SEQUENTIAL CONFIDENCE SETS FOR COVERAGE-CONSTRAINED CONFORMAL MODEL SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern conformal forecasting systems often maintain several adaptive pipelines that differ in base forecasters, conformity scores, calibration windows, and update rules. Comparing them is difficult because coverage is a hard constraint, whereas efficiency should be optimized only among feasible pipelines. We formulate this problem as sequential inference for a stochastic constrained argmin. At each time step, the target is the set of minimum-cost pipelines satisfying multiple prefix-average conditional miscoverage constraints. We introduce Coverage-Constrained Sequential Model Confidence Sets (CC-SMCS), which separate certifiably feasible, possibly feasible, and possibly constrained-optimal pipelines. Using simultaneous martingale confidence sequences, CC-SMCS projects a rectangular confidence region onto the constrained argmin and admits an exact closed-form rule. With probability at least $1-\delta$, it contains every constrained-optimal pipeline simultaneously over all times. This finite-sample guarantee requires no stationarity or mixing assumptions and remains valid under data-dependent stopping. We also establish an impossibility result for safe certification at the coverage boundary and extend the construction to delayed multi-horizon feedback and outcome-dependent efficiency objectives.

1 INTRODUCTION

Conformal prediction (CP) turns point or distributional predictors into set-valued predictions with explicit coverage targets (Vovk et al., 2005; Angelopoulos & Bates, 2023). In practice, however, a forecasting system rarely relies on a single conformal pipeline. Different base forecasters, conformity scores, calibration windows, nominal levels, and online update rules can all produce prediction sets with different coverage and efficiency, and their relative performance may change under distribution shift. The resulting problem is how to compare a library of adaptive conformal procedures while accounting for uncertainty about both coverage and efficiency.

The main difficulty is that coverage and efficiency play different roles. Coverage is a constraint: a narrow prediction set is not useful if it systematically undercovers. Efficiency, by contrast, should be optimized only among pipelines that satisfy the required coverage constraints. A scalar score that combines width and undercoverage replaces this constrained problem by a different objective and can favor an under-covering method. A two-stage rule that first screens methods for coverage and then selects the narrowest survivor has a different problem: failure to reject infeasibility is not evidence of feasibility. Near the coverage boundary, an infeasible but narrow method may survive because there is not yet enough evidence to exclude it, while a feasible boundary method may be difficult to certify.

This motivates a different inferential target. Rather than selecting a single pipeline, we ask which pipelines can still be the optimizer of a coverage-constrained problem at each time. At time t, every candidate pipeline has several prefix-average conditional miscoverage constraints and an observed prequential efficiency cost. A pipeline is feasible when all monitored coverage constraints are satisfied, and the target $\mathcal{M}_t^\star$ is the set of feasible pipelines with minimum cost. Because both the feasible set and the identity of the optimizer may change over time, we seek a sequence of random model sets $(\widehat{\mathcal{M}}_t^{\mathrm{proj}})_{t\geq 1}$ satisfying

$$\mathbb{P}\Big(\forall t \geq 1 :\ \mathcal{M}_t^\star \subseteq \widehat{\mathcal{M}}_t^{\mathrm{proj}}\Big) \geq 1-\delta. \tag{1}$$

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.

We introduce a *Coverage-Constrained Sequential Model Confidence Set* (CC-SMCS) for this target. It separates certified and possible feasibility and projects simultaneous coverage confidence sequences onto the constrained argmin. The rectangular projection has a closed-form rule computable in $O(MR)$ time. Under prequential measurability and bounded miscoverage increments, (1) holds without stationarity, mixing, or exchangeability assumptions and remains valid under data-dependent stopping.

The target is method comparison: membership in $\widehat{\mathcal{M}}_t^{\mathrm{proj}}$ does not imply certified feasibility or predictive coverage after selection.

Contributions.

1. We formulate conformal pipeline comparison as sequential inference for a stochastic constrained argmin with an unknown, time-varying feasible set. The framework supports multiple coverage constraints, with extensions to delayed multi-horizon feedback and outcome-dependent efficiency objectives.
2. We introduce CC-SMCS, which separates certifiably feasible, possibly feasible, and possibly constrained-optimal pipelines. For an observed prequential objective, we derive an exact rectangular projection rule and prove finite-sample, time-uniform oracle coverage using martingale confidence sequences.
3. We give sufficient conditions for oracle recovery and a finite-sample lower bound near the coverage boundary. At the boundary, we show that even a known strict cost advantage need not permit high-probability singleton identification under uniform oracle coverage.

2 Related Work

Model confidence sets and argmin inference. The classical Model Confidence Set (MCS) contains the best models with asymptotic confidence for a fixed evaluation sample (Hansen et al., 2011). Sequential MCS provides nonasymptotic time-uniform coverage for models ranked by scalar conditional-average losses (Arnold et al., 2026). Related methods test sequential predictive ability using betting processes (Choe & Ramdas, 2024). Model Prediction Sets target the empirically best model for the next period under a long-run model-set coverage criterion (Li & Zheng, 2025). Other work studies confidence sets for the index of the minimum component of a noisy mean vector (Zhang et al., 2024; Kim & Ramdas, 2025). We instead minimize an observed objective over an unknown, time-varying feasible set.

Online conformal prediction and sequential risk control. Conformal methods address temporal dependence and distribution shift (Barber et al., 2023; Xu & Xie, 2021; Zaffran et al., 2022). Adaptive conformal inference controls long-run coverage error (Gibbs & Candès, 2021; 2024), while strongly adaptive methods consider multiple time scales (Bhatnagar et al., 2023). Risk-control extensions address non-exchangeable data (Farinhas et al., 2024) and arbitrary inspection times (Hultberg et al., 2026). Multi-model methods learn a predictor to deploy (Hajihashemi & Shen, 2024), combine conformity scores through Bayesian model averaging (Bhagwat et al., 2025), or weight residuals using gating similarities (Kong et al., 2026). These methods construct or deploy prediction sets. CC-SMCS instead compares pipelines under monitored coverage constraints.

Risk-controlled selection and constrained identification. Learn-then-test treats risk control as multiple hypothesis testing (Angelopoulos et al., 2025), and Pareto Testing combines risk certification with optimization (Laufer-Goldshtein et al., 2023). Conformal risk control calibrates prediction sets for general monotone losses (Angelopoulos et al., 2024). Constrained best-arm identification uses adaptive sampling to identify a fixed optimizer under unknown feasibility constraints (Lardy et al., 2025; Cai & Kandasamy, 2026). We instead evaluate adaptive pipelines on a common outcome stream and cover all current constrained optimizers over time, allowing excluded pipelines to re-enter.

Conformal model selection and coverage assessment. Efficiency-oriented selection motivates procedures that preserve predictive validity after selection (Liang et al., 2024; Hegazy et al., 2025; Wang & Wang, 2026). Conformal Prediction Assessment evaluates conditional coverage and develops

a selection criterion using a learned reliability estimator (Zhou et al., 2026). Our target is sequential method comparison under prefix-average conditional coverage constraints.

3 Setup and Target

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space with an increasing filtration $(\mathcal{G}_t)_{t\geq 0}$. The sigma-field $\mathcal{G}_{t-1}$ contains all information available when the prediction for outcome Y_t is issued.

There are $M < \infty$ candidate conformal pipelines indexed by $[M]$ and $R < \infty$ monitored constraints indexed by $[R]$. A candidate is the complete conformal procedure, including its base forecaster, conformity score, calibration sample or rolling window, nominal level, update rule, hyperparameters, and randomization. Thus, candidates may share the same base forecaster while differing in their conformal layers. Each pipeline may adapt arbitrarily to the past, but its internal state and output at time t must be $\mathcal{G}_{t-1}$-measurable.

For every method $m \in [M]$ and constraint $r \in [R]$, let C_{mrt} be the prediction set for Y_t. A conformal pipeline is designed to produce prediction sets at a prescribed nominal coverage level. Our comparison framework evaluates these sets through the monitored coverage constraints defined below. For notational simplicity, the index r may encode a nominal level, a tail, a forecast horizon after the delayed-feedback construction, or a predictable subgroup.

Define the miscoverage indicator as

$$B_{mrt} = \mathbf{1}\{Y_t \notin C_{mrt}\}. \tag{2}$$

Let $\tau_r \in (0,1)$ be the tolerated miscoverage rate, and let $a_{rt} \in [0,1]$ be a $\mathcal{G}_{t-1}$-measurable exposure weight. For example, $a_{rt} = 1$ gives an overall constraint, while $a_{rt} = \mathbf{1}\{S_t = s\}$ selects a regime S_t known at prediction time. Define

$$X_{mrt} = a_{rt}(B_{mrt} - \tau_r), \qquad \mu_{mrt} = \mathbb{E}[X_{mrt} \mid \mathcal{G}_{t-1}],$$
$$A_{r,t} = \sum_{s=1}^{t} a_{rs}, \qquad \Gamma_{mr,t} = \begin{cases} A_{r,t}^{-1} \sum_{s=1}^{t} \mu_{mrs}, & A_{r,t} > 0, \\ 0, & A_{r,t} = 0. \end{cases}$$

The inequality $\Gamma_{mr,t} \leq 0$ requires the exposure-weighted prefix average of the conditional miscoverage probabilities to be at most τ_r. For a subgroup indicator, it is the average of the conditional risks over the periods when that subgroup is active. The target differs from usual marginal conformal coverage and is weaker than pointwise conditional coverage.

In practice, we allow for a prespecified tolerance

$$\tau_r = \alpha_r + \varepsilon_r, \qquad \varepsilon_r \geq 0, \tag{3}$$

where $1 - \alpha_r$ is the nominal conformal level. The quantity ε_r is a prespecified practical tolerance that relaxes the coverage requirement. Section 5.3 explains its role in certification.

3.1 Observed prequential efficiency

Let $c_{mt} \geq 0$ be a $\mathcal{G}_{t-1}$-measurable set cost, such as interval length, Lebesgue measure, classification-set cardinality, or a prespecified weighted average of these quantities across levels and horizons. Let $w_t \geq 0$ be predictable weights and assume $W_t = \sum_{s=1}^{t} w_s > 0$. Define the weighted prequential cost as

$$Q_{m,t} = \frac{1}{W_t} \sum_{s=1}^{t} w_s c_{ms}. \tag{4}$$

The objective is observed exactly because the set cost is known when the prediction is issued. This design isolates uncertainty about feasibility. Appendix G.2 extends the method to outcome-dependent efficiency losses by adding confidence sequences for pairwise objective differences.

Definition 3.1 (Dynamic feasible and oracle sets)**.** At time t, define

$$\mathcal{V}_t^\star = \{m \in [M] : \Gamma_{mr,t} \leq 0 \text{ for every } r \in [R]\}, \qquad \mathcal{M}_t^\star = \arg\min_{m \in \mathcal{V}_t^\star} Q_{m,t},$$

with $\mathcal{M}_t^\star = \varnothing$ if $\mathcal{V}_t^\star = \varnothing$.

Both sets are random and may change with t. Multiple pipelines may tie for the constrained optimum. Importantly, the target does not assume that any candidate is a true data-generating model or that the process is stationary.

Definition 3.2 (Coverage-constrained sequential model confidence set)**.** A sequence $(\widehat{\mathcal{M}}_t^{\mathrm{proj}})_{t\geq 1}$ is a level-$1-\delta$ CC-SMCS for $(\mathcal{M}_t^\star)_{t\geq 1}$ if

$$\mathbb{P}\Big(\forall t \geq 1 :\ \mathcal{M}_t^\star \subseteq \widehat{\mathcal{M}}_t^{\mathrm{proj}}\Big) \geq 1-\delta. \tag{5}$$

We call (5) *finite-sample time-uniform coverage*. Its operational consequence is anytime validity: the analyst may inspect the set repeatedly and stop according to the observed data without invalidating the guarantee.

4 Coverage-Constrained Sequential Model Confidence Sets

4.1 Simultaneous confidence sequences for coverage risk

Let

$$S_{mr,t} = \sum_{s=1}^{t} X_{mrs}. \tag{6}$$

Following the empirical-Bernstein construction of Howard et al. (2021), define the predictable estimate

$$\widehat{B}_{mrt} = \begin{cases} \tau_r, & A_{r,t-1} = 0, \\ \dfrac{\sum_{s=1}^{t-1} a_{rs} B_{mrs}}{A_{r,t-1}}, & A_{r,t-1} > 0, \end{cases} \qquad \widehat{X}_{mrt} = a_{rt}(\widehat{B}_{mrt} - \tau_r). \tag{7}$$

By construction, $\widehat{B}_{mrt}$ is $\mathcal{G}_{t-1}$-measurable and belongs to $[0,1]$. Since $a_{rt} \in [0,1]$, both X_{mrt} and $\widehat{X}_{mrt}$ belong to the fixed interval $[-\tau_r, 1-\tau_r]$, whose length is one. Define the empirical intrinsic-time process

$$V_{mr,t} = \sum_{s=1}^{t} (X_{mrs} - \widehat{X}_{mrs})^2. \tag{8}$$

For $\rho > 0$ and a two-sided error probability $\eta \in (0,1)$, let $b_{\eta,\rho}$ denote the gamma-exponential mixture boundary defined in Appendix B.1. The empirical-Bernstein construction gives

$$\mathbb{P}\left(\forall t \geq 1 : \left| S_{mr,t} - \sum_{s=1}^{t} \mu_{mrs} \right| \leq b_{\eta,\rho}(V_{mr,t}) \right) \geq 1-\eta. \tag{9}$$

Lemma B.2 in Appendix B.1 proves (9). Choose fixed weights $\pi_{mr} > 0$ satisfying $\sum_{m,r} \pi_{mr} = 1$, and set $\eta_{mr} = \delta \pi_{mr}$. For $A_{r,t} > 0$, define

$$L_{mr,t} = \frac{S_{mr,t} - b_{\eta_{mr},\rho}(V_{mr,t})}{A_{r,t}}, \qquad U_{mr,t} = \frac{S_{mr,t} + b_{\eta_{mr},\rho}(V_{mr,t})}{A_{r,t}}. \tag{10}$$

If $A_{r,t} = 0$, set $L_{mr,t} = -\infty$ and $U_{mr,t} = +\infty$. A union bound over the method–constraint pairs yields

$$\mathbb{P}\left(\forall t, m, r :\ L_{mr,t} \leq \Gamma_{mr,t} \leq U_{mr,t}\right) \geq 1-\delta. \tag{11}$$

Lemma B.3 in Appendix B.1 proves (11).

4.2 Certified and possibly feasible methods

A method m is ruled infeasible if there exists at least one constraint r such that $L_{mr,t} > 0$. On the joint confidence event, this implies $\Gamma_{mr,t} > 0$, so method m violates at least one coverage constraint. The set of methods that have not been ruled infeasible is

$$\widehat{\mathcal{V}}_t^{\mathrm{pos}} = \{m : L_{mr,t} \leq 0 \text{ for every } r\}. \tag{12}$$

Among these methods, the certified feasible set is

$$\widehat{\mathcal{V}}_t^{\mathrm{cert}} = \{m : U_{mr,t} \le 0 \text{ for every } r\}. \tag{13}$$

A method in $\widehat{\mathcal{V}}_t^{\mathrm{pos}}$ is possibly feasible because every confidence interval contains at least one nonpositive value, while a method in $\widehat{\mathcal{V}}_t^{\mathrm{cert}}$ is certified feasible because every value in each confidence interval is nonpositive. Therefore, $\widehat{\mathcal{V}}_t^{\mathrm{cert}} \subseteq \widehat{\mathcal{V}}_t^{\mathrm{pos}}$. On the joint confidence event, $\widehat{\mathcal{V}}_t^{\mathrm{cert}} \subseteq \mathcal{V}_t^\star \subseteq \widehat{\mathcal{V}}_t^{\mathrm{pos}}$ holds for every t. If a constraint has received no exposure, its confidence interval is $(-\infty, +\infty)$. It does not rule out feasibility, since its lower bound is nonpositive, but it prevents certification since its upper bound is positive.

4.3 Projection onto the constrained argmin

At time t, define the rectangular confidence region $\mathcal{C}_t = \times_{m=1}^{M} \times_{r=1}^{R} [L_{mr,t}, U_{mr,t}]$. For $\gamma = (\gamma_{mr}) \in \mathcal{C}_t$, let

$$\mathcal{V}(\gamma) = \{m \in [M] : \gamma_{mr} \le 0 \text{ for every } r \in [R]\}, \qquad \mathcal{M}^\star(\gamma) = \operatorname*{arg\,min}_{m \in \mathcal{V}(\gamma)} Q_{m,t}$$

when $\mathcal{V}(\gamma) \neq \varnothing$. Define the projection confidence set by

$$\widehat{\mathcal{M}}_t^{\mathrm{proj}} = \bigcup_{\substack{\gamma \in \mathcal{C}_t \\ \mathcal{V}(\gamma) \neq \varnothing}} \mathcal{M}^\star(\gamma). \tag{14}$$

The rectangular structure yields the following closed-form characterization.

Proposition 4.1. *Fix any time t. Suppose that the coordinate intervals in $\mathcal{C}_t$ are nonempty and each contains at least one finite value, and that $Q_{m,t} \in \mathbb{R}$ is fixed for every m. Define*

$$q_t^{\mathrm{cert}} = \min_{j \in \widehat{\mathcal{V}}_t^{\mathrm{cert}}} Q_{j,t}, \qquad \min \varnothing = +\infty. \tag{15}$$

Then

$$\widehat{\mathcal{M}}_t^{\mathrm{proj}} = \left\{ m \in \widehat{\mathcal{V}}_t^{\mathrm{pos}} : \; Q_{m,t} \le q_t^{\mathrm{cert}} \right\}. \tag{16}$$

The result is deterministic and uses only the rectangular confidence region and the fixed objective values. Appendix B.2 gives the proof. We therefore define the operational CC-SMCS by the right-hand side of (16), and Algorithm 1 summarizes the full procedure.

5 Theoretical Guarantees

The following assumption collects the measurability and design conditions used for the confidence sequences and the main coverage result.

Assumption 5.1. The integers M, R, tolerances τ_r, confidence level $\delta \in (0,1)$, allocations $\pi_{mr} > 0$ with $\sum_{m,r} \pi_{mr} = 1$, objective definition, and mixture parameter $\rho > 0$ are fixed before evaluation. For every t, m, r, the set C_{mrt}, exposure a_{rt}, cost c_{mt}, and weight w_t are $\mathcal{G}_{t-1}$-measurable; Y_t and the resulting miss indicators are $\mathcal{G}_t$-measurable. The costs are finite and $\sum_{s=1}^{t} w_s > 0$ whenever $Q_{m,t}$ is reported.

5.1 Time-uniform coverage and optional stopping

Theorem 5.2. *Under Assumption 5.1, construct the confidence sequences in (10) with fixed allocations $\eta_{mr} = \delta \pi_{mr}$. Then*

$$\mathbb{P}\Big(\forall t \ge 1 : \; \widehat{\mathcal{V}}_t^{\mathrm{cert}} \subseteq \mathcal{V}_t^\star \subseteq \widehat{\mathcal{V}}_t^{\mathrm{pos}}, \quad \mathcal{M}_t^\star \subseteq \widehat{\mathcal{M}}_t^{\mathrm{proj}}\Big) \ge 1 - \delta. \tag{17}$$

Because Theorem 5.2 holds simultaneously over all times, no additional correction is needed for repeated inspection or data-dependent stopping. For any stopping time T, the probability that $T < \infty$ and $\mathcal{M}_T^\star \not\subseteq \widehat{\mathcal{M}}_T^{\mathrm{proj}}$ is at most δ. Examples include stopping when $|\widehat{\mathcal{M}}_t^{\mathrm{proj}}| = 1$, when a prespecified method is certified, or when a computational budget is exhausted.

5.2 CONTRACTION OF THE ORACLE CONFIDENCE SET

For $A_{r,t} > 0$, define $\mathrm{rad}_{mr,t} := b_{\eta_{mr},\rho}(V_{mr,t})/A_{r,t}$, and set $\mathrm{rad}_{mr,t} = +\infty$ when $A_{r,t} = 0$. Under persistent exposure $A_{r,t} \asymp t$ almost surely for every r and balanced allocations $\eta_{mr} \asymp \delta/(MR)$, Proposition B.5 gives

$$\mathrm{rad}_{mr,t} = O\left(\sqrt{\frac{\log t + \log(MR/\delta)}{t}}\right) \longrightarrow 0. \tag{18}$$

On the joint confidence event, method m is excluded at time t whenever either

$$\begin{aligned} &\exists r: \quad \Gamma_{mr,t} > 2\mathrm{rad}_{mr,t}, \qquad \text{or} \\ &\exists j: \quad Q_{j,t} < Q_{m,t} \quad \text{and} \quad \Gamma_{jr,t} + 2\mathrm{rad}_{jr,t} \le 0 \quad \text{for every } r. \end{aligned}$$

Thus $\widehat{\mathcal{M}}_t^{\mathrm{proj}} = \mathcal{M}_t^\star$ whenever every non-oracle method satisfies either condition. Under persistent exposure, this equality holds eventually if there is $\zeta > 0$ such that, for all sufficiently large t, each non-oracle method either violates a constraint by at least ζ or has a strictly cheaper competitor satisfying $\Gamma_{jr,t} \le -\zeta$ for every r. Appendix B.4 proves these statements.

5.3 THE COVERAGE BOUNDARY

Section 5.2 shows how feasibility margins allow CC-SMCS to remove non-oracle methods. At the boundary, even a known strict cost advantage need not suffice. Consider $Y_t = B_t$ with $(B_t)_{t\ge1}$ i.i.d. Bernoulli(p), a single coverage constraint with threshold $\tau \in (0,1)$, and $a_{1t} = w_t = 1$. The fixed intervals $C_{11t} = [0, 1/2]$ and $C_{21t} = [0,1]$ have width costs $1/2 < 1$ and miscoverage probabilities p and 0. The oracle is therefore $\{1\}$ when $p \le \tau$ and $\{2\}$ when $p > \tau$: the cost ordering is known, but the identity of the oracle depends on whether the cheaper candidate is feasible.

Proposition 5.3. *Let T_{cert} be a stopping time for the natural filtration of $(B_t)_{t\ge1}$, where $T_{\mathrm{cert}} < \infty$ means that $p \le \tau$ is certified. For any $\eta \in (0,1)$,*

$$\sup_{p>\tau} \mathbb{P}_p(T_{\mathrm{cert}} < \infty) \le \eta \quad \Longrightarrow \quad \mathbb{P}_\tau(T_{\mathrm{cert}} < \infty) \le \eta.$$

For any model-set rule $(\widehat{\mathcal{S}}_t)_{t\ge1}$ adapted to this filtration and satisfying (5) for every $p \in (0,1)$, the first exclusion of candidate 2 is a safe certification time with $\eta = \delta$. Consequently,

$$\mathbb{P}_\tau\Big(\exists t \ge 1 : \widehat{\mathcal{S}}_t = \{1\}\Big) \le \delta.$$

Thus, a unique oracle with strictly lower observed cost does not guarantee high-probability singleton identification under uniform oracle coverage. Appendix B.5 proves these claims and gives a finite-sample lower bound: for fixed threshold and error levels, high-probability identification requires at least order $(\tau - p)^{-2}$ observations as $p \uparrow \tau$.

The prespecified tolerance in (3) creates a strict feasibility margin: if $p \le \alpha_r$ and $\tau_r = \alpha_r + \varepsilon_r$ with $\varepsilon_r > 0$, then $p - \tau_r \le -\varepsilon_r$. Under persistent exposure, this relaxed constraint is eventually certified on the joint confidence event.

6 EXPERIMENTS

6.1 SIMULATIONS

To evaluate the validity and power of CC-SMCS, we design three simulations that share the same variation process and point predictor but differ in oracle dynamics. These settings allow us to assess both oracle retention and adaptation to time-varying oracle models. Each simulation uses $T = 3000$ observations and is replicated N=1,000 times.

6.1.1 DATA GENERATING PROCESS

To obtain the population oracle for evaluating CC-SMCS, we directly generate the point forecast as $\widehat{Y}_t = Y_{t-1}$ and the outcome as $Y_t = Y_{t-1} + \sigma_t Z_t$, where $Z_t \overset{\mathrm{iid}}{\sim} \mathcal{N}(0,1)$. Following Hansen et al.

(2011), we generate $H_t = -\frac{\phi}{2(1+\phi)} + \phi H_{t-1} + \sqrt{\phi}\varepsilon_t$, where $\varepsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ and $\phi \in (0,1)$, with stationary initialization $H_0 \sim \mathcal{N}\left(-\frac{\phi}{2(1-\phi^2)}, \frac{\phi}{1-\phi^2}\right)$. We set $\phi = 0.5$. The predictable baseline variation for forecasting period t is defined as $\sigma_t^0 = \exp\{H_{t-1}\}$.

Let $h_t = H_{t-1} + \frac{\phi}{2(1-\phi^2)}$, and $\sigma_t = \sigma_t^0 \exp\{l_t|h_t|\}$. We design six methods. For M2–M6, we construct the candidate prediction intervals as

$$C_{mr,t} = \left[\widehat{Y}_t - z\sigma_{mr,t}, \widehat{Y}_t + z\sigma_{mr,t}\right], \qquad \sigma_{mr,t} = \sigma_t^0 \exp\{\theta_{mr}|h_t|\}$$

For M1, $C_{1r,t} = \left[\widehat{Y}_t + \kappa\sigma_{4r,t} - z\sigma_{4r,t}, \widehat{Y}_t + \kappa\sigma_{4r,t} + z\sigma_{4r,t}\right]$, $\kappa = 0.5$, where we instead use a shifted center interval so that M1 has the same interval width as M4 but exhibits undercoverage due to the shifted center.

For M2–M6, we set $(\theta_{2r}, \ldots, \theta_{6r}) = \xi(-2,-1,0,1,2)$, with $\xi = 0.15$ and $z = \Phi^{-1}(0.975)$. In Simulation 1, M2 and M3 undercover, M4 is calibrated, and M5 and M6 overcover with wider intervals. Under the true scale σ_t, the conditional miscoverage probability of a centered candidate is

$$P\left\{Y_t \notin C_{mr,t} \mid \mathcal{G}_{t-1}\right\} = 2\left[1 - \Phi\left(z\frac{\sigma_{mr,t}}{\sigma_t}\right)\right].$$

Simulation 1 (Stationary). We set $l_t = 0$ for all t, so M4 lies exactly on the coverage boundary and is the fixed oracle.

Simulation 2 (Structural Change). We introduce a permanent structural break after $t_0 = 400$ by setting:

$$l_t = \begin{cases} 0, & t \le 400, \\ 1.75\xi, & t > 400. \end{cases}$$

Simulation 3 (Linear Drift). We let $l_t = 2\xi(t-1)/(T-1)$, $t = 1, \ldots, T$, so that the true forecast-error scale changes gradually over time.

Further details of the simulation designs are provided in Appendix E.

6.1.2 Simulation Results

We evaluate finite-horizon oracle retention using

$$P_{\text{TU}} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{1}\left\{\mathcal{M}_{n,t}^{\star} \subseteq \widehat{\mathcal{M}}_{n,t}^{\text{proj}} \text{ for all } t \le T\right\},$$

which estimates the retention probability in Theorem 5.2 over the evaluation horizon. We also report the mean confidence radius across methods, $\overline{\text{rad}}_t$, at three time points to illustrate contraction. For the boundary method $M_\tau = M_4$ in Simulation 1, we report the fraction of runs with certification by T,

$$\widehat{p}_{\tau,T} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{1}\left\{\exists\, t \le T : M_\tau \in \widehat{\mathcal{V}}_{n,t}^{\text{cert}}\right\}.$$

Table 1 shows oracle-retention rates close to one, decreasing mean radii, and rare boundary certification. The population oracle set is nonempty at every evaluated time in all replications.

We summarize the model-set paths by the inclusion frequency $\text{PV}_{m,t}$ and mean set size PI_t:

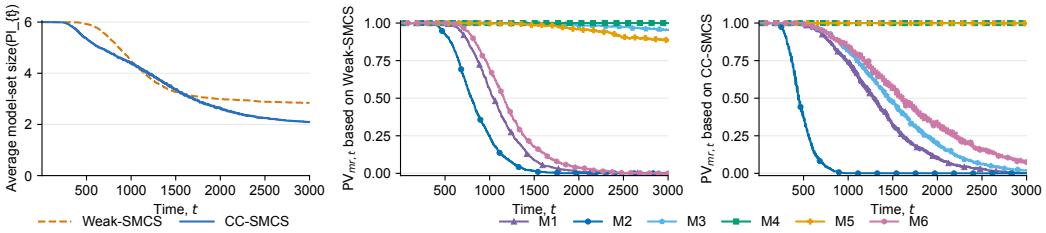
$$\text{PV}_{m,t} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{1}\left\{m \in \widehat{\mathcal{M}}_{n,t}^{\text{proj}}\right\}, \qquad \text{PI}_t = \frac{1}{N}\sum_{n=1}^{N} \left|\widehat{\mathcal{M}}_{n,t}^{\text{proj}}\right|. \tag{19}$$

Figure 1 compares CC-SMCS with Weak-SMCS (Arnold et al., 2026) at $\delta = 0.10$. Benchmark details and additional results are given in Appendices C and E.4, respectively.

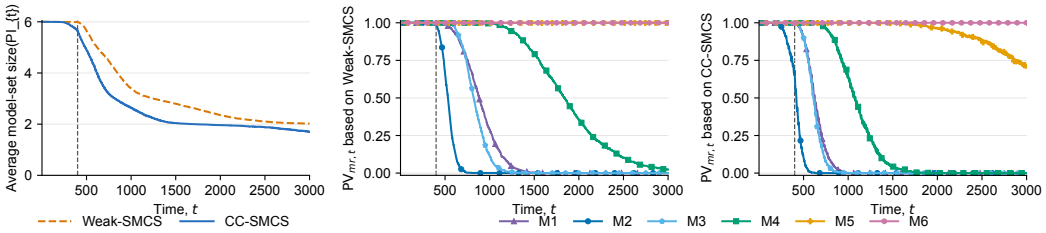
Table 1: Simulation results

Simulations	δ	P_{TU}	$\bar{\text{rad}}_{t_{0.25}}$	$\bar{\text{rad}}_{t_{0.50}}$	$\bar{\text{rad}}_{t_{0.75}}$	$\widehat{p}_{\tau,T}$
Simulation 1	0.10	1.000	0.0361	0.0234	0.0186	0.001
	0.25	1.000	0.0326	0.0211	0.0168	0.001
Simulation 2	0.10	1.000	0.0402	0.0283	0.0231	–
	0.25	1.000	0.0363	0.0255	0.0209	–
Simulation 3	0.10	0.999	0.0372	0.0252	0.0209	–
	0.25	0.999	0.0336	0.0227	0.0188	–

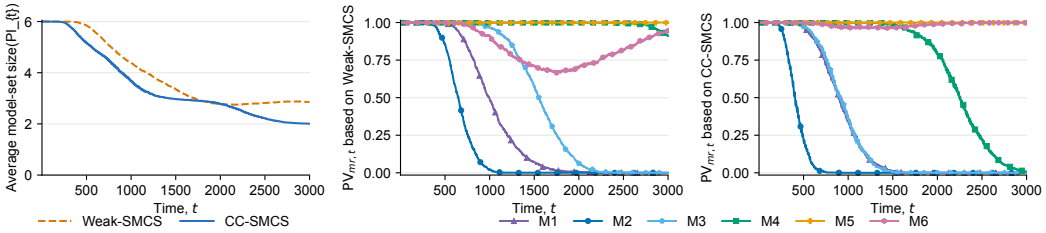
Note: $t_q = qT$ for $q \in \{0.25, 0.50, 0.75\}$. $\widehat{p}_{\tau,T}$ is reported only for Simulation 1, where M_4 remains on the coverage boundary throughout the evaluation.



(a) Simulation 1



(b) Simulation 2



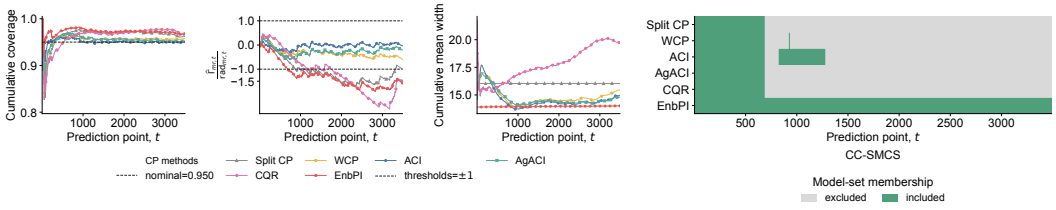
(c) Simulation 3

Figure 1: The performance of simulations with $\delta = 0.10$. The vertical dashed line in Figure 1b indicates the structural break.

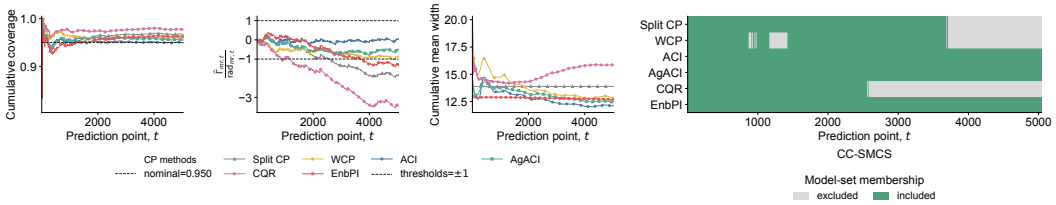
In Simulation 1, CC-SMCS reduces the set while retaining M_4. The continued inclusion of M_5 is consistent with the difficulty of certifying the boundary oracle M_4. In Simulations 2 and 3, the oracle moves from M_4 to M_5 and then M_6 after a structural break or under gradual drift. CC-SMCS excludes M_4 as evidence of undercoverage accumulates. Under gradual drift, the inclusion frequency of M_6 first decreases and then increases as it becomes the constrained oracle. Weak-SMCS follows different paths because it compares standardized Winkler-score differences rather than interval width subject to coverage constraints.

6.2 Empirical Study

In this section, we present empirical applications of CC-SMCS using four real-world time-series datasets: French Electricity Price, Electricity Load, Weather, and Wind Power, which were used by Zaffran et al. (2022), Hong et al. (2016), Wu et al. (2023), and Xu & Xie (2021), respectively. The Electricity Load and Weather datasets were also utilized by Lin et al. (2022) and Chen et al. (2024), respectively. For each dataset, we employ DLinear (Zeng et al., 2023), PatchTST (Nie et al., 2023), and TimesNet (Wu et al., 2023) as forecasting backbones, then construct prediction intervals by classical split conformal prediction (Split CP) (Lei et al., 2018), weighted conformal prediction (WCP) (Barber et al., 2023), Adaptive Conformal Inference (ACI) (Gibbs & Candès, 2021), Aggregated ACI (AgACI) (Zaffran et al., 2022), conformalized quantile regression (CQR) (Romano et al., 2019), and EnbPI (Xu & Xie, 2021). Implementation details are provided in Appendix F.



(a) French Electricity Price



(b) Electricity Load

Figure 2: Empirical results for DLinear at 95% nominal coverage with $\delta = 0.10$.

Figure 2 presents the DLinear results for French Electricity Price and Electricity Load. Additional results are in Appendix F.4. The constrained oracle is unobservable, so coverage and width help explain the reported model sets. The ratio $\widehat{\Gamma}_{mr,t}/\mathrm{rad}_{mr,t}$, with $\widehat{\Gamma}_{mr,t} = S_{mr,t}/A_{r,t}$, summarizes the coverage evidence: values above 1 rule out feasibility, while values at most -1 certify it. On both datasets, CQR is excluded despite high coverage when $Q_{\mathrm{CQR},t} > q_t^{\mathrm{cert}}$, illustrating how CC-SMCS separates coverage certification from efficiency comparison.

7 Conclusion

We introduced CC-SMCS to compare conformal pipelines under coverage constraints. The method separates feasibility from efficiency and provides finite-sample, time-uniform oracle coverage. Our theory gives conditions for oracle recovery and identifies limits at the coverage boundary. Simulations and real data examples illustrate how the model sets respond to changes in coverage and interval width.

Reproducibility Statement

The prequential protocol, target parameters, assumptions, and operational rule are stated explicitly in Sections 3 and 4. Algorithm 1 in Appendix A gives the full pseudocode. Appendix B provides complete proofs for the empirical-Bernstein confidence sequence, the exact projection formula, the main time-uniform guarantee, and the boundary result. Appendix G.1 gives the delayed-feedback construction and its proof. Section 6, Appendix E and Appendix F illustrate the simulation process, random seed, and reproducibility conditions, while providing the results of the empirical study and relative reproducibility details.

AI USE STATEMENT

A large language model was used to assist with language editing. The research questions, definitions, theoretical claims, proofs, experimental design, citation verification, and final wording remain the responsibility of the authors. The authors independently reviewed and verified all content included in the submission.

REFERENCES

Anastasios Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, volume 2024, pp. 55198–55218, 2024.

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *Foundations and Trends in Machine Learning*, 16(4): 494–591, 2023.

Anastasios N Angelopoulos, Stephen Bates, Emmanuel J Candès, Michael I Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025.

Sebastian Arnold, Georgios Gavrilopoulos, Benedikt Schulz, and Johanna F. Ziegel. Sequential model confidence sets. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2026. doi: 10.1093/jrsssb/qkag066.

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

Pankaj Bhagwat, Linglong Kong, and Bei Jiang. Cbma: Improving conformal prediction through bayesian model averaging. In *International Conference on Learning Representations*, volume 2025, pp. 240–258, 2025.

Aadyot Bhatnagar, Huan Wang, Caiming Xiong, and Yu Bai. Improved online conformal prediction via strongly adaptive online learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2337–2363, 2023.

Ting Cai and Kirthevasan Kandasamy. Constrained best arm identification with tests for feasibility. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40(24), pp. 19808–19816, 2026. doi: 10.1609/aaai.v40i24.39063.

Baiting Chen, Zhimei Ren, and Lu Cheng. Conformalized time series with semantic features. *Advances in Neural Information Processing Systems*, 37:121449–121474, 2024.

Yo Joong Choe and Aaditya Ramdas. Comparing sequential forecasters. *Operations Research*, 72(4): 1368–1387, 2024.

António Farinhas, Chrysoula Zerva, Dennis Ulmer, and Andre Martins. Non-exchangeable conformal risk control. In *International Conference on Learning Representations*, volume 2024, pp. 50952–50966, 2024.

Isaac Gibbs and Emmanuel J. 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.

Erfan Hajihashemi and Yanning Shen. Multi-model ensemble conformal prediction in dynamic environments. *Advances in Neural Information Processing Systems*, 37:118678–118700, 2024.

Peter R. Hansen, Asger Lunde, and James M. Nason. The model confidence set. *Econometrica*, 79 (2):453–497, 2011.

Mahmoud Hegazy, Liviu Aolaritei, Michael I. Jordan, and Aymeric Dieuleveut. Valid selection among conformal sets. In *Advances in Neural Information Processing Systems*, volume 38, pp. 193550–193581, 2025.

Tao Hong, Pierre Pinson, Shu Fan, Hamidreza Zareipour, Alberto Troccoli, and Rob J Hyndman. Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond, 2016.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021.

Bror Hultberg, Dave Zachariah, and Antônio H. Ribeiro. Anytime-valid conformal risk control. *arXiv preprint arXiv:2602.04364*, 2026.

Ilmun Kim and Aaditya Ramdas. Locally minimax optimal confidence sets for the best model. *arXiv preprint arXiv:2503.21639*, 2025.

Jingsen Kong, Wenlu Tang, Dezheng Kong, Linglong Kong, Guangren Yang, and Bei Jiang. Adaptive conformal prediction via mixture-of-experts gating similarity. In *International Conference on Learning Representations*, volume 2026, pp. 94169–94191, 2026.

Tyron Lardy, Christina Katsimerou, and Wouter M. Koolen. Constrained best arm identification. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Bracha Laufer-Goldshtein, Adam Fisch, Regina Barzilay, and Tommi S. Jaakkola. Efficiently controlling multiple risks with pareto testing. In *International Conference on Learning Representations*, 2023.

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.

Shibo Li and Yao Zheng. Online conformal model selection for nonstationary time series. *arXiv preprint arXiv:2506.05544*, 2025.

Ruiting Liang, Wanrong Zhu, and Rina Foygel Barber. Conformal prediction after efficiency-oriented model selection. *arXiv preprint arXiv:2408.07066*, 2024.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Conformal prediction with temporal quantile adjustments. *Advances in Neural Information Processing Systems*, 35:31017–31030, 2022.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023.

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

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

Yuhao Wang and Tengyao Wang. Localized conformal model selection. *arXiv preprint arXiv:2602.19284*, 2026.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *International Conference on Learning Representations*, 2023.

Chen Xu and Yao Xie. Conformal prediction interval for dynamic time-series. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11559–11569, 2021.

Margaux Zaffran, Olivier Féron, Yannig Goude, Julie Josse, and Aymeric Dieuleveut. Adaptive conformal predictions for time series. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 25834–25866, 2022.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 11121–11128, 2023.

Tianyu Zhang, Hao Lee, and Jing Lei. Winners with confidence: Discrete argmin inference with an application to model selection. *arXiv preprint arXiv:2408.02060*, 2024.

Zheng Zhou, Xiangfei Zhang, Chongguang Tao, and Yuhong Yang. Conformal prediction assessment: A framework for conditional coverage evaluation and selection. *arXiv preprint arXiv:2603.27189*, 2026.

Table 2: Main notation.

Symbol	Meaning
C_{mrt}	Prediction set from method m under constraint r at round t
B_{mrt}	Miscoverage indicator $\mathbf{1}\{Y_t \notin C_{mrt}\}$
a_{rt}	Predictable exposure or subgroup weight
$\Gamma_{mr,t}$	Prefix-average conditional excess miscoverage
$Q_{m,t}$	Observed prefix-average set cost
$\mathcal{V}_t^\star$	True feasible method set
$\mathcal{M}_t^\star$	Minimum-cost subset of $\mathcal{V}_t^\star$
$\widehat{\mathcal{V}}_t^{\mathrm{cert}}$	Methods certified feasible
$\widehat{\mathcal{V}}_t^{\mathrm{pos}}$	Methods not ruled infeasible
$\widehat{\mathcal{M}}_t^{\mathrm{proj}}$	Coverage-constrained sequential model confidence set

A Notation and Algorithm

Table 2 summarizes the main objects. The distinction between the three reported sets is essential: $\widehat{\mathcal{V}}_t^{\mathrm{cert}}$ is an inner confidence set for feasibility, $\widehat{\mathcal{V}}_t^{\mathrm{pos}}$ is an outer confidence set for feasibility, and $\widehat{\mathcal{M}}_t^{\mathrm{proj}}$ is an outer confidence set for the constrained optimizer.

The main event in Theorem 5.2 simultaneously implies

$$\widehat{\mathcal{V}}_t^{\mathrm{cert}} \subseteq \mathcal{V}_t^\star \subseteq \widehat{\mathcal{V}}_t^{\mathrm{pos}}, \qquad \mathcal{M}_t^\star \subseteq \widehat{\mathcal{M}}_t^{\mathrm{proj}} \quad \text{for every } t.$$

It does not imply $\widehat{\mathcal{M}}_t^{\mathrm{proj}} \subseteq \mathcal{V}_t^\star$.

Algorithm 1 summarizes the CC-SMCS procedure. If no method is certified, then $q_t^{\mathrm{cert}} = +\infty$ and the output equals $\widehat{\mathcal{V}}_t^{\mathrm{pos}}$. A positive practical tolerance ε_r creates a strict feasibility margin for methods satisfying the nominal target $p \le \alpha_r$, which permits eventual certification under persistent exposure. Section 5.3 gives the precise statement.

B Proofs

B.1 Empirical-Bernstein confidence sequence

For $\rho > 0$ and $z, v \ge 0$, define the gamma-exponential mixture function

$$\mathfrak{m}_\rho(z, v) = \frac{\rho^\rho e^{z+v} \gamma(v+\rho, z+v+\rho)}{\gamma(\rho,\rho)(z+v+\rho)^{v+\rho}}, \tag{20}$$

where $\gamma(a, h) = \int_0^h y^{a-1} e^{-y}\, dy$ is the lower incomplete gamma function. The details of the selection of ρ are provided in Appendix D. For a two-sided error probability $\eta \in (0, 1)$, define

$$b_{\eta,\rho}(v) = \inf\left\{ z \ge 0 : \mathfrak{m}_\rho(z, v) \ge \frac{2}{\eta} \right\}. \tag{21}$$

The threshold $2/\eta$ allocates error probability $\eta/2$ to each tail.

Lemma B.1. *Let $\mu_{mrt} := \mathbb{E}[X_{mrt} \mid \mathcal{G}_{t-1}]$, $D_{mrt} := X_{mrt} - \mu_{mrt}$, and $\Delta_{mrt} := (X_{mrt} - \widehat{X}_{mrt})^2$. Suppose that $a_{rt} \in [0, 1]$ is $\mathcal{G}_{t-1}$-measurable, $\tau_r \in [0, 1]$, $B_{mrt} \in \{0, 1\}$, and $\widehat{B}_{mrt} \in [0, 1]$ is $\mathcal{G}_{t-1}$-measurable. Then*

$$\mathbb{E}[D_{mrt} \mid \mathcal{G}_{t-1}] = 0.$$

Moreover, for every $\lambda \in [0, 1)$, with $\psi_E(\lambda) := -\log(1-\lambda) - \lambda$,

$$\mathbb{E}\left[\exp\{\lambda D_{mrt} - \psi_E(\lambda)\Delta_{mrt}\} \mid \mathcal{G}_{t-1}\right] \le 1. \tag{22}$$

Proof. By definition,

$$\mathbb{E}[D_{mrt} \mid \mathcal{G}_{t-1}] = \mathbb{E}[X_{mrt} - \mu_{mrt} \mid \mathcal{G}_{t-1}] = \mathbb{E}[X_{mrt} \mid \mathcal{G}_{t-1}] - \mu_{mrt} = 0.$$

Algorithm 1 Empirical-Bernstein CC-SMCS

Require: Candidate prediction sets $\{C_{mrt}\}_{m\in[M],r\in[R]}$, thresholds $\{\tau_r\}_{r\in[R]}$, predictable exposures $\{a_{rt}\}_{r\in[R]}$, efficiency costs $\{Q_{m,t}\}_{m\in[M]}$, confidence level δ, allocations $\{\pi_{mr}\}$ with $\sum_{m,r}\pi_{mr}=1$, and mixture parameter $\rho>0$.

Set $A_{r,0}=0$, $S_{mr,0}=0$, and $V_{mr,0}=0$ for all m,r.
Set $\eta_{mr}=\delta\pi_{mr}$ for all m,r.
for $t=1,2,\dots$ **do**
  for $r=1,\dots,R$ **do**
    for $m=1,\dots,M$ **do**
      if $A_{r,t-1}=0$ **then**
        Set $\widehat{B}_{mrt}=\tau_r$.
      else
        Set $\widehat{B}_{mrt}=\dfrac{\sum_{s=1}^{t-1}a_{rs}B_{mrs}}{A_{r,t-1}}$.
      end if
      Set $\widehat{X}_{mrt}=a_{rt}(\widehat{B}_{mrt}-\tau_r)$.
    end for
  end for
  Observe Y_t.
  for $r=1,\dots,R$ **do**
    Update $A_{r,t}=A_{r,t-1}+a_{rt}$.
    for $m=1,\dots,M$ **do**
      Compute $B_{mrt}=\mathbf{1}\{Y_t\notin C_{mrt}\}$ and $X_{mrt}=a_{rt}(B_{mrt}-\tau_r)$.
      Update $S_{mr,t}=S_{mr,t-1}+X_{mrt}$.
      Update $V_{mr,t}=V_{mr,t-1}+(X_{mrt}-\widehat{X}_{mrt})^2$.
      if $A_{r,t}>0$ **then**
        Compute $b_{\eta_{mr},\rho}(V_{mr,t})$ from $\mathfrak{m}_\rho(b_{\eta_{mr},\rho}(V_{mr,t}),V_{mr,t})=2/\eta_{mr}$.
        Set $L_{mr,t}=\dfrac{S_{mr,t}-b_{\eta_{mr},\rho}(V_{mr,t})}{A_{r,t}}$ and $U_{mr,t}=\dfrac{S_{mr,t}+b_{\eta_{mr},\rho}(V_{mr,t})}{A_{r,t}}$.
      else
        Set $L_{mr,t}=-\infty$ and $U_{mr,t}=+\infty$.
      end if
    end for
  end for
  Construct $\widehat{\mathcal{V}}_t^{\mathrm{pos}}=\{m: L_{mr,t}\le 0 \text{ for every } r\}$.
  Construct $\widehat{\mathcal{V}}_t^{\mathrm{cert}}=\{m: U_{mr,t}\le 0 \text{ for every } r\}$.
  Set $q_t^{\mathrm{cert}}=\min_{j\in\widehat{\mathcal{V}}_t^{\mathrm{cert}}}Q_{j,t}$, with $\min\varnothing:=+\infty$.
  Construct $\widehat{\mathcal{M}}_t^{\mathrm{proj}}=\{m\in\widehat{\mathcal{V}}_t^{\mathrm{pos}}: Q_{m,t}\le q_t^{\mathrm{cert}}\}$.
  Output $\widehat{\mathcal{V}}_t^{\mathrm{pos}}$, $\widehat{\mathcal{V}}_t^{\mathrm{cert}}$, and $\widehat{\mathcal{M}}_t^{\mathrm{proj}}$.
end for

Since $B_{mrt}\in\{0,1\}$, $\widehat{B}_{mrt}\in[0,1]$, and $a_{rt}\in[0,1]$, both $X_{mrt}=a_{rt}(B_{mrt}-\tau_r)$ and $\widehat{X}_{mrt}=a_{rt}(\widehat{B}_{mrt}-\tau_r)$ belong to $[-\tau_r,1-\tau_r]$. Hence

$$X_{mrt}-\widehat{X}_{mrt}=a_{rt}(B_{mrt}-\widehat{B}_{mrt})\in[-a_{rt},a_{rt}]\subseteq[-1,1]$$

almost surely. Following the self-normalization argument in the proof of Theorem 4 of Howard et al. (2021), for every $\lambda\in[0,1)$ and $x\ge -1$,

$$\exp\{\lambda x-\psi_E(\lambda)x^2\}\le 1+\lambda x. \tag{23}$$

Applying (23) to $x=X_{mrt}-\widehat{X}_{mrt}$ gives

$$\exp\{\lambda(X_{mrt}-\widehat{X}_{mrt})-\psi_E(\lambda)(X_{mrt}-\widehat{X}_{mrt})^2\}\le 1+\lambda(X_{mrt}-\widehat{X}_{mrt}).$$

Multiplying both sides by $\exp\{\lambda(\widehat{X}_{mrt}-\mu_{mrt})\}$ yields

$$\exp\{\lambda(X_{mrt}-\mu_{mrt})-\psi_E(\lambda)(X_{mrt}-\widehat{X}_{mrt})^2\}\le\exp\{\lambda(\widehat{X}_{mrt}-\mu_{mrt})\}\{1+\lambda(X_{mrt}-\widehat{X}_{mrt})\}.$$

Taking conditional expectations with respect to $\mathcal{G}_{t-1}$ and using the $\mathcal{G}_{t-1}$-measurability of $\widehat{X}_{mrt}$ gives

$$\begin{aligned}&\mathbb{E}\left[\exp\{\lambda(X_{mrt}-\mu_{mrt})-\psi_E(\lambda)(X_{mrt}-\widehat{X}_{mrt})^2\}\mid\mathcal{G}_{t-1}\right]\\&\le\exp\{\lambda(\widehat{X}_{mrt}-\mu_{mrt})\}\{1+\lambda(\mu_{mrt}-\widehat{X}_{mrt})\}\\&\le\exp\{\lambda(\widehat{X}_{mrt}-\mu_{mrt})\}\exp\{\lambda(\mu_{mrt}-\widehat{X}_{mrt})\}=1,\end{aligned}$$

where the last inequality uses $1-y\le e^{-y}$. □

Lemma B.2. *Let*

$$Z_{mr,t}:=\sum_{s=1}^{t}D_{mrs},\qquad\Delta_{mrt}:=(X_{mrt}-\widehat{X}_{mrt})^2,\qquad V_{mr,t}:=\sum_{s=1}^{t}\Delta_{mrs}=\sum_{s=1}^{t}(X_{mrs}-\widehat{X}_{mrs})^2.$$

For every $\eta\in(0,1)$ and $\rho>0$, with $b_{\eta,\rho}$ defined in (21),

$$\mathbb{P}\left(\forall t\ge1:\ |Z_{mr,t}|\le b_{\eta,\rho}(V_{mr,t})\right)\ge1-\eta.\tag{24}$$

Proof. Fix $\lambda\in[0,1)$ and define

$$M_t(\lambda):=\exp\{\lambda Z_{mr,t}-\psi_E(\lambda)V_{mr,t}\}.$$

Because D_{mrt} and Δ_{mrt} are $\mathcal{G}_t$-measurable, $M_t(\lambda)$ is $\mathcal{G}_t$-measurable. By Lemma B.1,

$$\mathbb{E}\left[\exp\{\lambda D_{mrt}-\psi_E(\lambda)\Delta_{mrt}\}\mid\mathcal{G}_{t-1}\right]\le1.$$

Using $Z_{mr,t}=Z_{mr,t-1}+D_{mrt}$ and $V_{mr,t}=V_{mr,t-1}+\Delta_{mrt}$, we have

$$\begin{aligned}M_t(\lambda)&=\exp\{\lambda Z_{mr,t-1}+\lambda D_{mrt}-\psi_E(\lambda)V_{mr,t-1}-\psi_E(\lambda)\Delta_{mrt}\}\\&=M_{t-1}(\lambda)\exp\{\lambda D_{mrt}-\psi_E(\lambda)\Delta_{mrt}\}.\end{aligned}$$

Taking conditional expectations and using that $M_{t-1}(\lambda)$ is $\mathcal{G}_{t-1}$-measurable gives

$$\begin{aligned}\mathbb{E}[M_t(\lambda)\mid\mathcal{G}_{t-1}]&=M_{t-1}(\lambda)\mathbb{E}\left[\exp\{\lambda D_{mrt}-\psi_E(\lambda)\Delta_{mrt}\}\mid\mathcal{G}_{t-1}\right]\\&\le M_{t-1}(\lambda).\end{aligned}\tag{25}$$

Since $M_t(\lambda)\ge0$, taking expectations and iterating the preceding inequality yields

$$\mathbb{E}[M_t(\lambda)]\le\mathbb{E}[M_{t-1}(\lambda)]\le\cdots\le\mathbb{E}[M_0(\lambda)]=1.$$

Thus $M_t(\lambda)$ is integrable, and $(M_t(\lambda))_{t\ge0}$ is a nonnegative $(\mathcal{G}_t)_{t\ge0}$-supermartingale with $M_0(\lambda)=1$.

Since

$$e^{\lambda z-\psi_E(\lambda)v}=e^{\lambda(z+v)}(1-\lambda)^v,$$

following the gamma-exponential mixture construction in Proposition 9 of Howard et al. (2021), for $\rho>0$ define

$$f_\rho(\lambda)=\frac{\rho^\rho}{\gamma(\rho,\rho)}(1-\lambda)^{\rho-1}e^{-\rho(1-\lambda)},\qquad0\le\lambda<1.$$

The factor $\rho^\rho/\gamma(\rho,\rho)$ is the normalizing constant, since

$$\int_0^1(1-\lambda)^{\rho-1}e^{-\rho(1-\lambda)}\,d\lambda=\rho^{-\rho}\gamma(\rho,\rho).$$

Define the mixture process

$$M_t:=\int_0^1M_t(\lambda)f_\rho(\lambda)\,d\lambda.$$

Since the integrand is nonnegative, Tonelli's theorem gives

$$\begin{aligned}\mathbb{E}[M_t]&=\mathbb{E}\left[\int_0^1M_t(\lambda)f_\rho(\lambda)\,d\lambda\right]\\&=\int_0^1\mathbb{E}[M_t(\lambda)]f_\rho(\lambda)\,d\lambda\\&\le\int_0^1f_\rho(\lambda)\,d\lambda=1.\end{aligned}$$

Thus M_t is integrable. By conditional Tonelli's theorem and (25),

$$\begin{aligned}\mathbb{E}[M_t \mid \mathcal{G}_{t-1}] &= \int_0^1 \mathbb{E}[M_t(\lambda) \mid \mathcal{G}_{t-1}] f_\rho(\lambda)\, d\lambda \\ &\le \int_0^1 M_{t-1}(\lambda) f_\rho(\lambda)\, d\lambda = M_{t-1}.\end{aligned}$$

Moreover,

$$M_0 = \int_0^1 M_0(\lambda) f_\rho(\lambda)\, d\lambda = \int_0^1 f_\rho(\lambda)\, d\lambda = 1.$$

Hence $(M_t)_{t\ge 0}$ is a nonnegative $(\mathcal{G}_t)_{t\ge 0}$-supermartingale. By Ville's inequality,

$$\mathbb{P}\left(\exists t \ge 1 :\ M_t \ge \frac{2}{\eta}\right) \le \frac{\eta}{2}. \tag{26}$$

For $z \ge 0$ and $v \ge 0$, the corresponding mixture integral is

$$\int_0^1 e^{\lambda z - \psi_E(\lambda) v} f_\rho(\lambda)\, d\lambda = \frac{\rho^\rho}{\gamma(\rho,\rho)} \int_0^1 e^{\lambda(z+v)} (1-\lambda)^{v+\rho-1} e^{-\rho(1-\lambda)}\, d\lambda.$$

Set $y = (1-\lambda)(z+v+\rho)$. Then

$$\begin{aligned}\int_0^1 e^{\lambda z - \psi_E(\lambda) v} f_\rho(\lambda)\, d\lambda &= \frac{\rho^\rho e^{z+v}}{\gamma(\rho,\rho)} \int_0^1 (1-\lambda)^{v+\rho-1} e^{-(1-\lambda)(z+v+\rho)}\, d\lambda \\ &= \frac{\rho^\rho e^{z+v}}{\gamma(\rho,\rho)(z+v+\rho)^{v+\rho}} \int_0^{z+v+\rho} y^{v+\rho-1} e^{-y}\, dy \\ &= \frac{\rho^\rho e^{z+v} \gamma(v+\rho, z+v+\rho)}{\gamma(\rho,\rho)(z+v+\rho)^{v+\rho}} \\ &= \mathfrak{m}_\rho(z,v),\end{aligned}$$

which is the expression used in (20). For fixed $v \ge 0$,

$$\frac{\partial}{\partial z} \mathfrak{m}_\rho(z,v) = \int_0^1 \lambda e^{\lambda z - \psi_E(\lambda) v} f_\rho(\lambda)\, d\lambda > 0.$$

Thus $\mathfrak{m}_\rho(z,v)$ is continuous and strictly increasing in $z \ge 0$. By the definition of $b_{\eta,\rho}(v)$, if $Z_{mr,t} \ge b_{\eta,\rho}(V_{mr,t})$, then $Z_{mr,t} \ge 0$ and $M_t \ge 2/\eta$. Therefore (26) gives

$$\mathbb{P}\left(\exists t \ge 1 :\ Z_{mr,t} \ge b_{\eta,\rho}(V_{mr,t})\right) \le \frac{\eta}{2}. \tag{27}$$

For the lower tail, consider the reflected variables

$$X^-_{mrt} := -X_{mrt}, \qquad \widehat{X}^-_{mrt} := -\widehat{X}_{mrt}, \qquad \mu^-_{mrt} := -\mu_{mrt}.$$

Then

$$X^-_{mrt} - \widehat{X}^-_{mrt} = -(X_{mrt} - \widehat{X}_{mrt}) \in [-1, 1],$$

and

$$(X^-_{mrt} - \widehat{X}^-_{mrt})^2 = (X_{mrt} - \widehat{X}_{mrt})^2.$$

Moreover,

$$X^-_{mrt} - \mu^-_{mrt} = -(X_{mrt} - \mu_{mrt}),$$

so the corresponding centered cumulative process and empirical intrinsic time satisfy

$$Z^-_{mr,t} := \sum_{s=1}^t (X^-_{mrs} - \mu^-_{mrs}) = -Z_{mr,t}, \qquad V^-_{mr,t} := \sum_{s=1}^t (X^-_{mrs} - \widehat{X}^-_{mrs})^2 = V_{mr,t}.$$

Applying (23) to $X^-_{mrt} - \widehat{X}^-_{mrt}$ and repeating the preceding one-step argument gives, for every $\lambda \in [0, 1)$,

$$M^-_t(\lambda) := \exp\{-\lambda Z_{mr,t} - \psi_E(\lambda) V_{mr,t}\} \tag{28}$$

as a nonnegative $(\mathcal{G}_t)_{t\geq 0}$-supermartingale. Mixing over the same density f_ρ gives

$$M_t^- := \int_0^1 M_t^-(\lambda) f_\rho(\lambda)\, d\lambda. \tag{29}$$

By the same Tonelli argument, $(M_t^-)_{t\geq 0}$ is a nonnegative supermartingale with $M_0^- = 1$. Hence Ville's inequality gives

$$\mathbb{P}\left(\exists t \geq 1 :\ M_t^- \geq \frac{2}{\eta}\right) \leq \frac{\eta}{2}. \tag{30}$$

If $-Z_{mr,t} \geq b_{\eta,\rho}(V_{mr,t})$, then $-Z_{mr,t} \geq 0$ and $M_t^- \geq 2/\eta$. Therefore

$$\mathbb{P}\left(\exists t \geq 1 :\ -Z_{mr,t} \geq b_{\eta,\rho}(V_{mr,t})\right) \leq \frac{\eta}{2}. \tag{31}$$

A union bound over (27) and (31) yields

$$\mathbb{P}\left(\forall t \geq 1 :\ |Z_{mr,t}| \leq b_{\eta,\rho}(V_{mr,t})\right) \geq 1 - \eta. \tag{32}$$

This proves (24). □

Lemma B.3. *Choose fixed weights $\pi_{mr} > 0$ with $\sum_{m,r} \pi_{mr} = 1$ and set $\eta_{mr} = \delta\pi_{mr}$. For $A_{r,t} > 0$, let $L_{mr,t}$ and $U_{mr,t}$ be the lower and upper bounds defined in (10). If $A_{r,t} = 0$, set $L_{mr,t} = -\infty$ and $U_{mr,t} = +\infty$. Then*

$$\mathbb{P}\left(\forall t, m, r :\ L_{mr,t} \leq \Gamma_{mr,t} \leq U_{mr,t}\right) \geq 1 - \delta. \tag{33}$$

Proof. For each (m, r), apply Lemma B.2 with error level $\eta_{mr} = \delta\pi_{mr}$. A union bound over all method–constraint pairs gives

$$\begin{aligned}
&\mathbb{P}\left(\forall m, r,\ \forall t \geq 1 :\ |Z_{mr,t}| \leq b_{\eta_{mr},\rho}(V_{mr,t})\right) \\
&\quad \geq 1 - \sum_{m=1}^{M}\sum_{r=1}^{R} \mathbb{P}\left(\exists t \geq 1 :\ |Z_{mr,t}| > b_{\eta_{mr},\rho}(V_{mr,t})\right) \\
&\quad \geq 1 - \sum_{m=1}^{M}\sum_{r=1}^{R} \eta_{mr} = 1 - \delta.
\end{aligned} \tag{34}$$

On this joint event, if $A_{r,t} > 0$, then by the definition of $\Gamma_{mr,t}$,

$$Z_{mr,t} = S_{mr,t} - \sum_{s=1}^{t} \mu_{mrs} = S_{mr,t} - A_{r,t}\Gamma_{mr,t}.$$

Hence

$$|S_{mr,t} - A_{r,t}\Gamma_{mr,t}| \leq b_{\eta_{mr},\rho}(V_{mr,t}),$$

which is equivalent to

$$-b_{\eta_{mr},\rho}(V_{mr,t}) \leq S_{mr,t} - A_{r,t}\Gamma_{mr,t} \leq b_{\eta_{mr},\rho}(V_{mr,t}).$$

Rearranging gives

$$S_{mr,t} - b_{\eta_{mr},\rho}(V_{mr,t}) \leq A_{r,t}\Gamma_{mr,t} \leq S_{mr,t} + b_{\eta_{mr},\rho}(V_{mr,t}).$$

Since $A_{r,t} > 0$, division by $A_{r,t}$ preserves both inequalities:

$$\frac{S_{mr,t} - b_{\eta_{mr},\rho}(V_{mr,t})}{A_{r,t}} \leq \Gamma_{mr,t} \leq \frac{S_{mr,t} + b_{\eta_{mr},\rho}(V_{mr,t})}{A_{r,t}}.$$

By the definition of the confidence bounds in (10), this is exactly

$$L_{mr,t} \leq \Gamma_{mr,t} \leq U_{mr,t}.$$

If $A_{r,t} = 0$, the same inclusion holds by construction. The result follows from (34). □

B.2 EXACT RECTANGULAR PROJECTION

Proof of Proposition 4.1. We prove both inclusions.

First take $m \in \widehat{\mathcal{M}}_t^{\mathrm{proj}}$. There exists $\gamma \in \mathcal{C}_t$ such that $m \in \mathcal{M}^\star(\gamma)$. Because m is feasible under γ, for every r we have $\gamma_{mr} \le 0$ and $L_{mr,t} \le \gamma_{mr}$; hence $m \in \widehat{\mathcal{V}}_t^{\mathrm{pos}}$. Next let $j \in \widehat{\mathcal{V}}_t^{\mathrm{cert}}$. Then $U_{jr,t} \le 0$ for every r, so every $\gamma \in \mathcal{C}_t$ makes j feasible. Optimality of m under the selected γ implies $Q_{m,t} \le Q_{j,t}$. Therefore, $Q_{m,t} \le q_t^{\mathrm{cert}}$, proving membership in the right-hand side of (16).

Conversely, suppose $m \in \widehat{\mathcal{V}}_t^{\mathrm{pos}}$ and $Q_{m,t} \le q_t^{\mathrm{cert}}$. We construct a matrix $\gamma \in \mathcal{C}_t$ under which m is optimal. Since $m \in \widehat{\mathcal{V}}_t^{\mathrm{pos}}$, for every r we have $L_{mr,t} \le 0$, so the interval $[L_{mr,t}, U_{mr,t}]$ intersects $(-\infty, 0]$. Choose a finite value $\gamma_{mr} \in [L_{mr,t}, U_{mr,t}] \cap (-\infty, 0]$ for every r. Then $\gamma_{mr} \le 0$ for every r, and therefore $m \in \mathcal{V}(\gamma)$.

Now consider any competitor j with $Q_{j,t} < Q_{m,t}$. Such a competitor cannot belong to $\widehat{\mathcal{V}}_t^{\mathrm{cert}}$, because otherwise the definition of q_t^{cert} would give $q_t^{\mathrm{cert}} \le Q_{j,t} < Q_{m,t}$, contradicting $Q_{m,t} \le q_t^{\mathrm{cert}}$. Hence $j \notin \widehat{\mathcal{V}}_t^{\mathrm{cert}}$, so there exists at least one coordinate $r(j)$ such that $U_{j,r(j),t} > 0$. The interval $[L_{j,r(j),t}, U_{j,r(j),t}]$ therefore contains a finite positive value. Choose

$$\gamma_{j,r(j)} \in [L_{j,r(j),t}, U_{j,r(j),t}] \cap (0, \infty).$$

For every remaining coordinate $r \neq r(j)$ of this competitor, choose any finite $\gamma_{jr} \in [L_{jr,t}, U_{jr,t}]$. Then $\gamma_{j,r(j)} > 0$, so $j \notin \mathcal{V}(\gamma)$.

For every competitor j with $Q_{j,t} \ge Q_{m,t}$, choose an arbitrary finite value $\gamma_{jr} \in [L_{jr,t}, U_{jr,t}]$ for every r. Because $\mathcal{C}_t$ is a Cartesian product of the coordinate intervals, all of these choices can be made independently and together define a matrix $\gamma \in \mathcal{C}_t$.

Under this matrix, m is feasible, and every method with a strictly smaller cost than $Q_{m,t}$ is infeasible. Therefore every feasible competitor j satisfies $Q_{j,t} \ge Q_{m,t}$, so

$$m \in \arg\min_{j \in \mathcal{V}(\gamma)} Q_{j,t} = \mathcal{M}^\star(\gamma).$$

Hence $m \in \widehat{\mathcal{M}}_t^{\mathrm{proj}}$. □

The proof relies on coordinatewise freedom inside a rectangle. With a nonrectangular confidence region, setting one competitor infeasible can restrict the feasible values for other methods, and the operational formula need not remain exact.

B.3 MAIN TIME-UNIFORM THEOREM

Proof of Theorem 5.2. Define the simultaneous confidence event

$$\mathcal{E} := \{\forall t \ge 1,\ \forall m, r :\ L_{mr,t} \le \Gamma_{mr,t} \le U_{mr,t}\}.$$

By Lemma B.3, $\mathbb{P}(\mathcal{E}) \ge 1 - \delta$.

Suppose that $\mathcal{E}$ holds and fix any time t. If $m \in \widehat{\mathcal{V}}_t^{\mathrm{cert}}$, then $U_{mr,t} \le 0$ for every r. Since $\Gamma_{mr,t} \le U_{mr,t}$ on $\mathcal{E}$, we have $\Gamma_{mr,t} \le 0$ for every r, and hence $m \in \mathcal{V}_t^\star$. Therefore

$$\widehat{\mathcal{V}}_t^{\mathrm{cert}} \subseteq \mathcal{V}_t^\star.$$

Conversely, if $m \in \mathcal{V}_t^\star$, then $\Gamma_{mr,t} \le 0$ for every r. Since $L_{mr,t} \le \Gamma_{mr,t}$ on $\mathcal{E}$, we have $L_{mr,t} \le 0$ for every r, so $m \in \widehat{\mathcal{V}}_t^{\mathrm{pos}}$. Thus

$$\mathcal{V}_t^\star \subseteq \widehat{\mathcal{V}}_t^{\mathrm{pos}}.$$

Now take any $m^\star \in \mathcal{M}_t^\star$. The preceding inclusion gives $m^\star \in \widehat{\mathcal{V}}_t^{\mathrm{pos}}$. Moreover, every $j \in \widehat{\mathcal{V}}_t^{\mathrm{cert}}$ is truly feasible, so the optimality of $m^\star$ over $\mathcal{V}_t^\star$ gives

$$Q_{m^\star,t} \le Q_{j,t}, \qquad \forall j \in \widehat{\mathcal{V}}_t^{\mathrm{cert}}.$$

Hence $Q_{m^\star,t} \le q_t^{\mathrm{cert}}$, where $q_t^{\mathrm{cert}} = +\infty$ if $\widehat{\mathcal{V}}_t^{\mathrm{cert}} = \varnothing$. By the definition of $\widehat{\mathcal{M}}_t^{\mathrm{proj}}$, we have $m^\star \in \widehat{\mathcal{M}}_t^{\mathrm{proj}}$. Therefore

$$\mathcal{M}_t^\star \subseteq \widehat{\mathcal{M}}_t^{\mathrm{proj}}.$$

Since t was arbitrary, all three inclusions hold for every t on $\mathcal{E}$. Because $\mathbb{P}(\mathcal{E}) \ge 1-\delta$, the result follows. □

Corollary B.4. *Let T be any stopping time with respect to $(\mathcal{G}_t)$ that is finite almost surely. Under the conditions of Theorem 5.2,*

$$\mathbb{P}\Big(\mathcal{M}_T^\star \subseteq \widehat{\mathcal{M}}_T^{\mathrm{proj}}\Big) \ge 1-\delta. \tag{35}$$

For a possibly infinite stopping time, the corresponding statement is $\mathbb{P}(T<\infty,\ \mathcal{M}_T^\star \not\subseteq \widehat{\mathcal{M}}_T^{\mathrm{proj}}) \le \delta$.

Proof of Corollary B.4. Let $\mathcal{E}$ denote the time-uniform event in Theorem 5.2. On $\mathcal{E}$, the inclusion $\mathcal{M}_t^\star \subseteq \widehat{\mathcal{M}}_t^{\mathrm{proj}}$ holds for every finite deterministic index t. Hence, on $\mathcal{E} \cap \{T<\infty\}$, it also holds at the realized index $T(\omega)$. Therefore

$$\Big\{T<\infty,\ \mathcal{M}_T^\star \not\subseteq \widehat{\mathcal{M}}_T^{\mathrm{proj}}\Big\} \subseteq \mathcal{E}^c,$$

and consequently

$$\mathbb{P}\left(T<\infty,\ \mathcal{M}_T^\star \not\subseteq \widehat{\mathcal{M}}_T^{\mathrm{proj}}\right) \le \mathbb{P}(\mathcal{E}^c) \le \delta.$$

If $T<\infty$ almost surely, this reduces to

$$\mathbb{P}\left(\mathcal{M}_T^\star \subseteq \widehat{\mathcal{M}}_T^{\mathrm{proj}}\right) \ge 1-\delta.$$

□

B.4 Contraction rate

Proposition B.5. *Suppose that $a_{rt} \in [0,1]$ and the exposure is persistent almost surely, in the sense that $A_{r,t} \asymp t$. For fixed $\rho>0$ and $\eta_{mr} \in (0,1)$, the empirical-Bernstein confidence sequence based on the gamma-exponential mixture boundary satisfies*

$$\mathrm{rad}_{mr,t} = O\left(\sqrt{\frac{\log t + \log(1/\eta_{mr})}{t}}\right) \to 0.$$

If $\eta_{mr} \asymp \delta/(MR)$, then

$$\mathrm{rad}_{mr,t} = O\left(\sqrt{\frac{\log t + \log(MR/\delta)}{t}}\right) \to 0.$$

Proof of Proposition B.5. Recall that

$$\mathrm{rad}_{mr,t} = \frac{b_{\eta_{mr},\rho}(V_{mr,t})}{A_{r,t}}, \qquad V_{mr,t} = \sum_{s=1}^{t}(X_{mrs} - \widehat{X}_{mrs})^2.$$

Since

$$X_{mrs} - \widehat{X}_{mrs} = a_{rs}(B_{mrs} - \widehat{B}_{mrs}),$$

with $B_{mrs} \in \{0,1\}$, $\widehat{B}_{mrs} \in [0,1]$, and $a_{rs} \in [0,1]$, we have

$$V_{mr,t} = \sum_{s=1}^{t} a_{rs}^2(B_{mrs} - \widehat{B}_{mrs})^2 \le \sum_{s=1}^{t} a_{rs}^2 \le \sum_{s=1}^{t} a_{rs} = A_{r,t}.$$

Because $A_{r,t} \asymp t$, there exist constants $c_A, C_A > 0$ and t_0 such that, for all $t \ge t_0$,

$$c_A t \le A_{r,t} \le C_A t.$$

Hence,

$$0 \le V_{mr,t} \le C_A t.$$

Since $V_{mr,t}$ is nondecreasing, it either remains bounded or tends to infinity. For the gamma-exponential mixture boundary, Proposition 2 of Howard et al. (2021) gives

$$b_{\eta_{mr},\rho}(v) = O\left(\sqrt{v\,[\log v + \log(1/\eta_{mr})]}\right)$$

as $v \to \infty$, for fixed $\rho > 0$ and $\eta_{mr} \in (0,1)$. Therefore, if $V_{mr,t} \to \infty$,

$$b_{\eta_{mr},\rho}(V_{mr,t}) = O\left(\sqrt{t\,[\log t + \log(1/\eta_{mr})]}\right),$$

where we have used $V_{mr,t} \le C_A t$. Since $A_{r,t} \ge c_A t$,

$$\mathrm{rad}_{mr,t} = \frac{b_{\eta_{mr},\rho}(V_{mr,t})}{A_{r,t}} = O\left(\sqrt{\frac{\log t + \log(1/\eta_{mr})}{t}}\right) \to 0.$$

If instead $V_{mr,t}$ remains bounded, then $b_{\eta_{mr},\rho}(V_{mr,t}) = O(1)$ for fixed η_{mr} and ρ, while $A_{r,t} \asymp t$. Hence,

$$\mathrm{rad}_{mr,t} = O(t^{-1}) \to 0.$$

Thus, in either case,

$$\mathrm{rad}_{mr,t} = O\left(\sqrt{\frac{\log t + \log(1/\eta_{mr})}{t}}\right) \to 0.$$

If $\eta_{mr} \asymp \delta/(MR)$, then

$$\log(1/\eta_{mr}) = \log(MR/\delta) + O(1),$$

and therefore

$$\mathrm{rad}_{mr,t} = O\left(\sqrt{\frac{\log t + \log(MR/\delta)}{t}}\right) \to 0.$$

On the joint confidence event, the confidence bounds satisfy

$$U_{mr,t} \le \Gamma_{mr,t} + 2\mathrm{rad}_{mr,t}, \qquad L_{mr,t} \ge \Gamma_{mr,t} - 2\mathrm{rad}_{mr,t}.$$

If $\Gamma_{mr,t} > 2\mathrm{rad}_{mr,t}$ for some r, then $L_{mr,t} > 0$, so $m \notin \widehat{\mathcal{V}}_t^{\mathrm{pos}}$. Alternatively, suppose that a competitor j satisfies $Q_{j,t} < Q_{m,t}$ and $\Gamma_{jr,t} + 2\mathrm{rad}_{jr,t} \le 0$ for every r. Then $U_{jr,t} \le 0$ for every r, so $j \in \widehat{\mathcal{V}}_t^{\mathrm{cert}}$ and

$$q_t^{\mathrm{cert}} \le Q_{j,t} < Q_{m,t}.$$

The projection rule therefore excludes m in either case. Since the same event guarantees $\mathcal{M}_t^\star \subseteq \widehat{\mathcal{M}}_t^{\mathrm{proj}}$, excluding every non-oracle method gives exact recovery.

Finally, $M, R < \infty$ and persistent exposure imply $\max_{m,r} \mathrm{rad}_{mr,t} \to 0$. Under the uniform margins stated in Section 5.2, every non-oracle method therefore satisfies an exclusion condition for all sufficiently large t. □

B.5 Certification and oracle identification

Proof of Proposition 5.3. Let $\mathcal{F}_n := \sigma(B_1, \dots, B_n)$ and $E_n := \{T_{\mathrm{cert}} \le n\}$. Since $E_n \in \mathcal{F}_n$, its probability under the Bernoulli product law is a finite sum of terms of the form $p^k(1-p)^{n-k}$, and is therefore continuous in p. The assumption gives $\mathbb{P}_p(E_n) \le \eta$ for every $p > \tau$. Taking $p \downarrow \tau$ yields $\mathbb{P}_\tau(E_n) \le \eta$. Finally, since $E_n \uparrow \{T_{\mathrm{cert}} < \infty\}$,

$$\mathbb{P}_\tau(T_{\mathrm{cert}} < \infty) = \lim_{n\to\infty} \mathbb{P}_\tau(E_n) \le \eta.$$

□

Implication for oracle identification. In the two-candidate construction of Section 5.3, candidate 2 is the unique oracle whenever $p > \tau$. Let $(\widehat{\mathcal{S}}_t)_{t\geq 1}$ be an $(\mathcal{F}_t)$-adapted rule satisfying (5) for every $p \in (0,1)$, and define

$$T_2 := \inf\{t \geq 1 : 2 \notin \widehat{\mathcal{S}}_t\}.$$

For $p > \tau$, excluding candidate 2 violates oracle coverage, so $\sup_{p>\tau} \mathbb{P}_p(T_2 < \infty) \leq \delta$. Proposition 5.3 therefore gives $\mathbb{P}_\tau(T_2 < \infty) \leq \delta$. Since reporting $\{1\}$ implies excluding candidate 2,

$$\mathbb{P}_\tau\Big(\exists t \geq 1 : \widehat{\mathcal{S}}_t = \{1\}\Big) \leq \mathbb{P}_\tau(T_2 < \infty) \leq \delta.$$

Finite-sample lower bound near the boundary. Now let $p = \tau - \Delta$ with $\Delta \in (0, \tau)$, so that candidate 1 is strictly feasible and uniquely optimal. The difficulty of distinguishing this risk from the boundary risk is measured by the Bernoulli relative entropy

$$\mathrm{kl}(a,b) := a \log \frac{a}{b} + (1-a)\log \frac{1-a}{1-b}, \qquad a, b \in (0,1).$$

For any $\beta \in (0, 1-\delta)$ and integer $n \geq 1$,

$$\mathbb{P}_{\tau-\Delta}(T_2 \leq n) \geq 1-\beta \quad \Longrightarrow \quad n \geq \frac{\mathrm{kl}(1-\beta, \delta)}{\mathrm{kl}(\tau-\Delta, \tau)}.$$

Indeed, let $E_n := \{T_2 \leq n\}$. The boundary result gives $\mathbb{P}_\tau(E_n) \leq \delta$. Applying the log-sum inequality to E_n and its complement yields

$$n\,\mathrm{kl}(\tau-\Delta, \tau) \geq \mathrm{kl}(\mathbb{P}_{\tau-\Delta}(E_n), \mathbb{P}_\tau(E_n)) \geq \mathrm{kl}(1-\beta, \delta),$$

where the last inequality uses $\mathbb{P}_{\tau-\Delta}(E_n) \geq 1-\beta > \delta \geq \mathbb{P}_\tau(E_n)$. This proves the bound.

For fixed $\tau \in (0,1)$,

$$\mathrm{kl}(\tau-\Delta, \tau) = \frac{\Delta^2}{2\tau(1-\tau)} + o(\Delta^2) \qquad \text{as } \Delta \downarrow 0.$$

Hence the lower bound is of order Δ^{-2} for fixed δ and β. Reporting the correct singleton $\{1\}$ by time n implies $T_2 \leq n$, so the same lower bound applies to high-probability singleton identification.

C The Weak Sequential Model Confidence Set

The Weak Sequential Model Confidence Set (Weak-SMCS) of Arnold et al. (2026) is a score-based procedure for sequentially identifying weakly superior forecasting methods. Let $\ell_{m,t}$ denote the loss of method m at time t, and define the pairwise loss difference between methods m and j as

$$d_{mj,t} = \ell_{m,t} - \ell_{j,t}.$$

The average conditional expected loss difference up to time t is

$$\Delta_{mj,t} = \frac{1}{t}\sum_{s=1}^{t} \mathbb{E}\left[d_{mj,s} \mid \mathcal{G}_{s-1}\right].$$

The set of weakly superior methods at time t is defined as

$$\mathcal{M}_t^{w,\star} = \{m : \Delta_{mj,t} \leq 0,\ \forall j \neq m\}.$$

Unlike uniformly weak superiority, this target is allowed to vary over time, so a method excluded at an earlier stage may become weakly superior again later.

In our implementation, the loss is the Winkler score. For a prediction interval $[\underline{Y}_{m,t}, \overline{Y}_{m,t}]$ with nominal miscoverage level α, the Winkler score is

$$\ell_{m,t} = (\overline{Y}_{m,t} - \underline{Y}_{m,t}) + \frac{2}{\alpha}(\underline{Y}_{m,t} - Y_t)\mathbf{1}\{Y_t < \underline{Y}_{m,t}\} + \frac{2}{\alpha}(Y_t - \overline{Y}_{m,t})\mathbf{1}\{Y_t > \overline{Y}_{m,t}\}.$$

Although the Winkler score itself is unbounded as a function of Y_t, the difference between the scores of two fixed prediction intervals is conditionally bounded. Following Section 3.4 of Arnold et al. (2026), for each ordered pair (m, j) and time t, we compute the predictable bound

$$b_{mj,t} = \sup_{y \in \mathbb{R}} |WS_{m,t}(y) - WS_{j,t}(y)| .$$

Since $b_{mj,t}$ depends only on the prediction intervals issued before observing Y_t, it is predictable. We then standardize the pairwise loss difference as

$$\widetilde{d}_{mj,t} = \frac{d_{mj,t}}{b_{mj,t}},$$

with $\widetilde{d}_{mj,t} = 0$ when $b_{mj,t} = 0$. By construction,

$$|\widetilde{d}_{mj,t}| \leq 1.$$

This predictable normalization converts the conditionally bounded loss differences into uniformly bounded ones. Accordingly, the Weak-SMCS used in our experiments targets weak superiority with respect to the scaled loss differences. Define

$$\widetilde{\Delta}_{mj,t} = \frac{1}{t} \sum_{s=1}^{t} \mathbb{E}\left[\widetilde{d}_{mj,s} \mid \mathcal{G}_{s-1}\right] = \frac{1}{t} \sum_{s=1}^{t} \frac{\mathbb{E}[d_{mj,s} \mid \mathcal{G}_{s-1}]}{b_{mj,s}}.$$

The corresponding weakly superior set is

$$\widetilde{\mathcal{M}}_t^{w,\star} = \left\{ m : \widetilde{\Delta}_{mj,t} \leq 0,\ \forall j \neq m \right\}.$$

For the standardized differences, we use the bounded-difference construction of Proposition 3.6 in Arnold et al. (2026). Since $|\widetilde{d}_{mj,t}| \leq 1$, we set $c = 2$ and $\lambda = 1/4$. The predictable centering sequence is given by the running mean of previous standardized differences,

$$\gamma_{mj,1} = 0, \qquad \gamma_{mj,t} = \frac{1}{t-1} \sum_{s=1}^{t-1} \widetilde{d}_{mj,s}, \quad t \geq 2,$$

with $\gamma_{mj,t}$ restricted to $[-1, 1]$. Define

$$V_{mj,t} = \sum_{s=1}^{t} \left(\widetilde{d}_{mj,s} - \gamma_{mj,s}\right)^2,$$

and

$$\psi_{E,c}(\lambda) = \frac{-\log(1 - c\lambda) - c\lambda}{c^2}.$$

For each ordered pair (m, j) and candidate value $x \in \mathbb{R}$, define

$$M_{mj,t}(x) = \exp\left\{ \lambda \sum_{s=1}^{t} \widetilde{d}_{mj,s} - \lambda t x - \psi_{E,c}(\lambda) V_{mj,t} \right\}.$$

In particular, evaluating the process at the boundary value $x = 0$ gives

$$M_{mj,t}(0) = \exp\left\{ \lambda \sum_{s=1}^{t} \widetilde{d}_{mj,s} - \psi_{E,c}(\lambda) V_{mj,t} \right\}.$$

Following the joint confidence-region construction of Arnold et al. (2026), let $X = (x_{mj})_{m \neq j}$ denote a candidate matrix of pairwise conditional-average loss differences and define

$$M_t(X) = \frac{1}{M(M-1)} \sum_{m \neq j} M_{mj,t}(x_{mj}), \qquad \mathcal{C}_t = \left\{ X : M_t(X) \leq \frac{1}{\delta} \right\}.$$

The set $\mathcal{C}_t$ is a joint confidence region for the matrix of scaled pairwise conditional-average loss differences. The dynamic Weak-SMCS is obtained by checking, for each method m and every

competitor $j \neq m$, whether the joint confidence region still contains a candidate matrix under which method m is not worse than method j:

$$\widehat{\mathcal{M}}_t^w = \{m : \mathcal{C}_t \cap \{X : x_{mj} \leq 0\} \neq \varnothing,\ \forall j \neq m\}.$$

Since $M_{mj,t}(x)$ is nonnegative, convex, and decreasing in x, the monotonicity argument of Arnold et al. (2026) reduces the intersection condition to the boundary value $x_{mj} = 0$. For fixed (m, j), the remaining coordinates of X can be taken arbitrarily large, making their contributions to $M_t(X)$ arbitrarily close to zero. Hence,

$$\inf_{X:\, x_{mj} \leq 0} M_t(X) = \frac{M_{mj,t}(0)}{M(M-1)}.$$

$$\mathcal{C}_t \cap \{X : x_{mj} \leq 0\} \neq \varnothing \quad \Longleftrightarrow \quad M_{mj,t}(0) < \frac{M(M-1)}{\delta}.$$

Equivalently, method m is excluded at time t if there exists a competitor $j \neq m$ such that

$$M_{mj,t}(0) \geq \frac{M(M-1)}{\delta}.$$

Thus, the estimated dynamic Weak-SMCS can be written as

$$\widehat{\mathcal{M}}_t^w = \left\{ m : M_{mj,t}(0) < \frac{M(M-1)}{\delta},\ \forall j \neq m \right\}.$$

The resulting sequence satisfies the time-uniform guarantee

$$\Pr\left(\widetilde{\mathcal{M}}_t^{w,\star} \subseteq \widehat{\mathcal{M}}_t^w,\ \forall t\right) \geq 1 - \delta.$$

Because the Weak-SMCS is defined pointwise at each time t, previously excluded methods are allowed to re-enter when the accumulated evidence changes. We use this dynamic version rather than its running-intersection counterpart.

We include Weak-SMCS as a score-based benchmark in simulations. Its inferential target differs from that of CC-SMCS: Weak-SMCS compares methods through predictably standardized Winkler-score differences, whereas CC-SMCS treats coverage as an explicit constraint and interval width as the efficiency criterion. Therefore, the two procedures target different superior sets and may exhibit different model-set paths even when applied to the same sequence of prediction intervals.

D THE SETTING OF ρ

The parameter ρ determines the intrinsic-time region over which the mixture boundary in equation (21) is relatively tight. Following Howard et al. (2021), we also use the tuning rule derived for the normal mixture boundary to guide the choice of ρ for the gamma-exponential mixture boundary.

In our construction, the total two-sided crossing probability assigned to a confidence sequence is η, so each one-sided boundary is assigned crossing probability $\eta/2$. Following the normal-mixture tuning rule of Howard et al. (2021), for a prespecified target intrinsic time $w > 0$, we set

$$\rho(w) = \frac{w}{-W_{-1}\left(-\dfrac{\eta^2}{el_0^2}\right) - 1}.$$

Here, $w > 0$ denotes the target intrinsic time, $\rho > 0$ is the mixture tuning parameter, $\eta \in (0, 1)$ is the corresponding two-sided crossing probability, $l_0 \geq 1$ is the initial supermartingale constant, and $W_{-1}(\cdot)$ denotes the lower branch of the Lambert W function satisfying $W(x)e^{W(x)} = x$. In our scalar setting, we take $l_0 = 1$.

To determine an appropriate target intrinsic time w, recall that

$$A_{r,t} = \sum_{s=1}^{t} a_{rs}$$

denotes the cumulative exposure. For method m and constraint r, the empirical-Bernstein intrinsic time is

$$V_{mr,t} = \sum_{s=1}^{t} \left(X_{mrs} - \widehat{X}_{mrs} \right)^2,$$

and the corresponding one-sided confidence radius is

$$R_{mr,t} = \frac{b_{\eta,\rho}(V_{mr,t})}{A_{r,t}}.$$

Because $V_{mr,t}$ depends on the realized observations, it is unknown before the forecasting period begins and therefore cannot be used directly to prespecify ρ.

To address this issue, we follow Howard et al. (2021), who use a Hoeffding variance process to provide a deterministic intrinsic-time scale for bounded observations, and adopt the same idea in our weighted setting. Since $B_{mrs} \in \{0, 1\}$,

$$B_{mrs} - \tau_r \in [-\tau_r,\, 1 - \tau_r].$$

Hence, the weighted increment $a_{rs}(B_{mrs} - \tau_r)$ lies in

$$a_{rs}(B_{mrs} - \tau_r) \in [-a_{rs}\tau_r,\, a_{rs}(1 - \tau_r)],$$

whose range length is

$$a_{rs}(1 - \tau_r) - (-a_{rs}\tau_r) = a_{rs}.$$

By Hoeffding's lemma, the corresponding centered increment is conditionally sub-Gaussian with variance upper bound

$$\frac{a_{rs}^2}{4}.$$

Following the Hoeffding variance construction of Howard et al. (2021), we therefore use

$$V_{r,t}^{H} = \frac{1}{4} \sum_{s=1}^{t} a_{rs}^2$$

as the reference intrinsic time for tuning ρ. Under the unit-exposure setting used in our experiments, this reference path is deterministic and can be computed before the forecasting period begins.

Let T denote the forecasting horizon. We prespecify three target forecasting times corresponding to the 25%, 50%, and 75% points of the forecasting horizon,

$$t_q^\star = qT, \qquad q \in \{0.25, 0.50, 0.75\}.$$

Under the unit-exposure setting adopted in this paper,

$$a_{rs} = 1, \quad A_{r,t} = t, \quad V_{r,t}^{\mathrm{H}} = \frac{t}{4}.$$

The corresponding target intrinsic times are therefore

$$w_{0.25} = \frac{T}{16}, \qquad w_{0.50} = \frac{T}{8}, \qquad w_{0.75} = \frac{3T}{16}.$$

These target intrinsic times generate three candidate tuning parameters,

$$\rho_q = \rho(w_q), \qquad q \in \{0.25, 0.50, 0.75\}.$$

To select among these candidates, we evaluate the corresponding normal-mixture reference radius along the deterministic Hoeffding path

$$V_t^{\mathrm{H}} = \frac{t}{4}.$$

Let $u_{\eta,\rho}(v)$ denote the two-sided normal mixture boundary used for tuning. For each candidate ρ_q, we compute

$$u_{\eta,\rho}\big(V_{r,t}^{\mathrm{H}}\big) = \sqrt{\big(V_{r,t}^{\mathrm{H}} + \rho\big) \log\left(\frac{V_{r,t}^{\mathrm{H}} + \rho}{\eta^2 \rho} \right)}, \quad \overline{R}^{\mathrm{ref}}(\rho_q) = \frac{1}{T} \sum_{t=1}^{T} \frac{u_{\eta,\rho_q}(t/4)}{t},$$

and select

$$\rho^\star = \arg \min_{\rho_q:\, q\in\{0.25,0.50,0.75\}} \overline{R}^{\mathrm{ref}}(\rho_q).$$

The selected $\rho^\star$ is fixed before the forecasting outcomes are observed.

The Hoeffding reference intrinsic time is used only for the prespecified choice of ρ. Once $\rho^\star$ has been selected, the actual confidence sequence is constructed using the realized empirical-Bernstein intrinsic time $V_{mr,t}$. Under the unit-exposure setting, the resulting one-sided confidence radius is

$$R_{mr,t} = \frac{b_{\eta,\rho^\star}(V_{mr,t})}{t}.$$

Thus, for a fixed error allocation η, the prespecified choice of $\rho^\star$ depends only on the forecasting horizon T. Since the three simulation settings share the same forecasting horizon, they have the same candidate values of ρ and the same selected $\rho^\star$. We therefore report a single $\rho^\star$ for the three simulations. Table 3 summarizes the selected values of $\rho^\star$ for the simulation setting and the four empirical datasets.

Table 3: Selected values of ρ

The selection of ρ	δ	Simulation Data Set	Data Set 1	Data Set 2	Data Set 3	Data Set 4
$\rho^\star$	0.10	17.616	20.494	29.584	30.946	8.151
	0.25	21.752	25.305	36.529	38.211	10.064

Note: Data Sets 1, 2, 3, 4 respectively represent the data of French electricity price, load, weather, and wind

E More details of Simulations

E.1 The computation of coverage and width for each method

In simulations, we design six methods. Specifically, the coverage rates for M2-M6 are computed by:

$$\begin{aligned}
P\left\{Y_t \in C_{mr,t} \mid \mathcal{G}_{t-1}\right\} &= P\left\{\widehat{Y}_t - \sigma_{mr,t}z \le Y_t \le \widehat{Y}_t + \sigma_{mr,t}z \mid \mathcal{G}_{t-1}\right\} \\
&= P\left\{-\sigma_{mr,t}z \le Y_t - \widehat{Y}_t \le \sigma_{mr,t}z \mid \mathcal{G}_{t-1}\right\} \\
&= P\left\{-\frac{\sigma_{mr,t}z}{\sigma_t} \le \frac{Y_t - \widehat{Y}_t}{\sigma_t} \le \frac{\sigma_{mr,t}z}{\sigma_t} \,\middle|\, \mathcal{G}_{t-1}\right\} \\
&= \Phi(\frac{\sigma_{mr,t}z}{\sigma_t}) - \Phi(-\frac{\sigma_{mr,t}z}{\sigma_t}) \\
&= 2\Phi(\frac{\sigma_{mr,t}z}{\sigma_t}) - 1 \\
&= 2\Phi(\exp\{(\theta_{mr} - l_t)|h_t|\}z) - 1
\end{aligned}$$

The last equality follows from $\sigma_{mr,t} = \sigma_t^0 \exp\{\theta_{mr}|h_t|\}$. Hence, the conditional miscoverage probability is:

$$\Pr\left\{Y_t \notin C_{mr,t} \mid \mathcal{G}_{t-1}\right\} = 2\left[1 - \Phi\left(\exp\{(\theta_{mr} - l_t)|h_t|\}z\right)\right].$$

This conditional miscoverage probability can be related to our framework as follows:

$$\Gamma_{mr,t} = \begin{cases} A_{r,t}^{-1} \sum_{s=1}^{t} \mu_{mrs}, & A_{r,t} > 0, \\ 0, & A_{r,t} = 0. \end{cases}$$

where $\mu_{mrt} = \mathbb{E}[X_{mrt} \mid \mathcal{G}_{t-1}] = a_{rt}\left[\mathbb{E}[B_{mrt} \mid \mathcal{G}_{t-1}] - \tau_r\right]$ and $B_{mrt} = \mathbf{1}\{Y_t \notin C_{mr,t}\}$. Hence, when $A_{r,t} > 0$,

$$\begin{aligned}\Gamma_{mr,t} &= A_{r,t}^{-1} \sum_{s=1}^{t} a_{rs}\left[\mathbb{E}[B_{mrs} \mid \mathcal{G}_{s-1}] - \tau_r\right] \\ &= A_{r,t}^{-1} \sum_{s=1}^{t} a_{rs}\left[\Pr\left\{Y_s \notin C_{mr,s} \mid \mathcal{G}_{s-1}\right\} - \tau_r\right] \\ &= A_{r,t}^{-1} \sum_{s=1}^{t} a_{rs}\left[2\left\{1 - \Phi\left(\exp\{(\theta_{mr} - l_s)|h_s|\}z\right)\right\} - \tau_r\right].\end{aligned}$$

Similarly, for M1, whose prediction interval is $C_{1r,t} = [\widehat{Y}_t + \kappa\sigma_{4r,t} - z\sigma_{4r,t},\ \widehat{Y}_t + \kappa\sigma_{4r,t} + z\sigma_{4r,t}]$, the conditional miscoverage probability is

$$\Pr\{Y_t \notin C_{1r,t} \mid \mathcal{G}_{t-1}\} = 1 - \left[\Phi\left(\frac{(\kappa + z)\sigma_{4r,t}}{\sigma_t}\right) - \Phi\left(\frac{(\kappa - z)\sigma_{4r,t}}{\sigma_t}\right)\right].$$

Therefore, $\Gamma_{1r,t}$ is given by:

$$\Gamma_{1r,t} = A_{r,t}^{-1} \sum_{s=1}^{t} a_{rs}\left[1 - \left[\Phi\left(\exp\{(\theta_{4r} - l_s)|h_s|\}(\kappa + z)\right)\right] + \left[\Phi\left(\exp\{(\theta_{4r} - l_s)|h_s|\}(\kappa - z)\right)\right] - \tau_r\right]$$

In the stationary case, the relationship between θ_{mr} and l directly determines the feasibility of M2–M6. With constant $l_t = l$ and $\tau = \alpha$, if $\theta_{mr} < l$, then $\sigma_{mr,t} < \sigma_t$, the conditional miscoverage probability exceeds α, and hence $\Gamma_{mr,t} > 0$; if $\theta_{mr} = l$, then $\Gamma_{mr,t} = 0$; and if $\theta_{mr} > l$, then $\sigma_{mr,t} > \sigma_t$, the conditional miscoverage probability is below α, and hence $\Gamma_{mr,t} < 0$. In our construction, M4 corresponds to the case $\theta_{4r} = l$ and therefore attains the nominal coverage level in the stationary setting. M1 is constructed using M4 as a reference: it has the same interval width as M4, but its center is shifted away from $\widehat{Y}_t$. This shift reduces the coverage probability without changing the interval width. The parameter κ controls the magnitude of the shift: $\kappa = 0$ recovers M4, while any nonzero κ lowers the coverage relative to M4. We set $\kappa = 0.5$, which yields $\Gamma_{1r,t} > 0$ and therefore makes M1 infeasible under the coverage constraint while preserving the same width as M4.

In the dynamic settings, as in Simulations 2 and 3, l_t changes over time. For M2–M6, the relation between θ_{mr} and l_t determines only the instantaneous contribution to the cumulative excess miscoverage, while the sign of $\Gamma_{mr,t}$ also depends on the coverage history accumulated at previous time points. M1 is affected differently because its interval is shifted relative to the predictive center. Its instantaneous miscoverage therefore depends jointly on the scale ratio between M4 and the true forecast-error scale and on the fixed center-shift parameter κ. Consequently, the feasibility of M1 in the dynamic settings is determined from its own accumulated excess miscoverage $\Gamma_{1r,t}$ rather than from the relation between θ_{4r} and l_t alone. According to Definition 3.1, we determine the oracle from the cumulative coverage risk together with the cumulative interval width. For M2–M6, the interval width $2z\sigma_{mr,t}$ increases monotonically with $\sigma_{mr,t}$, while M1 has the same width as M4 by construction. The constrained oracle is therefore the method with the smallest cumulative interval width among those satisfying the cumulative coverage constraint.

E.2 Parameter Settings

We specify the candidate interval scale as $\sigma_{mr,t} = \sigma_t^0 \exp\{\theta_{mr}|h_t|\}$, where $\sigma_t^0 = \exp\{H_{t-1}\}$ and $\phi = 0.5$. Given $\sigma_{mr,t}$, the corresponding coverage probability and interval width can be computed directly. Since h_t is fixed by the data-generating process, it remains only to specify θ_{mr}. For convenience, we introduce a common scaling parameter ξ and, for $r = 1$, set

$$\theta_{mr} = \xi(-2, -1, 0, 1, 2), \qquad m = 2, 3, 4, 5, 6$$

corresponding to M2–M6, respectively. We set $\xi = 0.15$. For M1, we construct the interval as $[\widehat{Y}_t + \kappa\sigma_{4r,t} - z\sigma_{4r,t}, \widehat{Y}_t + \kappa\sigma_{4r,t} + z\sigma_{4r,t}]$, where κ controls the magnitude of the center shift. We set $\kappa = 0.5$. Thus, M1 has the same interval width as M4 but a deliberately shifted center, which induces undercoverage. For all methods, we set $z = \Phi^{-1}(1 - \alpha/2)$ with $\alpha = 0.05$.

For the CC-SMCS parameters used in the simulations, we set $R = 1$, $a_{rt} = 1$, and $\epsilon_r = 0$, so that $\tau_r = \alpha_r = 0.05$. Since there are six candidate methods and one monitored constraint, we use a balanced error allocation with $\pi_{mr} = 1/6$, giving $\eta_{mr} = \delta/6$, where $\delta \in \{0.10, 0.25\}$. We set $w_t = 1$, so that the efficiency objective is the cumulative mean interval width. The mixture tuning parameter $\rho^\star$ is fixed before evaluation according to the procedure described in Appendix D. The selected values are $\rho^\star = 17.616$ for $\delta = 0.10$ and $\rho^\star = 21.752$ for $\delta = 0.25$, as reported in Table 3.

E.3 Connection Between the Parameters and the Three Simulations

Across the three simulations, the candidate-specific parameters θ_{mr} remain unchanged. Instead, we modify the true forecast-error scale to generate stationary, structural-break, and gradual-drift environments. We set the master random seed to 42 and generate independent replication-specific seeds using NumPy's SeedSequence. The same replication-specific seeds are reused across the three simulation designs to implement paired common random numbers.

Simulation 1 The stationary setting assumes that the true scale σ_t follows its baseline process, $\sigma_t^0 = \exp\{H_{t-1}\}$, throughout the forecasting horizon. Under this setting, M4 is the constrained oracle at all time points because $\theta_{4r} = 0$. Since θ_{mr} determines the conditional miscoverage probability through $\sigma_{mr,t}$ and $R = 1$, we have

$$\Gamma_{1r,t} > 0, \qquad \Gamma_{2r,t} > 0, \qquad \Gamma_{3r,t} > 0, \qquad \Gamma_{4r,t} = 0, \qquad \Gamma_{5r,t} < 0, \qquad \Gamma_{6r,t} < 0.$$

Hence, M1–M3 violate the coverage constraint, M4 lies exactly on the coverage boundary, and M5–M6 are conservative. Under the width specification used in the selection step,

$$Q_{2r,t} < Q_{3r,t} < Q_{1r,t} = Q_{4r,t} < Q_{5r,t} < Q_{6r,t}, \qquad \forall t \in T.$$

Therefore, M4 is the narrowest method satisfying the population coverage constraint and is the unique constrained oracle throughout Simulation 1.

Simulation 2 This simulation introduces a permanent structural break in the true forecast-error scale while keeping all candidate procedures unchanged. Specifically, after $t = 400$, the true scale changes from σ_t^0 to $\sigma_t^0 \exp\{1.75\xi|h_t|\}$. The candidate interval constructions remain unchanged, whereas their conditional coverage probabilities shift after the break. The forecast-error scales then satisfy

$$\sigma_{4r,t} = \sigma_t^0 < \sigma_{5r,t} = \sigma_t^0 \exp\{\xi|h_t|\} < \sigma_t = \sigma_t^0 \exp\{1.75\xi|h_t|\} < \sigma_{6r,t} = \sigma_t^0 \exp\{2\xi|h_t|\}.$$

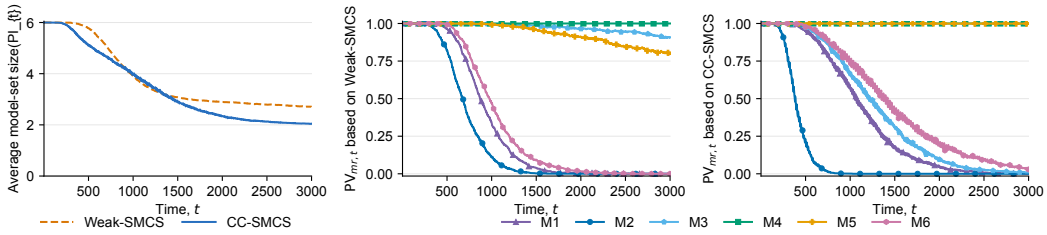
M4 lies on the coverage boundary before the break and loses feasibility at $t = 401$. After the break, the oracle is M5 when $\Gamma_{5r,t} \le 0$ and M6 otherwise. The transition therefore depends on the coverage margin accumulated by M5 before the break.

Simulation 3 Instead of the abrupt break, this simulation sets a gradual linear drift in the true forecast-error scale. Specifically, the coefficient governing the scale distortion increases linearly from 0 to 2ξ over the evaluation period, so that the true scale is given by $\sigma_t = \sigma_t^0 \exp 2\xi \frac{t-1}{T-1}|h_t|$. As the scale changes continuously over time, the conditional coverage probabilities of the candidate methods also evolve gradually. Since M4 is initially located on the coverage boundary, it loses feasibility shortly after the drift begins, and the oracle therefore moves quickly from M4 to M5. In theory, $\sigma_{5r,t}$ remains larger than σ_t until approximately $t = 1500$, while $\sigma_{6r,t}$ remains larger than σ_t until $t = 3000$. This setting therefore provides a more challenging environment for evaluating whether CC-SMCS can maintain validity and adapt to a smoothly evolving oracle.

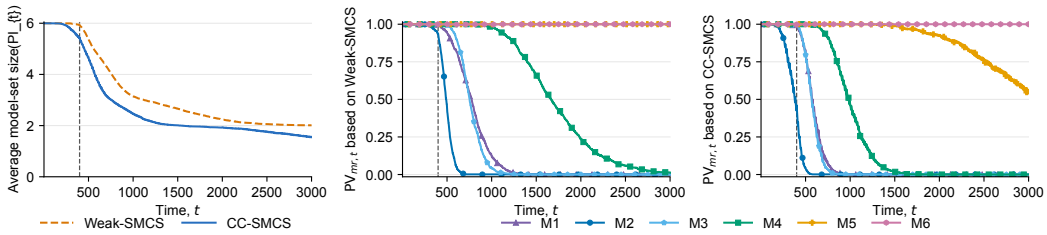
E.4 Additional Results

Figure 3 presents the simulation results for $\delta = 0.25$. In Simulation 1, where M4 is the fixed oracle, CC-SMCS retains M4 together with M5 in the final model set, consistent with the oracle-retention guarantee in Theorem 5.2. In Simulation 2, the constrained oracle changes from M4 to M5 and subsequently to M6. Figure 3b shows that CC-SMCS adapts to these oracle changes while continuously retaining the current oracle in the model set. Simulation 3 exhibits a pattern similar to that obtained for $\delta = 0.10$ with a linear drift. M4 becomes undercovered from $t \ge 2$, after

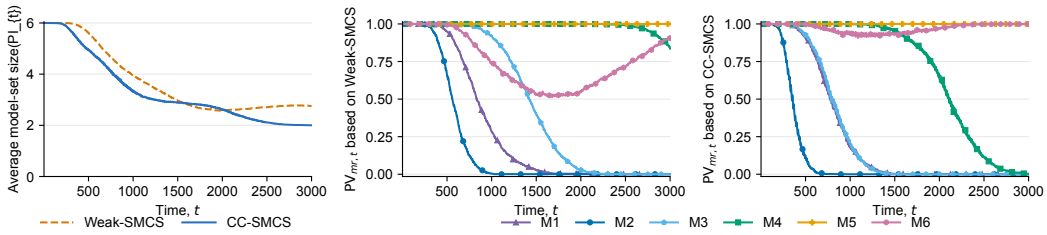
which M5 becomes the constrained oracle. Although $\sigma_{5r,t}$ remains larger than the true scale σ_t until approximately $t = 1500$, M5 does not immediately cease to be feasible once $\sigma_{5r,t} < \sigma_t$, because feasibility is determined by cumulative rather than instantaneous miscoverage. M5 therefore remains the oracle until its accumulated coverage margin is exhausted and $\Gamma_{5r,t}$ becomes positive, after which M6 becomes the constrained oracle. In Figure 3c, M5 remains in the CC-SMCS throughout the evaluation period. The inclusion probability of M6 first decreases and then increases, and M4 is gradually excluded as the drift strengthens. Weak-SMCS, which targets the Winkler score rather than coverage-constrained interval width, exhibits different model-set dynamics. In Simulation 1, it retains the oracle throughout the evaluation period, whereas in Simulations 2 and 3 it tends to retain methods with undercoverage for longer after the oracle changes.



(a) Simulation 1



(b) Simulation 2



(c) Simulation 3

Figure 3: The performance of simulations with $\delta = 0.25$. The vertical dashed line in Figure 3b indicates the structural break.

F Additional Details of the Empirical Study

F.1 Datasets

French electricity price. The first dataset is the French electricity price series used by Zaffran et al. (2022), covering the period from January 8, 2016 to December 31, 2019. It contains 34,896 hourly observations. We focus on the variable `Spot` and use only its historical values to construct one-step-ahead forecasts. It can be directly downloaded from `https://github.com/mzaffran/AdaptiveConformalPredictionsTimeSeries`.

Electricity load. The second dataset is the electricity load series of Hong et al. (2016) from January 1, 2005, to September 30, 2010, containing 50,376 consecutive

hourly observations, which can be found at http://blog.drhongtao.com/2017/03/gefcom2014-load-forecasting-data.html. This dataset was also used by Lin et al. (2022) for conformal prediction. Unlike their setting, we consider a univariate forecasting task and use only the historical values of LOAD to train the forecasting models and predict future load.

Weather. The third dataset is the Weather dataset of Wu et al. (2023), and also used by Chen et al. (2024). It can be found at https://huggingface.co/datasets/thuml/Time-Series-Library/blob/main/weather/weather.csv, covering the year 2020 with observations recorded at 10-minute intervals. We use T (degC) as the target variable, which contains no missing values among the originally observed records. We find one duplicated timestamp and nine missing timestamps in the original time index. Since only nine timestamps are missing over the entire sample period, we remove the duplicated record, regularize the series to a complete 10-minute grid, and fill each missing value using the temperature observed at the same time on the previous day. This preprocessing preserves the regular sampling frequency while using only past information. The resulting series contains 52,704 regularly spaced observations. We then use only the historical values of T (degC) for model training and one-step-ahead prediction.

Wind power generation. The final dataset is the Hackberry Wind Farm dataset used by Xu & Xie (2021), which can be downloaded from https://github.com/hamrel-cxu/EnbPI/blob/main/Data/Wind_Hackberry_Generation_2019_2020.csv. While their original experiment focuses on the 2019 sample, we use the extended dataset from January 1, 2019, to July 31, 2020, containing 13,871 hourly observations. The original records preserve local daylight-saving-time transitions, which explains why the number of recorded observations differs from the number obtained by simply multiplying the number of calendar days by 24. We focus on the univariate variable MWH and use only its historical values for model training and one-step-ahead forecasting.

F.2 Point Forecasting

Forecasting framework. To construct conformal prediction intervals, we first train point forecasting models to obtain one-step-ahead forecasts. We consider three representative time-series forecasting backbones: DLinear (Zeng et al., 2023), PatchTST (Nie et al., 2023), and TimesNet (Wu et al., 2023). For the common point-forecasting framework, observations in each dataset are divided chronologically according to a $12:4:2:2$ ratio, corresponding approximately to 60% for training, 20% for validation, 10% for conformal calibration, and the final 10% for out-of-sample evaluation. This design is adapted from the $12:4:4$ training–validation–test ratio commonly used in long-term forecasting benchmarks, with the original test portion further divided equally into calibration and evaluation samples. The method-specific adaptations for CQR and EnbPI are described below.

All three models perform one-step-ahead univariate forecasting using only historical observations of the corresponding target variable. For a target observation at time t, the input window contains observations strictly before t. For the common point-forecasting framework, each series is standardized using the mean and standard deviation estimated from the training sample only. The same transformation is subsequently applied to the validation, calibration, and test samples, and all predictions are transformed back to the original scale before interval construction and evaluation.

Model configurations. For the model-specific hyperparameters, we adopt the benchmark configurations reported in the original papers and their official implementations of DLinear (Zeng et al., 2023), PatchTST (Nie et al., 2023), and TimesNet (Wu et al., 2023), while changing the forecasting horizon to one to match our sequential prediction setting. DLinear uses an input sequence length of 336 and a moving-average kernel size of 25, with a batch size of 32, an initial learning rate of 5×10^{-3}, and zero weight decay. PatchTST utilizes an input sequence length of 336, a patch length of 16, a stride of 8, three encoder layers, four attention heads, $d_{\text{model}} = 16$, $d_{\text{ff}} = 128$, dropout of 0.3, fully connected dropout of 0.3, and RevIN normalization with the affine transformation disabled. Its batch size is 128, and its initial learning rate is 10^{-4}. TimesNet adopts an input sequence length of 96, two TimesBlocks, $d_{\text{model}} = 16$, $d_{\text{ff}} = 32$, the top five detected periods, six inception kernels, and dropout of 0.1, with a batch size of 32 and an initial learning rate of 10^{-4}.

Training and model selection. All point forecasting models are optimized using the Adam optimizer with the mean squared error (MSE) loss and random seed 2021. Model selection is based

exclusively on the validation sample, and the model parameters corresponding to the lowest validation MSE are retained for subsequent prediction. DLinear and TimesNet are trained for at most 10 epochs with an early-stopping patience of three epochs, while PatchTST is trained for at most 100 epochs with a patience of 100 epochs. We use type-1 learning-rate adjustment for DLinear and TimesNet and type-3 adjustment for PatchTST. In each case, the learning rate is updated at the end of each training epoch and the adjusted value is used in the following epoch. The same model-specific training configuration is applied across all four empirical datasets.

Adaptations for CQR and EnbPI. Two interval construction methods require modifications to the common point-forecasting framework. For CQR (Romano et al., 2019), as detailed in Subsection F.3, DLinear, PatchTST, and TimesNet are retained as forecasting backbones, while their scalar output layers are modified to produce two quantile outputs. For each backbone, a separate CQR model is trained for the 95% nominal coverage level, using the sum of the lower- and upper-quantile pinball losses instead of MSE. EnbPI (Xu & Xie, 2021), also detailed in Subsection F.3, trains each bootstrap model independently, with early stopping determined by the training MSE on its own bootstrap sample rather than by validation loss.

F.3 Conformal Prediction Methods

We consider six conformal prediction methods with nominal coverage levels of 95%. They differ in how the conformity scores are constructed, how the calibration information is weighted or updated, and whether the prediction interval is centered at a point forecast or obtained from conditional quantile estimates. Let n_{cal} denote the number of observations in the designated calibration sample.

Split Conformal Prediction. The first method is classical split conformal prediction (Split CP) (Lei et al., 2018). Given a fitted point forecasting model with prediction $\widehat{Y}_i$, we use the absolute prediction error as the conformity score,

$$S_i = \left| Y_i - \widehat{Y}_i \right|.$$

For a calibration sample of size n_{cal}, let $k = \lceil (n_{\text{cal}} + 1)(1 - \alpha) \rceil$, and let $\widehat{q}_{1-\alpha}$ be the kth order statistic of the calibration scores. The prediction interval at time t is

$$\widehat{C}_t^{\text{Split}} = \left[\widehat{Y}_t - \widehat{q}_{1-\alpha},\ \widehat{Y}_t + \widehat{q}_{1-\alpha} \right].$$

The calibration scores and the resulting radius remain fixed throughout the entire test period. Therefore, newly observed test residuals do not affect subsequent Split CP intervals.

Weighted Conformal Prediction. Weighted conformal prediction (WCP) (Barber et al., 2023) modifies the conformal quantile by assigning different weights to past conformity scores. We use the same absolute prediction errors as in Split CP but assign decaying recency weights exponentially:

$$w_{i,t} = \lambda^{t-i}, \qquad \lambda = 0.99.$$

Thus, more recent observations receive greater weight. Following the weighted conformal construction, the test point is assigned unit mass at infinity when determining the weighted $(1 - \alpha)$ quantile. Let $\widehat{q}^{w}_{1-\alpha,t}$ denote the resulting weighted conformal quantile. The WCP interval is

$$\widehat{C}_t^{\text{WCP}} = \left[\widehat{Y}_t - \widehat{q}^{w}_{1-\alpha,t},\ \widehat{Y}_t + \widehat{q}^{w}_{1-\alpha,t} \right].$$

WCP utilizes a rolling residual window of fixed length. At the beginning of the test period, the residual window consists of the n_{cal} calibration residuals, and after Y_t is revealed, the new absolute prediction residual is inserted, and only the most recent n_{cal} residuals are retained.

Adaptive Conformal Inference. Adaptive Conformal Inference (ACI) (Gibbs & Candès, 2021) dynamically adjusts the effective miscoverage level according to previous coverage errors. Let α denote the target miscoverage rate and α_t the level used to construct the interval at time t. Starting from $\alpha_1 = \alpha$, ACI updates

$$\alpha_{t+1} = \alpha_t + \gamma \left[\alpha - \mathbf{1} \left\{ Y_t \notin \widehat{C}_t(\alpha_t) \right\} \right], \qquad \gamma = 0.01. \tag{36}$$

At time t, the empirical $(1-\alpha_t)$ quantile of the current absolute-residual window determines the interval radius. If Y_t is not covered, α_t decreases, and the subsequent interval becomes more conservative; if Y_t is covered, α_t increases. As in our WCP implementation, the residual window is initialized using the calibration residuals and updated after each observed test response while retaining only the most recent n_{cal} residuals. We do not truncate the recursive value of α_t to $[0,1]$; when the implied quantile level falls outside this range, we use the corresponding maximum or minimum residual in the current window, resulting in a finite-radius implementation of ACI.

Aggregated Adaptive Conformal Inference. Aggregated Adaptive Conformal Inference (AgACI) (Zaffran et al., 2022) reduces the sensitivity of ACI to the choice of a single learning rate by running multiple ACI experts in parallel. We employ the 30 learning rates adopted in the official AgACI implementation. Each $\gamma \in \Gamma$ defines one ACI expert that updates its adaptive level $\alpha_{t,\gamma}$ according to Equation (36) and consequently produces its own lower and upper prediction endpoints. Rather than directly averaging these intervals, the lower and upper endpoints are aggregated separately using Bernstein Online Aggregation (BOA). The BOA updates are based on pinball-loss gradients at levels $\alpha/2$ for the lower endpoint and $1-\alpha/2$ for the upper endpoint. After Y_t is revealed, each ACI expert updates its adaptive level, and the common absolute-residual window is updated while retaining the most recent n_{cal} observations. The same finite-radius convention described above is applied to each ACI expert when its implied quantile level falls outside $[0,1]$.

Conformalized Quantile Regression. Conformalized Quantile Regression (CQR) (Romano et al., 2019) first estimates conditional quantiles rather than constructing a symmetric interval around a point forecast. For nominal coverage $1-\alpha$, the forecasting backbone jointly estimates

$$\widehat{q}_{\alpha/2}(X_t) \quad \text{and} \quad \widehat{q}_{1-\alpha/2}(X_t),$$

using the pinball loss. The corresponding raw quantile interval is

$$\left[\widehat{q}_{\alpha/2}(X_t),\, \widehat{q}_{1-\alpha/2}(X_t)\right].$$

For each observation in the calibration sample, the CQR conformity score is

$$S_i = \max\left\{\widehat{q}_{\alpha/2}(X_i) - Y_i,\, Y_i - \widehat{q}_{1-\alpha/2}(X_i)\right\}.$$

Let $\widehat{q}^{\text{CQR}}_{1-\alpha}$ denote the finite-sample conformal quantile of these scores. The final interval is

$$\widehat{C}^{\text{CQR}}_t = \left[\widehat{q}_{\alpha/2}(X_t) - \widehat{q}^{\text{CQR}}_{1-\alpha},\, \widehat{q}_{1-\alpha/2}(X_t) + \widehat{q}^{\text{CQR}}_{1-\alpha}\right].$$

For 95% nominal coverage, we jointly train the lower and upper quantile outputs at levels 0.025 and 0.975. DLinear, PatchTST, and TimesNet are used as the corresponding quantile-regression backbones with their scalar output heads replaced by two-quantile output heads. The two raw outputs are ordered at prediction time to avoid quantile crossing. The conformal correction is computed once from the fixed calibration sample and remains unchanged throughout the test period. Since the CQR conformity scores are signed relative to the raw interval boundaries, the resulting conformal correction may be negative and can therefore contract an initially conservative quantile interval. If such contraction makes the two final endpoints cross, we project the resulting empty interval to its midpoint rather than exchanging the lower and upper endpoints.

Ensemble Batch Prediction Intervals. Finally, Ensemble Batch Prediction Intervals (EnbPI) (Xu & Xie, 2021) constructs sequential prediction intervals through bootstrap ensemble forecasting and out-of-bag (OOB) residuals without requiring a separate conformal calibration set. We use $B = 30$ bootstrap models, mean aggregation, and batch size $s = 1$. Let S_b denote the bootstrap sample used to train model $\widehat{f}^{\,b}$. For each observation i in the initial EnbPI sample, only bootstrap models for which $i \notin S_b$ are aggregated to form its OOB prediction,

$$\widehat{f}_{-i}(X_i) = \frac{1}{|\mathcal{B}_i|} \sum_{b \in \mathcal{B}_i} \widehat{f}^{\,b}(X_i), \qquad \mathcal{B}_i = \{b : i \notin S_b\}.$$

The corresponding absolute OOB residual is

$$S^{\text{OOB}}_i = \left|Y_i - \widehat{f}_{-i}(X_i)\right|.$$

For a future input X_t, the same OOB model sets are used to produce leave-one-out ensemble predictions $\{\widehat{f}_{-i}(X_t)\}$ across the initial observations. Following Algorithm 1 of Xu & Xie (2021), the prediction center is the empirical $(1-\alpha)$ quantile of these leave-one-out ensemble predictions,

$$\widehat{Y}_t^{\mathrm{EnbPI}} = \widehat{Q}_{1-\alpha}\left(\{\widehat{f}_{-i}(X_t)\}_{i=1}^{T_0}\right),$$

where T_0 denotes the number of initial EnbPI prediction–response pairs. The interval radius is obtained from the empirical $(1-\alpha)$ quantile of the current absolute-residual window,

$$w_t = \widehat{Q}_{1-\alpha}\left(\{S_i^{\mathrm{OOB}}\}\right)$$

and the resulting symmetric prediction interval is

$$\widehat{C}_t^{\mathrm{EnbPI}} = \left[\widehat{Y}_t^{\mathrm{EnbPI}} - w_t,\ \widehat{Y}_t^{\mathrm{EnbPI}} + w_t\right].$$

Because $s = 1$, once Y_t is revealed, the new absolute prediction residual is inserted into the residual window and the oldest residual is removed, so the residual distribution is updated after every test observation.

Unlike Split CP, WCP, ACI, AgACI, and CQR, EnbPI does not use the designated 10% calibration segment as a separate conformal calibration sample and does not require a separate validation set for training its bootstrap models. Instead, each bootstrap model is trained independently, with early stopping determined by the training MSE on its own bootstrap sample. To ensure an identical evaluation horizon across all six methods, the initial EnbPI history is constructed entirely from observations available before the common test period and has the same raw length as the training portion used in the corresponding point forecasting experiment. The maximum number of epochs and the early-stopping patience are set according to the corresponding forecasting backbone.

F.4 Additional results

This section reports the remaining empirical results for the four datasets under DLinear, PatchTST, and TimesNet with $\delta \in \{0.10, 0.25\}$. Throughout the empirical analysis, we set $R = 1$, $a_{rt} = 1$, and $\epsilon_r = 0$, so that $\tau_r = \alpha_r = 0.05$. For the six candidate methods, we use the balanced allocation $\eta_{mr} = \delta/6$. The values of the mixture tuning parameter $\rho^\star$ are reported in Table 3. As in the main analysis, we examine cumulative coverage, the ratio $\widehat{\Gamma}_{mr,t}/\mathrm{rad}_{mr,t}$, and the cumulative mean interval width of each method to explain the evolution of the CC-SMCS model set over time.

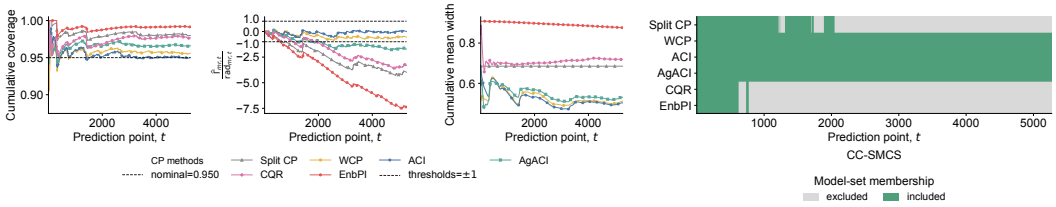
For the French electricity price dataset, PatchTST and TimesNet exhibit patterns broadly similar to those observed under DLinear. We focus on the time at which the certified set first becomes nonempty. Under PatchTST, Split CP and EnbPI are the first methods to enter the certified set because their cumulative coverage performance becomes sufficiently favorable. Since Split CP has a larger cumulative mean interval width than EnbPI, it is excluded from the final model set earlier. As the cumulative coverage performance of CQR improves, CQR also enters the certified set, but is subsequently excluded because of its relatively large interval width. CC-SMCS also allows a method to re-enter the final model set when its relative efficiency improves. WCP provides such an example: while remaining in the possible set, it is excluded when its cumulative mean interval width exceeds the minimum cumulative mean width among the certified methods, q_t^{cert}, and is re-included once its cumulative mean interval width falls below q_t^{cert} (Figures 6a and 7a). Similar patterns are observed for DLinear with $\delta = 0.25$ (Figure 5a) and for TimesNet under both $\delta = 0.10$ and $\delta = 0.25$ (Figures 8a and 9a).

For the Electricity Load dataset, the results under DLinear with $\delta = 0.25$ are similar to those observed for the French electricity price dataset (Figure 5b). TimesNet also exhibits similar dynamics under both $\delta = 0.10$ and $\delta = 0.25$, with the model set evolving as the coverage performance and interval widths of the candidate methods change over time (Figures 8b and 9b). In contrast, a different pattern emerges under PatchTST. CQR becomes infeasible because its cumulative coverage remains substantially below the nominal level of 95%, causing $\widehat{\Gamma}_{mr,t}/\mathrm{rad}_{mr,t}$ to exceed 1. As a result, CQR leaves the possible set and is therefore excluded from the final model set (Figures 6b and 7b).

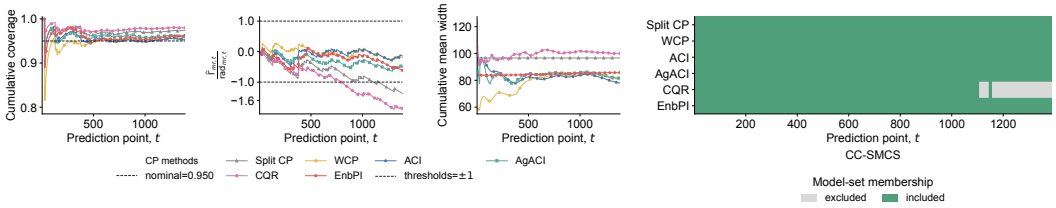
For the Weather dataset, CC-SMCS generally excludes methods with relatively wide intervals once a narrower method becomes certified feasible, while retaining methods with adequate coverage and

relatively narrow intervals. This pattern is consistent across DLinear, PatchTST, and TimesNet, as shown in Figures 4a, 6c, 7c, 8c, and 9c.

For the Wind Power dataset, the model-set paths differ across the three forecasting backbones. Under DLinear and PatchTST, the certified set remains empty until near the end of the prediction period. Split CP is the first method to enter the certified set but also has the largest cumulative mean interval width, and it is excluded only after EnbPI becomes certified with a smaller width (Figures 4b, 5d, 6d, and 7d). This reflects relatively weak separation in the coverage–width profiles of the candidate pipelines, so the available evidence is insufficient to reduce the model set earlier. In contrast, under TimesNet, Split CP and EnbPI become certified earlier and the differences in interval width are more pronounced, leading to earlier exclusions of Split CP and CQR (Figures 8d and 9d). The contrast across forecasting backbones illustrates that the informativeness of CC-SMCS depends not only on the conformal calibration method but also on the underlying forecasting backbone, since changes in the base forecaster can alter the coverage–efficiency profiles of the resulting candidate pipelines.



(a) Weather



(b) Wind Power

Figure 4: Empirical results for DLinear at 95% nominal coverage with $\delta = 0.10$.

G More Discussions and Extensions

G.1 Delayed multi-horizon feedback

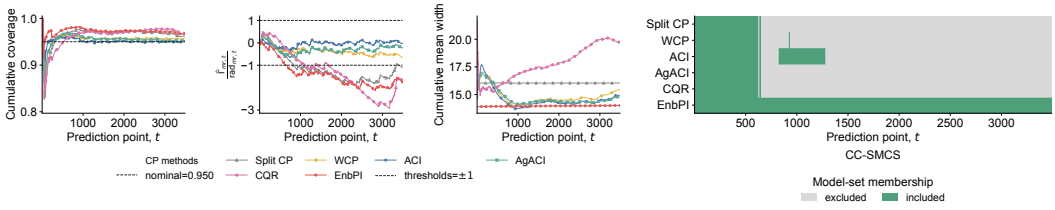
For an h-step forecast issued at origin s, C_{mrsh} is $\mathcal{G}_s$-measurable and is evaluated using Y_{s+h}. If its increment is inserted at calendar time $s+h$ into the usual filtration, conditioning on $\mathcal{G}_{s+h-1}$ need not recover the forecast-origin conditional mean given $\mathcal{G}_s$. We therefore partition forecast origins into residue classes

$$\mathcal{I}_{h,k} = \{k, k+h, k+2h, \ldots\}, \qquad k = 1, \ldots, h. \tag{37}$$

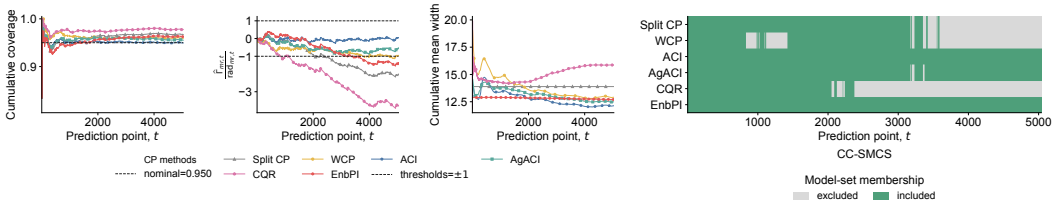
Within each class, an increment issued at one origin is resolved by the next origin in the same class. We run a separate confidence sequence for every (m, r, h, k) stream and allocate the error probability across all streams.

Corollary G.1. *If the streamwise allocations satisfy $\sum_{m,r,h,k} \eta_{mrhk} \leq \delta$, then Theorem 5.2 holds simultaneously over all calendar times and monitored horizons, with each horizon-specific target defined over forecast origins whose outcomes have been resolved.*

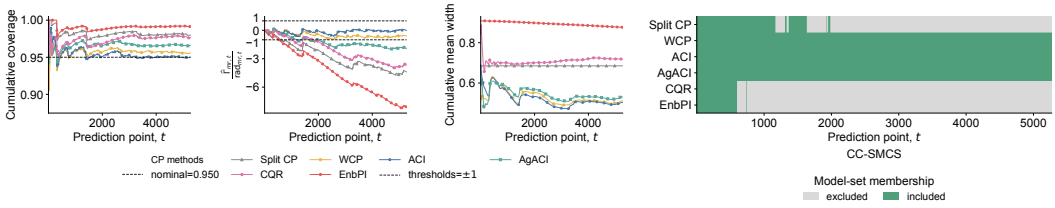
Before proving Corollary G.1, we reformulate the problem and derive the following preliminary results first. For each monitored horizon h, a forecast issued at origin s is $\mathcal{G}_s$-measurable and targets



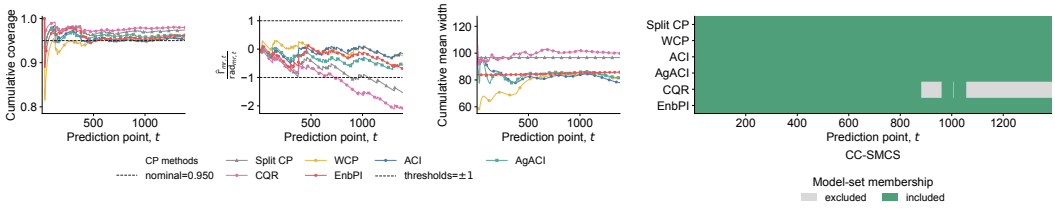
(a) French Electricity Price



(b) Electricity Load

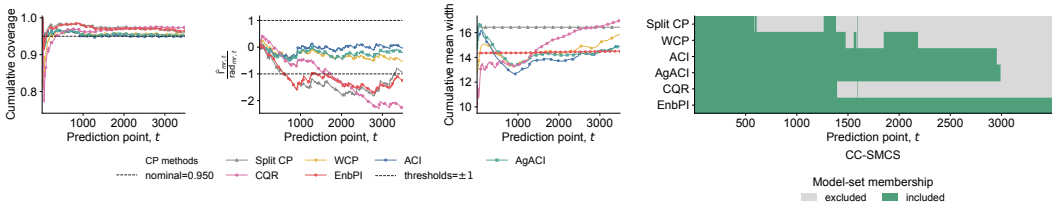


(c) Weather

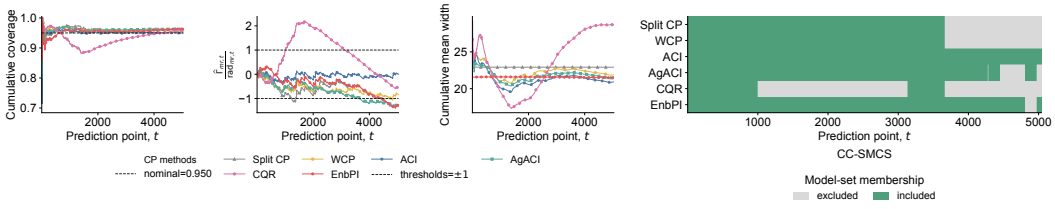


(d) Wind Power

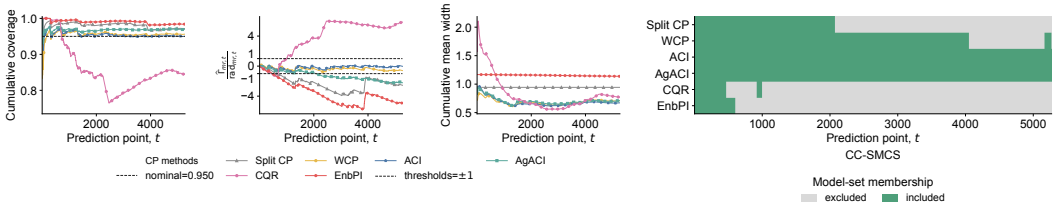
Figure 5: Empirical results for DLinear at 95% nominal coverage with $\delta = 0.25$.



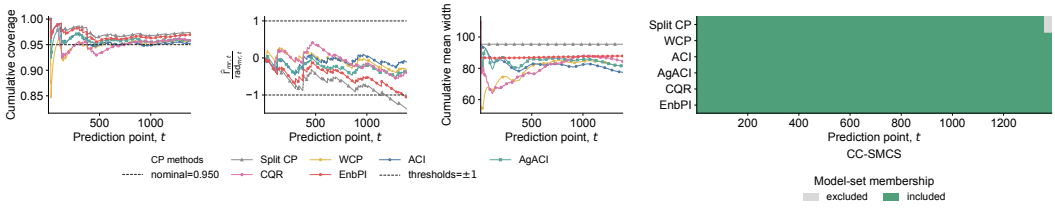
(a) French Electricity Price



(b) Electricity Load

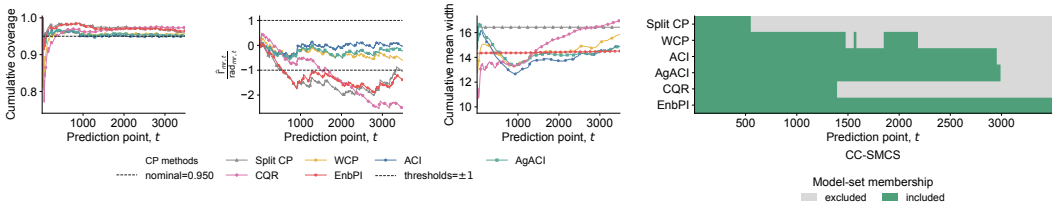


(c) Weather

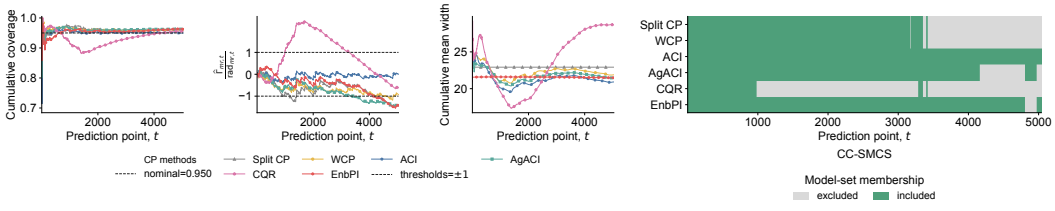


(d) Wind Power

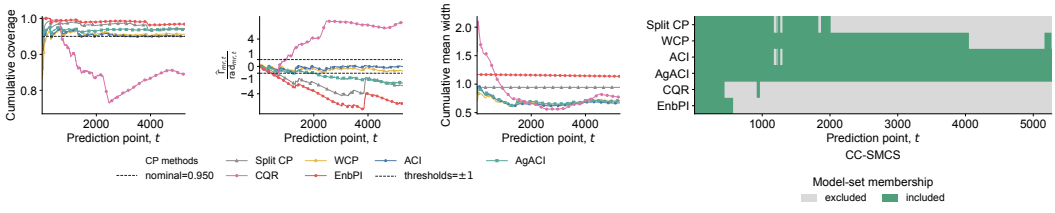
Figure 6: Empirical results for PatchTST at 95% nominal coverage with $\delta = 0.10$.



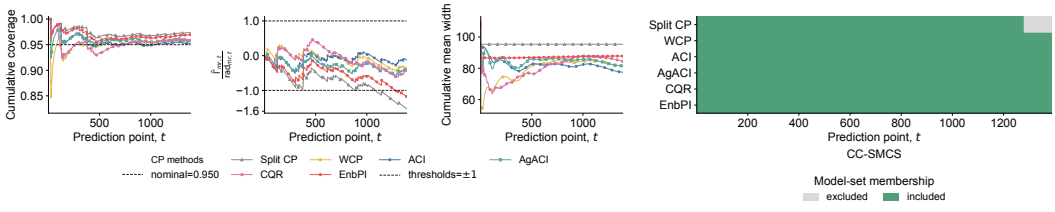
(a) French Electricity Price



(b) Electricity Load

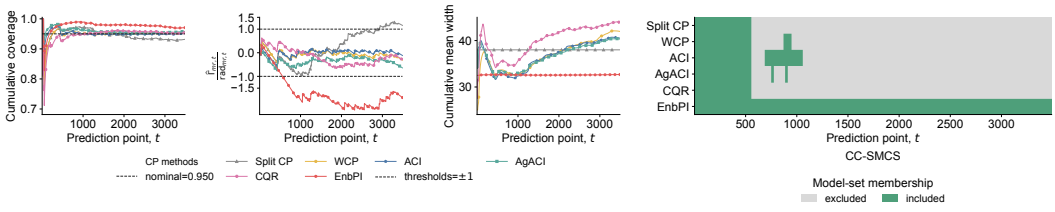


(c) Weather

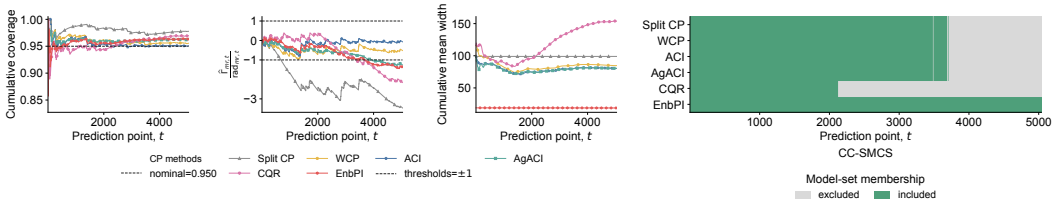


(d) Wind Power

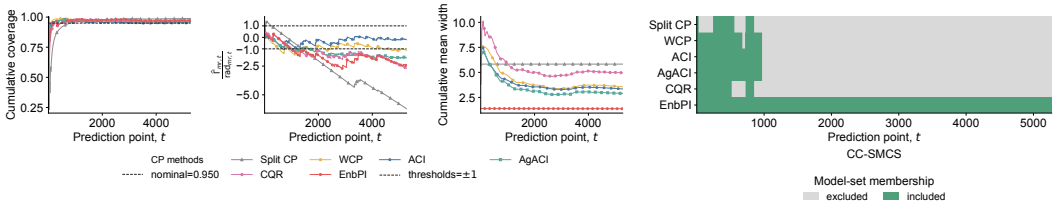
Figure 7: Empirical results for PatchTST at 95% nominal coverage with $\delta = 0.25$.



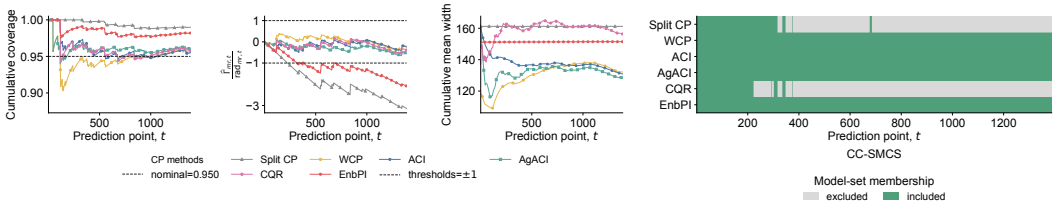
(a) French Electricity Price



(b) Electricity Load

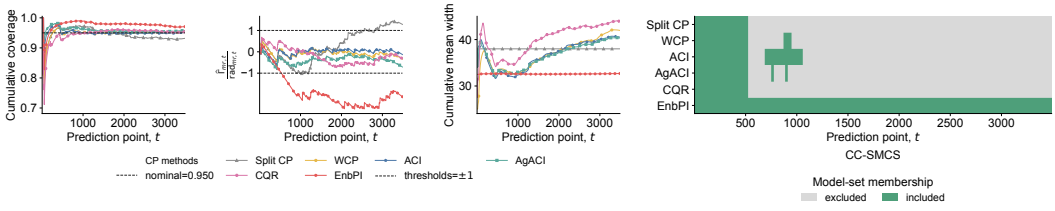


(c) Weather

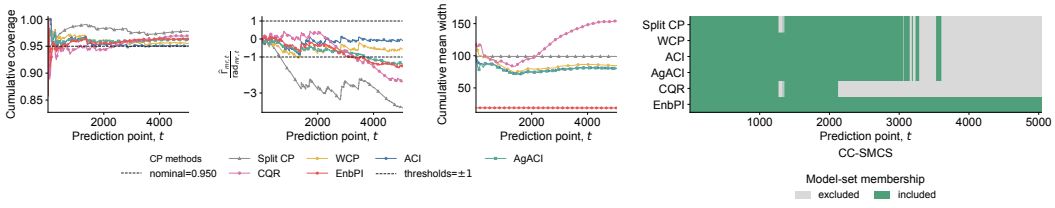


(d) Wind Power

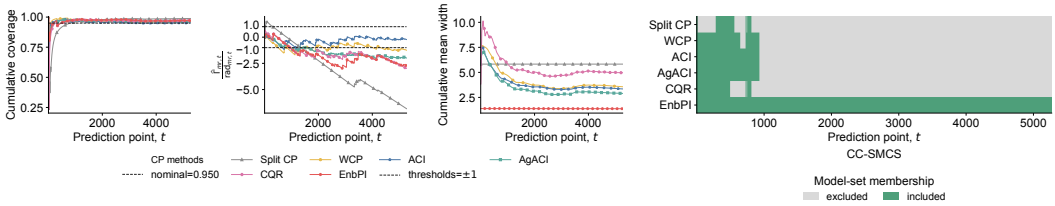
Figure 8: Empirical results for TimesNet at 95% nominal coverage with $\delta = 0.10$.



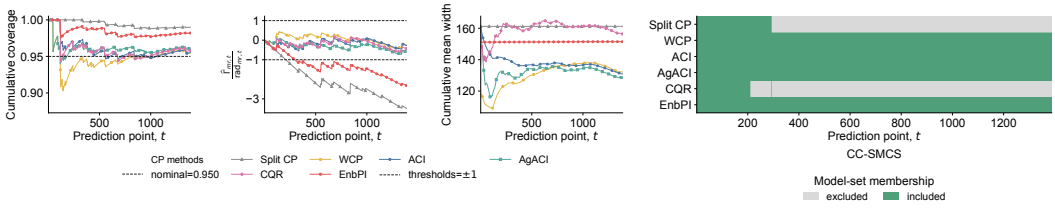
(a) French Electricity Price



(b) Electricity Load



(c) Weather



(d) Wind Power

Figure 9: Empirical results for TimesNet at 95% nominal coverage with $\delta = 0.25$.

Y_{s+h}. Define

$$B_{mrsh} := \mathbf{1}\{Y_{s+h} \notin C_{mrsh}\}, \qquad X_{mrsh} := a_{rsh}(B_{mrsh} - \tau_{rh}), \tag{38}$$

$$\widehat{B}_{mrsh} := \begin{cases} \tau_{rh}, & A_{rh,s-h} = 0, \\ \dfrac{\sum_{u=1}^{s-h} a_{ruh} B_{mruh}}{A_{rh,s-h}}, & A_{rh,s-h} > 0, \end{cases} \tag{39}$$

$$\widehat{X}_{mrsh} := a_{rsh}(\widehat{B}_{mrsh} - \tau_{rh}), \qquad \mu_{mrsh} := \mathbb{E}[X_{mrsh} \mid \mathcal{G}_s], \tag{40}$$

where $a_{rsh} \in [0,1]$ is $\mathcal{G}_s$-measurable and $A_{rh,s-h} := \sum_{u=1}^{s-h} a_{ruh}$, with sums over empty index sets understood as zero.

Fix $k \in \{1, \dots, h\}$ and enumerate the origins in the kth residue stream by

$$s_{k,n} := k + nh, \qquad n = 0, 1, 2, \dots.$$

Define

$$\mathcal{H}_{k,n} := \mathcal{G}_{s_{k,n}}, \qquad D_{mrh,k,n+1} := X_{mr,s_{k,n},h} - \mu_{mr,s_{k,n},h}.$$

Since $s_{k,n} + h = s_{k,n+1}$, we have

$$X_{mr,s_{k,n},h} \in \mathcal{H}_{k,n+1}, \qquad \mu_{mr,s_{k,n},h} \in \mathcal{H}_{k,n}, \qquad D_{mrh,k,n+1} \in \mathcal{H}_{k,n+1}.$$

Moreover, since $\mathcal{H}_{k,n} = \mathcal{G}_{s_{k,n}}$ and $\mu_{mr,s_{k,n},h} = \mathbb{E}[X_{mr,s_{k,n},h} \mid \mathcal{G}_{s_{k,n}}]$,

$$\begin{aligned} \mathbb{E}[D_{mrh,k,n+1} \mid \mathcal{H}_{k,n}] &= \mathbb{E}[X_{mr,s_{k,n},h} - \mu_{mr,s_{k,n},h} \mid \mathcal{G}_{s_{k,n}}] \\ &= \mathbb{E}[X_{mr,s_{k,n},h} \mid \mathcal{G}_{s_{k,n}}] - \mu_{mr,s_{k,n},h} \\ &= 0. \end{aligned}$$

Conditionally on $\mathcal{H}_{k,n}$, define

$$\Delta_{mrh,k,n+1} := (X_{mr,s_{k,n},h} - \widehat{X}_{mr,s_{k,n},h})^2.$$

Since $B_{mr,s_{k,n},h} \in \{0,1\}$, $\widehat{B}_{mr,s_{k,n},h} \in [0,1]$, and $a_{r,s_{k,n},h} \in [0,1]$,

$$X_{mr,s_{k,n},h} - \widehat{X}_{mr,s_{k,n},h} \in [-a_{r,s_{k,n},h}, a_{r,s_{k,n},h}] \subseteq [-1,1],$$

By Lemma B.1, for every $\lambda \in [0,1)$,

$$\mathbb{E}\left[\exp\{\lambda D_{mrh,k,n+1} - \psi_E(\lambda)\Delta_{mrh,k,n+1}\} \middle| \mathcal{H}_{k,n}\right] \le 1.$$

Thus, Lemma B.2 applies to each residue stream with

$$V_{mrh,k,N} := \sum_{n=0}^{N-1} (X_{mr,s_{k,n},h} - \widehat{X}_{mr,s_{k,n},h})^2 = \sum_{n=0}^{N-1} \Delta_{mrh,k,n+1}.$$

For every (m, r, h, k),

$$\mathbb{P}\left(\forall N \ge 1 : \left|\sum_{n=0}^{N-1} D_{mrh,k,n+1}\right| \le b_{\eta_{mrhk},\rho}(V_{mrh,k,N})\right) \ge 1 - \eta_{mrhk}.$$

At calendar time t, define the number of resolved origins in stream k by

$$N_{h,k}(t) := \#\{n \ge 0 : s_{k,n} + h \le t\}.$$

With empty sums understood as zero, define

$$S_{mrh,k}(t) := \sum_{n=0}^{N_{h,k}(t)-1} X_{mr,s_{k,n},h},$$

$$A_{rh,k}(t) := \sum_{n=0}^{N_{h,k}(t)-1} a_{r,s_{k,n},h},$$

$$V_{mrh,k}(t) := \sum_{n=0}^{N_{h,k}(t)-1} (X_{mr,s_{k,n},h} - \widehat{X}_{mr,s_{k,n},h})^2.$$

Taking $N = N_{h,k}(t)$ in the streamwise confidence sequence gives

$$\left| S_{mrh,k}(t) - \sum_{n=0}^{N_{h,k}(t)-1} \mu_{mr,s_{k,n},h} \right| \le b_{\eta_{mrhk},\rho}(V_{mrh,k}(t))$$

for all stream resolution times; when $N_{h,k}(t) = 0$, both sums are zero and the inequality holds trivially.

Let $\mathcal{E}_{mrhk}$ denote the time-uniform event for stream (m, r, h, k). Then

$$\mathbb{P}\left(\bigcap_{m,r,h,k} \mathcal{E}_{mrhk} \right) \ge 1 - \sum_{m,r,h,k} \mathbb{P}(\mathcal{E}_{mrhk}^c) \ge 1 - \sum_{m,r,h,k} \eta_{mrhk} \ge 1 - \delta.$$

Define

$$A_{rh}^{\text{delay}}(t) := \sum_{k=1}^{h} A_{rh,k}(t).$$

When $A_{rh}^{\text{delay}}(t) > 0$, define

$$\Gamma_{mrh,t}^{\text{delay}} := \frac{\sum_{k=1}^{h} \sum_{n=0}^{N_{h,k}(t)-1} \mu_{mr,s_{k,n},h}}{A_{rh}^{\text{delay}}(t)}.$$

On the joint stream event, for every k,

$$\left| S_{mrh,k}(t) - \sum_{n=0}^{N_{h,k}(t)-1} \mu_{mr,s_{k,n},h} \right| \le b_{\eta_{mrhk},\rho}(V_{mrh,k}(t)).$$

Therefore, by the triangle inequality,

$$\begin{aligned}
\left| \sum_{k=1}^{h} S_{mrh,k}(t) - \sum_{k=1}^{h} \sum_{n=0}^{N_{h,k}(t)-1} \mu_{mr,s_{k,n},h} \right| &= \left| \sum_{k=1}^{h} \left(S_{mrh,k}(t) - \sum_{n=0}^{N_{h,k}(t)-1} \mu_{mr,s_{k,n},h} \right) \right| \\
&\le \sum_{k=1}^{h} \left| S_{mrh,k}(t) - \sum_{n=0}^{N_{h,k}(t)-1} \mu_{mr,s_{k,n},h} \right| \\
&\le \sum_{k=1}^{h} b_{\eta_{mrhk},\rho}(V_{mrh,k}(t)).
\end{aligned}$$

For brevity, write

$$B_{mrh}^{\text{delay}}(t) := \sum_{k=1}^{h} b_{\eta_{mrhk},\rho}(V_{mrh,k}(t)).$$

The preceding inequality becomes

$$-B_{mrh}^{\text{delay}}(t) \le \sum_{k=1}^{h} S_{mrh,k}(t) - A_{rh}^{\text{delay}}(t)\Gamma_{mrh,t}^{\text{delay}} \le B_{mrh}^{\text{delay}}(t).$$

Rearranging,

$$\sum_{k=1}^{h} S_{mrh,k}(t) - B_{mrh}^{\text{delay}}(t) \le A_{rh}^{\text{delay}}(t)\Gamma_{mrh,t}^{\text{delay}} \le \sum_{k=1}^{h} S_{mrh,k}(t) + B_{mrh}^{\text{delay}}(t).$$

Since $A_{rh}^{\text{delay}}(t) > 0$, dividing by $A_{rh}^{\text{delay}}(t)$ gives

$$L_{mrh,t}^{\text{delay}} \le \Gamma_{mrh,t}^{\text{delay}} \le U_{mrh,t}^{\text{delay}},$$

where

$$L^{\mathrm{delay}}_{mrh,t} := \frac{\sum_{k=1}^{h} S_{mrh,k}(t) - \sum_{k=1}^{h} b_{\eta_{mrhk},\rho}(V_{mrh,k}(t))}{A^{\mathrm{delay}}_{rh}(t)},$$

$$U^{\mathrm{delay}}_{mrh,t} := \frac{\sum_{k=1}^{h} S_{mrh,k}(t) + \sum_{k=1}^{h} b_{\eta_{mrhk},\rho}(V_{mrh,k}(t))}{A^{\mathrm{delay}}_{rh}(t)}.$$

If $A^{\mathrm{delay}}_{rh}(t) = 0$, set

$$\Gamma^{\mathrm{delay}}_{mrh,t} := 0, \qquad L^{\mathrm{delay}}_{mrh,t} := -\infty, \qquad U^{\mathrm{delay}}_{mrh,t} := +\infty.$$

Therefore

$$\mathbb{P}\left(\forall t, m, r, h : \; L^{\mathrm{delay}}_{mrh,t} \le \Gamma^{\mathrm{delay}}_{mrh,t} \le U^{\mathrm{delay}}_{mrh,t}\right) \ge 1 - \delta.$$

Proof of Corollary G.1. On the preceding joint event, regard each pair (r, h) as a coverage-constraint index and apply the definitions of Section 4 with the delayed quantities. If a method is certified, then for every monitored (r, h),

$$\Gamma^{\mathrm{delay}}_{mrh,t} \le U^{\mathrm{delay}}_{mrh,t} \le 0,$$

so it is truly feasible. Conversely, true feasibility gives

$$L^{\mathrm{delay}}_{mrh,t} \le \Gamma^{\mathrm{delay}}_{mrh,t} \le 0$$

for every monitored (r, h), so the method is possibly feasible. Thus, for every t,

$$\widehat{\mathcal{V}}^{\mathrm{cert}}_t \subseteq \mathcal{V}^{\star}_t \subseteq \widehat{\mathcal{V}}^{\mathrm{pos}}_t.$$

Now let $m^{\star} \in \mathcal{M}^{\star}_t$. Since every $j \in \widehat{\mathcal{V}}^{\mathrm{cert}}_t$ is truly feasible,

$$Q_{m^{\star},t} \le Q_{j,t}, \qquad \forall j \in \widehat{\mathcal{V}}^{\mathrm{cert}}_t.$$

Therefore $Q_{m^{\star},t} \le q^{\mathrm{cert}}_t$, with $q^{\mathrm{cert}}_t = +\infty$ when $\widehat{\mathcal{V}}^{\mathrm{cert}}_t = \varnothing$. Since $m^{\star} \in \widehat{\mathcal{V}}^{\mathrm{pos}}_t$,

$$m^{\star} \in \left\{ m \in \widehat{\mathcal{V}}^{\mathrm{pos}}_t : Q_{m,t} \le q^{\mathrm{cert}}_t \right\} = \widehat{\mathcal{M}}^{\mathrm{proj}}_t.$$

Hence $\mathcal{M}^{\star}_t \subseteq \widehat{\mathcal{M}}^{\mathrm{proj}}_t$ simultaneously over all calendar times and monitored horizons. The joint event has probability at least $1 - \delta$, which proves the corollary. □

G.2 Outcome-Dependent Efficiency Objectives

The main text uses a predictable set-size objective. Suppose instead that method m incurs an adapted loss ℓ_{mt} after Y_t is observed, and the target is

$$\overline{Q}_{m,t} = \frac{1}{W_t} \sum_{s=1}^{t} \mathbb{E}[w_s \ell_{ms} \mid \mathcal{G}_{s-1}].$$

For each ordered pair (m, j), define

$$D_{mj,t} := w_t(\ell_{mt} - \ell_{jt}), \qquad \Delta_{mj,t} := \overline{Q}_{m,t} - \overline{Q}_{j,t} = \frac{1}{W_t} \sum_{s=1}^{t} \mathbb{E}[D_{mj,s} \mid \mathcal{G}_{s-1}].$$

Suppose that simultaneous lower confidence sequences $L^{\Delta}_{mj,t}$ can be constructed such that

$$\mathbb{P}\big(\forall t, m, j : \; L^{\Delta}_{mj,t} \le \Delta_{mj,t}\big) \ge 1 - \delta_Q.$$

Such confidence sequences can be obtained, for example, when the centered pairwise increments admit predictable bounds or an appropriate conditional sub-ψ process. Combine this event with the feasibility confidence event using total error $\delta_V + \delta_Q \le \delta$, and define

$$\widehat{\mathcal{M}}^{\mathrm{loss}}_t = \left\{ m \in \widehat{\mathcal{V}}^{\mathrm{pos}}_t : L^{\Delta}_{mj,t} \le 0 \text{ for every } j \in \widehat{\mathcal{V}}^{\mathrm{cert}}_t \right\}. \tag{41}$$

If $m^{\star}$ is a true constrained optimizer and j is certified, then j is truly feasible and $\Delta_{m^{\star}j,t} \le 0$; the lower confidence bound cannot be positive on the joint event. Therefore $m^{\star} \in \widehat{\mathcal{M}}^{\mathrm{loss}}_t$ for all t. Pairwise differences exploit the fact that all methods are evaluated on the same outcomes, but (41) is generally an outer set rather than the exact projection of a coherent joint objective region.