P$ is the capacity factor recorded directly by grid telemetry. y^{avail} is the capacity factor the resource would have realised absent curtailment. y^{curt} is the fraction of that potential withheld by dispatch, and zero where no potential is available. Equation 1 holds exactly, but it inherits the differing standing of its inputs. Realised generation g is metered. The withheld quantity c is the Geração Não Realizada Apurada that ONS publishes for the half hours in which it limits a plant. ONS defines it as reference generation minus realised generation, floored at zero (Operador Nacional do Sistema Elétrico (ONS), 2026). Reference generation is an estimate, as ONS states. It is a productivity curve fitted from the metered history of that plant, revised annually and evaluated at the anemometry or irradiance that the plant streams in real time (Operador Nacional do Sistema Elétrico, 2025). So y^{avail} is not a second meter reading. What separates it from a simulated counterfactual is provenance rather than precision. A simulated response surface is chosen by whoever builds the benchmark and can match the assumptions of the estimators it scores. Curth et al. (2021) identify this circularity. RO-AO.BR.13 predates this benchmark by years, and the wind speed or irradiance it reads at the plant is withheld from every method. It settles compensation payments, so a counterparty with money at stake contests it. Section 7 states what follows for interpretation.

Curtailment induces censoring that is informative, not random. Dispatch curtails a resource not as a deterministic function of weather but when transmission is congested, stability limits bind, or maintenance and load require curtailment. A model fitted to predict y^{act} from weather therefore risks fitting the dispatch policy rather than the physical response of the resource. The two are conflated in every hour with $c_{i,t} > 0$. Figure 2b in Appendix B shows how far apart they are. On $736{,}128$ wind cluster-hours, weather explains a cross-validated 0.1804 of the variance in y^{curt}; on shuffled labels it explains -0.0001. Weather leaves most of the variance in dispatch unexplained, and that gap is what a structured decomposition has to exploit. Recovering y^{avail} from y^{act} under this censoring, scored against the published value, is the estimation problem the rest of the paper addresses.

4 The Benchmark

How rare an adjudicated counterfactual is. A counterfactual of this kind is not unique to ONS, but its form differs sharply across the operators that disclose curtailment below the system level. Table 16 in Appendix N compares the three most relevant. Of the three, only ONS publishes a per-unit, per-period estimate of the unrestricted quantity itself. The ERCOT figure must be derived from a High Sustained Limit whose tracking of weather-available output is untested here (Electric Reliability Council of Texas (ERCOT), 2026), and the CAISO figure cannot be attributed to a resource at

Table 2: The benchmark, spanning 2021-10-01 to 2026-05-31, split at 2025-01-05 on wind and 2025-10-06 on solar. A cluster-hour is one (hour, cluster) cell. Rates are computed over cells whose cluster has positive installed capacity, so a cluster not yet built does not dilute them, and the curtailed share is of available rather than realised energy. The two technologies have different windows: wind is scored from 2021-10 and solar from 2024-04, when its record opens.

	Wind	Solar
Resource clusters	18	22
Hours	40,896	18,984
Held-out cluster-hours	220,842	125,290
of which censored	83,833	35,614
Curtailment record opens	2021-10	2024-04
Censored share of live cells, train	16.1%	26.6%
held out	38.0%	30.2%
Curtailed share of available energy, train	2.64%	20.16%
held out	15.43%	24.25%
Installed capacity growth across the split	1.07×	1.15×

all (California Independent System Operator (CAISO), 2026). Four further systems publish related figures: the unconstrained intermittent generation forecast from AEMO, the German *Ausfallarbeit* under redispatch, the EirGrid dispatch-down and the British bid-offer acceptances. None of these figures has been measured against the same criteria. Measuring them and checking a derived counterfactual against the adjudicated one is the future work this paper prioritises.

Contract. Table 2 states the shape of the benchmark. The benchmark is built by the pipeline of a predecessor paper (Anonymous, 2026), whose published curtailment figures rest on the same field redefined in Section 3 and are affected accordingly. Its 79 resource clusters are formed separately by technology with capacity-weighted k-means over plant coordinates matched to substations, and 40 of them are scored.[1] The end date is a data-integrity decision, not an arbitrary cut. For the reason given below, it is the last date on which every weather covariate the contract depends on is published.

Build-time guards. Every invariant the contract depends on is checked at build time, not merely documented. The four guards inherited unchanged from the pipeline of the predecessor are time-zone alignment (UTC timestamps from NASA POWER align with local-time records from ONS), fold disjointness (chronological, non-overlapping segments), target range (capacity factors and curtailment ratios, which the build clips into $[0, 1]$, are finite and stay there) and reconstruction identity (Equation 1 holds to floating-point tolerance on every row). This project adds a fifth guard. No weather column may be forward-filled for more than 168 consecutive hours. The threshold is set with a 2.3x margin over 72 hours, the longest legitimate run of missing hours elsewhere in the contract.

On a real rebuild this guard flags 2,283 consecutive hours that the build would forward-fill in a single column, the all-sky irradiance of one solar cluster. All four inherited guards pass in the same build, and a manual inspection of those hours misses the forward fill. The cause is a publication lag rather than a caching defect. NASA POWER publishes irradiance roughly three months after its other variables, so the tail of a recent contract repeats the last real value instead of staying missing. This lag sets the end date above. The defect that matters is the one no existing check covers.

5 Evaluation Protocol

Ordinary forecasting evaluation assumes the predicted quantity is missing at random and that any competent metric averaged over enough hours ranks models correctly. Neither assumption holds here. Each departure below is found through a result that otherwise appears reversed in this paper.

[1] The constrained-off record covers only centrally dispatched plants, so no distributed-generation cluster appears in it at any hour. The 38 distributed-generation clusters and one utility cluster whose plants never enter the archive have no counterfactual at any hour and are excluded from every metric rather than scored as uncurtailed. What remains is 18 wind and 22 solar clusters.

Observation setting. Every deployable arm trains and predicts from the same restricted view and never from the published c, y^{avail} or y^{curt} defined in Section 3. The view is y^{act} with weather, calendar features and grid-state proxies. Evaluation then scores the prediction of y^{avail} against the published value. Two arms are labelled exceptions: one is given y^{curt} as well, to measure what perfect curtailment supervision contributes, and one is fitted directly on y^{avail}.

The grid-state proxy. The oversupply and network limits behind curtailment are not published per resource, so every arm is given the same observable proxy for cluster i at hour t. It is the capacity-weighted mean realised output of every *other* wind or solar cluster in the contract, $s_{i,t} = \sum_{j \neq i} P_{j,t} y^{\text{act}}_{j,t} \big/ \sum_{j \neq i} P_{j,t}$. Because every deployable arm observes y^{act} for its own cluster, the proxy adds neighbour telemetry and nothing withheld; if built from y^{avail}, it leaks. It is deliberately contemporaneous. A deployed model reads current telemetry, which a day-ahead forecast cannot supply. The curtailment head of the structured decomposition reads nine features, none withheld from a deployed arm: $s_{i,t}$, the unweighted mean output of the other clusters of the same technology, the mean own output of the cluster over the preceding day, and six calendar terms. Measured forward across the split used by the arms, the nine features raise R^2 over weather for y^{curt} by 0.1344 on wind and 0.2818 on solar, and for y^{avail} by 0.0867 on wind and 0.1762 on solar. The own trailing output alone reproduces most of the latter on wind and all of it on solar.

The measurements that follow come from a controlled identifiability check that uses the covariates of the real contract but a synthetic censoring process. Its decomposition is known under independently seeded repetition, whereas that of a single historical record is not. The synthetic censoring process is calibrated against an earlier revision of the contract and censors 20.98% of held-out hours, compared with 37.96% for wind in the present revision. The check is thus a diagnostic in a comparable regime and not a replica of the record.

Paired metrics, not one average. Most held-out hours of the check have no censoring, so a metric averaged over every hour is dominated by hours where, by Equation 1, $y^{\text{act}} = y^{\text{avail}}$ exactly and recovery is trivial. Figure 4 in Appendix D shows the distortion. The two metrics order the same four arms differently, and the arm fitted directly on y^{avail} has the *worst* global MAE and the *best* censored-hour MAE. On the four hours in five with no censoring, an arm reporting y^{act} unchanged is exact by construction. For the arm fitted on y^{avail}, the irreducible error summed over that majority outweighs what it gains on the censored minority. A global average ranks the best-informed arm below the do-nothing one. Censored-hour MAE alone has the opposite blind spot. It scores only hours on which the target exceeds realised output, so an arm that corrects upward everywhere is never charged for the hours that needed no correction. The protocol therefore pairs censored-hour MAE, for reconstruction where the operator adjudicates a suppression, with uncensored-hour MAE, for false corrections where realised output already equals the adjudicated availability. Global MAE weights the two by prevalence, and Appendix K also reports their mean.

Information-set discipline. The permitted inputs of each arm are declared in code and checked, not assumed. An executable guard permutes y^{curt} and y^{avail} against a fixed y^{act} and grid state, refits the ratio estimator and asserts its estimate is bitwise unchanged. The reproducibility statement gives what the guard does and does not cover. The ratio estimator receives the scalar proxy. The structured decomposition receives a nine-feature expansion, added as one of the eliminations below and kept. Both arms see y^{act}, weather and grid state, differing only in the form of the proxy. Given the same nine features, the ratio estimator falls further behind on solar and stays ahead on wind, and adding only the six calendar terms changes neither ordering (Appendix J). The discipline is not bookkeeping. The ratio estimator can be built against y^{curt} directly or, as in Figure 1, against y^{act} alone. Only the second is deployable, since recovering y^{curt} from observables *is* the task. On the check it also has the lower censored-hour MAE, 0.0516 against 0.0814.

The policy-shift split. Train and test windows are chosen by regime, not at random. Training uses the earlier 70% of the record of each technology and testing uses the later 30%, with boundaries at 2025-01-05 on wind and 2025-10-06 on solar. Each technology is split within its own window because the records open at different dates. The two boundaries matter, since they separate very different amounts of policy. On wind the curtailed share of available energy rises from 2.64% to 15.43% across the boundary, and the intensity on censored cells rises from 0.135 to 0.364; on solar

Table 3: Recovery on the published counterfactual. Mean absolute error over censored hours, where the operator adjudicates a suppression, and over uncensored hours, where realised output already equals the adjudicated availability (lower better), over five seeds. Arms are ordered by information, not by result. The quantile ratio estimator fits the 0.75 quantile of realised output with calendar inputs, and Table 4 gives the full grid. Bold marks the lowest censored-hour MAE among the deployable arms shown. The floored decomposition is the unfloored fit with $y^{\text{avail}} \geq y^{\text{act}}$ imposed on its output. Seed standard deviations of censored-hour MAE are at most 0.0006 except for the two structured arms, at 0.0068 and 0.0086 on wind and 0.0061 and 0.0024 on solar.

		Wind		Solar	
Arm	Observes	Cens.	Uncens.	Cens.	Uncens.
Reports y^{act} unchanged	y^{act}	0.1920	0	0.1923	0
Least-squares ratio est.	+ weather, grid state	0.1400	0.0538	0.1557	0.0090
Quantile ratio est., 0.75	+ weather, grid state, calendar	**0.1388**	0.1072	**0.1237**	0.0198
Structured decomposition	+ weather, grid state	0.1865	0.2367	0.1586	0.2586
+ availability floor	+ weather, grid state	0.1713	0.1949	0.1327	0.2483
Ratio est., y^{curt}-supervised	+ y^{curt}	0.1666	0.0035	0.1335	0.0082
Fit directly on y^{avail}	weather, grid state, y^{avail}	0.1401	0.1405	0.1273	0.0319

it does not move (0.325 against 0.322). Table 2 gives what each boundary separates and Figure 1 shows its shape. Neither boundary is a stress test invented to make the benchmark harder. Censoring is a dispatch action whose rate has grown as oversupply and reliability limits bind more often (Vieira et al., 2026). Splitting by date therefore reproduces the shift a deployed model meets. Section 6 shows that drift moves the wind margin but not the gap between technologies.

6 What Works and What Does Not

The headline. Every arm is refit and rescored against the y^{avail} of the contract rather than that of a simulator, over the censored cells counted in Table 2. The structured fits in this section share one configuration, fixed before any test score is seen: hidden width 64 and sparsity weight 0.01. Appendix H shows what moves with it.

Results differ by technology. On wind the decomposition loses to the ratio estimator by 0.0464 against a paired spread of 0.0069 (6.7 standard deviations, ahead on 0 of 5 seeds). On solar the two tie at 0.0029 against 0.0060 (0.5 standard deviations, ahead on 1 of 5 seeds). The availability floor of the next paragraph widens the gap between the technologies. Wind still loses by 0.0312 (3.6 standard deviations, ahead on 0 of 5 seeds), whereas on solar the decomposition now wins by 0.0230 (9.9 standard deviations, ahead on 5 of 5 seeds) and comes within 0.0054 of the arm fitted directly on y^{avail}. Seeds 6 to 15 reproduce all four outcomes, and splits at 0.6 and 0.8 keep both signs (Appendix J).

The availability floor. Since $y^{\text{curt}} \in [0, 1]$, $y^{\text{avail}} \geq y^{\text{act}}$ holds identically. The ratio estimator enforces it by clipping its curtailment estimate at zero, whereas the sigmoid availability head of the structured decomposition does not enforce it. On a metric scored only where $y^{\text{curt}} > 0$, the identity has value but does not reflect a difference in structure. The floored decomposition imposes it, raising the fitted output to $\max(\hat{y}^{\text{avail}}, y^{\text{act}})$ without refitting. The gain is 0.0152 on wind and 0.0259 on solar. It leaves the loss to least squares on wind standing and turns the solar tie into a win.

One-sided estimation without structure. The headline ratio estimator fits the conditional mean of realised output by least squares, and curtailment pulls that mean below availability. Fitting an upper quantile (Koenker & Bassett, 1978) instead is a zero-label way to estimate what a plant could produce. Table 4 reports the ratio estimator with its first stage fitted at the quantiles 0.75, 0.9 and 0.95, a grid fixed in advance. Holding the inputs fixed isolates the loss. With the headline inputs, the least-squares fit loses to the floored decomposition on solar, whereas the fit at 0.75 beats it on censored hours on 14 of 15 seeds (0.1294 against 0.1327). With calendar terms added, the least-squares fit still loses on solar on every seed (0.1463), and the fit at 0.75 beats the decomposition on

Table 4: Censored-hour MAE of the ratio estimator with its first stage fitted at an upper quantile of realised output, over five seeds. The inputs are those of the headline, the nine features of the structured decomposition, or the headline inputs with six calendar terms added. The grid was fixed before it was run. The floored decomposition scores 0.1713 on wind and 0.1327 on solar, and the least-squares ratio estimator 0.1400 and 0.1557.

	Wind			Solar		
Inputs	0.75	0.9	0.95	0.75	0.9	0.95
Headline	0.1407	0.1701	0.3795	0.1294	0.1335	0.1519
Nine features	0.1250	0.1384	0.3795	0.1422	0.1320	0.1354
Headline and calendar	0.1388	0.1676	0.3795	0.1237	0.1352	0.1508

all 15 seeds of both technologies. Table 3 shows this cell, the best of the grid on every solar seed. With the nine features of the decomposition, both fits lose to it on solar. A validation block inside the training window, scored against its own adjudications, picks the 0.75 quantile on both technologies, with the headline inputs on solar and the nine features on wind, and on the test block these cells beat the decomposition on 14 and 15 of 15 seeds. At 0.9 the quantile fit comes within 0.003 of the decomposition on solar, and at 0.95 it loses on solar and degenerates on wind. Over all clusters, one-sided estimation with no latent factors therefore reproduces the solar gain of the decomposition over least squares. The quantile fit of Table 3 reproduces nearly all of it at the early solar split and about three fifths of a larger gain on the solar clusters live throughout. There the decomposition keeps a lead of 0.0204 on restricted hours and trails by 0.1946 on unrestricted ones (Appendix K).

What the paired metrics show. The floored decomposition corrects many hours that needed no correction. Its uncensored-hour MAE is 0.1949 on wind and 0.2483 on solar, against 0.0538 and 0.0090 for the least-squares ratio estimator, so over all hours it falls behind reporting y^{act} unchanged (0.2154 against 0.0547 on solar). The quantile ratio estimator of Table 3 corrects much less where no restriction applies, at 0.1072 on wind and 0.0198 on solar, so it is below the floored decomposition in both paired errors on every seed of both technologies. It is also below the arm fitted directly on y^{avail} in both errors, with 0.1237 against 0.1273 on solar and 0.1388 against 0.1401 on wind on censored hours. That arm never reads realised output, whereas the ratio estimator keeps it wherever it exceeds the first stage. Against the least-squares ratio estimator it lowers censored-hour error and raises uncensored-hour error on both technologies, and on wind the rise is far larger than the fall. The y^{curt}-supervised ratio estimator reads labels no deployment has and still finishes behind the least-squares one on wind and the floored decomposition on solar.

The model family and which constraints help. The structured model writes Equation 1 into the likelihood instead of applying it afterwards as a correction. Figure 3 in Appendix C shows the full architecture. Three constraints are meant to make the two factors separately identifiable: only the curtailment head reads the grid-state proxies, a sparsity prior pushes $\hat{y}^{\mathrm{curt}}$ towards zero over the majority of hours with no curtailment, and a sigmoid inside the model bounds $\hat{y}^{\mathrm{avail}}$ into $[0, 1]$. Removing the prior or the sigmoid sends every seed to a degenerate optimum near censored-hour MAE 0.40 to 0.46 on both technologies, with predictions still clipped into $[0, 1]$ for scoring. The comparison with the least-squares ratio estimator is sensitive to regularisation. A sparsity prior 10 times stronger reverses the wind result and cuts the solar margin to about one seed standard deviation, and the criterion free of the counterfactual that Appendix H tests picks different weights on the two technologies. Against the quantile ratio estimator over all clusters, no sparsity weight or width tested puts the floored decomposition ahead on solar on any seed, whereas on wind a sparsity weight of 0.1 or more puts it ahead on every seed at width 64, as does 0.1 at width 32 (Appendix K). The paper therefore makes no general claim about when structure pays. The structural asymmetry has opposite effects on wind and solar. Removing it *improves* wind (from 0.1865 to 0.1686, against a paired spread of 0.0099) and *costs* solar (from 0.1586 to 0.1972, against a paired spread of 0.0160). Seeds 6 to 15 reproduce every sign (Appendix J). The structural asymmetry helps only where the decomposition beats least squares, not where the information it withholds predicts that it should. The nine proxy features inform availability about twice as much on solar as on wind (Section 5), yet removing the structural asymmetry costs solar. Appendix G locates the effect among the features read by the availability head, with its architecture fixed.

Seven explanations and what settles each. On the check, Figure 6 in Appendix F and Table 17 give the elimination ladder. At the shared sparsity weight, non-identification does not explain the loss, since identification succeeds and the loss remains. Revealing curtailment labels raises the correlation of the curtailment estimate with the truth to 0.81 and improves censored-hour MAE. Yet with every label revealed, a 0.0180 gap of 7.7 seed standard deviations remains (Figure 5 in Appendix E). What defeats the decomposition on wind survives full curtailment supervision.

Drift and the technology split. Within a technology, the margin of the decomposition over least squares moves with how far censoring drifts between training and test. Drift does not account for the difference between the technologies (Appendix I).

7 LIMITATIONS

The counterfactual is an estimate, and it carries error. Reference generation is a productivity curve evaluated at plant anemometry or irradiance, so every number scored against it inherits that error. The curve is fitted on unrestricted hours, so applying it to restricted ones extrapolates across the shift studied here. The curve is not constant across the window. RO-AO.BR.13 Rev.08 changed the calculation on 2025-08-01, inside the wind test period (Operador Nacional do Sistema Elétrico, 2025). All arms share that uncertain target. Appendix M scales the adjudicated part of it up or down by a fifth, moves each cell across the restricted boundary with probability a tenth, splits the window at the revision and rebuilds the benchmark on the figure the operator settles on after its own verification. On both technologies the margin over least squares keeps its sign throughout. Re-scored against the perturbed targets and on both sides of the revision, the quantile arm of Table 3 stays ahead on wind throughout and on solar whenever only the size of an adjudication changes. Moving a tenth of cells across the boundary leaves it ahead on 7 of 15 solar seeds and 0.0013 ahead in the mean (Appendix K). The label is also a function of wind speed or irradiance measured at the plant, whereas the arms predict from reanalysis at a centroid. The gap between the two instruments keeps the evaluation from circularity but is not a guarantee.

One grid, and a synthetic diagnosis. Every result here rests on the record of one system operator. Section 4 names six others, none measured. Figures 4 to 6 and Table 17 diagnose the mechanism on a calibrated censoring process over real weather, for wind only. Its record-like variant in Appendix H carries only the fingerprints named there. On the record, removing the proxy cuts most of the wind margin at 0.1 while helping the arm fitted on y^{avail} about three quarters as much.

No criterion free of the counterfactual selects the quantile. Over all clusters, the grid shows that one-sided estimation reproduces the solar gain without latent structure, but on the clusters live throughout it reproduces only part of a larger one. A deployment would still have to choose a quantile without the adjudicated target, and the highest quantile degenerates on wind.

One architecture family. The result covers a recurrent encoder with two heads and a multiplicative likelihood. Other families are untested, and one setting of this family moves the result (Appendix H).

8 CONCLUSION

A grid operator bases settlements on a counterfactual and so records what the potential-outcome literature usually simulates. The record is an estimate of the suppressed outcome that no evaluator chooses. This paper turns it into a benchmark with a protocol that declares what each method reads and scores restricted and unrestricted hours separately. Over all clusters, a structured latent decomposition at a configuration fixed in advance beats a least-squares ratio estimator on restricted solar hours and loses on wind. A ratio estimator that fits the 0.75 quantile of realised output, with no latent structure, beats the decomposition on both, and on solar at every sparsity weight and width tested. On the solar clusters live throughout, the decomposition keeps a lead over the quantile fit on restricted hours. The benchmark therefore does not support a rule for when latent structure pays. It shows instead that a comparison with least squares can credit structure with a gain that one-sided estimation reproduces in whole or in part. An adjudicated counterfactual, a declared information set and paired errors on restricted and unrestricted hours expose that confound.

Reproducibility statement

Every numeric claim in this paper is traced to a file in the artefact that will be released with the benchmark, and the rule that produces each is stated where the number is used rather than left implicit. Dispersion across seeds is the sample standard deviation. On the record, seeds change the fits and not the data, so a seed standard deviation measures how far the fits vary on one fixed record, not sampling uncertainty over dates or clusters. A margin between two arms is measured against the standard deviation of the paired per-seed difference, because the two arms share seeds. Section 4 describes the construction of the contract, including all five build-time assertions and the 168-hour forward-fill threshold with its measured margin. The pipeline that implements this construction will be released with the benchmark, together with a checksummed snapshot of the directories it reads. The supplementary material carries the code, the result files and the built contract except its reanalysis, with a checksum for every contract file. An anonymous copy of the whole contract will be shared with reviewers during the discussion period. The snapshot stores the constrained-off files under the names the operator publishes, the generation under the month each file covers and the reanalysis cache under the names the fetcher gives it, so a rebuild can be checked against the same material. The assertions run on every build, so a rebuild that violates one aborts rather than producing a silently wrong contract. The recorded-mode evaluation reads the contract files `y_actual`, `y_avail` and `y_curtail`, asserts the reconstruction identity at load time and reports the split boundary and per-arm cell counts in its output. The information set each arm is permitted to read is declared in code and checked by the permutation guard of Section 5. Its coverage has two limits. The permutation guard refits the ratio estimator end-to-end but checks the structured model only at its proxy-assembly step, where a leak into that arm has to enter. It holds the proxy fixed rather than permuting it, so it tests for leakage of the withheld arrays rather than dependence on the proxy. The nine proxy features of the structured model, listed in Section 5, are the sine and cosine of hour, day of year and day of week, the mean of up to the previous 24 hourly values of own output with the current hour excluded, the unweighted mean output of the other clusters of the same technology, and $s_{i,t}$. Each is built separately on either side of the split, so no training history enters the test block and the own-output mean restarts there. Every figure is regenerated from the result files by five figure scripts that will be released with the code. The selection rule for the window of Figure 2 is stated in its caption so the panel cannot be a hand-picked episode. Seeds, the split fraction and the training configuration shared by every structured fit are recorded in the output of each run alongside its metrics. No headline structured fit exhausts its step budget. Each stops between $1{,}652$ and $2{,}977$ steps of $10{,}000$ on a training-loss plateau, and these fits have no validation split. Except for the readings in a second software environment below and the quantile readings of Appendix K under target error, under the other conditions and on the validation block, every number reported here comes from python 3.12.3, numpy 2.3.2, torch 2.8.0+cu128 and lightgbm 4.7.0 on one 32 GB consumer GPU. Those quantile readings come from python 3.13.5 and numpy 2.1.3 with the same lightgbm, and Appendix K reports how closely they reproduce the recorded runs. The runs behind Appendix M record those versions in a field of their own output, but the earlier ones predate that field. The versions matter to the digits and not to the ordering. In that environment, recorded by each run with numpy 2.5.2 and torch 2.14.0+cu130 among its packages, the arms fitted without the decomposition reproduce to within 5.8×10^{-11} and each structured fit moves by up to 0.0117 on a seed. Appendix M reports this comparison beside the rest of the sensitivity readings.

Ethics statement

This work uses only public data and no personal data of any kind: aggregate generation and curtailment at the level of a plant, which a public system operator publishes in the ordinary course of its statutory disclosure, and a public weather reanalysis. The clustering that forms the resource clusters of the benchmark aggregates plants by location and capacity and is documented in Section 4. The paper notes one dual-use consideration and judges it minor. A model that predicts curtailment can in principle inform bidding strategy in an energy market. The finding here, however, is that the most accurate deployable estimators read publicly available realised output, which confers no informational advantage over any other market participant reading the same public feed.

AI USE STATEMENT

Generative AI tools are used to implement methods, build the data pipeline, write and edit code, produce the scripts behind every figure and table, search and summarise related work, and draft and revise this manuscript. No model supplies a number. Every reported value is computed by code that will be released with the benchmark, from public data, and is traced to the file and key it comes from.

All AI-assisted work has been reviewed, and one episode is worth reporting because it bounds what such review is worth. An earlier revision builds the target by adding a field the data provider documents as a real-time dispatch limit to realised generation, treating a set-point as withheld energy. Every build assertion passes, every number traces to a file, and the cross-check run by the pipeline itself against published national curtailment figures agrees because that check is calibrated with the same wrong field. The defect is found by reading the data dictionary of the provider and testing the field against the data. A check cannot catch an error it shares. Responsibility for the final content, including text, claims and artefacts produced with AI assistance, rests with the authors.

REFERENCES

Anonymous. Title withheld to preserve anonymity, 2026. Manuscript under review at another venue; shares authorship with this submission, so its title is withheld. Available to reviewers through the programme chairs on request.

California Independent System Operator (CAISO). Renewables curtailment reporting (system-level, hourly, by fuel type). CAISO, 2026.

Mathieu Chevalley, Yusuf H. Roohani, Arash Mehrjou, Jure Leskovec, and Patrick Schwab. A large-scale benchmark for network inference from single-cell perturbation data. *Communications Biology*, 8, 2025.

David R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B*, 34(2):187–202, 1972.

Alicia Curth, David Svensson, James Weatherall, and Mihaela van der Schaar. Really doing great at estimating CATE? a critical look at ML benchmarking practices in treatment effect estimation. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021.

Peng Ding. *A First Course in Causal Inference*. Chapman and Hall/CRC, 2024. doi: 10.1201/9781003484080.

Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott, and Dan Cervone. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34(1):43–68, 2019.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *International Conference on Machine Learning (ICML)*, pp. 1097–1104, 2011.

Electric Reliability Council of Texas (ERCOT). 60-day SCED disclosure reports: Generation resource data in SCED (NP3-965-ER). ERCOT Market Information List, 2026.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640:647–653, 2025. doi: 10.1038/s41586-025-08744-2.

Jennifer L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

International Energy Agency. Renewables 2025. IEA, Paris, 2025. Section "Curtailment is rising with VRE expansion".

Fredrik D. Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International Conference on Machine Learning (ICML)*, 2016.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Roger Koenker and Gilbert Bassett. Regression quantiles. *Econometrica*, 46(1):33–50, 1978. doi: 10.2307/1913643.

Christos Louizos, Uri Shalit, Joris Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Paulo Nascimento, William Cossich, Lais Araujo, Isabela Santos, Kevin Almeida, and André Marcato. An integrated methodology for assessing wind power curtailment using anemometric measurements and operational data in the Brazilian context. *Atmosphere*, 17(4):333, 2026. doi: 10.3390/atmos17040333.

Operador Nacional do Sistema Elétrico. RO-AO.BR.13: Apuração de restrição de operação por constrained-off de usinas eólicas e fotovoltaicas. Manual de Procedimentos da Operação, Submódulo 5.13, Rev. 08, 2025. In force from 2025-08-01.

Operador Nacional do Sistema Elétrico (ONS). Geração de usinas (realised generation) and restrição de geração (constrained-off / curtailed generation) – plant-level, half-hourly public disclosure. ONS Open Data, 2026. Fields `val_geracao` and `val_geracaolimitada`, 2018–2026.

James L. Powell. Censored regression quantiles. *Journal of Econometrics*, 32(1):143–155, 1986. doi: 10.1016/0304-4076(86)90016-3.

Pablo Reyes Cerda and Kerven Cea Morales. An errata-corrected, plant-level dataset of solar, wind and hydropower curtailment in Chile's National Electricity System (2022–2026). Zenodo, data paper, version 1.1, 2026.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

Yuta Saito, Shunsuke Aihara, Megumi Matsutani, and Yusuke Narita. Open bandit dataset and pipeline: Towards realistic and reproducible off-policy evaluation. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021.

Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: Generalization bounds and algorithms. In *International Conference on Machine Learning (ICML)*, 2017.

R. Tawn and J. Browell. A review of very short-term wind and solar power forecasting. *Renewable and Sustainable Energy Reviews*, 153:111758, 2022. doi: 10.1016/j.rser.2021.111758.

James Tobin. Estimation of relationships for limited dependent variables. *Econometrica*, 26(1): 24–36, 1958.

Giovani T. T. Vieira, Walquiria N. Silva, Luís F. N. Lourenção, Renato M. Monaro, Mauricio B. C. Salles, and Carlos F. M. Almeida. Characterizing wind and solar curtailment in Brazil: An evidence-based analysis of operational drivers. *Electric Power Systems Research*, 254:112672, 2026. doi: 10.1016/j.epsr.2025.112672.

Yang Wang and Haiyun Wang. Learning deliverable wind-power capability from censored SCADA records for curtailment-aware energy management. *Electric Power Systems Research*, 265: 114140, 2027. doi: 10.1016/j.epsr.2026.114140.

Simon Wiegrebe, Philipp Kopper, Raphael Sonabend, Bernd Bischl, and Andreas Bender. Deep learning for survival analysis: a review. *Artificial Intelligence Review*, 57:65, 2024. doi: 10.1007/s10462-023-10681-3.

A THE SEARCH BEHIND THE NOVELTY CLAIM

The claim in Section 2 is negative, so the search that supports it is described rather than asserted. Four sweeps are run in 2026-09, each restricted to 2022 onward except where a benchmark predates that year: arXiv cs.LG, stat.ME and stat.AP for datasets pairing a realised outcome with an operator-supplied or separately recorded counterfactual; Semantic Scholar for such datasets, following citations from the benchmarks in Table 1; the energy-systems literature for per-unit curtailment disclosure beyond the seven systems that Section 4 names; and a cross-model adversarial review asked specifically to refute the claim with URLs.

The closest hits confirm the claim rather than refute it. Interventional screens realise both potential outcomes on different cells, sibling designs realise them on different units, and off-policy evaluation recovers the suppressed outcome in expectation. The two 2026 studies on the same ONS records (Nascimento et al., 2026; Vieira et al., 2026) estimate the withheld quantity and characterise its drivers without releasing an evaluation benchmark. The first works from SCADA anemometry at the plant. The benchmark withholds that instrument from every arm, so that study supplies a second estimate of the counterfactual rather than a baseline. The closest hit outside Brazil is a plant-level dataset of curtailment in the Chilean system, consolidated from monthly reports of its system operator (Reyes Cerda & Cea Morales, 2026). It records curtailed energy per plant and hour, but it is not an evaluation benchmark, and whether that energy is an adjudicated estimate of what the plants could have produced is not established.

The search cannot establish that no such benchmark exists, only that its sweeps find none, and the claim is scoped accordingly.

B THE DECOMPOSITION ON THE RECORD

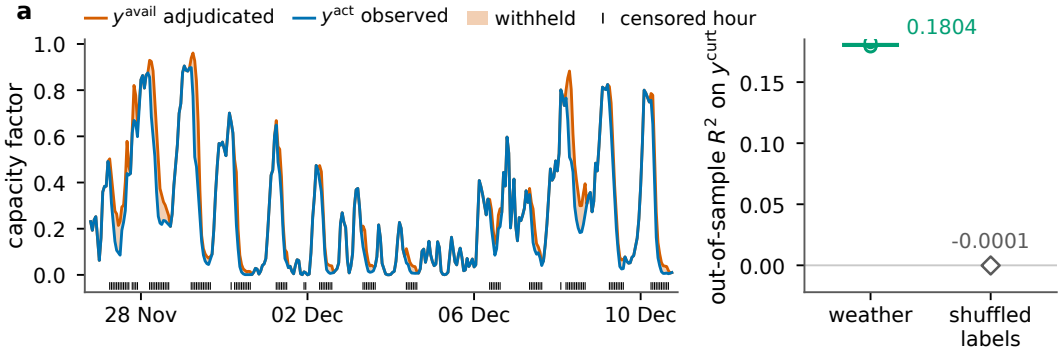


Figure 2: The censoring is published per plant and is not a deterministic function of weather. **(a)** The window shown in miniature in Figure 1, here at full size, covers 14 days of one wind cluster in the held-out split. The rug marks the hours under a dispatch instruction. Both the cluster and the window are chosen by rule rather than by eye. The cluster has the median censored share, and the window has a censored share of 39.0%, which is the closest to the median of 38.6% over the split. The script that applies this rule will be released with the code. **(b)** Five-fold cross-validated R^2 of a gradient-boosted regression of y^{curt} on weather, and of the same regression on shuffled labels (open circles show folds and the bar shows the mean). Random folds on an autocorrelated panel overstate predictability, so the R^2 of 0.1804 is an upper bound.

C THE MODEL UNDER TEST

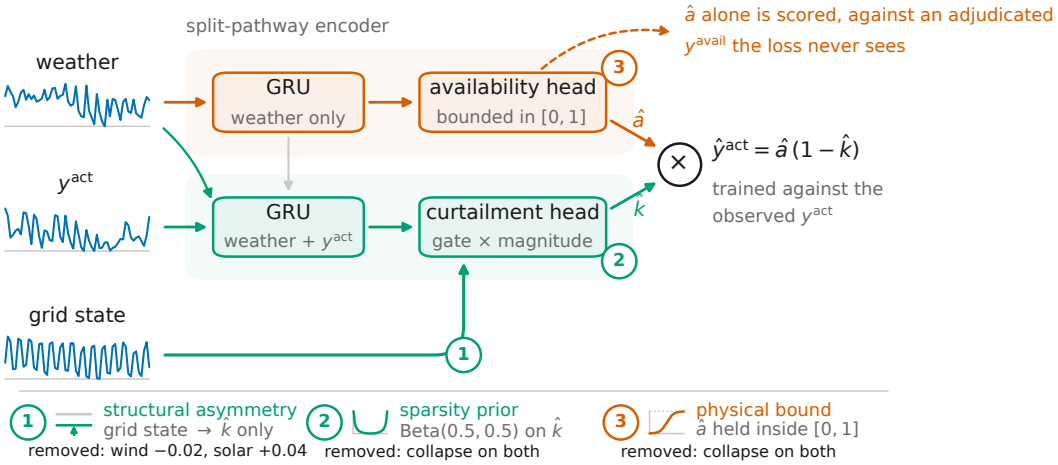


Figure 3: The structured decomposition of Section 6, in the configuration the paper reports. Sparklines show the three inputs as observed over the same 14 days as Figure 2, each on its own scale. Colour marks the pathway rather than the quantity, so weather appears in two colours because it feeds both pathways. Badges mark the three constraints ablated in Section 6, at the points where they act. Solid arrows carry gradient. The dashed arrow carries none and is the one that matters. The reconstruction term uses only the product, whereas the result is scored on $\hat{a}$ alone against a published y^{avail} that training never observes.

D THE METRIC INVERSION

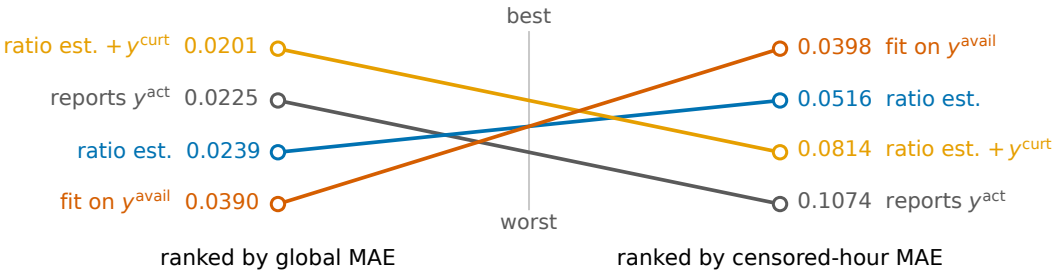


Figure 4: The metric inversion in the identifiability check of Section 5. The same four arms are ranked by global MAE and censored-hour MAE, averaged over five seeds. Every line crosses. The uncensored majority does not merely compress the order but reverses it.

E THE SUPERVISION SWEEP

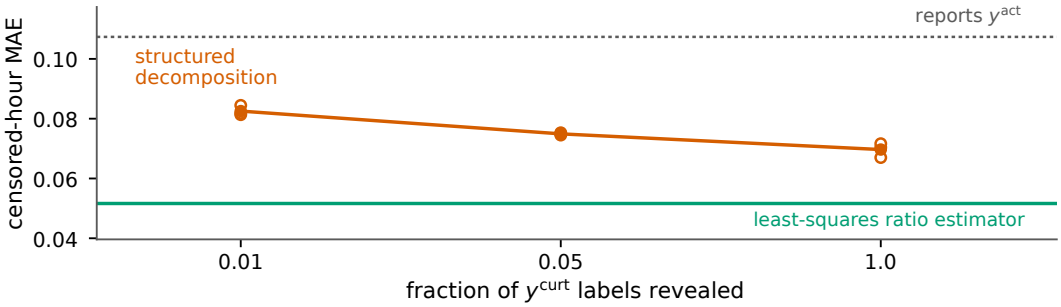


Figure 5: What curtailment labels add under the shift. Censored-hour MAE of the structured decomposition against the fraction of training y^{curt} labels revealed in the identifiability check of Section 5 (open circles show three seeds and filled circles show their mean). Fractions are plotted equally spaced, not to scale. Error falls monotonically but never reaches the line for the least-squares ratio estimator, including at a fraction of 1.0.

F THE ELIMINATION LADDER

The rungs of Table 17 are measured on a censoring process calibrated to an earlier revision of the contract, which realises less censoring on the held-out side than the present one does. When the process is refitted so that censored share, event size and the drift between the two sides all match the revision used throughout this paper, the comparison gives the same result as on the ladder. The structured decomposition beats the ratio estimator (0.0494) on 0 of 5 seeds. Two seeds converge normally at censored-hour MAE 0.0857 and 0.0897, whereas the other three sit at a global MAE between 0.7155 and 0.7167 with a correlation of at most 0.22 to the truth. The values for these three seeds are consistent with predicting availability at the ceiling of its range, which costs one minus the mean availability of the series. No constraint excludes that solution, since it lies on the imposed boundary rather than outside it. Revealing curtailment labels removes the degeneracy without changing the verdict. Across three seeds at each of three fractions, every fit converges normally. With every training label revealed at $p = 1.0$, the decomposition reaches 0.0811, whereas the ratio estimator reaches 0.0494. No fit at any fraction beats the ratio estimator (0 of 9 fits). The refit is not a controlled comparison with the earlier rungs, which run on a longer record with one more cluster, so the change in margin is not attributed here.

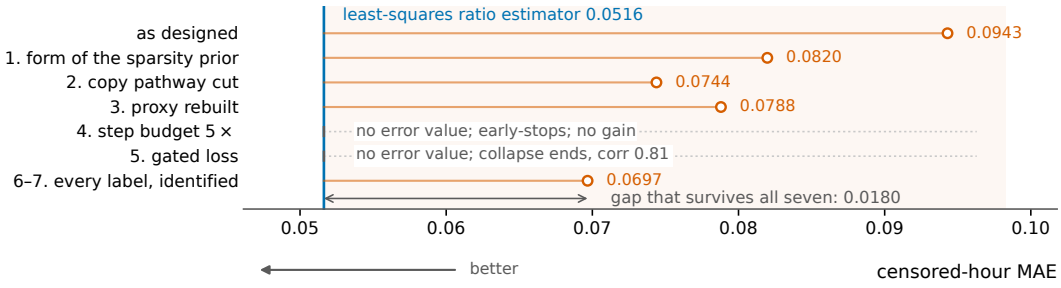


Figure 6: Seven explanations for why the structured decomposition loses, and the gap that survives all of them. Censored-hour MAE after each elimination, compared with the least-squares ratio estimator on the same seeds and split. Rungs 4 and 5 are closed by a step count and a correlation rather than by error, so they have no point on this axis. Rungs 6 and 7 share one set of fits, read in two ways.

G WHICH WITHHELD FEATURES CARRY THE ABLATION

Removing the structural asymmetry gives the availability head all nine proxy features at once. Table 5 reports this ablation repeated with the head reading no group, the whole block, one group at a time or every group but one. The head keeps one architecture in every row, and the columns that it may not read are set to zero. All rows therefore start from the same weights and differ only in what the head can read. Narrowing the head instead would change the initial weights of every layer built after it. The whole-block row is the no-asymmetry ablation of the main text, seed for seed. Against the no-access row, it still improves wind by 0.0180 at 4.7 seed standard deviations and costs solar 0.0412 at 5.1 seed standard deviations. The gain and the cost each hold on all five seeds, so the effect is not an artefact of initialisation. On wind the gain is carried by the own trailing output of the cluster. Alone it gives 0.0205 at 2.8 seed standard deviations on all five seeds, whereas the rest of the block without it gives 0.0020 on average, with one seed losing. On solar, every group except the calendar terms costs on its own on all five seeds, and one fit for the own trailing output collapses. Whichever group is removed, a cost remains on all five seeds and averages between 0.0332 and 0.0427, so no group is needed for the cost. Seeds 6 to 15 reproduce the wind gain from the own trailing output and the solar cost of the whole block and of every block missing one group (Appendix J). Against the marginal measurements of Section 5, the groups that cost solar inform its censoring about as much as or more than its availability, the calendar terms inform little of either and cost little, and the group that carries the wind gain informs wind availability more than wind censoring. The pattern fits the account of Section 6, in which the structural asymmetry helps only where the decomposition beats least squares. The evidence is an ablation read against marginal measurements, not a manipulation of the mechanism.

Table 5: Censored-hour MAE over five seeds with the availability head given no group of the grid-state proxy, the whole block, one group or every group but one. The head keeps one architecture in every row, and the columns that it may not read are set to zero, so all rows start from the same weights. Differences are paired against the no-access row, with the sample standard deviation of the paired per-seed difference in parentheses. The curtailment head reads the whole block in every row.

	Wind			Solar		
Availability head reads	MAE	Difference (SD)	Improved	MAE	Difference (SD)	Improved
No group (baseline)	0.1866			0.1560		
The whole block	0.1686	−0.0180 (0.0038)	5/5	0.1972	+0.0412 (0.0080)	0/5
One group only						
Calendar terms	0.1883	+0.0016 (0.0123)	2/5	0.1616	+0.0056 (0.0060)	1/5
Own trailing output	0.1661	−0.0205 (0.0074)	5/5	0.2078	+0.0518 (0.0855)	0/5
Other-cluster aggregate	0.1777	−0.0089 (0.0090)	4/5	0.1843	+0.0282 (0.0121)	0/5
Base scalar $s_{i,t}$	0.1774	−0.0092 (0.0063)	5/5	0.1866	+0.0305 (0.0152)	0/5
All groups except						
Calendar terms	0.1670	−0.0196 (0.0051)	5/5	0.1987	+0.0427 (0.0061)	0/5
Own trailing output	0.1847	−0.0020 (0.0188)	4/5	0.1911	+0.0351 (0.0080)	0/5
Other-cluster aggregate	0.1639	−0.0227 (0.0036)	5/5	0.1893	+0.0332 (0.0084)	0/5
Base scalar $s_{i,t}$	0.1593	−0.0274 (0.0055)	5/5	0.1937	+0.0377 (0.0075)	0/5

H THE STRENGTH OF THE SPARSITY PRIOR

Every structured fit in the body uses one shared configuration with hidden width 64 and a sparsity weight of 0.01 on the prior that pushes $\hat{y}^{\mathrm{curt}}$ towards zero, and this configuration is fixed before any test score is seen. Table 6 reports the whole headline grid at six other weights on seeds 1 to 5 and 6 to 10. The floored decomposition turns from a loss into a win on wind between 0.05 and 0.1 and wins on every seed from there to 1.0. On solar, it wins clearly at 0.01 and 0.03 and falls into an unstable near-tie from 0.05 onward. The unfloored decomposition loses on wind at every weight. At a weight of 0.001, the prior is too weak to constrain the fit, and the floored decomposition loses on every seed of both technologies by margins between 0.0922 and 0.3145. Removing the prior produces censored-hour MAE in the range 0.40 to 0.46 (Section 6), and 16 of the 20 fits at 0.001 sit in this range. The other 4 are all on solar seeds 6 to 10 and sit between 0.2480 and 0.3921. Hidden

width does not move the technology split at the shared weight. At widths 32 and 128, wind still loses and solar still wins on every seed of both sets. On wind, the margins are $+0.0593$ and $+0.0486$ at width 32 and $+0.0218$ and $+0.0191$ at width 128. On solar, they are -0.0257 and -0.0262 at width 32 and -0.0247 and -0.0230 at width 128. On wind at a weight of 0.1, width 32 keeps the win on every seed of both sets, with margins of -0.0126 and -0.0112, whereas at width 128 the floored margin is negative on 8 of 10 seeds. Yet the other two seeds lose by $+0.0392$ and $+0.0390$, so the means are $+0.0010$ and -0.0001. On solar at width 128, the means are -0.0028 and -0.0051, and the floored margin is negative on 7 of 10 seeds. The win on every seed at 0.1 therefore holds at widths 32 and 64 but not at 128.

Table 6: Floored decomposition minus least-squares ratio estimator in censored-hour MAE as the sparsity weight varies, with all other settings as in the headline fit. Each entry gives the mean paired difference, followed in parentheses by its absolute value divided by the sample standard deviation, and then the seeds with a negative difference.

	Wind		Solar	
Sparsity weight	Seeds 1 to 5	Seeds 6 to 10	Seeds 1 to 5	Seeds 6 to 10
0.001	$+0.3115$ (62.7) 0/5	$+0.3108$ (82.5) 0/5	$+0.2478$ (551.4) 0/5	$+0.1919$ (2.9) 0/5
0.01 (shared)	$+0.0312$ (3.6) 0/5	$+0.0270$ (2.5) 0/5	-0.0230 (9.9) 5/5	-0.0236 (17.5) 5/5
0.03	$+0.0196$ (1.2) 1/5	$+0.0088$ (0.6) 2/5	-0.0174 (2.5) 5/5	-0.0277 (9.4) 5/5
0.05	$+0.0188$ (0.7) 2/5	$+0.0213$ (1.2) 1/5	-0.0039 (0.3) 3/5	-0.0107 (0.5) 4/5
0.1	-0.0124 (3.1) 5/5	-0.0126 (4.5) 5/5	-0.0066 (1.0) 4/5	-0.0116 (1.2) 4/5
0.3	-0.0129 (5.5) 5/5	-0.0104 (3.2) 5/5	$+0.0562$ (1.1) 1/5	-0.0104 (1.6) 5/5
1.0	-0.0112 (9.5) 5/5	-0.0125 (4.9) 5/5	$+0.0052$ (0.5) 2/5	$+0.0033$ (0.1) 3/5

The result of the drift analysis in Section 6 also changes with the prior (Table 7). At a weight of 0.1, the floored decomposition is ahead of the ratio estimator in every condition of Figure 7 on both seed sets. The wind margin barely changes with drift. On seeds 1 to 5 it is -0.0124 on the full window and -0.0136 on the levelled one, and on seeds 6 to 10 it is -0.0126 and -0.0131. By comparison, the same margins at the shared prior on seeds 1 to 5 are $+0.0312$ and -0.0001. Within solar, the early split still wins by more than the default one, with margins of -0.0152 against -0.0066 and of -0.0173 against -0.0116. On the solar window (from 2024-04-01), where wind loses on every seed at the shared prior, the wind margin at 0.1 is -0.0028 and -0.0061 and is below zero on 9 of 10 seeds. At the shared prior, the wind loss follows the drift, whereas at 0.1 the decomposition wins on both the full and the levelled windows by about the same margin. The transport account therefore describes the loss at the shared prior rather than a cost paid at every prior strength.

Table 7: The rows are the conditions of Figure 7 together with wind restricted to the solar window, the other split positions and solar clusters that are live throughout. Each condition is run at the shared sparsity weight and at 0.1, with all other settings as in the headline fit. The headline fit uses the full window and the default split at 0.7 of each record. The split rows move this split to 0.6 or 0.8. Floored decomposition minus least-squares ratio estimator in censored-hour MAE, with entries as in Table 6. The drift is $2.69\times$ on the full wind window, $1.28\times$ on the levelled one, $0.99\times$ on the default solar split and $1.42\times$ on the early one.

	Weight 0.01 (shared)		Weight 0.1	
Condition	Seeds 1 to 5	Seeds 6 to 10	Seeds 1 to 5	Seeds 6 to 10
Wind				
Full window	$+0.0312$ (3.6) 0/5	$+0.0270$ (2.5) 0/5	-0.0124 (3.1) 5/5	-0.0126 (4.5) 5/5
Levelled window	-0.0001 (0.0) 3/5	-0.0005 (0.1) 2/5	-0.0136 (4.4) 5/5	-0.0131 (2.6) 5/5
Solar window	$+0.0133$ (2.9) 0/5	$+0.0148$ (2.1) 0/5	-0.0028 (0.8) 4/5	-0.0061 (2.3) 5/5
Split 0.6	$+0.0448$ (6.4) 0/5	$+0.0487$ (5.3) 0/5	-0.0011 (0.2) 3/5	$+0.0106$ (0.6) 1/5
Split 0.8	$+0.0218$ (2.0) 0/5	$+0.0443$ (1.5) 0/5	-0.0105 (2.3) 5/5	-0.0085 (1.1) 4/5
Solar				
Default split	-0.0230 (9.9) 5/5	-0.0236 (17.5) 5/5	-0.0066 (1.0) 4/5	-0.0116 (1.2) 4/5
Early split	-0.0331 (13.6) 5/5	-0.0309 (12.4) 5/5	-0.0152 (2.6) 5/5	-0.0173 (1.7) 5/5
Split 0.6	-0.0284 (8.6) 5/5	-0.0313 (10.4) 5/5	$+0.0074$ (0.1) 3/5	-0.0186 (2.8) 5/5
Split 0.8	-0.0251 (13.2) 5/5	-0.0268 (7.2) 5/5	-0.0031 (0.6) 3/5	-0.0065 (1.1) 4/5
Live clusters	-0.0525 (9.3) 5/5	-0.0502 (28.1) 5/5	-0.0317 (4.7) 5/5	-0.0310 (3.1) 5/5

Table 8 reports the two within-technology comparisons across the whole sweep. The wind gap ranges between $+0.0101$ and $+0.0313$ at weights up to 0.05 and between -0.0002 and $+0.0042$ at weights from 0.1 on. The dependence on drift therefore becomes small between 0.05 and 0.1, where the wind loss itself turns into a win (Table 6). The solar gap is negative at every weight on both seed sets. For solar, the higher-drift split is always better for the decomposition.

Table 8: The two within-technology comparisons of Figure 7 across the sparsity weight, with all other settings as in the headline fit. Each entry is the mean floored margin in the higher-drift condition minus that in the lower-drift one. On wind the entry is the full window ($2.69\times$) minus the levelled one ($1.28\times$), and on solar it is the early split ($1.42\times$) minus the default one ($0.99\times$). The transport account predicts a positive entry.

	Wind		Solar	
Sparsity weight	Seeds 1 to 5	Seeds 6 to 10	Seeds 1 to 5	Seeds 6 to 10
0.01 (shared)	$+0.0313$	$+0.0274$	-0.0101	-0.0073
0.03	$+0.0160$	$+0.0101$	-0.0177	-0.0049
0.05	$+0.0140$	$+0.0284$	-0.0215	-0.0155
0.1	$+0.0012$	$+0.0005$	-0.0087	-0.0057
0.3	$+0.0014$	$+0.0033$	-0.0775	-0.0117
1.0	$+0.0042$	-0.0002	-0.0060	-0.0064

The split position and cluster set also affect the stronger-prior result. At 0.1, the floored decomposition wins on wind on 9 of 10 seeds with the split at 0.8, where the margin is -0.0105 and -0.0085. By contrast, it does not win with the split at 0.6, where the margin is -0.0011 and $+0.0106$ with 4 of 10 seeds below zero. Over all solar clusters, the stronger prior shrinks the win at every split position, from between -0.0313 and -0.0230 at the shared prior to between -0.0186 and $+0.0074$. At the 0.6 split on seeds 1 to 5 the mean crosses zero. Over the three split positions and both seed sets at 0.1, 23 of the 30 fits stay below zero. Over the solar clusters that are live throughout the record, the floored decomposition still wins on every seed, with margins of -0.0317 and -0.0310 compared with -0.0525 and -0.0502 at the shared prior. It wins there on every seed in both sets at every weight from 0.01 to 0.3, by between 0.0256 and 0.0525. It ties only at 1.0, with margins of $+0.0065$ and $+0.0003$. The prior therefore moves the ordering the same way under every setting tested, towards the decomposition on wind and away from it on solar. However, whether wind wins depends on the width and split.

Table 9 reports what a criterion free of the counterfactual selects. On uncurtailed hours, realised output equals availability, so error there needs no counterfactual value. Here, uncurtailed hours are identified from the curtailment labels of the benchmark, which the deployable arms never see. The table therefore diagnoses that criterion, whereas using it in deployment instead requires the restriction flag of the operator, a mask that need not coincide with the curtailment labels. On wind, the criterion is lowest at 0.1, where censored-hour error is also near its lowest. On solar, it is lowest at 0.05, whereas censored-hour error is lowest at 0.01. At 0.05 the floored margin keeps 17% of the win it has at the shared weight. The inversion between error on uncurtailed hours and censored-hour error is the one that Section 5 documents between global and censored-hour metrics, and here it decides a hyperparameter.

The synthetic identifiability check behind the elimination ladder is refitted to this revision of the contract (Appendix F). On this check, the decomposition does not overtake the ratio estimator at any tested prior. It trails on every seed at 0.01, 0.1, 0.3 and 1.0, both on seeds 1 to 5 and on seeds 6 to 10. Each stronger prior makes 6 to 9 of 10 fits converge, more than the 3 that converge under the shared one. The converged floored fits lie between 0.0660 and 0.1023, against 0.0493 to 0.0498 for the ratio estimator. The calibrated process differs from the record in the shape of the censoring. Of its censored cells, 1.1% carry a curtailed fraction below 0.05 and none reaches 0.6, whereas the corresponding shares on the recorded wind series are 37.8% and 15.4%. The recorded shape is imposed by mapping the censored magnitudes on each side of the split onto the recorded quantiles, and nothing else about the series changes. This mapping leaves the verdict unchanged. The decomposition still trails the ratio estimator on all 40 fits, though the median gap of a converged fit at the stronger priors falls from 0.0238 to 0.0048. The shape narrows the gap without closing it.

Table 9: Choosing the sparsity weight with and without the counterfactual. Censored-hour MAE of the floored decomposition against its MAE on hours the benchmark marks uncurtailed. Censored-hour MAE needs the benchmark labels, whereas on uncurtailed hours the realised output is the availability. Seeds 1 to 5. On uncensored hours the ratio estimator scores 0.0538 on wind and 0.0090 on solar.

	Wind		Solar	
Sparsity weight	Censored	Uncensored	Censored	Uncensored
0.001	0.4515	0.5544	0.4035	0.9142
0.01 (shared)	0.1713	0.1949	0.1327	0.2483
0.03	0.1596	0.1660	0.1383	0.2675
0.05	0.1588	0.1575	0.1518	0.0670
0.1	0.1277	0.0783	0.1491	0.0881
0.3	0.1271	0.0844	0.2119	0.1952
1.0	0.1289	0.0902	0.1609	0.1130

When the synthetic check is made like the record in four measured respects, it changes with the prior as the record does (Table 10). Besides the recorded shape, the process then carries availability noise sized so that the features explain 60% of availability variance on held-out hours. The configuration as a whole also leaves this share at 60%, against 66% on the record and 94% in the calibrated process. The process also carries a curtailment driver shared across clusters, which lifts the correlation of the curtailment indicator between clusters from 0.12 to 0.32, against 0.49 on the record. Like the recorded proxy, the grid-state proxy of the process is built from the realised output of the other clusters, but it is unweighted whereas the record weights by capacity. The calibrated process instead observes the latent state that drives its curtailment. Mean availability stays unmatched, at 0.29 against 0.47 on the record. The process has the properties named here and not the joint distribution of the record. The floored decomposition then trails on every seed at the shared weight, whereas at 0.1 every fit that converges leads. It leads in 5 of 5 fits on seeds 1 to 5, by a median of 0.0065, and in 4 of 5 on seeds 6 to 10, by a median of 0.0047. The fits that fail to converge are counted rather than dropped, and the one on seeds 6 to 10 leaves the mean over all five seeds at $+0.0726$ against that median. With autocorrelated noise, the three fits that converge at 0.1 on seeds 1 to 5 lead as well, and with independent noise the two that converge at each of 0.3 and 1.0 also lead.

Table 10: Synthetic check in which the calibrated process is made like the recorded wind series, for seeds 1 to 5 unless stated. The columns for the ratio estimator and the oracle give censored-hour MAE averaged over seeds, and the column for the floored decomposition at a sparsity weight of 0.1 gives it averaged over the fits that converge. At each weight, an entry lists the seeds on which the floored decomposition scores below the ratio estimator. It then gives the median difference of the floored decomposition from the ratio estimator over the fits that converge, with the number of those fits in parentheses. A fit converges when its global MAE is below 0.1. With availability noise the threshold is 0.3, because that noise raises the global MAE of a regular fit. The second threshold, set after the runs with noise, changes no count at 0.1. Rows 2 to 4 each add to the row above. The indented rows repeat row 4 on seeds 6 to 10 or with autocorrelated noise. Row 7 keeps only the two ingredients needed for the win at 0.1. The last four rows each remove one ingredient of row 4 and are reported at 0.1 only.

Process	Ratio estimator	Oracle	Floored, 0.1	Weight 0.01	Weight 0.1
Calibrated	0.0494	0.0403	0.0750	0/5, $+0.0358$ (2)	0/5, $+0.0260$ (4)
+ recorded shape	0.0641	0.0403	0.0697	0/5, $+0.0150$ (1)	0/5, $+0.0059$ (4)
+ availability noise	0.1100	0.1220	0.1173	0/5, $+0.0258$ (4)	0/5, $+0.0072$ (5)
+ shared driver, output proxy	0.1229	0.1221	0.1170	0/5, $+0.0161$ (4)	5/5, -0.0065 (5)
seeds 6 to 10	0.1237	0.1218	0.1184	0/5, $+0.0111$ (4)	4/5, -0.0047 (4)
autocorrelated noise	0.1230	0.1224	0.1177	0/5, $+0.0125$ (4)	3/5, -0.0051 (3)
Shape and output proxy only	0.0799	0.0403	0.0681	3/5, -0.0072 (3)	5/5, -0.0127 (5)
Without availability noise	0.0858	0.0403	0.0724		5/5, -0.0158 (5)
Without the recorded shape	0.1090	0.1221	0.1290		0/5, $+0.0202$ (4)
Without the shared driver	0.1187	0.1220	0.1170		5/5, -0.0019 (5)
Latent proxy	0.1099	0.1220	0.1173		0/5, $+0.0069$ (5)

Removing ingredients separates the two parts of the pattern. The win at 0.1 needs the output-built proxy and the recorded shape. When the latent state is observed in place of the proxy, or when the calibrated shape is used, the decomposition trails on every seed, both on seeds 1 to 5 and on seeds 6 to 10. The loss at the shared weight needs the noise or the shared driver. With the recorded shape and the output-built proxy alone, the decomposition already leads at the shared weight on the 3 of 5 fits that converge there, and without the noise the win at 0.1 is wider, with a median of 0.0158. Both results hold on seeds 6 to 10. The win without the noise is wider there, and with the two ingredients alone the decomposition leads on 3 of 5 seeds at the shared weight and on every seed at 0.1. On each seed set the other two fits fail at the shared weight. With them counted, the means are $+0.1974$ on seeds 1 to 5 and $+0.1980$ on seeds 6 to 10, so the lead applies to the fits that converge. The shared driver widens the margin at 0.1 rather than deciding its sign. The median margin is 0.0019 without the driver and 0.0065 with it. On seeds 6 to 10, the decomposition without the driver leads by a median of 0.0004 on 3 of the 4 fits that converge. With all four ingredients the process both loses at the shared weight and wins at 0.1, as the record does.

In the synthetic crossing of the proxy with the shared driver, the proxy affects the ratio estimator but leaves the decomposition unchanged. With the recorded shape and the noise, the floored decomposition at 0.1 averages between 0.1170 and 0.1173 in all four combinations (Table 10, rows 3 and 4 and the last two). The ratio estimator averages 0.1100 and 0.1099 with the latent state, 0.1187 with the output-built proxy alone and 0.1229 with the proxy and the shared driver together. The oracle is the same regressor trained on availability itself rather than on realised output, and it scores between 0.1218 and 0.1223 in all four. The oracle uses no realised output, whereas the ratio estimator keeps the realised output wherever that output exceeds the first-stage regression. The ratio estimator can therefore score below the oracle, and with the latent state it does. The output-built proxy removes most of that margin and the shared driver removes the remainder, but the driver acts through the proxy. Under the latent state, the driver moves the ratio estimator from 0.1100 to 0.1099. Without the noise, the same proxy costs the ratio estimator more, raising its average from 0.0641 to 0.0799.

On the record, replacing the scalar proxy with a constant in every arm that reads it produces the same asymmetry at 0.1 on all three seed sets. This removes the capacity-weighted proxy of Section 5. The decomposition keeps the unweighted mean of realised output across the other clusters, a mean that its own head builds. Without the proxy, the ratio estimator improves by 0.0084, 0.0086 and 0.0084, and the oracle improves by 0.0063, 0.0062 and 0.0065. The floored decomposition moves by $+0.0018$, $+0.0016$ and $+0.0017$ at 0.1 and by -0.0058, $+0.0036$ and $+0.0098$ at the shared weight. The ordering at the shared weight does not reverse. The decomposition trails by $+0.0337$, $+0.0392$ and $+0.0463$, compared with $+0.0312$, $+0.0270$ and $+0.0281$ with the proxy. At 0.1, the remaining win is -0.0023, -0.0024 and -0.0005, against -0.0124, -0.0126 and -0.0107 with the proxy. It is below zero for every seed of the first set and for 3 of 5 seeds in each of the other two sets. The proxy carries most of the margin at 0.1 on the record, but the remainder does not hold seed by seed. The oracle improves by about three quarters as much as the ratio estimator when the scalar proxy is removed, so the removal does not separate a proxy that curtailment depresses from a feature whose removal helps any regression fitted here.

The contrast between the record and the calibrated process is therefore consistent with how the grid state is observed and with the shape of the curtailment. A proxy that the curtailment depresses costs the ratio estimator its advantage, and the decomposition with a stronger sparsity prior can then overtake it. On the record, the proxy carries most of the margin at 0.1, and the intervention leaves the rest unexplained. The synthetic check in Appendix F gives the same verdict at the shared prior in both processes. The ladder itself is not rerun on the record-like process, so its conclusion is limited to its own process and the shared prior.

I Drift within a technology

This appendix compares the floored decomposition with the least-squares ratio estimator. An arm predicting a censoring magnitude learns it in training and applies it in test, so its error should grow with how far censoring moves. That distance differs by technology. Over censored cells the mean curtailed fraction rises from 0.135 to 0.364 on wind but not on solar (0.325 against 0.322). A regressor fitted on y^{curt} undershoots the held-out censored mean by $9.1\times$ on wind and $2.0\times$ on solar.

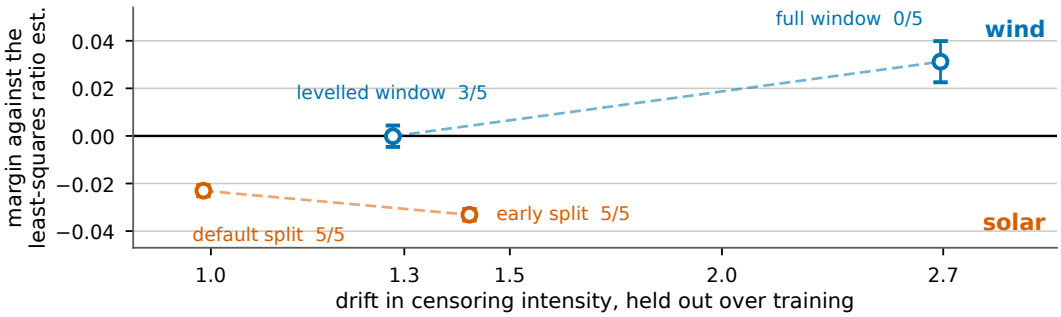


Figure 7: The drift manipulation, run in both directions. Drift is the mean curtailed fraction over held-out censored cells divided by the same quantity in training. The margin is the floored decomposition minus the least-squares ratio estimator on censored-hour MAE, so a point below the rule is a win for the decomposition. Bars give the seed standard deviation over five seeds, and the fraction beside each label is the proportion of seeds the decomposition wins. The manipulation changes the drift but not the technology. It restricts wind to the window where its censoring has levelled off and splits solar early so that its censoring does move. At the shared prior, the margin follows the drift within wind, as the transport account predicts; within solar it moves the other way. Between the technologies, the account fails. Solar wins at drift $1.42\times$ whereas wind only ties at $1.28\times$.

The transport account predicts the effect of a manipulation, run here in both directions. Figure 7 reports four conditions ordered by drift, each reproduced on seeds 6 to 15 (Appendix J). Within wind, the prediction holds. Restricting wind to the window from 2024-01-01, where censoring has levelled off, changes the drift from $2.69\times$ to $1.28\times$, leaves the technology unchanged and turns a loss of 0.0312 (0 of 5 seeds ahead) into a tie of -0.0001 (3 of 5 ahead). Across the technologies, the prediction fails. Splitting solar early, at a training fraction of 0.35, so that its censoring does move (drift $1.42\times$) keeps the win at -0.0331 (13.6 seed standard deviations). Solar wins at every drift level tested and wind at best ties, so drift changes the margin within a technology without accounting for the difference between them.

One observation may bear on the difference between the technologies. From the calendar alone, the out-of-sample R^2 of availability is 0.77 on solar and -0.13 on wind (cluster medians, both on the solar window). At the 0.75 quantile, calendar inputs also improve the quantile ratio estimator more on solar than on wind (Table 4). With two technologies, this is an association and not an explanation.

J Further seeds and controls

Figure 8 shows the seed spread behind the censored-hour columns of Table 3.

Except for the prior-sensitivity analyses above, which report seeds 1 to 5 and 6 to 10, the five-seed results come from seeds 1 to 5. Table 11 reports the headline and the listed diagnostics, rerun on seeds 6 to 10 and 11 to 15. The comparison with least squares reproduces on both further sets. On wind the decomposition loses to the ratio estimator on every seed, floored or not. On solar the floored decomposition wins on every seed, whereas the unfloored one again ties. The constraint ablations also reproduce. Lifting the structural asymmetry helps wind and costs solar on every seed in both sets. Removing the prior or the sigmoid again sends every fit to the degenerate optimum. With the head architecture fixed, every variant carrying the own trailing output improves wind on every seed in both sets, and the block without it gives no reliable gain. The whole block and every block missing one group cost solar on every seed, and over the three seed sets their costs lie between 0.0282 and 0.0459. On seeds 1 to 5 the aggregate and the scalar alone give wind small gains, but the gains do not recur. The conditions of Figure 7 and the live-cluster check of Appendix L also reproduce. With the floored decomposition, wind ties on its levelled window and loses on every seed on the solar window, and solar wins on every seed on the early split. On solar clusters that are live throughout, the decomposition wins on every seed, floored or not. On wind the live-cluster check is the headline itself because every wind cluster is live.

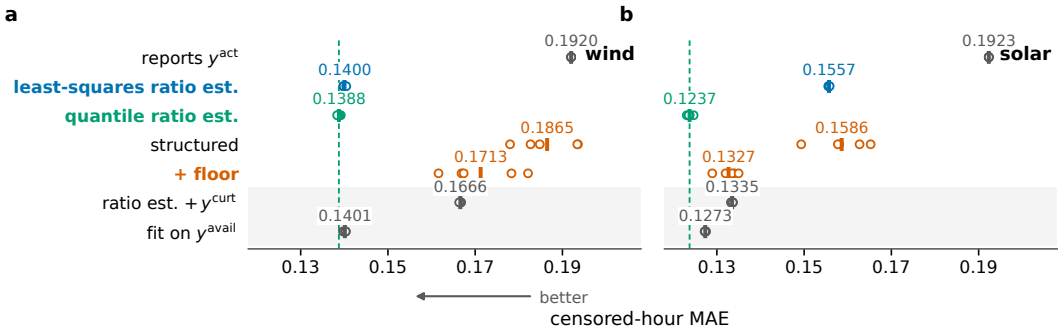


Figure 8: Censored-hour MAE of the arms of Table 3 against the published y^{avail} on **(a)** wind and **(b)** solar, by seed. Each bar is the mean over five seeds (open circles). The structured decomposition (vermilion) loses to the least-squares ratio estimator (blue) on wind but, with the floor, beats it on solar. The quantile ratio estimator (green) beats both on this metric on every seed, and the dashed line marks its mean. Shaded rows use quantities unavailable to a deployment and are diagnostics, not baselines.

Table 11: The five-seed results on seeds 1 to 5, 6 to 10 and 11 to 15. Each entry is the difference in censored-hour MAE from the reference named in its heading. It is followed in parentheses by its absolute value divided by the sample standard deviation of the paired per-seed difference, and then by the seeds on which it is negative.

	Seeds 1 to 5	Seeds 6 to 10	Seeds 11 to 15
Headline, decomposition minus ratio estimator			
Wind	+0.0464 (6.7) 0/5	+0.0418 (4.9) 0/5	+0.0439 (6.6) 0/5
Wind, floored	+0.0312 (3.6) 0/5	+0.0270 (2.5) 0/5	+0.0281 (4.3) 0/5
Solar	+0.0029 (0.5) 1/5	+0.0004 (0.2) 2/5	+0.0035 (0.7) 1/5
Solar, floored	−0.0230 (9.9) 5/5	−0.0236 (17.5) 5/5	−0.0227 (11.2) 5/5
Structural asymmetry lifted, against the constrained fit			
Wind	−0.0179 (1.8) 5/5	−0.0183 (2.1) 5/5	−0.0198 (2.7) 5/5
Solar	+0.0386 (2.4) 0/5	+0.0390 (7.0) 0/5	+0.0369 (7.8) 0/5
Head access with the architecture fixed, against no access			
Wind, whole block	−0.0180 (4.7) 5/5	−0.0162 (1.1) 5/5	−0.0269 (1.2) 5/5
Wind, own trailing output only	−0.0205 (2.8) 5/5	−0.0178 (1.9) 5/5	−0.0292 (1.8) 5/5
Wind, all but own trailing output	−0.0020 (0.1) 4/5	+0.0016 (0.1) 1/5	−0.0069 (0.4) 3/5
Solar, whole block	+0.0412 (5.1) 0/5	+0.0349 (4.2) 0/5	+0.0459 (6.1) 0/5
Conditions, floored decomposition minus ratio estimator			
Wind, levelled window	−0.0001 (0.0) 3/5	−0.0005 (0.1) 2/5	−0.0012 (0.2) 3/5
Wind, solar window	+0.0133 (2.9) 0/5	+0.0148 (2.1) 0/5	+0.0108 (2.2) 0/5
Solar, early split	−0.0331 (13.6) 5/5	−0.0309 (12.4) 5/5	−0.0386 (10.0) 5/5
Solar, live clusters	−0.0525 (9.3) 5/5	−0.0502 (28.1) 5/5	−0.0529 (7.7) 5/5
Solar, live clusters, unfloored	−0.0325 (4.1) 5/5	−0.0292 (6.3) 5/5	−0.0326 (4.9) 5/5

The split position does not decide the comparison with least squares either. When the chronological split moves from 0.7 of each record to 0.6 or 0.8 with every other setting unchanged, the decomposition, floored or not, still loses on wind on every seed of both sets at every split. At 0.6, 0.7 and 0.8, the floored decomposition loses by 0.0448, 0.0312 and 0.0218 on seeds 1 to 5, and by 0.0487, 0.0270 and 0.0443 on seeds 6 to 10. On solar it wins on every seed at every split, by 0.0284, 0.0230 and 0.0251 on seeds 1 to 5, and by 0.0313, 0.0236 and 0.0268 on seeds 6 to 10. The unfloored decomposition on solar wins at 0.6 on both seed sets and ties at 0.7 and 0.8.

Sample size does not decide the comparison with least squares either. Scored on the window used for solar, wind has the same 13,289 training and 5,695 held-out hours as solar, and it has more censored cells to score (40,931 against 35,614). The floored decomposition still loses there on every seed. On that window, the arm fitted directly on y^{avail} beats the ratio estimator on every seed of both

technologies, scoring 0.1203 against 0.1486 on wind and 0.1273 against 0.1557 on solar. The gain is 0.0284 on each technology. Wind therefore does not lose for lack of data.

The difference in inputs does not decide the comparison with least squares either. The ratio estimator reads the scalar proxy, whereas the structured decomposition reads the nine-feature expansion. Refitted with the same nine features and every other setting unchanged, the feature-matched ratio estimator scores 0.1652 on solar and 0.1459 on wind. The calendar-augmented one adds only the six calendar terms to the scalar and scores 0.1463 and 0.1388. The floored decomposition beats both variants on solar on every seed of all three seed sets. Its margin lies between 0.0320 and 0.0330 against the first variant and between 0.0136 and 0.0143 against the second. On wind it loses to both on every seed. The arm fitted directly on y^{avail} improves with the same nine features, to 0.1057 on solar, so the features carry information about availability that a labelled fit can use. Run with the headline inputs, the same code reproduces the reported ratio estimator and arm fitted on y^{avail} to within 6.2×10^{-8}. Refitted with y^{avail} and y^{curt} permuted by row, the new ratio arms return the same censored-hour MAE to within 1.2×10^{-8}, which is the variation between repeated fits on identical inputs.

K UPPER-QUANTILE RATIO ESTIMATORS

Table 12 reports every arm of the quantile grid with the errors of Section 5 on seeds 1 to 5. These are the censored-hour, uncensored-hour and global MAE, and the mean of the first two. The table also counts the seeds of the three sets on which each quantile arm beats the floored decomposition on censored hours. The arm with calendar inputs at 0.75 beats it on all 15 seeds of both technologies. The arm with the headline inputs at 0.75 beats it on every seed on wind and on 14 of 15 on solar, and the arm with the nine features at 0.75 beats it on every seed on wind and on none on solar. At 0.9 the arms are close to the decomposition on solar and, except for the nine-feature arm, on wind. At 0.95 every arm loses on solar, and on wind every arm degenerates to the same error on every seed. On solar every quantile arm keeps the uncensored-hour error below 0.04, against 0.2483 for the floored decomposition. On wind the quantile arms make more false corrections than the least-squares ratio estimator, 0.0853 at best against 0.0538. The grid was fixed before it was run, and the cell that Table 3 shows is its best on solar by censored-hour MAE.

A validation block cut from the training window makes the choice among the nine cells without the test block. Each cell is fitted on the earlier 70% of the training window and scored on the rest, against the published adjudication of those hours. On solar the 0.75 quantile with the headline inputs leads, at 0.1275 against 0.1311 for the calendar inputs, and on wind the 0.75 quantile with the nine features leads the grid. The same cells lead the grid when the mean of censored-hour and uncensored-hour MAE is the criterion. On the test block they beat the floored decomposition on 14 and 15 of 15 seeds. The validation block reads adjudications of training hours, so it checks selection by the test block and is not a criterion free of the counterfactual.

The quantile arms read no curtailment label either, so the perturbations of Appendix M apply to them exactly. Table 13 takes the two arms at 0.75 that beat the floored decomposition on solar, with calendar inputs and with the headline inputs. Each is refitted once per seed and scored against the perturbed targets and either side of the revision of those runs. The refits reproduce Table 12 to within 10^{-16}. Under every condition, the arm that reports y^{act} unchanged reproduces its recorded scores to within 10^{-16}, and a refit of the least-squares ratio estimator reproduces its recorded scores to within 2×10^{-7}, as both must if every arm is scored against the same target. On wind both arms stay below the decomposition on every seed under every condition. On solar the arm with calendar inputs stays below it on at least 14 of 15 seeds whenever only the size of an adjudication changes, and the arm with the headline inputs on at least 11. Moving cells across the restricted boundary narrows both leads. At a probability of 0.10 each way, the arm with calendar inputs is below the decomposition on 7 of 15 seeds, with a mean margin of 0.0013 over the 15, and the arm with the headline inputs on 3. Such a move gives unrestricted hours the mean adjudication of their cluster, so it rewards an arm that corrects upward everywhere. On uncensored hours both quantile arms stay ahead of the decomposition on every seed under every condition.

The quantile arms have no sparsity weight or width, so their scores also serve every configuration of the structured sweep in Appendix H, which keeps the default split and cluster set. On solar the arm with calendar inputs is below the floored decomposition on every seed at every configuration,

Table 12: Every arm of the quantile grid beside four arms of Table 3, on seeds 1 to 5. C and U are the censored-hour and uncensored-hour MAE. G is the global MAE and B the mean of C and U. The column headed n counts the seeds, of the 15 in seeds 1 to 15, on which the arm beats the floored decomposition on censored hours. Q stands for the quantile ratio estimator, followed by the quantile and the inputs: the headline inputs, the nine features, or the headline inputs with calendar terms.

	Wind					Solar				
Arm	C	U	G	B	n	C	U	G	B	n
Reports y^{act} unchanged	0.1920	0	0.0729	0.0960	–	0.1923	0	0.0547	0.0962	–
Least-squares ratio est.	0.1400	0.0538	0.0865	0.0969	–	0.1557	0.0090	0.0507	0.0823	–
Floored decomposition	0.1713	0.1949	0.1859	0.1831	–	0.1327	0.2483	0.2154	0.1905	–
Fit directly on y^{avail}	0.1401	0.1405	0.1404	0.1403	–	0.1273	0.0319	0.0590	0.0796	–
Q 0.75, headline	0.1407	0.1093	0.1212	0.1250	15	0.1294	0.0190	0.0504	0.0742	14
Q 0.75, nine	0.1250	0.0853	0.1004	0.1051	15	0.1422	0.0176	0.0530	0.0799	0
Q 0.75, calendar	0.1388	0.1072	0.1192	0.1230	15	0.1237	0.0198	0.0494	0.0718	15
Q 0.9, headline	0.1701	0.1829	0.1780	0.1765	4	0.1335	0.0318	0.0607	0.0826	4
Q 0.9, nine	0.1384	0.1508	0.1461	0.1446	15	0.1320	0.0298	0.0589	0.0809	10
Q 0.9, calendar	0.1676	0.1807	0.1757	0.1741	5	0.1352	0.0324	0.0616	0.0838	0
Q 0.95, headline	0.3795	0.4552	0.4265	0.4174	0	0.1519	0.0394	0.0714	0.0956	0
Q 0.95, nine	0.3795	0.4552	0.4265	0.4174	0	0.1354	0.0381	0.0658	0.0868	0
Q 0.95, calendar	0.3795	0.4552	0.4265	0.4174	0	0.1508	0.0397	0.0713	0.0953	0

Table 13: The two quantile arms at 0.75 that beat the floored decomposition on solar, under the perturbed targets and the date split of Table 15. Each margin is the floored decomposition minus the quantile ratio estimator in censored-hour MAE on seeds 1 to 5, so a positive value favours the quantile ratio estimator. The column headed n counts the seeds, of the 15 in seeds 1 to 15, on which the quantile ratio estimator is below the decomposition. Cal. and Head. stand for the calendar and headline inputs. The window held out for solar lies wholly after the revision.

	Wind				Solar			
Target	Cal.	n	Head.	n	Cal.	n	Head.	n
as reported	$+0.0325$	15	$+0.0306$	15	$+0.0089$	15	$+0.0033$	14
wedge $\times 0.8$	$+0.0380$	15	$+0.0357$	15	$+0.0087$	15	$+0.0068$	15
wedge $\times 0.9$	$+0.0350$	15	$+0.0329$	15	$+0.0085$	15	$+0.0046$	14
wedge $\times 1.1$	$+0.0293$	15	$+0.0276$	15	$+0.0097$	15	$+0.0026$	11
wedge $\times 1.2$	$+0.0264$	15	$+0.0249$	15	$+0.0104$	14	$+0.0024$	11
per-cell error, sd 0.1	$+0.0323$	15	$+0.0304$	15	$+0.0091$	15	$+0.0036$	14
per-cell error, sd 0.2	$+0.0324$	15	$+0.0305$	15	$+0.0098$	15	$+0.0047$	15
each cell moved with probability 0.05	$+0.0325$	15	$+0.0307$	15	$+0.0055$	14	$+0.0005$	7
each cell moved with probability 0.10	$+0.0323$	15	$+0.0307$	15	$+0.0026$	7	-0.0019	3
held out before 2025-08-01	$+0.0197$	15	$+0.0169$	15	–	–	–	–
held out from 2025-08-01	$+0.0392$	15	$+0.0377$	15	–	–	–	–

over all seven sparsity weights and the widths 32 and 128, which is 120 fits. On wind it is below on every seed at the shared weight and at 0.001 and on 7 of 10 at 0.03 and 0.05. It is below on none at 0.1, 0.3 and 1.0 or at 0.1 with width 32, and at 0.1 the floored decomposition scores 0.1277 against 0.1388. At width 128 and 0.1 it is below on 2 of 10.

The two arms are also refitted under the other conditions of the appendices, at the shared configuration (Table 14). Each condition is rebuilt from the contract, and its runs must match the rebuild in seeds, split, row counts and training configuration. The arm that reports y^{act} unchanged and a refit of the least-squares ratio estimator then reproduce those runs to within 10^{-16} and 4×10^{-7}. On wind both arms stay below the floored decomposition on every seed of every condition. On solar the arm with calendar inputs stays below it on 19 of 20 seeds at the splits 0.6 and 0.8, and at the early split it is below on 8 of 15, with a mean margin of -0.0017 over the 15. On the solar clusters live throughout the record, where the decomposition gains most over least squares, both arms stay above it on every seed. There the arm with calendar inputs reproduces about three fifths of the gain of 0.0525 and the arm with the headline inputs about half, and the decomposition keeps the rest. On

uncensored hours the arm with calendar inputs stays ahead of the decomposition on every seed of every condition.

Table 14: The two quantile arms at 0.75 under the other conditions of the appendices, at the shared configuration. The first margin is the floored decomposition minus the least-squares ratio estimator, as in Table 7. The others are the floored decomposition minus the quantile ratio estimator, so a positive value favours the quantile ratio estimator. Margins are in censored-hour MAE on seeds 1 to 5, except in the last column, which is in uncensored-hour MAE. Each column headed n counts, over every seed the condition is run on, the seeds on which the quantile ratio estimator is below the decomposition. Cal. and Head. stand for the calendar and headline inputs.

	Least squares	Quantile, Cal.		Quantile, Head.		Cal., uncensored
Condition	margin	margin	n	margin	n	margin
Wind, split 0.6	$+0.0448$	$+0.0407$	10/10	$+0.0390$	10/10	$+0.0811$
Wind, split 0.8	$+0.0218$	$+0.0304$	10/10	$+0.0289$	10/10	$+0.0691$
Wind, levelled window	-0.0001	$+0.0140$	15/15	$+0.0124$	15/15	$+0.0442$
Wind, solar window	$+0.0133$	$+0.0204$	15/15	$+0.0183$	15/15	$+0.0678$
Solar, split 0.6	-0.0284	$+0.0082$	10/10	$+0.0079$	9/10	$+0.2637$
Solar, split 0.8	-0.0251	$+0.0059$	9/10	$+0.0013$	8/10	$+0.2433$
Solar, early split	-0.0331	-0.0007	8/15	-0.0033	2/15	$+0.2253$
Solar, live clusters	-0.0525	-0.0204	0/15	-0.0260	0/15	$+0.1946$

L Clusters commissioned mid-series

A cluster carries a zero capacity factor before it is built, so an arm can train partly on structural zeros. Restricting to clusters live throughout the record, with positive capacity at every hour from 2021-10-01, leaves wind untouched because all 18 wind clusters are live across it. On solar it keeps 15 of the 22 clusters, since it also drops the four built after the record starts but before the solar window opens, and it strengthens solar. The floored decomposition improves from 0.1327 to 0.1076 against 0.1601 for the least-squares ratio estimator. The unfloored decomposition turns a tie into a win of 0.0325 at 4.1 seed standard deviations on 5 of 5 seeds. At the shared configuration the solar margin over least squares is not an artefact of the structural zeros. On the clusters free of them it is larger.

M Error in the adjudicated target

The counterfactual is estimated by the operator, so an ordering measured against it is only as good as that estimate. The identity $y^{\mathrm{act}} = y^{\mathrm{avail}}(1 - y^{\mathrm{curt}})$ separates what is metered from what is adjudicated. Realised output is measured, whereas the wedge $y^{\mathrm{avail}} - y^{\mathrm{act}}$ is what the operator says the restriction costs. The wedge is a capacity factor, as are the two terms from which it is built. Table 15 reports the margin of the floored decomposition over the least-squares ratio estimator, scored against the target with that wedge scaled by 0.8, 0.9, 1.1 and 1.2 or multiplied cell by cell by $\max(0, 1 + \varepsilon)$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ and $\sigma = 0.1$ or 0.2. Under all those perturbations, the censored cells remain those marked by the record, so only the size of an adjudication changes, not which hours are adjudicated. In the two membership rows of the table, the set of censored cells changes instead. An error in that set is the other way an adjudication can be wrong. Neither arm on which the comparison turns reads a curtailment label, so what they predict cannot depend on the target. The re-scoring is therefore exact and requires no rerun. The two label-supervised arms are re-scored in the same way. This re-scoring understates their exposure to error in the target, because a different reconstruction also changes the labels they train on. In the runs behind the table, the headline fits are unchanged to within 9.3×10^{-15}.

A matched comparison within the record shows how far the adjudication is from output in comparable unrestricted hours, and it is a consistency check rather than an independent measurement. Within one cluster, one month of the year and one decile of the variable that sets output (wind speed at 50 m or global horizontal irradiance), unrestricted hours show what the plant does under those conditions. When the strata are pooled by restricted cells, the adjudicated availability of restricted hours is 1.0014 times the corresponding unrestricted output on wind and 1.0203 times it on solar. The $2{,}130$

Table 15: The ordering under error in the adjudicated target. Every row reports the floored decomposition minus the least-squares ratio estimator on censored-hour MAE, the sample standard deviation of the paired per-seed difference in parentheses, and the seeds on which the decomposition is below the ratio estimator. A negative value favours the decomposition. The rows differ in what changes. The wedge and per-cell rows report one set of predictions re-scored against an altered target. The membership rows report those predictions re-scored against a target whose censored cells have moved. In the two date rows, the target is as recorded and the window held out for wind divides at the revision of the reference-generation rule. The window held out for solar lies wholly after it. In the last row, every arm is refitted under a second software environment on the same target and window.

	wind		solar	
target	margin (sd)	below	margin (sd)	below
as reported	$+0.0312$ (0.0087)	0/5	-0.0230 (0.0023)	5/5
wedge $\times 0.8$	$+0.0546$ (0.0123)	0/5	-0.0047 (0.0055)	4/5
wedge $\times 0.9$	$+0.0417$ (0.0103)	0/5	-0.0152 (0.0031)	5/5
wedge $\times 1.1$	$+0.0225$ (0.0074)	0/5	-0.0287 (0.0032)	5/5
wedge $\times 1.2$	$+0.0150$ (0.0064)	0/5	-0.0331 (0.0044)	5/5
per-cell error, sd 0.1	$+0.0321$ (0.0088)	0/5	-0.0220 (0.0024)	5/5
per-cell error, sd 0.2	$+0.0344$ (0.0091)	0/5	-0.0189 (0.0024)	5/5
each cell moved with probability 0.05	$+0.0311$ (0.0086)	0/5	-0.0233 (0.0066)	5/5
each cell moved with probability 0.10	$+0.0311$ (0.0088)	0/5	-0.0237 (0.0113)	5/5
held out before 2025-08-01	$+0.0248$ (0.0086)	0/5	–	–
held out from 2025-08-01	$+0.0346$ (0.0087)	0/5	–	–
second software environment	$+0.0277$ (0.0125)	0/5	-0.0226 (0.0026)	5/5

wind strata cover 100.0% of the restricted cells, and the $1{,}408$ solar strata cover 99.3% of them. The rest fall in strata with no unrestricted hour for comparison. The pooled ratio uses per-stratum weights, whereas the median across strata does not. Splitting the unrestricted hours in half and pooling them with the same weights gives 1.0002 on wind and 0.9973 on solar. Both values measure the sampling stability of the construction and say nothing about bias. Relative to the wedge, which the rows above scale, the pooled discrepancies are 0.6% on wind and 6.3% on solar, and those of the median stratum are 7.6% and 22.7%. The comparison holds only where a restriction is unrelated to what the plant would have produced within a stratum. A congestion-driven policy need not meet that requirement, so the comparison cannot exclude an error that depends on the same conditions as the restriction. What it does say is that, with the strata pooled, the adjudication is close to matched behaviour. The perturbations above cover both pooled discrepancies and the wind median stratum but stop short of the solar median stratum.

The margin keeps its sign on both technologies. On wind the floored decomposition stays above the ratio estimator on every seed at every perturbation, by 0.0150 where the wedge is a fifth larger and 0.0546 where it is a fifth smaller. On solar it stays below the ratio estimator on every seed except at the smallest wedge, where the margin narrows to 0.0047 and one of five seeds crosses. The one perturbation tested that puts the solar result inside the seed spread is a counterfactual that is systematically a fifth too generous. The change of method by the operator in RO-AO.BR.13 Rev.08 on 2025-08-01 falls inside the window held out for wind. It divides that window into 4,973 hours before it and 7,296 after. The wind loss holds on both sides, at 0.0248 before the change and 0.0346 after it. The window held out for solar opens 2025-10-06, wholly after the revision, so the solar result rests on the revised calculation alone. The operator also corrects itself in public. Over the months the benchmark covers, the operator publishes a settled reference generation for 9.5% of restricted wind rows and 7.3% of solar ones. That figure differs from the first estimate on 3.9% of all restricted rows on wind and 2.4% on solar. Where it differs, the median row on either technology has its reference set to zero. The restriction on such a row turns out to have cost nothing. Taking the settled figure wherever it exists lowers the curtailed energy in the window by 3.4% on wind and 1.8% on solar. The operator decides where to publish that figure, so what follows is not an estimate of the error on the rows it leaves alone. Rebuilding the benchmark on that figure and running the headline command against it puts the wind margin at $+0.0356$ against $+0.0312$. The floored decomposition stays above the ratio estimator on every seed under both contracts. The solar margin is -0.0227 against -0.0230. The floored decomposition stays below the ratio estimator on every seed under both contracts. A settled figure below realised output closes a wedge, so the second

contract also has fewer censored cells to score on. It has $83{,}087$ censored cells against $83{,}833$ on wind and $35{,}431$ against $35{,}614$ on solar. This comparison is between two contracts rather than one set of cells under two targets. Of the alternative targets tested, the settled figure is the only one the operator publishes rather than the paper constructs, and the ordering under it is the ordering the paper reports. The ordering reproduces on seeds 6 to 10, with margins of $+0.0310$ on wind and -0.0231 on solar. Scoring one set of predictions against both targets separates the change of target from the change of population. On the $83{,}087$ cells both contracts censor, the wind margin reads $+0.0309$ under the first estimate and $+0.0356$ under the settled one. On the $35{,}431$ shared solar cells the solar margin reads -0.0234 and -0.0227 respectively. The second contract changes the margin mainly through its target rather than through the hours it drops. The same holds on seeds 6 to 10, where the wind margin reads $+0.0267$ against $+0.0310$ and the solar margin reads -0.0240 against -0.0231. The last perturbation moves cells across the boundary on which the metric is defined rather than changing a magnitude. Each adjudicated hour is scored as an unrestricted one with a stated probability, and each unrestricted hour is given the mean adjudication of its own cluster with the same probability. At 0.05 and 0.10 each way, the wind margin reads $+0.0311$ and $+0.0311$, and the solar margin reads -0.0233 and -0.0237. Neither sign moves. One reading does move, and it is the one already at zero. The levelled wind window of the drift figure has a margin of -0.0001, with the floored decomposition below the ratio estimator on three of five seeds. That margin reads $+0.0139$ where the wedge is a fifth smaller and -0.0096 where the wedge is a fifth larger, with the same sign on every seed in each direction. A margin inside the spread of its own seeds is decided by a target error of that size, whereas the two margins that carry the result are not. The early solar split of the same figure has a reported margin of -0.0331 and keeps its sign at every perturbation. The levelled window reproduces on seeds 6 to 10, where its margin is -0.0005 as reported and $+0.0153$ against -0.0111 under the two perturbations that move it. On seeds 6 to 10 the wind margin runs from $+0.0112$ to $+0.0503$ across the conditions above, and the solar margin runs from -0.0350 to -0.0031. Neither sign changes at any condition, although the smallest wedge again leaves one solar seed above the ratio estimator. At the stronger prior of Appendix H, the floored decomposition wins on wind. Under the same conditions at that prior, the wind margin runs from -0.0169 to -0.0052, with the floored decomposition below the ratio estimator on every seed at every condition. At that prior the solar margin runs from -0.0092 to -0.0012 and stays inside one seed standard deviation throughout, with the floored decomposition below the ratio estimator on three to four of five seeds. Seeds 6 to 10 at that prior give -0.0182 to -0.0033 on wind and -0.0155 to -0.0039 on solar. In the last row of the table, the software changes rather than the target. That row keeps the same device, data and seeds but uses newer numpy and torch versions. The four arms fitted without the decomposition move by 5.8×10^{-11}, and each structured fit moves by up to 0.0117 on a seed because the optimisation stops at a different point. The ordering does not move. Wind stays above the ratio estimator on every seed at $+0.0277$, and solar stays below it on every seed at -0.0226. The digits reported everywhere else in this paper belong to the environment the reproducibility statement names.

N SUPPORTING TABLES

The tables below are referenced from the main text and placed here rather than in the body. In each case, the surrounding prose states the comparison the table tabulates. Table 17 is the elimination ladder, a diagnosis measured on the calibrated censoring process of Section 5. Table 3 in the body reports the result measured on the record.

Table 16: Curtailment disclosure across three grid operators. All three quantities are derived rather than metered. They differ in who derives them, under what published rule and at what granularity.

System	Granularity	Nature
ONS (Brazil)	Per plant, half-hourly	Adjudicated by the operator under RO-AO.BR.13, from anemometry or irradiance at the plant through a fitted productivity curve, for financial settlement
ERCOT	Per resource, SCED interval ($\sim$5 min), 60-day lag	Inferred as High Sustained Limit minus Base Point (60-Day SCED Disclosure Reports, NP3-965-ER)
CAISO	System-level, hourly, by fuel type	Aggregate; cannot be resolved to an individual resource

Table 17: Seven explanations for the failure of the structured decomposition to recover y^{avail}, and the measurement that closes each. *Collapse* is the degenerate solution $\hat{y}^{\text{curt}} \approx 0$ everywhere. Censored-hour MAE is compared throughout against 0.0516 for the least-squares ratio estimator on the same seeds and split. Lower is better. Rows 2–4 report two seeds because a pre-registered collapse diagnostic stops those configurations before the full grid. The ladder is cumulative, not a set of independent ablations. Each row retains the changes above it. Rows 3 and 4 come from a single run that raises the step budget while rebuilding the proxy, so those two explanations are closed jointly rather than separately.

Explanation	Change that tests it	Measurement that closes it
1. The form of the sparsity prior causes the collapse	L_1 prior replaced by $\text{Beta}(0.5, 1)$, plus a 200-step warm-up of its weight	Collapse deepens, with $\hat{y}^{\text{curt}} < 10^{-6}$ on 100.0% of cells. MAE improves $0.0943 \to 0.0820$, still 0 of 5 seeds
2. The availability head reads y^{act} and copies it	A split encoder leaves the availability head structurally unable to see y^{act}, and the prior becomes $\text{Beta}(0.5, 0.5)$	Copying stops, hours where $\hat{y}^{\text{avail}}$ is within 0.01 of y^{act} fall from 25.9% to 8.5% and 10.6%, and MAE does not improve (0.0767, 0.0721, 0 of 2)
3. The grid-state proxy starves the curtailment head	Proxy rebuilt from one raw scalar to nine features: the scalar, calendar, own-cluster lagged output, a same-technology aggregate of the other clusters	No effect on identification; the collapse on seed 1 becomes total (100.0% of cells below 10^{-6}, from 85.9%). MAE 0.0829, 0.0747
4. The model is undertrained	Step budget raised $2{,}000 \to 10{,}000$, patience unchanged	Both seeds now early-stop, at $4{,}891$ and $4{,}873$ steps, and both land slightly worse than at the smaller budget
5. The supervision loss cannot express a zero-inflated target	Squared error on revealed labels replaced by a Bernoulli gate on $\mathbf{1}[y^{\text{curt}} > 0]$ plus squared error on censored cells only	Under squared error the curtailment estimate is uninformative at every fraction including full ($\lvert\text{corr}\rvert \leq 0.044$, over 99.99% of cells below 10^{-6}); under the gated loss it reaches 0.81
6. Curtailment labels are too scarce	Labels revealed on a fraction $p \in \{0.01, 0.05, 1.0\}$ of training cells, three seeds each	0 of 9 fits beat the ratio estimator. At $p = 1.0$ the error is 0.0697 (seeds 0.0670–0.0716) against 0.0516
7. The censoring process is never identified	The same nine fits, read for identification rather than for error	At $p = 1.0$, $\text{corr}(\hat{y}^{\text{curt}}, y^{\text{curt}}) = 0.813$ on censored hours, where mean $\hat{y}^{\text{curt}}$ is 0.190 against 0.073 elsewhere. The process *is* identified, and recovery still loses