REQUIRED-LEVEL RISK CONTROL FOR TIME SERIES FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable time series forecasting requires risk control at the temporal scale of deployment, yet long-run average coverage can mask frequent window-level failures. We propose Required-Level Risk Control (ReqRC), which replaces outcome-based updates with feedback on the least conservative level that would have satisfied each completed window's target. We uniquely decompose window shortfall into predictor shortfall and selection shortfall, identifying the latter as the component addressed by level adjustment. We establish an exact worst-case lower bound on the window success rate and derive reliability bounds for delayed order statistics, adjusted for distribution shift and infeasibility. ReqRC uses these statistics to raise the level of any causal base method without training or tuning. Across eight real-world datasets, ReqRC achieves the lowest mean shortfall and highest window success rate against ten alternatives, with improvements across all 40 predictor–method pairs. On the test set, alternatives matched to ReqRC's validation cost are less reliable, whereas those matching its success rate incur higher costs. By turning retrospective feasibility into feedback for future decisions, ReqRC provides a principled approach to cost-efficient window-level risk control.

1 INTRODUCTION

Uncertainty quantification for time series is crucial across domains including climate science, energy, compute scheduling, and resource allocation, especially in high-stakes settings. An ideal method should meet its coverage target in every forecast window, not only on average: the forecast for a window must be fixed before that window begins, so a forecast that proves too low cannot be corrected afterwards, and one that is too high wastes resources. Online conformal prediction (Vovk et al., 2005, Shafer & Vovk, 2008, Lei et al., 2018) instead optimizes average coverage over the whole horizon, and we question that objective: the averaging can hide almost every window failure. Proposition 1 gives the exact worst-case window success rate in closed form: at the window length and target used in this paper, the long-run guarantee collapses to 4%.

When a window fails, existing methods (Gibbs & Candès, 2021, Angelopoulos et al., 2024) do not consider whether any candidate forecast could have met the target. A window misses either because the predictor offered no forecast high enough or because the method did not select one that was; only the second can be corrected by level selection. We prove in Section 4.2 that the shortfall v_t splits uniquely into $g_t + r_t$: g_t is the smallest shortfall attainable by any candidate forecast, and r_t the excess left by the selected level. The observed shortfall alone therefore cannot tell a better predictor from a better method.

We separate the two to show what reliability in every window costs. Given only an upper bound ρ on the highest level's failure probability and the window success target β, Theorem 1 shows that any method that must work for every admissible distribution must select that level with probability at least $\beta/(1-\rho)$. Reliability is bought with conservatism.

We next use delayed history to derive a finite-sample bound on the current window's success probability: each completed window reveals exactly how conservative the forecast would have had to be, information that existing methods discard, since they feed back only the outcome of the level they deployed. Theorem 2 turns the j-th smallest of the last K such minimum feasible levels into that bound, accounting for delayed feedback, distribution shift, and possible infeasibility. Based on these two results, we propose Required-Level Risk Control (ReqRC), which selects the smallest level

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.

satisfying both the base level and the historical threshold, so a base level that already exceeds the threshold is preserved; Figure 1 shows the three steps.

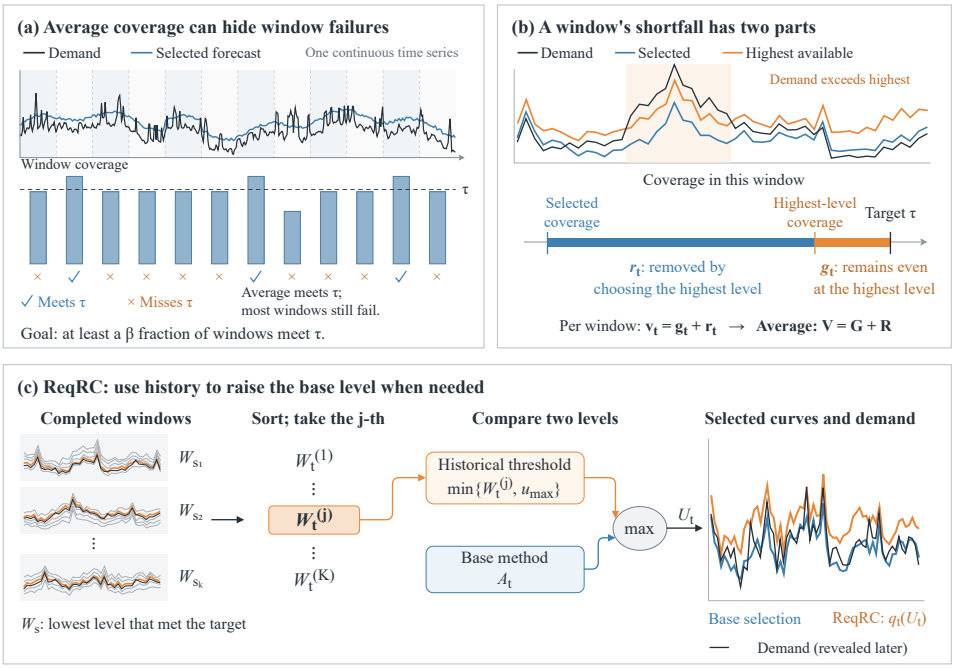


Figure 1: **Window-level coverage and the ReqRC update.** (a) Average coverage reaches the target while most windows fall below it. (b) Selecting the highest level eliminates r_t, leaving g_t. (c) The required-level feedback loop: completed windows return the level that would have met the target, its order statistic sets the threshold, and the base level is retained.

We summarize our contributions:

- We establish the exact worst-case separation between long-run and window-level coverage: a method can meet the long-run target while 96% of windows fail (Proposition 1). We uniquely decompose window shortfall into predictor and selection components (Proposition 2). We then characterize the minimum probability of selecting the highest level required for uniform reliability (Theorem 1), and derive a finite-sample window success bound from delayed historical levels (Theorem 2). Neither theorem assumes anything about the data-generating process.
- The two theorems lead directly to a simple algorithm. We propose ReqRC, which, unlike existing updates that feed back the outcome of the level they deployed, calibrates a given predictor's forecast by required-level feedback. We prove that among all levels that retain the base level and meet the historical threshold its update is the smallest, and therefore the cheapest (Corollary 1). It is one order statistic per window, with no parameters trained and no learning rate to set, and it wraps any causal base method with an ordered family of candidate forecasts without touching the predictor (Section 5, Algorithm 1).
- Extensive experiments on eight real-world time-series datasets show that ReqRC attains the lowest mean shortfall and the highest window success rate. Under matched comparisons, it is the safest at equal cost and the least costly at equal safety (Section 6).

2 Related Work

Online risk control and window-level coverage. Conformal Risk Control (CRC) extends conformal calibration from miscoverage to monotone losses (Angelopoulos et al., 2024). Its finite-sample guarantee presupposes that the calibration parameter can drive the loss below the target; when no admissible parameter does, the bound is silent, which our setting makes explicit through ρ_t. Adaptive Conformal Inference (ACI) moves a scalar threshold after each coverage error; AgACI, DtACI, and SF-OGD aggregate or shorten these updates (Gibbs & Candès, 2021, Zaffran et al., 2022, Gibbs & Candès, 2024, Bhatnagar et al., 2023). Rolling RC handles general monotone risks, Conformal PID

combines quantile tracking with score prediction, and L-ARC localizes calibration (Feldman et al., 2023, Angelopoulos et al., 2023, Zecchin & Simeone, 2024). ECI incorporates error magnitude, NCC learns the calibration step, and COP estimates the score distribution (Wu et al., 2025, Li & Rodríguez, 2025, Hu et al., 2026). Vaze (Vaze, 2026) and Cost-Aware ACI (Wu et al., 2026) report simultaneous coverage-and-efficiency guarantees and severity-weighted updates. These methods expand the uncertainty set until the risk falls below the target, which requires the threshold to reach the whole label space. Calibrating the smallest member of a nested family of sets is standard (Gupta et al., 2022), but such families are unbounded above; our level family is fixed instead, so a window may be infeasible. Their local guarantees either localize over the covariate space, as in L-ARC, or average per-step coverage over intervals of the observed stream, as in the strongly adaptive regret of SF-OGD (Zecchin & Simeone, 2024, Bhatnagar et al., 2023, Gibbs & Candès, 2024); ours bounds the frequency with which individual windows meet the target. All of them provably control long-run coverage under arbitrary distribution shift, yet none bounds the coverage of an individual window. Their updates are driven by the outcome of the level already deployed — a miscoverage indicator, its magnitude, a predicted score, or an estimated score distribution — whereas ReqRC feeds back the required level of each completed window, which does not depend on the deployed level. It needs no step size, and it separates the shortfall no level can remove from the shortfall a better level can.

Decision post-processing and nonexchangeability. Conformal decision methods, safety filters, and risk-controlled post-processors modify an existing policy or select a fallback under a risk constraint (Strawn et al., 2023, Chee et al., 2024, Lekeufack et al., 2024, Kiyani et al., 2025). Nonexchangeable analyses quantify coverage changes under distribution shift and temporal dependence (Tibshirani et al., 2019, Barber et al., 2023, Barber & Pananjady, 2026). Distribution-free conditional coverage is impossible without assumptions (Lei & Wasserman, 2014, Foygel Barber et al., 2021); our window target is a quantifiable relaxation. ReqRC instead builds its threshold from historical minimum feasible levels.

Prediction-driven allocation. Quantile regression and probabilistic forecasting produce ordered demand bounds under asymmetric shortage and overprovisioning costs, and conformalized quantile regression calibrates such intervals with finite samples (Koenker & Bassett, 1978, Romano et al., 2019, Salinas et al., 2020, Lim et al., 2021). EnbPI and SPCI instead build time-series intervals from point forecasts and sequential residuals (Xu & Xie, 2021; 2023); ReqRC takes the ordered forecasts as given candidates and calibrates the prediction level online.

3 Problem Formulation

We study online calibration for time-series forecasting: before each window an existing predictor supplies an upper quantile forecast of the observations in that window, and the method chooses how conservative to make it. The goal is to reduce coverage shortfalls within windows while limiting the cost of a forecast that is too conservative. Throughout, the risk we control is this shortfall. In capacity allocation Y_t is demand and the cost is excess reservation; otherwise the formulation is general.

We partition the deployment horizon into T nonoverlapping windows of H time points. Let $Y_t = (y_{t,1}, \dots, y_{t,H}) \in \mathbb{R}^H$ denote the quantity to be forecast in window t, and let $\mathcal{F}_t$ contain the information available before that window. Given $\mathcal{F}_t$, the predictor supplies an ordered family $\{q_t(u) : u \in \mathcal{U}\}$ with $\mathcal{U} = [u_{\min}, u_{\max}]$. Each trajectory $q_t(u) = (q_{t,1}(u), \dots, q_{t,H}(u)) \in \mathbb{R}^H$ gives the forecast at level u; forecasts increase with u, so higher levels are more conservative. Throughout, forecasts are supplied at six prespecified levels $\{0.85, 0.90, 0.95, 0.975, 0.99, 0.995\}$ and interpolated linearly between adjacent levels, so $u_{\min} = 0.85$ and $u_{\max} = 0.995$.

Assumption 1 (Available information). The method selects $U_t \in \mathcal{U}$ from the available history and deploys $q_t(U_t)$. We assume that the conditional distribution of Y_t is invariant to the deployed level. For randomized selection, the method's internal randomization is independent of Y_t conditional on $\mathcal{F}_t$.

For target $\tau \in (0, 1)$, define

$$\mathrm{Cover}_t(u) := \tfrac{1}{H} \textstyle\sum_{h=1}^{H} \mathbf{1}\{y_{t,h} \le q_{t,h}(u)\}, \qquad v_t(u) := [\tau - \mathrm{Cover}_t(u)]_+.$$

Here, $\mathbf{1}\{\cdot\}$ is the indicator function, and $[x]_+ := \max\{x, 0\}$ denotes the positive part.

The coverage target τ is the fraction of observations a window must cover, and β is the frequency with which windows should attain it. A window is feasible when some $u \in \mathcal{U}$ reaches τ, and ρ_t is the probability that it is not. Mean shortfall V measures the magnitude of coverage deficits; $1 - \widehat{p}_\tau$ records their frequency.

A level u meets the window target when $\mathrm{Cover}_t(u) \geq \tau$. The empirical window success rate is $\widehat{p}_\tau := \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\{\mathrm{Cover}_t(U_t) \geq \tau\}$.

We measure the cost of an overly conservative forecast as the amount by which it exceeds the realized observations, using positive mean absolute error (PMAE), defined as $\mathrm{PMAE} := \frac{1}{TH}\sum_{t=1}^{T}\sum_{h=1}^{H}[q_{t,h}(U_t) - y_{t,h}]_+$. For cross-dataset comparisons, we normalize each channel by its training interquartile range before computing PMAE.

4 Shortfall Decomposition and the Price of Reliability

In this section, we prove three properties of window-level reliability. We first show that long-run average coverage can be satisfied while 96% of windows fail (Proposition 1). We then show that each window's shortfall decomposes uniquely into predictor and selection components (Proposition 2). Finally, we show that uniform reliability forces a specific top-level frequency when the candidate family is bounded (Theorem 1).

4.1 Long-run coverage and window-level coverage. We first quantify how far the two objectives can diverge. **Proposition 1 (Long-run coverage does not ensure window-level coverage).** Coverage in every window implies long-run average coverage, whereas long-run coverage permits window failures. A successful window covers at least $\lceil \tau H \rceil$ time points, with $\lceil x \rceil$ denoting the ceiling function. The largest coverage rate of a failed window is therefore $(\lceil \tau H \rceil - 1)/H$. These observations yield the sharp lower bound

$$\frac{1}{T}\sum_{t=1}^{T} \mathrm{Cover}_t(U_t) \geq \tau \quad \Longrightarrow \quad \widehat{p}_\tau \geq \frac{\tau - (\lceil \tau H \rceil - 1)/H}{1 - (\lceil \tau H \rceil - 1)/H}. \tag{1}$$

For $H = 96$ and $\tau = 0.85$, equation (1) gives $\widehat{p}_\tau \geq 0.04$, and the bound cannot be improved in general: in words, a method can meet the long-run target while failing 96% of windows. Appendix A.1 proves this and constructs a distribution attaining mean window shortfall $\tau(1-\tau)$, its maximum under the long-run target.

4.2 Exact shortfall decomposition. We next separate the shortfall by what determines it. Define predictor shortfall as $g_t := \min_{u \in \mathcal{U}} v_t(u)$ and selection shortfall as $r_t(u) := v_t(u) - g_t$, with $r_t := r_t(U_t)$ and U_t the deployed level. For each candidate family and realized Y_t, $\boxed{v_t(U_t) = g_t + r_t}$, $\quad g_t, r_t \geq 0$. **Proposition 2 (Uniqueness of the shortfall decomposition).** Three conditions uniquely determine the decomposition: predictor shortfall is independent of the selected level, both components are nonnegative, and selection shortfall vanishes at every optimal level. It holds whenever v_t attains a minimum on $\mathcal{U}$, not only for ordered forecasts, and for ordered forecasts $g_t = [\tau - \mathrm{Cover}_t(u_{\max})]_+$, so $g_t > 0$ precisely when the window is infeasible. Averaging over windows gives

$$V := \frac{1}{T}\sum_{t=1}^{T} v_t(U_t), \qquad G := \frac{1}{T}\sum_{t=1}^{T} g_t, \qquad R := \frac{1}{T}\sum_{t=1}^{T} r_t, \qquad \boxed{V = G + R}. \tag{2}$$

In words, zero selection shortfall means the deployed forecast attains the best coverage its candidate family offers; in an infeasible window only r_t can be removed.

Remark 1 (Forecast expansion and deployed risk). Expanding the candidate family while preserving existing forecasts weakly decreases predictor shortfall. If the deployed forecast stays fixed, v_t is constant and each decrease in g_t becomes an equal increase in r_t, so predictive improvements lower risk only when the added forecasts are deployed. Appendix A.1 proves uniqueness, the expansion identity, and the corresponding comparisons within a common candidate family.

A causal method and one that observes Y_t before selecting differ only in information, and their gap measures the value of the current observation. Appendix A.3 separates this information gap from the shortfall reducible through causal level selection.

4.3 The minimax price of reliability. We now show what uniform reliability costs.

Theorem 1 (Exact price of reliability over all admissible distributions). Let $\beta \in (0, 1]$ be the window success target and $u_{\min} < u_{\max}$ the least and most conservative admissible levels. Fix a realized history $\mathcal{F}_t = f_t$. For each $\rho \in [0, 1]$, consider all conditional distributions of the minimum required level whose highest-level failure probability is at most ρ. Uniform reliability means achieving window success probability at least β for every history and every distribution in this class. Such reliability is attainable by deployable levels exactly when $\rho \leq 1 - \beta$. In this case, every uniformly reliable causal rule satisfies

$$\boxed{\mathbb{P}(U_t = u_{\max} \mid \mathcal{F}_t = f_t) \geq \frac{\beta}{1 - \rho}} \tag{3}$$

The probability is conditional on the observed history $\mathcal{F}_t = f_t$, and the bound cannot be improved in general: equality is attainable. A deterministic uniformly reliable method therefore selects $u_{\max}$ at every history. Appendix A.2 specifies the distribution class, a sharp two-point construction, an attaining randomized rule, and the associated cost lower bound. In words, the ratio $\beta/(1 - \rho)$ increases with both the requested success probability and the failure bound: at $\rho = 0$ it equals β, and as $\rho \to 1 - \beta$ it approaches one. The literature has avoided this price: it targets long-run averages (Wu et al., 2025, Hu et al., 2026), and Proposition 1 shows how little that choice permits. Theorem 2 bounds the marginal success probability instead, charging an explicit shift correction ε_t for the departure from exchangeability and allowing $u_{\max}$ to be selected far less often; together with Theorem 1 it yields ReqRC. Section 5 develops its threshold and quantifies its reliability.

5 ReqRC and Its Reliability Bound

In this section, we show what delayed history guarantees. We first derive the minimum feasible level of a completed window and use its order statistics as a historical threshold (Section 5.1). We then bound the current window's success probability in finite samples (Theorem 2). Finally, we show how to choose the historical rank (Section 5.3).

Design principle. Where existing updates feed back an error, such as a miscoverage indicator or its magnitude, ReqRC feeds back the required level: the least conservative level that would have met the completed window's target. Selection becomes an order-statistic problem rather than a step-size calibration: the update costs $O(K \log K)$ by sorting, or $O(K)$ by selection, and carries no learning rate (Appendix B.7). Its bound is finite-sample. Of the windows where the base action fell short, 79% admitted a level below $u_{\max}$ that would have met the target and 21% were infeasible (Appendix C.6).

5.1 Required-level feedback and the capped rank threshold. We now derive the feedback that ReqRC uses. Before each window, the base predictor supplies ordered forecasts and the base method a level A_t; each completed window then reveals the minimum level that would have met its target.

For window t, define the minimum feasible level $W_t := \min\{u \in \mathcal{U} : \mathrm{Cover}_t(u) \geq \tau\}$, $\min \varnothing := +\infty$. Assume candidate forecasts are nondecreasing in u and $\mathrm{Cover}_t(u)$ right-continuous. Then $W_t \leq u \iff \mathrm{Cover}_t(u) \geq \tau$. A finite W_t identifies a feasible window, while $W_t = +\infty$ identifies an infeasible one. The value becomes available after the observations arrive and the prespecified feedback delay has elapsed.

Building on the standard exchangeable rank argument, the current minimum feasible level lies below the j-th historical order statistic with probability at least $j/(K + 1)$. We denote the distribution-shift correction between the historical and current windows by ε_t, specified in Appendix A.4. Appendices A.4–A.5 define two discrepancy measures, use their minimum, and establish sharpness of the resulting bounds. The rank argument is standard; its object is not: ReqRC scores a window's required level rather than a single label.

Let $W_t^{(j)}$ be the j-th smallest of the latest K available minimum feasible levels, with $W_t^{(K+1)} := +\infty$. ReqRC caps this threshold at $u_{\max}$ and selects the larger of it and the base level A_t,

$$U_t = \max\left\{A_t,\ \min\left(W_t^{(j)}, u_{\max}\right)\right\}.$$

An infinite threshold gives $U_t = u_{\max}$. Theorem 2 provides the reliability bound for $j \in \{1, \dots, K\}$. The **ReqRC** update maintains a length-K buffer and requires one sort and one maximum.

The maximum preserves the base method's response whenever A_t exceeds the historical threshold; otherwise the threshold determines the selected level. Since forecasts are ordered, the update weakly increases coverage and weakly decreases shortfall relative to the base selection, at the cost of the additional allocation PMAE measures.

Algorithm 1: ReqRC online calibration rule.

Require: target $\tau \in (0, 1)$, history length K, prespecified $j \in \{1, \dots, K+1\}$, ordered candidate forecasts, and base levels A_t from any causal base method.

State: buffer B of revealed (s, W_s) and stored forecasts for windows awaiting feedback.

1: **for** $t = 1, \dots, T$ **do**

2: before observing Y_t, process each newly arrived feedback Y_s from a completed window:
$W_s = \min\left\{u \in \mathcal{U} : \frac{1}{H}\sum_h \mathbf{1}\{y_{s,h} \le q_{s,h}(u)\} \ge \tau\right\}, \qquad \min \varnothing := +\infty$

3: insert each newly computed (s, W_s) into B

4: before observing Y_t, obtain the candidate forecasts and $A_t \in \mathcal{U}$

5: take the latest K revealed values W_s in B

6: if fewer than K values are available, initialize them by the prespecified rule in Appendix B

7: let $W_t^{(j)}$ be their j-th order statistic, read as $+\infty$ when $j = K + 1$

8: deploy $U_t = \max\{A_t, \min(W_t^{(j)}, u_{\max})\}$ and $q_t(U_t)$

9: **end for**

Output: selected levels (U_t), deployed forecasts $(q_t(U_t))$, updated buffer B.

ReqRC selects the smallest deployable level satisfying both the base level and the historical threshold (Appendix A.6): within that class the threshold is determined by the reliability target, and the base level is the only remaining degree of freedom. Its effect on risk is the selection component R of Section 4.2, with G fixed.

Corollary 1 (Minimality within the base-retaining, threshold-meeting class). Fix $j = \lceil (K+1)\beta \rceil$. Among all $\mathcal{H}_t^D$-measurable levels that retain the base level A_t and meet the historical threshold $W_t^{(j)}$, the level $\max\{A_t, \min(W_t^{(j)}, u_{\max})\}$ is the smallest (Appendix A.6).

5.2 Rank reliability. We now bound the success probability this threshold delivers. Recall that ρ_t is the probability that the current window is infeasible and ε_t the shift penalty between history and the current window.

Theorem 2 (Window reliability from delayed history). For each prespecified historical rank $j \in \{1, \dots, K\}$, ReqRC attains zero selection shortfall with probability at least $\left[\frac{j}{K+1} - \varepsilon_t\right]_+$. The result also holds for any causal method satisfying the same historical threshold. The current-window success probability obeys

$$\boxed{\mathbb{P}\left(\mathrm{Cover}_t(U_t) \ge \tau\right) \ge \left[\frac{j}{K+1}(1 - \rho_t) - \varepsilon_t\right]_+} \tag{4}$$

In words, the bound starts from the nominal rank probability $j/(K+1)$ and accounts for infeasibility and distribution shift. It gives the marginal success probability of the current window. When the history and the current window are exchangeable, $\varepsilon_t = 0$, and the bound reduces to $\frac{j}{K+1}(1 - \rho_t)$. Neither ρ_t nor ε_t is needed to run ReqRC, which uses only the revealed minimum feasible levels. Appendices A.4–A.5 define the shift correction and establish the bounds and their sharpness. Appendix A.7 gives a certification procedure that replaces ε_t by a reference bound estimable from revealed history, given a valid bound on future deterioration.

5.3 Choosing j using Theorem 2. We now show how to choose the rank without tuning. In the main experiments, we set it directly from the target window success probability β, $j = \lceil (K+1)\beta \rceil$. This selects the smallest rank whose nominal probability $j/(K+1)$ reaches β: with $K = 8$ and $\beta = \tau = 0.85$ the rank is $j = 8$.

For $K = j = 8$, the nominal probability $8/9$ exceeds 0.85 by 0.0389. Theorem 2 gives a window success probability of at least 0.85 when $\varepsilon_t + \frac{8}{9}\rho_t \le 8/9 - 0.85$. Increasing j raises the threshold and the window success rate $\widehat{p}_\tau$; decreasing it lowers the threshold and the cost. Upper bounds valid jointly across candidate ranks support choosing the smallest rank achieving the corrected target (Appendices A.6–A.7); Appendix D evaluates it on controlled data.

6 Experiments

We evaluate window failures under long-run coverage, predictor shortfall, and ReqRC's effects on reliability and resource cost across multiple predictors, base methods, and matched comparisons, selecting tuned configurations on validation data.

6.1 Experimental Setup. We evaluate eight datasets — Weather, SMD, a CPU trace, Ali, and four ETT splits — comprising 206 channels and 16,316 test windows, with a chronological 56%/14%/30% training/validation/test split. The default configuration uses $H = 96$, $\tau = \beta = 0.85$, $K = j = 8$, the level range of Section 3, and a one-window feedback delay.

We combine TFT (Lim et al., 2021), DeepAR (Salinas et al., 2020), DistPred (Liang, 2025), and a six-level quantile ensemble with ReqRC and ten alternative methods (ACI, Rolling RC, Conformal PID, DtACI, AgACI, SF-OGD, L-ARC, ECI, NCC, and COP), yielding 44 primary combinations. Additional validation-selected configurations produce the 181 operating points in Figure 4. Methods share each predictor's frozen candidate forecasts and apply their updates under the causal protocol in Section 3. For each channel, ReqRC uses common base levels A_t across the four predictors (Appendix B.2). We report V, G, R from equation (2) and PMAE from Section 3. Appendices C.1–C.3 specify the datasets, metrics, aggregation, and model selection, and Appendix B the implementations, Appendix E the artifact, environment, and claim checklist.

6.2 Window failures and predictor shortfall. All 11 methods exceeded the long-run coverage target of 0.85, yet only three of the ten alternatives met the same threshold window by window, their success rates spanning 73.9%–85.6% (Table 1A).

Table 1: Window failures and shortfall decomposition. (A) Method means across four predictors over all 206 channels. (B) Coverage, window failures, and predictor shortfall at the highest level.

A. Method means across four predictors

Method	Avg. cov. ↑	Meeting target $\widehat{p}_\tau$ ↑	Shortfall V ↓	Predictor shortfall G ↓	Selection shortfall R ↓
ReqRC	**94.84%**	**90.04%**	**0.02429**	0.00837	**0.01592**
L-ARC	92.47%	85.13%	0.03616	0.00837	0.02779
COP	92.71%	85.48%	0.03628	0.00837	0.02791
ECI	92.76%	85.57%	0.03639	0.00837	0.02802
SF-OGD	92.38%	84.96%	0.03652	0.00837	0.02814
AgACI	91.92%	84.03%	0.03900	0.00837	0.03063
Conformal PID	91.97%	83.95%	0.03990	0.00837	0.03153
DtACI	91.54%	83.25%	0.04167	0.00837	0.03330
NCC	89.38%	78.76%	0.05433	0.00837	0.04596
ACI	88.02%	75.08%	0.06131	0.00837	0.05294
Rolling RC	87.51%	73.91%	0.06436	0.00837	0.05598

B. Highest-level analysis

Predictor	$u_{\max}$	Max. cov. ↑	Still fails ↓	Predictor shortfall G ↓
TFT	0.995	96.81%	6.54%	0.01724
DeepAR	0.995	98.86%	2.11%	0.00570
DistPred	0.995	99.22%	1.51%	**0.00358**
Quantile ensemble	0.995	98.39%	3.13%	0.00697

Panel B reports direct deployment at $u_{\max} = 0.995$. Across predictors, 1.51%–6.54% of windows fell below the coverage target. That residual is the predictor shortfall no level selection can remove.

The shared candidate forecasts give every method the same mean predictor shortfall, $G = 0.00837$. Under the decomposition in Section 4.2, $R = V - G$, so differences in V are differences in R. Table 1A uses all 206 channels; normalized comparisons use the 203 with nonzero training IQR, where ReqRC's V is 0.02456 rather than 0.02429. Appendices C.4–C.5 detail the configurations and selection procedure.

6.3 Across predictors and designs. ReqRC achieved the lowest mean shortfall ($V = 0.02456$) and the highest mean window success rate ($\widehat{p}_\tau = 0.9003$), averaging the four predictors equally (Table 2). Its shortfall was 32.8% lower than L-ARC's, and its window success rate was 4.50 percentage points higher than ECI's. It was the only method whose mean window success rate exceeded 90%. ReqRC's

Table 2: Window-level coverage and resource cost, averaged across four predictors over the 203 channels with nonzero training interquartile ranges.

Method	Global cov. ↑	V ↓	$\widehat{p}_\tau$ ↑	PMAE ↓	Method	Global cov. ↑	V ↓	$\widehat{p}_\tau$ ↑	PMAE ↓
ReqRC	**0.9482**	**0.02456**	**0.9003**	1.4308	Conformal PID	0.9193	0.04034	0.8390	1.3136
L-ARC	0.9243	0.03655	0.8510	1.0320	DtACI	0.9149	0.04214	0.8320	1.0025
COP	0.9266	0.03669	0.8543	1.3739	NCC	0.8930	0.05499	0.7868	0.8215
ECI	0.9272	0.03681	0.8553	1.4469	ACI	0.8792	0.06212	0.7492	0.8190
SF-OGD	0.9234	0.03693	0.8491	1.0316	Rolling RC	0.8739	0.06525	0.7372	**0.7991**
AgACI	0.9188	0.03939	0.8403	0.9899					

PMAE was 1.4308, compared with 1.0320 for L-ARC; lowering the rank j reduces cost at the expense of window success (Figure 4).

Matched cost and coverage comparisons. One direction matches each baseline's validation cost to ReqRC's and compares held-out safety (Figure 2); the other matches its validation window success rate and compares held-out cost (Figure 2). Appendix C.6 reports both in full.

No baseline matched ReqRC's safety at its cost: all 40 cost-matched mixtures had higher held-out shortfall and lower window success, so the shaded region of Figure 2 is empty. In the other direction, 14 mixtures retained ReqRC's held-out success rate, and all 14 were more expensive.

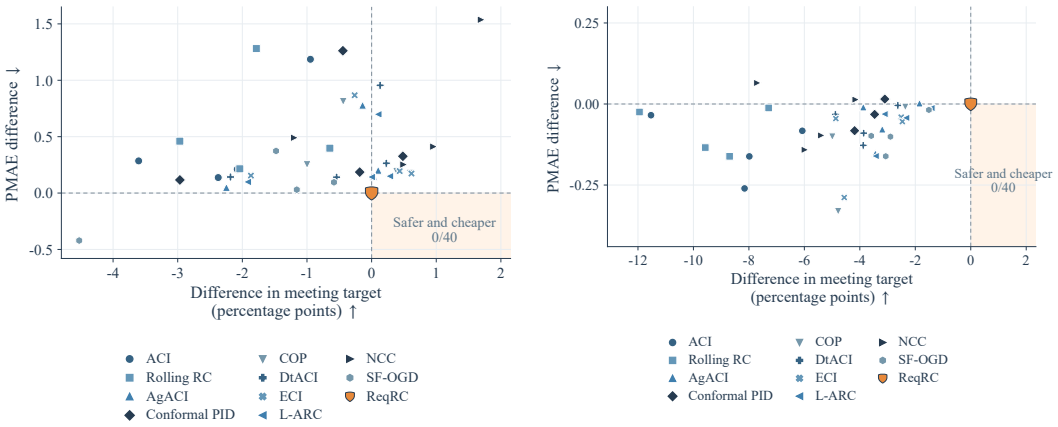
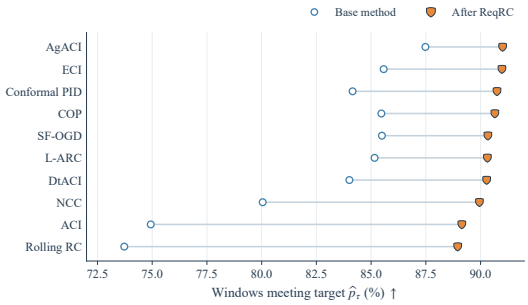


Figure 2: **Matched comparisons in both directions. Left:** baselines meet or exceed ReqRC's validation window success rate; points show test differences, so a point above the axis costs more than ReqRC at equal safety. **Right:** baselines match ReqRC's validation PMAE; the shaded region holds 0 of the 40 matched configurations. In both panels blue markers denote methods across four predictors, the orange shield marks ReqRC at the origin, and horizontal differences are in percentage points. Matching uses validation data, so held-out values need not coincide.



ReqRC improves coverage across base methods. Applying ReqRC to ten methods and four predictors reduced V and increased $\widehat{p}_\tau$ in all 40 combinations, averaged here across predictors. Predictor shortfall G remained constant, attributing the reduction in V to improved level selection.

Figure 3: **Window success rates before and after ReqRC.** Each row averages four predictors for one base method. Blue open circles denote the base method, and orange shields denote the result with ReqRC ($j = 8$); the ten base methods span 11.8 points before ReqRC and 1.95 after.

The paired analysis used 10,000 dataset-stratified channel-bootstrap draws. The weakest base method with ReqRC reached 88.9%, above the 85.5% of the strongest without it (Figure 3). Applied to each base method, ReqRC increased window success by 5.01 points for L-ARC and 15.23 for Rolling RC

on average, and reduced mean shortfall significantly for each of the ten base methods: every 95% interval excluded zero. ReqRC improves window success while selecting the highest level in only 19%–35% of windows. Appendix C.6 reports the complete paired results.

Method design comparisons. Against a window-failure ACI variant (WF-ACI), which updates from binary window success rather than the coverage error and targets $8/9$ success, ReqRC achieved 20.0% lower mean shortfall and a 2.60-percentage-point higher window success rate. This comparison tests the feedback signal and update rule together. To isolate the feedback signal, we fix the base method, history length, delay, and rank rule, and replace historical minimum feasible levels with binary failure records. This variant selects the highest level more often: at the same rank, it is safer but more expensive. At matched cost, required-level feedback reduces mean shortfall by 27.3%. The minimum feasible level records how high the forecast needed to be, rather than only whether it failed.

Finally, we deploy the historical threshold directly to test the effect of retaining the base level. Taking the maximum of the two reduces mean shortfall by 10.3% and increases window success by 1.12 percentage points, at approximately 3.0% higher PMAE. Appendix C.6 reports all three comparisons.

6.4 Risk–cost trade-off: choosing j. Among the 181 test operating points in Figure 4, ReqRC with the six-level quantile ensemble reached $(V, \mathrm{PMAE}) = (0.02172, 1.162)$, a new Pareto point dominating 38 configurations. Along the rank path $j = 1, \ldots, 9$ in Figure 4, increasing j reduced R and increased both $\widehat{p}_\tau$ and cost. Appendices C.7–C.8 report the numerical path that Section 5.3 prescribes and its sensitivity to the target and endpoint. ReqRC's risk and cost fell between fixed levels 0.95 and 0.975, with the highest level selected in 19.31% of windows.

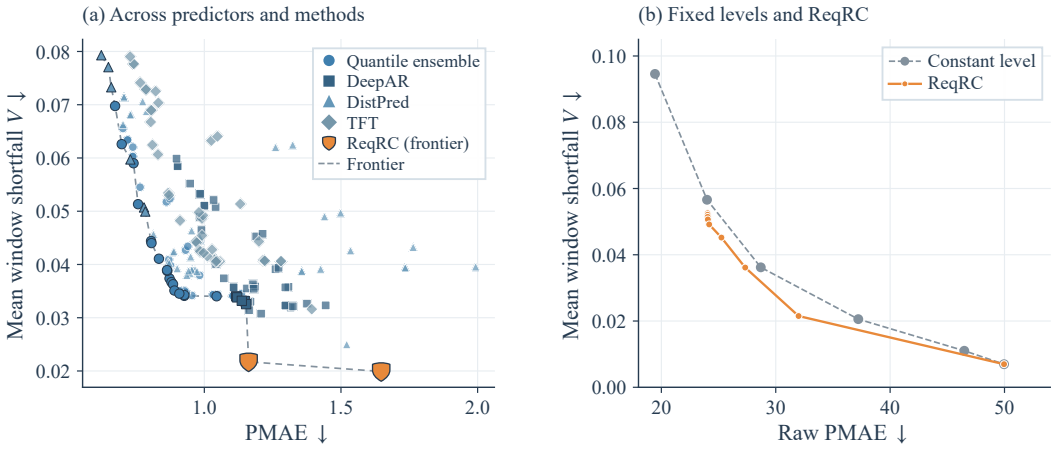


Figure 4: **Risk and cost along the ReqRC rank path. Left:** 181 test configurations with normalized PMAE; blue markers are predictors, the gray dashed line is the pooled-test frontier, and shields mark ReqRC points on it. **Right:** six fixed levels (gray dashed) and ReqRC ranks $j = 1, \ldots, 9$ (orange). Both panels share one predictor and window set, with $G = 0.00697$.

Under a fixed reservation limit, ReqRC raised window success from 83.16% to 91.41% and cut uncovered demand by 34%, at 4.2% higher mean reservations (Appendix C.9).

7 Conclusion

Long-run coverage can hold while individual windows keep falling short, and existing updates correct only the error of the level they already deployed. We propose ReqRC, which wraps any causal base method and feeds back the required level of each completed window, raising the base level to an order statistic of the delayed minimum feasible levels. Our theory uniquely decomposes shortfall into predictor and selection components and characterizes the highest-level selection probability required for uniform reliability. It also gives a finite-sample window success bound for the historical rank threshold, with a certification procedure developed in Appendix A.7. Empirically, ReqRC achieved the lowest mean shortfall and the highest window success rate against all ten alternatives and improved all 40 predictor–method pairs. On the test set every alternative matched to ReqRC's validation cost is less safe, and every alternative that matches its success rate is more expensive.

AI USE STATEMENT

We used generative AI tools to assist with Chinese–English translation, manuscript drafting and editing, literature exploration, discussion of theoretical formulations and proofs, experimental design, implementation and debugging of experimental code, interpretation of results, and preparation of figures and tables. The authors made the final decisions on the research claims, experimental protocols, and manuscript content.

The authors checked the mathematical arguments and cited sources, tested the experimental code, and verified the reported numerical results against the saved experimental outputs. The authors take responsibility for the final content of this work, including all AI-assisted text, claims, code, and figures.

REFERENCES

Anastasios N. Angelopoulos, Emmanuel J. Candès, and Ryan J. Tibshirani. Conformal PID control for time series prediction. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024.

Rina Foygel Barber and Ashwin Pananjady. Predictive inference for time series: Why is split conformal effective despite temporal dependence? In *Algorithmic Learning Theory*, volume 313 of *Proceedings of Machine Learning Research*, pp. 1–24, 2026.

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

Aadyot Bhatnagar, Huan Wang, Caiming Xiong, and Yu Bai. Improved online conformal prediction via strongly adaptive online learning. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2337–2363, 2023.

Kong Yao Chee, Thales C. Silva, M. Ani Hsieh, and George J. Pappas. Uncertainty quantification and robustification of model-based controllers using conformal prediction. In *Learning for Dynamics and Control*, volume 242 of *Proceedings of Machine Learning Research*, pp. 528–540, 2024.

Shai Feldman, Liran Ringel, Stephen Bates, and Yaniv Romano. Achieving risk control in online learning settings. *Transactions on Machine Learning Research*, 2023.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 2021.

Isaac Gibbs and Emmanuel Candès. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems*, volume 34, pp. 1660–1672, 2021.

Isaac Gibbs and Emmanuel J. Candès. Conformal inference for online prediction with arbitrary distribution shifts. *Journal of Machine Learning Research*, 25(162):1–36, 2024.

Jing Guo, Zihao Chang, Sa Wang, Haiyang Ding, Yihui Feng, Liang Mao, and Yungang Bao. Who limits the resource efficiency of my datacenter: An analysis of alibaba datacenter traces. In *2019 IEEE/ACM International Symposium on Quality of Service*, pp. 1–10, 2019.

Chirag Gupta, Arun K. Kuchibhotla, and Aaditya Ramdas. Nested conformal prediction and quantile out-of-bag ensemble methods. *Pattern Recognition*, 127:108496, 2022. doi: 10.1016/j.patcog.2021.108496.

Dongjian Hu, Junxi Wu, Shu-Tao Xia, and Changliang Zou. Distribution-informed online conformal prediction. In *International Conference on Learning Representations*, 2026.

Shayan Kiyani, George J. Pappas, Aaron Roth, and Hamed Hassani. Decision theoretic foundations for conformal prediction: Optimal uncertainty quantification for risk-averse agents. In *International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 30943–30965, 2025.

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

Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B*, 76(1):71–96, 2014.

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

Jordan Lekeufack, Anastasios N. Angelopoulos, Andrea Bajcsy, Michael I. Jordan, and Jitendra Malik. Conformal decision theory: Safe autonomous decisions from imperfect predictions. In *IEEE International Conference on Robotics and Automation*, pp. 11668–11675, 2024.

Ruipu Li and Alexander Rodríguez. Neural conformal control for time series forecasting. In *AAAI Conference on Artificial Intelligence*, 2025.

Daojun Liang. DistPred: A distribution-free probabilistic inference method for regression and forecasting. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 753–764, 2025. doi: 10.1145/3690624.3709286.

Bryan Lim, Sercan O. Arik, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4): 1748–1764, 2021.

Yaniv Romano, Evan Patterson, and Emmanuel Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

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

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421, 2008.

Kegan J. Strawn, Nora Ayanian, and Lars Lindemann. Conformal predictive safety filter for RL controllers in dynamic environments. *IEEE Robotics and Automation Letters*, 8(11):7833–7840, 2023. doi: 10.1109/LRA.2023.3322644.

Ya Su, Youjian Zhao, Chenhao Niu, Rong Liu, Wei Sun, and Dan Pei. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2828–2837, 2019.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Rahul Vaze. Simultaneous coverage and efficiency guarantee in online conformal prediction. *arXiv preprint arXiv:2607.26577*, 2026. doi: 10.48550/arXiv.2607.26577.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Junxi Wu, Dongjian Hu, Yajie Bao, Shu-Tao Xia, and Changliang Zou. Error-quantified conformal inference for time series. In *International Conference on Learning Representations*, 2025.

Taoran Wu, Jingduo Pan, Luke Ong, and Bai Xue. Cost-aware adaptive conformal inference for runtime assurance in dynamic environments. *arXiv preprint arXiv:2605.24463*, 2026. doi: 10.48550/arXiv.2605.24463.

Chen Xu and Yao Xie. Conformal prediction interval for dynamic time-series. In *International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 2021.

Chen Xu and Yao Xie. Sequential predictive conformal inference for time series. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 2023.

Margaux Zaffran, Olivier Feron, Yannig Goude, Julie Josse, and Aymeric Dieuleveut. Adaptive conformal predictions for time series. In *International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 25834–25866, 2022.

Matteo Zecchin and Osvaldo Simeone. Localized adaptive risk control. In *Advances in Neural Information Processing Systems*, 2024.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11106–11115, 2021.

APPENDIX A PROOFS

Appendices A–E contain proofs, algorithm and implementation details, full experimental results, synthetic studies, and reproducibility information. Proofs of the main results are in A.1–A.4; deployment certification is developed in A.6–A.7.

NOTATION

Notation is introduced where it is first used. Quantities defined in the main text retain their meanings.

$W_t = +\infty$ means that none of the current candidate forecasts satisfies the window-level coverage target; $+\infty$ is not a deployable resource level. Unless stated otherwise, every randomized method satisfies the causal-randomization condition in Section 3: conditional on information available before the decision, its internal random seed is independent of the current Y_t. Because W_t is a function of Y_t once the candidate forecasts are fixed, this condition also prevents method randomization from peeking at the current W_t.

A.1 PROOF OF PROPOSITION 1 AND THE SHORTFALL DECOMPOSITION

Exact worst-case bounds behind Proposition 1 The main text states the direct implication and its failure in reverse. Here we sharpen that statement by characterizing the largest possible local shortfall and the lowest possible window success rate under a satisfied long-run target.

Choose a window uniformly from the T test windows and write Z for its empirical coverage. Then

$$\mathbb{E}[Z] = \overline{\mathrm{Cover}}, \qquad \mathbb{E}[(\tau - Z)_+] = V, \qquad \mathbb{P}(Z \geq \tau) = \widehat{p}_\tau.$$

Because $x \mapsto (\tau - x)_+$ is convex, Jensen's inequality gives

$$V_{\mathrm{long}} = (\tau - \mathbb{E}[Z])_+ \leq \mathbb{E}[(\tau - Z)_+] = V.$$

The pointwise inequality $(\tau - z)_+ \leq \tau(1 - z)$ holds for every $z \in [0, 1]$. Hence, whenever $\overline{\mathrm{Cover}} \geq \tau$,

$$V \leq \tau(1 - \overline{\mathrm{Cover}}) \leq \tau(1 - \tau).$$

This bound is attained, over distributions on the empirical coverage grid, by placing probability $1 - \tau$ at $Z = 0$ and probability τ at $Z = 1$.

For the success bound, let $p := \mathbb{P}(Z \geq \tau)$ and $\tau_H^- = (\lceil H\tau \rceil - 1)/H$. Because Z lies on the grid $\{0, 1/H, \ldots, 1\}$, every window that misses the target has $Z \leq \tau_H^-$, whereas every window that meets the target has $Z \leq 1$. Therefore

$$\overline{\mathrm{Cover}} = \mathbb{E}[Z] \leq (1 - p)\tau_H^- + p,$$

and $\overline{\mathrm{Cover}} \geq \tau$ implies

$$p \geq \frac{\tau - \tau_H^-}{1 - \tau_H^-}.$$

The bound is attained by a two-point law on τ_H^- and 1 with the displayed mixing weight. For $H = 96$ and $\tau = 0.85$, an integer construction attains it without rounding: a window succeeds only if it covers 82 of the 96 points, since 81 gives $81/96 < 0.85$. Combining 96 such windows with four fully covered ones yields average coverage 0.85 and success rate 0.04. On a finite path, the required mixture weights may need integer rounding, which changes the attainable $\widehat{p}_\tau$ by at most $1/T$. Finally, $V = 0$ forces $Z \geq \tau$ for every window and thus implies the long-run target, while the sharp constructions prove that the converse fails.

Exact decomposition and comparisons Fix a nonempty level set $\mathcal{U}$ on which the minimum exists. By definition, $g_t = \min_{u\in\mathcal{U}} v_t(u) \geq 0$ and $r_t(u) = v_t(u) - g_t \geq 0$, so $v_t(u) = g_t + r_t(u)$. Setting $u = U_t$ and averaging gives $V = G + R$. Methods A and B using the same candidate forecasts and realized demand share g_t; subtracting their decompositions gives

$$v_t(U_t^A) - v_t(U_t^B) = r_t(U_t^A) - r_t(U_t^B).$$

These identities prove the decomposition, aggregation, and same-candidate comparison in Section 4.2.

Ordered forecasts and uniqueness When each $q_{t,h}(u)$ is nondecreasing, coverage is nondecreasing and the largest level minimizes shortfall. Thus

$$\begin{aligned} g_t &= v_t(u_{\max}) = [\tau - \mathrm{Cover}_t(u_{\max})]_+, \\ r_t(u) &= [\tau - \mathrm{Cover}_t(u)]_+ - [\tau - \mathrm{Cover}_t(u_{\max})]_+ \\ &= [\min\{\tau, \mathrm{Cover}_t(u_{\max})\} - \mathrm{Cover}_t(u)]_+. \end{aligned}$$

In particular, $g_t > 0$ if and only if $\mathrm{Cover}_t(u_{\max}) < \tau$, equivalently $W_t = +\infty$.

For uniqueness, suppose $v_t(u) = a_t + b_t(u)$ for every u, where a_t is independent of the selected level, both terms are nonnegative, and $b_t(u^\star) = 0$ at every minimizing level. Evaluating at any such $u^\star$ fixes both terms:

$$a_t = v_t(u^\star) = g_t, \qquad b_t(u) = v_t(u) - a_t = r_t(u).$$

Uniqueness is relative to the fixed window, realized demand, target, loss, and candidate set; $g_t > 0$ and $r_t(u) > 0$ may occur together.

Nested prediction families Suppose that an expanded candidate family indexed by $\mathcal{U}'$ preserves every forecast indexed by $\mathcal{U}$, and assume that the minima exist for both families. Minimizing over the larger family gives

$$g_t(\mathcal{U}') \leq g_t(\mathcal{U}).$$

If the same original level $u \in \mathcal{U}$ is deployed before and after expansion, its deployed shortfall is unchanged. Applying the exact decomposition in both spaces and subtracting gives

$$r_t(u;\mathcal{U}') - r_t(u;\mathcal{U}) = g_t(\mathcal{U}) - g_t(\mathcal{U}').$$

Thus, adding candidate forecasts lowers the predictor shortfall. If the method does not use any newly added forecast, however, the realized shortfall does not decrease.

A.2 Proof of Theorem 1: Feasibility and the Price of Reliability over All Admissible Distributions

Throughout this section, $\overline{\mathcal{U}} := [u_{\min}, u_{\max}] \cup \{+\infty\}$ is the extended set of minimum feasible levels.

Fix a realized history $\mathcal{F}_t = f_t$, and let $c_t(u, Y_t)$ denote a nonnegative cost that is nondecreasing in u for every realized Y_t. We suppress f_t in conditional expressions when no ambiguity arises.

Known conditional distribution Since $W_t \leq u$ is equivalent to meeting the coverage target, the highest level attains the largest possible success probability:

$$\sup_{u\in\mathcal{U}} \mathbb{P}(W_t \leq u \mid f_t) = \mathbb{P}(W_t < +\infty \mid f_t) = 1 - \mathbb{P}(W_t = +\infty \mid f_t).$$

A reliable finite deterministic level therefore exists exactly when $\mathbb{P}(W_t = +\infty \mid f_t) \leq 1 - \beta$. In that case, the conditional quantile

$$Q_\beta(W_t \mid f_t) := \inf\{u \in \overline{\mathcal{U}} : \mathbb{P}(W_t \leq u \mid f_t) \geq \beta\}$$

is finite and is the smallest reliable level. Every other reliable deterministic level u is at least this quantile, so monotonicity gives $c_t(u, Y_t) \geq c_t(Q_\beta(W_t \mid f_t), Y_t)$ almost surely. This characterizes the optimum when the conditional distribution is known; an online method need not know this quantile.

Endpoint-probability lower bound for reliability over all admissible distributions To formalize the admissible distributions in Theorem 1, consider all marginal laws of the minimum feasible level on $\overline{\mathcal{U}}$:

$$\mathcal{M}_\rho(f_t) = \left\{P \in \mathcal{P}(\overline{\mathcal{U}}) : P(\{+\infty\}) \le \rho\right\}.$$

The uniform statement ranges over this entire class of laws of W_t. In particular, it includes the two-point law used below; it is not restricted to the laws realizable by one fixed candidate family. If such a restriction is imposed instead, the same necessity argument requires that the restricted class contain that law. The cost conclusions below apply to compatible joint extensions with Y_t.

Let ν_t be the distribution of the level selected by a causal randomized method under history f_t, and write

$$\omega_t := \nu_t(\{u_{\max}\}).$$

If $\rho > 1 - \beta$, consider any admissible distribution that assigns mass ρ to $+\infty$. No finite level can reach the window coverage target when $W_t = +\infty$, so the worst-case window-level coverage probability is at most $1 - \rho < \beta$. Thus, no method restricted to finite deployable levels can be valid for all admissible distributions.

Now suppose that $\rho \le 1 - \beta$. Consider

$$P^\star = \rho\delta_{+\infty} + (1-\rho)\delta_{u_{\max}} \in \mathcal{M}_\rho(f_t).$$

Under this distribution, success is possible only when $W_t = u_{\max}$ and $U_t = u_{\max}$. Causal randomization makes U_t independent of the current W_t conditional on f_t, and therefore

$$\mathbb{P}_{P^\star}(W_t \le U_t \mid f_t) = (1-\rho)\omega_t.$$

Any policy whose actual window-level coverage probability is at least β for every distribution in $\mathcal{M}_\rho(f_t)$ must consequently satisfy

$$\omega_t \ge \frac{\beta}{1-\rho}.$$

Why the lower bound is sharp Let $\omega^\star := \beta/(1-\rho)$ and consider the randomized endpoint policy

$$U_t = \begin{cases} u_{\max}, & \text{with probability } \omega^\star, \\ u_{\min}, & \text{with probability } 1 - \omega^\star. \end{cases}$$

For every $P \in \mathcal{M}_\rho(f_t)$, retaining only the branch in which $u_{\max}$ is selected and W_t is finite gives

$$\mathbb{P}_P(W_t \le U_t \mid f_t) \ge \omega^\star P(W_t < +\infty) \ge \omega^\star(1-\rho) = \beta.$$

The lower bound is therefore achieved, yielding the exact characterization

$$\inf_{\pi:\, \inf_{P\in\mathcal{M}_\rho(f_t)} \mathbb{P}_P(W_t \le U_t^\pi \mid f_t) \ge \beta} \mathbb{P}(U_t^\pi = u_{\max} \mid f_t) = \frac{\beta}{1-\rho}.$$

Cost consequence under a compatible joint distribution The reliability lower bound requires only the marginal distribution of W_t. To analyze cost, take any conditional joint distribution of (W_t, Y_t) that is compatible with some $P \in \mathcal{M}_\rho(f_t)$, and define

$$\bar{c}_t(u) := \mathbb{E}[c_t(u, Y_t) \mid f_t].$$

By causal randomization, the selected level is conditionally independent of the current Y_t. If cost is nondecreasing in the selected level, every policy that is valid for all admissible distributions satisfies

$$\begin{aligned}\mathbb{E}[c_t(U_t, Y_t) \mid f_t] &\geq \omega_t \overline{c}_t(u_{\max}) + (1-\omega_t)\overline{c}_t(u_{\min}) \\ &\geq \frac{\beta}{1-\rho}\overline{c}_t(u_{\max}) + \left(1 - \frac{\beta}{1-\rho}\right)\overline{c}_t(u_{\min}).\end{aligned}$$

This conditional cost lower bound holds separately for every compatible joint extension; it does not incorporate the distribution of Y_t into the definition of $\mathcal{M}_\rho(f_t)$.

Deterministic methods For a deterministic method, ω_t can only be zero or one. When $\beta > 0$ and $\rho \leq 1 - \beta$, the necessary condition above forces $\omega_t = 1$, meaning that the method must select $u_{\max}$. This proves the deterministic consequence stated after Theorem 1.

A.3 Causal Decisions and Hindsight Benchmarks

This section explains why a level selected after observing the current Y_t can serve only as a diagnostic lower bound. Let $\ell_t(u, Y_t)$ be an integrable loss. The composite loss used in the main text may be written as

$$\ell_{t,\lambda}(u, Y_t) = c_t(u, Y_t) + \lambda v_t(u), \qquad \lambda \geq 0.$$

Assume that the relevant minima exist and admit measurable optimizing selections. Define the conditional hindsight risk

$$\mathrm{Risk}_t^{\mathrm{or}} := \mathbb{E}\left[\min_{u \in \mathcal{U}} \ell_t(u, Y_t) \,\middle|\, \mathcal{F}_t\right],$$

and the smallest conditional risk achievable using only information available before the decision,

$$\mathrm{Risk}_t^{\mathrm{ca}} := \operatorname*{ess\,inf}_{U \in L^0(\mathcal{F}_t; \mathcal{U})} \mathbb{E}[\ell_t(U, Y_t) \mid \mathcal{F}_t].$$

Here the essential infimum is taken over $\mathcal{F}_t$-measurable levels. Equivalently, after fixing a realized history f_t, it is the pointwise infimum of $\mathbb{E}[\ell_t(u, Y_t) \mid \mathcal{F}_t = f_t]$ over $u \in \mathcal{U}$. Independent randomization cannot lower this linear conditional expectation, so deterministic measurable level choices suffice.

Define the causal information gap and method suboptimality by

$$\begin{aligned}\mathrm{IG}_t &:= \mathrm{Risk}_t^{\mathrm{ca}} - \mathrm{Risk}_t^{\mathrm{or}}, \\ \mathrm{CR}_t^{\pi} &:= \mathbb{E}[\ell_t(U_t^{\pi}, Y_t) \mid \mathcal{F}_t] - \mathrm{Risk}_t^{\mathrm{ca}}.\end{aligned}$$

Proposition A.1 (Causal information-gap decomposition). For every causal policy with finite conditional risks, $\mathrm{IG}_t \geq 0$, $\mathrm{CR}_t^{\pi} \geq 0$, and the policy's conditional risk equals the hindsight risk plus these two gaps.

For every fixed level, the pointwise minimum loss is no greater than the loss at that level. Taking conditional expectations and then the infimum over levels chosen using the available information yields $\mathrm{Risk}_t^{\mathrm{or}} \leq \mathrm{Risk}_t^{\mathrm{ca}}$, so $\mathrm{IG}_t \geq 0$. The definition of $\mathrm{Risk}_t^{\mathrm{ca}}$ also gives $\mathrm{CR}_t^{\pi} \geq 0$. Adding and subtracting $\mathrm{Risk}_t^{\mathrm{ca}}$ in the deployed conditional risk gives

$$\mathbb{E}[\ell_t(U_t^{\pi}, Y_t) \mid \mathcal{F}_t] = \mathrm{Risk}_t^{\mathrm{or}} + \mathrm{IG}_t + \mathrm{CR}_t^{\pi}.$$

For this general loss, defining the pathwise regret

$$\mathrm{Reg}_t^{\pi} = \ell_t(U_t^{\pi}, Y_t) - \min_{u \in \mathcal{U}} \ell_t(u, Y_t),$$

we obtain

$$\mathbb{E}[\mathrm{Reg}_t^{\pi} \mid \mathcal{F}_t] = \mathrm{IG}_t + \mathrm{CR}_t^{\pi} .$$

If an $\mathcal{F}_t$-measurable level $u_t^{\star}$ achieves the pointwise minimum loss for almost every current outcome, then $\mathrm{IG}_t = 0$. Conversely, when a level minimizing the causal risk exists, $\mathrm{IG}_t = 0$ implies that the nonnegative random variable

$$\ell_t(u_t^{\star}, Y_t) - \min_{u \in \mathcal{U}} \ell_t(u, Y_t)$$

has conditional expectation zero and is therefore zero almost surely.

A convenient sufficient condition for strict positivity is the following. If $\mathcal{U}$ is finite and, for every $u \in \mathcal{U}$,

$$\mathbb{P}\left(\ell_t(u, Y_t) > \min_{a \in \mathcal{U}} \ell_t(a, Y_t) \,\middle|\, \mathcal{F}_t \right) > 0,$$

then every fixed decision has a strictly positive conditional expected gap from the hindsight optimum. The minimum of finitely many positive values remains positive, so $\mathrm{IG}_t > 0$. When the decisions form a continuum, analogous compactness, continuity, and uniform positive-separation conditions are required; variation of the apparent optimizer with Y_t alone does not establish strict positivity.

For the pure local-shortfall loss in the main text, monotonicity makes $u_{\max}$ uniformly risk-optimal for every Y_t. Hence

$$\mathrm{IG}_t = 0.$$

Knowing Y_t lets the hindsight rule choose, for the current window, the least costly level that satisfies the risk target. This advantage must not be interpreted as an improvement achievable by an online method under the original information set.

A.4 Proof That the Γ Bound in Theorem 2 Is Sharp

To keep the dependence on history length and feedback delay explicit, this appendix writes the main text's $W_t^{(j)}$ and ε_t as $W_{t,(j)}^{K,D}$ and $\varepsilon_{t,K,D}$, respectively. Recall $\overline{\mathcal{U}} = [u_{\min}, u_{\max}] \cup \{+\infty\}$. Define the vector of delayed historical levels and the current level, together with its joint law, by

$$\mathbf{W}_{t,K,D} := (W_{t-D-K+1}, \ldots, W_{t-D}, W_t) \in \overline{\mathcal{U}}^{K+1}, \qquad P_{t,K,D} := \mathcal{L}(\mathbf{W}_{t,K,D}).$$

Let $W_{t,(j)}^{K,D}$ denote the jth order statistic of the first K coordinates of $\mathbf{W}_{t,K,D}$, namely the delayed historical levels; the current coordinate W_t is not included in this order statistic.

Let Ex_{K+1} be the set of exchangeable laws on $\overline{\mathcal{U}}^{K+1}$, and define

$$\Gamma_{t,K,D} := \inf_{Q \in \mathsf{Ex}_{K+1}} d_{\mathrm{TV}}(P_{t,K,D}, Q).$$

For $i = 1, \ldots, K+1$, let σ_i swap coordinate i with the current coordinate $K+1$, with σ_{K+1} the identity. The average coordinate-swap distance and the deviation term used in Theorem 2 are

$$\Delta_{t,K,D}^{\mathrm{swap}} := \frac{1}{K+1} \sum_{i=1}^{K+1} d_{\mathrm{TV}}\left(P_{t,K,D}, P_{t,K,D} \circ \sigma_i^{-1}\right), \qquad \varepsilon_{t,K,D} := \min\{\Gamma_{t,K,D}, \Delta_{t,K,D}^{\mathrm{swap}}\}.$$

Let

$$E_j := \{W_t \le W_{t,(j)}^{K,D}\}, \qquad p_j := \frac{j}{K+1}.$$

For the appendix proofs only, define the rank-threshold class

$$\mathfrak{C}_{t,j} := \left\{ U_t \in \mathcal{U} : \begin{array}{ll} U_t \ge W_{t,(j)}^{K,D}, & W_{t,(j)}^{K,D} < +\infty, \\ U_t = u_{\max}, & W_{t,(j)}^{K,D} = +\infty \end{array} \right\},$$

where level choices are required to satisfy the causal protocol in the main text.

Event inclusion Fix any $U_t \in \mathfrak{C}_{t,j}$. If $W_{t,(j)}^{K,D} < +\infty$ and E_j occurs, then

$$U_t \ge W_{t,(j)}^{K,D} \ge W_t,$$

so the selected level satisfies the window-level coverage target and $g_t = r_t(U_t) = 0$. If the historical threshold equals $+\infty$, a method in $\mathfrak{C}_{t,j}$ selects $u_{\max}$ by definition. When $W_t < +\infty$, the highest level reaches the window coverage target; when $W_t = +\infty$, it achieves the minimum shortfall among the candidate forecasts. Both branches therefore imply

$$E_j \subseteq \{r_t(U_t) = 0\}.$$

Exchangeable rank bound and TV transfer For every $Q \in \mathsf{Ex}_{K+1}$, exchangeability gives

$$Q(E_j) \ge \frac{j}{K+1} = p_j.$$

Equality holds when all $K+1$ values are continuous, finite, and distinct. Ties or atoms at $+\infty$ can only increase the probability when the event is defined using "$\le$' '. Leveraging the standard total-variation transfer inequality for events gives

$$P_{t,K,D}(E_j) \ge Q(E_j) - d_{\mathrm{TV}}(P_{t,K,D}, Q) \ge p_j - d_{\mathrm{TV}}(P_{t,K,D}, Q).$$

This inequality holds for every exchangeable Q. Taking the infimum in the distance term and using the event inclusion gives

$$\mathbb{P}(r_t(U_t) = 0) \ge [p_j - \Gamma_{t,K,D}]_+ .$$

Why the bound is sharp over a TV neighborhood The argument concerns pure rank rules, the subclass of continuous and finite minimum feasible levels, and the setting in which a lower bound imposed by a base prediction level is inactive. Choose a continuous exchangeable distribution Q such that $Q(E_j) = p_j$ and the pure rank rule satisfies

$$\{r_t = 0\} = E_j \qquad Q\text{-a.s.}$$

For $0 \le \delta \le p_j$, define a tilted distribution relative to Q by

$$\frac{\mathrm{d}P_\delta}{\mathrm{d}Q} = \left(1 - \frac{\delta}{p_j}\right) \mathbf{1}_{E_j} + \left(1 + \frac{\delta}{1-p_j}\right) \mathbf{1}_{E_j^c}.$$

The density is nonnegative and integrates to one, and it satisfies

$$d_{\mathrm{TV}}(P_\delta, Q) = \delta, \qquad P_\delta(E_j) = p_j - \delta.$$

Define the TV neighborhood

$$\mathfrak{P}(\gamma) := \left\{ P : \inf_{Q' \in \mathsf{Ex}_{K+1}} d_{\mathrm{TV}}(P, Q') \le \gamma \right\}.$$

The Γ branch of Theorem 2 gives the uniform lower bound $[p_j - \gamma]_+$ over this neighborhood. For $\gamma \le p_j$, set $\delta = \gamma$; for $\gamma > p_j$, set $\delta = p_j$. These constructions attain $p_j - \gamma$ and zero, respectively. Hence

$$\inf_{P \in \mathfrak{P}(\gamma)} \mathbb{P}_P(r_t = 0) = [p_j - \gamma]_+.$$

Probability of meeting the window target When E_j occurs and the current W_t is finite, the level at the delayed historical threshold satisfies the window coverage target. Therefore,

$$E_j \cap \{W_t < +\infty\} \subseteq \{\mathrm{Cover}_t(U_t) \ge \tau\}.$$

Let $m = K + 1$, $p_j = j/m$, and $X = (X_1, \dots, X_m)$ be the delayed-history/current vector. For each coordinate i, let $X_{-i,(j)}$ be the j-th order statistic with coordinate i excluded, and define

$$S_i(X) := \mathbf{1}\{X_i < +\infty,\ X_i \le X_{-i,(j)}\}, \qquad H_i(X) := S_i(X) + p_j \mathbf{1}\{X_i = +\infty\}.$$

If n coordinates are finite, at least $\min(j, n)$ satisfy the event in S_i, including ties. Hence

$$\sum_{i=1}^{m} H_i \ge \min(j, n) + p_j(m - n) \ge j.$$

Since $0 \le H_i \le 1$, every exchangeable law Q satisfies $\mathbb{E}_Q H_m \ge p_j$. TV transfer and the definition of $\Gamma_{t,K,D}$ give $\mathbb{E}_P H_m \ge p_j - \Gamma_{t,K,D}$. Also, swapping coordinates i, m sends H_m to H_i, so

$$\mathbb{E}_P H_m \ge \frac{1}{m} \sum_i \mathbb{E}_P H_i - \frac{1}{m} \sum_i d_{\mathrm{TV}}(P, P \circ \sigma_i^{-1}) \ge p_j - \Delta^{\mathrm{swap}}_{t,K,D}.$$

Combining the two inequalities and using $\mathbb{E}_P H_m = \mathbb{P}_P(S_m = 1) + p_j \rho_t$, together with the event inclusion above, proves

$$\mathbb{P}(\mathrm{Cover}_t(U_t) \ge \tau) \ge [p_j(1 - \rho_t) - \varepsilon_{t,K,D}]_+.$$

No independence assumption is added. The earlier union-bound expression remains valid but is weaker by $(1 - p_j)\rho_t$ before clipping.

For sharpness over the abstract joint-law class, put mass ρ on the all-infinite vector and distribute the remaining mass uniformly over permutations of m distinct finite levels. For $0 \le \gamma \le p_j(1 - \rho)$, transfer total mass γ uniformly from permutations with current rank at most j to those with current rank greater than j. The resulting success probability for the pure capped rank rule is $p_j(1 - \rho) - \gamma$. Its distance to the original exchangeable law is γ; the H_m bound gives the matching lower bound on its distance to every exchangeable law, so $\Gamma = \gamma$. Directly counting swaps across the two rank groups gives $\Delta^{\mathrm{swap}} = \gamma$. At $\gamma = p_j(1 - \rho)$, success is zero, also attaining the clipped bound in any larger defect neighborhood. This sharpness concerns the pure threshold rule with inactive base protection; extra base protection may improve success.

A.5 Coordinate-Swap and Deletion–Switch Bounds for Time Series

This section reuses the event inclusion and TV transfer in Appendix A.4. We give the coordinate-swap argument explicitly, then state the deletion–switch extension and time-series specializations.

Let $m := K + 1$, $X := \mathbf{W}_{t,K,D}$, and $E_j := \{X_m \le X_{-m,(j)}\}$, where $X_{-i,(j)}$ is the jth order statistic of all coordinates except i. Section A.4 proves $E_j \subseteq \{r_t(U_t) = 0\}$ for every $U_t \in \mathfrak{C}_{t,j}$.

Coordinate-swap bound. Let σ_i swap coordinates i and m, $P^{(i)} := P \circ \sigma_i^{-1}$, and

$$\Delta^{\mathrm{swap}}_{t,K,D} := \frac{1}{m} \sum_{i=1}^{m} d_{\mathrm{TV}}(P, P^{(i)}).$$

Then

$$\mathbb{P}(r_t(U_t) = 0) \geq \mathbb{P}(E_j) \geq \left[\frac{j}{m} - \Delta^{\text{swap}}_{t,K,D}\right]_+ .$$

Proof. For any fixed vector x, moving any of its j smallest coordinates to the current position makes E_j hold; ties can only add successes. Thus $\sum_{i=1}^m \mathbf{1}_{E_j}(\sigma_i x) \geq j$, and integration gives $\sum_i P^{(i)}(E_j) \geq j$. Averaging $P(E_j) \geq P^{(i)}(E_j) - d_{\text{TV}}(P, P^{(i)})$ yields the bound. The event inclusion from A.4 then gives the result for $r_t(U_t)$. □

Combining this bound with the Γ branch gives the main-text penalty

$$\varepsilon_{t,K,D} = \min\{\Gamma_{t,K,D}, \Delta^{\text{swap}}_{t,K,D}\}.$$

Sharp bound without structure Without further structure the best uniform defect is

$$\sup_P \min\{\Gamma(P), \Delta^{\text{swap}}(P)\} = \frac{K}{K+1}. \tag{A.23}$$

Proof. The identity swap contributes zero and every other TV distance is at most one, so $\Delta^{\text{swap}} \leq (m-1)/m$. For equality, take $P = \delta_x$, where the current coordinate differs from the other $m-1$ identical coordinates. Its permutation orbit has m points; an exchangeable law assigns mass at most $1/m$ to x, attained by the uniform law on that orbit. Hence $\Gamma(P) = 1 - 1/m$. Each nonidentity swap also moves the point mass, giving $\Delta^{\text{swap}}(P) = 1 - 1/m$. □ Thus certifying a method before deployment requires identifiable temporal, contextual, or transport structure.

Deletion–switch bound. We use the deletion maps of Barber & Pananjady (2026, Definitions 1–2). For a vector x of length m, write $x_{a:b} = (x_a, \ldots, x_b)$, with an empty block omitted. For $0 \leq \ell \leq j-1$ and $1 \leq k \leq m-1-\ell$, define

$$D^0_{k,\ell} x = (x_{1:m-\ell-k}, x_{m-k+1:m}), \qquad D^1_{k,\ell} x = (x_{k+\ell+1:m}, x_{1:k}).$$

For $m - \ell \leq k \leq m$, define $D^0_{k,\ell} x = x_{\ell+1:m}$ and $D^1_{k,\ell} x = x_{k-m+\ell+1:k}$. Both vectors have $m-\ell$ entries; their last entries are x_m and x_k, respectively. Set

$$\overline{\Psi}_\ell(X) := \frac{1}{m}\sum_{k=1}^m d_{\text{TV}}\left(\mathcal{L}(D^0_{k,\ell} X), \mathcal{L}(D^1_{k,\ell} X)\right).$$

Then

$$\mathbb{P}(r_t(U_t) = 0) \geq \left[\frac{j}{m} - \min_{0 \leq \ell \leq j-1}\left\{\frac{\ell}{m} + \overline{\Psi}_\ell(X)\right\}\right]_+ .$$

Proof. For a reduced vector z, let E' be the event that its last entry is no greater than the $(j-\ell)$th smallest of its other entries. The current coordinate is excluded from that order statistic. If E' holds for $D^0_{k,\ell} x$, at most $j-\ell-1$ retained entries and ℓ deleted entries are strictly below x_m; hence E_j holds in the full vector. Moreover, each of the $j-\ell$ smallest entries of x, when placed last by its corresponding D^1, satisfies E'. Thus $\sum_k \mathbf{1}_{E'}(D^1_{k,\ell} x) \geq j - \ell$, including ties. Apply the TV event inequality to each pair (D^0, D^1), average over k, and obtain $P(E_j) \geq (j-\ell)/m - \overline{\Psi}_\ell(X)$. Optimize over ℓ and use the event inclusion from A.4. □

Conditioning on information available before calibration. Let $\mathcal{C}_t \subseteq \mathcal{F}_{t-D-K+1}$ be information fixed before the earliest calibration window begins. For almost every $\mathcal{C}_t = c$, apply A.4 and the coordinate-swap proof above to the conditional law $\mathcal{L}(X \mid \mathcal{C}_t = c)$. Their pointwise event inclusion and counting argument are unchanged, giving

$$\mathbb{P}(r_t(U_t) = 0 \mid \mathcal{C}_t = c) \geq \left[\frac{j}{K+1} - \min\{\Gamma_{t,K,D}(c), \Delta^{\text{swap}}_{t,K,D}(c)\}\right]_+ .$$

This is weaker than the history-wise uniform guarantee of Section 4.3, which conditions on the complete realized history.

Stationary mixing and independent drift. For a stationary process of minimum feasible levels, use the deleted-gap convention

$$\beta_W(\ell) := \sup_a d_{\text{TV}}(\mathcal{L}(W_{s:s\leq a}, W_{s:s\geq a+\ell+1}), \mathcal{L}(W_{s:s\leq a}) \otimes \mathcal{L}(W_{s:s\geq a+\ell+1})).$$

Thus $\ell = 0$ separates adjacent blocks. For a contiguous vector of length $M = K + D$, the first $M - 1 - \ell$ pairs of deletion maps each compare two blocks separated by ℓ deleted entries. Replacing each joint law by the product of its block marginals costs at most $\beta_W(\ell)$; stationarity makes the two products identical. The remaining pairs have identical laws by stationarity. Consequently $\overline{\Psi}_\ell \le 2(M - 1 - \ell)\beta_W(\ell)/M$, as in Barber & Pananjady (2026, Proposition 1).

Apply this bound to $(W_{t-D-K+1}, \dots, W_t)$. Its $M - 1$ historical entries include the K available entries and $D - 1$ unavailable recent entries. The jth order statistic of all $M - 1$ entries is no larger than that of the available K. Success of the full-block rank event therefore implies success of the delayed rank event, giving

$$\mathbb{P}(r_t(U_t) = 0) \ge \left[\frac{j}{K+D} - \min_{0 \le \ell \le j-1} \left\{ \frac{\ell}{K+D} + 2\frac{K+D-1-\ell}{K+D}\beta_W(\ell) \right\} \right]_+ .$$

Under independent coordinates with adjacent marginal TV distance at most $d_{\max}^{\mathrm{adj}}$,

$$\mathbb{P}(r_t(U_t) = 0) \ge \left[\frac{j}{K+1} - \frac{2d_{\max}^{\mathrm{adj}}}{K+1} \left(KD + \frac{K(K-1)}{2} \right) \right]_+ .$$

Indeed, swapping two independent coordinates changes their product law by at most twice the TV distance between their marginals. Summing their time separations $D, D+1, \dots, D+K-1$ in the coordinate-swap bound proves the displayed drift bound.

Separating dependence and marginal drift. For historical indices $s_1 < \cdots < s_K$ and $s_m = t$, let $X_i = W_{s_i}$, $P_i = \mathcal{L}(X_i)$, and

$$B_{\mathrm{dep}} := d_{\mathrm{TV}} \left(\mathcal{L}(X), \bigotimes_{i=1}^m P_i \right), \qquad B_{\mathrm{drift}} := \inf_\mu \sum_{i=1}^m d_{\mathrm{TV}}(P_i, \mu).$$

The product law $\mu^{\otimes m}$ is exchangeable. The triangle and product-TV inequalities give $\Gamma(\mathcal{L}(X)) \le B_{\mathrm{dep}} + B_{\mathrm{drift}}$, so A.4 gives $P(r_t = 0) \ge [j/m - B_{\mathrm{dep}} - B_{\mathrm{drift}}]_+$. This bound also applies to sparse historical indices. In the stationary mixing case,

$$B_{\mathrm{dep}} \le \sum_{i=1}^{m-1} \beta_W(s_{i+1} - s_i - 1), \qquad B_{\mathrm{drift}} = 0.$$

To see this, split off the last coordinate, bound its dependence on the retained prefix by the indicated mixing coefficient, and repeat on the prefix. Data processing bounds each retained-prefix distance by that of the full past and future blocks. All these are bounds on the law of the minimum feasible levels, not estimates from a realized sequence.

A.6 Choosing j, Fallback, and Minimum-Update Proofs

Throughout this section, $\beta_{\mathrm{sel}} \in (0, 1]$ denotes the target probability for the event $\{r_t = 0\}$. We state the three results needed by the adaptive extension and give short proofs.

Choosing j from jointly valid bounds Let $\mathcal{H}_t^{\mathrm{sel}}$ be a preselection information sigma-field fixed before deployment. Suppose the $\mathcal{H}_t^{\mathrm{sel}}$-measurable bounds $\underline{p}_{t,j}^{(b)}$ satisfy, simultaneously over all constructions b and ranks j,

$$\mathbb{P}\big(r_t(\Phi_{t,j}(A_t)) = 0 \mid \mathcal{H}_t^{\mathrm{sel}}\big) \ge \underline{p}_{t,j}^{(b)} \qquad \text{a.s.},$$

and set $\underline{p}_{t,j}^{\mathrm{sel}} := \max_b \underline{p}_{t,j}^{(b)}$. A fixed-rank guarantee is not generally reusable after selecting j from the same block; simultaneous validity over all candidate ranks is the required correction. If an $\mathcal{H}_t^{\mathrm{sel}}$-measurable upper bound satisfies $\mathbb{P}(W_t = +\infty \mid \mathcal{H}_t^{\mathrm{sel}}) \le \overline{\rho}_t$, define

$$j_t^{\mathrm{cert}} := \min \left\{ j \in \{1, \dots, K\} : [\underline{p}_{t,j}^{\mathrm{sel}} - \overline{\rho}_t]_+ \ge \beta \right\}.$$

The subtraction is the conservative union-bound form. Appendix A.4 also derives a coupled form that is larger by $(1 - p_j)\overline{\rho}_t$ whenever a drift bound is available, where $p_j = j/(K+1)$ is the nominal rank

probability. For each fixed j, reuse the event subtraction in A.4 conditional on $\mathcal{H}_t^{\mathrm{sel}}$. The resulting bounds hold on one common probability-one set, so the measurable choice j_t^{cert} can be substituted without changing that set. Whenever a finite rank passes, this gives

$$\mathbb{P}\big(\mathrm{Cover}_t(\Phi_{t,j_t^{\mathrm{cert}}}(A_t)) \geq \tau \mid \mathcal{H}_t^{\mathrm{sel}}\big) \geq \beta.$$

If no finite rank passes, the rule uses the maximum-level branch, which certifies the target when $1 - \overline{\rho}_t \geq \beta$; otherwise no guarantee is issued.

Randomized maximum-level fallback An optional extension, not used by main ReqRC, deploys $u_{\max}$ with independent probability ω and otherwise uses the rank-j update. Because the maximum level always gives $r_t = 0$, conditioning on the independent coin gives the lower bound $\omega + (1-\omega)\underline{p}_{t,j}^{\mathrm{sel}}$. Solving for this bound to reach β_{sel} gives the minimum sufficient probability

$$\omega_j^\star = \begin{cases} 0, & \underline{p}_{t,j}^{\mathrm{sel}} \geq \beta_{\mathrm{sel}}, \\ \dfrac{\beta_{\mathrm{sel}} - \underline{p}_{t,j}^{\mathrm{sel}}}{1 - \underline{p}_{t,j}^{\mathrm{sel}}}, & \underline{p}_{t,j}^{\mathrm{sel}} < \beta_{\mathrm{sel}}. \end{cases}$$

For a window-level target, set $\underline{p}_{t,j}^{\mathrm{win}} := [\underline{p}_{t,j}^{\mathrm{sel}} - \overline{\rho}_t]_+$ and $\underline{p}_{t,\max}^{\mathrm{win}} := 1 - \overline{\rho}_t$; if $\underline{p}_{t,j}^{\mathrm{win}} < \beta \leq \underline{p}_{t,\max}^{\mathrm{win}}$, the analogous mixture probability is $(\beta - \underline{p}_{t,j}^{\mathrm{win}})/(\underline{p}_{t,\max}^{\mathrm{win}} - \underline{p}_{t,j}^{\mathrm{win}})$.

Smallest deployed level and rank ordering Let $\mathcal{H}_t^D$ denote the information available before deployment with delay D, including the base level A_t and the available historical levels. Appendix A.7 gives an explicit filtration construction.

The update $\Phi_{t,j}(A_t)$ equals $\max\{A_t, W_{t,(j)}^{K,D}\}$ when the threshold is finite and $u_{\max}$ when it is $+\infty$. For a finite threshold, any admissible level must be at least as large as both entries of this maximum; for an infinite threshold, the requirement fixes the level at $u_{\max}$. The update is therefore the smallest $\mathcal{H}_t^D$-measurable level satisfying both requirements, and minimizes any nondecreasing cost within this class.

For any $j_1 \leq j_2 \leq K$, the order statistics satisfy $W_{t,(j_1)}^{K,D} \leq W_{t,(j_2)}^{K,D}$, so

$$\Phi_{t,j_1}(A_t) \leq \Phi_{t,j_2}(A_t).$$

Write $\Phi_1 = \Phi_{t,j_1}(A_t)$ and $\Phi_2 = \Phi_{t,j_2}(A_t)$. Monotonicity of the candidate family and cost then gives, pathwise,

$$v_t(\Phi_2) \leq v_t(\Phi_1), \qquad r_t(\Phi_2) \leq r_t(\Phi_1), \qquad c_t(\Phi_2, Y_t) \geq c_t(\Phi_1, Y_t).$$

The predictor shortfall g_t is identical across ranks. This establishes the ordering of the $j = 1, \ldots, K$ risk–cost path; it does not claim global optimality among all policies.

Minimality of the rank threshold over the exchangeable class The preceding paragraph establishes minimality inside the class that retains the base level and meets the historical threshold. Corollary 1 records that result. Outside that class the claim fails: a rule that uses fewer than K historical levels can be both smaller and still valid, so the rank threshold is not the pointwise cheapest history-measurable choice.

We consider the class of $\mathcal{H}_t^D$-measurable levels $U_t \in \mathcal{U}$ satisfying $Q(W_t \leq U_t) \geq \beta$ for every $Q \in \mathsf{Ex}_{K+1}$, with the probability taken over the joint law of the delayed history and the current value, and assume that the $K+1$ delayed-history/current values are continuous and distinct.

Proof of Corollary 1. *Admissibility.* Appendix A.4 gives $Q(E_j) = p_j \geq \beta$ for every $Q \in \mathsf{Ex}_{K+1}$, and Appendix A.4 gives $E_j \subseteq \{r_t(U_t) = 0\}$ for every $U_t \in \mathfrak{C}_{t,j}$. Since $W_{t,(j)}^{K,D} \in \mathfrak{C}_{t,j}$ whenever it is finite, it is admissible; the value $W_{t,(j)}^{K,D} = +\infty$ deploys $u_{\max}$, which is the infeasibility event quantified in Theorem 2. *Necessity within the class.* Every admissible level in $\mathfrak{C}_{t,j}$ is at least $W_{t,(j)}^{K,D}$ by the definition of $\mathfrak{C}_{t,j}$, so $\max\{A_t, W_{t,(j)}^{K,D}\}$ is the smallest element of the class. □

Combining both directions, $\max\{A_t, W^{K,D}_{t,(j)}\}$ is the cheapest admissible level among the rules that retain the base level A_t and meet the historical threshold $W^{(j)}_t$: within that class the threshold is determined by the reliability target, and the base level is the only remaining degree of freedom. A smaller level can be admissible only by using less information, such as dropping a historical sample, or by knowing the conditional law, as in the conditional β-quantile $(F^D_t)^{-1}(\beta)$ of Appendix A.7, which presumes knowledge of the conditional law that the exchangeable class does not supply.

A.7 Conditional Window Coverage Under a Drift Bound

The idea is to estimate a reference success probability from revealed historical windows, then subtract an allowance for how much the current distribution may differ from that reference. The historical estimate and the drift allowance require separate justification. This section states their validity conditions and proves the resulting conditional guarantee.

Let $\mathcal{F}^-_s \subseteq \mathcal{F}^+_s \subseteq \mathcal{F}^-_{s+1}$ be a filtration with W_s measurable with respect to $\mathcal{F}^+_s$. It describes the mathematical evolution of the process, not just the delayed observations available to the method. Let $\mathcal{D}_0 \subseteq \mathcal{F}^-_1$ contain the design choices frozen before the calibration observations, and let Z_t collect the covariates, candidate forecasts, and base level available before window t. Define the method's information by

$$\mathcal{H}^D_t := \sigma(\mathcal{D}_0, Z_{1:t}, W_{1:(t-D)_+}) \subseteq \mathcal{F}^-_t, \qquad F^D_t(u) := \mathbb{P}(W_t \le u \mid \mathcal{H}^D_t).$$

Set $\rho^D_t := 1 - F^D_t(u_{\max})$. Let $\mathcal{U}_{\mathrm{grid}}$ be a finite calibration grid and $\mathfrak{Q}_t$ a finite prespecified query family. For each query q and grid point z, the reference lower bound $\underline{F}^{\mathrm{ref}}_{tq}(z)$, the gate $I^{\mathrm{tr}}_{tq} \in \{0,1\}$, and the nonnegative drift allowance $\overline{\kappa}^{\mathrm{tr}}_{tq}(z)$ are $\mathcal{H}^D_t$-measurable.

Information available before deployment and common validity events Let $F^{\mathrm{ref}}_{tq}(z)$ be a reference CDF, with a concrete construction below. Assume that, simultaneously over all protected times, streams, queries, and grid points, the events

$$\mathcal{E}_{\mathrm{stat}} := \{\underline{F}^{\mathrm{ref}}_{tq}(z) \le F^{\mathrm{ref}}_{tq}(z) \text{ for all } t, q, z\},$$
$$\mathcal{E}_{\mathrm{tr}} := \{I^{\mathrm{tr}}_{tq} = 1 \implies F^D_t(z) \ge F^{\mathrm{ref}}_{tq}(z) - \overline{\kappa}^{\mathrm{tr}}_{tq}(z) \text{ for all } t, q, z\}$$

have probability at least $1 - \delta_{\mathrm{stat}}$ and $1 - \delta_{\mathrm{tr}}$, respectively. Stream indices are suppressed. It follows on their intersection that, simultaneously over levels,

$$F^D_t(u) \ge \underline{p}^{\mathrm{win}}_t(u), \qquad \rho^D_t \le 1 - \underline{p}^{\mathrm{win}}_t(u_{\max}), \tag{A.24}$$

where the monotone left-grid extension is

$$\underline{p}^{\mathrm{win}}_t(u) := \min\Big\{1, \max\Big(\{0\} \cup \{[\underline{F}^{\mathrm{ref}}_{tq}(z) - \overline{\kappa}^{\mathrm{tr}}_{tq}(z)]_+ : q \in \mathfrak{Q}_t, I^{\mathrm{tr}}_{tq} = 1, z \le u\}\Big)\Big\}.$$

Theorem A.7.1 (Conditional window-level coverage guarantee under one-sided drift). For any finite family of $\mathcal{H}^D_t$-measurable candidates $U_{t,1}, \dots, U_{t,J}$, on the common event and simultaneously over j,

$$\mathbb{P}\big(\mathrm{Cover}_t(U_{t,j}) \ge \tau \mid \mathcal{H}^D_t\big) = F^D_t(U_{t,j}) \ge \underline{p}^{\mathrm{win}}_t(U_{t,j}). \tag{A.25}$$

Selecting the smallest candidate with $\underline{p}^{\mathrm{win}}_t(U_{t,j}) \ge \beta$, or falling back to $u_{\max}$, yields a certificate Cert_t satisfying

$$\mathbb{P}\big(\forall t : \mathsf{Cert}_t = 1 \Longrightarrow \mathbb{P}(\mathrm{Cover}_t(U_t) \ge \tau \mid \mathcal{H}^D_t) \ge \beta\big) \ge 1 - \delta_{\mathrm{stat}} - \delta_{\mathrm{tr}}. \tag{A.26}$$

Proof. On the common event, each enabled pair (q, z) with $z \le u$ satisfies

$$F^D_t(u) \ge F^D_t(z) \ge \underline{F}^{\mathrm{ref}}_{tq}(z) - \overline{\kappa}^{\mathrm{tr}}_{tq}(z).$$

Taking the maximum over these lower bounds and clipping at zero gives equation A.24. Since it holds simultaneously in u, it also holds at each candidate $U_{t,j}$ chosen using information available before the decision, proving equation A.25. As in A.6, set $\mathsf{Cert}_t = 1$ only when the selected candidate, including any maximum-level fallback, passes the lower-bound test. The implication in equation A.26 then holds on the common event; a union bound on its complement completes the proof. □

Computable reference bound For grid point z, set $X_s(z) := \mathbf{1}\{W_s \le z\}$ and $F_s^{\mathrm{pre}}(z) := \mathbb{E}[X_s(z) \mid \mathcal{F}_s^-]$. For each (t,q), choose a nonnegative weight vector $(a_{tq,s})_s$, measurable with respect to $\mathcal{D}_0$, with support in $s \le t-D$ and sum one. Thus its entire support and weights are fixed before the calibration outcomes, rather than selected after inspecting them. Define

$$\widehat{F}_{tq}^{\mathrm{ref}}(z) := \sum_s a_{tq,s} X_s(z), \qquad F_{tq}^{\mathrm{ref}}(z) := \sum_s a_{tq,s} F_s^{\mathrm{pre}}(z),$$

and

$$\underline{F}_{tq}^{\mathrm{ref}}(z) := \left[\widehat{F}_{tq}^{\mathrm{ref}}(z) - \sqrt{\frac{\sum_s a_{tq,s}^2}{2} \log \frac{1}{\delta_{tqz}}} \right]_+, \tag{A.27}$$

where the positive δ_{tqz} are allocated in advance to sum to at most δ_{stat}, including every protected stream. For $\lambda > 0$, conditional Hoeffding gives

$$\mathbb{E}\big[\exp\{\lambda a_{tq,s}(X_s(z) - F_s^{\mathrm{pre}}(z))\} \mid \mathcal{F}_s^-\big] \le \exp\{\lambda^2 a_{tq,s}^2/8\}.$$

Iterating conditional expectations through the filtration and applying Chernoff's bound yields, conditional on $\mathcal{D}_0$,

$$\mathbb{P}\Big(\widehat{F}_{tq}^{\mathrm{ref}}(z) - F_{tq}^{\mathrm{ref}}(z) \ge r \mid \mathcal{D}_0\Big) \le \exp\left\{-\frac{2r^2}{\sum_s a_{tq,s}^2}\right\}.$$

Substitute the radius in equation A.27 and take a union bound to establish $\mathcal{E}_{\mathrm{stat}}$. Selecting among the prespecified queries is permitted because all their bounds hold jointly. Choosing uncorrected weights after observing the outcomes is not covered by this argument. The drift event $\mathcal{E}_{\mathrm{tr}}$ still requires the separate assumption above.

Pairwise drift and delay correction **Corollary A.7.2 (Pairwise drift and telescoping).** If, on an event of probability at least $1 - \delta_{\mathrm{tr}}$,

$$F_t^D(z) \ge F_s^{\mathrm{pre}}(z) - \kappa_{s\to t}^{\mathrm{pair}}(z)$$

for every enabled query and every s in its support, then the correction

$$\overline{\kappa}_{tq}^{\mathrm{tr}}(z) := \sum_s a_{tq,s} \kappa_{s\to t}^{\mathrm{pair}}(z) \tag{A.28}$$

validates $\mathcal{E}_{\mathrm{tr}}$: multiply each pairwise inequality by its nonnegative weight and sum, using $\sum_s a_{tq,s} = 1$. Telescoping adjacent one-sided drift envelopes gives $\kappa_{s\to t}^{\mathrm{pair}}(z) \le (t-s)\kappa_{\max}^{\mathrm{step}}(z)$, and averaging gives the delay correction

$$\overline{\kappa}_{tq}^{\mathrm{tr}}(z) \le \kappa_{\max}^{\mathrm{step}}(z) \sum_s a_{tq,s}(t-s), \tag{A.29}$$

where $t-s$ is the delay of historical window s. For uniform weights on $s = t-D-n+1, \dots, t-D$, the sum equals $D + (n-1)/2$, where n is the number of weights in the query support.

Impossibility under an unobserved change **Proposition A.7.3 (Impossibility under delayed observation of a change).** Let P_0 and P_ν agree on all information available before the decision for the first D windows after a distribution change and then place mass one on $W_t = +\infty$ under P_ν. Any certification rule whose probability of issuing a false certificate is at most δ_{cert} satisfies

$$\mathbb{P}_{P_0}(\mathsf{Cert}_t = 1) \le \delta_{\mathrm{cert}} \tag{A.30}$$

during those delayed windows. Proof sketch: under P_ν no finite level succeeds, so a certificate is false whenever issued; observational equivalence transfers the same event probability to P_0. □

Why the pointwise correction is sharp Fix a finite $u_0 \in \mathcal{U}$ and $\ell, \kappa_{\mathrm{KS}} \in [0,1]$. Consider all pairs of CDFs (F^0, F) of laws on $\overline{\mathcal{U}}$ satisfying $F^0(u_0) \ge \ell$ and $\sup_{u\in\mathcal{U}}[F^0(u) - F(u)]_+ \le \kappa_{\mathrm{KS}}$. The pointwise minimax lower bound over this class is

$$\inf F(u_0) = [\ell - \kappa_{\mathrm{KS}}]_+, \tag{A.31}$$

Proof. The neighborhood condition gives $F(u_0) \ge \ell - \kappa_{\mathrm{KS}}$, and a CDF is nonnegative. For equality, put reference mass ℓ at $u_{\min}$ and the remainder at $+\infty$; move mass $\min\{\ell, \kappa_{\mathrm{KS}}\}$ from $u_{\min}$ to $+\infty$

for the current law. Their one-sided CDF difference at every finite level is exactly the moved mass, and $F(u_0) = [\ell - \kappa_{\mathrm{KS}}]_+$. □

The left-grid maximum in Theorem A.7.1 loses no information supplied by the grid lower bounds and CDF monotonicity alone. To verify this, merge the enabled bounds at each grid point by taking their maximum. Place the increments of their running maximum at the corresponding grid points, and put the remaining mass at $+\infty$. This CDF attains the left-grid lower envelope at every finite level. Additional distributional assumptions can, of course, yield a larger bound.

APPENDIX B ALGORITHM AND IMPLEMENTATION

B.1 REQRC CAUSAL PROTOCOL

Main Algorithm 1 gives the procedure; the feedback and initialization conventions are as follows.

1. **No look-ahead.** After a window ends, its stored candidate forecasts and demand may be used to compute W_t only when the delayed feedback arrives. Before each decision, newly revealed values are inserted into the buffer; information still awaiting feedback is excluded.
2. **Buffer initialization.** The real-data experiments initialize the buffer with revealed validation-window values, as specified in Appendix B.4. For deployment without K revealed values, the startup convention is to use $u_{\max}$ until enough feedback is available.

The extension that selects the rank using valid bounds also records two status flags: *threshold infinite* and *certified*. Main ReqRC outputs only the selected levels, the deployed forecasts, and the updated buffer.

B.2 HOW PREDICTORS AND BASE LEVELS ARE CONNECTED

One experimental predictor–method pair is a lightweight quantile system. Given target τ, it uses

$$u_{\mathrm{lo}} := \tau, \qquad u_{\mathrm{anc}} := 1 - \eta_{\mathrm{tail}}(1-\tau), \qquad \eta_{\mathrm{tail}} \in (0,1).$$

Let A_t^{anc} denote the reference base level supplied by this two-endpoint predictor. Thus the upper reference level retains the fraction η_{tail} of the original tail probability: $1 - u_{\mathrm{anc}} = \eta_{\mathrm{tail}}(1-\tau)$. The predictor generates these two low-cost endpoints, while a separately fitted base calibration rule selects a level from their interval. For each univariate channel, a `StandardScaler` is fitted on the initial 56% of time points. We then train two phase-aware spline quantile models:

$$q_t^{\mathrm{anc}}(u_{\mathrm{lo}}), \qquad q_t^{\mathrm{anc}}(u_{\mathrm{anc}}).$$

The two experts use a lookback length and forecasting horizon of 96, periodic phase features, and level-equivariant spline features. Each minimizes the pinball loss at its respective quantile with ℓ_1 regularization. The current implementation uses $\tau = 0.85$, $\eta_{\mathrm{tail}} = 1/3$, and hence $u_{\mathrm{anc}} = 0.95$, together with a regularization coefficient of 50 and default phase periods of 24 and 168 samples. The anchor endpoint is projected pointwise to be no lower than the lower endpoint.

Both models are then frozen and used to generate out-of-sample predictions on the 14% calibration segment. A separate phase-conditioned coverage rule is fitted only on this segment to select $A_t^{\mathrm{base}} \in [u_{\mathrm{lo}}, u_{\mathrm{anc}}]$. Its coverage target is $\tau + 0.01$, and the ℓ_2 regularization coefficient for its phase terms is 0.01. The final 30% is used only for testing.

During testing, the fitted phase coefficients remain fixed. Write ϕ_s for the phase design, θ for the fitted phase coefficients, σ for the logistic function, b and a for the lower-endpoint and anchor mean forecasts over a window, and y for the observed demand. Over the latest K revealed windows the pooled empirical coverage at additive offset $\delta \geq 0$ is

$$\mathrm{QRE}(\delta) := \frac{1}{KHP} \sum_{j=1}^{K} \sum_{h=1}^{H} \sum_{p=1}^{P} \mathbf{1}\Big\{ b_{jh}^{(p)} + \sigma\big(\phi_{s_j}^{\top}\theta + \delta\big)\big(a_{jh}^{(p)} - b_{jh}^{(p)}\big) \geq y_{jh}^{(p)} \Big\},$$

and the offset is the smallest nonnegative value reaching the target,

$$\delta_t := \min\{\delta \geq 0 : \mathrm{QRE}(\delta) \geq \tau\} \text{ on } [0, 35],$$

computed by 70 bisection steps, with $\delta_t = 35$ when $\mathrm{QRE}(35) < \tau$. The resulting logistic weight is mapped linearly to $[u_{\mathrm{lo}}, u_{\mathrm{anc}}]$ to obtain the base level,

$$A_t^{\mathrm{base}} = u_{\mathrm{lo}} + \big(u_{\mathrm{anc}} - u_{\mathrm{lo}}\big)\, \sigma\big(\phi_{s_t}^\top \theta + \delta_t\big).$$

The offset uses revealed data only. For each channel, this same causal sequence of base levels is reused across TFT, DeepAR, DistPred, and the six-level quantile ensemble in the main ReqRC comparison. The method-wrapping experiment instead takes A_t from the method being wrapped.

These models and the separately frozen base calibration rule form one input pair to ReqRC. They are experimental components, not part of the ReqRC method definition.

B.3 Sparse Quantile Ensemble Used in Experiments

The fourth predictor is a lightweight six-level quantile ensemble. It is included only as an experimental predictor; ReqRC does not require this architecture.

For a predeclared high-tail set $\mathcal{U}_{\mathrm{tail}} \subset (u_{\mathrm{anc}}, 1)$, the general construction preserves u_{lo} and u_{anc}, optionally inserts interpolated nodes between them, and adds $\mathcal{U}_{\mathrm{tail}}$. The experimental instantiation is

$$\mathcal{U}_{\mathrm{knot}} = \{0.85, 0.90, 0.95, 0.975, 0.99, 0.995\}.$$

They lie in the continuous level interval $\mathcal{U} = [0.85, 0.995]$, and predictions between adjacent nodes are defined by interpolation.

Here $\tau = 0.85$, $\eta_{\mathrm{tail}} = 1/3$, and $\mathcal{U}_{\mathrm{tail}} = \{0.975, 0.99, 0.995\}$. The ensemble preserves u_{lo} and u_{anc} exactly. Writing $u_{\mathrm{mid}} = (u_{\mathrm{lo}} + u_{\mathrm{anc}})/2$, the midpoint is obtained by linear interpolation:

$$q_t(u_{\mathrm{mid}}) = q_t(u_{\mathrm{lo}}) + \frac{u_{\mathrm{mid}} - u_{\mathrm{lo}}}{u_{\mathrm{anc}} - u_{\mathrm{lo}}}\big(q_t(u_{\mathrm{anc}}) - q_t(u_{\mathrm{lo}})\big).$$

The endpoints in $\mathcal{U}_{\mathrm{tail}}$ use the same phase-aware, level-equivariant spline architecture as the base quantile predictor and are fitted independently on the expert-training segment with their respective pinball losses. Let $q_{t,h}^{\mathrm{raw}}(u_\ell)$ denote a raw upper-tail output. The reference-endpoint-preserving monotone projection is

$$\widetilde{q}_{t,h}(u_\ell) = \max\{q_{t,h}^{\mathrm{raw}}(u_\ell), \widetilde{q}_{t,h}(u_{\ell-1})\}, \qquad u_\ell > u_{\mathrm{anc}}.$$

Adjacent nodes are connected by piecewise-linear interpolation. This construction ensures that the prediction at every position in the window is nondecreasing with the level, while leaving the two reference endpoints unchanged when correcting crossings among upper-tail endpoints.

B.4 ReqRC Method

ReqRC follows the update in Section 5.1. The main experiments use $K = 8$, $D = 1$, $j = 8$, and $u_{\max} = 0.995$; these are experimental settings, not restrictions on the method definition.

Before the first test decision, the replay computes minimum feasible levels from the stored validation forecasts and their revealed demands. It initializes the history with these values and takes the latest K for the first threshold. An infeasible validation window contributes $+\infty$, rather than being discarded. At each subsequent decision, newly available test-window values enter the history, and the threshold uses the latest K available values. The explicit-index replay checks that validation history is sufficient and raises an error otherwise; it does not silently generate independent samples or fill missing entries. Thus the reported experiments start with observed pre-test history, rather than the no-history fallback in B.1.

Appendix A.6 gives the optional rule for choosing j using valid bounds, and Appendix D.4 evaluates it when the data-generating mechanism provides these bounds.

B.5 ReQRC with Fixed j and the Full Path

The experiment varies $j = 1, \ldots, K + 1$ while keeping the predictor, base rule, candidate forecasts, and history protocol fixed. With $K = 8$, $j = 1, \ldots, 8$ selects historical order statistics, while $j = 9$ uses the maximum-level fallback. Appendix C.7 reports the complete path, including the main setting $j = 8$.

B.6 Baseline Implementation and Adaptation

All predictors produce frozen candidate forecasts on the same continuous calibration interval and finite calibration nodes, and every method must select its level before the ground-truth demand is revealed. The predictor-side protocol is as follows.

Table 3: Predictors and their evaluated output protocol.

Predictor	Paper and implementation source	Evaluated output and adaptation
TFT	Lim et al. (2021); Time-Series-Library backbone	Native multi-quantile pinball training at the common calibration nodes; three-seed arithmetic mean followed by cumulative-maximum projection
DeepAR	Salinas et al. (2020); official GluonTS kernel	Student-t NLL checkpoints; empirical level quantiles extracted from 4,096 trajectories per test forecast
DistPred	Liang (2025); official repository, commit `bcc9115b6c87`	Official-derived Minusformer-96M architecture with PWM-CRPS and 100 native outputs; the 0.995 endpoint is not treated as an exact population quantile
Six-level quantile ensemble	Sections B.2–B.3; implementation supplied in the artifact	Lightweight predictor producing six forecasts ordered from least to most conservative

All methods operate on the same frozen forecasts and continuous level interval, using the adaptations described below.

Table 4: Methods and their one-sided adaptations.

Method	Published source	Implementation used here
ACI	Gibbs & Candès (2021)	Artifact implementation of its scalar online update, clipped to the shared ordered calibration interval
Rolling RC	Feldman et al. (2023)	Artifact implementation with pointwise miscoverage loss and the paper's online risk-update principle
Conformal PID	Angelopoulos et al. (2023)	Artifact one-sided adapter with a causal persistence scorecaster
DtACI	Gibbs & Candès (2024)	Artifact dynamically tuned ACI adapter with a fixed-share step-size bank and randomized expert selection
AgACI	Zaffran et al. (2022)	Artifact one-sided BOA aggregation over an ACI step-size bank
SF-OGD	Bhatnagar et al. (2023)	Artifact scale-free OGD update, clipped to the common calibration interval
L-ARC	Zecchin & Simeone (2024)	Artifact finite-feature localized risk-control adapter with causal phase features
ECI	Wu et al. (2025)	Artifact one-sided error-quantified update over the common calibration interval

Method	Published source	Implementation used here
NCC	Li & Rodríguez (2025)	Artifact causal neural calibration method using only the shared online state and fixed candidate forecasts
COP	Hu et al. (2026)	Artifact distribution-informed optimistic update with causal empirical score history

Their exact experimental adaptations are standardized as follows.

ACI. An upper empirical-quantile rule computes the 0.85 quantile of the minimum feasible levels on the calibration segment and uses it to initialize the prediction level:

$$u_{t+1} = \Pi_{[0.85,0.995]}\left[u_t + \eta(\tau - \mathrm{Cover}_t(u_t))\right].$$

For the additional step-size scans, we prespecify $\eta \in \{0.01, 0.025, 0.05, 0.10\}$. Each candidate is initialized from the upper empirical 0.85-quantile on the first half of the routing segment; the second half determines the validation Pareto set. This adaptation uses strictly causal feedback and does not use the coverage of the current maximum endpoint, observed only after the window, to modify the target.

Rolling RC. With pointwise miscoverage loss $\ell_t^{\mathrm{RRC}} = H^{-1}\sum_h \mathbf{1}\{q_{t,h}(u_t) < y_{t,h}\}$, the update is

$$u_{t+1} = \Pi_{[0.85,0.995]}\left[u_t + \eta\big(\ell_t^{\mathrm{RRC}} - (1-\tau)\big)\right], \qquad \eta \in \{0.01, 0.025, 0.05, 0.10\}.$$

The same routing split and validation-Pareto rule are used for all four step sizes.

In this adaptation, ACI and Rolling RC use the same feedback update but start from different levels. ACI computes its initial test level from the calibration data as described above; Rolling RC starts at 0.85, updates its level over the calibration windows, and continues from that level on the test data. Their main-experiment step sizes also differ: ACI uses $(0.995 - 0.85)/\sqrt{8} \approx 0.0513$, and Rolling RC uses 0.05. The four-step-size grid above is used for the additional scans.

Window-failure ACI (WF-ACI). For the method design comparison in Appendix C.6, we additionally evaluate an ACI variant driven only by whether a completed window meets the target. Like ACI and ReqRC, it permits continuous levels between the six forecast nodes and evaluates forecasts by linear interpolation; its projection clips to the interval rather than rounding to a node. Appendix C.6 specifies its update, initialization, and validation-only step-size selection. This is a purpose-built comparison, not an additional published method in the main set of ten alternatives.

ReqRC. For every predictor, we reuse the same historical order-statistic update with $K = 8, D = 1$ and prespecified $j = 8$. ACI and Rolling RC use the same fixed candidate forecasts and causal information cutoff, making the comparison one between calibration-level update rules under a shared protocol.

Additional modern methods. All eight methods use the same fixed prediction levels, target 0.85, one-window feedback delay, and initial level 0.85. The first half of the calibration segment initializes state; the second half is used only for validation selection as specified in Appendix C.3. Conformal PID scans $k_p \in \{0, 0.02\}$ and $k_i \in \{0.002, 0.005, 0.01\}$ with a persistence scorecaster. DtACI uses step bank $\{0.002, 0.004, 0.008, 0.016, 0.032, 0.064, 0.128\}$, aggregation rates $\{1, 2\}$, fixed shares $\{0.001, 0.01\}$, and randomized expert selection. AgACI uses one-sided BOA aggregation over $\{0.002, 0.004, 0.008, 0.016, 0.032, 0.064\}$. SF-OGD scans scale multipliers $\{0.5, 1, 2\}$. L-ARC scans bandwidths $\{0.5, 1, 2\}$ and $\eta \in \{0.1, 0.25, 0.5\}$. ECI scans learning rates $\{0.05, 0.10, 0.50, 1.00\}$; COP crosses learning rates $\{0.05, 0.10, 0.50, 1.00\}$ with history lengths $\{8, 16, 100\}$; NCC scans the prespecified hidden-width, history-length, learning-rate, and adjustment-factor configurations recorded in the released result table. All methods retain continuous outputs clipped to $[0.85, 0.995]$, and forecasts are evaluated by the same linear interpolation between adjacent nodes. DtACI and AgACI also retain continuous expert outputs. L-ARC keeps its rank-coordinate parameterization but computes failure feedback from the actual deployed level, without rounding it to a node.

B.7 Computational Complexity and Resources

Let M_{knot} be the number of calibration nodes and H the window length. Storing the candidate forecasts requires $O(HM_{\mathrm{knot}})$ memory per pending window; this is a predictor-side cost rather than ReqRC overhead. Because $\mathrm{Cover}_t(u)$ is a step function of u determined by the H crossing levels at which the candidates meet the observation, ReqRC itself needs only those H values per pending window, so its buffer grows as $O(DH)$ rather than $O(DHM_{\mathrm{knot}})$ with feedback delay D. ReqRC computes one order statistic from K historical minimum feasible levels. Direct sorting costs $O(K \log K)$, whereas selection can achieve $O(K)$; the method buffer requires $O(K)$ space. With $K = 8$, this update is small relative to predictor inference.

The recorded logs and the hardware used for verification are summarized below. Reported runtimes are direct aggregates of the available logs; unrecorded jobs are not extrapolated into total GPU-hours.

Table 5: Recorded training and inference time.

		Training time		Inference time	
Component	Logged runs	Total	Median per run	Total	Median per run
Logged 0.85–0.99 components of the six-level quantile ensemble	206	4.043 h	8.72 s	0.009 h	0.095 s
TFT	24	4.082 h	343.01 s	0.110 h	11.06 s
DeepAR	24	0.827 h	113.40 s	2.309 h	168.51 s
DistPred	24	2.098 h	235.95 s	0.004 h	0.63 s

The verification machine runs Windows 11 and has an Intel Core i9-14900HX CPU, 31.73 GiB of memory, and an NVIDIA RTX 4060 Laptop GPU (8,188 MiB; driver 560.94). No separate complete timing log is available for the standalone 0.995 upper-tail head. The first row therefore covers only the traceable 0.85–0.99 components and must not be interpreted as the total end-to-end training budget of the six-level quantile ensemble.

Appendix C Full Experimental Results

C.1 Datasets and Temporal Splits

The experiments cover 8 multivariate time-series datasets. Each variable is treated as a separate univariate upper-forecast channel, yielding 206 channels and 16,316 held-out test windows in total. The 157 nonnegative channels in Ali, SMD, and the independent CPU dataset receive a literal resource-allocation interpretation; Weather and ETT are used as generic upper-forecast tasks. The capacity replay in Appendix C.9 uses the 156 normalized CPU/memory channels from Ali and SMD only.

Table 6: Datasets, upstream sources, and held-out test counts.

Dataset	Channels	Test windows	Source and resolution
Weather	21	3,444	Max Planck Institute for Biogeochemistry; 2020 archive, 10-minute observations
SMD	56	8,288	Server Machine Dataset, released with OmniAnomaly (Su et al., 2019)
CPU	1	90	CPU usage from a park of Huawei Technologies Co., Ltd.; one-minute observations, May 8–27, 2023
Ali	100	700	Alibaba Cluster Trace v2018 (Guo et al., 2019)
ETTh1	7	378	ETT-small, hourly
ETTh2	7	378	ETT-small, hourly
ETTm1	7	1,519	ETT-small, 15-minute

Dataset	Channels	Test windows	Source and resolution
ETTm2	7	1,519	ETT-small, 15-minute
Total	**206**	**16,316**	

Weather uses the 21-variable series in `weather_official.csv`; missing-value markers are repaired causally by a prespecified forward-fill rule. SMD uses CPU and memory series from one fixed random set of machines because materializing the full archive was impractical during preparation; every method uses the same extract. Ali uses a fixed 100-machine CPU/memory extract resampled to the common tabular format before the chronological split. The four ETT files (Zhou et al., 2021) each contribute their seven numeric forecasting channels.

Each channel is split chronologically into a 56% expert-training segment, a 14% routing and validation segment, and a 30% held-out test segment. Both the forecasting horizon and the predictor lookback length are 96 time points, and test windows do not overlap. A `StandardScaler` is fitted only on the first 56% of time points, and model outputs are transformed back to their original scale before resource metrics are computed. Test ground truth is never used for training, checkpoint selection, $u_{\max}$ selection, or the current level decision.

The other datasets receive neither outlier deletion nor imputation. Calendar features are generated when a time column is present. The one exactly duplicated Weather row is retained in temporal order because its removal was not prespecified.

C.2 Metrics, Aggregation, and Uncertainty

For each window t, the experiments first compute

$$\mathrm{Cover}_t(U_t) = \frac{1}{H}\sum_{h=1}^{H} \mathbf{1}\{y_{t,h} \le q_{t,h}(U_t)\},$$

$$v_t = [\tau - \mathrm{Cover}_t(U_t)]_+, \quad g_t = [\tau - \mathrm{Cover}_t(u_{\max})]_+, \quad r_t = v_t - g_t.$$

The per-window overprovisioning cost and its aggregate are

$$c_t(U_t, Y_t) = \frac{1}{H}\sum_{h=1}^{H}[q_{t,h}(U_t) - y_{t,h}]_+, \qquad \mathrm{PMAE} = \frac{1}{T}\sum_{t=1}^{T} c_t(U_t, Y_t).$$

The pooled-window aggregation used in the main text assigns equal weight to every test window; the main-text risk symbol is

$$V = \frac{1}{T}\sum_{t=1}^{T} v_t,$$

and the remaining metrics use the same window weights. The equal-channel aggregation first averages within each channel and then weights all 206 channels equally. The equal-dataset aggregation first obtains an equal-channel result for each dataset and then weights all 8 datasets equally. These aggregations therefore answer different questions, and their values must not be used interchangeably.

Risk–cost Pareto dominance is defined as follows: system A dominates system B if A is no worse than B in both V and PMAE and is strictly better in at least one of them. Raw PMAE retains each channel's measurement scale, which need not share a physical unit across datasets; its aggregate is therefore used only for systems evaluated on the same fixed test mixture. Scale-free comparisons divide each channel's PMAE by its training-only IQR, while the literal resource interpretation is restricted to Ali, SMD, and CPU.

Channel scope of the reported results. Main Table 1, the constant-level comparison, and the rank path use all 206 channels and 16,316 test windows. Main Table 2, the Pareto comparison with normalized PMAE, the matched-cost/matched-success comparisons, and the ReqRC comparisons across all base methods use the 203 channels with nonzero training IQR, comprising 15,840 test windows. Shortfall, the window success rate, and cost in these comparisons are all computed on the same channel subset. The separate maximum-level-use summary in Appendix C.6 uses all 206 channels because it does not require cost normalization. The excluded channels are SMD's `CPU_m_2-2` (148 windows) and Weather's `rain (mm)` and `raining (s)` (164 windows each), whose training IQRs are zero. Main Tables 1A and 2 both average the four predictors equally; ReqRC has $V = 0.02429$ and $V = 0.02456$, respectively, because of the channel scope, not a configuration change. The quantile-ensemble–ReqRC combination alone has $V = 0.02151$ over all windows and $V = 0.02172$ on the subset used for the normalized Pareto comparison.

Paired uncertainty is estimated with two bootstrap procedures that resample channels within each dataset. Both preserve all windows within a sampled channel and the pairing between the two systems, and use 10,000 replicates. The primary intervals weight sampled channels by their number of windows, so their estimand matches the pooled-window main tables. The supplementary intervals give sampled channels equal weight and correspond to the equal-channel estimand. The two types of intervals are stored separately; the main text does not mix them.

C.3 Training and Model-Selection Protocol

The formal level interval, finite nodes, targets, and control parameters are as follows.

Table 7: Experimental targets, forecast levels, and calibration settings.

Item	Fixed value
Level interval $\mathcal{U}$	$[0.85, 0.995]$
Level nodes $\mathcal{U}_{\text{knot}}$	$\{0.85, 0.90, 0.95, 0.975, 0.99, 0.995\}$
Window coverage target τ	0.85
Reference-level factor η_{tail}	$1/3$, giving $u_{\text{anc}} = 0.95$
Window-level coverage probability target β	0.85
Within-space selection target β_{sel}	0.85
Number of historical windows K	8
Feedback delay D	1
Data split	56% / 14% / 30%
Neural-predictor random seeds	2024, 2025, 2026
DeepAR inference trajectories	4,096
Native DistPred outputs	100
Bootstrap replicates	10,000

Model selection for TFT, DeepAR, and DistPred uses only the training and validation segments. Test predictions from the three random seeds are first averaged arithmetically and then subjected to a cumulative-maximum projection over the calibration nodes. All methods are replayed causally on the same frozen candidate forecasts, with each selected configuration frozen before test evaluation. The held-out test set is used only for final evaluation. For each predictor and each of the eight additional method families, the primary comparison selects the configuration with the smallest validation V over all 206 channels, breaking ties by raw PMAE. ACI and Rolling RC retain their prespecified primary settings. For the additional points in Main Figure 4, we retain each family's validation-nondominated configurations in (V, PMAE) on the 203 channels with nonzero training IQR. Together with all 44 primary combinations, this gives 181 distinct configurations. All these decisions precede test evaluation.

Scope of the Predictor Comparison We select TFT, DeepAR, and DistPred as external predictor baselines because they natively produce multiple quantiles, a conditional predictive distribution, and predictive samples, respectively. Each can therefore produce the same six prediction levels without altering its model definition. A six-level quantile ensemble provides the fourth predictor. All four are fixed before method evaluation; none is part of ReqRC's method definition.

C.4 Existence of Window-Level Coverage Failures

Main Table 1 reports two findings that motivate window-level control:

1. With a fixed set of candidate forecasts, every method satisfies the long-run average coverage target, yet no method other than ReqRC exceeds 0.86 window success and seven of the ten fall below the 0.85 target; their shared G in Main Table 1A shows that the entire gap comes from level selection.
2. Even the highest forecast fails on 1.51%–6.54% of windows, so some windows have no feasible level and no method can remove that shortfall.

Appendix C.3 gives the training and selection protocol; Appendix B.6 describes the method adaptations.

C.5 Method Comparisons across Predictors

Shortfall decomposition for methods using the same candidates Main Table 1A reports this decomposition for all methods. With shared candidate forecasts, every method has the same G, so all differences in V are differences in R.

Validation-Only Method Scans Table 8 lists the ACI and Rolling RC configurations retained by the validation protocol and their held-out test results. Both use four prespecified step sizes. ReqRC's prespecified rank is used directly rather than selected from this scan.

Table 8: Validation-only ACI and Rolling RC scan on the quantile ensemble; frontier status is within this scan and ReqRC.

Family	Configuration	Val. V	Val. raw PMAE	Test V	Test raw PMAE	Test frontier
ReqRC	Prespecified rank	–	–	0.02151	31.996	Yes
Rolling RC	$\eta = 0.010$	0.07554	17.666	0.06863	19.947	Yes
Rolling RC	$\eta = 0.025$	0.06856	18.068	0.06452	20.225	Yes
Rolling RC	$\eta = 0.050$	0.06505	18.282	0.06241	20.531	Yes
Rolling RC	$\eta = 0.100$	0.06236	18.505	0.06106	20.936	Yes
ACI	$\eta = 0.010$	0.05264	22.903	0.05382	22.062	Yes
ACI	$\eta = 0.025$	0.05622	21.150	0.05818	21.064	Yes
ACI	$\eta = 0.050$	0.05861	19.709	0.05938	20.969	Yes
ACI	$\eta = 0.100$	0.05953	19.265	0.05967	21.131	No

Additional Modern Methods For the eight additional modern methods, the validation-retained endpoints primarily extend the lower-cost, higher-risk region; none dominates ReqRC's low-shortfall point. Main Figure 4 plots all 181 held-out points over the four predictors, and Main Table 2 reports the method-level means.

Per-channel fraction meeting the target

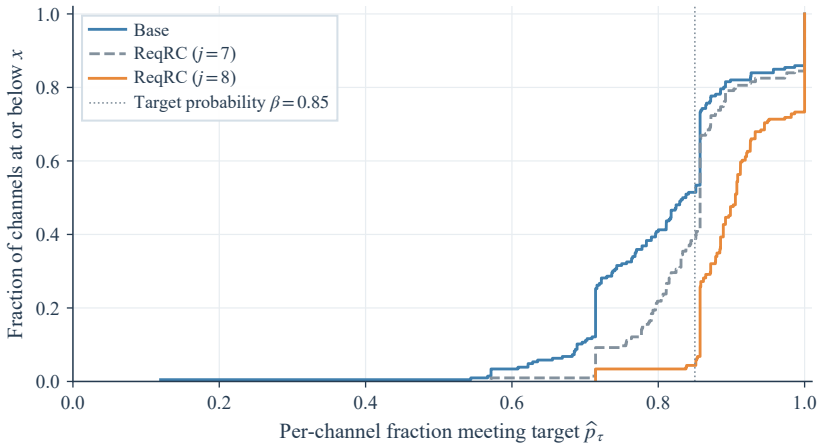


Figure 5: Empirical distribution of the per-channel window success rate $\hat{p}_\tau$ over 206 channels. The dotted line marks the target probability 0.85. The curves compare Base with ReqRC at $j = 7$ and $j = 8$, using the frozen results for $j = 7$. ReqRC at $j = 8$ shifts the distribution to the right: its median is 0.9054 and only 9/206 channels fall below the target, compared with a median of 0.8356 and 106/206 below target for Base.

Figure 5 directly checks whether pooling all windows conceals persistently unreliable channels.

C.6 Supporting Predictor and Method Analyses

When the base rule retains the same deployed level, expanding the predictor reduces G from 0.03621 to 0.00697, and the entire reduction is transferred to R, leaving V unchanged. ReqRC makes the newly introduced levels available for deployment. This supporting ablation separates a predictor-side expansion from the ReqRC method's contribution; the main method comparison is the cross-predictor experiment in C.5.

Using pooled-window averages and 10,000 channel-bootstrap replicates, adding ReqRC to the expanded predictor lowers V by 0.03125 (95% interval $[-0.03709, -0.02624]$), and replacing the base predictor with the six-level quantile ensemble lowers V by 0.01855 ($[-0.02401, -0.01408]$). Equal-channel effects have the same directions.

The first comparison fixes the six-level quantile predictions and adds only ReqRC, so the decrease in V exactly equals the decrease in R. The second fixes ReqRC and changes only the predictor: G decreases, R changes because its reference lower bound changes, and the net V decreases.

Prediction-component check. With the ReqRC method, windows, and $u_{\max}$ fixed, removing the reference-endpoint projection changes V from 0.02151 to 0.02196 and raw PMAE from 31.996 to 31.865. Removing the upper-tail heads instead raises V to 0.04006 and G to 0.03621. The projection restores ordered forecasts, while the upper-tail heads determine how much shortfall remains at the highest level. Full results and paired intervals are in `evidence/submission_audits/ote_ablation_summary.csv` and `evidence/submission_audits/ote_ablation_bootstrap.csv`.

We next apply the same $j = 8$ update to all 10 alternative methods and aggregate over the four fixed predictors. To report scale-free cost in the same table, this analysis uses the 203 channels with nondegenerate training IQR. Every method configuration is still fixed on validation data; test outcomes do not select configurations.

Table 9: Applying the same ReqRC update to all 10 alternative methods, averaged over four fixed predictors.

Base method	Base V	+ReqRC V	ΔV [95% interval]	$\Delta\widehat{p}_\tau$ [95% interval]	ΔPMAE [95% interval]
Rolling RC	0.06525	0.02717	-0.03808 [-0.03956,-0.03654]	+15.23 pp [14.63,15.83]	+0.593 [0.520,0.677]
ACI	0.06212	0.02666	-0.03546 [-0.03693,-0.03406]	+14.21 pp [13.64,14.80]	+0.576 [0.503,0.658]
NCC	0.05499	0.02533	-0.02966 [-0.03144,-0.02792]	+11.06 pp [10.60,11.52]	+0.580 [0.499,0.673]
DtACI	0.04214	0.02471	-0.01744 [-0.01840,-0.01647]	+6.79 pp [6.52,7.07]	+0.424 [0.365,0.492]
Conformal PID	0.04034	0.02302	-0.01732 [-0.01827,-0.01638]	+6.84 pp [6.58,7.09]	+0.324 [0.282,0.371]
AgACI	0.03939	0.02490	-0.01449 [-0.01531,-0.01367]	+5.79 pp [5.57,6.02]	+0.418 [0.357,0.487]
ECI	0.03681	0.02268	-0.01413 [-0.01490,-0.01337]	+5.38 pp [5.17,5.60]	+0.240 [0.205,0.280]
COP	0.03669	0.02320	-0.01349 [-0.01427,-0.01272]	+5.20 pp [5.00,5.40]	+0.248 [0.212,0.286]
SF-OGD	0.03693	0.02446	-0.01247 [-0.01313,-0.01181]	+5.11 pp [4.92,5.30]	+0.385 [0.326,0.452]
L-ARC	0.03655	0.02419	-0.01235 [-0.01315,-0.01161]	+5.01 pp [4.80,5.22]	+0.382 [0.322,0.454]

All differences are "+ReqRC minus Base," with intervals from 10,000 dataset-stratified paired channel-bootstrap draws. All 40 predictor–method combinations have nonincreasing V and a higher window success rate, and every method-level ΔV interval excludes zero. Because the candidate forecasts are unchanged, $\Delta G = 0$ and $\Delta R = \Delta V$; the positive cost changes quantify the resources paid for the common historical threshold. This experiment tests the Section 5.1 update with different causal base levels A_t.

Table 10 reports the same 40 predictor–method pairs individually, which is the disaggregation behind the claim that ReqRC improves every combination.

Table 10: The 40 predictor–method pairs behind the claim that ReqRC improves every combination. Each row applies the same $j = 8$ update to that base method on that predictor's frozen forecasts, over the 203 channels with nondegenerate training IQR. V and PMAE follow Main Table 2, with PMAE normalized by the training interquartile range. All differences are "+ReqRC minus Base," so a negative ΔV and a positive $\Delta\widehat{p}_\tau$ are improvements; all 40 rows improve on both. Paired bootstrap intervals are reported at the method level in Table 9.

Predictor	Base method	Base V	+ReqRC V	ΔV	$\Delta\widehat{p}_\tau$ pp	ΔPMAE
Quantile ens.	Rolling RC	0.06339	0.02477	-0.03862	+15.25 pp	+0.421
	ACI	0.06029	0.02433	-0.03596	+14.27 pp	+0.405
	NCC	0.05132	0.02254	-0.02878	+10.54 pp	+0.399
	DtACI	0.03980	0.02213	-0.01767	+6.76 pp	+0.290
	Conformal PID	0.03796	0.02102	-0.01695	+6.49 pp	+0.245
	AgACI	0.03737	0.02228	-0.01509	+5.98 pp	+0.283
	ECI	0.03454	0.02044	-0.01411	+5.24 pp	+0.150
	COP	0.03410	0.02113	-0.01297	+5.00 pp	+0.158
	SF-OGD	0.03511	0.02232	-0.01279	+5.18 pp	+0.241
	L-ARC	0.03418	0.02163	-0.01255	+4.81 pp	+0.226
	Rolling RC	0.05326	0.02346	-0.02979	+11.95 pp	+0.558

Predictor	Base method	Base V	+ReqRC V	ΔV	$\Delta\widehat{p}_\tau$ pp	ΔPMAE
DistPred	ACI	0.05103	0.02298	-0.02805	+11.29 pp	+0.544
	NCC	0.04646	0.02174	-0.02472	+9.14 pp	+0.567
	DtACI	0.03550	0.02092	-0.01457	+5.81 pp	+0.420
	Conformal PID	0.03580	0.01964	-0.01617	+6.16 pp	+0.353
	AgACI	0.03327	0.02121	-0.01206	+4.82 pp	+0.410
	ECI	0.03232	0.01892	-0.01340	+5.11 pp	+0.274
	COP	0.03264	0.01970	-0.01294	+4.82 pp	+0.295
	SF-OGD	0.03076	0.02071	-0.01005	+4.26 pp	+0.360
	L-ARC	0.03302	0.02078	-0.01224	+4.77 pp	+0.388
	Rolling RC	0.07149	0.02707	-0.04443	+17.32 pp	+0.799
	ACI	0.06818	0.02670	-0.04149	+16.12 pp	+0.777
	NCC	0.05973	0.02534	-0.03439	+12.85 pp	+0.779
	DtACI	0.04465	0.02445	-0.02020	+7.59 pp	+0.578
	Conformal PID	0.04325	0.02212	-0.02113	+8.39 pp	+0.405
	AgACI	0.04146	0.02471	-0.01675	+6.55 pp	+0.578
	ECI	0.03959	0.02138	-0.01821	+6.74 pp	+0.297
	COP	0.03941	0.02271	-0.01670	+6.23 pp	+0.317
	SF-OGD	0.03904	0.02410	-0.01495	+5.90 pp	+0.571
	L-ARC	0.03840	0.02420	-0.01420	+5.71 pp	+0.573
TFT	Rolling RC	0.07286	0.03336	-0.03950	+16.41 pp	+0.594
	ACI	0.06899	0.03265	-0.03634	+15.18 pp	+0.577
	NCC	0.06245	0.03171	-0.03075	+11.70 pp	+0.575
	DtACI	0.04863	0.03133	-0.01730	+7.01 pp	+0.410
	Conformal PID	0.04432	0.02931	-0.01501	+6.31 pp	+0.292
	AgACI	0.04545	0.03140	-0.01405	+5.82 pp	+0.400
	ECI	0.04080	0.02997	-0.01082	+4.43 pp	+0.239
	COP	0.04062	0.02926	-0.01136	+4.77 pp	+0.220
	SF-OGD	0.04279	0.03071	-0.01207	+5.09 pp	+0.368
	L-ARC	0.04059	0.03015	-0.01043	+4.73 pp	+0.342

Feedback-object ablation. Holding the AgACI base action, the history length $K = 8$, the delay, and the rank rule fixed, we vary only the object retained from each completed window: the exact minimum feasible level W_t, or the binary fact that the base action failed, encoded as $u_{\min}$ on success and $u_{\max}$ on failure. The base action fell short in $10{,}425$ of $65{,}264$ windows. Of those, $8{,}256$ (79.19%) admitted a strictly cheaper finite level that would have met the target, $2{,}169$ (20.81%) were infeasible, and none was repairable only at $u_{\max}$. A binary indicator therefore cannot separate the repairable failures from the infeasible ones; the required level can.

At the prespecified rank eight the required-level arm attains $V = 0.02490$, $\widehat{p}_\tau = 0.8982$, and 1.408 normalized PMAE, saturating 19.32% of windows; the binary arm attains $V = 0.01688$, $\widehat{p}_\tau = 0.9350$, and 2.379, saturating 76.44%. The binary arm is thus safer at matched rank, but it buys that safety by saturating four times as many windows: its shortfall is 0.00794 lower than the required-level arm's (95% interval $[0.00727, 0.00865]$), and its saturation rate is 57.1 points higher ($[54.7, 59.5]$). At matched cost the required-level arm is the safer one and at matched safety the cheaper one: at 1.408 normalized PMAE it attains a 27.3% lower V, and at $V = 0.02490$ it needs 1.408 normalized PMAE against 1.988. Rank curves for both objects are reported as diagnostics; the rank-eight comparison is prespecified and no rank is selected on held-out outcomes.

Maximum-level use. Across all 206 channels and 16,316 test windows per predictor, the fraction of windows using $u_{\max}$, averaged over the four predictors for each base method, ranges from 19.31% (AgACI or NCC with ReqRC) to 35.05% (ECI with ReqRC). This is the rounded 19%–35% range reported in Section 6.3; always selecting the highest level would give 100%. The rates before and after ReqRC for all ten methods are in `evidence/continuous_controllers/wrapper_maximum_level_use.csv`.

Validation-matched cost and success comparisons For each of the 40 predictor–method pairs behind Main Table 2, we compare ReqRC with a two-point mixture of that method's own validation configurations and the fixed highest-level endpoint. Each endpoint runs as an independent causal shadow policy; mixture metrics are expectations over the frozen mixture, not a feedback-coupled averaged action. The mixture weights are chosen on validation data only. In the cost comparison, the mixture matches ReqRC's validation PMAE and, among all such mixtures, minimizes validation V. In the success comparison, it reaches at least ReqRC's validation $\widehat{p}_\tau$ and, among all such choices,

minimizes validation PMAE. We evaluate the same weights on held-out data and report the mixture's expected metrics. Because both endpoints follow the causal protocol, the mixture does too.

On held-out data, every cost-matched baseline has higher V and a lower $\widehat{p}_\tau$ than ReqRC. In the success comparison, 14 of the 40 baselines retain at least ReqRC's held-out $\widehat{p}_\tau$, and all 14 are more expensive. Thus validation matching does not enforce equal test costs or success rates; no selected baseline is simultaneously at least as safe and cheaper than ReqRC.

Table 11: Validation-matched comparisons for the 40 predictor–method pairs behind Main Table 2. Cost-match: each mixture matches ReqRC's validation PMAE. Success-match: each mixture reaches at least ReqRC's validation $\widehat{p}_\tau$. $\Delta\widehat{p}_\tau$ is reported in percentage points. Deltas are baseline minus ReqRC on held-out data; positive ΔV and $\Delta\mathrm{PMAE}$ mean worse and more expensive for the baseline, and negative $\Delta\widehat{p}_\tau$ means less safe.

		Matched cost			Matched success	
Predictor	Method	ΔV	$\Delta\widehat{p}_\tau$ (pp)	$\Delta\mathrm{PMAE}$	$\Delta\widehat{p}_\tau$ (pp)	$\Delta\mathrm{PMAE}$
TFT	ACI	+0.01383	-6.07	-0.0827	-2.08	+0.2108
TFT	AgACI	+0.00443	-1.85	+0.0027	+0.10	+0.2024
TFT	SF-OGD	+0.00277	-1.51	-0.0182	-0.58	+0.0955
TFT	L-ARC	+0.00239	-1.42	-0.0125	-0.00	+0.1422
TFT	COP	+0.00521	-2.36	-0.0093	+0.39	+0.1922
TFT	ECI	+0.00591	-2.51	-0.0406	+0.43	+0.1948
TFT	Conformal PID	+0.00708	-3.10	+0.0152	+0.48	+0.3262
TFT	DtACI	+0.00672	-2.63	-0.0044	+0.23	+0.2640
TFT	NCC	+0.01144	-4.15	+0.0135	+0.96	+0.4117
TFT	Rolling RC	+0.01847	-7.29	-0.0121	-0.65	+0.3979
DeepAR	ACI	+0.02053	-8.16	-0.2604	-3.60	+0.2853
DeepAR	AgACI	+0.00828	-3.44	-0.1528	-2.25	+0.0500
DeepAR	SF-OGD	+0.00705	-3.07	-0.1612	-4.52	-0.4210
DeepAR	L-ARC	+0.00846	-3.44	-0.1605	-1.92	+0.0985
DeepAR	COP	+0.01209	-4.78	-0.3310	-1.00	+0.2523
DeepAR	ECI	+0.01170	-4.56	-0.2884	-1.87	+0.1549
DeepAR	Conformal PID	+0.01117	-4.19	-0.0823	-2.97	+0.1162
DeepAR	DtACI	+0.01011	-3.88	-0.1276	-2.18	+0.1424
DeepAR	NCC	+0.01629	-5.98	-0.1413	-1.19	+0.4904
DeepAR	Rolling RC	+0.02176	-8.70	-0.1614	-2.97	+0.4601
DistPred	ACI	+0.03010	-11.53	-0.0345	-0.95	+1.1861
DistPred	AgACI	+0.00974	-3.87	-0.0094	-0.14	+0.7792
DistPred	SF-OGD	+0.00890	-3.59	-0.0984	-1.48	+0.3732
DistPred	L-ARC	+0.00744	-3.12	-0.0311	+0.10	+0.6982
DistPred	COP	+0.01296	-4.99	-0.1010	-0.44	+0.8116
DistPred	ECI	+0.01349	-4.87	-0.0449	-0.26	+0.8672
DistPred	Conformal PID	+0.01646	-6.52	+0.3498	-0.44	+1.2617
DistPred	DtACI	+0.01306	-4.88	-0.0321	+0.13	+0.9552
DistPred	NCC	+0.02052	-7.69	+0.0649	+1.70	+1.5367
DistPred	Rolling RC	+0.03114	-11.94	-0.0247	-1.78	+1.2813
Quantile ensemble	ACI	+0.02180	-7.98	-0.1616	-2.37	+0.1375
Quantile ensemble	AgACI	+0.00859	-3.19	-0.0776	+0.03	+0.1501
Quantile ensemble	SF-OGD	+0.00801	-2.90	-0.1008	-1.16	+0.0301
Quantile ensemble	L-ARC	+0.00706	-2.34	-0.0428	+0.28	+0.1495
Quantile ensemble	COP	+0.00731	-2.48	-0.0554	+0.60	+0.1751
Quantile ensemble	ECI	+0.00741	-2.47	-0.0543	+0.62	+0.1730
Quantile ensemble	Conformal PID	+0.01055	-3.47	-0.0321	-0.18	+0.1857
Quantile ensemble	DtACI	+0.01100	-3.86	-0.0900	-0.54	+0.1403

		Matched cost			Matched success	
Predictor	Method	ΔV	$\Delta\widehat{p}_\tau$ (pp)	ΔPMAE	$\Delta\widehat{p}_\tau$ (pp)	ΔPMAE
Quantile ensemble	NCC	+0.01632	-5.39	-0.0970	+0.50	+0.2526
Quantile ensemble	Rolling RC	+0.02509	-9.57	-0.1344	-2.04	+0.2152

Method design comparisons We also apply the same validation-matching protocol to WF-ACI and the direct historical threshold across all four predictors. All eight cost-matched test points have higher shortfall and a lower fraction meeting the target than ReqRC. After success matching, two of the eight retain at least ReqRC's test fraction, and both cost more. Full configurations, mixture weights, and test differences for these eight and the 40 published-baseline pairs are in `evidence/continuous_controllers/matched_cost_all12.csv` and `evidence/continuous_controllers/matched_success_all12.csv`. The standalone WF-ACI comparison below retains its original target-based step-size selection; it is distinct from this mixture comparison.

This comparison examines the feedback used to update the level and the effect of retaining the base level. Window-failure ACI (WF-ACI) receives only the binary indicator that the completed window missed its target, not the shortfall magnitude or the minimum feasible level. With one-window feedback delay, its update is

$$u_{t+1} = \Pi_{[0.85,0.995]}\left[u_t + \eta\left(\mathbf{1}\{\mathrm{Cover}_t(u_t) < \tau\} - (1-\beta_c)\right)\right],$$

where Π clips to the interval, $\tau = 0.85$, and $\beta_c \in \{0.85, 8/9\}$ is the desired probability of meeting the window target. The value $8/9$ matches ReqRC's nominal rank baseline at $K = j = 8$, not its actual cost or achieved reliability. The direct historical threshold $U_t = \min(W_t^{(j)}, u_{\max})$ uses the same historical order statistic as ReqRC but omits the maximum with A_t. All these rules use the same frozen forecasts with linear interpolation between the six nodes, the same feedback delay, and the same held-out windows. Table 12 reports their results on the 203 channels with nonzero training interquartile range (15,840 windows per predictor); the final block averages the four predictors equally.

For WF-ACI, we scan $\eta \in \{0.001, 0.0025, 0.005, 0.01, 0.025, 0.05, 0.10, 0.145\}$. Each candidate starts at 0.85 and uses the first half of the routing segment to initialize its state. For each predictor and target, the second half selects the smallest validation PMAE among candidates meeting β_c, breaking ties by lower V and smaller η. If none meets the target, it selects the highest validation fraction meeting the target, then lower V, lower PMAE, and smaller η. The selected method is initialized on the full routing segment before testing; no test result selects a step size. Only DeepAR has a candidate meeting the validation target at $\beta_c = 0.85$; none does at $8/9$. In the predictor order TFT, DeepAR, DistPred, and quantile ensemble, the selected step sizes are $(0.025, 0.05, 0.05, 0.10)$ at 0.85 and $(0.025, 0.05, 0.025, 0.05)$ at $8/9$.

At $\beta_c = 0.85$, WF-ACI reduces mean V relative to the original ACI from 0.06212 to 0.03957 and raises the fraction meeting the target from 74.92% to 84.33%, with PMAE increasing from 0.8190 to 1.0765. At $\beta_c = 8/9$, ReqRC still has 20.0% lower V and a 2.60-percentage-point higher fraction meeting the target, with 22.1% higher PMAE. For this latter comparison, WF-ACI minus ReqRC differences are $+0.00615$ in V (95% interval $[0.00551, 0.00678]$), -2.60 percentage points in the fraction meeting the target ($[-2.82, -2.40]$), and -0.2585 in PMAE ($[-0.3037, -0.2187]$). These intervals use 10,000 paired channel-bootstrap replicates, resampling within each predictor–dataset block. WF-ACI improves substantially on the original ACI but remains less safe and less expensive than ReqRC.

Relative to the direct historical threshold, ReqRC lowers mean V by 10.3%, raises the window success rate by 1.12 percentage points, and increases PMAE by 0.041 (approximately 3.0%). Paired per-channel differences favor ReqRC on all four predictors; 95% bootstrap intervals exclude zero for V, the fraction meeting the target, and PMAE.

Table 12: Method design comparisons on held-out data. WF-ACI uses binary window-failure feedback; Direct uses the historical threshold without the maximum with the base level. PMAE is normalized by the training interquartile range.

Predictor	Method	$V \downarrow$	$\widehat{p}_\tau \uparrow$	PMAE $\downarrow$
TFT	WF-ACI (0.85)	0.044184	82.43%	1.0258
	WF-ACI $(8/9)$	0.036478	85.22%	1.1614
	Direct	0.033582	86.77%	1.3803
	ReqRC	0.031606	87.51%	1.3930
DeepAR	WF-ACI (0.85)	0.031685	87.52%	1.2697
	WF-ACI $(8/9)$	0.026026	89.56%	1.3608
	Direct	0.023690	90.78%	1.5375
	ReqRC	0.019907	92.34%	1.6472
DistPred	WF-ACI (0.85)	0.041884	84.01%	1.0810
	WF-ACI $(8/9)$	0.030815	87.77%	1.1744
	Direct	0.027272	89.13%	1.5017
	ReqRC	0.024992	90.08%	1.5210
Quantile ensemble	WF-ACI (0.85)	0.040536	83.36%	0.9293
	WF-ACI $(8/9)$	0.029508	87.16%	0.9925
	Direct	0.024922	88.98%	1.1387
	ReqRC	0.021723	90.20%	1.1620
Mean	WF-ACI (0.85)	0.039572	84.33%	1.0765
	WF-ACI $(8/9)$	0.030707	87.43%	1.1723
	Direct	0.027367	88.91%	1.3895
	ReqRC	0.024557	90.03%	1.4308

C.7 Cost of the Prespecified Rank

Prespecified Rank Versus Constant Levels All configurations use the same fixed six-level quantile predictions, test windows, and cost definition, and therefore share $G = 0.00697$. Constant prediction levels ignore both current and historical outcomes. All ReqRC ranks and constant levels use exactly the causal rules in Section 5.

Paired intervals obtained by resampling channels compare each ReqRC configuration with its nearest lower and higher fixed-level neighbors. Relative to fixed 0.95, ReqRC lowers V by 0.01470 (95% interval $[-0.02037, -0.01006]$) and raises $\widehat{p}_\tau$ by 5.52 points ($[3.79, 7.75]$); the raw-PMAE interval crosses zero. Relative to fixed 0.975, ReqRC's risk interval crosses zero but its raw PMAE is lower by 5.220 ($[-9.661, -1.694]$). These comparisons show how ReqRC trades shortfall against cost relative to nearby fixed levels.

In the full rank path, the empirical $\Pr(r_t = 0)$ and the 3.13% infeasible-window rate are held-out diagnostics rather than unconditional validation of Theorem 2, because valid distribution-shift bounds are unavailable. Relative to the maximum-level endpoint, ReqRC has higher selection shortfall by 0.01454 (95% interval $[0.01327, 0.01584]$), a 6.61-point lower fraction meeting the target ($[6.10, 7.12]$), and lower raw PMAE by 17.964 ($[9.448, 28.354]$).

We next scan $j = 1, \ldots, K+1$. The candidate forecasts and base prediction level remain fixed, and j is not selected using test results; $j = 9$ denotes the maximum-level branch.

Table 13: Empirical risk–cost path as ReqRC's j varies.

j	Shortfall V	Selection shortfall R	Meeting target $\widehat{p}_\tau$	Overprov. cost (raw PMAE)	Maximum-level use
1	0.05252	0.04555	78.54%	24.043	0.01%
2	0.05206	0.04509	78.63%	24.045	0.02%

j	Shortfall V	Selection shortfall R	Meeting target $\widehat{p}_\tau$	Overprov. cost (raw PMAE)	Maximum-level use
3	0.05143	0.04446	78.82%	24.048	0.09%
4	0.05065	0.04368	79.03%	24.065	0.16%
5	0.04915	0.04218	79.57%	24.181	0.29%
6	0.04519	0.03822	81.06%	25.251	1.20%
7	0.03614	0.02917	84.25%	27.331	4.31%
8	0.02151	0.01454	90.25%	31.996	19.31%
9 (highest level)	0.00697	0	96.87%	49.960	100%

This curve illustrates the empirical price of increasing the prespecified rank: higher ranks reduce selection shortfall and increase resource cost, while the predictor shortfall remains fixed.

C.8 Sensitivity Analyses

The analyses below vary the evaluation target, temporal aggregation, method settings, endpoint design, or diagnostic assumptions while leaving the main comparison unchanged. The certificate-budget calculation and real-sequence proxies are reported here because they diagnose the deployed configuration rather than prove the distributional bounds in A.5.

Only aggregate summary tables are retained in this section.

Parameter sensitivity. The scans below either hold the deployed level path fixed and vary the evaluation standard, or re-run ReqRC at a different setting; none retrains a model or uses test labels to reselect levels. The target-τ rows rescore one deployment under three targets and the block-length rows re-block the same deployment, so raw PMAE is almost unchanged across them (31.996 at all three targets and at block lengths 48 and 96, 32.154 at 192) and the maximum-level-use column is left blank. The β_{sel}, K, and D rows re-run ReqRC, so that column is reported.

Table 14: Summary of parameter-sensitivity scans.

Setting	Value	Rank used	Shortfall V	Meeting target $\widehat{p}_\tau$	Overprov. cost (raw PMAE)	Maximum-level use
Target τ	0.80	–	0.01726	92.67%	31.996	–
Target τ	0.85	–	0.02151	90.25%	31.996	–
Target τ	0.90	–	0.02728	86.25%	31.996	–
Block length	48	–	0.02586	90.83%	31.996	–
Block length	96	–	0.02151	90.25%	31.996	–
Block length	192	–	0.01287	89.20%	32.154	–
β_{sel}	0.70	7/8	0.03614	84.25%	27.331	4.31%
β_{sel}	0.80	8/8	0.02151	90.25%	31.996	19.31%
β_{sel}	0.85	8/8	0.02151	90.25%	31.996	19.31%
β_{sel}	0.90	9/8	0.00697	96.87%	49.960	100%
History length K	4	5/4	0.00697	96.87%	49.960	100%
History length K	16	15/16	0.02232	89.49%	30.849	11.42%
History length K	32	29/32	0.02435	88.78%	29.208	7.53%
Feedback delay D	2	8/8	0.02167	90.18%	32.099	19.28%
Feedback delay D	4	8/8	0.02188	89.95%	32.191	19.14%
Feedback delay D	8	8/8	0.02052	90.18%	31.759	18.86%

$\beta_{\text{sel}} = 0.80$ and 0.85 select the same integer rank and therefore yield identical deployments. For $K = 4$, $\lceil \beta_{\text{sel}}(K+1) \rceil = 5$, so the procedure enters the maximum-level branch directly. This scan varies the method's rank target, not the window-level coverage probability target β. The effect of changing D is small and nonmonotonic on these real datasets; this only reports empirical sensitivity for the present sequences and parameter range.

Maximum-level selection and higher-endpoint stress test. The endpoint is selected without test labels. On routing validation, moving from 0.99 to 0.995 changes V by -0.003222, whose paired-bootstrap upper endpoint is -0.002258, while PMAE rises by 7.11%, below the prespecified 10% budget. The predeclared rule therefore selects $u_{\max} = 0.995$.

Under fixed-maximum deployment, raising the endpoint from 0.995 to 0.999 lowers $V = G$ from 0.00697 to 0.00453 and raises the fraction meeting the target from 96.87% to 97.95%, while raw PMAE rises from 49.960 to 53.596. The median training-tail count falls from 48 to 9, and the number of channels with fewer than five tail observations rises from 0 to 100 out of 206. Full endpoint results are in `evidence/core/umax_pressure_endpoints.csv`.

For projected 0.999 minus projected 0.995, the channel-bootstrap differences are $\Delta V = -0.000123$ (95% interval $[-0.000216, -0.000048]$), $\Delta G = -0.001284$ ($[-0.001826, -0.000848]$), and $\Delta\mathrm{PMAE} = 0.004$ ($[0.001, 0.009]$). The higher endpoint is therefore informative as a stress test, but its sparse training-tail support prevents it from replacing the formal 0.995 endpoint.

Real-sequence diagnostics. On the 206 channels, the mean marginal-TV proxy rises from 0.0455 at lag 1 to 0.1807 at lag 8, and rank-event swap disagreement rises from 0.1718 to 0.2293. Deleting one versus seven recent historical values gives event-disagreement rates of 0.0086 and 0.1308. These descriptive quantities are not valid upper bounds on the joint-distribution defects or on the probability that no level works, so they do not establish the 0.0389 certification condition of Section 5.3. The diagnostic values and bootstrap intervals are in `evidence/submission_audits/exchange_diagnostic_summary.csv`.

Predictor Tail Resolution With each checkpoint fixed, we uniformly sample at most eight test start points per dataset and random seed, yielding 189 start points and 18,144 forecast coordinates. Increasing DeepAR's nested trajectory sample from 2,048 to 4,096 changes G only from 0.01135 to 0.01064. For DistPred, replacing the empirical 0.995 quantile of its 100 native samples with the sample maximum reduces G from 0.00029 to zero on this subsample, but increases the endpoint magnitude, with a mean absolute change of 0.5677. The main experiments retain the empirical quantile; the sample maximum is only a sensitivity check. The full four-configuration results, including endpoint coverage, are in `evidence/submission_audits/high_tail_resolution_summary.csv`.

C.9 Capacity-Constrained Replay

The capacity-constrained replay interprets $q_{t,h}(U_t)$ as resources reserved in advance and retrospectively computes reserved capacity and unmet demand under a normalized capacity limit. It is not a reproduction of a production scheduler and does not use capacity-limited service to overwrite the potential demand Y_t.

The replay uses exactly the same per-window results as the main study and covers 156 normalized CPU/memory resource channels from Ali and SMD, with the capacity limit set to 1.0 on the normalized scale. The independent one-channel CPU dataset is part of the main forecasting study but was not included in this replay. The capacity value has no common physical unit across the two replay datasets.

Table 15: Capacity-constrained replay on 156 Ali and SMD resource channels.

Method	Shortfall V	Meeting target $\widehat{p}_\tau$	Mean reserved	Mean idle	Mean unmet demand	Demand-weighted shortfall
Base	0.04267	83.16%	0.6638	0.0983	0.00459	0.01793
ReqRC	**0.02283**	**91.41%**	0.6918	0.1247	**0.00303**	**0.01011**
Fixed $u_{\max}$	0.00440	98.38%	0.7847	0.2154	0.00070	0.00306

Relative to Base, ReqRC has an equal-channel V difference of -0.01984, with a 95% interval of $[-0.02329, -0.01666]$; a mean-unmet-demand difference of -0.00156, with an interval of $[-0.00192, -0.00125]$; and an increase in mean reserved capacity of 0.02796, with an interval

of $[0.02258, 0.03372]$. This result maps the risk–cost tradeoff to an explicit capacity constraint, but it remains distinct from a production scheduling experiment with migration, queues, or dynamic prices.

APPENDIX D SYNTHETIC STUDIES

This appendix reports synthetic experiments regenerated under the current theoretical parameters. All probabilities are specified directly by the data-generating mechanisms. These experiments check the exact thresholds and failure directions predicted by the theory.

D.1 TWO-ATOM LOWER-BOUND CONSTRUCTION

We simulate the least-favorable distribution in Theorem 1. For prespecified ρ and β, we draw W_t from

$$P^\star = \rho\delta_{+\infty} + (1-\rho)\delta_{u_{\max}},$$

and let the policy deploy $u_{\max}$ with probability ω. The theory predicts an actual window-level coverage probability of $(1-\rho)\omega$ and a minimum endpoint probability of $\omega^\star = \beta/(1-\rho)$ for reliability over all admissible distributions.

We sweep $\rho \in \{0, 0.05, 0.10, 0.15\}$, $\beta \in \{0.80, 0.85, 0.90\}$, and, over the allowed range, $\omega/\omega^\star \in \{0.8, 1.0, 1.2\}$. Each configuration uses 200,000 independent draws. Across all 33 configurations, the maximum absolute error between the empirical probability and $(1-\rho)\omega$ is 0.00187, and the mean absolute error is 0.00064.

For example, when $\beta = 0.85$ and $\omega = 0.8\omega^\star$, the theoretical success probability is always 0.68; the empirical values across the four values of ρ range from 0.6796 to 0.6812. At $\omega = \omega^\star$, the theoretical value is 0.85 and the empirical values range from 0.8496 to 0.8514. Policies below this probability systematically miss the target, whereas the policy exactly at this probability reaches it within Monte Carlo error. The empirical probabilities agree with the sharp boundary $\beta/(1-\rho)$.

D.2 LIKELIHOOD-RATIO PERTURBATION AND WHY THE Γ BOUND IS SHARP

We start from a continuous, finite, and exchangeable minimum-feasible-level baseline Q and construct P_δ as in Appendix A.4, so that both the total-variation distance and the probability of the order-statistic event are known by construction. The theory predicts

$$P_\delta(E_j) = \frac{j}{K+1} - \delta.$$

We use $(K, j) \in \{(4,4), (8,8), (16,15)\}$ and $\delta \in \{0, 0.02, 0.05, 0.10, 0.20\}$, with 200,000 draws per configuration. Across the 15 configurations, the maximum absolute error between the empirical success probability and $j/(K+1) - \delta$ is 0.00216, and the mean absolute error is 0.00075. The construction and sharpness proof remain in Appendix A.4; the individual results and Monte Carlo intervals are in `evidence/submission_audits/tv_tilt_tightness.csv`.

D.3 DELAY AND NON-EXCHANGEABILITY STRESS TESTS

This experiment illustrates how stale feedback, abrupt shifts, and gradual drift affect empirical rank behavior, and checks the directions predicted by the independent-drift and stationary-mixing corollaries in Appendix A.5.

The generator directly produces a seven-state minimum-feasible-level sequence. States $0, \ldots, 5$ correspond in order to the six finite calibration nodes, and state 6 corresponds to $+\infty$; rank comparisons use only this ordering. The initial innovation distribution is

$$\pi^{(0)} = (0.18, 0.17, 0.17, 0.15, 0.13, 0.10, 0.10).$$

Let the path length be $N_{\text{path}} = 564$. Gradual drift uses $\phi_t = t/(N_{\text{path}} - 1)$, whereas abrupt drift uses $\phi_t = \mathbf{1}\{t \geq N_{\text{path}}/2\}$. For drift mass m, let $s_t = m\phi_t$ and move masses $0.6s_t$ and $0.4s_t$ from states 0 and 1 to states 6 and 5, respectively. The time-varying innovation distribution is therefore

$$\pi_t = \pi^{(0)} + s_t(-0.6, -0.4, 0, 0, 0, 0.4, 0.6).$$

The restriction $m \leq 0.10$ keeps every component nonnegative. A one-step persistence parameter a introduces dependence through

$$\mathbb{P}(S_t = s \mid S_{t-1} = r) = a\,\mathbf{1}\{s = r\} + (1 - a)\pi_t(s).$$

When $a = 0$, every state is independently sampled from the current innovation distribution; larger a makes the previous state persist longer. The first 64 states form completed routing history, and the next 500 states are test decision points. At a test point t with delay D, the method may use only states revealed through $t - D$. It selects the most recent $K = 8$ states and uses their $j = 8$ order statistic as the threshold. Success means that the current state does not exceed this threshold. When both are state 6, this event corresponds to $r_t = 0$; it does not mean that the window coverage target is reachable.

We sweep $a \in \{0, 0.5, 0.9\}$, $m \in \{0, 0.05, 0.10\}$, abrupt and gradual change, and $D \in \{1, 2, 4, 8\}$. Each configuration uses 300 prespecified random seeds, with seed identifiers shared across configurations to reduce comparison noise. For each seed, we first compute the success fraction over the 500 test points on its path; each cell then reports the mean and empirical 2.5%/97.5% quantiles across the 300 paths. The table below additionally averages equally over the three values of m and both change types solely to show the interaction between dependence and delay. The result file also records $\max_t d_{\text{TV}}(\pi_t, \pi_{t-1})$; this is a known adjacent-step change in the innovation distributions, not Γ for the generated state sequence or a defect measure for real data.

Table 16: Delay stress test under persistent and drifting latent states.

Persistence a	$D = 1$	$D = 2$	$D = 4$	$D = 8$
0	0.9385	0.9385	0.9382	0.9376
0.5	0.9306	0.8954	0.8680	0.8597
0.9	0.9669	0.9372	0.8860	0.8114

Entries are empirical values of $\Pr(r_t = 0)$. Delay has little effect for independent sequences, but performance degrades markedly as feedback becomes stale under persistence: at high persistence, it falls from 0.9669 for $D = 1$ to 0.8114 for $D = 8$. For abrupt drift with mass 0.10 and $a = 0.5$, the corresponding values are 0.9327, 0.8984, 0.8716, and 0.8631. These fixed-$j = 8$ results illustrate the separate effects of drift and dependence in Appendix A.5; they do not estimate real-data distributional bounds. Choosing j using valid bounds is evaluated separately in D.4.

D.4 Choosing j with Valid Bounds

This experiment directly instantiates the adaptive rank extension in Section 5.3. Minimum feasible levels are mutually independent over time, and the complete time-varying marginal schedule and the fixed probability ρ of having no feasible level are known before deployment. Thus the product-swap bound, adjacent-drift bound, and ρ bound are deterministic common inputs that hold simultaneously for every candidate rank. The method still computes thresholds only from levels revealed at least D windows earlier and reads the current level only after committing the level.

We set $K = 8$, $D = 1$, and $\beta = \tau = 0.85$. We cross $\rho \in \{0, 0.02, 0.05\}$, finite-state drift mass in $\{0, 0.02, 0.05, 0.10\}$, and gradual versus abrupt change, yielding 24 prespecified scenarios. Each scenario uses 300 independent paths with 64 routing windows and 500 test windows. The exact-bound variant uses the complete known marginals to compute the product-swap bound at every decision. The uniform-$d_{\max}^{\text{adj}}$ variant uses only the scenario's maximum adjacent TV drift and is consequently more conservative. Synthetic PMAE retains the positive-part overprovisioning definition but is measured in generator-specific units.

Policy	Mean rank	Fallback	Min. Success LB	$\widehat{p}_\tau$	V	syn. PMAE
ReqRC (rank 8)	8.000	0.00%	0.6780	93.07%	0.04853	0.19428
Certificate extension (exact)	8.336	33.59%	**0.8516**	94.26%	0.04021	0.19483
Certificate extension (uniform $d^{\mathrm{adj}}_{\max}$)	8.583	58.33%	**0.8675**	95.50%	0.03149	0.19540
Always maximum	9.000	100.00%	0.9500	97.66%	0.01638	0.19640

Both rules using valid bounds have a theoretical window-level coverage lower bound above 0.85 at every decision in all 24 scenarios. Relative to ReqRC, the exact-bound variant raises empirical $\widehat{p}_\tau$ by 1.19 points and lowers V by 17.1% at 0.284% higher synthetic PMAE; the coarser uniform-$d^{\mathrm{adj}}_{\max}$ bound invokes the maximum-level branch more often and pays more. With $K = 8$, $j = 8$ already selects the largest historical value, so the extension chooses between $j = 8$ and maximum-level fallback in this experiment. This controlled study demonstrates choosing j, using the fallback, and reporting guarantees with known bounds; tighter bounds require fewer maximum-level fallbacks. It is not an estimator of real-sequence ε or ρ.

APPENDIX E REPRODUCIBILITY

E.1 CODE AND DATA

The anonymous artifact includes the ReqRC implementation and prespecified configurations. Frozen method results, validation selections, mixture weights, and plotting data for the continuous-level comparison are provided in `evidence/continuous_controllers`; its README maps these files to the main tables and figures.

E.2 SOFTWARE ENVIRONMENT

The verification environment uses Python 3.12 and PyTorch 2.5 on a single NVIDIA RTX 4060 Laptop GPU. The runtimes in Appendix B.7 are direct log aggregates and do not claim that all models were rerun in one container.

E.3 CLAIMS AND SUPPORTING EVIDENCE

Table 17: Where to find the main results and their supporting evidence.

Result	Location
Long-run versus window-level coverage	Proposition 1; proof in A.1; real-data comparison in Table 1 and C.4.
Predictor/selection shortfall and transfer under expansion	Section 4.2; proofs in A.1; causal information gap in A.3; component analysis in C.6.
Uniform reliability and maximum-level probability	Theorem 1; proof and cost bound in A.2; two-atom simulation in D.1.
Delayed ranks under nonexchangeability	Theorem 2; definitions, proofs, time-series cases, and the no-structure boundary in A.4–A.5; controlled studies in D.2–D.3.
ReqRC's minimum update, distribution-free minimality within the base-retaining class, and effect across base methods	Section 5.1; minimality in Corollary 1, proof in A.6; 40 paired comparisons, window-failure ACI, and direct historical threshold in C.6.
Risk–cost comparisons and sensitivity	Figures 2 and 4; matched comparisons in C.6; aggregation alternatives in C.5; scans of j and other parameters in C.7–C.8; capacity replay in C.9.

Result	Location
Valid-bound rank selection and certification limits	Section 5.3; selection, conditional guarantees, and unobservable-change impossibility in A.6–A.7; known-generator selection experiment in D.4.
Reproducibility and scope	Predictors and method adaptations in B; datasets, metrics, and configuration in C.1–C.3; code and environment in E.1–E.2; limitations in E.4.

E.4 Limitations and scope

1. **Closed-loop demand.** The present theory treats Y_t as potential demand whose distribution the current level does not change. Extending the shortfall analysis to settings where deployment changes demand requires a model of those effects.
2. **Beyond ordered forecasts.** The decomposition in Section 4.2 needs only a minimum, but the minimum-feasible-level equivalence, the highest-level closed form, and rank minimality all use a right-continuous monotone ordering.
3. **Bounds available before deployment on real data.** Theorem 2 can use global Γ, coordinate-swap bounds for the event of interest, or explicit drift and mixing specializations. Appendix A.7 separates a reference lower bound estimable from revealed history from a one-sided drift condition that must hold before acting; Appendix D.4 demonstrates rank selection using known bounds. On the eight real datasets these bounds are unavailable, so the main evaluation uses the prespecified $j = 8$, which follows from the target alone; the reported results there are empirical.
4. **Upper-tail estimation.** The 0.995 endpoint depends on finite training data and model extrapolation: DistPred has 100 native samples and DeepAR's empirical upper tail depends on the number of trajectories. C.8 therefore reports endpoint-resolution sensitivity rather than treating either as an exact population quantile.
5. **Aggregation dependence.** Different aggregate weightings are different cost estimands and are reported separately. Under PMAE normalized by the training interquartile range, ReqRC's six-level quantile predictor is nondominated under all three.
6. **Method adaptation.** Replaying 11 methods over the same ordered forecasts isolates update-rule differences under one online protocol; reproducing each paper's complete end-to-end system would answer a different question that also varies the predictor.
7. **Statistical uncertainty.** The central two-factor mechanism effects and capacity-replay differences have stable stratified channel-bootstrap intervals, while smaller differences can remain sensitive to the dataset mixture and aggregation weights; additional datasets remain more informative than further resampling.
8. **Resource interpretation.** PMAE and the normalized capacity replay translate the theory into reservation–shortage tradeoffs under a static cap; production validation should additionally model migration, startup, queuing, dynamic prices, and physical units.
9. **Method positioning.** ReqRC is the method implied by the minimum-level result: it preserves the base prediction level whenever that level already exceeds the historical order-statistic threshold. Cross-predictor experiments show that this single update rule remains competitive with four predictive families, while final risk–cost performance belongs to the predictor–method pair.