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.

Risk-Constrained Energy Storage Scheduling Under Electricity Price Forecast Uncertainty

Kangqian Huang

ecbbig9777345@outlook.com

South China Agricultural University

Xin Hu

Changsha University of Science and Technology

Jingxiang Shang

Hong Kong Polytechnic University

Fan Yang

North China Electric Power University

Zhiqing Wang

Guangzhou Xinhua University

Article

Keywords:

Posted Date: November 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11576188/v1

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

Additional Declarations: No competing interests reported.

Risk-Constrained Energy Storage Scheduling Under Electricity Price Forecast Uncertainty

Kangqian Huang[1,*], Xin Hu[2], Jingxiang Shang[3], Fan Yang[4], and Zhiqing Wang[5]

[1]South China Agricultural University, Guangzhou, China
[2]Changsha University of Science and Technology, Changsha, China
[3]The Hong Kong Polytechnic University, Hong Kong, China
[4]North China Electric Power University, Beijing, China
[5]Xinhua College of Sun Yat-sen University, Guangzhou, China
[*]Correspondence: Kangqian Huang (email: ecbbig9777345@outlook.com)

ABSTRACT

Electricity price uncertainty can turn anticipated storage arbitrage gains into concentrated financial losses when forecast errors coincide with scheduled charging or discharging. Improved prediction scores alone therefore provide insufficient evidence that a forecasting and scheduling system delivers reliable economic value. A unified residual scenario interface connects Ridge, quantile LightGBM, and quantile TCN to identical battery constraints and scenario CVaR scheduling. Controlled ablations vary state conditioning and temporal pairing while preserving forecast means or empirical marginals. Chronological evaluation combines NSW1 testing, regional retraining, probability diagnostics, paired bootstrap intervals, and sensitivity analyses across budgets and battery configurations. Under the reference NSW1 configuration, Ridge CVaR scheduling increased annual net profit from 99,931 to 116,254 AUD and reduced realized daily loss CVaR_{95} from 608.40 to 504.98 AUD. Stronger predictors did not consistently dominate economically. State conditioning barely improved high volatility coverage, and complete paths offered no observed advantage. The 1 MWh configuration increased tail risk relative to deterministic scheduling. The evidence supports conditional economic benefits rather than guaranteed loss protection. VIC1 and SA1 extend regional scheduling validation, while DE and PJM assess forecasting only.

1 Introduction

Electricity price volatility creates opportunities for battery energy storage while exposing market participants to substantial operating uncertainty. Charging during low-price intervals and discharging when prices rise can generate arbitrage revenue, but the value of this flexibility depends on the timing and magnitude of price differences, conversion losses, and storage limitations. Negative prices can make charging financially attractive, whereas short-lived price spikes can dominate the profitability of an entire trading day. These characteristics are particularly relevant in the Australian National Electricity Market, where probabilistic forecasting must accommodate highly volatile prices and changing market conditions[1]. Evidence from European markets also demonstrates that storage arbitrage value varies across regions and periods, with efficiency and storage duration influencing achievable returns[2]. Such variation makes average historical profitability an incomplete basis for operational planning. An engineering assessment must determine whether a forecast-informed schedule captures useful price spreads while controlling exposure to adverse realizations. This requirement connects electricity price forecasting with uncertainty representation, battery constraints, and the economic evaluation of scheduling decisions across changing daily market conditions.

The decision setting examined here involves a price-taking battery that fixes its next-day charging and discharging schedule using information available before the operating day. The resulting actions are evaluated against realized prices, so forecast errors can affect revenues even when the planned schedule satisfies every physical constraint. Underestimating an expensive interval may encourage charging at an unexpectedly high cost, while overestimating a selling opportunity may allocate limited stored energy to an unprofitable discharge. Previous stochastic arbitrage research has explicitly considered uncertainty in electricity prices when determining storage operations[3]. Price uncertainty and financial risk represent economically different quantities. A large forecasting error matters economically only through the battery's exposure, operating constraints, and associated actions. Even moderate errors can produce substantial losses when they reverse the anticipated ordering of buying and selling prices over successive intervals. Research comparing risk measures and degradation costs further shows that introducing risk management need not improve arbitrage outcomes relative to deterministic strategies[4]. Assessing a scheduling policy therefore requires observed monetary outcomes alongside forecast statistics, rather than treating a numerical risk constraint as evidence of realized protection.

Electricity price forecasting has developed from statistical and regression-based approaches toward increasingly flexible machine learning architectures. Reviews describe a broad range of point forecasting methods and emphasize the distinctive statistical properties of electricity prices[5]. Probabilistic forecasting extends this task by representing uncertainty through quantiles, prediction intervals, or predictive distributions, allowing the analysis to address outcomes beyond a single expected price[6]. Deep learning comparisons have examined whether nonlinear architectures outperform traditional forecasting algorithms[7], while distributional neural networks directly estimate parameters of predictive price distributions[8]. These developments provide useful forecasting tools, but predictive flexibility alone does not establish operational value. Point errors average discrepancies across intervals without accounting for the battery's eventual trading exposure. Marginal quantile scores evaluate distributional forecasts at individual times but do not fully describe the financial implications of a complete daily price trajectory. Decision-focused electricity price prediction has consequently sought to connect forecasting with downstream storage optimization[9]. Recent research evaluating both statistical and economic forecast quality reinforces the need to examine characteristics beyond aggregate error measures[10]. For battery scheduling, the relevant question is whether a forecasting improvement changes feasible actions in ways that improve realized revenue and downside risk. These effects require direct evaluation within an operational model.

Storage scheduling under price uncertainty is commonly formulated using deterministic forecasts, scenario-based stochastic optimization, or explicitly risk-aware objectives and constraints. Deterministic scheduling optimizes operations against a predicted price trajectory and provides a transparent reference for assessing the value of uncertainty treatment. Scenario-based formulations instead evaluate candidate schedules across possible price realizations, with the resulting decisions depending on both scenario values and their joint organization. Conditional value-at-risk offers a tractable way to summarize adverse outcomes beyond a selected loss quantile, but its practical interpretation depends on the loss definition, confidence level, and treatment of operating costs. Applications in electricity retailing have distinguished daily and hourly risk assessments[11], and integrated risk-control research has examined how alternative risk preferences affect energy trading decisions[12]. Studies of storage participation in day-ahead and balancing markets further illustrate the interaction between uncertainty and market-specific operating opportunities[13]. Degradation-aware risk formulations also recognize that additional trading activity can alter the cost of storage participation[14]. These established approaches motivate the present formulation without implying that CVaR itself is a new contribution. Their evaluation nevertheless requires a distinction between risk computed under the assumed scenario distribution and risk measured from subsequent market observations. A schedule can satisfy its scenario-based budget while experiencing substantial realized losses if the assumed distribution misrepresents economically important outcomes. These assumptions warrant scrutiny when interpreting realized performance.

The evidence needed for practical model selection extends beyond demonstrating that one forecasting architecture or risk formulation performs well in a particular case study. Multivariate electricity price forecasting research has shown how marginal predictions can be combined with temporal dependence to construct daily predictive distributions[15], and normalizing-flow methods provide another route to coherent price scenarios[16]. Preserving an observed empirical dependence pattern does not automatically establish that the resulting scenarios are suitable for a particular battery or risk budget. Forecasts can exhibit similar marginal coverage while inducing different assessments of losses from the same actions. Research on probabilistic forecasting for electricity trading also links distributional prediction with economic evaluation[17], while conformal interval methods address uncertainty quantification in power markets[18]. A broader assessment of multivariate forecasting models highlights the importance of evaluating predictive distributions beyond isolated point predictions[19]. The gap addressed here is therefore a controlled empirical comparison of how these representations translate into storage decisions under a shared optimization setting and consistent physical battery assumptions. Forecast means, marginal distributions, battery constraints, and evaluation periods must be held comparable whenever the tested mechanism permits, so that changes in realized outcomes can be interpreted without conflating forecasting accuracy, scenario construction, and risk tolerance.

This study examines three connected questions through a sequence of forecasting, scheduling, and diagnostic experiments. One question concerns whether better point prediction and marginal probability calibration correspond to higher storage revenue and lower realized tail losses. Ridge regression, quantile LightGBM, and a basic quantile temporal convolutional network are evaluated through a common residual-based scenario interface and the same battery scheduler, allowing forecast-model differences to be assessed at both predictive and economic levels. Another question concerns how state-conditioned uncertainty estimates and temporal scenario pairing affect risk evaluation and charging decisions. Historical volatility states are determined from information available before scheduling, and their residual banks are compared with an unconditional bank. A dependence experiment preserves each time point's empirical scenario values while changing their cross-time pairing, isolating the role of joint organization within that scenario bank. The remaining question concerns whether observed advantages persist across risk budgets, storage parameters, scenario counts, and regional markets. Budget curves, capacity-normalized risk controls, efficiency and degradation-cost variations, and regional retraining provide assessments. Loss diagnostics inspect adverse trading days, forecast discrepancies, state of charge, and charging or discharging exposure. Australian regional tests evaluate scheduling outcomes, whereas supplementary European and US benchmark data evaluate forecasting only. These distinctions ensure that

each question is matched to the evidence produced, without treating prediction validation as proof of transferable arbitrage performance.

The contribution is an auditable empirical framework for examining the decision value and limitations of forecast uncertainty in risk-constrained storage scheduling. Its controlled forecasting comparison separates raw prediction outputs from the calibrated mean and scenario distributions supplied to optimization, reducing ambiguity about what the scheduler receives. Its marginal-preserving dependence experiment changes temporal pairing without changing the scenario values at individual horizons, enabling an examination of how joint distributions influence risk assessments and selected actions. The accompanying fixed-action comparison distinguishes changes in estimated risk from changes in actual monetary outcomes. Its diagnostic and sensitivity analyses identify conditions under which aggregate performance masks concentrated losses or parameter-dependent reversals. The framework retains unfavorable evidence, including incomplete calibration, revenue-risk trade-offs, and cases where a nominally risk-constrained policy does not reduce realized tail losses. Regional validation evaluates locally retrained models under a shared protocol, rather than claiming that weights learned in one market transfer unchanged to another. Supplementary forecasting benchmarks provide a separate check on predictive behavior across different price environments. These contributions concern evaluation design and empirical boundaries rather than a new CVaR algorithm or a universally superior forecasting model. The repeated use of the test period for diagnosis is acknowledged as exploratory, and the resulting evidence is interpreted as a basis for independent validation rather than an unconditional guarantee of operational risk protection.

Section II defines the decision setting, forecasting models, calibration variants, scenario construction, and battery scheduling formulation. Section III describes the datasets, chronological partitions, comparison protocols, implementation settings, and statistical evaluation. Section IV reports forecasting performance, realized scheduling outcomes, loss diagnostics, calibration and dependence ablations, budget sensitivity, battery variations, and external validation. Section V interprets the findings, compares their implications with related research, and discusses engineering limitations and future validation requirements. Section VI concludes by summarizing the supported decision-level findings and the conditions that limit their applicability in the evaluated settings.

2 Forecasting and Risk-Constrained Scheduling Framework

2.1 Decision Setting and Forecasting Models

The scheduling problem concerns a battery that participates as a price taker in a regional electricity spot market. Its operating decisions are assumed to have no effect on market prices, and its revenue therefore depends on the realized price trajectory rather than an endogenous clearing mechanism. A decision is made at midnight before each operating day, indexed by n, and determines charging and discharging power for every subsequent interval. The Australian scheduling experiments use a half-hourly grid with H equal to 48 intervals and an interval duration of 0.5 hours. The complete daily action vector is fixed before any price from that operating day becomes available. No adjustment is made after observing intermediate prices, so uncertainty affects the schedule through its initial construction. Charging purchases electricity and discharging sells electricity, with conversion losses, power limits, energy limits, and terminal energy requirements imposed by the battery model. Realized operating profit is evaluated using observed regional prices and the specified throughput cost. This setting represents a forecast-informed arbitrage backtest, rather than a simulation of submitted bids, market clearing, or balancing participation. Comparing policies within this setting isolates the consequences of their forecast and uncertainty inputs while holding their physical operating opportunities comparable.

$$\mathbf{u}_n = \left(p^{\mathrm{ch}}_{n,1}, \ldots, p^{\mathrm{ch}}_{n,H}, p^{\mathrm{dis}}_{n,1}, \ldots, p^{\mathrm{dis}}_{n,H}\right)^{\top}, \qquad H = 48, \qquad \Delta t = 0.5\,\mathrm{h}. \tag{1}$$

Let the forecast origin for day n be midnight, denoted by τ_n, and let h index intervals within the next operating day. The target is the vector of realized prices π_n, while the model input $\mathbf{x}_n$ is constructed from historical observations and known calendar information. The nominal information cutoff precedes the origin by one observation interval, corresponding to 30 minutes in the Australian experiments. Observations from the operating day are excluded from all forecast inputs. Historical features comprise the latest two days of available price and demand observations and the corresponding daily trajectories from one and two weeks earlier. Calendar features encode weekly and annual position using sine and cosine functions. A monotone inverse-hyperbolic-sine transformation is applied to prices, retaining negative observations while reducing the influence of extreme magnitudes. Demand is divided by a fixed scale, and any statistical normalization is estimated using training observations only. These operations define the feature mapping from the information set to the forecasting input. The supplementary hourly forecasting experiments retain the same history arrangement but use historical prices and calendar features only, with the demand channel set to zero. Their horizon contains 24 intervals, and no storage scheduling is performed

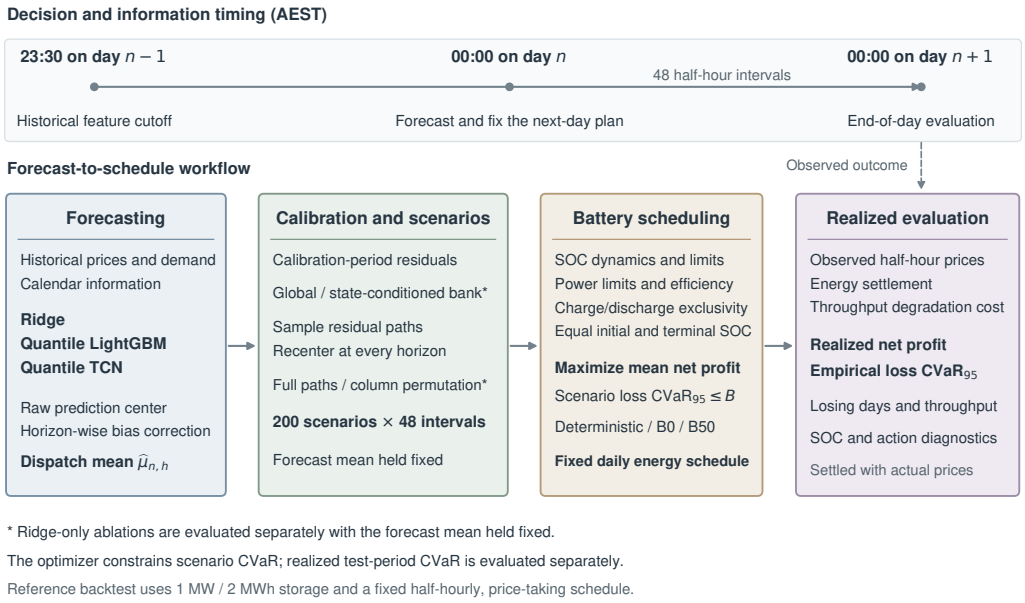


Figure 1. Forecasting, scenario construction, and risk-constrained storage scheduling with the daily information timeline. Historical features end 30 minutes before the midnight decision. A fixed schedule covers 48 half-hour intervals and is settled using realized prices. State conditioning and column permutation are separate Ridge-only ablations. The optimizer constrains scenario CVaR, while realized test-period CVaR is evaluated separately.

on those benchmark data.

$$\begin{aligned} &\boldsymbol{\pi}_n = (\pi_{n,1}, \ldots, \pi_{n,H})^\top, \\ &y_{n,h} = \operatorname{asinh}\left(\frac{\pi_{n,h}}{a_\pi}\right), \qquad v_{n,h} = \frac{D_{n,h}}{a_D}, \\ &\mathbf{x}_n = \Phi\left(\mathscr{F}_{\tau_n-\delta}, \mathbf{c}_n\right), \qquad \delta = \Delta t, \\ &a_\pi = 100\,\mathrm{AUD/MWh}, \qquad a_D = 10{,}000\,\mathrm{MW}. \end{aligned} \tag{2}$$

The forecasting comparison contains a regularized linear model, a tree-based quantile model, and a convolutional quantile model. Ridge regression fits separate horizon-specific linear relationships to transformed prices, with coefficient shrinkage controlling sensitivity to correlated predictors[20]. Its output is a vector of transformed point predictions. LightGBM uses gradient-boosted decision trees[21], with a separate model for each requested quantile. Training observations are pooled across forecast horizons, and a normalized horizon index is appended to the common daily feature vector. This construction allows each quantile model to learn horizon-dependent relationships without requiring an independent tree ensemble for every interval. The temporal convolutional network follows the causal, dilated convolutional architecture associated with sequence modeling[22]. A shared encoder processes the recent history and each weekly reference segment separately, and their terminal representations are combined with calendar features. A direct output layer produces all horizon-quantile predictions simultaneously. Residual connections support representation learning within each historical segment, while causal padding prevents convolutional access to later positions within that segment. Quantile models minimize the asymmetric quantile loss[23]. All model outputs refer to transformed prices before inversion. Model sizes, regularization settings, optimization parameters, and early-stopping procedures are reported in the experimental configuration rather than embedded in the mathematical definition.

$$\rho_q(e) = e\,(q - \mathbb{I}\{e < 0\}), \qquad e = y_{n,h} - \widehat{y}_{n,h}(q), \qquad q \in \mathscr{Q} \subset (0,1). \tag{3}$$

Fig. 1 summarizes the forecasting and scheduling workflow, including the information cutoff, scenario construction, fixed daily decisions, and settlement using realized prices.

The interface supplied to scheduling separates native prediction outputs from the scenario distribution used for risk evaluation. For Ridge, the raw center is obtained by applying the inverse price transformation to its point prediction. This inversion does not make the result an unbiased conditional mean on the original price scale. For LightGBM and the temporal

convolutional network, quantile outputs are ordered across quantile levels, and the inverse-transformed median defines the raw center. Their remaining inverse-transformed quantiles are retained as native probabilistic outputs. A horizon-specific additive correction is estimated from raw-center errors on a separate calibration period and applied to every subsequent forecast. The corrected center is denoted by $\widehat{\mu}_{n,h}^{(m)}$ and serves as the mean of the constructed scenario distribution, not as a guaranteed estimate of the true conditional expectation. Each forecasting model has its own calibration residuals, but all models use the same residual-sampling and centering protocol. Sampled residuals are recentered separately at each horizon, ensuring that the scenario average equals the corrected center. Native quantiles are evaluated separately from residual-derived forecast quantiles. This distinction prevents improvements in a model's native distribution from being incorrectly attributed to a scheduler that receives an empirical residual distribution centered on its corrected forecast.

$$\widehat{c}_{n,h}^{(m)} = \begin{cases} a_\pi \sinh\left(\widehat{y}_{n,h}^{(\mathrm{R})}\right), & m = \mathrm{R}, \\ a_\pi \sinh\left(\widehat{y}_{n,h}^{(m)}(0.5)\right), & m \in \{\mathrm{L},\mathrm{T}\}, \end{cases} \tag{4}$$

$$\begin{aligned} b_h^{(m)} &= \frac{1}{|\mathscr{C}|} \sum_{j \in \mathscr{C}} \left(\pi_{j,h} - \widehat{c}_{j,h}^{(m)}\right), \\ \widehat{\mu}_{n,h}^{(m)} &= \widehat{c}_{n,h}^{(m)} + b_h^{(m)}, \qquad m \in \{\mathrm{R},\mathrm{L},\mathrm{T}\}, \end{aligned} \tag{5}$$

2.2 Calibration and Price Scenario Construction

The calibration stage constructs empirical forecast-error trajectories on the original price scale. Let $\mathscr{C}$ denote the calibration days and let M denote their number, with all observations preceding the evaluation period. For model m, the horizon-specific correction defined above removes the average raw-center error over these days. The remaining residual for day j and horizon h is the realized price minus the corrected forecast center. Stacking daily residual vectors produces an M-by-H matrix whose rows retain the observed ordering of errors within each operating day. The unconditional path bank contains these complete residual trajectories without subdividing them by market state. Its full-sample residual mean is zero at every horizon because the correction is estimated from the same calibration observations. This property is an empirical centering identity, not evidence that future errors have zero conditional expectation. Calibration-window research motivates explicit reporting of the reference period used to characterize forecast uncertainty[24]. The reference period remains fixed during evaluation, and neither its residual matrix nor the horizon corrections are updated using realized test prices. This separation limits potential leakage through the uncertainty-construction stage.

$$\begin{aligned} r_{j,h}^{(m)} &= \pi_{j,h} - \widehat{\mu}_{j,h}^{(m)}, \qquad j \in \mathscr{C}, \\ \mathbf{R}^{(m)} &= \left[r_{j,h}^{(m)}\right]_{j \in \mathscr{C},\, h=1,\ldots,H}, \qquad M = |\mathscr{C}|, \\ \mathscr{B}_{\mathrm{G}}^{(m)} &= \left\{ \left(r_{j,1}^{(m)}, \ldots, r_{j,H}^{(m)}\right) \,\middle|\, j \in \mathscr{C} \right\}, \\ \frac{1}{M} \sum_{j \in \mathscr{C}} r_{j,h}^{(m)} &= 0, \qquad h = 1, \ldots, H. \end{aligned} \tag{6}$$

A finite scenario bank is obtained by sampling calibration-day indices uniformly with replacement. Each selected index supplies an entire residual row, so errors from different horizons remain paired as observed on the selected day. The sampled bank is constructed once for a given experimental configuration and reused throughout its evaluation period. Sampling can introduce a nonzero average even when the full residual matrix is centered. The sampled residuals are therefore recentered separately at every horizon before they are added to the corrected forecast. The resulting scenarios have equal weights, and their arithmetic mean exactly reproduces the forecast center at each time point. Repeated selection of a calibration day changes its representation within the finite bank but does not create additional independent historical observations. Increasing the scenario count consequently reduces sampling variability without expanding the underlying calibration information. Matching the scenario mean also does not guarantee marginal coverage, sharpness, or joint calibration, which must be assessed against subsequent observations[25]. The constructed distribution is an empirical approximation intended for controlled scheduling

comparisons, rather than a distribution-free guarantee concerning future prices or profits.

$$
\begin{aligned}
\iota_s &\overset{\text{iid}}{\sim} \text{Uniform}(\mathscr{C}), \qquad s = 1, \ldots, S, \\
e_{s,h}^{(\text{G},m)} &= r_{\iota_s,h}^{(m)} - \frac{1}{S}\sum_{\ell=1}^{S} r_{\iota_\ell,h}^{(m)}, \\
\widetilde{\pi}_{n,s,h}^{(\text{G},m)} &= \widehat{\mu}_{n,h}^{(m)} + e_{s,h}^{(\text{G},m)}, \\
w_s &= \frac{1}{S}, \qquad \sum_{s=1}^{S} w_s \widetilde{\pi}_{n,s,h}^{(\text{G},m)} = \widehat{\mu}_{n,h}^{(m)}.
\end{aligned}
\tag{7}
$$

The state-conditioned variant tests whether uncertainty representation benefits from conditioning on recent market volatility. For each operating day, a sample standard deviation is calculated from the 48-hour price-history window ending at the information cutoff. A threshold is estimated as the eightieth percentile of these volatility values within the training period. Days exceeding the threshold are assigned to the high-volatility state, while the remaining days form the ordinary state. Calibration trajectories are partitioned using their own preceding volatility histories, ensuring that realized volatility within a target day never determines its state. The scenario bank selected for an evaluation day is sampled from the calibration subset with the matching state. A subset containing fewer than 30 calibration days is replaced by the unconditional pool to limit extreme fragmentation. Each sampled state bank is recentered at every horizon using the same procedure as the unconditional bank. Conditioning therefore changes the empirical residual distribution but leaves the forecast center unchanged. The variant does not estimate a state-specific mean correction and is treated as a hypothesis to be tested, not as an established calibration improvement.

$$
\begin{aligned}
\overline{\pi}_n^{\text{hist}} &= \frac{1}{K}\sum_{k=1}^{K} \pi_{n,k}^{\text{hist}}, \qquad K = 96, \\
\sigma_n^{\text{hist}} &= \left[\frac{1}{K-1}\sum_{k=1}^{K}\left(\pi_{n,k}^{\text{hist}} - \overline{\pi}_n^{\text{hist}}\right)^2\right]^{1/2}, \\
\kappa &= \text{Quantile}_{0.8}\left(\left\{\sigma_j^{\text{hist}} \,\middle|\, j \in \mathscr{T}\right\}\right), \\
z_n &= \mathbb{I}\left\{\sigma_n^{\text{hist}} > \kappa\right\}.
\end{aligned}
\tag{8}
$$

$$
\begin{aligned}
\mathscr{C}_k &= \left\{j \in \mathscr{C} \,\middle|\, z_j = k\right\}, \qquad k \in \{0,1\}, \\
\mathscr{C}_k^{*} &= \begin{cases} \mathscr{C}_k, & |\mathscr{C}_k| \geq 30, \\ \mathscr{C}, & |\mathscr{C}_k| < 30, \end{cases} \\
\widetilde{\pi}_{n,s,h}^{(\text{SC})} &= \widehat{\mu}_{n,h}^{(\text{R})} + e_{s,h}^{(z_n,\text{R})}, \qquad \frac{1}{S}\sum_{s=1}^{S} e_{s,h}^{(z_n,\text{R})} = 0.
\end{aligned}
\tag{9}
$$

The dependence experiment modifies scenario pairing while preserving every univariate empirical distribution. Starting from the centered unconditional bank, an independent random permutation of scenario indices is generated for each horizon. Applying these permutations separately to the columns rearranges which residual values appear together in a daily trajectory. Every column retains exactly the same values and multiplicities, so its mean, variance, empirical quantiles, and coverage implications are unchanged. This distinction between marginal postprocessing and ensemble dependence also underlies ensemble copula coupling[26]. The permuted construction is described as an independent-coupling approximation because column permutations are selected independently, but a finite bank can retain nonzero empirical cross-horizon correlation. Correlations are therefore measured rather than assumed to vanish. The permutations are generated before evaluation and remain fixed across evaluation days for a given seed. The original and permuted banks use identical forecast centers, scenario counts, and probability weights. Differences in scheduling can consequently arise from changed joint price combinations and their effect on the distribution of daily profit. They cannot be attributed to improved marginal prediction accuracy, since the

marginal scenario distributions are identical.

$$
\begin{aligned}
e^{(\mathrm{I})}_{s,h} &= e^{(\mathrm{G,R})}_{\psi_h(s),h},\\
\widetilde{\pi}^{(\mathrm{I})}_{n,s,h} &= \widehat{\mu}^{(\mathrm{R})}_{n,h} + e^{(\mathrm{I})}_{s,h},\\
\widehat{F}^{(\mathrm{I})}_{n,h}(x) &= \frac{1}{S}\sum_{s=1}^{S} \mathbb{I}\left\{\widetilde{\pi}^{(\mathrm{I})}_{n,s,h} \le x\right\}\\
&= \frac{1}{S}\sum_{s=1}^{S} \mathbb{I}\left\{\widetilde{\pi}^{(\mathrm{G,R})}_{n,s,h} \le x\right\} = \widehat{F}^{(\mathrm{G,R})}_{n,h}(x), \qquad x \in \mathbb{R}.
\end{aligned}
\tag{10}
$$

The experimental contrasts are defined by which components remain fixed and which are allowed to change. Comparisons between forecasting models use a calibration and sampling protocol, although their forecast centers and residual values differ because their prediction errors differ. State conditioning holds the Ridge forecast center constant while changing the residual pool, and therefore changes marginal uncertainty as well as the sampled dependence structure. It should not be interpreted as an isolated dependence intervention. The column-permutation comparison instead holds the empirical marginal distributions constant and changes only their joint pairing. Both mechanism experiments retain the battery constraints, operating costs, confidence level, and selected risk budget of their reference policy. For any fixed battery schedule, equal scenario means imply equal expected linear trading revenues, while the joint pairing can change estimated tail losses. Fixed-action risk evaluation is therefore performed alongside comparisons of independently optimized schedules. The mechanism experiments concern the Ridge setting, whereas the common forecasting interface is used in model comparisons. Parameter and regional extensions evaluate the reference scheduling framework without implying that every calibration variant has been externally validated.

2.3 Battery Constraints and CVaR Formulation

Battery operation is represented by stored energy at the boundaries of each scheduling interval. Charging power increases stored energy after accounting for charging efficiency, while discharging reduces stored energy by the energy delivered to the grid divided by discharging efficiency. The energy trajectory is bounded by fractions of nominal capacity, and charging and discharging powers have separate upper limits. A binary variable indicates whether charging is permitted, preventing simultaneous charging and discharging within an interval. This exclusivity is important when negative prices and efficiency losses could otherwise create artificial trading opportunities. Initial and terminal stored energy are fixed at the same prescribed fraction of capacity for every operating day. The daily boundary condition removes differences in terminal inventory value from policy comparisons and prevents schedules from earning apparent profits by exhausting the battery. Charging and discharging efficiencies are set to the square root of the specified round-trip efficiency. The reference configuration uses a 1 MW power rating, 2 MWh nominal capacity, energy limits between 10 percent and 90 percent, and an initial energy fraction of 50 percent. Parameter experiments change the relevant physical settings while retaining the same constraint structure. The formulation does not include self-discharge or intraday schedule revisions.

$$
E_{n,h} = E_{n,h-1} + \eta_{\mathrm{ch}} p^{\mathrm{ch}}_{n,h}\Delta t - \frac{p^{\mathrm{dis}}_{n,h}\Delta t}{\eta_{\mathrm{dis}}}, \qquad h = 1,\dots,H.
\tag{11}
$$

$$
\begin{aligned}
\underline{s}E^{\mathrm{cap}} \le E_{n,h} \le \overline{s}E^{\mathrm{cap}}, &\qquad h = 0,\dots,H,\\
0 \le p^{\mathrm{ch}}_{n,h} \le \gamma_{n,h}\overline{P}_{\mathrm{ch}},&\\
0 \le p^{\mathrm{dis}}_{n,h} \le (1-\gamma_{n,h})\overline{P}_{\mathrm{dis}},&\\
\gamma_{n,h} \in \{0,1\}, &\qquad h = 1,\dots,H,\\
E_{n,0} = E_{n,H} = s_0 E^{\mathrm{cap}},&\\
\eta_{\mathrm{ch}} = \eta_{\mathrm{dis}} = \sqrt{\eta_{\mathrm{rt}}},&\\
\underline{s} = 0.1, \qquad \overline{s} = 0.9, &\qquad s_0 = 0.5.
\end{aligned}
\tag{12}
$$

Scenario profit is defined as electricity sales revenue minus electricity purchase expenditure and a throughput-based degradation charge. Net grid exchange equals discharging power minus charging power, so negative electricity prices can produce positive charging revenue without requiring a change in the accounting convention. The degradation term applies to both charging and discharging energy and therefore penalizes total grid-side throughput. This linear operating-cost representation supports consistent comparison across schedules but does not reproduce detailed battery aging, cycle-depth effects, or calendar

degradation. Reviews of degradation implementation emphasize the variety of representations available for storage operation[27]. The present analysis accordingly interprets this coefficient as a simplified cost-of-use assumption and tests its sensitivity rather than claiming an electrochemical lifetime model. The objective maximizes the weighted average of net operating profit across the price scenarios. Equal scenario weights are used throughout, and the centering procedure ensures that expected linear trading revenue depends only on the corrected forecast center for any fixed schedule. Differences between uncertainty variants with the same center therefore enter through risk feasibility rather than through a changed expected-price objective. Capital expenditure, network charges, taxes, and ancillary-service income are excluded. Reported profitability consequently concerns the defined arbitrage operation rather than complete investment viability.

$$
\begin{aligned}
T_n(\mathbf{u}_n) &= \Delta t \sum_{h=1}^{H} \left(p_{n,h}^{\mathrm{ch}} + p_{n,h}^{\mathrm{dis}} \right), \\
\Pi_{n,s}(\mathbf{u}_n) &= \Delta t \sum_{h=1}^{H} \widetilde{\pi}_{n,s,h} \left(p_{n,h}^{\mathrm{dis}} - p_{n,h}^{\mathrm{ch}} \right) - c_{\mathrm{deg}} T_n(\mathbf{u}_n), \\
\overline{\Pi}_n(\mathbf{u}_n) &= \sum_{s=1}^{S} w_s \Pi_{n,s}(\mathbf{u}_n) \\
&= \Delta t \sum_{h=1}^{H} \widehat{\mu}_{n,h} \left(p_{n,h}^{\mathrm{dis}} - p_{n,h}^{\mathrm{ch}} \right) - c_{\mathrm{deg}} T_n(\mathbf{u}_n).
\end{aligned}
\tag{13}
$$

Financial risk is formulated using signed loss, defined as the negative of net scenario profit. Profitable scenarios therefore have negative losses, and losses are not truncated at zero before risk calculation. CVaR at a confidence level of 0.95 summarizes the adverse tail of this loss distribution. Its optimization representation introduces an unrestricted threshold variable and nonnegative auxiliary variables for losses exceeding that threshold[28]. The general-loss formulation permits the same definition for empirical discrete distributions and does not require normally distributed price errors[29]. The resulting linear inequalities impose an upper bound B on the scenario-based daily CVaR while preserving the expected-profit objective. This budget is a monetary risk tolerance per operating day, not a maximum allowable loss in every scenario. Larger budgets relax the feasible risk restriction for a fixed scenario bank, but realized revenue and realized tail losses need not vary monotonically with that relaxation. The deterministic reference removes the risk constraint and optimizes net profit against the corrected forecast trajectory. Because the scenarios are centered, its objective coincides with the unconstrained scenario-average objective under identical battery constraints. Differences between deterministic and risk-constrained schedules can thus be attributed to the imposed risk restriction within each fixed forecasting and scenario configuration.

$$
\begin{aligned}
L_{n,s}(\mathbf{u}_n) &= -\Pi_{n,s}(\mathbf{u}_n), \qquad \alpha = 0.95, \\
\mathrm{CVaR}_{\alpha}^{\mathrm{scen}}(\mathbf{u}_n) &= \min_{\vartheta_n \in \mathbb{R}} \left[\vartheta_n + \frac{1}{1-\alpha} \sum_{s=1}^{S} w_s \left[L_{n,s}(\mathbf{u}_n) - \vartheta_n \right]_+ \right], \\
[x]_+ &= \max(x, 0).
\end{aligned}
\tag{14}
$$

$$
\begin{aligned}
\max_{\mathbf{u}_n, \vartheta_n, \zeta_n} \quad & \overline{\Pi}_n(\mathbf{u}_n) \\
\text{subject to} \quad & \mathbf{u}_n \in \mathscr{U}, \\
& \zeta_{n,s} \geq L_{n,s}(\mathbf{u}_n) - \vartheta_n, \qquad s = 1, \ldots, S, \\
& \zeta_{n,s} \geq 0, \qquad s = 1, \ldots, S, \\
& \vartheta_n + \frac{1}{1-\alpha} \sum_{s=1}^{S} w_s \zeta_{n,s} \leq B, \\
& \vartheta_n \in \mathbb{R},
\end{aligned}
\tag{15}
$$

The physical constraints and CVaR inequalities form a mixed-integer linear program whose binary decisions enforce charging exclusivity. Each operating day is solved independently using the forecast center and the scenario bank available under the experimental protocol. HiGHS supplies the optimization backend through the SciPy interface, with revised-simplex methodology documented in the solver literature[30]. Solver settings and termination criteria are reported with the experimental configuration. Schedules are checked against energy dynamics, capacity bounds, power limits, charging exclusivity, terminal energy, and the computed scenario-risk budget where applicable. Failed solves are treated as execution errors rather than omitted

trading days. The accepted schedule is then evaluated using realized prices, with the same efficiency and throughput-cost assumptions retained. Daily realized profits are accumulated for profitability evaluation, while realized losses across the test days define a separate empirical CVaR statistic. This statistic measures the historical tail of realized daily outcomes and is not constrained by the optimization problem. Satisfying the scenario budget therefore does not establish that test-period CVaR will satisfy the same numerical bound. The comparison reports both quantities to expose distributional mismatch, alongside loss frequency, worst-day outcomes, and throughput. No realized test price is used to revise the schedule before evaluating its performance.

$$
\begin{aligned}
\Pi_n^{\mathrm{real}} &= \Delta t \sum_{h=1}^{H} \pi_{n,h} \left(\widehat{p}_{n,h}^{\mathrm{dis}} - \widehat{p}_{n,h}^{\mathrm{ch}} \right) - c_{\mathrm{deg}} T_n(\widehat{\mathbf{u}}_n), \\
L_n^{\mathrm{real}} &= -\Pi_n^{\mathrm{real}}, \\
\widehat{\mathrm{CVaR}}_{\alpha}^{\mathrm{test}} &= \min_{v \in \mathbb{R}} \left[v + \frac{1}{(1-\alpha)N} \sum_{n \in \mathscr{D}_{\mathrm{test}}} \left[L_n^{\mathrm{real}} - v \right]_+ \right], \\
N &= |\mathscr{D}_{\mathrm{test}}|.
\end{aligned}
\tag{16}
$$

3 Data and Experimental Design

3.1 Datasets and Chronological Protocol

The primary data source is the Australian Energy Market Operator's public historical archive for the National Electricity Market[31]. Monthly files covering January 2022 through December 2025 provide regional prices and demand at five-minute resolution. Prices are extracted from the `RRP` field of `DISPATCHPRICE`, while regional demand is extracted from the `TOTALDEMAND` field of `DISPATCHREGIONSUM`. Records are selected using the region identifier and an `INTERVENTION` value of zero, retaining the non-intervention series for the defined backtest. NSW1 provides the development setting for forecasting comparisons, scheduling analysis, loss diagnostics, and mechanism experiments. VIC1 and SA1 provide regional extensions under the same data-processing and chronological protocols. Each regional series contains 420,768 five-minute intervals over the four-year period, with price and demand observations aligned on a common time index. Negative prices and extreme positive prices are retained rather than removed as outliers, because they represent financially relevant operating conditions. Monetary calculations use prices in AUD per MWh, and demand is represented in MW. Regional forecasting models are trained on their own historical observations, while model configurations are retained from the primary setting. The extensions consequently examine performance under comparable procedures across regions, rather than unchanged model weights transferred directly from NSW1 to other markets.

Australian timestamps are interpreted in fixed Australian Eastern Standard Time, corresponding to UTC plus ten hours without daylight-saving adjustment. Dispatch timestamps identify interval endings and are shifted backward by five minutes to obtain interval starts. Price and demand records are aligned, and any duplicate timestamps are resolved using the latest `LASTCHANGED` value. Half-hour prices and demand are arithmetic averages of six constituent observations, with complete intervals required and no additional interpolation applied. The extracted regional series contain no missing price or demand observations on the expected five-minute grid. Forecast examples begin after a fourteen-day history warmup. Eligible days in 2022 and 2023 form the training set, January through June 2024 forms validation, July through December 30, 2024 forms calibration, and 2025 forms testing. These partitions contain 716, 182, 183, and 365 daily examples, respectively. December 31, 2024 is excluded from fitting and calibration, although its observations can enter subsequent historical inputs. The files represent final historical archives, not reconstructed data releases available at each forecast origin. Historical publication versions and reporting delays have not been audited, so the nominal feature cutoff does not establish operational availability. Dataset roles and chronological partitions, including all retained daily counts, are summarized in Table 1.

Supplementary forecasting validation uses the German and PJM COMED price series distributed with the open-access electricity price forecasting benchmark[32,33]. The German series spans January 9, 2012 through December 31, 2017, while the PJM series spans January 1, 2013 through December 24, 2018. Their supplied hourly indices and repository preprocessing are retained, and the prediction horizon contains 24 hourly prices. Only historical prices and calendar features are used, excluding the external load and renewable forecasts included in the source files. For Germany, eligible observations before 2015 form training, the first half of 2015 forms validation, and July through December 30, 2015 forms calibration. Testing covers 2016 and 2017, comprising 731 days. For PJM, the corresponding partition year is 2016, with training before that year, validation in its first half, and calibration ending December 30. Testing covers January 2017 through December 24, 2018, comprising 723 days. The fourteen-day history warmup and separation of calibration from evaluation are maintained. These experiments evaluate locally retrained forecasting models under a controlled historical price-only protocol. They neither reproduce the complete benchmark's original modeling configuration nor evaluate storage arbitrage in these markets. Results are compared

Table 1. Chronological data partitions and daily forecasting horizons. Dates and day counts refer to valid modeling samples after the 14-day history warm-up.

Dataset	Time step	H	Training	Validation	Calibration	Test
AEMO NSW1	30 min	48	2022-01-15 to 2023-12-31 (716 days)	2024-01-01 to 2024-06-30 (182 days)	2024-07-01 to 2024-12-30 (183 days)	2025-01-01 to 2025-12-31 (365 days)
AEMO VIC1	30 min	48	2022-01-15 to 2023-12-31 (716 days)	2024-01-01 to 2024-06-30 (182 days)	2024-07-01 to 2024-12-30 (183 days)	2025-01-01 to 2025-12-31 (365 days)
AEMO SA1	30 min	48	2022-01-15 to 2023-12-31 (716 days)	2024-01-01 to 2024-06-30 (182 days)	2024-07-01 to 2024-12-30 (183 days)	2025-01-01 to 2025-12-31 (365 days)
EPF DE	1 h	24	2012-01-23 to 2014-12-31 (1074 days)	2015-01-01 to 2015-06-30 (181 days)	2015-07-01 to 2015-12-30 (183 days)	2016-01-01 to 2017-12-31 (731 days)
EPF PJM	1 h	24	2013-01-15 to 2015-12-31 (1081 days)	2016-01-01 to 2016-06-30 (182 days)	2016-07-01 to 2016-12-30 (183 days)	2017-01-01 to 2018-12-24 (723 days)

H is the number of forecast intervals in a complete next day. NSW1 is the primary scheduling dataset; VIC1 and SA1 provide regional scheduling validation. AEMO inputs use RRP from DISPATCHPRICE and TOTALDEMAND from DISPATCHREGIONSUM, with six complete five-minute observations averaged into each half-hour interval. Times use AEST without daylight saving adjustment. EPF DE and PJM retain the repository-supplied hourly preprocessing and evaluate forecasting only. AEMO predictors use historical prices, demand, and calendar variables, while EPF predictors use prices and calendar variables without the supplied exogenous forecast fields. December 31 of each calibration year is excluded from fitting and calibration but can contribute to subsequent historical inputs. Final AEMO archives are used; a historical publication-version audit remains incomplete.

within each market without equating native price units across different monetary environments.

3.2 Compared Methods and Implementation

Experiments use Python 3.12.14, scikit-learn 1.8.0[34], LightGBM 4.7.0, SciPy 1.17.0, and CPU-based PyTorch 2.8.0. LightGBM uses two configured worker threads, while TCN training uses two intra-operation threads and one inter-operation thread. Ridge selects a penalty of 1000 from candidates 1, 10, 100, and 1000 using validation MAE, with feature standardization fitted on training observations. LightGBM selects 31 leaves over 15 using original-scale validation Pinball loss. Each quantile model permits 400 boosting iterations, with learning rate 0.05, minimum leaf sample count 50, L_2 coefficient 1, and early-stopping patience 30 under its transformed-scale loss. TCN selects 32 channels over 16 using original-scale validation Pinball loss. Its five residual blocks use dilations 1, 2, 4, 8, and 16, two kernel-three convolutions per block, and dropout 0.1. Training uses Adam[35], learning rate 0.001, weight decay 0.0001, batch size 32, gradient-norm clipping at 1, and at most 80 epochs. Early stopping uses patience 12 and a minimum run of 15 epochs, retaining the best observed validation checkpoint. Nine quantile levels are used for both quantile predictors. The primary seed is 1507. Forecast parameters remain fixed during testing. Model settings and comparison roles are summarized in Table 2. These configurations support reproducibility without implying exhaustive hyperparameter optimization.

$$\mathscr{Q} = \{0.05, 0.10, 0.20, 0.30, 0.50, 0.70, 0.80, 0.90, 0.95\}. \tag{17}$$

Table 2. Forecasting, scenario construction, scheduling, and comparison configurations. Predictor configurations are selected using NSW1 validation data and frozen for locally retrained external models.

Component	Configuration
A. Forecasting models and statistical targets	
Input and target processing	Target transformation $y = \operatorname{asinh}(\pi/100)$ and inverse transformation to the original price scale. Ridge input scaling and TCN channel scaling use training data only.
Ridge	Regularization $\lambda = 1000$, selected in NSW1 from $\{1, 10, 100, 1000\}$ using original-scale validation MAE.
Quantile LightGBM	31 leaves; learning rate 0.05; at most 400 trees; minimum child samples 50; L_2 regularization 1; early stopping patience 30 rounds. NSW1 candidates have 15 or 31 leaves and are selected by original-scale validation pinball loss.
Quantile TCN	32 channels; five residual blocks with dilations $\{1, 2, 4, 8, 16\}$ and two kernel-3 causal convolutions per block; dropout 0.1; direct $H \times 9$ output. Adam uses learning rate 10^{-3}, weight decay 10^{-4}, batch size 32, and gradient clipping at norm 1. Training uses at most 80 epochs, at least 15 epochs, and patience 12. NSW1 candidates have 16 or 32 channels and are selected by original-scale validation pinball loss.
Probability outputs	Nine quantiles $\{0.05, 0.10, 0.20, 0.30, 0.50, 0.70, 0.80, 0.90, 0.95\}$. Native quantiles are evaluated separately from the common residual scenario interface; raw centers and corrected dispatch means are distinct targets.
B. Reference scenarios and scheduling	
Common scenario interface	Horizon-wise calibration bias correction followed by 200 complete residual paths sampled with replacement and recentered at each horizon. Seed 1507; residual banks remain fixed across test dates.
Reference battery	Power 1 MW; capacity 2 MWh; round-trip efficiency 0.90; charging and discharging efficiencies $\sqrt{0.90}$; SOC limits 10%–90%; initial and terminal SOC 50%; throughput degradation charge 5 AUD/MWh.
Decision protocol	Midnight decision with historical inputs cut off one interval earlier. A fixed next-day charging/discharging plan is settled at realized prices. Deterministic, B0, and B50 policies share the battery constraints.
CVaR scheduling	Maximize scenario mean net profit subject to daily signed loss CVaR at confidence level 0.95 not exceeding B. B0 and B50 use budgets of 0 and 50 AUD, respectively.
C. Ablations, sensitivity, and validation	
State calibration	Ridge only. States use the preceding 48-hour price standard deviation and the training-period 80th percentile threshold, 86.2812 AUD/MWh. State banks retain the forecast mean and revert to the global bank below 30 calibration days; budgets B0 and B50.
Dependence ablation	Ridge only. Independently permute the 200 scenario rows within each horizon, preserving empirical marginal values and multiplicities. Main seed 1507 uses B0 and B50; seeds 21 and 42 repeat B50.
Risk budget grid	All three predictors use $B \in \{0, 10, 25, 50, 100, 200, 500\}$ AUD per day with unchanged scenario banks.
Battery sensitivity	Ridge only. Capacity $\{1, 2, 4\}$ MWh with fixed 1 MW power; round-trip efficiency $\{0.85, 0.90, 0.95\}$; throughput charge $\{0, 5, 10, 20\}$ AUD/MWh. One factor varies at a time. Capacity-normalized comparisons use B25 for 1 MWh and B100 for 4 MWh.
Scenario count	Ridge only. $S \in \{50, 100, 200, 500, 1000\}$ with seed 1507, common sampling prefixes, and separate horizon recentering at each count.
Additional TCN seeds	Seeds 21, 42, 2026, and 999 evaluate prediction sensitivity only. Main economic comparisons use seed 1507.
External validation	VIC1, SA1, DE, and PJM use local retraining with frozen NSW1 configurations and local early stopping. AEMO uses $H = 48$; EPF uses $H = 24$ and performs no storage scheduling.
Statistical comparison	Date-paired seven-day circular block bootstrap with 2000 resamples and pointwise 95% percentile intervals. Comparisons are exploratory and are not adjusted for multiplicity.

The price transformation scale of 100 uses the native price units of each dataset. Price and demand histories comprise the preceding two days and corresponding days one and two weeks earlier. State and dependence variants are separate ablations. MILP solves use SciPy/HiGHS, relative gap 10^{-6}, and time limits of 15 s for NSW1 reference and ablation solves and 45 s for battery sensitivity and regional solves. Independent audits verify SOC dynamics, operating limits, charge/discharge exclusivity, terminal SOC, and scenario risk feasibility. The recorded CPU environment uses Python 3.12.14, scikit-learn 1.8.0, SciPy 1.17.0, LightGBM 4.7.0, and PyTorch 2.8.0. LightGBM uses two threads; TCN uses two intra-operation threads and one inter-operation thread.

The main scheduling comparison crosses three forecasting models with deterministic scheduling and CVaR budgets of 0 and 50 AUD per day. All policies use the same reference battery, a 95 percent risk confidence level, and the model-specific corrected forecast center. Risk-constrained policies use 200 centered unconditional residual paths. The state-conditioning comparison replaces the Ridge unconditional bank with the matching state bank at both budgets, while retaining its forecast center. The dependence comparison permutes the Ridge unconditional bank columns at those budgets, preserving every empirical marginal distribution. These interventions are evaluated separately rather than combined into a factorial experiment. Additional permutation seeds 21 and 42 are evaluated at budget 50, whereas additional TCN seeds 21, 42, 2026, and 999 assess prediction sensitivity only. The primary scheduling results retain the TCN trained with seed 1507. Auxiliary scaled-residual interfaces are excluded from the main comparisons. MILP solves use relative optimality gap 10^{-6} and time limits of 15 seconds for NSW1 comparisons or 45 seconds for the parameter and regional extensions. Tighter feasibility tolerances of 10^{-9} are applied in the stronger-model, budget, and external-validation solvers. Independent checks evaluate returned actions and scenario risks, so solver tolerances are not treated as substitutes for verifying feasibility.

Risk-budget sensitivity evaluates 0, 10, 25, 50, 100, 200, and 500 AUD per day for all three forecasting models, retaining their original scenario banks. The deterministic reference remains unconstrained. Physical-parameter and scenario-count scans use Ridge only, changing one factor at a time from the reference configuration. Nominal capacities are 1, 2, and 4 MWh with power fixed at 1 MW. Round-trip efficiencies are 0.85, 0.90, and 0.95, while throughput charges are 0, 5, 10, and 20 AUD per MWh. The physical reference uses efficiency 0.90, charge 5, and 200 scenarios, with budget 50. Capacity-normalized comparisons use budgets 25 and 100 for capacities 1 and 4 MWh, maintaining a daily budget of 25 AUD per MWh of nominal capacity. Scenario counts are 50, 100, 200, 500, and 1000, using a common seeded index-stream prefix and separate recentering at each count. VIC1 and SA1 evaluate locally retrained models under deterministic scheduling and budgets 0 and 50. Regional and EPF extensions retain the selected NSW1 model settings, refit weights locally, and adapt output dimensions to the prediction horizon. Local validation controls early stopping without selecting configurations from test outcomes. These extensions therefore assess repeated procedures rather than zero-shot transfer of fitted model weights.

3.3 Evaluation and Statistical Analysis

Point prediction is evaluated on the original price scale using mean absolute error and root mean squared error. MAE is reported separately for raw centers and corrected forecast means, while the main RMSE comparison refers to corrected means. Probabilistic evaluation uses average Pinball loss, consistent with proper scoring principles[36], together with empirical coverage, mean width, and interval score for 90 percent prediction intervals. Native quantiles and residual-derived quantiles are evaluated separately. The main forecasting comparison uses nine quantile levels, whereas state-calibration diagnostics use five. Pinball averages are therefore compared only within a common quantile grid. Coverage is also evaluated within the historical volatility states defined before scheduling. Financial evaluation reports total realized net profit, empirical daily-loss CVaR95, loss-day count, worst-day profit, and total grid-side throughput. Realized CVaR is calculated from the complete signed-loss distribution rather than positive losses alone. Loss-day counts use a numerical tolerance of 10^{-5} AUD to exclude negligible negative values caused by roundoff. This counting tolerance does not alter profits or the losses entering CVaR. Scenario-based CVaR is reported independently from realized CVaR, preserving the distinction between optimization feasibility and observed risk. All daily summaries retain the original evaluation dates and the same units within each market.

$$
\begin{aligned}
\mathrm{MAE} &= \frac{1}{NH} \sum_{n \in \mathscr{D}_{\mathrm{test}}} \sum_{h=1}^{H} \left| \pi_{n,h} - \widehat{\pi}_{n,h} \right|, \\
\mathrm{RMSE} &= \left[\frac{1}{NH} \sum_{n \in \mathscr{D}_{\mathrm{test}}} \sum_{h=1}^{H} \left(\pi_{n,h} - \widehat{\pi}_{n,h} \right)^2 \right]^{1/2}, \\
\mathrm{PB} &= \frac{1}{NH|\mathscr{Q}_{\mathrm{eval}}|} \sum_{n \in \mathscr{D}_{\mathrm{test}}} \sum_{h=1}^{H} \sum_{q \in \mathscr{Q}_{\mathrm{eval}}} \rho_q \left(\pi_{n,h} - \widehat{q}_{n,h}(q) \right), \\
\mathscr{Q}_{\mathrm{SC}} &= \{0.05, 0.10, 0.50, 0.90, 0.95\}.
\end{aligned}
\tag{18}
$$

$$
\begin{aligned}
l_{n,h} &= \widehat{q}_{n,h}(0.05), \qquad u_{n,h} = \widehat{q}_{n,h}(0.95), \qquad \beta = 0.10, \\
\mathrm{Coverage}_{90} &= \frac{1}{NH} \sum_{n \in \mathscr{D}_{\mathrm{test}}} \sum_{h=1}^{H} \mathbb{I}\left\{ l_{n,h} \le \pi_{n,h} \le u_{n,h} \right\}, \\
\mathrm{Width}_{90} &= \frac{1}{NH} \sum_{n \in \mathscr{D}_{\mathrm{test}}} \sum_{h=1}^{H} \left(u_{n,h} - l_{n,h} \right), \\
\mathrm{IS}_{90} &= \frac{1}{NH} \sum_{n \in \mathscr{D}_{\mathrm{test}}} \sum_{h=1}^{H} \left[u_{n,h} - l_{n,h} + \frac{2}{\beta} \left(l_{n,h} - \pi_{n,h} \right)_+ + \frac{2}{\beta} \left(\pi_{n,h} - u_{n,h} \right)_+ \right].
\end{aligned}
\tag{19}
$$

$$
\begin{aligned}
P_{\mathrm{test}} &= \sum_{n \in \mathscr{D}_{\mathrm{test}}} \Pi_n^{\mathrm{real}}, \\
N_{\mathrm{loss}} &= \sum_{n \in \mathscr{D}_{\mathrm{test}}} \mathbb{I}\left\{ \Pi_n^{\mathrm{real}} < -\varepsilon_{\mathrm{num}} \right\}, \qquad \varepsilon_{\mathrm{num}} = 10^{-5}\,\mathrm{AUD}, \\
P_{\mathrm{worst}} &= \min_{n \in \mathscr{D}_{\mathrm{test}}} \Pi_n^{\mathrm{real}}, \\
T_{\mathrm{test}} &= \sum_{n \in \mathscr{D}_{\mathrm{test}}} T_n(\widehat{\mathbf{u}}_n).
\end{aligned}
\tag{20}
$$

Statistical comparisons are paired by operating date so that candidate and reference policies face the same observed market conditions. Uncertainty is assessed using 2000 circular block bootstrap replicates with a fixed block length of seven days, following the block-resampling approach for dependent observations[37]. Block starting positions are sampled uniformly, consecutive dates wrap around the evaluation-period boundary, and concatenated blocks are truncated to the original number of days. The same sampled indices are applied to both members of every comparison. Forecast-score differences are averaged across sampled days, while total-profit differences are summed. CVaR differences are calculated by recomputing each policy's tail statistic within every replicate, rather than averaging daily risk estimates. Confidence limits are the 2.5 and 97.5 percentiles of the replicate differences. Differences are defined as candidate minus reference, making positive profit differences favorable and negative Pinball or CVaR differences favorable. Forecast models, calibration banks, and operating schedules are not refitted during resampling. State-specific coverage comparisons retain the resampled state labels, and seed-specific permutation comparisons preserve their own paired daily outcomes. These intervals quantify uncertainty conditional on the completed experiment.

$$
\begin{aligned}
\Delta_P^{(b)} &= \sum_{i \in \mathscr{I}^{(b)}} \left(\Pi_{i,\mathrm{cand}}^{\mathrm{real}} - \Pi_{i,\mathrm{ref}}^{\mathrm{real}} \right), \\
\Delta_{\mathrm{CVaR}}^{(b)} &= \widehat{\mathrm{CVaR}}_\alpha \left(\left\{ L_{i,\mathrm{cand}}^{\mathrm{real}} \right\}_{i \in \mathscr{I}^{(b)}} \right) - \widehat{\mathrm{CVaR}}_\alpha \left(\left\{ L_{i,\mathrm{ref}}^{\mathrm{real}} \right\}_{i \in \mathscr{I}^{(b)}} \right), \\
\mathrm{CI}_{95}(\Delta) &= \left[\mathrm{Quantile}_{0.025} \left(\left\{ \Delta^{(b)} \right\}_{b=1}^{2000} \right), \mathrm{Quantile}_{0.975} \left(\left\{ \Delta^{(b)} \right\}_{b=1}^{2000} \right) \right],
\end{aligned}
\tag{21}
$$

The analyses are exploratory because the NSW1 test year was repeatedly inspected during loss diagnosis and subsequent mechanism development. Confidence intervals are pointwise for individual comparisons and are not adjusted to provide simultaneous coverage across budgets, models, or sensitivity settings. Limited training and permutation seeds characterize selected sources of variability without constituting independent market replications. Scenario-count experiments use one sampling stream and cannot establish general convergence. NSW1, VIC1, and SA1 share an interconnected market and contemporaneous shocks, limiting their interpretation as independent confirmations. EPF evaluations provide forecasting evidence rather than financial validation. Intervals containing zero indicate unresolved differences, not demonstrated equivalence. These boundaries are retained when interpreting apparent advantages and identifying requirements for new-year validation under a frozen analysis protocol.

4 Results

4.1 Forecasting Performance and Calibration

Table 3 reports forecasting results for the 365 NSW1 test days using the chronological protocol. The raw prediction centers yield mean absolute errors of 50.30, 50.44, and 50.74 AUD/MWh for Ridge, LightGBM, and TCN, respectively. These values

Table 3. NSW1 forecasting performance over the 365 test days in 2025. Point errors distinguish the raw prediction center from the corrected dispatch mean. Probability metrics use the common global residual scenario interface.

Model	Raw center MAE	Dispatch mean MAE	Dispatch mean RMSE	Calibrated pinball loss	90% coverage (%)	90% interval width	90% interval score
Ridge	50.30	75.08	378.20	22.6204	88.78	271.52	692.86
LightGBM	50.44	77.13	379.76	22.3477	88.53	273.02	673.16
TCN	50.74	77.84	379.27	22.4257	88.30	273.88	679.26

All metrics except coverage are expressed in AUD/MWh. Calibrated pinball loss is averaged over the nine quantiles $\{0.05, 0.10, 0.20, 0.30, 0.50, 0.70, 0.80, 0.90, 0.95\}$. Raw centers are uncorrected model outputs, whereas dispatch means include the horizon-wise calibration bias correction. Native quantile outputs are evaluated separately. LightGBM has native pinball loss of 19.8680 and coverage of 83.32%, while TCN has native pinball loss of 19.5380 and coverage of 83.05%.

describe the uncorrected model outputs and must be distinguished from the means supplied to the scheduling optimization. After the calibration bias correction, the corresponding dispatch mean errors increase to 75.08, 77.13, and 77.84 AUD/MWh. This separation matters because the correction estimates an average calibration residual at each horizon rather than minimizing absolute prediction error in the test period. The corrected means have root mean squared errors of 378.20, 379.76, and 379.27 AUD/MWh, indicating that squared error is strongly influenced by relatively large deviations. The common residual scenario interface produces average pinball losses of 22.6204, 22.3477, and 22.4257, with empirical coverage of 88.78%, 88.53%, and 88.30% for nominal 90% intervals. Ridge therefore has the lowest raw center error and the highest aggregate coverage, whereas LightGBM has the lowest calibrated pinball loss. The rankings already differ across metrics, providing no single forecasting winner that can be inferred from all measures simultaneously. Each comparison uses identical dates and evaluation horizons to avoid differences caused by sample selection.

The stronger forecasting baselines provide modest improvements in calibrated distributional accuracy relative to Ridge. LightGBM reduces pinball loss by 0.2726, corresponding to approximately 1.21%, and its paired bootstrap 95% interval for the difference is $[-0.5015, -0.0645]$. TCN reduces the same score by 0.1947, approximately 0.86%, but its interval of $[-0.4938, 0.0507]$ includes zero. These intervals support a directional improvement for LightGBM under the specified resampling protocol while leaving the TCN comparison uncertain. They do not establish performance superiority across alternative test years or model configurations. The raw center errors increase by 0.14 and 0.44 AUD/MWh for LightGBM and TCN, and their corrected mean errors increase by 2.05 and 2.76 AUD/MWh. Distributional gains thus coexist with slightly weaker point predictions in this test sample. Average 90% interval widths are 271.52, 273.02, and 273.88 AUD/MWh for Ridge, LightGBM, and TCN. Their corresponding interval scores are 692.86, 673.16, and 679.26, indicating that the stronger baselines improve this combined measure of width and exceedance penalties despite slightly wider intervals. The evidence favors a limited improvement in probability scores rather than a broad improvement in every forecasting dimension. The confidence intervals remain exploratory, pointwise comparisons conditional on the fitted models and the observed test sequence used here.

Fig. 2 separates the 302 normal volatility days from the 63 high volatility days identified using information available before each scheduling decision. Ridge achieves 90.02% interval coverage in the normal state but only 82.84% in the high volatility state. LightGBM exhibits corresponding coverage of 89.91% and 81.88%, while TCN achieves 89.60% and 82.08%. The aggregate coverage values therefore conceal a substantial conditional deficit shared by all three forecasting interfaces. For Ridge, lower and upper interval exceedances are 2.19% and 7.79% during normal conditions, compared with 7.87% and 9.29% during high volatility conditions. Both tails exceed their nominal 5% frequencies in the high state. LightGBM also exhibits breaches in both tails, with lower and upper rates of 9.89% and 8.23%, whereas the corresponding TCN rates are 8.86% and 9.06%. These patterns do not support attributing conditional undercoverage exclusively to unexpectedly high prices. The unconditional residual banks generate the same average interval width in both states for each model. High volatility interval scores consequently increase to approximately 1193–1226, compared with 565–582 under normal conditions. The deterioration reflects the interaction between the realized price distribution and the fixed residual interface, motivating the subsequent controlled evaluation of state conditioning without presuming that conditioning improves decisions.

The native quantile outputs and the residual scenarios represent different probability interfaces and require interpretation. LightGBM and TCN achieve native pinball losses of 19.8680 and 19.5380, which are lower than their residual scenario scores on the nine quantiles. Their native 90% coverage is nevertheless only 83.32% and 83.05%, respectively. The residual interface raises aggregate coverage to 88.53% and 88.30% while increasing pinball loss, demonstrating that improved coverage alone does not imply a uniformly better predictive distribution. Neither interface achieves nominal coverage across the volatility states. The corrected scenario mean remains an empirical estimate constructed from historical residuals rather than a verified conditional expectation. These distinctions constrain how forecasting results can be interpreted economically. Point errors measure discrepancies without incorporating scheduled energy quantities, while probability scores summarize marginal

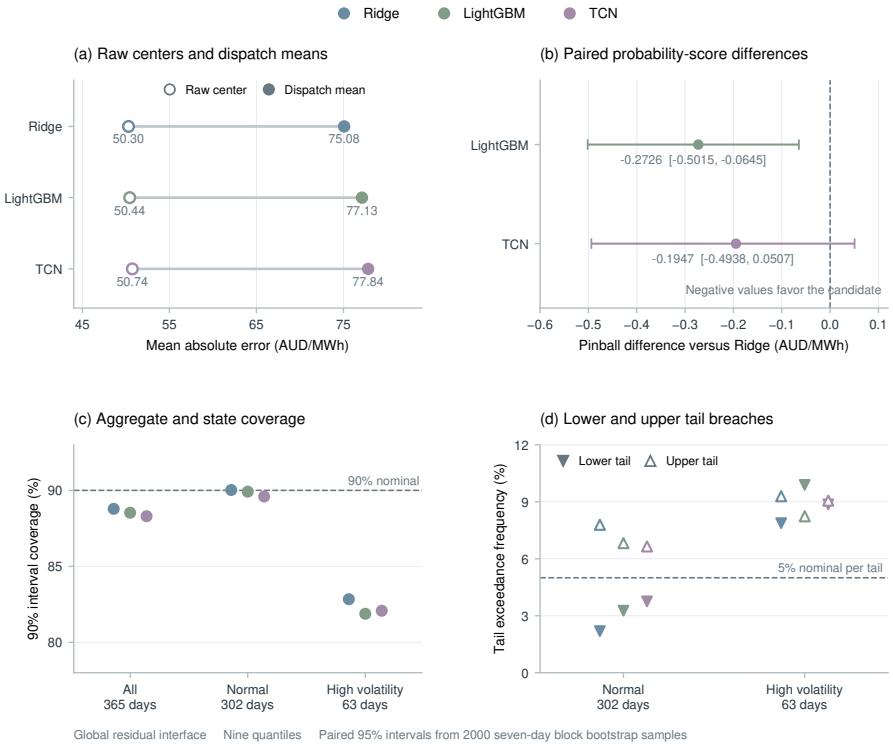


Figure 2. NSW1 forecasting performance and conditional calibration diagnostics in 2025. Panel (a) distinguishes raw center MAE from corrected dispatch mean MAE. Panel (b) reports calibrated pinball differences relative to Ridge with pointwise 95% intervals from 2000 paired seven-day circular block bootstrap samples. Panel (c) compares aggregate, normal-state, and high-volatility coverage. Panel (d) separates lower and upper interval exceedances. Dashed reference lines indicate nominal 90% coverage and 5% exceedance per tail. Probability diagnostics use the common global residual scenario interface.

predictive performance without evaluating the financial consequences of a joint price path. A charging error and a discharging error can produce different profit effects even when their absolute price deviations are identical. The next analysis therefore evaluates realized profit and signed loss risk using the same battery constraints and scheduling formulation. Forecasting metrics serve as diagnostic evidence for explaining those decisions, rather than substitutes for economic performance or guarantees of risk budget compliance.

4.2 Scheduling Performance and Loss Diagnostics

Table 4 compares realized scheduling outcomes under the reference battery configuration. Ridge deterministic scheduling earns 99,931.11 AUD over the test year, with an empirical daily signed loss CVaR at the 95% level of 608.40 AUD and 42 losing days. The B0 policy earns 110,016.18 AUD, reduces realized CVaR to 484.11 AUD, and records ten losing days. The B50 policy earns 116,254.42 AUD, with realized CVaR of 504.98 AUD and 14 losing days. Both risk constrained policies improve annual profit and reduce empirical tail loss relative to the deterministic comparator in this sample. Relaxing the scenario risk budget from zero to 50 AUD raises profit while increasing realized CVaR and the number of losing days. Annual throughput falls from 3537.58 MWh under deterministic scheduling to 2152.16 MWh under B0 and 2621.33 MWh under B50, indicating that the risk constraint changes the extent of annual trading activity. The positive realized CVaR under B0 does not indicate a violation of the optimization constraint, which applies to the estimated scenario distribution at each decision origin. It instead demonstrates a discrepancy between modeled risk and subsequent realized outcomes. The budget labels consequently describe optimization inputs and cannot be interpreted as guaranteed upper bounds on observed losses.

Applying the same scheduling formulation to all predictors changes the ranking suggested by forecasting scores. Under B50, Ridge earns 116,254.42 AUD, followed by LightGBM at 113,272.80 AUD and TCN at 110,471.77 AUD. Their realized CVaR values are 504.98, 539.70, and 395.58 AUD, respectively, with 14, 17, and 18 losing days. TCN therefore has the lowest tail loss despite producing the lowest annual profit among these three policies. Its greater number of losing days also shows that loss frequency and tail severity describe different aspects of performance. Under B0, Ridge remains the profit leader at 110,016.18 AUD, whereas TCN earns 101,971.78 AUD and LightGBM earns 97,459.10 AUD. TCN again has the lowest

Table 4. Realized NSW1 scheduling outcomes over the 365 test days in 2025 under the reference battery configuration.

Model	Policy	Annual net profit (AUD)	Realized loss $CVaR_{95}$ (AUD)	Losing days	Worst daily profit (AUD)	Annual throughput (MWh)
Ridge	Deterministic	99,931.11	608.40	42	−5512.29	3,537.58
Ridge	CVaR B0	110,016.18	484.11	10	−4491.74	2,152.16
Ridge	CVaR B50	116,254.42	504.98	14	−4834.36	2,621.33
LightGBM	Deterministic	89,604.06	686.76	53	−6336.71	3,463.64
LightGBM	CVaR B0	97,459.10	465.03	6	−4284.30	1,686.66
LightGBM	CVaR B50	113,272.80	539.70	17	−4677.91	2,374.85
TCN	Deterministic	92,343.66	633.53	43	−6486.33	3,649.08
TCN	CVaR B0	101,971.78	375.37	12	−4903.94	2,323.04
TCN	CVaR B50	110,471.77	395.58	18	−4803.21	2,712.51

The reference battery has power of 1 MW, capacity of 2 MWh, round-trip efficiency of 0.90, and a throughput degradation charge of 5 AUD/MWh. SOC remains between 10% and 90%, with initial and terminal SOC fixed at 50%. CVaR policies use 200 complete-path scenarios. B0 and B50 constrain daily scenario signed loss CVaR at the 95% level to 0 and 50 AUD, respectively. Realized CVaR is calculated from the 365 observed daily signed losses and is distinct from the scenario constraint. A losing day has realized net profit below -10^{-5} AUD. Throughput includes charging and discharging energy, and annual net profit deducts the throughput degradation charge.

realized CVaR at 375.37 AUD, compared with 465.03 AUD for LightGBM and 484.11 AUD for Ridge. LightGBM records only six losing days under B0, compared with ten for Ridge and twelve for TCN. The deterministic profit ranking is Ridge, TCN, and LightGBM, while the B50 ranking places LightGBM ahead of TCN. These reversals demonstrate that economic rankings depend on the scheduling policy as well as the forecasting model. LightGBM's improvement in calibrated pinball loss does not translate into the highest realized profit or lowest realized CVaR under either budget.

The loss diagnostics rank all 365 trading days separately for each Ridge policy by ascending realized net profit and retain the ten lowest observations. This retrospective selection uses realized outcomes for diagnosis without changing the forecasts, scenario banks, risk budgets, or scheduled actions. The selected dates are therefore policy dependent rather than a common subset imposed before evaluation. Gross annual loss is defined as the sum of loss magnitudes over losing days, without offsetting them against profitable days. It amounts to 11,607.48 AUD for deterministic scheduling, 8835.02 AUD for B0, and 9244.99 AUD for B50. The ten worst days account for 91.44%, 100.00%, and 99.65% of these totals, respectively. All three policies share the same three worst dates, November 26, November 25, and December 19, 2025. Together, these dates contribute 84.90%, 98.01%, and 96.34% of gross loss under deterministic scheduling, B0, and B50. For B50, their realized profits are −4834.36, −2872.80, and −1199.79 AUD. This concentration explains why a small number of adverse episodes can dominate empirical tail risk despite substantial positive annual earnings. It also makes the estimated risk sensitive to which extreme episodes occur in the observed year, limiting conclusions drawn from loss frequency or aggregate coverage alone.

Fig. 3 examines price, charging, discharging, and SOC trajectories for representative loss days. On November 26, the Ridge B50 policy schedules charging at 1 MW during the 11:00–11:30 AEST interval. Its corrected price forecast is 10.26 AUD/MWh and the upper boundary of the 90% price interval is 95.22 AUD/MWh, whereas the realized half hourly price reaches 10,342.49 AUD/MWh. Purchasing 0.5 MWh during this interval produces an adverse charging prediction error of 5166.11 AUD. SOC increases from approximately 49.42% to 73.13%, consistent with the specified charging efficiency. The following interval includes discharging at approximately 0.554 MW, but this later transaction does not eliminate the preceding price exposure. The daily SOC trajectory remains within the 10%–90% operating bounds, and the operational checks identify no simultaneous charging and discharging. On November 25, another scheduled 1 MW charging interval coincides with a realized price of 5984.09 AUD/MWh against a forecast of −2.18 AUD/MWh. That day was classified as normal volatility before the decision, illustrating that the historical state indicator does not anticipate every subsequent jump. The observed trajectories connect severe forecast discrepancies to executed energy quantities, while confirming that the resulting financial losses arise within the feasible operating schedule rather than through violated battery constraints.

The combined diagnostics connect prediction error, insufficient tail representation, and scheduled exposure without treating them as independent causal components. For the B50 actions on November 26, the signed charging cost underestimate is 5007.28 AUD and the signed discharging revenue overestimate is 345.16 AUD. Their sum equals the 5352.44 AUD gap between forecast mean profit of 518.08 AUD and realized profit of −4834.36 AUD. This accounting identity permits favorable and unfavorable interval errors to offset each other and differs from an aggregation retaining only adverse errors. It measures the financial effect of price discrepancies conditional on the chosen schedule, without identifying how an alternative forecasting or optimization procedure would have changed those actions. The realized profit also falls far below the scenario profit first

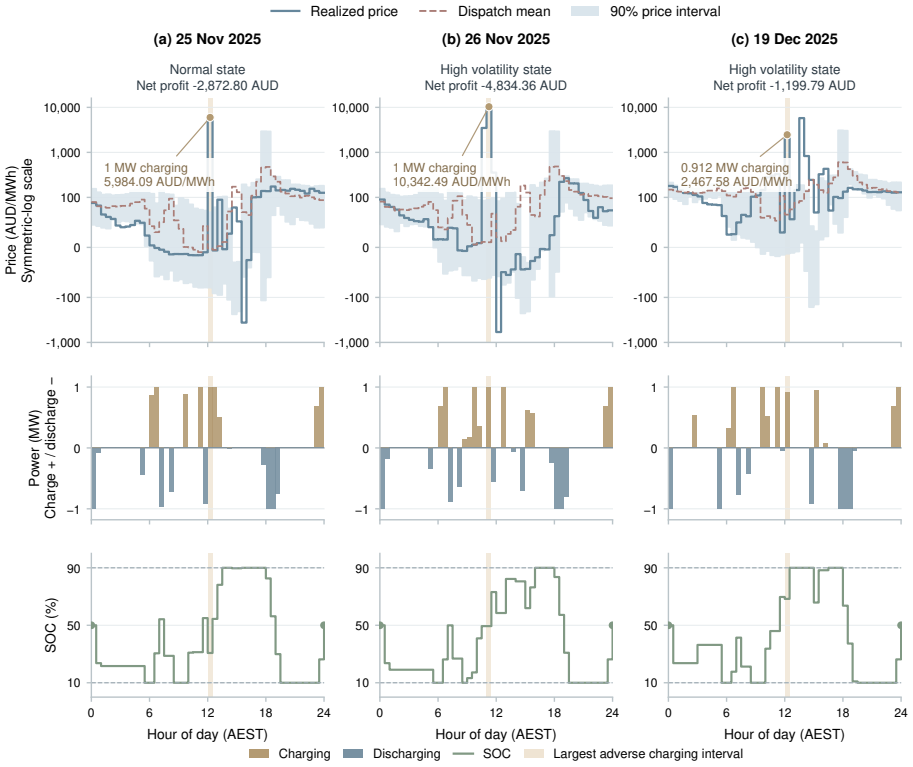


Figure 3. Price and battery trajectories for the three lowest-profit days of the Ridge B50 policy, selected retrospectively. Columns show November 25, November 26, and December 19, 2025. The upper row compares realized half-hour prices, corrected dispatch means, and 90% prediction intervals using a symmetric logarithmic price axis. The middle row shows charging as positive power and discharging as negative power. The lower row shows SOC with the 10% and 90% operating bounds and the terminal state at 24:00. Vertical shading identifies the interval with the largest positive charging-cost forecast error on each day. Volatility states are determined before the midnight decision.

percentile of -54.28 AUD, indicating substantial underrepresentation of this episode in the estimated lower tail. B50's annual 90% profit interval coverage is nevertheless 95.07%, demonstrating that aggregate coverage can coexist with severe extreme loss misspecification. The first percentile is estimated from only 200 scenarios and serves as a diagnostic rather than a precise tail estimate. The evidence supports an exposure based explanation of losses, but does not establish separate causal percentages attributable to forecasts, tails, or actions.

4.3 Calibration and Dependence Ablations

Table 5 reports the Ridge calibration comparisons with the forecast mean held fixed. State conditioning increases high volatility coverage from 82.84% to 82.87%, providing essentially no improvement toward the nominal 90% level. The average interval width increases from 271.52 to 301.41 AUD/MWh, while the interval score worsens from 1225.53 to 1245.06. Widening the high state intervals therefore does not establish successful calibration. In normal conditions, coverage decreases from 90.02% to 84.39%, accompanied by narrower intervals of 217.71 AUD/MWh and an improved interval score of 537.23, compared with 581.74 for global calibration. Lower tail exceedances rise to 10.58% in this state, indicating that the improvement in interval score coexists with substantial undercoverage. Using the same five quantiles within this ablation, high state pinball loss increases from 35.33 to 36.75, whereas normal state loss decreases from 17.34 to 16.67. These scores should not be directly compared with the preceding nine quantile results. The state variant consequently changes predictive sharpness and tail allocation without resolving the principal high volatility coverage deficit or delivering consistent improvements across both states and all reported probability evaluation metrics.

State conditioning changes the economic tradeoff despite its limited calibration improvement. Under B50, annual profit decreases from 116,254.42 to 106,539.67 AUD, while realized CVaR decreases from 504.98 to 329.55 AUD. The paired profit difference is -9714.75 AUD, with a 95% interval of $[-21132.18, 2792.32]$, and the CVaR difference is -175.43 AUD, with an interval of $[-626.74, 80.73]$. Both intervals include zero, leaving the direction of the underlying differences unresolved

Table 5. Ridge calibration and dependence ablations in NSW1 during 2025. Calibration statistics, realized scheduling outcomes, and selected paired economic differences are reported separately.

A. State calibration diagnostics

Bank	State	90% coverage (%)	90% width (AUD/MWh)	90% interval score	Pinball loss	Lower breach (%)	Upper breach (%)
Global	Normal	90.02	271.52	581.74	17.3427	2.19	7.79
Global	High	82.84	271.52	1225.53	35.3266	7.87	9.29
State	Normal	84.39	217.71	537.23	16.6662	10.58	5.04
State	High	82.87	301.41	1245.06	36.7472	8.40	8.73

B. Realized scheduling outcomes

Variant	Budget (AUD)	Seed	Annual net profit (AUD)	Loss CVaR_{95} (AUD)	Losing days	Worst daily profit (AUD)
Global full paths	0	1507	110,016.18	484.11	10	−4491.74
Global full paths	50	1507	116,254.42	504.98	14	−4834.36
State full paths	0	1507	107,510.22	255.40	40	−2616.47
State full paths	50	1507	106,539.67	329.55	41	−2616.47
Column permutation	0	1507	123,484.30	267.27	17	−3394.11
Column permutation	50	1507	125,640.04	283.33	13	−3954.03
Column permutation	50	21	121,221.67	461.57	13	−6015.70
Column permutation	50	42	122,284.95	259.05	14	−2131.98

C. B50 paired differences relative to global full paths

Variant / seed	Profit difference (AUD)	Profit difference 95% interval (AUD)	CVaR difference (AUD)	CVaR difference 95% interval (AUD)
State / 1507	−9714.75	$[-21132.18, 2792.32]$	−175.43	$[-626.74, 80.73]$
Permutation / 1507	9385.62	$[3494.90, 17501.08]$	−221.64	$[-652.67, 10.34]$
Permutation / 21	4967.25	$[332.12, 8772.67]$	−43.41	$[-166.81, 61.07]$
Permutation / 42	6030.53	$[-4998.08, 19605.56]$	−245.93	$[-786.98, 42.24]$

All variants retain the Ridge forecast mean and reference battery constraints. State conditioning and column permutation are separate ablations. Seeds denote residual sampling for full-path banks and column permutation for dependence variants. Panel A evaluates 90% intervals with nominal 5% exceedance per tail and uses five quantiles $\{0.05, 0.10, 0.50, 0.90, 0.95\}$. Its pinball losses are not directly comparable with the nine-quantile scores in Table 3. Interval scores and pinball losses are in AUD/MWh. Column permutation preserves empirical marginal distributions exactly and approximates independent coupling. Risk budgets constrain scenario CVaR, whereas Panel B reports realized daily signed loss CVaR. A losing day has net profit below -10^{-5} AUD. Panel C reports pointwise 95% intervals from 2000 paired seven-day circular block bootstrap samples. Positive profit differences and negative CVaR differences favor the variant.

under this exploratory protocol. Losing days increase from 14 to 41, although the worst daily loss decreases from 4834.36 to 2616.47 AUD. Reduced extreme loss severity therefore coexists with more frequent losses and lower annual earnings. Under B0, the state variant earns 107,510.22 AUD with realized CVaR of 255.40 AUD and 40 losing days, compared with 110,016.18 AUD, 484.11 AUD, and ten losing days for global calibration. These results provide evidence of a changed operating policy rather than a uniformly improved one. The observed point estimates suggest a tradeoff between extreme loss severity and profitability, while the confidence intervals prevent treating the apparent tail reduction as a statistically established benefit of state conditioning.

The dependence ablation preserves every empirical marginal distribution by independently permuting scenario rows within each forecasting horizon. Price coverage, quantiles, and marginal probability scores remain identical, so any resulting scheduling differences cannot be attributed to improved marginal forecasts. Under B50 and permutation seed 1507, the column permuted construction earns 125,640.04 AUD, compared with 116,254.42 AUD for complete paths. Realized CVaR decreases from 504.98 to 283.33 AUD, and losing days decrease from 14 to 13. Annual throughput decreases from 2621.33 to 2182.07 MWh, showing that changed scenario pairing alters trading exposure. The profit improvement is 9385.62 AUD, with a paired 95% interval of $[3494.90, 17501.08]$. The CVaR difference is −221.64 AUD, but its interval of $[-652.67, 10.34]$ includes zero. Under B0, column permutation also increases profit to 123,484.30 AUD and lowers the CVaR point estimate to 267.27 AUD, while increasing losing days to 17. The main comparisons therefore provide no observed advantage for retaining the current

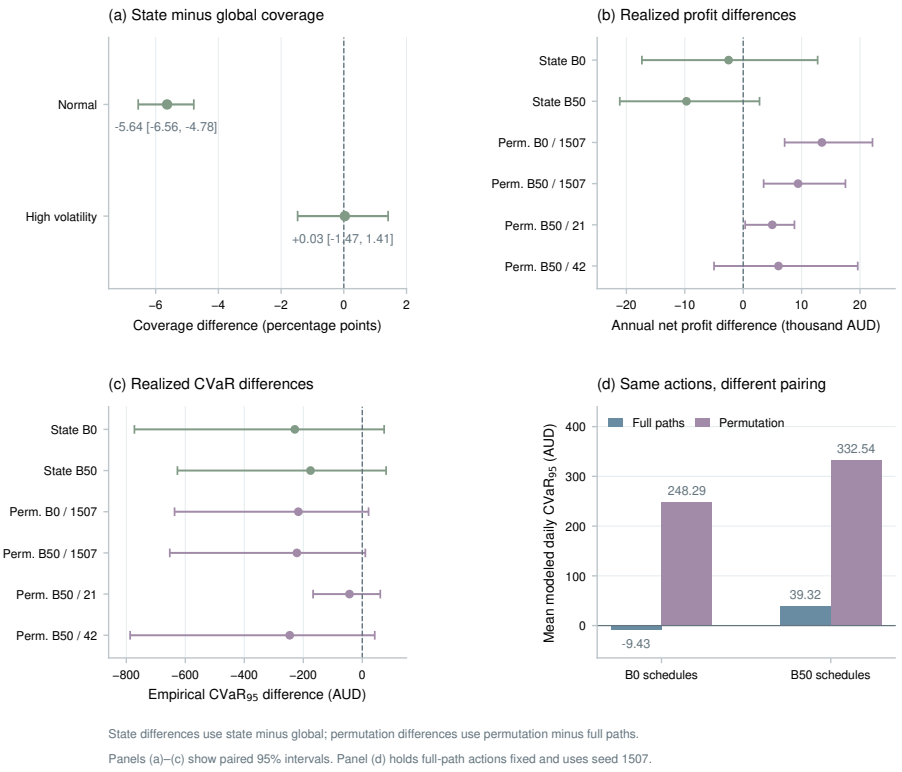


Figure 4. Calibration and dependence ablations for Ridge in NSW1 during 2025. Panel (a) reports state-conditioned minus global coverage differences. Panels (b) and (c) report realized profit and signed loss CVaR differences relative to global full paths. Error bars indicate pointwise 95% intervals from 2000 paired seven-day circular block bootstrap samples. Dashed lines indicate zero difference. Panel (d) evaluates unchanged full-path schedules under original and column-permuted scenario pairing using seed 1507. It reports average modeled daily CVaR, while actions and realized settlement profits remain unchanged. State conditioning and column permutation are separate ablations.

empirical complete paths. They test one residual bank and an independent coupling approximation, rather than demonstrating that temporal dependence is generally unimportant or should always be removed.

Fig. 4 isolates the effect of pairing by evaluating complete path schedules under both joint constructions. For the B50 schedules, average modeled daily CVaR is 39.32 AUD under complete paths and 332.54 AUD under the column permuted scenarios from seed 1507. The B0 values are −9.43 and 248.29 AUD. A negative modeled CVaR is admissible because the loss variable is signed and profitable outcomes are not truncated at zero. Forecast means, actions, degradation charges, and realized profits remain unchanged here. The altered values consequently represent different model assessments of the same exposure rather than different observed financial outcomes. Scenario profit standard deviations also change, but do not determine the direction of CVaR because that measure depends on the lower profit tail. The pairing modification can therefore make an accepted schedule exceed the modeled budget and lead the optimizer to select different actions upon reoptimization. This controlled comparison identifies a mechanism through which joint scenario construction affects decisions. It does not identify either construction as the true conditional price distribution or establish that its risk estimates are better calibrated out of sample.

Permutation seeds limit the strength of the dependence conclusion. Under B50, seeds 21 and 42 produce profits of 121,221.67 and 122,284.95 AUD, with realized CVaR of 461.57 and 259.05 AUD. These point estimates retain the main direction, but all three CVaR difference intervals include zero, and the profit interval for seed 42 also includes zero. Seed 21 produces a worst daily loss of 6015.70 AUD, exceeding the complete path loss of 4834.36 AUD despite its lower aggregate CVaR. Benefits on an annual tail statistic therefore do not guarantee protection against every extreme episode. The state calibration comparison likewise uses only 50 distinct high volatility calibration days, and drawing 200 scenarios with replacement does not increase that independent historical support. Finite column permutations preserve marginal samples exactly while providing only an approximate independent coupling, without guaranteeing zero residual correlations. The test year has already informed the preceding diagnostics, so these mechanism analyses remain exploratory. The combined evidence establishes sensitivity to

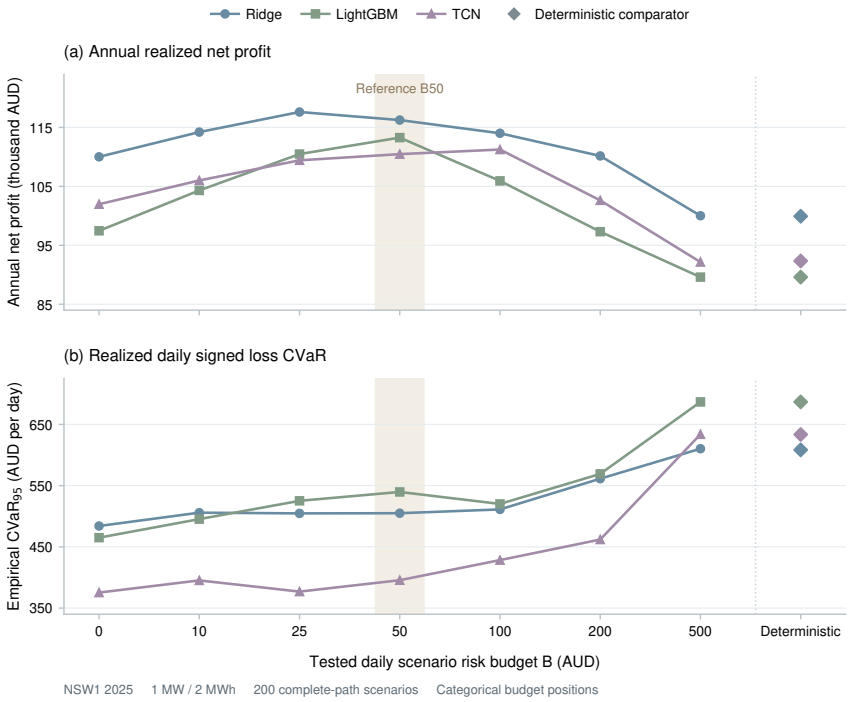


Figure 5. NSW1 scheduling performance across daily scenario risk budgets in 2025. Panel (a) reports annual realized net profit, and panel (b) reports empirical daily signed loss CVaR at the 95% level. Budget positions are categorical, and connecting lines guide the eye between tested settings. Diamonds indicate the deterministic comparators without a CVaR constraint. Shading identifies the reference B50 setting. All comparisons use the reference battery configuration and 200 complete-path scenarios. Points represent descriptive test-year estimates.

uncertainty construction and identifies failures of the tested variants, while leaving their superiority across other years, markets, and dependence models unresolved.

4.4 Risk Budgets and Battery Sensitivity

Fig. 5 shows realized profit and tail risk across daily scenario budgets of 0, 10, 25, 50, 100, 200, and 500 AUD. Ridge earns the highest annual profit among the three predictors at every finite budget tested. TCN has the lowest realized CVaR from B0 through B200, while Ridge has the lowest value at B500. These rankings separate profitability from tail protection and extend the B50 comparison beyond a single operating point. Ridge profit rises from 110,016.18 AUD at B0 to 117,619.50 AUD at B25, then falls to 100,018.07 AUD at B500. Its realized CVaR varies from 484.11 to 610.32 AUD across these settings. The realized curves are not monotonic, although relaxing the modeled constraint expands the feasible set and raises the optimized scenario mean objective. Model based expected profit and subsequent realized financial performance therefore require distinct interpretation. The B25 maximum describes this test year and cannot justify selecting an optimal deployment budget from the same observations.

Table 6 summarizes the parameter comparisons and paired uncertainty intervals, including the small-capacity and limited-scenario counterexamples. The battery sensitivity experiments use Ridge with B50 and vary one parameter at a time. Increasing round trip efficiency from 0.85 to 0.95 raises annual profit from 110,462.34 to 121,529.45 AUD and decreases realized CVaR from 523.45 to 497.81 AUD. Across these settings, B50 retains higher profit and lower CVaR than the corresponding deterministic schedules, with paired intervals supporting both differences. Increasing the throughput degradation charge from zero to 20 AUD/MWh reduces B50 profit from 125,912.99 to 87,345.26 AUD. Profit remains above the deterministic comparator at every tested charge, but realized CVaR does not decrease monotonically with the charge. The paired profit intervals exclude zero at charges of zero, ten, and 20 AUD/MWh. At the highest charge, the CVaR difference is -168.17 AUD with a 95% interval of $[-385.00, 13.58]$, leaving the risk improvement uncertain. These comparisons establish partial parameter stability within the tested configurations rather than invariance to battery operating costs or efficiency losses under arbitrary economic conditions or schedules.

Table 6. Ridge parameter sensitivity and key counterexamples in NSW1 during 2025. Each risk-constrained policy is compared with deterministic scheduling under the same physical and cost configuration.

Setting	Budget (AUD)	Deterministic annual profit daily $CVaR_{95}$	CVaR policy annual profit daily $CVaR_{95}$	Profit difference 95% interval	CVaR difference 95% interval
A. Efficiency and throughput degradation charge					
		99,931.11	116,254.42	16323.31	−103.42
Reference	50	608.40	504.98	[5491.12, 26240.17]	[−177.91, −48.01]
		93,424.36	110,462.34	17037.98	−111.49
$\eta_{rt} = 0.85$	50	634.94	523.45	[6243.89, 27486.91]	[−182.31, −56.59]
		106,344.22	121,529.45	15185.23	−88.98
$\eta_{rt} = 0.95$	50	586.79	497.81	[3727.67, 26434.89]	[−154.33, −40.41]
		115,093.05	125,912.99	10819.94	−90.54
$c_{deg} = 0$ AUD/MWh	50	565.90	475.37	[1763.26, 19197.95]	[−186.86, −28.14]
		87,032.29	107,444.54	20412.25	−123.38
$c_{deg} = 10$ AUD/MWh	50	652.35	528.97	[9263.48, 31218.62]	[−186.52, −71.94]
		66,550.58	87,345.26	20794.68	−168.17
$c_{deg} = 20$ AUD/MWh	50	641.81	473.64	[6639.35, 32937.79]	[−385.00, 13.58]
B. Capacity and capacity-normalized budgets					
		52,845.36	60,125.19	7279.83	242.27
Capacity 1 MWh	50	165.63	407.90	[−7321.09, 17494.79]	[−95.82, 890.09]
Capacity 1 MWh		52,845.36	62,037.82	9192.46	192.95
normalized budget	25	165.63	358.58	[−4438.04, 19678.89]	[−98.98, 747.79]
		195,895.50	205,869.97	9974.47	−207.78
Capacity 4 MWh	50	752.11	544.33	[−5586.76, 24945.64]	[−419.66, −51.79]
Capacity 4 MWh		195,895.50	208,548.39	12652.88	−189.07
normalized budget	100	752.11	563.04	[−376.80, 25401.19]	[−401.09, −35.26]
C. Number of complete-path scenarios					
		99,931.11	90,094.01	−9837.10	−15.27
50 scenarios	50	608.40	593.13	[−27769.94, 7887.31]	[−102.97, 58.74]
		99,931.11	108,642.02	8710.92	−175.67
100 scenarios	50	608.40	432.73	[−115.45, 17076.57]	[−433.17, −14.71]
		99,931.11	116,254.42	16323.31	−103.42
200 scenarios	50	608.40	504.98	[5491.12, 26240.17]	[−177.91, −48.01]
		99,931.11	116,049.36	16118.25	−171.01
500 scenarios	50	608.40	437.39	[10364.20, 23574.77]	[−510.55, 21.15]
		99,931.11	121,894.01	21962.90	−215.64
1000 scenarios	50	608.40	392.77	[12304.24, 32062.50]	[−559.23, 3.92]

All financial amounts are in AUD. Outcome cells report annual net profit on the first line and realized daily signed loss CVaR at the 95% level on the second line. Difference cells report the estimate above its pointwise 95% interval. Differences equal the risk-constrained policy minus its matching deterministic comparator and use 2000 paired seven-day circular block bootstrap samples. The reference configuration uses power of 1 MW, capacity of 2 MWh, round-trip efficiency of 0.90, a throughput degradation charge of 5 AUD/MWh, 200 complete-path scenarios, and budget B50. Physical and scenario scans vary one factor at a time; normalized capacity comparisons additionally scale the budget. Capacity remains subject to fixed 1 MW power; normalized budgets preserve 25 AUD per MWh of capacity. Scenario-count comparisons use one sampling seed and do not establish convergence. The reference row is repeated as the 200-scenario setting for comparison. Positive profit differences and negative CVaR differences favor the risk-constrained policy.

Capacity provides a counterexample to uniform risk improvement. With 1 MWh capacity and fixed 1 MW power, deterministic scheduling earns 52,845.36 AUD with realized CVaR of 165.63 AUD. B50 increases profit to 60,125.19 AUD but raises CVaR to 407.90 AUD. The risk difference is 242.27 AUD, with a 95% interval of $[-95.82, 890.09]$, so the observed deterioration is not statistically established. Setting the budget to B25 preserves the reference allowance of 25 AUD per MWh of capacity. This normalized configuration earns 62,037.82 AUD but still has higher realized CVaR of 358.58 AUD than deterministic scheduling. The counterexample consequently persists in point estimates after this budget adjustment. At 4 MWh, both B50 and normalized B100 produce higher profit and lower CVaR point estimates, although their profit difference intervals include zero. Holding power fixed changes storage duration as capacity varies, so budget normalization removes only one scale difference. The findings delimit the tested framework's performance rather than establish universal effects of capacity alone.

Scenario count also affects the realized outcomes nonmonotonically. Using 50, 100, 200, 500, and 1000 scenarios produces B50 profits of 90,094.01, 108,642.02, 116,254.42, 116,049.36, and 121,894.01 AUD, respectively. Corresponding realized CVaR values are 593.13, 432.73, 504.98, 437.39, and 392.77 AUD. With only 50 scenarios, profit falls below the deterministic benchmark of 99,931.11 AUD, eliminating the baseline profit advantage. The paired difference is -9837.10 AUD, with an interval of $[-27769.94, 7887.31]$, which includes zero. Increasing the count from 100 to 200 raises both profit and realized CVaR, while increasing it from 200 to 500 slightly reduces profit. More scenarios therefore do not ensure monotonic improvement in either economic endpoint. These runs use one sampling seed with common sequence prefixes and horizon recentering at each count. Repeated draws from 183 calibration days increase Monte Carlo representation without supplying new historical episodes. The observed 1000 scenario result cannot establish convergence or identify a preferred count without repeated sampling and independent validation of these patterns.

4.5 Regional Validation and EPF Benchmarks

Table 7 reports scheduling comparisons within VIC1 and SA1 using locally trained predictors and the frozen experimental configuration. In VIC1, Ridge deterministic scheduling earns 93,055.03 AUD with realized CVaR of 67.45 AUD, whereas B50 earns 98,241.09 AUD with CVaR of 33.62 AUD. The profit difference is 5186.06 AUD, with a paired 95% interval of $[1564.30, 11659.70]$, and the CVaR difference is -33.83 AUD, with an interval of $[-49.47, -15.81]$. In SA1, deterministic scheduling earns 97,156.02 AUD with CVaR of 170.50 AUD, compared with 106,662.32 AUD and 71.02 AUD under B50. The corresponding profit improvement is 9506.31 AUD, with an interval of $[6075.08, 13175.37]$, and the CVaR difference is -99.48 AUD, with an interval of $[-202.67, -36.78]$. Both regions therefore support higher profit and lower empirical tail loss for the reference Ridge B50 policy relative to their own deterministic comparators. These market internal differences extend the NSW1 result without assuming that absolute profit levels should match across regions. They validate the standard global residual scenario formulation under the specified battery and settlement assumptions, rather than demonstrate the effectiveness of the state calibration or dependence variants, which were not evaluated in these additional external regional scheduling experiments.

The economic model rankings show both consistency and regional dependence. Under B50, Ridge has the highest annual profit in all three AEMO regions. VIC1 profits are 98,241.09 AUD for Ridge, 93,195.78 AUD for LightGBM, and 89,170.71 AUD for TCN. Their realized CVaR values are 33.62, 61.52, and 61.30 AUD, respectively. SA1 profits follow the same ranking at 106,662.32, 100,657.69, and 94,112.44 AUD, with corresponding CVaR values of 71.02, 161.98, and 95.08 AUD. TCN's lowest tail risk in NSW1 therefore does not persist in VIC1 or SA1, where Ridge has the lowest CVaR among the B50 policies. LightGBM nevertheless improves SA1 calibrated pinball loss relative to Ridge by 1.0073, with a 95% interval of $[-2.5424, -0.1057]$. This predictive improvement again does not establish the strongest economic outcome under the common scheduler. In VIC1, neither LightGBM nor TCN has a calibrated pinball difference interval excluding zero relative to Ridge. Model selection based on one regional probability score would consequently miss differences in scheduled exposure and realized economic performance. The findings support evaluating predictors with the intended scheduling policy and local market conditions rather than transferring a forecasting or risk ranking unchanged across regions.

Table 7 presents the separate DE and PJM locally retrained forecasting comparisons using the common residual probability interface. For DE, calibrated pinball losses are 2.0481, 2.0429, and 2.0366 for Ridge, LightGBM, and TCN. TCN improves the score relative to Ridge by 0.0115, approximately 0.56%, but the paired 95% interval of $[-0.0761, 0.0470]$ includes zero. LightGBM's difference is -0.0052, with an interval of $[-0.0824, 0.0687]$, leaving its small point improvement uncertain as well. For PJM, the corresponding scores are 1.3486, 1.3442, and 1.3144. TCN achieves an improvement of 0.0342, approximately 2.53%, with an interval of $[-0.0785, 0.0072]$. LightGBM's difference is -0.0043, with an interval of $[-0.0445, 0.0290]$. The larger relative TCN improvement in PJM therefore remains insufficient to establish a directional advantage under the specified bootstrap analysis. The comparisons use 731 DE test days and 723 PJM test days, respectively, and differences are calculated within each dataset using identical paired test dates. Raw error magnitudes are not compared between these markets because their price units and distributions differ. The evidence suggests modest distributional score improvements for the stronger predictors, while retaining uncertainty about whether those gains would persist in independent periods or under alternative training configurations.

Table 7. Regional scheduling validation and EPF forecasting benchmarks. AEMO outcomes use the 2025 test year and reference battery configuration; EPF results are forecasting comparisons without storage scheduling.

A. Regional deterministic and B50 scheduling

Region	Model	Annual profit Det. / B50	Daily $CVaR_{95}$ Det. / B50	B50 losing days	Profit difference 95% interval	CVaR difference 95% interval
		99,931.11	608.40		16323.31	−103.42
NSW1	Ridge	116,254.42	504.98	14	[5491.12, 26240.17]	[−177.91, −48.01]
		89,604.06	686.76		23668.74	−147.06
NSW1	LightGBM	113,272.80	539.70	17	[17872.65, 29847.81]	[−337.48, 5.85]
		92,343.66	633.53		18128.11	−237.96
NSW1	TCN	110,471.77	395.58	18	[10840.70, 25272.05]	[−510.21, −29.77]
		93,055.03	67.45		5186.06	−33.83
VIC1	Ridge	98,241.09	33.62	18	[1564.30, 11659.70]	[−49.47, −15.81]
		88,377.40	90.90		4818.37	−29.38
VIC1	LightGBM	93,195.78	61.52	35	[901.79, 11687.36]	[−42.32, −14.38]
		85,598.98	88.14		3571.73	−26.84
VIC1	TCN	89,170.71	61.30	37	[361.83, 9309.36]	[−49.48, −7.57]
		97,156.02	170.50		9506.31	−99.48
SA1	Ridge	106,662.32	71.02	35	[6075.08, 13175.37]	[−202.67, −36.78]
		94,188.15	163.92		6469.55	−1.95
SA1	LightGBM	100,657.69	161.98	40	[−11659.78, 27858.58]	[−49.21, 63.04]
		81,951.34	216.84		12161.10	−121.76
SA1	TCN	94,112.44	95.08	46	[6859.80, 18992.57]	[−303.27, −20.97]

B. EPF forecasting metrics

Market	Model	Raw center MAE	Corrected mean MAE	Corrected mean RMSE	Calibrated pinball loss	90% coverage (%)
DE	Ridge	6.0602	5.8996	9.5022	2.0481	85.54
DE	LightGBM	6.1274	5.9923	9.6484	2.0429	85.07
DE	TCN	6.0596	5.9720	9.4410	2.0366	85.97
PJM	Ridge	4.0889	4.0781	6.3664	1.3486	81.28
PJM	LightGBM	3.8229	4.1143	6.3822	1.3442	79.06
PJM	TCN	3.8487	4.0020	6.1565	1.3144	79.92

C. EPF paired pinball differences relative to local Ridge

Market	Model	Pinball difference	95% interval
DE	LightGBM	−0.0052	[−0.0824, 0.0687]
DE	TCN	−0.0115	[−0.0761, 0.0470]
PJM	LightGBM	−0.0043	[−0.0445, 0.0290]
PJM	TCN	−0.0342	[−0.0785, 0.0072]

Panel A reports AUD amounts. In the profit and CVaR outcome cells, deterministic values appear above B50 values. Economic differences equal B50 minus the matching deterministic policy; estimate cells place the difference above its pointwise 95% interval. A losing day has net profit below -10^{-5} AUD. Realized daily signed loss CVaR is distinct from the scenario constraint. Each AEMO region has 365 test days and locally retrained models with frozen NSW1 configurations. EPF uses 731 DE days and 723 PJM days. DE prediction errors and pinball scores are in EUR/MWh, while PJM values are in USD/MWh; raw magnitudes are not compared between markets. Probability metrics use the common global residual interface and nine quantiles. All intervals use 2000 paired seven-day circular block bootstrap samples. Negative pinball differences favor the candidate model. EPF evaluates price and calendar forecasting only and does not validate storage profits or arbitrage risk.

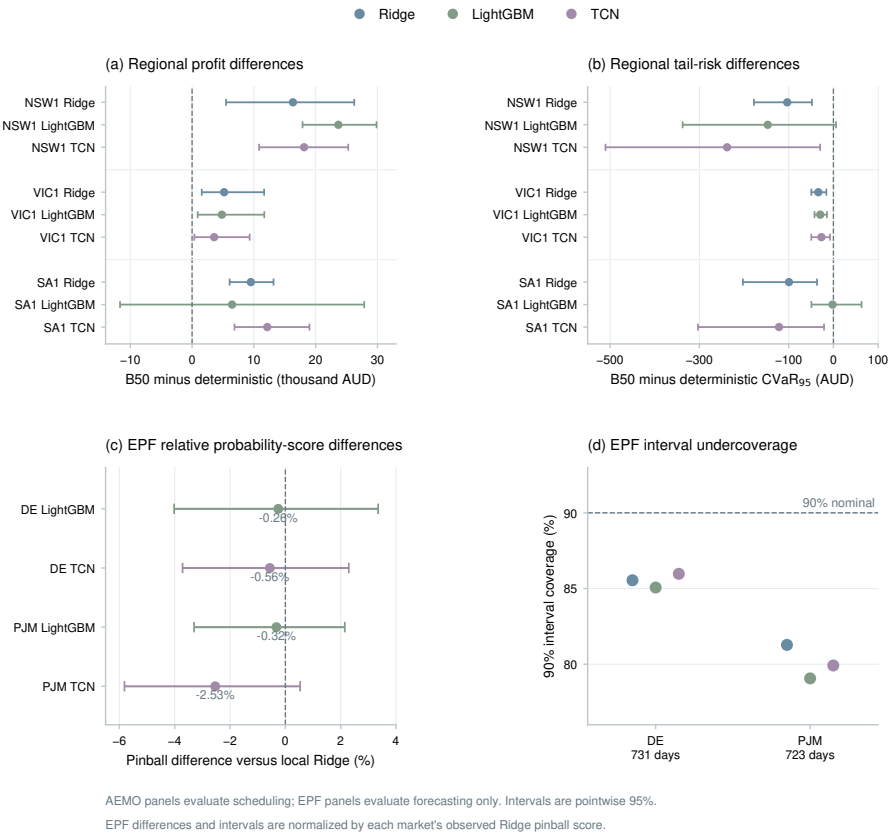


Figure 6. Regional scheduling validation and external forecasting diagnostics. Panels (a) and (b) report B50 minus deterministic differences in annual net profit and realized daily signed loss CVaR within each AEMO region. Panel (c) reports EPF pinball differences relative to the local Ridge baseline, with estimates and intervals normalized by each market's observed Ridge score. Panel (d) shows empirical coverage against the nominal 90% target. Error bars indicate pointwise 95% intervals from 2000 paired seven-day circular block bootstrap samples. AEMO comparisons evaluate storage scheduling, whereas EPF comparisons evaluate forecasting only.

The EPF intervals remain undercovered despite the favorable pinball point estimates. DE coverage is 85.54% for Ridge, 85.07% for LightGBM, and 85.97% for TCN against a nominal 90% target. PJM coverage is lower at 81.28%, 79.06%, and 79.92%, respectively. These deficits constrain any claim that the residual interface provides broadly calibrated uncertainty across markets. The EPF experiments evaluate hourly forecasts from historical prices and calendar variables, without using the exogenous forecast fields or reproducing an official toolbox leaderboard. They include no battery optimization, realized storage profit, or arbitrage risk evaluation, and therefore provide prediction validation rather than cross market scheduling validation. The AEMO regional experiments do evaluate scheduling, but NSW1, VIC1, and SA1 belong to the same interconnected market and share the same test year. They cannot be treated as independent replications across market institutions or historical periods. Local retraining with frozen configurations also tests procedural transfer rather than transfer of NSW1 model weights without adaptation. The combined evidence extends the empirical scope while preserving separate conclusions for regional scheduling performance, external forecasting accuracy, and the unresolved calibration limitations observed. Fig. 6 summarizes regional scheduling differences and EPF forecasting diagnostics as distinct validation tasks.

5 Discussion

5.1 Interpretation of the Main Findings

The main findings answer the research question by showing that predictive accuracy, probability calibration, and scheduling value are related but noninterchangeable evaluation targets. LightGBM improves the common calibrated pinball score in NSW1, yet Ridge delivers the highest profit under the reference risk budgets. TCN produces lower realized tail loss in NSW1 without consistently reducing loss frequency, and its risk ranking changes across regions. These differences indicate that a

forecasting improvement cannot be assigned a fixed economic value independently of the decision rule. A price discrepancy becomes financially consequential through its coincidence with charging or discharging, the energy quantity exposed, and the opportunities available elsewhere in the day. This interpretation is consistent with decision focused learning research that evaluates prediction through the decisions it induces rather than prediction error alone[38]. The present experiments do not implement that training approach, but demonstrate why its distinction matters for evaluation. Separating raw prediction centers from corrected dispatch means also prevents differences in statistical targets from being mistaken for contradictory findings. The evidence consequently supports joint forecasting and scheduling assessment while leaving open whether training directly on economic objectives would improve this particular framework. That possibility requires a controlled comparison beyond the experiments reported here.

The calibration and dependence ablations identify mechanisms without resolving their causes. State conditioning changes interval widths and tail allocation but barely improves high volatility coverage, while reducing normal state coverage. Splitting the calibration sample leaves only 50 high volatility days, and scenario replication cannot create additional independent episodes. Reduced historical support could contribute to unstable tail estimates, but the experiments do not isolate sample scarcity from distribution shift or limitations of the volatility indicator. The normal state loss on November 25 also shows that recent volatility does not reliably identify every subsequent jump. These observations support a limitation of the tested conditioning scheme, while explaining its failure through inadequate state information remains a hypothesis. The dependence ablation provides more direct mechanism evidence because marginal samples and forecast means are unchanged. Evaluating identical actions under different scenario pairings changes modeled CVaR, establishing that the joint construction affects risk judgments. Reoptimization then changes trading exposure and realized outcomes. The complete path bank's lack of advantage does not imply that temporal dependence is absent or irrelevant. Historical dependence may inadequately represent future adverse combinations, but that explanation requires further validation against alternative joint models, tail stress scenarios, and periods not used for development.

The results align with earlier storage research in treating risk management as conditional on the modeled trading opportunities and operating assumptions. Abramova and Bunn formulate daily battery trading through predictive arbitrage spread distributions, linking uncertainty to the profitability of specific purchase and sale combinations[39]. Our fixed schedule experiments instead use full price scenarios, but similarly show that economic exposure depends on combinations of prices across time rather than isolated hourly errors. Research on battery cost of use also examines how ageing and efficiency representations affect arbitrage operation[40]. The present linear throughput charge provides a controlled simplification, so its sensitivity results should not be interpreted as validation of a degradation model. The small capacity counterexample further demonstrates that a risk constraint does not guarantee lower realized risk under every feasible battery configuration. Scenario CVaR controls an estimated distribution, and its financial protection depends on how that distribution represents the losses induced by available actions. Regional improvements therefore support the reference formulation within its tested conditions, while calibration failures and uncertain ablation effects define its boundaries. The contribution is evidence about these conditions and mechanisms, rather than a new CVaR algorithm or proof that one forecasting architecture provides universally superior storage decisions.

5.2 Engineering Implications

Ridge provides an economic reference because its scheduling profit remains competitive against the forecasting architectures under the tested protocol. This finding supports retaining a simple predictor in comparisons rather than assuming that architectural complexity will improve storage operation. A candidate model should be assessed through the same uncertainty interface, battery constraints, cost assumptions, and risk budgets as the reference before its economic value is inferred. The comparison should distinguish native quantile quality from the residual distribution actually supplied to the optimizer and report raw center errors separately from corrected mean errors. These controls help identify whether an apparent gain comes from forecasting, uncertainty representation, or a changed decision setting. The experiments do not establish a computational efficiency advantage for Ridge, since deployment latency and operational resource requirements were not evaluated systematically. Its practical role here is a transparent benchmark against which additional complexity can justify itself through decision outcomes. Model selection should consequently consider profit, tail severity, loss frequency, and uncertainty together rather than prioritize a single predictive accuracy score.

Risk budgets require a monetary and temporal interpretation. The daily allowance constrains modeled signed loss CVaR and should not be presented as a guaranteed maximum realized loss. Capacity normalization can make comparisons interpretable, but fixing the allowance per MWh does not equalize storage duration when power remains unchanged. The small battery counterexample suggests that each physical configuration should receive its own economic evaluation rather than inherit a budget solely through proportional scaling. Scenario count requires similar care. Increasing the number of sampled paths changes the estimated joint distribution and the optimization problem, but does not increase the number of distinct calibration days. The high volatility bank contains only 50 historical days, so repeated sampling cannot supply missing extreme episodes.

Engineering reports should therefore disclose both scenario count and historical support, including how scenarios are centered and paired. The nonmonotonic sampling results also suggest evaluating several seeds before choosing a computational setting. Such checks would assess sampling sensitivity without treating larger scenario sets as automatic evidence of convergence or better protection.

A prospective implementation should select risk budgets on a development period independent of performance evaluation and then freeze the selection procedure before examining subsequent outcomes. The highest profit observed in the present test year should not determine the deployment budget. Monitoring should combine marginal coverage with upper and lower exceedance frequencies, exceedance magnitudes, and the energy quantities exposed during those intervals. Realized profit, loss frequency, throughput, and empirical tail loss should be tracked separately from the optimizer's scenario risk estimate. Conditional diagnostics could identify periods in which aggregate coverage masks concentrated exposure, although their operational thresholds require independent validation. Any proposed recalibration or rescheduling response should use information available at the intervention time and undergo a separate backtest reflecting its altered decision protocol. These measures are recommendations derived from the observed limitations, rather than demonstrated capabilities of an operating platform. Their value would need to be established through prospective evaluation, including whether they reduce losses without sacrificing excessive trading opportunities.

5.3 Limitations and Future Work

The empirical conclusions are limited by the evaluation history and data protocol. NSW1 outcomes from 2025 were repeatedly examined during loss diagnosis, mechanism development, and sensitivity analysis. Although future test observations were not used to fit forecasting parameters, calibration banks, or volatility thresholds, this repeated inspection prevents treating the year as an untouched confirmatory sample. The paired bootstrap intervals quantify uncertainty conditional on the observed sequence and fitted methods, without removing development related selection or adjusting for all reported comparisons. VIC1 and SA1 extend the regional scope, but their contemporaneous observations originate from the same interconnected market and cannot provide independent evidence across institutional settings or years. The analysis also uses final historical archives, and an audit of the data versions available at each original decision time remains incomplete. Revision differences could therefore affect the interpretation of historical information availability despite the imposed feature cutoff. Aggregating five minute observations into half hourly prices smooths short lived extremes and restricts the findings to the selected temporal resolution. The results do not establish equivalent performance for a five minute dispatch protocol. These constraints warrant interpreting the regional comparisons as conditional retrospective evidence and the mechanism analyses as exploratory findings requiring independent confirmation.

The operational model introduces boundaries. A complete next day schedule is fixed at midnight, excluding responses to information arriving during the trading day. The price taker assumption also treats storage actions as having no effect on market prices and does not represent strategic bidding, market clearing, or competition among storage participants. Profits include energy settlement and a linear charge applied to charging and discharging throughput, but exclude capital expenditure, network charges, and ancillary service revenues. They consequently describe operating arbitrage outcomes rather than investment returns or total asset value. The degradation charge does not capture cycle depth, temperature, calendar ageing, or changes in usable capacity and efficiency over battery life. Parameter scans vary one factor at a time and cannot identify interactions among these operating characteristics. Sampling and training variation are also incompletely characterized. The scenario count comparison uses one sampling seed, while additional TCN seeds evaluate prediction rather than repeat the complete economic experiment. The dependence comparison includes three permutation seeds, which remain alternative constructions on the same test year rather than independent market replications. These simplifications limit extrapolation to operational portfolios and make it necessary to distinguish verified optimization feasibility from financial reliability under unmodeled disturbances or costs.

Future work should freeze the state calibration and dependence configurations before evaluating them on regions and years not used to develop these mechanisms. Such comparisons should retain the reported global versus state and complete path versus permutation contrasts, so that new improvements are assessed against the boundaries documented here. Annual periods would help determine whether concentrated losses and model rankings persist under different price regimes, while repeated forecasting and scenario seeds would distinguish training variation from uncertainty in scenario construction. State conditioning could be examined through partial pooling or shrinkage, and tail stress scenarios could test whether rare adverse exposures are inadequately represented. These are proposed extensions rather than explanations established by the current experiments. A historical release audit should reconstruct information availability wherever archived versions permit, alongside a comparison at the native five minute resolution. Rolling scheduling should then be evaluated as a separate decision protocol with explicit update times, information cutoffs, and terminal energy treatment. Its potential benefit cannot be inferred from the present fixed schedule results. Budget selection and intervention thresholds should be determined on independent development data, followed by prospective evaluation of profit, tail loss, feasibility, and computational requirements under the frozen operating

rules and configurations.

6 Conclusion

This study evaluated whether electricity price forecasting quality translates into profitable and less risky storage scheduling under a fixed daily decision protocol. Controlled comparisons of Ridge, quantile LightGBM, and quantile TCN showed that predictive scores, realized profit, and tail loss do not produce a common ranking. Under the reference battery configuration, Ridge CVaR scheduling improved profit and reduced realized tail loss relative to deterministic scheduling in NSW1, VIC1, and SA1. The stronger predictors did not universally outperform Ridge, and TCN's tail risk advantage in NSW1 did not persist across the other regions. State conditioning failed to improve high volatility coverage, while preserving the tested complete paths provided no observed scheduling advantage over column permutation. These findings answer the research questions by demonstrating that decision value depends on the uncertainty interface, scenario pairing, operating constraints, and risk allowance as well as the forecasting model. They support evaluating prediction and optimization rather than treating probability accuracy as sufficient evidence of economic superiority.

The contribution lies in a controlled evaluation that distinguishes marginal calibration, modeled risk, and realized financial outcomes. Fixed action comparisons established that changing joint scenario pairing alters the optimizer's risk assessment without changing settlement profit, while loss diagnostics connected severe price discrepancies to scheduled charging exposure. These mechanisms explain sensitivity without identifying independent causal shares of the losses. The benefits remain conditional on the tested configurations. The 1 MWh case retained higher tail loss point estimates after budget normalization, and the 50 scenario configuration lost the baseline profit advantage. Scenario count effects were nonmonotonic, and several risk difference intervals remained inconclusive. Regional scheduling comparisons extended the reference formulation within one interconnected market, whereas DE and PJM validated forecasting only and retained undercoverage. Repeated inspection of NSW1 also limits confirmatory interpretation. Independent periods, frozen mechanism comparisons, and repeated sampling are therefore needed to establish the durability of these findings before drawing broader conclusions about storage operation or selecting deployment risk budgets.

Data Availability

The research outputs, analysis code, experiment configurations, and reproducibility documentation supporting this study are available on Zenodo at `https://doi.org/10.5281/zenodo.23388643`. The archive documents the original AEMO and EPF data sources and provides instructions for retrieving upstream files that are not redistributed. Third-party data remain subject to their respective terms of use.

Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

Kangqian Huang contributed to conceptualization, methodology, formal analysis, and writing the original draft. Xin Hu contributed to data curation, software implementation, and experimental analysis. Jingxiang Shang contributed to methodology refinement, validation, and interpretation of the results. Fan Yang contributed to visualization, experimental evaluation, and manuscript revision. Zhiqing Wang contributed to supervision and manuscript review and editing.

Funding

This research received no external funding.

References

1. Cornell, C., Dinh, N. T. & Pourmousavi, S. A. A probabilistic forecast methodology for volatile electricity prices in the Australian National Electricity Market. *Int. J. Forecast.* **40**, 1421–1437, DOI: 10.1016/j.ijforecast.2023.12.003 (2024).
2. Mercier, T., Olivier, M. & De Jaeger, E. The value of electricity storage arbitrage on day-ahead markets across Europe. *Energy Econ.* **123**, 106721, DOI: 10.1016/j.eneco.2023.106721 (2023).

3. Krishnamurthy, D., Uckun, C., Zhou, Z., Thimmapuram, P. R. & Botterud, A. Energy Storage Arbitrage Under Day-Ahead and Real-Time Price Uncertainty. *IEEE Transactions on Power Syst.* **33**, 84–93, DOI: 10.1109/TPWRS.2017.2685347 (2018).

4. Camuñas García-Miguel, P. L., Alonso-Martinez, J., Arnaltes Gómez, S. & Rodríguez-Amenedo, J. L. Impact of risk measures and degradation cost on the optimal arbitrage schedule for battery energy storage systems. *Int. J. Electr. Power & Energy Syst.* **157**, 109883, DOI: 10.1016/j.ijepes.2024.109883 (2024).

5. Weron, R. Electricity price forecasting: A review of the state-of-the-art with a look into the future. *Int. J. Forecast.* **30**, 1030–1081, DOI: 10.1016/j.ijforecast.2014.08.008 (2014).

6. Nowotarski, J. & Weron, R. Recent advances in electricity price forecasting: A review of probabilistic forecasting. *Renew. Sustain. Energy Rev.* **81**, 1548–1568, DOI: 10.1016/j.rser.2017.05.234 (2018).

7. Lago, J., De Ridder, F. & De Schutter, B. Forecasting spot electricity prices: Deep learning approaches and empirical comparison of traditional algorithms. *Appl. Energy* **221**, 386–405, DOI: 10.1016/j.apenergy.2018.02.069 (2018).

8. Marcjasz, G., Narajewski, M., Weron, R. & Ziel, F. Distributional neural networks for electricity price forecasting. *Energy Econ.* **125**, 106843, DOI: 10.1016/j.eneco.2023.106843 (2023).

9. Sang, L., Xu, Y., Long, H., Hu, Q. & Sun, H. Electricity Price Prediction for Energy Storage System Arbitrage: A Decision-Focused Approach. *IEEE Transactions on Smart Grid* **13**, 2822–2832, DOI: 10.1109/TSG.2022.3166791 (2022).

10. Maciejowska, K., Lipiecki, A. & Uniejewski, B. Statistical and economic evaluation of forecasts in electricity markets: Beyond RMSE and MAE. *Energy Convers. Manag.* **356**, 121408, DOI: 10.1016/j.enconman.2026.121408 (2026).

11. do Prado, J. C. & Chikezie, U. A Decision Model for an Electricity Retailer With Energy Storage and Virtual Bidding Under Daily and Hourly CVaR Assessment. *IEEE Access* **9**, 106181–106191, DOI: 10.1109/ACCESS.2021.3100815 (2021).

12. Xiao, D., Chen, H., Cai, W., Wei, C. & Zhao, Z. Integrated risk measurement and control for stochastic energy trading of a wind storage system in electricity markets. *Prot. Control. Mod. Power Syst.* **8**, 60, DOI: 10.1186/s41601-023-00329-3 (2023).

13. Wessel, E., Smets, R. & Delarue, E. Risk-aware participation in day-ahead and real-time balancing markets for energy storage systems. *Electr. Power Syst. Res.* **235**, 110741, DOI: 10.1016/j.epsr.2024.110741 (2024).

14. Pareek, P. *et al.* Degradation-infused energy portfolio allocation framework: Risk-averse fair storage participation. *Energy* **313**, 133688, DOI: 10.1016/j.energy.2024.133688 (2024).

15. Grothe, O., Kächele, F. & Krüger, F. From point forecasts to multivariate probabilistic forecasts: The Schaake shuffle for day-ahead electricity price forecasting. *Energy Econ.* **120**, 106602, DOI: 10.1016/j.eneco.2023.106602 (2023).

16. Hilger, H. *et al.* Multivariate scenario generation of day-ahead electricity prices using normalizing flows. *Appl. Energy* **367**, 123241, DOI: 10.1016/j.apenergy.2024.123241 (2024).

17. Maciejowska, K., Serafin, T. & Uniejewski, B. Probabilistic forecasting with a hybrid Factor-QRA approach: Application to electricity trading. *Electr. Power Syst. Res.* **234**, 110541, DOI: 10.1016/j.epsr.2024.110541 (2024).

18. Kath, C. & Ziel, F. Conformal prediction interval estimation and applications to day-ahead and intraday power markets. *Int. J. Forecast.* **37**, 777–799, DOI: 10.1016/j.ijforecast.2020.09.006 (2021).

19. Mashlakov, A., Kuronen, T., Lensu, L., Kaarna, A. & Honkapuro, S. Assessing the performance of deep learning models for multivariate probabilistic energy forecasting. *Appl. Energy* **285**, 116405, DOI: 10.1016/j.apenergy.2020.116405 (2021).

20. Hoerl, A. E. & Kennard, R. W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* **12**, 55–67, DOI: 10.1080/00401706.1970.10488634 (1970).

21. Ke, G. *et al.* LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems*, vol. 30 (Curran Associates, Inc., 2017).

22. Bai, S., Kolter, J. Z. & Koltun, V. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling, DOI: 10.48550/arXiv.1803.01271 (2018). ArXiv preprint, 1803.01271.

23. Koenker, R. & Bassett, G. Regression Quantiles. *Econometrica* **46**, 33–50, DOI: 10.2307/1913643 (1978).

24. Marcjasz, G., Serafin, T. & Weron, R. Selection of Calibration Windows for Day-Ahead Electricity Price Forecasting. *Energies* **11**, 2364, DOI: 10.3390/en11092364 (2018).

25. Gneiting, T., Balabdaoui, F. & Raftery, A. E. Probabilistic Forecasts, Calibration and Sharpness. *J. Royal Stat. Soc. Ser. B: Stat. Methodol.* **69**, 243–268, DOI: 10.1111/j.1467-9868.2007.00587.x (2007).

26. Schefzik, R., Thorarinsdottir, T. L. & Gneiting, T. Uncertainty Quantification in Complex Simulation Models Using Ensemble Copula Coupling. *Stat. Sci.* **28**, 616–640, DOI: 10.1214/13-STS443 (2013).

27. García-Miguel, P. L. C., Alonso-Martínez, J., Arnaltes Gómez, S., García Plaza, M. & Asensio, A. P. A Review on the Degradation Implementation for the Operation of Battery Energy Storage Systems. *Batteries* **8**, 110, DOI: 10.3390/batteries8090110 (2022).

28. Rockafellar, R. T. & Uryasev, S. Optimization of conditional value-at-risk. *The J. Risk* **2**, 21–41, DOI: 10.21314/JOR.2000.038 (2000).

29. Rockafellar, R. T. & Uryasev, S. Conditional value-at-risk for general loss distributions. *J. Bank. & Finance* **26**, 1443–1471, DOI: 10.1016/S0378-4266(02)00271-6 (2002).

30. Huangfu, Q. & Hall, J. A. J. Parallelizing the dual revised simplex method. *Math. Program. Comput.* **10**, 119–142, DOI: 10.1007/s12532-017-0130-5 (2018).

31. Australian Energy Market Operator. Historical data. Market Management System historical data archive. Accessed: 2026-10-14; DISPATCHPRICE and DISPATCHREGIONSUM data for 2022–2025.

32. Open-access benchmark dataset for day-ahead electricity prices. Zenodo dataset, DOI: 10.5281/zenodo.4624805 (2021). Accessed: 2026-10-14.

33. Lago, J., Marcjasz, G., De Schutter, B. & Weron, R. Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark. *Appl. Energy* **293**, 116983, DOI: 10.1016/j.apenergy.2021.116983 (2021).

34. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

35. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations* (2015).

36. Gneiting, T. & Raftery, A. E. Strictly Proper Scoring Rules, Prediction, and Estimation. *J. Am. Stat. Assoc.* **102**, 359–378, DOI: 10.1198/016214506000001437 (2007).

37. Kunsch, H. R. The Jackknife and the Bootstrap for General Stationary Observations. *The Annals Stat.* **17**, DOI: 10.1214/aos/1176347265 (1989).

38. Elmachtoub, A. N. & Grigas, P. Smart "Predict, then Optimize". *Manag. Sci.* **68**, 9–26, DOI: 10.1287/mnsc.2020.3922 (2022).

39. Abramova, E. & Bunn, D. Optimal Daily Trading of Battery Operations Using Arbitrage Spreads. *Energies* **14**, 4931, DOI: 10.3390/en14164931 (2021).

40. García-Miguel, P. L. C., Asensio, A. P., Merino, J. L. & Plaza, M. G. Analysis of cost of use modelling impact on a battery energy storage system providing arbitrage service. *J. Energy Storage* **50**, 104203, DOI: 10.1016/j.est.2022.104203 (2022).