Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Continuation-specific reliability assessment of receding-horizon wind–battery–reserve dispatch

Xin He[1]
Shihezi University
Yingshuo Qi[1]
Shihezi University
Xinrong Xie[1]
Shihezi University
Yongpeng Tong[1]
tongyp@shzu.edu.cn

Shihezi University

Article

Keywords: model predictive control, wind power, battery storage, reliability assessment, policy evaluation, information timing

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11527625/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Continuation-specific reliability assessment of receding-horizon wind–battery–reserve dispatch

Xin He[1]†, Yingshuo Qi[1]†, Xinrong Xie[1]† and Yongpeng Tong[1]*

[1] Department of Mathematics, School of Science, Shihezi University, Shihezi 832003, Xinjiang, China (Xin He: Mathematics and Applied Mathematics; Yingshuo Qi and Xinrong Xie: Information and Computing Science)

† These authors contributed equally to this work.

* *Correspondence: tongyp@shzu.edu.cn*

Abstract

Reliability estimates for receding-horizon control are frequently evaluated under a planned action sequence, although only its first command is executed before replanning. We examine how this difference changes the shortfall event attributed to an initial wind–battery–reserve dispatch decision. In paired replays, the pre-action state, first command, realised wind and synthetic delivery obligation are held fixed while the subsequent control policy varies. Across 9,010 UK development origins and 2,586 Australian transfer origins, the deployed-minus-frozen-FULL shortfall differences for deterministic model predictive control were −6.25 and −12.34 percentage points; differences relative to its own fixed plan were +5.58 and +4.18 points. An exploratory 14-day UK extension (1,985 paired origins) yielded −0.35 and +6.80 points, respectively; the near-zero FULL contrast changed sign under origin thinning. Identical probability forecasts also produced different Brier scores when evaluated against deployed rather than planned events. Fixed-endpoint replanning and dependence-aware sensitivity analyses further delineated the sources and stability of these differences. Reliability comparisons therefore require an explicitly specified continuation and scoring target; neither the direction of the event shift nor a cross-target score improvement is universal.

Keywords: model predictive control; wind power; battery storage; reliability assessment; policy evaluation; information timing

Introduction

Wind-power dispatch must reconcile uncertain generation with limited operational flexibility. Batteries and directional reserve can offset deviations between scheduled and realised output, but each intervention changes the energy, headroom and reserve budget available later in the delivery window. Scheduling decisions are therefore coupled across time, motivating stochastic model predictive control (MPC) and scenario-based storage operation rather than independent interval-by-interval decisions [1,2].

Receding-horizon implementation introduces a separate complication. MPC computes a sequence of future actions but implements only the first before updating the system state and solving a new optimisation problem. Closed-loop stochastic MPC accordingly studies recursive feasibility and risk over the executed policy, rather than only over a nominal open-loop trajectory [3,4]. Predictive safety filters likewise assess an implementable future policy [5]. Consequently, a reliability claim is incomplete unless its assumed continuation is specified.

Related uncertainty-quantification methods increasingly connect probabilistic guarantees to decisions and policies. Conformal decision theory and conformal risk control target decision-relevant losses [6,7], including under non-exchangeability [8]. Sequence-level risk control has been studied for open- and closed-loop MPC [9], while shrinking-horizon conformal MPC considers feasibility as information arrives [10]. Off-policy conformal prediction addresses policy changes between observation and evaluation [11]; Conformal Policy Control regulates policy changes relative to a reference under stated assumptions [12]. Distributionally robust screening has also been considered for closed-loop voltage control [13]. These approaches establish important guarantees, but do not remove the need to identify the event against which an empirical risk score is computed.

A policy-specific guarantee and a retrospective deployment audit address distinct questions. The former states a property under an identified policy and assumptions; the latter examines whether the event used to evaluate a decision corresponds to the event produced by the policy subsequently executed. We address this empirical alignment problem using a transparent dispatch replay with explicitly defined information availability.

In practice, a risk screen may evaluate an initial command followed by a frozen backup sequence or by the remainder of the plan optimised at the current origin. Deployment then proceeds through repeated forecasts and optimisations. Evaluating the original probability against an outcome produced by this different continuation conflates forecast error with a change in the event being predicted. Replanning may increase or decrease the probability of shortfall, depending on the controller, reference continuation and operating conditions; its sign cannot be assumed.

We develop a paired, continuation-specific audit that records the shortfall label under each named continuation, both directions of disagreement between paired labels and the probability score against the event for which that probability was constructed. Each comparison shares the same pre-action state, exact initial command, realised wind trajectory and delivery obligation (Fig. 1). An additional replay that retains the original endpoint distinguishes reoptimisation within the original task from differences that emerge when the rolling planning horizon extends. The audit isolates changes in the event target without changing the initial decision.

We first analyse an historically exposed UK development series and a frozen 18-day Australian transfer stress test. We then add a separate, exploratory 14-day UK temporal extension in 2022, using the unchanged UK-fitted predictor, residual scenarios and controller parameters (Table 1; Supplementary Tables S1 and S21). Across these cohorts, the comparison with the controller's own fixed plan is positive, whereas the contrast with a frozen FULL reference varies markedly in magnitude. Our contribution is an auditable distinction between continuation-induced event changes and probability-scoring changes, together with evidence of their dependence on the reference policy, observation window and statistical design.

Results

Continuation choice changes the event attributed to the same first command

The primary endpoint is a directional delivery shortfall: at least one interval in the assessed horizon exceeds the 0.01 per-unit (pu) tolerance in the shortfall direction. Each horizon spans 7–12 ten-minute intervals, comprising the remainder of the current clock hour and the following full hour. For every eligible origin, the exact first command is held fixed and the remaining window is replayed under either a named reference continuation or subsequent rolling deployment (Table 2).

In the UK development cohort, the full-resource point-forecast controller (F_FULL) had shortfall-event rates of 73.82% under its frozen-origin FULL continuation and 86.54% under rolling deployment, a +12.72-percentage-point (pp) difference. The scenario menu controller (F_BR0) instead yielded 77.46% and 69.61% (−7.85 pp), while deterministic MPC (DMPC) yielded 75.85% and 69.60% (−6.25 pp). Thus, the direction of the deployment–reference contrast differed across controllers even within the same data period (Table 3).

The reference policy also changed the direction of DMPC's estimated contrast. Relative to the frozen-origin FULL continuation, rolling deployment yielded 6.25 pp fewer shortfall events in the UK and 12.34 pp fewer in Australia; relative to DMPC's own origin-fixed plan, it yielded 5.58 and 4.18 pp more, respectively (Fig. 2a). The archived seven-day ordinary moving-block percentile ranges were [−12.82, −1.80], [−25.48, −8.00], [4.15, 6.73] and [2.55, 5.03] pp in the corresponding cohort–reference comparisons. These ranges are exploratory conditional summaries rather than validated population confidence intervals, as the dependence diagnostics below make clear.

The opposing directions of paired disagreement were larger than their net differences. Against FULL in the UK, DMPC produced 462 deployed-only and 1,025 reference-only failures: 16.50% of labels disagreed despite a net −6.25 pp contrast. Against its own fixed plan, the counts were 552 and 49. Corresponding Australian counts were 84 and 403 against FULL, and 114 and 6 against the own plan (Fig. 2b; Supplementary Table S2). Reporting both directions avoids concealing substantial disagreement through cancellation.

These differences were neither controller-invariant nor interpretable as observed field failure rates. For example, F_FULL's Australian deployment–FULL contrast was +3.29 pp, with an archived seven-day exploratory range crossing zero [−5.26, 6.34] pp. The high absolute event rates reflect the deliberately stressful synthetic obligation and the chosen 0.01 pu threshold, not the commercial reliability of either wind farm.

Changing the event target can improve an apparent probability score

For the origin-fixed DMPC plan, the mean scenario probabilities, target event rates and target-matched Brier scores were 66.52%, 64.02% and 0.1866 in the UK, and 65.67%, 53.09% and 0.1844 in Australia. For the frozen FULL continuation, the corresponding values were 77.03%, 75.85% and 0.1740 in the UK, and 78.08%, 69.61% and 0.2052 in Australia (Fig. 2c; Supplementary Table S3). Neither these aggregate comparisons nor their Brier scores establish calibration, and the probabilities were not designed to predict the rolling-deployment event.

Replacing the own-plan labels with deployment labels, while holding the probability vector unchanged, reduced the Brier score from 0.1866 to 0.1669 in the UK and from 0.1844 to 0.1729 in Australia. The analogous substitution for frozen-FULL probabilities increased the scores from 0.1740 to 0.2075 and from 0.2052 to 0.2652, respectively (Fig. 2d; Supplementary Table S18). All four differences satisfied the binary-score identity in equation (4), with a maximum numerical residual of 1.39×10^{-17}. A lower cross-target score cannot establish improved prediction of the original event because the event itself has changed; moreover, the score difference depends on the predicted probabilities at discordant origins, not merely on the aggregate event-rate gap.

Fixing the original endpoint separates two continuation contrasts

We introduced a fixed-original-endpoint continuation, S, at each archived DMPC origin. After executing the identical saved first command, S reoptimises as new forecasts arrive and its own resource state evolves, but

retains the original endpoint and outstanding obligation. Unlike rolling deployment D, it does not append the next delivery hour after the hour boundary. All objective coefficients and the terminal energy target are unchanged.

Shortfall-event rates for the UK cohort were 64.02% under the original fixed plan P, 67.57% under S and 69.60% under D; the Australian rates were 53.09%, 56.11% and 57.27%, respectively (Fig. 4a; Supplementary Table S19). The sequential contrasts S−P and D−S were +3.55 and +2.03 pp in the UK, and +3.02 and +1.16 pp in Australia. D−S contained 183 deployed-only failures in the UK and 30 in Australia, with no S-only failures in these cohorts. Holding the endpoint fixed therefore did not eliminate the discrepancy between the original plan and deployment. Although the sequential contrasts sum exactly to D−P, their ordering does not constitute an additive decomposition into independent causal effects.

All 40,581 pre-rollover command comparisons between S and D agreed exactly, locating their divergence after rolling deployment could extend the planning endpoint. Figure 4c,d presents the earliest qualifying D-only shortfall in each cohort, chosen by a fixed chronological rule. In these illustrative cases, D introduced charging after rollover whereas S did not, and the resulting realised trajectory crossed the shortfall threshold. Full command, energy and reserve paths are provided in Supplementary Figs. S4 and S5. These examples demonstrate a possible resource-allocation pathway, not its frequency or a general advantage of fixed-endpoint operation.

Operational frequency, severity and resource use do not move together

Paired shortfall labels do not, by themselves, rank controllers by engineering merit (Table 4). Under actual deployment, UK complete-hour shortfall rates were 78.84% for F_FULL, 58.02% for F_BR0 and 58.68% for DMPC; cumulative excess shortfall energies were 35.20, 41.28 and 39.04 pu h, respectively. In Australia, the corresponding frequencies were 60.65%, 47.45% and 47.22%, whereas excess shortfall energies were 9.70, 14.78 and 13.76 pu h. Controllers with fewer hours containing a shortfall could therefore still incur a larger total shortfall.

Resource utilisation further qualified the comparison. In the UK, battery throughput and directional reserve use were 90.84/10.15 pu h for F_FULL, 10.64/25.74 pu h for F_BR0 and 19.74/28.33 pu h for DMPC; corresponding Australian totals were 27.93/2.59, 2.36/7.03 and 4.01/8.04 pu h. These are normalised physical quantities, not market costs. Figure 3 therefore displays shortfall frequency alongside mean excess shortfall power rather than presenting a single composite controller ranking (Supplementary Table S9).

Exploratory temporal transfer to UK observations from 2022

To assess temporal transportability without refitting, we applied the 2021-frozen predictor, residual bank and dispatch parameters to 1–14 January 2022 at Kelmarsh. A continuous, quality-screened source segment supplied 1,991 chronological ten-minute execution origins, 331 complete hours and 1,985 complete paired assessment origins; none of the three controllers required solver fallback. The 2022 interval was chosen at the beginning of the calendar year, without selection based on its controller outcomes, but its data had been used for earlier quality assurance. Accordingly, this is an exploratory within-site temporal extension, not a blinded annual test (Table 1; Supplementary Table S21).

For DMPC, the deployed shortfall-event rate was 76.52%, compared with 76.88% under the frozen FULL continuation and 69.72% under the origin-fixed plan. Thus the deployment–FULL contrast attenuated to −0.35 pp, whereas the deployment–own-plan contrast was +6.80 pp (Table 3). Against FULL, 124 origins failed only under deployment and 131 only under the reference, giving 12.85% discordance despite a nearly zero mean difference. Against the own plan, the corresponding counts were 139 and 4. F_FULL had a deployment–FULL contrast of +16.98 pp and F_BR0 of −1.46 pp. These results confirm that neither magnitude nor direction of the continuation contrast is controller- or period-invariant.

The 2022 data also reproduced the target-switching score diagnostic. The same own-plan probability vector yielded Brier scores of 0.1701 against its intended fixed-plan label and 0.1491 against deployment; the FULL-target vector yielded 0.1658 and 0.1807, respectively (Supplementary Table S22). The own-plan score reduction therefore does not establish improved calibration. Under chronological execution, complete-hour shortfall frequencies were 82.78%, 62.84% and 64.05% for F_FULL, F_BR0 and DMPC (Table 4). As

in the original cohorts, event frequency and excess shortfall energy did not provide identical controller rankings.

Point-estimate directions persist under the tested perturbations

Post-freeze event relabelling changed the scoring rule while leaving controller actions, realised trajectories, obligations and the physical executor unchanged. At thresholds from 0.005 to 0.03 pu, DMPC's deployment–FULL contrasts ranged from −6.36 to −4.66 pp in the UK and −12.34 to −4.76 pp in Australia; deployment–own-plan contrasts ranged from +5.22 to +5.79 and +2.78 to +4.99 pp, respectively. The all-origin directions were also retained when outcomes were scored over the first 30 min, first 60 min or full 70–120 min window (Supplementary Tables S10 and S11; Supplementary Fig. S1). These checks concern event definition, not redesigned actuators or controllers.

Complete DMPC reruns at synthetic-obligation multipliers of 0.90, 1.00 and 1.10 retained negative deployment–FULL and positive deployment–own-plan contrasts in both original cohorts. The FULL contrasts were −11.74, −6.25 and −4.36 pp in the UK and −13.73, −12.34 and −11.25 pp in Australia; corresponding own-plan contrasts were +2.78, +5.58 and +5.84 pp, and +2.90, +4.18 and +4.99 pp (Supplementary Tables S12 and S13). These tests alter obligation magnitude, not the issuing mechanism or contractual validity of that obligation.

A separate synthetic-obligation sensitivity replaced the single-observation persistence rule with the mean of six power observations labelled 130–80 min before delivery, retaining the h−60 min issuance time and the 20-min availability gate. Full chronological DMPC reruns over the original cohorts produced deployment–FULL and deployment–own-plan contrasts of −9.27 and +4.31 pp in the UK, and −11.72 and +3.79 pp in Australia (Supplementary Table S20). All constituent observations were available by issuance. This alternative construction supports sensitivity assessment but does not establish agreement with a historical commercial commitment.

The primary information convention makes a ten-minute-labelled observation available 20 min after its timestamp. Rebuilding the original cohorts at assumed delays of 10, 30 and 40 min preserved negative deployment–FULL and positive deployment–own-plan point contrasts: −8.46 to −4.88 and +3.80 to +7.49 pp in the UK, and −12.72 to −11.21 and +3.94 to +5.53 pp in Australia. Because changing the gate also affects eligible origins, these are sensitivity analyses rather than a monotonic latency-response experiment (Supplementary Table S5).

Seven one-factor battery and reserve configurations likewise retained the all-origin direction in the original cohorts. Deployment–FULL contrasts ranged from −6.39 to −5.99 pp in the UK and −12.53 to −12.18 pp in Australia; corresponding own-plan contrasts ranged from +5.32 to +5.84 and +3.63 to +4.29 pp (Supplementary Table S6). These configurations were not selected to maximise the observed contrasts.

Origin selection and boundary weighting qualify statistical stability

Origin selection materially affected the statistical evidence. Reproduction of all 12 archived seven-day thinning analyses showed that the UK deployment–FULL contrast changed from −6.25 pp across 9,010 origins to −5.20 pp across 1,500 hourly phase-zero origins and −3.72 pp across 752 non-overlapping origins. The respective ordinary moving-block ranges were [−12.82, −1.80], [−11.53, +0.07] and [−10.11, +2.39] pp. The latter two cross zero. Own-plan contrasts in the UK and both Australian reference contrasts retained their observed signs under these sampling schemes (Supplementary Table S14; Supplementary Fig. S2). Thinning changes the phase distribution and horizon composition as well as the effective sample size; these schemes are related but distinct descriptive estimands.

Ordinary moving-block resampling need not centre on the observed statistic in finite samples. For Australia's deployment–FULL comparison, the exact expected seven-day resampled estimate was −16.89 pp, compared with the observed −12.34 pp, a −4.55 pp shift attributable to boundary weighting and truncated final draws. A within-run circular sensitivity, which is centred at the empirical mean in expectation, yielded [−23.43, −3.98] pp versus the archived ordinary range [−25.48, −8.00] pp (Supplementary Tables S15 and S16). Circular wrapping introduces artificial endpoint adjacency and does not yield a validated population confidence interval. Correlations between paired-event differences remained 0.369 after 24 h and 0.149 after

seven days in the Australian FULL comparison (Supplementary Table S17; Supplementary Fig. S3), precluding an assumption of seven-day independence.

The 14-day UK ordinary analysis retained 16.84% of origins in short, fixed runs; its narrower resampling range therefore offers no stronger population-level evidence. Leave-one-day-out sensitivity retained the all-origin signs, with UK deployment–FULL contrasts from −6.58 to −5.21 pp and Australian contrasts from −13.80 to −10.97 pp, while own-plan contrasts remained positive. Daily effects were not uniformly signed (Supplementary Tables S4 and S7). Omitting a day or increasing Monte Carlo repetitions cannot create independent wind farms or independent weeks.

Additional design-based checks across three evaluated periods

The additional statistical audit recovered the original paired labels and repeated the same three origin-selection schemes on the UK 2021 development, Australian 2026 transfer and UK 2022 exploratory cohorts. For DMPC, deployment–FULL estimates under all, hourly and non-overlapping origins were −6.25, −5.20 and −3.72 pp (UK 2021); −12.34, −13.23 and −12.50 pp (Australia); and −0.35, +1.52 and +1.81 pp (UK 2022). The UK 2022 comparison therefore changed sign when overlapping windows were thinned. By contrast, deployment–own-plan estimates remained positive: +5.58/+5.27/+5.98 pp, +4.18/+5.57/+5.09 pp and +6.80/+8.18/+7.23 pp, respectively (Supplementary Table S23). Neither the selection schemes nor their sample counts can be treated as interchangeable independent replications.

For the 2022 extension, a 2,000-replicate within-run circular seven-day moving-block analysis yielded exploratory percentile ranges of [−5.19, +4.53] pp for deployment–FULL and [+4.84, +8.87] pp for deployment–own-plan. Additional block-length checks at one and three days, and parallel analyses of the original cohorts, are reported in Supplementary Table S24. The 14-day 2022 window is particularly short relative to the seven-day blocks. Circular endpoint wrapping, overlapping origins and previously inspected source data preclude interpreting these intervals as population-level confidence bounds or as prospective validation.

Timing and replay audits support the declared information set

The UK pipeline comprised 3,995 model-fitting origins and a subsequent, temporally separated bank of 4,020 residual-bank origins, both preceding the development evaluation. The six ten-minute phases each supplied 670 residual trajectories; at a new origin, 32 same-phase nearest-neighbour trajectories were selected using only features deemed available by that time. Input-perturbation tests confirmed that modifying not-yet-available observations did not change the planned commands in the tested cases. Independent physical replay reproduced every evaluated step and paired label in the two original cohorts, with no solver fallback.

The Australian source consisted of 30 daily AEMO archives spanning 30 August–28 September 2026, with the scoring window fixed to market dates 11–28 September. Its 2,592 ten-minute proxies were derived from pairs of five-minute SCADA snapshots. The source-file audit showed that the archived payload-creation timestamps preceded the assumed 10-min sensitivity gate for all evaluated proxies, by at least 262 s. This establishes a limited source-provenance check; it does not establish the receipt time of operator telemetry or the time when an external user could retrieve the file. The 20-min availability gate remains an explicit simulation assumption (Supplementary Table S8).

Discussion

The central result is a distinction between reliability attributed to a contemplated continuation and reliability observed after the controller is repeatedly reoptimised. With the initial state and command matched, fixed-plan, fixed-endpoint and rolling continuations can generate different shortfall labels. Recording both directions of discordance reveals changes that disappear in net averages, while target-matched and cross-target probability scores clarify whether a numerical score change reflects different predictions or a different event.

The fixed-original-endpoint experiment provides a more precise interpretation of the subsequent control sequence. Reoptimisation within the original task already increased the own-plan shortfall frequency in both

original cohorts; extending the rolling task yielded a further difference. Exact agreement of commands before rollover identifies when the two reoptimising policies began to diverge. These sequential contrasts cannot identify an isolated causal effect of updated forecasts, nor demonstrate that extending the horizon worsens performance over the longer task, because newly introduced obligations lie outside the original scoring window.

The method is an empirical audit rather than an alternative to policy-specific reliability guarantees. Work on sequence-level conformal risk control, shrinking-horizon MPC, off-policy prediction and conformal policy regulation establishes guarantees under stated policies and assumptions [9–12]; power-system screening has also addressed distributional and policy shifts [13]. Our contribution is different: an executable fixed-state, fixed-first-command comparison of event targets, paired discordance and probability scores in a wind–storage dispatch model. No theorem-level guarantee or superiority over those methods is claimed.

The probability results illustrate why target specification matters. A probability estimated for a fixed plan may receive a lower numerical Brier score when evaluated against outcomes generated by rolling deployment, despite remaining unchanged. The opposite change occurred for FULL-based probabilities. Equation (4) attributes these differences to the predicted probabilities at origins where the event labels disagree. Cross-target scoring is thus a diagnostic of target mismatch, not evidence of improved calibration. Learning and prospectively validating a probability predictor for the deployed policy remain separate research tasks.

Operational performance further cautions against equating event-frequency improvement with overall benefit. In both original cohorts, F_BR0 and DMPC failed in fewer complete hours than F_FULL but accumulated more excess shortfall energy, alongside different battery and reserve usage. The exploratory 2022 extension also displayed controller-specific frequency–severity and resource trade-offs (Table 4). Neither the paired-label contrasts nor the fixed-endpoint decomposition establishes lower dispatch cost, improved battery lifetime or superior field performance.

The added 2022 temporal extension materially qualifies the generality of the original comparison. Its deployment–FULL contrast for DMPC was only −0.35 pp, compared with −6.25 pp in the 2021 UK development period, and reversed sign under hourly or non-overlapping origin selection. This weakens any broad claim that rolling deployment systematically reduces shortfall relative to FULL. By contrast, deployment–own-plan contrasts remained positive across all three evaluated cohorts and the corresponding thinning designs, although the time periods and overlapping origins cannot support a universal policy-level conclusion. The extension was selected from previously quality-reviewed data and lasted only 14 days; it is an exploratory transfer, not a blinded annual validation.

Inference and engineering interpretation remain conditional on the replay model. UK development data had prior project exposure, the Australian transfer covers a single 18-day interval, and the 2022 UK extension is a short contiguous segment from the same site. Overlapping windows, temporally dependent errors, origin-thinning changes and resampling boundary effects limit claims of population-level uncertainty. Delivery obligations are synthetic; the observation gate is assumed; Australian power is a proxy from instantaneous SCADA snapshots; and curtailment, battery and reserve dynamics are simplified. Absolute shortfall rates should not be extrapolated to field reliability, market profitability or equipment degradation. Broader inference requires additional independent periods and sites, operational commitments and verified timing and actuator data.

A practical audit should therefore identify the continuation underlying each probability and event, preserve the common initial decision when comparing policies, and report deployed-only and reference-only failures in addition to net differences. Where the planning endpoint changes, a fixed-endpoint replay helps locate the source of the discrepancy. Probability scores should first be interpreted against their stated targets; cross-target results must be labelled explicitly. The combined analyses reveal a reproducible, context-dependent assessment issue in wind–battery–reserve scheduling rather than a universally beneficial dispatch strategy.

Methods

Study design and evidence hierarchy

We conducted a chronological UK development analysis, a frozen Australian transfer stress test and a subsequent exploratory UK temporal extension. The development source was an aggregate of six Senvion MM92 turbines from the publicly released Kelmarsh dataset [14]. Because the 2021 series had been used in earlier project iterations, all UK 2021 results are designated developmental. The predefined time-ordered pipeline allocated 28 days to model fitting, the following 28 days to residual-bank construction and later eligible 2021 observations to development assessment. 'Time-causal' here denotes respect for the declared information clock, not causal identification. Evaluation began at 16:00 UTC on 7 October 2021 and included 9,028 executed ten-minute intervals in three contiguous episodes, 1,503 complete hours and 9,010 complete paired origins.

The external stress test used the public AEMO DISPATCH_UNIT_SCADA records for COOPGWF1, Coopers Gap Wind Farm. AEMO defines SCADAVALUE as instantaneous megawatts at the start of a five-minute dispatch interval [15]. We used 453 MW, the operator-reported nameplate capacity [16], as a frozen normalisation constant; AEMO's enhanced locational information separately lists 440 MW as maximum capacity [17]. Neither figure is treated as a validated time-varying export limit. After an exposure audit identified earlier assets for 1 and 8–10 September, market dates 11–28 September 2026 were fixed before computing the tested controller outcomes. The resulting 2,592 ten-minute proxy observations comprised 432 complete hours and 2,586 paired origins. The UK predictor, scaler, residual bank, resource constraints and controllers were transferred without local refitting or outcome-based tuning. This is a documented frozen stress test rather than a pristine multi-site blinded holdout.

Exploratory UK 2022 temporal extension

We used an archived, quality-screened aggregate from the same six-turbine Kelmarsh source to evaluate the first 14 complete calendar days of 2022 (UTC). This period was chosen by its calendar position rather than by observed dispatch outcomes, within a contiguous segment passing the archived data-quality rules. The source was normalised by the fixed 12.3 MW six-turbine capacity. The 2021 ridge coefficients, training scaler, same-phase residual scenario bank, 20-min information gate, synthetic commitment rule, battery and reserve limits, controller objective and initial physical state were held unchanged. Controller execution proceeded chronologically through the uninterrupted segment; the battery state was initialised once at the beginning, not independently at each assessment origin. After constructing available-history features, 1,991 eligible steps were executed and 1,985 complete paired origins evaluated. The extension covers 331 complete hours. No 2022 observation was used to refit or choose controller parameters. Archived quality review of the 2022 source, however, means this cohort is not strictly blinded (Supplementary Methods; Supplementary Data 4).

The source-processing adapter was restricted to dataset paths and validation-date checks; the optimisation implementation and design configuration retained the same file hashes as the archived baseline. Historical observations required for warm-up were not used before the declared availability time. The 2022 labels were generated with the unchanged physical executor and paired replay, and target-matched or cross-target scenario probabilities were computed as in the original analyses. Source checksums, execution logs, event arrays and per-origin statistical outputs are included in Supplementary Data 4. Gaps elsewhere in 2022 were not simulated, and these results must not be presented as a full-year, continuous-state replay.

Non-anticipative forecasting and synthetic delivery obligations

The forecasting model is intentionally simple because the paper concerns reliability assessment rather than point-model novelty. At each ten-minute origin, the feature vector contains 24 lagged normalized-power values at lags 2–25 and sine/cosine time-of-day terms, for 26 features in total. With the primary 20-min availability convention, lag 2 is the most recent power label assumed available. Features are standardized using the training set only. A multioutput ridge model with an unpenalized intercept and penalty $\alpha = 10$ predicts 12 future ten-minute power values. Predictions are clipped to [0,1]. Ridge regression is used as a

transparent fixed baseline rather than an algorithmic contribution; transfer to Australia is intentionally frozen and is not used to claim local forecast adequacy.

The forecasting set contained 3,995 eligible fitting origins and ended when all associated targets had become available by 9 September 2021, 16:00 UTC. A subsequent 4,020 origins formed the residual bank, frozen at 7 October 2021, 16:00 UTC. Each residual is the realised 12-step trajectory minus the clipped ridge forecast. The six ten-minute phases contain 670 trajectories each. At an assessment origin, the 32 nearest same-phase residual paths are selected using Euclidean distance on train-standardised, currently available features, added to the new point forecast and clipped to [−0.01, 1.01]. The selection procedure never accesses subsequently realised wind.

The hourly delivery obligation is synthetic and must not be interpreted as a recorded market commitment. For delivery hour h, the primary obligation equals the clipped observation labelled h−80 min, with an assumed issuance at h−60 min under the primary 20-min information gate. The optimisation horizon covers the remainder of the current UTC clock hour and the following full hour, yielding 7–12 active ten-minute steps.

Information timing and Australian source semantics

The main analysis assumes a ten-minute-labelled observation can be used 20 min after its timestamp; this is a modelling convention, not a measured telemetry latency. Sensitivity analyses at 10, 30 and 40 min were undertaken after the original outputs had been frozen. In AEMO's data dictionary, SCADAVALUE denotes an instantaneous start-of-interval measurement, while example dispatch-solution SETTLEMENTDATE fields refer to intervals ending at the recorded timestamp [15]. We therefore assign a snapshot time of SETTLEMENTDATE minus five minutes and convert fixed UTC+10 Australian Eastern Standard Time to UTC.

For each ten-minute period starting at t, the power proxy is the mean of the SCADA snapshots at t and t+5 min, divided by 453 MW. It represents two five-minute zero-order-hold cells, not settlement-metered energy. Available finite readings are used without outcome-dependent clipping or interpolation. Thirty official daily archives spanning 30 August–28 September contained 8,640 unique five-minute records, with no missing, duplicate or non-numeric timestamps, forming 4,320 ten-minute pairs. Only the 2,592 proxies in the frozen evaluation interval contribute to the primary Australian results.

The provenance audit reparsed LASTCHANGED and payload-creation metadata for the 5,184 five-minute records contributing to the 18-day stress test. Creation metadata preceded the assumed 20-min information gate for every proxy and also preceded the 10-min sensitivity gate. These records establish that the audited archived payloads existed before the declared availability times; they do not establish wind-farm-to-AEMO telemetry receipt or external NEMWeb retrieval time.

Physical ledger and controllers

All powers are normalized by the selected farm capacity and the time step is $\Delta t = \frac{1}{6}$ h. Battery internal energy is constrained to 0.11–0.99 pu h with initial energy 0.55 pu h. Charge and discharge powers are each limited to 0.4 pu. Charge and discharge efficiencies are both $\eta = \sqrt{0.85}$. Upward and downward reserve powers are each limited to 0.2 pu, with separate energy budgets of 0.05 pu h per direction that reset at clock-hour boundaries. Grid charging is permitted. A disconnected episode restarts the simulated energy and reserve state after the required history warm-up; the reset is an explicit replay convention, not physical charging.

$$E_{t+1} = E_t + \Delta t \left(\eta\, c_t - \frac{d_t}{\eta}\right) \tag{1}$$

For realized wind w_t, local curtailment q_t, battery discharge d_t, charge c_t, upward reserve u_t and downward reserve v_t, delivered normalized power is

$$p_t = w_t - q_t + d_t - c_t + u_t - v_t \tag{2}$$

Curtailment is the only same-interval feedback allowed to read the realized wind before the interval ledger is closed; battery and reserve commands are held. Curtailment is bounded and is chosen minimally to enforce the precommitted export ceiling when feasible. The shortfall excess at interval t is $[s_t - p_t - \varepsilon]_+$ with $\varepsilon = 0.01$ pu and an additional numerical event margin of 10^{-7} pu. The primary horizon event equals one if any active interval has positive shortfall excess.

Three controllers use the same physical executor (Table 2). F_FULL constructs a greedy command sequence from the origin point forecast using full resource caps. F_BR0 evaluates a Cartesian menu of 225 first-command cap combinations against 32 wind scenarios, with common battery and reserve commands and scenario-dependent curtailment. Candidates are selected lexicographically by two-sided window failures, failing steps, a resource-and-violation proxy, descending cap sum and candidate index. DMPC solves a continuous linear program for battery charging and discharging, directional reserve, predicted curtailment and violation slack over the active horizon. Its objective weights shortfall and overdelivery excess by 100, battery throughput by 1 and reserve activation by 0.25, with a terminal energy deviation penalty of weight 1 about 0.55 pu h. Only the initial command is executed before replanning. The objective coefficients are engineering proxies, not electricity prices. Complete actuator, LP and paired-replay definitions appear in Supplementary Methods.

Continuation-specific paired labels

Let x_t denote the observed pre-action resource state, a_t the exact first command selected by the deployed method, $w_{t:t+H}$ the realized future wind path, and $s_{t:t+H}$ the frozen obligation over the active horizon. For a continuation c, define $Y_t(c) = 1$ if replaying the common first command a_t followed by c from x_t on the same w and s path produces at least one directional shortfall event. Three continuations are used: (i) B, a frozen-origin FULL continuation after the first command; (ii) P, the method's entire origin-fixed command sequence, used as an additional DMPC target; and (iii) D, the actual rolling deployed continuation reconstructed from subsequent controller actions. The deployed continuation intentionally bundles later state updates, forecast revisions and reoptimization; the primary B/P/D comparison measures their joint effect. The additional S diagnostic below fixes the original endpoint but does not isolate forecast updates from reoptimization.

$$\Delta(c) = \frac{100}{N} \sum_t [Y_t(D) - Y_t(c)] \tag{3}$$

The primary statistic $\Delta(c)$ is reported in percentage points. A positive value means more events under rolling deployment than under the named reference continuation. Paired labels are constructed only when the complete active horizon remains inside the same episode. Terminal executed rows with incomplete future labels remain in the operational ledger but are excluded from complete-horizon paired statistics. Exact replay of an identical fixed command sequence serves as the negative control and produced zero label discrepancies in the independent audit.

Fixed-original-endpoint and construction diagnostics

In the additional S replay, the DMPC first command and initial resource state at every eligible archived origin are fixed to those used in the baseline comparison. At subsequent relative steps, the same solver optimises over the original remaining obligation and endpoint, with newly available forecasts restricted to that residual horizon. S evolves its own resource state under the unchanged hourly reserve resets and terminal-energy objective. Future realised wind is used only in physical replay and outcome scoring, never in the optimiser. Every original complete paired origin is retained; an invalid command or failed solve terminates the calculation rather than being silently excluded. Pre-rollover commands are compared directly with those saved under D.

For the alternative synthetic-obligation construction, the hourly commitment is the clipped mean of the six observed power values labelled h−130, h−120, …, h−80 min, issued at h−60 min. All six observations must be finite, belong to the same continuous source episode and be available before issuance. Forecasts, residual scenarios, resource limits and objective weights remain fixed; chronological DMPC and all paired B/P/D outcomes are recomputed because changing the obligation changes the chosen commands. The original

single-observation rule remains primary. The added analyses were designed after the initial outcomes were known and must not be interpreted as prospectively specified confirmation tests.

Audit reporting procedure

For each eligible origin, we archive the pre-action state, first command, available-information set, obligation vector and original endpoint. The first command and realised disturbance path are shared across replays; only the named continuation varies. Both directions of event-label disagreement are tabulated with the net difference. Each probability is first scored against its own continuation-specific event. If rolling execution extends the task, an additional fixed-endpoint replay is used to aid interpretation. Operational frequency, shortfall magnitude and resource use are calculated from a single chronological deployment ledger rather than from overlapping assessment windows.

Probability scoring and operational endpoints

A scenario-based event probability is the fraction of the 32 paths on which the designated shortfall occurs under the named fixed continuation. The target-matched Brier score uses the corresponding realised label. For the cross-target diagnostic, only the realised label is replaced by that from rolling deployment; no probability is recalculated. For binary outcomes, the difference in squared errors reduces to equation (4), verified for all original cohort–target pairs. A matched or cross-target Brier score alone is not a calibration guarantee. Operational endpoints are computed from the chronological deployment: complete-hour and interval shortfall frequency, shortfall energy, battery and reserve use, curtailment and terminal energy. Under the permitted local-curtailment executor, recorded baseline overdelivery events are zero.

$$BS_D(p_c) - BS_c(p_c) = \frac{1}{N}\sum_t [Y_t(D) - Y_t(c)][1 - 2p_{t,c}] \quad (4)$$

Dependence-aware uncertainty and robustness

The frozen primary analysis used 2,000 paired ordinary moving-block replicates, sampling blocks independently within each contiguous run of complete assessment origins [18]. The seven-day duration was a retained analysis choice rather than a demonstrated decorrelation scale; one- and fourteen-day alternatives were examined for sensitivity. Runs no longer than a block were held fixed. Otherwise, full blocks were drawn uniformly over allowable starts, and the final block was truncated to preserve the original run length. Reported 2.5th–97.5th percentiles are exploratory resampling ranges conditional on the observed runs, predictors and controllers. They do not establish population coverage or include between-site, seasonal, selection or training uncertainty.

Archived comparisons were reproduced under all eligible origins, hourly phase-zero origins and a greedy non-overlapping-window design. For the latter, after retaining origin i, the next eligible origin is at least i+H_i, where H_i denotes that origin's horizon length. The resampling duration after thinning uses the within-run median selection cadence. The ordinary moving-block expectation was also calculated exactly from the set of allowable full-block starts and truncated prefixes. Separately, circular sampling selects starts uniformly across each valid run and wraps within that run before the final truncation. This removes finite-sample start-position weighting in expectation but introduces artificial endpoint adjacency, so it is not treated as an automatic confidence-interval correction. Lag-pair correlations and leave-one-day-out analyses remain descriptive diagnostics. Additional algorithms and frozen seeds are documented in Supplementary Methods.

The original latency and one-factor resource sweeps, as well as leave-one-calendar-day-out analyses, were retained in full rather than filtered for favourable results. Removing a date recalculates the paired contrast on all remaining eligible origins; it does not generate a confidence interval. Neither a resampling method nor a block length was selected post hoc to obtain a desired sign or narrower range.

Additional origin-design and circular resampling diagnostics

The post-freeze statistical audit recomputed paired deployment–reference differences for the three available cohorts under all complete origins, the first ten-minute phase of each clock hour and greedy non-overlapping origins. Within-cohort comparisons used the archived reference/deployment labels without refitting. For each cohort and reference, 2,000 circular moving-block draws were generated separately within continuous, fully scored runs at block durations of one, three and seven days; short runs were held fixed. The 2.5th and 97.5th

percentiles were used only as exploratory sensitivity ranges. Because origin selection changes horizon composition and circular sampling wraps endpoints, these results are descriptive robustness analyses, not validated population confidence intervals. The 2022 sample contains only 14 days and one evaluated farm-year segment. Daily and weekly descriptive contrasts, leave-one-period-out checks and identity checks are archived with the complete per-origin labels in Supplementary Data 4 (Supplementary Tables S23 and S24).

A separate post-freeze scoring-definition analysis reused the frozen DMPC trajectories. Controller actions, the 0.01 pu local-curtailment executor, state trajectories, realized wind and synthetic obligations were held fixed, and directional shortfall events were relabeled at scoring thresholds $\varepsilon \in \{0.005, 0.01, 0.02, 0.03\}$ pu. At $\varepsilon = 0.01$ pu, the same replays were also scored over the first three steps (30 min), the first six steps (60 min), and the full active 7–12-step window. Each condition used the same paired all-origin estimator and 2,000 seven-day moving-block resamples as the primary analysis. The archived 0.01 pu relabeling reproduced every stored shortfall label. Because changing the controller's design tolerance or the delivery obligation would alter the counterfactual trajectory itself, these analyses are explicitly scoring-definition sensitivities rather than alternative physical-control experiments.

A second post-freeze sensitivity reran DMPC under obligation scales $\kappa \in \{0.90, 1.00, 1.10\}$, replacing each synthetic obligation s by $\mathrm{clip}(\kappa s, 0, 1)$ while preserving the same issue times, forecasting pipeline, realized wind, resource settings and controller coefficients. For $\kappa = 0.90$ and 1.10, the controller and both paired continuations were recomputed from scratch; $\kappa = 1.00$ is the frozen primary run. The same all-origin estimator and 2,000 seven-day moving-block resamples were then applied. Because this changes the service level but not the rule that constructs or issues the obligation, it is reported as an obligation-level sensitivity rather than validation of a different market product.

Independent audit and software

Independent scripts reconstructed the training, residual-bank and development indices; re-solved the ridge coefficients; perturbed not-yet-available inputs to test non-anticipativity; checked battery and reserve prefix constraints; verified that no opposite-direction actions were simultaneously executed; replayed every saved execution step; rebuilt every paired label; and recomputed target-matched probability scores. Exact replay of the same fixed command sequence gave zero directional label mismatches in the audit. The 20-min parameterized builder reproduced archived development and Australian numerical inputs to within 3.5×10^{-15} maximum absolute difference across checked arrays. All full analyses had zero solver fallback. The frozen primary analyses used Python 3.12.14, NumPy 2.3.5, pandas 2.2.3 and SciPy 1.17.0; DMPC used scipy.optimize.linprog(method='highs') [19]. The final post-freeze obligation-level reruns were independently regenerated under Python 3.13.5 with NumPy 2.3.5 and SciPy 1.17.0, with zero solver fallback; the $\kappa = 1.00$ values are the unchanged frozen primary results.

Revision diagnostics are supplied in Supplementary Data 3. The fixed-original-endpoint comparison required 98,561 additional linear-program solves over 11,596 complete origins, with no solver failures or fallback. An independently implemented scalar executor reproduced 110,157 evaluated step transitions to a maximum absolute discrepancy of 2.22×10^{-16}. All 40,581 pre-rollover command comparisons between S and D agreed exactly. Alternative-obligation reruns separately verified 11,620 chronological steps and 34,788 paired B/P/D labels, again with no solver fallback and maximum step discrepancy of 2.22×10^{-16}. All 12 archived thinning rows and the four original Brier-score identities were numerically reproduced. These calculations test internal numerical consistency, not external field validity. The earlier full latency and resource sweeps were not rerun in the present revision; their archived scope remains identified in Supplementary Information.

The new 2022 extension produced 1,991 chronological execution origins, 1,985 complete paired origins and 331 complete hours, with no solver fallback for F_FULL, F_BR0 or DMPC. Independent recomputation of paired counts gave the DMPC deployed-only/reference-only totals of 124/131 against FULL and 139/4 against the origin-fixed plan. The additional analysis reproduced archived original-cohort contrasts, recalculated the three origin-selection designs and checked the 2022 probability-score changes against the stored event vectors. Complete output arrays, source checksums, adapter code, execution logs and analysis drivers are included in Supplementary Data 4. These internal checks have not been independently reproduced by all coauthors and do not validate physical field operation.

OpenAI ChatGPT and Codex assisted with research-code review, reproducibility checks and manuscript drafting and language editing. The authors are responsible for independent verification of the reported methods, results and final text.

Code availability

The primary frozen code and analysis files are supplied as Supplementary Data 1 (Supplementary_Data_1_Reproducibility_20261006.zip). Supplementary Data 3 (Supplementary_Data_3_Revision_Analyses.zip) contains the fixed-endpoint and alternative-obligation scripts, dependence diagnostics, result arrays, source figures and verification records; Supplementary Data 2 contains the archived AEMO raw source files. The 2022 UK temporal extension and additional three-cohort statistical analyses are provided separately in Supplementary Data 4 (Supplementary_Data_4_UK2022_and_Statistical_Validation.zip), including the adapted execution scripts, frozen-model provenance, per-origin labels, analysis outputs and source checksums. These archives are intended to be used together; computational and data requirements are documented within them. No public repository identifier is claimed.

Data availability

UK wind-power observations originate from the publicly released Kelmarsh dataset on Zenodo under CC BY 4.0 [14]. The Australian observations derive from AEMO's public DISPATCH_UNIT_SCADA archives (https://nemweb.com.au/Reports/Archive/Dispatch_SCADA/), with interpretation supported by the official data dictionary [15] and farm capacity sources [16,17]. Supplementary Data 1 supplies processed 2021 and Australian assessment inputs, chronological ledgers and frozen primary outputs; Supplementary Data 2 supplies 30 AEMO daily raw archives and provenance records; Supplementary Data 3 supplies revision analyses and supporting arrays. Supplementary Data 4 contains the 2022 UK extension input subset, execution and statistical outputs, scripts and file-level provenance. The synthetic obligations, modelling assumptions and simulation results are supplied for reproducibility, not as evidence of recorded commercial delivery contracts. No private operational dataset is needed for the reported analyses.

References

1. Kou, P., Gao, F. & Guan, X. Stochastic predictive control of battery energy storage for wind farm dispatching: using probabilistic wind power forecasts. *Renew. Energy* **80,** 286–300 (2015). https://doi.org/10.1016/j.renene.2015.02.001.
2. Toubeau, J.-F., Bottieau, J., De Grève, Z., Vallée, F. & Bruninx, K. Data-driven scheduling of energy storage in day-ahead energy and reserve markets with probabilistic guarantees on real-time delivery. *IEEE Trans. Power Syst.* **36,** 2815–2828 (2021). https://doi.org/10.1109/TPWRS.2020.3046710.
3. Hewing, L., Wabersich, K. P. & Zeilinger, M. N. Recursively feasible stochastic model predictive control using indirect feedback. *Automatica* **119,** 109095 (2020). https://doi.org/10.1016/j.automatica.2020.109095.
4. Wang, K. & Gros, S. Recursive feasibility of stochastic model predictive control with mission-wide probabilistic constraints. In 2021 60th IEEE Conference on Decision and Control (CDC), 2312–2317 (IEEE, 2021). https://doi.org/10.1109/CDC45484.2021.9683028.
5. Wabersich, K. P. & Zeilinger, M. N. A predictive safety filter for learning-based control of constrained nonlinear dynamical systems. *Automatica* **129,** 109597 (2021). https://doi.org/10.1016/j.automatica.2021.109597.
6. Lekeufack, J., Angelopoulos, A. N., Bajcsy, A., Jordan, M. I. & Malik, J. Conformal decision theory: safe autonomous decisions from imperfect predictions. In 2024 IEEE International Conference on Robotics and Automation (ICRA), 11668–11675 (IEEE, 2024). https://doi.org/10.1109/ICRA57147.2024.10610041.
7. Angelopoulos, A. N., Bates, S., Fisch, A., Lei, L. & Schuster, T. Conformal risk control. In International Conference on Learning Representations (2024). https://openreview.net/forum?id=33XGfHLtZg.
8. Farinhas, A., Zerva, C., Ulmer, D. & Martins, A. F. T. Non-exchangeable conformal risk control. In International Conference on Learning Representations (2024). https://openreview.net/forum?id=j511LaqEeP.
9. Zecchin, M., Park, S. & Simeone, O. Forking uncertainties: reliable prediction and model predictive control with sequence models via conformal risk control. *IEEE J. Sel. Areas Inf. Theory* **5,** 44–61 (2024). https://doi.org/10.1109/JSAIT.2024.3368229.
10. Stamouli, C., Lindemann, L. & Pappas, G. J. Recursively feasible shrinking-horizon MPC in dynamic environments with conformal prediction guarantees. *Proc. Mach. Learn. Res.* **242,** 1330–1342 (2024). https://proceedings.mlr.press/v242/stamouli24a.html.
11. Kuipers, T. et al. Conformal off-policy prediction for multi-agent systems. In 2024 IEEE 63rd Conference on Decision and Control (CDC), 1067–1074 (IEEE, 2024). https://doi.org/10.1109/CDC56724.2024.10886791.

12. Prinster, D. et al. Conformal policy control. *Proc. Mach. Learn. Res.* **306,** 99772–99824 (2026). https://proceedings.mlr.press/v306/prinster26a.html.
13. Bouchkati, S. et al. Safety screening for voltage control in active distribution grids via distributionally robust conformal screening. Preprint at https://doi.org/10.48550/arXiv.2608.30889 (2026).
14. Plumley, C. & Takeuchi, R. Kelmarsh wind farm data. *Zenodo*, version 4 (2025). https://doi.org/10.5281/zenodo.16807551.
15. Australian Energy Market Operator. MMS Data Model Report: DISPATCH_UNIT_SCADA and DISPATCHCASESOLUTION. https://visualisations.aemo.com.au/aemo/nemweb/MMSDataModelReport/Electricity/MMS%20Data%20Model%20Report_files/MMS_123.htm; https://visualisations.aemo.com.au/aemo/nemweb/MMSDataModelReport/Electricity/MMS%20Data%20Model%20Report_files/MMS_125.htm (accessed 3 October 2026).
16. Tilt Renewables. Coopers Gap Wind Farm. https://tiltrenewables.com/portfolio/coopers-gap-wind-farm/ (accessed 3 October 2026).
17. Australian Energy Market Operator. 2025 Enhanced Locational Information Report, Appendix A4: Queensland. https://www.aemo.com.au/-/media/files/electricity/nem/planning_and_forecasting/enhanced-locational-information/2025/2025-eli-report-appendix-a4-queensland.pdf (accessed 3 October 2026).
18. Künsch, H. R. The jackknife and the bootstrap for general stationary observations. *Ann. Stat.* **17,** 1217–1241 (1989). https://doi.org/10.1214/aos/1176347265.
19. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17,** 261–272 (2020). https://doi.org/10.1038/s41592-019-0686-2.

Acknowledgements

We thank Cubico Sustainable Investments for making the Kelmarsh wind-power observations available, the Australian Energy Market Operator for maintaining the public dispatch-data archive, and Tilt Renewables for public project metadata used only as a normalisation reference.

Author contributions

Xin He, Yingshuo Qi and Xinrong Xie contributed equally to study design, methodology, software, formal analysis, data curation, validation, visualisation, writing and manuscript revision. Yongpeng Tong supervised the research, contributed to validation and software review, and reviewed and edited the manuscript. All authors are responsible for approving the final submitted version.

Funding

No financial support was received for this research.

Additional Information

Competing interests

The authors declare no competing interests.

Table 1. Development, transfer and exploratory temporal-extension cohorts.

Cohort / role	Power source	Evaluation period	Executed 10-min intervals	Complete paired origins	Complete hours	Evidence status
UK development	Processed Kelmarsh six-turbine aggregate [14]	7 Oct–31 Dec 2021 (UTC)	9,028	9,010	1,503	Historically exposed development
Australian stress test	AEMO COOPGWF1 public SCADA [15]	11–28 Sep 2026 (AEST)	2,592	2,586	432	Frozen external stress test
UK temporal extension	Kelmarsh six-turbine aggregate [14]	1–14 Jan 2022 (UTC)	1,991	1,985	331	Exploratory, within-site; source previously quality-reviewed

The UK 2021 development period begins after the predictor and residual bank were frozen. The UK 2022 exploratory extension uses 1–14 January 2022 UTC with a single uninterrupted physical-state simulation and 1,985 complete paired origins. Six terminal Australian origins were executed but did not contain a complete future horizon and were excluded from paired assessments. The UK 2022 extension was not a blinded or annual holdout.

Table 2. Frozen controller definitions and continuation targets.

Controller	Planning rule	Deployed execution	Assessed continuation(s)
F_FULL	Point-forecast full-resource rule	Continuous held charge, discharge and reserve commands; replanned each 10 min	Frozen-origin forecast FULL tail
F_BR0	32 residual scenarios; 225 first-cap combinations	Scenario-shared held commands; lexicographic event/step/resource selection; replanned each 10 min	Selected first command followed by frozen-origin FULL
DMPC	Deterministic linear program over the full active horizon	Continuous charge, discharge, directional reserve and predicted curtailment; execute first command and reoptimize	Frozen-origin FULL continuation; additionally, its own origin-fixed optimized sequence

All controllers use the same physical resource constraints. F_BR0 is restricted to a menu of 225 first-command cap combinations; DMPC has no corresponding menu restriction. FULL is a designated comparison continuation, not a certified safe backup. Objective weights are engineering surrogates rather than monetary costs.

Table 3. Continuation-specific paired shortfall events in the evaluated cohorts.

Cohort	Controller	Reference continuation	Reference event rate (%)	Deployed event rate (%)	Gap (pp)	7-day 95% range (pp)	Label mismatch (%)
UK development	F_FULL	Frozen-origin FULL	73.82	86.54	+12.72	[9.48, 16.45]	18.27
UK development	F_BR0	Frozen-origin FULL	77.46	69.61	-7.85	[-13.37, -3.57]	15.73
UK development	DMPC	Frozen-origin FULL	75.85	69.60	-6.25	[-12.82, -1.80]	16.50
UK development	DMPC	Origin-fixed DMPC plan	64.02	69.60	+5.58	[4.15, 6.73]	6.67
Australian stress test	F_FULL	Frozen-origin FULL	65.85	69.14	+3.29	[-5.26, 6.34]	17.90
Australian stress test	F_BR0	Frozen-origin FULL	71.11	58.04	-13.07	[-25.79, -9.01]	18.10
Australian stress test	DMPC	Frozen-origin FULL	69.61	57.27	-12.34	[-25.48, -8.00]	18.83
Australian stress test	DMPC	Origin-fixed DMPC plan	53.09	57.27	+4.18	[2.55, 5.03]	4.64
UK 2022 extension	F_FULL	Frozen-origin FULL	74.91	91.89	+16.98	not estimated	19.19
UK 2022 extension	F_BR0	Frozen-origin FULL	78.49	77.03	−1.46	not estimated	12.14
UK 2022 extension	DMPC	Frozen-origin FULL	76.88	76.52	−0.35	[−5.19, +4.53]†	12.85
UK 2022 extension	DMPC	Origin-fixed DMPC plan	69.72	76.52	+6.80	[+4.84, +8.87]†	7.20

Gap = deployed minus named reference event rate (percentage points). Archived 2021/2026 ranges are 2.5th–97.5th percentiles from 2,000 ordinary seven-day within-run moving-block resamples, not validated population confidence intervals. UK 2022 ranges are from a separate circular seven-day sensitivity and are marked accordingly. Origin thinning alters the estimates; the 2022 FULL contrast changes sign with hourly or non-overlapping origin selection. Resampling excludes independent site, season, model-selection and training uncertainty (Supplementary Tables S14–S16 and S23–S24).

Table 4. Operational endpoints under chronological deployment.

Cohort	Controller	Failed hours (%)	Excess shortfall (pu h)	Mean excess power (pu)	Battery throughput (pu h)	Reserve use (pu h)	Curtailment (pu h)
UK development	F_FULL	78.84	35.20	0.0234	90.84	10.15	40.19
UK development	F_BR0	58.02	41.28	0.0274	10.64	25.74	66.60
UK development	DMPC	58.68	39.04	0.0259	19.74	28.33	66.57
Australian stress test	F_FULL	60.65	9.70	0.0224	27.93	2.59	9.48
Australian stress test	F_BR0	47.45	14.78	0.0342	2.36	7.03	21.39
Australian stress test	DMPC	47.22	13.76	0.0319	4.01	8.04	21.25
UK 2022 extension	F_FULL	82.78	11.34	0.0343	26.20	2.91	11.21
UK 2022 extension	F_BR0	62.84	13.78	0.0416	3.34	6.89	19.57
UK 2022 extension	DMPC	64.05	12.81	0.0387	6.52	7.63	19.07

Energy totals integrate the corresponding evaluation interval and are not directly comparable across cohorts of different duration. Mean excess shortfall power is time-normalised. No solver fallback occurred in any of the reported controller replays. Shortfall rates reflect synthetic obligations and model constraints, not field reliability.

Figure legends

Figure 1

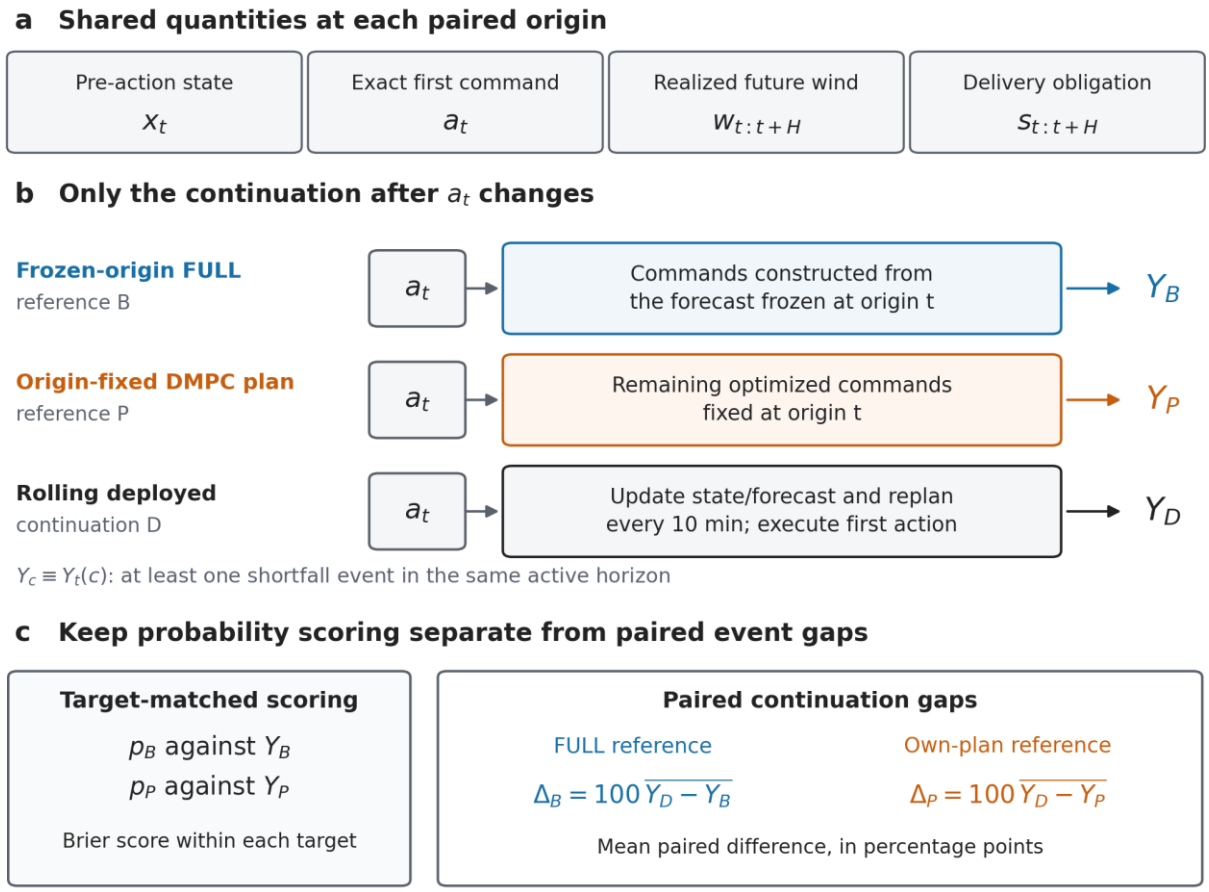


Figure 1. Continuation-specific reliability audit. (a) The pre-action state, exact first command, realised future wind and delivery obligation are matched at every assessment origin. (b) Only the policy executed after the first command varies: frozen-origin FULL (B), origin-fixed DMPC plan (P), or rolling deployment (D). (c) Probabilities are scored against the events induced by their named continuations; separate paired comparisons quantify the event shift on deployment. The procedure is an empirical audit, not a new control-theoretic guarantee.

Figure 2

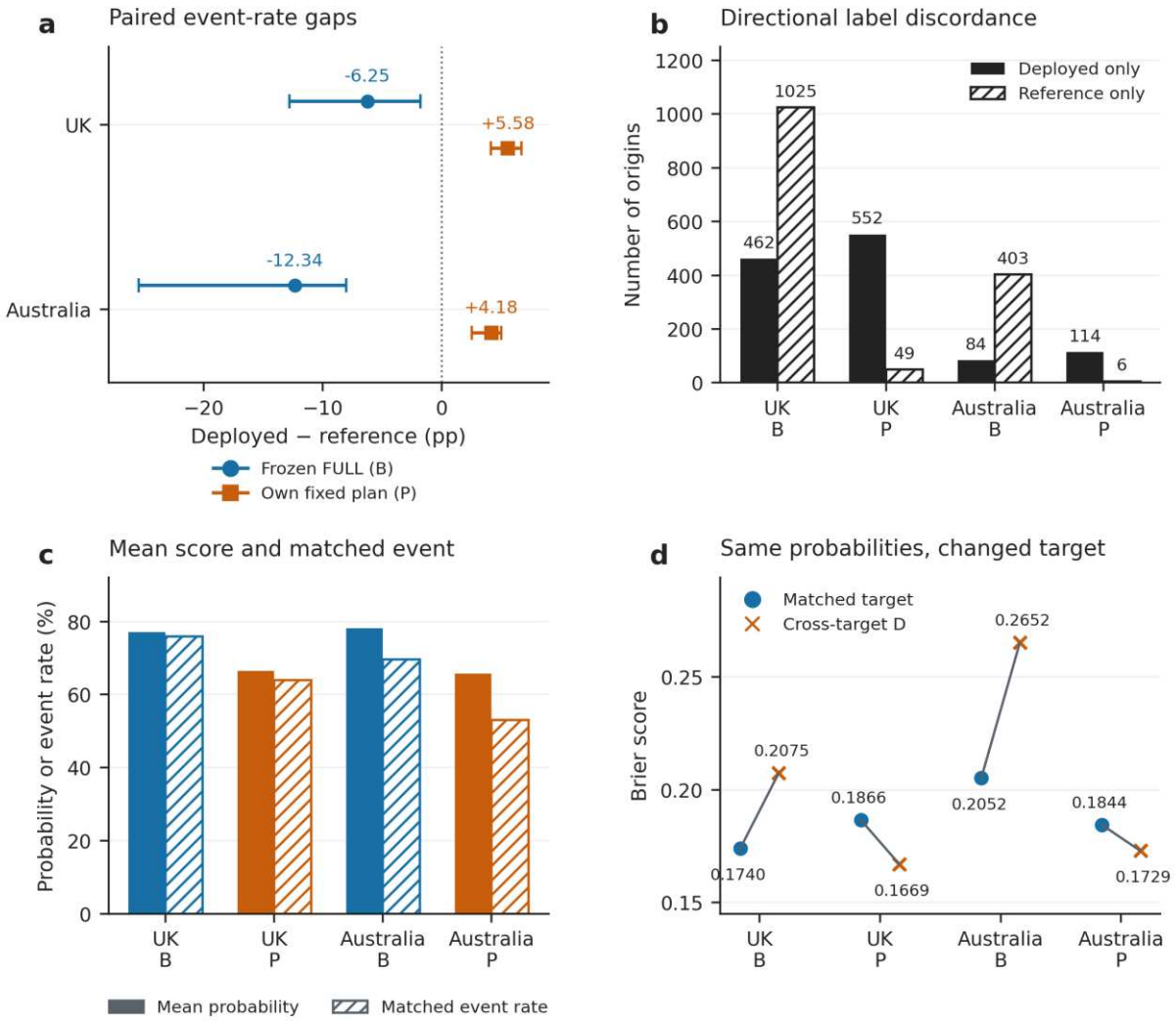


Figure 2. Reference-target alignment and probability-score differences for DMPC in the original cohorts. (a) Deployment–reference contrasts and exploratory ordinary seven-day block-resampling ranges, which are not validated population confidence intervals. (b) Deployed-only and reference-only event discordance. (c) Mean scenario probabilities and observed event frequencies under matched targets. (d) Identical probabilities scored against their named reference event or the different deployment event. A change in scoring target is not an improvement or deterioration on an unchanged prediction task. The exploratory 2022 extension is reported separately in Table 3 and Supplementary Tables S21–S24.

Figure 3

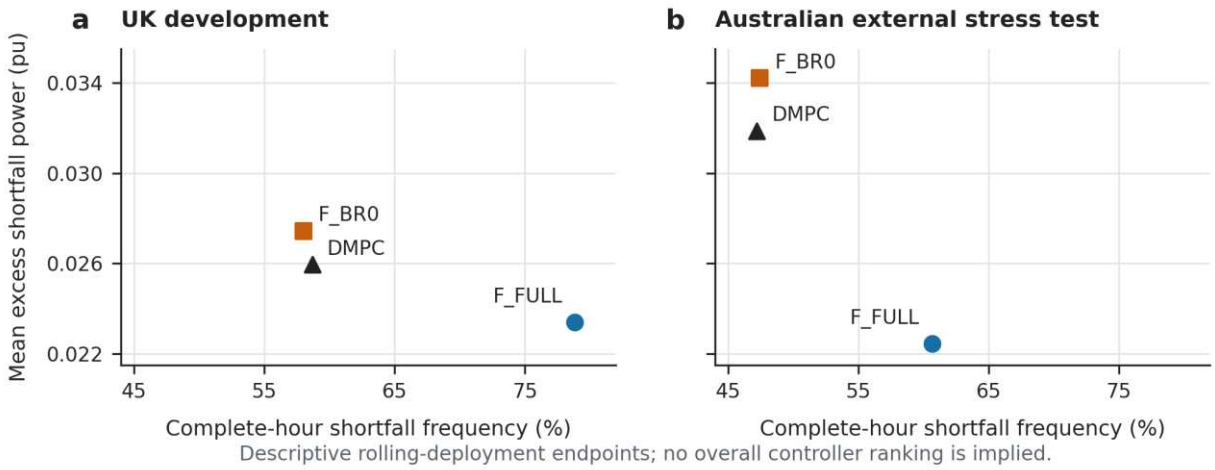


Figure 3. Shortfall frequency and severity under chronological deployment. (a) UK 2021 development cohort and (b) Australian transfer cohort. Each point denotes one controller. Fewer complete hours containing a shortfall can coincide with greater excess shortfall power, so the panels do not establish an overall controller ranking. Results for the 2022 UK extension are reported separately in Table 4.

Figure 4

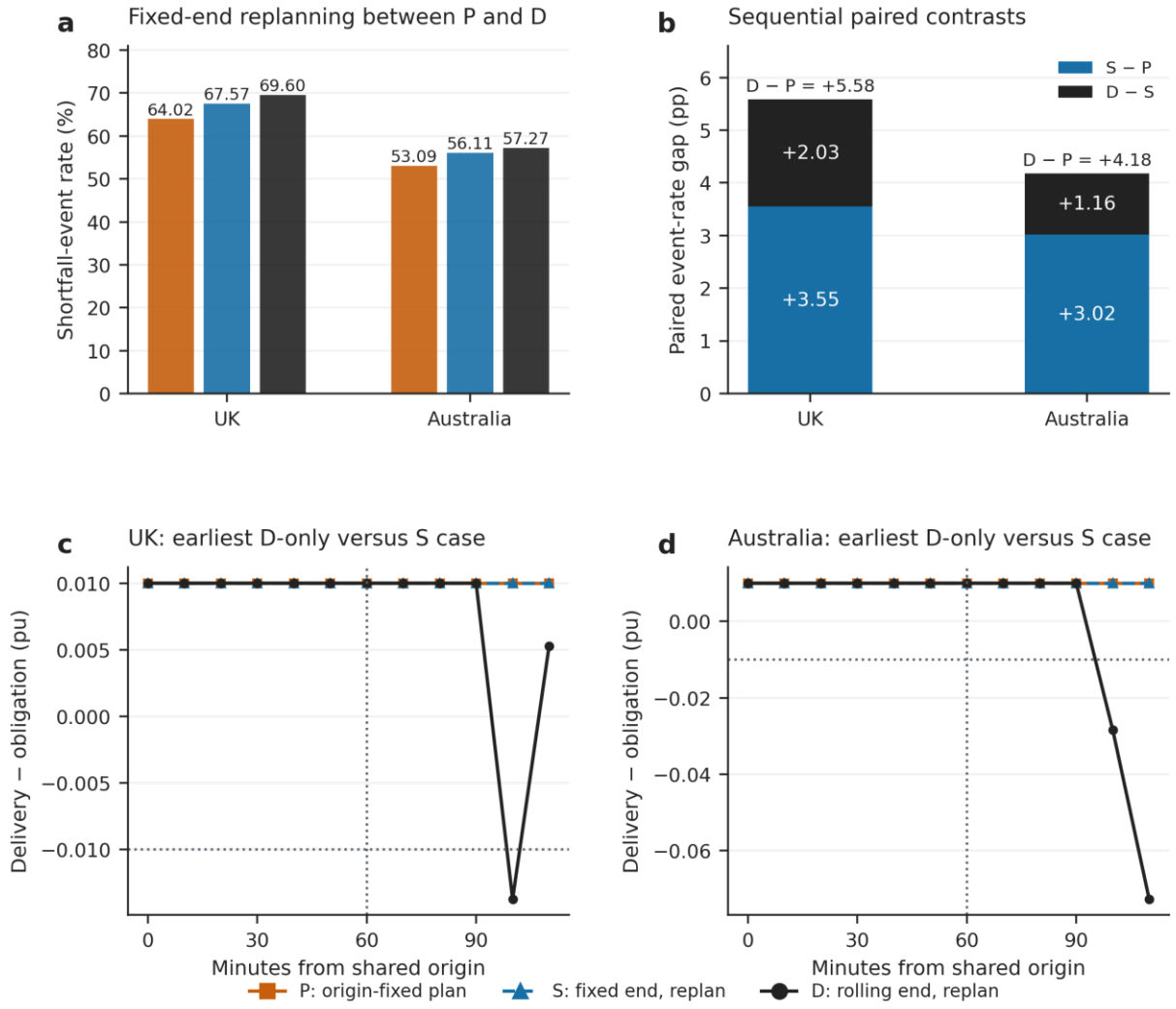


Figure 4. Fixed-endpoint reoptimisation as a mechanism diagnostic in the original cohorts. (a) Event rates under the origin-fixed plan P, fixed-original-endpoint reoptimisation S and rolling deployment D. (b) Sequential contrasts S−P and D−S sum to D−P but do not identify independent causal effects. (c,d) The earliest qualifying D-only-versus-S origin in each cohort, selected chronologically: 14 October 2021 at 20:00 UTC and 11 September 2026 at 05:00 UTC. Dashed lines indicate the first hour boundary; horizontal lines mark the −0.01 pu shortfall tolerance. The selected examples illustrate trajectories and are not representative cases. Complete command and resource paths appear in Supplementary Figs. S4 and S5.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ScientificReportsTables.zip
- ScientificReportsSupplementaryInformationUPLOAD.pdf
- SupplementaryData3RevisionAnalyses.zip
- SupplementaryData4UK2022andStatisticalValidation.zip
- SupplementaryData1Reproducibility20261006.zip
- SupplementaryData2AEMORawSources1.zip