Coverage debt: pricing the conformal coverage gap through a real electricity-market redesign

Phongsakon Mark Konrad[a], Tim Lukas Adam[a], Serkan Ayvaz[a,*]

[a] *SDU Centre for Industrial Software, The Mærsk Mc-Kinney Møller Institute, University of Southern Denmark, Sønderborg 6400, Denmark*

Abstract

Probabilistic price forecasts are useful to an electricity-market bidder only if their stated reliability holds when the market itself changes. We study what happens to distribution-free prediction intervals, and to a safety rule that turns them into a guaranteed worst-case loss limit, when a real balancing-market redesign expands price volatility. Across eight European bidding zones spanning one such redesign, a static calibration loses reliability in step with the volatility increase, and the certified loss limit is then breached at the same rate the intervals under-cover, so a statistical shortfall becomes a direct financial exposure we call coverage debt. Refreshing the calibration, on a schedule or continuously, repays the debt and restores the guarantee, with weighted variants giving the tightest intervals at equal reliability. We release the multi-zone benchmark, including non-reforming control markets and placebo dates, so the effect can be re-measured beyond our data.

Keywords: Conformal prediction, Electricity price forecasting, Imbalance settlement, Balancing markets, Distribution shift, Decision-focused evaluation, Uncertainty quantification, Probabilistic forecasting

1. Introduction

Two days after the Nordic balancing market changed its settlement rules in March 2025, the imbalance price in western Denmark reached more than three hundred times a normal level inside a single fifteen-minute window [1]. A forecasting model tuned on the old market had no reason to expect that, and neither did the safety margins a bidder would compute from it.

*Corresponding author

Email address: seay@mmmi.sdu.dk (Serkan Ayvaz)

This is the situation we study. An automated bidder in a balancing market does not act on a price forecast alone. It acts on an interval around that forecast, and it trusts the interval to bound how much a bad clearing price can cost. When the interval is well calibrated, the bidder can refuse any offer whose worst outcome inside the interval falls below an operator loss limit, and the refusal carries a distribution-free guarantee. The guarantee is only as good as the calibration behind it. If the market that generates the prices is redesigned, the calibration built before the redesign may no longer hold, and the guarantee it certifies can fail without any visible warning.

Conformal prediction gives calibrated intervals without distributional assumptions, and a growing body of work now feeds conformal intervals into electricity-market decisions. Two facts about this literature matter here. Methods that stay valid under distribution shift are tested on synthetic drift or on day-ahead prices, and their reported quantity is the interval, not a decision. Work that places conformal intervals inside a bidding rule reports realized returns and keeps its coverage claim on the price interval. Neither line asks what a real market redesign does to a decision-level guarantee, even though the European balancing markets have just been through one and the day-ahead markets are following.

We ask that question on the 2025 Nordic balancing-market redesign and answer it with a single quantity. When the redesign expands a zone's imbalance-price volatility, a static conformal calibration under-covers in proportion to the expansion, and a bidder's certified worst-case loss floor is then breached at the same rate the intervals under-cover. A coverage gap that reads as a harmless statistical shortfall becomes a priced financial exposure. We measure this across eight bidding zones, we show that either a periodic or an online refresh of the calibration repays the exposure, and we release the data and code so the relationship can be checked elsewhere. The mechanism itself is expected from conformal theory. The contribution is to price it on a real market at the level of the decision, and to give a reusable way to measure it.

This paper does not claim a new calibration method or a new guarantee. The methods are standard, and the coverage gap under shift is already bounded in theory. What is new is the decision-level, real-market quantification of the gap, the observation that whether the exposure is measurable at all depends on the settlement payoff the safety rule assumes, and the benchmark that supports both.

Contributions.

- We show that a conformal coverage gap through a real market redesign is a priced economic exposure rather than a cosmetic statistical shortfall, and that whether it is visible at all depends on the settlement payoff the safety rule assumes.

- We characterise how the exposure grows with reform-induced volatility, and which off-the-shelf refresh of the calibration repays it at what sharpness cost.
- We release a multi-zone European benchmark, with non-reforming control markets and placebo dates, for evaluating conformal calibration through real structural breaks.

2. Background and related work

Electricity price forecasting is a mature field, and the review of Weron [2] organises its point-forecasting methods while Lago et al. [3] fix an open benchmark of strong statistical and neural baselines. A bidder needs more than a point forecast, and the probabilistic branch surveyed by Nowotarski and Weron [4] supplies intervals and densities. Its workhorse is quantile regression averaging [5], extended by regularisation [6] and by distributional neural networks [7]. These methods calibrate well when the quantile model is correctly specified and the price process is stable. A market redesign breaks both conditions at once. A parametric interval then loses coverage with no warning, whereas a distribution-free interval carries a finite-sample guarantee from a calibration set to test data and degrades measurably, not silently, when the assumption behind it weakens. Conformalised quantile regression [8] bridges the two by wrapping a quantile forecaster in a conformal guarantee, and it is why we calibrate with conformal prediction rather than with a bare quantile-regression interval.

Conformal prediction gives finite-sample marginal coverage without distributional assumptions [9, 10, 11, 12]. Its validity rests on exchangeability, which a structural break removes, and two families of methods respond. Weighted and non-exchangeable conformal reweights the calibration set toward the recent past [13, 14], and online conformal adapts the working miscoverage level after each observation [15, 16]. Zaffran et al. [17] carry adaptive conformal prediction to electricity-price forecasting and remove the learning-rate choice by aggregating experts, and a line of time-series conformal methods tightens intervals by modelling the residual process [18, 19, 20, 21] or by controlling adaptive regret [22, 23]. Barber et al. [14]

bound the coverage gap that a static method incurs as the data leaves exchangeability, and that bound is the theory our volatility gradient realises on a real market.

Conformal prediction has reached energy forecasting. Kath and Ziel [24] apply it to day-ahead and intraday power prices and compare it against quantile regression averaging, and conformal methods now calibrate wind-power and load intervals [25, 26, 27]. The balancing market is the most volatile and least studied of the three price stages [28], and it has its own probabilistic forecasting work [29, 30]. O'Connor et al. [31] feed conformal balancing-market prices into a storage-trading rule and report realised returns. That work keeps its coverage claim on the price interval and does not ask what a market redesign does to it.

A separate line controls a decision loss rather than the interval. Risk-controlling prediction sets [32], conformal risk control [33], and its online extension [34] calibrate a bounded monotone loss with finite-sample or long-run guarantees, and conformal decision theory [35] carries the idea to sequential decisions. In energy, value-oriented and decision-focused learning trains the forecaster against the downstream market objective [36, 37]. These results state a guarantee where coverage holds. None measures what the guarantee costs when coverage fails on a real market.

We take the shift-robust calibration the first family studies, the market-decision setting the energy-conformal work uses, and the loss-control guarantee the third line proves, and we ask a question none of them answers. What does a coverage gap cost, at the level of an accepted bid, when a real redesign expands price volatility? We answer it on the 2025 Nordic redesign, with non-reforming control markets and a placebo date so the effect can be separated from a common shock, in the spirit of the difference-in-differences designs used to study imbalance-pricing reform [38, 39].

3. Data and the 2025 balancing-market redesign

On 4 March 2025 the four Nordic transmission system operators launched the mFRR Energy Activation Market together, replacing a sixty-minute manually dispatched balancing arrangement with a fifteen-minute automatically cleared one. The change is a redesign of the process that generates imbalance prices, and it is dated and exogenous to any single market participant. It arrives at the same instant in every Nordic zone, so a comparison across time within one zone cannot separate the redesign from other events that share the date. We therefore compare treated zones against zones that did not undergo the redesign, and we compare the redesign date against a placebo date in an earlier year.

The imbalance price settles the gap between what a participant schedules and what it delivers, and it is the most volatile of the three prices a balancing-market participant faces, after the day-ahead and intraday prices. Before the redesign the Nordic imbalance price was set hourly through a manually dispatched process. After it, the price clears every fifteen minutes through an automatically ordered merit list, a faster and more reactive mechanism on a finer grid. The imbalance price derives from the prices of activated balancing energy, so it is not held to the day-ahead technical limits of minus five hundred to four thousand euros per megawatt-hour and it moves over a wider range. In our DK1 sample the hourly price runs from about minus eight hundred to forty-eight hundred euros per megawatt-hour, wide but finite. The loss floor does not rest on that range in any case, since the worst-case profit the verifier certifies is read off the endpoints of the finite conformal interval, as Section 4.3 makes precise. The redesign left the price mechanism's limits unchanged. What it changed is how violently the price moves within them, which Figure 1 shows for DK1.

We study eight bidding zones. Five are Nordic and treated by the redesign, namely the two Danish zones DK1 and DK2, the Swedish zone SE3, the Norwegian zone NO2, and Finland. Three are continental and serve as controls that did not undergo the Nordic redesign on that date, namely the Netherlands, Belgium, and Austria. Table 1 lists the zones, their roles, data sources, and calibration and test sample sizes. The placebo date is 4 March 2024, which sits inside the pre-redesign regime and carries no reform of its own.

Danish imbalance prices come from the Energinet Energi Data Service [40], which is public, free, and licensed CC-BY, so the benchmark redistributes the Danish series. Prices for the other six zones come from the ENTSO-E Transparency Platform [41], which requires a personal access token and whose reuse terms for redistribution are ambiguous, so the benchmark ships download scripts rather than the raw data. We work at hourly resolution throughout. The pre-redesign Nordic imbalance price is published only hourly, because the fifteen-minute price does not exist before the redesign, so hourly is the only resolution common to both sides of the break. Aggregating to the hourly grid averages away the sharpest fifteen-minute excursions, including the spike that opens this paper, so the volatility expansion and the coverage gap we measure are a conservative lower bound on what a bidder settling every fifteen minutes now faces. Coverage of ENTSO-E imbalance data varies by zone and month, and we exclude any zone that

Table 1: Data provenance for the eight bidding zones. Role marks whether the zone underwent the 2025 Nordic redesign (treated) or not (control). The volatility ratio is the post-break over pre-break standard deviation of the daily price increment on the matched hourly window. The last two columns give the conformal calibration and test sample sizes in hours. NO2 is treated but its volatility fell over the window, which is why it does not under-cover.

Zone	Role	Source	Vol. ratio	n_{calib}	n_{test}
DK1	treated	Energinet	3.19	2928	2122
DK2	treated	Energinet	2.60	2928	2122
SE3	treated	ENTSO-E	2.64	817	1800
NO2	treated	ENTSO-E	0.67	1211	1800
FI	treated	ENTSO-E	2.85	2928	2136
NL	control	ENTSO-E	0.98	2928	2136
BE	control	ENTSO-E	1.05	2928	2136
AT	control	ENTSO-E	1.10	2928	2136

does not clear a minimum of five hundred calibration hours and one thousand post-break hours.

4. Methods

4.1. Base forecaster

The claim under test concerns the calibration layer, not the point forecast, so we hold the base forecaster simple and state it plainly. Let y_t be the realised imbalance price in hour t and $\hat{y}_t$ its point forecast. The point forecast is the imbalance price of the same hour on the previous day, a seasonal-naive predictor with a daily period. The nonconformity score is the absolute residual

$$s_t = |y_t - \hat{y}_t|. \tag{1}$$

This choice is deliberate rather than a shortcut. Split-conformal validity rests on the exchangeability of the scores, not on the accuracy of the point forecast, and a regime change that expands price volatility breaks that exchangeability whatever the forecaster. A stronger model from the state-of-the-art benchmark of Lago et al. [3] shrinks the residual magnitude but does not restore the exchangeability a fixed quantile needs, so it moves the numbers without changing the mechanism. The central comparison is in any case relative, between a static and a refreshed calibration on the same residual stream, so the base forecaster is held fixed across the two and cannot confound it. On DK1, where a learned forecaster was available, static split-conformal covers 0.74 and the online method 0.89, the same pattern the

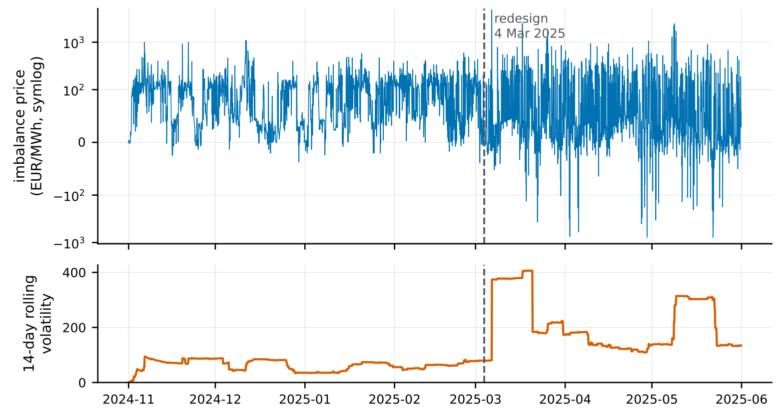

Figure 1: The DK1 imbalance price across the 2025 redesign, from the public Energinet series. The top panel shows the hourly price on a symmetric-log axis, and the bottom panel a fourteen-day rolling standard deviation of the hourly price change. Both widen at the go-live and stay wider, and the per-zone volatility ratio is reported in Table 1. Hourly aggregation smooths the sharpest fifteen-minute excursions, so the visible expansion understates what a bidder settling every fifteen minutes now faces.

seasonal-naive base produces, which we report as a cross-check rather than as a forecasting result. A simple base also keeps the eight zones comparable, since none receives zone-specific feature engineering.

4.2. Calibrators

A calibrator maps the score stream to an interval half-width q_t, which sets the symmetric prediction interval

$$[\ell_t,\, u_t] = [\,\hat{y}_t - q_t,\ \hat{y}_t + q_t\,], \tag{2}$$

evaluated against a target miscoverage $\alpha = 0.10$. Write $\widehat{Q}_{1-\alpha}(\cdot)$ for the empirical $(1-\alpha)$ quantile of a set of scores. Static split-conformal fixes the half-width once on the pre-break calibration set $\mathcal{C}$,

$$q_t \equiv \widehat{Q}_{1-\alpha}\big(\{s_i : i \in \mathcal{C}\}\big), \tag{3}$$

and never updates it, which is valid only while the post-break scores remain exchangeable with $\mathcal{C}$. Periodic split-conformal recomputes (3) on a trailing window, weekly and monthly, which lets the interval track a slow drift without any online machinery. Online adaptive conformal inference [15] instead holds the half-width construction fixed and adapts a working miscoverage level α_t after each step,

$$\alpha_{t+1} = \alpha_t + \gamma\big(\alpha - \mathbf{1}\{s_t > q_t\}\big), \qquad \gamma = 0.05, \tag{4}$$

reading the quantile at level $1-\alpha_t$ off a rolling score buffer. Aggregated adaptive conformal inference [17] runs several experts with different learning rates γ and combines them by online expert aggregation, which removes the choice of γ. The weighted, non-exchangeable calibrator [14] replaces the unweighted quantile in (3) with a weighted one whose weights decay into the past, which reweights rather than resets the calibration set. These five span the static, periodic, online, aggregated, and weighted responses to shift, which is the axis our result turns on. More elaborate time-series conformal methods that model the residual process or control adaptive regret [19, 22, 23] sit on the refreshed side of that axis, and we leave a sharpness comparison among them to future work.

4.3. Decision layer

A bidder commits to a position that settles at the realised imbalance price. Let c be the committed price, p the realised clearing price, v the committed volume in megawatts, and Δ the settlement-interval length in hours. A sell obligation and a buy obligation then earn

$$\pi_{\mathrm{sell}}(p) = (c-p)\,v\,\Delta, \qquad \pi_{\mathrm{buy}}(p) = (p-c)\,v\,\Delta. \tag{5}$$

Write π for whichever side the bidder holds. Each payoff in (5) is monotone in p, decreasing for a sell and increasing for a buy, so its worst value over a price interval is attained at an endpoint. This obligation payoff can lose money, which is what makes a loss floor meaningful. A pure limit order, by contrast, fills only on the favourable side of its limit and so can never realise a loss, which is why a loss floor stated on a limit order is vacuous.

The verifier gates an untrusted proposer, as Figure 2 sketches. Given a candidate bid and a conformal interval $[\ell_t, u_t]$ from (2), it computes the worst-case profit and accepts the bid only if that profit clears the operator loss floor τ,

$$\pi_{\min} = \min\{\pi(\ell_t),\, \pi(u_t)\}, \qquad \text{accept} \iff \pi_{\min} \geq \tau. \tag{6}$$

The coverage guarantee then transfers to the decision in one step, which we state to make the price of a coverage gap precise.

Lemma 1 (Coverage transfers to the loss floor)**.** *Let π be monotone in the clearing price p, and consider a bid accepted under* (6)*. If the interval covers, meaning $p \in [\ell_t, u_t]$, then $\pi(p) \geq \pi_{\min} \geq \tau$. Consequently, over accepted bids drawn from a calibrator whose adverse-side miscoverage is α, the loss floor is breached with probability at most α, and a proposer that commits at the accept boundary $\pi_{\min} = \tau$ attains a breach probability equal to the adverse-side miscoverage.*

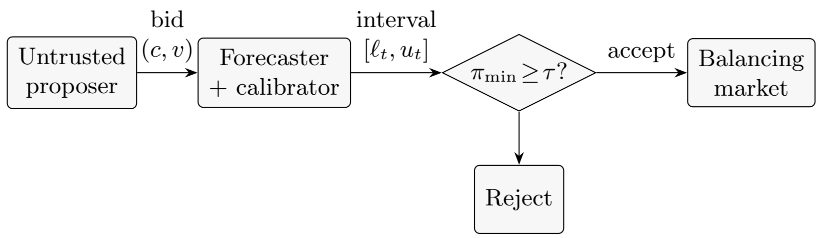

Figure 2: The verifier as a gate between an untrusted proposer and the market. The calibrator turns the point forecast into an interval $[\ell_t, u_t]$ at level $1 - \alpha$; the gate admits a bid only when its worst-case settlement over that interval clears the operator loss floor τ, per (6). By Lemma 1, an accepted bid breaches the floor with probability at most the calibrator's adverse-side miscoverage, so the interval's coverage is the only trusted quantity the guarantee rests on.

Proof. Monotonicity places the minimum of π over $[\ell_t, u_t]$ at an endpoint, so $\pi(p) \geq \min\{\pi(\ell_t), \pi(u_t)\} = \pi_{\min}$ for every p in the interval, and acceptance gives $\pi_{\min} \geq \tau$ by (6). A breach $\pi(p) < \tau$ therefore forces p outside the interval on the adverse side, an event of probability at most the adverse-side miscoverage. When the bid sits at the boundary $\pi_{\min} = \tau$, every adverse-side exceedance is itself a breach, so the two rates coincide. $\square$

Lemma 1 is a market specialisation of conformal risk control and conformal decision theory [33, 35], not a new guarantee. Its consequence is what matters here. A calibrator that under-covers by a given amount hands the operator a floor breach of the same order, so the coverage gap and the loss-floor breach are one quantity read in two units.

4.4. Metrics

Post-break coverage is reported against the target, with a ninety-five percent confidence interval from a moving-block bootstrap over the autocorrelated series, using a block length of forty-eight hours and two thousand resamples. Alongside coverage we report interval width and the Winkler interval score, which adds a miscoverage penalty to the width and so ranks a method by sharpness at fixed validity, since a method can otherwise buy coverage by widening. A zone's volatility change is the ratio of the post-break to the pre-break standard deviation of the daily price increment. The relationship between that volatility change and miscoverage across zones is summarized by a rank correlation.

5. Results

5.1. Coverage tracks the volatility of the redesign

Static split-conformal coverage after the redesign is not uniform across zones, and its variation is orderly. At a target of 0.90 it falls to 0.75 and 0.69 in the two Danish zones, where the redesign expanded imbalance-price volatility the most, and it holds between 0.85 and 0.98 in the calmer Nordic zones and between 0.90 and 0.92 in the continental controls (Table 2). The bootstrap intervals for the Danish zones sit well below the target, so the shortfall is not sampling noise. Figure 3 places every zone on one axis. The horizontal position is the ratio of post-redesign to pre-redesign price volatility, and the vertical position is the static miscoverage. The zones with the largest volatility increase sit far above the nominal line, the controls sit on it, and two placebo points from a non-reform date sit between the two. The rank correlation between volatility increase and miscoverage is 0.75 across the ten points, with a wide bootstrap interval that runs from 0.22 to 0.99, so the figure supports a monotone relationship rather than a precise slope. The online adaptive method, by contrast, holds between 0.89 and 0.90 in every zone with tight intervals. This is the behaviour conformal theory predicts for a fixed quantile once the scores stop being exchangeable [14], measured here on a real market rather than on synthetic drift.

The collapse is not specific to conformal calibration. A linear quantile-regression interval, the building block of the quantile-regression-averaging methods standard in probabilistic price forecasting [5], fit on the pre-redesign window on DK1 covers only 0.62 after the redesign, further below the target than the conformal interval, and refitting it on a trailing window recovers it to 0.81. What the break defeats is static calibration, not the conformal wrapper, and a distribution-free method recovers the target more completely than a re-estimated parametric one.

5.2. The coverage gap is a priced liability

The coverage gap does not stay a statistical quantity once a decision depends on it. We run the verifier of Section 4 against an untrusted proposer that commits at the accept boundary, the most exposed bid the rule will pass, and we record how often the realized settlement falls below the certified loss floor. Figure 4 shows the exposure in two units. Its left panel reports the breach rate under static and refreshed calibration. In the two Danish zones the static calibration breaches the floor at 0.13 and 0.16, above the promised limit of 0.10, so the guarantee the operator relies on does not hold there. The refreshed calibration holds the floor at about 0.06 in the same zones. In a calmer treated zone and in the control the two calibrations both keep the floor, because there the coverage gap is small. The breach rate follows the calibrator's adverse-side miscoverage, which is the transfer Lemma 1 makes precise.

Read as money, the same gap is a settlement shortfall below the floor, which the right panel of Figure 4 prices as the expected debt per megawatt-hour of committed volume for the same boundary-riding bidder. Under a stale calibration that debt is about 29 and 40 euros per megawatt-hour in the two Danish zones, against roughly 15 under a refreshed one. The continental control carries about 15 either way, because its volatility barely moved, so the excess a treated zone pays over the control, between 14 and 25 euros per megawatt-hour, is the coverage debt in the unit the operator settles in. The operator pays for stale calibration in realized losses rather than in a coverage statistic it might never inspect.

5.3. Refreshing the calibration repays the debt

The exposure is repaid by keeping the calibration current, and the choice among the working methods is a sharpness choice rather than a validity choice. On the worst-hit Danish zone a weekly refit of the static quantile lifts coverage from 0.75 back to 0.87, and a monthly refit reaches 0.85, so even a coarse schedule recovers most of the gap without any online machinery. The online and weighted calibrators recover it with no schedule at all. They differ in the width they pay to do so, which Figure 5 shows through the Winkler score across zones ordered by volatility. All the adaptive and weighted methods hold coverage, but the naive online method widens its intervals more than it needs to in the most volatile zones, where its score is the worst of the four. The weighted, non-exchangeable calibrator has the best score in five of the eight zones. We read this as a ranking rather than a tested per-zone dominance, and it restates on a real break the advantage that aggregation and reweighting are already known to give [17]. No single calibrator wins everywhere, and a comparison on coverage alone would rank the four as equal and miss this, since only the width-aware score separates them.

5.4. Pre-registration, controls, and placebo dates

The design guards against three ways the result could be an artifact rather than an effect. The first is selection. One zone, DK1, was used to choose the forecaster and the adaptive learning rate, so its result is reported as exploratory. Every other zone was analysed under an analysis plan, with the calibrators, the learning rate, the target level, and the pass rule fixed and

Table 2: Post-break coverage (target 0.90) with 95% moving-block-bootstrap confidence intervals and Winkler score per bidding zone. Static split-conformal coverage tracks the post-reform volatility increase and its confidence interval falls below the target in the treated Nordic zones, while the online adaptive method holds near the target in every zone. The Winkler score is marked per zone, best in bold and second best underlined, lower being sharper (↓); coverage is read against the 0.90 target through its confidence interval. The weighted variant has the best Winkler score in five of the eight zones.

		static split-CP		online adaptive		AgACI	nexCP
Zone	Role	Coverage [95% CI]	Wink.↓	Coverage [95% CI]	Wink.↓	Wink.↓	Wink.↓
DK1	trea	0.747 [0.71,0.78]	1405	0.891 [0.88,0.90]	1464	<u>1230</u>	**1217**
DK2	trea	0.693 [0.65,0.73]	1822	0.890 [0.88,0.90]	1532	<u>1452</u>	**1427**
SE3	trea	0.862 [0.83,0.89]	1165	0.894 [0.88,0.91]	1181	<u>1129</u>	**1126**
NO2	trea	0.980 [0.97,0.99]	194	0.897 [0.88,0.91]	**139**	<u>140</u>	<u>140</u>
FI	trea	0.846 [0.81,0.88]	1265	0.893 [0.88,0.90]	**1151**	<u>1211</u>	1221
NL	cont	0.912 [0.89,0.93]	907	0.898 [0.89,0.91]	910	<u>905</u>	**904**
BE	cont	0.900 [0.88,0.92]	648	0.896 [0.89,0.91]	**626**	**626**	<u>642</u>
AT	cont	0.915 [0.89,0.93]	<u>1180</u>	0.899 [0.89,0.91]	1307	1195	**1175**

committed to a public version-controlled record before the zone was run, so the confirmation zones cannot have been tuned to the outcome. The second is a common shock. If some Europe-wide event in early 2025 rather than the Nordic redesign drove the coverage loss, the continental control zones would show it too, and they do not. The Netherlands, Belgium, and Austria hold static coverage between 0.90 and 0.92 through the same window, with volatility ratios near one, so the loss is specific to the reformed markets. The third is the calendar. A placebo test repeats the analysis on the treated Danish zones at 4 March 2024, a date inside the old regime with no reform. There the static coverage is 0.83, mildly below target, with a volatility ratio near 1.3. This last result tempers the causal reading rather than confirming it. Static calibration in these zones drifts with ordinary volatility, so the 2025 redesign is best read as the largest instance of a volatility-driven effect rather than its sole cause, which is why we frame the relationship around volatility throughout.

6. Discussion

The central observation is that a coverage gap does not stay a statistical quantity once a decision depends on it. An operator who reads only the average coverage of a calibrator sees a number that drifts a little below target and looks tolerable. The same drift, passed through the safety rule, is a stream of accepted bids that settle below the loss floor the rule promised to

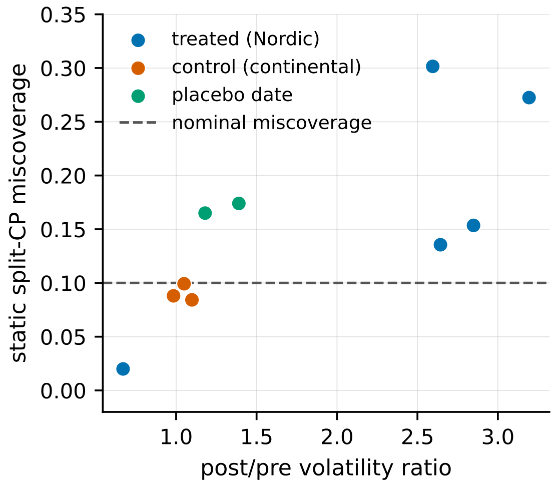

Figure 3: Static split-conformal miscoverage after the redesign against the post-to-pre volatility ratio, for eight bidding zones and two placebo dates. Treated Nordic zones with large volatility increases under-cover well beyond the nominal level, control zones stay near it, and placebo dates sit between the two. The relationship is what conformal theory predicts for a static quantile under a scale shift.

hold. That shortfall has a settlement value, and the operator pays it whether or not anyone is watching the coverage plot. This is why we call it debt. It accrues quietly while the calibration is stale, and it comes due in realized losses at exactly the rate the intervals under-cover.

The debt is not distributed evenly, and its size is legible in advance. It is large where the redesign expanded price volatility and small where volatility barely moved, which is what the coverage gradient records. A treated zone whose volatility roughly tripled carries a floor breach well above the promised limit, while a control zone at flat volatility carries none. Volatility change is therefore a leading indicator an operator can watch. A zone whose imbalance prices become more volatile is a zone whose static calibration is about to fall into debt, and the operator can act before the losses arrive rather than after.

Repaying the debt does not require the most elaborate method, and the choice among the working methods is a sharpness choice rather than a validity choice. A periodic refit recovers most of the lost coverage, and the online and weighted calibrators recover it without a schedule. All of them clear the loss floor. They differ in how wide their intervals grow to do so, and this difference is invisible to a coverage audit and visible only to a width-aware score. The naive online method holds coverage in the most volatile

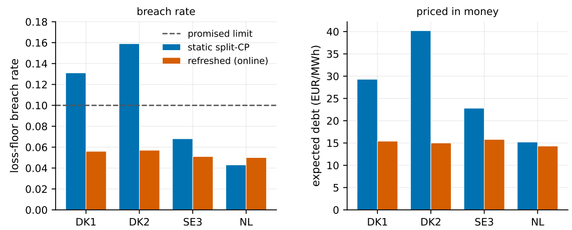

Figure 4: The coverage gap as a decision liability, in two units, for an untrusted bidder that commits at the safety rule's accept boundary. Left, the rate at which a certified worst-case loss floor is breached under static and refreshed calibration. The static calibration breaches the promised limit in the two highest-volatility zones, where refreshing holds it. Right, the same exposure priced as the expected settlement shortfall below the floor per megawatt-hour of committed volume. The treated Danish zones carry the larger euro debt under static calibration, refreshing removes the excess, and the continental control carries the same small debt either way.

zones by widening its intervals more than it needs to, which a bidder pays for in accepted volume, while the weighted calibrator holds the same coverage with tighter intervals in most zones. Practical guidance follows from this. Keep the calibration current, prefer a weighted or aggregated calibrator over a single-rate online one as the default, and judge the choice on a width-aware score rather than on coverage alone.

The result sits deliberately close to known theory, and the honesty of that placement is part of the contribution. Our coverage gradient is what the non-exchangeable bound of Barber et al. [14] predicts for a static quantile under a growing shift, and we do not present it as a new phenomenon. The advantage of the aggregated and weighted calibrators over the single-rate online one restates the finding of Zaffran et al. [17] on a new market rather than discovering it. Risk-controlling prediction sets, conformal risk control, and conformal decision theory own the step from coverage to a controlled decision loss [32, 33, 35]. What those results do not do is measure the gap on a real market redesign, price it at the level of the bid, and show the certified floor breaking at the miscoverage rate when the calibration is stale. The general theory is stated where coverage holds, and the contribution here is to watch what the decision does where coverage fails, on data a market actually produced, with a reusable way to repeat the measurement.

One modelling choice does the load-bearing work, and it is worth stating on its own because it is easy to get wrong. Whether the debt is even visible depends on the settlement payoff the verifier assumes. A limit order, the natural first model and the one a bidding study reaches for, fills only on the

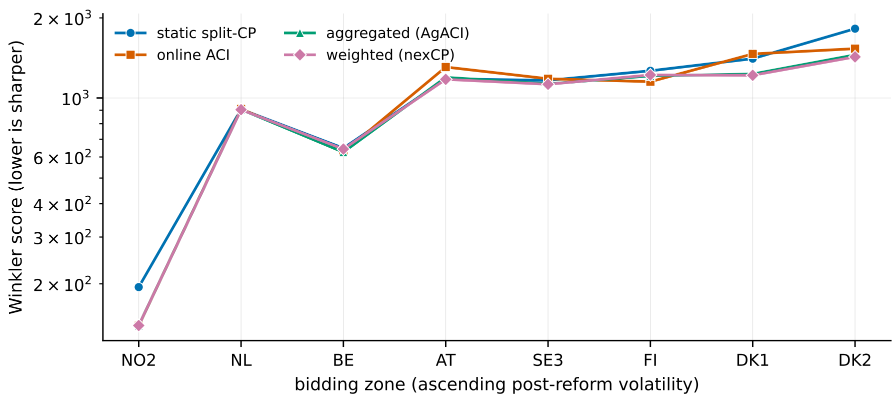

Figure 5: Winkler score per calibrator across zones ordered by post-reform volatility, lower is sharper. The adaptive and weighted methods hold coverage everywhere, the weighted variant is sharpest in most zones, and the naive online method over-widens in the most volatile zones.

favourable side of its limit and so can never realise a loss, which makes any loss floor stated on it vacuous and hides the exposure entirely. Only a loss-capable payoff that is monotone in the clearing price, such as the obligation settlement a balancing participant actually faces, turns the coverage gap into a number an operator can lose. The lesson is that a decision-level safety claim is only as meaningful as the payoff it is tested against, and the audit should use the payoff the market imposes rather than the one that flatters the guarantee.

This setting is not a one-off. The European balancing markets are converging on shared activation platforms and on fifteen-minute settlement, and the day-ahead market is moving to fifteen-minute resolution next, so reform-induced volatility shifts are a recurring feature of this decade rather than a single event. A bidder that trusts a static calibration through any of these transitions carries coverage debt it cannot see on a coverage plot. The benchmark we release lets an operator measure that exposure on its own zone and its own settlement rules before committing capital to the guarantee.

7. Limitations and future work

Four limitations bound the claims. The base forecaster is held simple so that the results speak to the calibration wrapper rather than to forecast quality, and a stronger forecaster is checked on one zone only, so the numbers are calibration-behaviour numbers and not a forecasting benchmark.

The volatility gradient rests on eight zones and two placebo dates, and its rank correlation carries a wide confidence interval, so it is evidence for a relationship rather than a precise law, and more zones and dates would tighten it. Causal attribution is limited, because a placebo date in the same zones shows a mild coverage loss without any reform, so the reform is best read as the largest instance of a volatility-driven effect rather than its sole cause. The loss floor is defined under an obligation-settlement payoff that can lose money, which replaces a limit-order payoff under which no bid can lose and the floor is vacuous, and this modelling choice should be matched to the settlement rules of a target market. A second, feature-side redesign of the day-ahead market is left for future work.

8. Conclusion

A distribution-free prediction interval is only as safe as the calibration behind it, and a real market redesign is a stress test of that safety. Across eight European bidding zones spanning the 2025 Nordic balancing-market redesign, a static calibration loses coverage in proportion to the reform-induced volatility increase, and a bidder's certified worst-case loss floor is then breached at the same rate the intervals under-cover. The coverage gap is a priced liability. Keeping the calibration current, by a periodic refit or an online update, repays it, and a weighted calibrator repays it with the least width. The mechanism is expected from conformal theory, so the contribution is the measurement on a real market at the level of the decision, together with a benchmark that lets an operator price the exposure on its own data before it trusts the guarantee.

Data availability

Raw imbalance prices for the Danish zones are available from Energinet Energi Data Service under CC-BY-4.0 and are included with the benchmark. Prices for the other zones are retrieved from the ENTSO-E Transparency Platform with a personal token, and the benchmark ships download scripts rather than redistributing that data. Code and download scripts are released as the accompanying benchmark.

CRediT author contributions

Phongsakon Mark Konrad: Conceptualization, Methodology, Software, Formal analysis, Data curation, Visualization, Writing – original draft.

Tim Lukas Adam: Conceptualization, Methodology, Software, Validation, Investigation, Writing – review & editing. **Serkan Ayvaz:** Conceptualization, Supervision, Project administration, Writing – review & editing. All authors read and approved the final manuscript.

Declaration of competing interests

The authors declare no competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Declaration of generative AI use

During the preparation of this work the authors used a large language model (Anthropic Claude) to assist with language editing, wording, and LaTeX formatting of the manuscript. The authors designed the study, implemented the benchmark, ran and interpreted all experiments, and drew all conclusions. After using this tool the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Prior work

This paper extends the authors' BSc thesis at the University of Southern Denmark with Danfoss A/S, which introduces the conformal verifier and reports a single-zone coverage collapse through the 2025 redesign on a model-selection window. The present paper differs in four ways that change the conclusions. It moves from one zone to eight with a pre-registered confirmation design, control markets, and placebo dates, which shows the single-zone collapse does not generalise and reframes the finding as a volatility-driven coverage gap. It replaces the thesis's limit-order safety model, under which no bid can lose money and the loss floor is vacuous, with an obligation-settlement payoff that makes the floor testable. It prices the coverage gap through the breach identity, which the thesis does not state. It releases a reusable multi-zone benchmark.

References

[1] Nordic Balancing Model, Confirmation of mFRR EAM go-live 4 march 2025, `https://nordicbalancingmodel.net`, 2025. Nordic transmission system operators.

[2] R. Weron, Electricity price forecasting: A review of the state-of-the-art with a look into the future, International Journal of Forecasting 30 (2014) 1030–1081. doi:10.1016/j.ijforecast.2014.08.008.

[3] J. Lago, G. Marcjasz, B. De Schutter, R. Weron, Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark, Applied Energy 293 (2021) 116983.

[4] J. Nowotarski, R. Weron, Recent advances in electricity price forecasting: A review of probabilistic forecasting, Renewable and Sustainable Energy Reviews 81 (2018) 1548–1568. doi:10.1016/j.rser.2017.05.234.

[5] J. Nowotarski, R. Weron, Computing electricity spot price prediction intervals using quantile regression and forecast averaging, Computational Statistics 30 (2015) 791–803. doi:10.1007/s00180-014-0523-0.

[6] B. Uniejewski, R. Weron, Regularized quantile regression averaging for probabilistic electricity price forecasting, Energy Economics 95 (2021) 105121. doi:10.1016/j.eneco.2021.105121.

[7] G. Marcjasz, M. Narajewski, R. Weron, F. Ziel, Distributional neural networks for electricity price forecasting, Energy Economics 125 (2023) 106843. doi:10.1016/j.eneco.2023.106843.

[8] Y. Romano, E. Patterson, E. Candès, Conformalized quantile regression, in: Advances in Neural Information Processing Systems, volume 32, 2019.

[9] V. Vovk, A. Gammerman, G. Shafer, Algorithmic Learning in a Random World, Springer, New York, 2005. doi:10.1007/b106715.

[10] G. Shafer, V. Vovk, A tutorial on conformal prediction, Journal of Machine Learning Research 9 (2008) 371–421.

[11] J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, L. Wasserman, Distribution-free predictive inference for regression, Journal of the American Statistical Association 113 (2018) 1094–1111. doi:10.1080/01621459.2017.1307116.

[12] A. N. Angelopoulos, S. Bates, A gentle introduction to conformal prediction and distribution-free uncertainty quantification, Foundations and Trends in Machine Learning 16 (2023) 494–591.

[13] R. J. Tibshirani, R. Foygel Barber, E. J. Candès, A. Ramdas, Conformal prediction under covariate shift, in: Advances in Neural Information Processing Systems, volume 32, 2019.

[14] R. F. Barber, E. J. Candès, A. Ramdas, R. J. Tibshirani, Conformal prediction beyond exchangeability, The Annals of Statistics 51 (2023) 816–845.

[15] I. Gibbs, E. Candès, Adaptive conformal inference under distribution shift, Advances in Neural Information Processing Systems 34 (2021) 1660–1672.

[16] I. Gibbs, E. J. Candès, Conformal inference for online prediction with arbitrary distribution shifts, Journal of Machine Learning Research 25 (2024) 1–36.

[17] M. Zaffran, O. Féron, Y. Goude, J. Josse, A. Dieuleveut, Adaptive conformal predictions for time series, in: International Conference on Machine Learning, 2022, pp. 25834–25866.

[18] C. Xu, Y. Xie, Conformal prediction interval for dynamic time-series, in: International Conference on Machine Learning, 2021, pp. 11559–11569.

[19] C. Xu, Y. Xie, Sequential predictive conformal inference for time series, in: International Conference on Machine Learning, volume 202 of *PMLR*, 2023, pp. 38707–38727.

[20] A. Auer, M. Gauch, D. Klotz, S. Hochreiter, Conformal prediction for time series with modern Hopfield networks, in: Advances in Neural Information Processing Systems, volume 36, 2023.

[21] K. Stankevičiūtė, A. M. Alaa, M. van der Schaar, Conformal time-series forecasting, in: Advances in Neural Information Processing Systems, volume 34, 2021.

[22] A. Bhatnagar, H. Wang, C. Xiong, Y. Bai, Improved online conformal prediction via strongly adaptive online learning, in: International Conference on Machine Learning, volume 202 of *PMLR*, 2023, pp. 2337–2363.

[23] A. N. Angelopoulos, E. J. Candès, R. J. Tibshirani, Conformal PID control for time series prediction, in: Advances in Neural Information Processing Systems, volume 36, 2023, pp. 23047–23074.

[24] C. Kath, F. Ziel, Conformal prediction interval estimation and applications to day-ahead and intraday power markets, International Journal of Forecasting 37 (2021) 777–799. doi:10.1016/j.ijforecast.2020.09.006.

[25] J. Jonkers, D. Nieves Avendano, G. Van Wallendael, S. Van Hoecke, A novel day-ahead regional and probabilistic wind power forecasting framework using deep CNNs and conformalized regression forests, Applied Energy 361 (2024) 122900. doi:10.1016/j.apenergy.2024.122900.

[26] W. Wang, B. Feng, G. Huang, C. Guo, W. Liao, Z. Chen, Conformal asymmetric multi-quantile generative transformer for day-ahead wind power interval prediction, Applied Energy 333 (2023) 120634. doi:10.1016/j.apenergy.2022.120634.

[27] V. Jensen, F. M. Bianchi, S. N. Anfinsen, Ensemble conformalized quantile regression for probabilistic time series forecasting, IEEE Transactions on Neural Networks and Learning Systems 35 (2024) 9014–9025. doi:10.1109/TNNLS.2022.3217694.

[28] R. Yu, et al., Deep learning for electricity price forecasting: A review of day-ahead, intraday, and balancing electricity markets, arXiv preprint arXiv:2602.10071 (2026).

[29] M. Narajewski, Probabilistic forecasting of German electricity imbalance prices, Energies 15 (2022) 4976. doi:10.3390/en15144976.

[30] S. Deng, J. Inekwe, V. Smirnov, A. Wait, C. Wang, Seasonality in deep learning forecasts of electricity imbalance prices, Energy Economics 137 (2024) 107770. doi:10.1016/j.eneco.2024.107770.

[31] C. O'Connor, M. Bahloul, R. Rossi, S. Prestwich, A. Visentin, Conformal prediction for electricity price forecasting in the day-ahead and real-time balancing market, arXiv preprint arXiv:2502.04935 (2025).

[32] S. Bates, A. Angelopoulos, L. Lei, J. Malik, M. I. Jordan, Distribution-free, risk-controlling prediction sets, Journal of the ACM 68 (2021) 1–34. doi:10.1145/3478535.

[33] A. N. Angelopoulos, S. Bates, A. Fisch, L. Lei, T. Schuster, Conformal risk control, in: International Conference on Learning Representations, 2024.

[34] S. Feldman, L. Ringel, S. Bates, Y. Romano, Achieving risk control in online learning settings, Transactions on Machine Learning Research (2023).

[35] J. Lekeufack, A. N. Angelopoulos, A. Bajcsy, M. I. Jordan, J. Malik, Conformal decision theory: Safe autonomous decisions from imperfect predictions, in: IEEE International Conference on Robotics and Automation (ICRA), 2024, pp. 11668–11675.

[36] T. Carrière, G. Kariniotakis, An integrated approach for value-oriented energy forecasting and data-driven decision-making applied to renewable energy trading, IEEE Transactions on Smart Grid 10 (2019) 6933–6944. doi:10.1109/TSG.2019.2914379.

[37] A. Stratigakos, S. Pineda, J. M. Morales, Decision-focused linear pooling for probabilistic forecast combination, International Journal of Forecasting 41 (2025) 1112–1125. doi:10.1016/j.ijforecast.2024.11.006.

[38] S. Clò, E. Fumagalli, The effect of price regulation on energy imbalances: A difference in differences design, Energy Economics 81 (2019) 754–764. doi:10.1016/j.eneco.2019.05.008.

[39] R. A. C. van der Veen, R. A. Hakvoort, The electricity balancing market: Exploring the design challenge, Utilities Policy 43 (2016) 186–194. doi:10.1016/j.jup.2016.10.008.

[40] Energinet, Energi data service, https://www.energidataservice.dk, 2026. Danish TSO open data, CC-BY-4.0.

[41] ENTSO-E, ENTSO-E transparency platform, https://transparency.entsoe.eu, 2026.