Energy and AI 21 (2025) 100571

Contents lists available at ScienceDirect

Energy and AI

journal homepage: www.elsevier.com/locate/egyai

Conformal Prediction for electricity price forecasting in the day-ahead and real-time balancing market

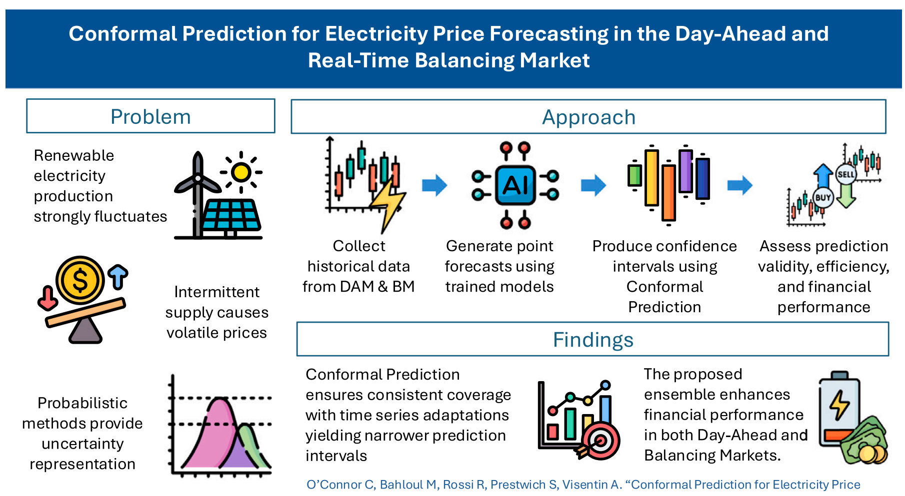
Ciaran O'Connor [a], Mohamed Bahloul [b], Roberto Rossi [c], Steven Prestwich [d], Andrea Visentin [d],*

[a] SFI CRT in Artificial Intelligence, School of Computer Science & IT, University College Cork, Ireland
[b] Water & Energy Transition Unit, Vlaamse Instelling voor Technologisch Onderzoek, Mol, Belgium
[c] Business School, University of Edinburgh, Edinburgh, UK
[d] SFI Insight Centre for Data Analytics, School of Computer Science & IT, University College Cork, Ireland

HIGHLIGHTS

- Applied time-series-adapted Conformal Prediction to electricity price forecasting.
- Developed Q-Ens ensemble for enhanced prediction accuracy and interval reliability.
- Achieved narrower prediction intervals while maintaining coverage guarantees.
- Assessed financial benefits of CP and QR models in Day-Ahead and Balancing Markets.
- Demonstrated increased returns in energy storage arbitrage using the proposed ensemble.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Probabilistic Electricity Price Forecasting
Conformal Prediction
Arbitrage trading
Machine learning

ABSTRACT

The integration of renewable energy into electricity markets poses significant challenges to price stability and increases the complexity of market operations. Accurate and reliable electricity price forecasting is crucial for effective market participation, where price dynamics can be significantly more challenging to predict. Probabilistic forecasting, through prediction intervals, efficiently quantifies the inherent uncertainties in electricity prices, supporting better decision-making for market participants. This study explores the enhancement of probabilistic price prediction using Conformal Prediction (CP) techniques, specifically Ensemble Batch Prediction Intervals and Sequential Predictive Conformal Inference. These methods provide precise and reliable prediction intervals, outperforming traditional models in validity metrics. We propose an ensemble approach that combines the efficiency of quantile regression models with the robust coverage properties of time series adapted CP techniques. This ensemble delivers both narrow prediction intervals and high coverage, leading to more reliable and accurate forecasts. We further evaluate the practical implications of CP techniques through a simulated trading algorithm applied to a battery storage system. The ensemble approach demonstrates

* Corresponding author.
E-mail addresses: 119226305@umail.ucc.ie (C. O'Connor), mohamed.bahloul@vito.be (M. Bahloul), roberto.rossi@ed.ac.uk (R. Rossi), s.prestwich@cs.ucc.ie (S. Prestwich), andrea.visentin@ucc.ie (A. Visentin).

https://doi.org/10.1016/j.egyai.2025.100571
Received 24 January 2025; Received in revised form 24 June 2025; Accepted 22 July 2025
Available online 4 August 2025

improved financial returns in energy trading in both the Day-Ahead and Balancing Markets, highlighting its practical benefits for market participants.

1. Introduction

Electricity Price Forecasting (EPF) is particularly challenging due to market volatility, sudden price fluctuations, and seasonal demand shifts. Accurate EPF is essential for all market participants in deregulated markets, providing a foundation for informed decision-making and financial planning. Power generation companies rely on forecasts to hedge against risks from deviations between day-ahead and real-time markets, thereby stabilising revenue streams and operational planning. Load-serving utilities use forecasts to secure energy at lower prices, optimising procurement strategies and ensuring cost-effective supply for consumers. Virtual traders, lacking actual generation power or load requirements, depend on EPF to estimate price spreads between spot and forward markets, exploiting arbitrage opportunities to generate profits. The increasing integration of renewable energy sources and the advent of smart grids further complicate EPF, exacerbating market volatility.

Spot markets such as the Day-Ahead Market (DAM), Intra-Day Market (IDM), and Balancing Market (BM) are crucial for maintaining grid stability. Yet, increasing market price volatility, driven by forecast errors in renewable electricity generation, presents significant challenges [1]. This underscores the critical importance of accurate forecasting in stabilising energy production planning, optimising trading strategies, and implementing risk-aware policies to ensure a reliable, cost-effective, and sustainable energy supply. Accurate EPF supports market participants' financial health and plays an important role in advancing energy policy goals surrounding renewable integration, grid stability, and reducing carbon emissions.

There are two major EPF schemes: point forecasts and probabilistic forecasts. Point forecasts provide single price predictions that are easy to interpret and establish parsimonious predictor-target relationships under the assumption of homoscedasticity. However, they cannot quantify uncertainties, as they typically link predictors to target values using the conditional mean function. Thus, a critical evolution in EPF has been the transition from traditional point forecasts to probabilistic forecasts which offer a range of potential outcomes, quantifying uncertainty, and generating prediction intervals (PIs) that reflect the confidence levels of forecasts. Probabilistic Electricity Price Forecasting (PEPF) can capture the full range of potential future prices in dynamic, non-linear markets [2,3] addressing uncertainties in smart grids, supply–demand dynamics, and price variations emphasising the operational impact of forecasts [4].

Key methods in probabilistic forecasting include Quantile Regression (QR), Quantile Regression Averaging (QRA), Bayesian, Historical Simulation, and Bootstrapped PIs. However, these methods fail to consistently provide validity, particularly in non-stationary settings, which limits their reliability in high-stakes applications. Conformal Prediction (CP) has gained significant attention for its ability to provide valid and adaptive PIs by ensuring predefined confidence levels are met, regardless of the underlying data distribution or non-stationarity [5,6]. Recent advancements, such as Ensemble Batch Prediction Intervals (EnbPI) [7] and Sequential Predictive Conformal Inference (SPCI) [8], have successfully extended CP methodologies to time series applications, addressing the unique challenges of electricity markets. This focus on validity, combined with adaptability, positions CP as a robust framework for managing uncertainty in highly volatile and dynamic markets, making it a critical tool for risk management.

1.1. Contributions

Building on the existing literature in the field, this paper makes several notable contributions to the area of PEPF in the DAM and BM. While the IDM and BM share structural similarities, we focus primarily on the BM due to its fundamental role in balancing real-time supply and demand. The key contributions of our research are outlined as follows:

- *Application of QRA and CP to BM*: While QRA and Split Conformal Prediction (SCP) have been applied to the DAM, their use in the BM remains unexplored. Our study address this gap by applying these methods to both markets, exploring their performance and how combining QRA with CP impacts model accuracy and reliability.
- *Introduction of Time Series Adapted CP Approaches*: We employ EnbPI and SPCI in both DAM and BM contexts. To our knowledge, these CP approaches have not been previously applied to any PEPF context. Our research demonstrates the potential benefits of these methods in providing reliable PIs.
- *Introduction of Q-Ens*: We propose a novel ensemble method that combines predictions from QR, EnbPI, and SPCI models. This approach aims to balance the superior efficiency of QR with the superior validity of CP methods, offering a robust solution in PEPF.
- *Evaluation and Trading Impact*: We assess the efficiency and validity of forecasts using various probabilistic metrics, and investigate the importance of coverage guarantees for trading outcomes in both DAM and BM. Our financial analysis includes both single-trade and high-frequency trading strategies, examining the impact of probabilistic forecasts on trading decisions and outcomes.

This research advances the field by introducing innovative CP and QRA applications, proposing a new ensemble approach, and comprehensively evaluating probabilistic forecasts in electricity markets. We provide the accompanying dataset and Python code for forecasting and trading, facilitating the adoption and replication of our methods. Our findings highlight the strengths of CP methods in ensuring coverage guarantees and benefiting trading strategies.

The structure of this paper is as follows: Section 2 provides an overview of recent advancements in PEPF within the context of the DAM and BM. Section 3, we detail our methodological framework, including our forecasting models and probabilistic approaches. Section 4 discusses probabilistic metrics used to evaluate the efficiency and validity of our forecasts. The effectiveness of each forecasting approach, especially the proposed ensemble, is discussed through case study results in Section 5. Finally, Section 6 summarises our findings.

2. Related work

EPF has emerged as a critical research area due to the increasing complexity and volatility of electricity markets. Accurate forecasting is essential for market participants to make informed decisions, mitigate financial risks, and optimise operations. Probabilistic forecasting methods, in particular, address the inherent uncertainties in these markets by providing PIs rather than single-point estimates, allowing stakeholders to assess the range of potential outcomes. Table 1 provides a comprehensive summary of the literature on probabilistic EPF methods, highlighting the evolution from traditional point forecasts to advanced probabilistic techniques, including QR and CP. This section critically examines the role of forecasting in electricity spot markets, foundational concepts of uncertainty in EPF, recent advancements in probabilistic methods, and the growing role of CP as an alternative framework, while evaluating their strengths, limitations, and practical implications.

2.1. Electricity spot markets

The DAM is a cornerstone of electricity trading, where market participants submit bids for electricity trades up to one day in advance, with the bidding process closing in the afternoon to determine market-clearing prices and volumes based on supply, demand, and grid constraints [9,10]. While the DAM enhances cross-border electricity trading and grid stability, it faces challenges from forecast uncertainties driven by the increasing penetration of intermittent renewable energy sources such as wind and solar [11]. These challenges necessitate advanced forecasting models capable of managing the variability introduced by weather-dependent renewables, such as privacy-preserving frameworks that enhance accuracy by integrating diverse data sources [12]. Initially dominated by traditional statistical methods, forecasting has shifted toward advanced machine learning (ML) techniques, which excel at capturing the complex, non-linear relationships inherent in electricity markets [13–16]. Hybrid approaches combining ML with other techniques have further improved precision, particularly in modelling supply–demand dynamics and adapting to varying market conditions [17,18].

The IDM is crucial for real-time electricity trading, bridging the DAM and BM by allowing continuous trading up to a few hours before delivery. This enables participants to adjust their positions in response to unforeseen events, such as plant outages or fluctuations in renewable energy output [19]. In contrast to the DAM's hourly settlement intervals, the IDM allows trades in 30 min blocks, notably in markets like the Irish Single Electricity Market. The IDM's real-time nature necessitates advanced forecasting models capable of accurately predicting short-term price movements. These challenges necessitate advanced forecasting models capable of managing the variability introduced by weather-dependent renewables [20], with explainable frameworks helping ensure adaptability and reliability in dynamic conditions [21]. Traditional statistical methods [22,23] have shown success in processing large datasets and adapting to the market's dynamic conditions. Additionally, in markets with a high penetration of renewable energy, the IDM optimises trading strategies by enabling participants to refine their actions in near real-time, thereby mitigating risks associated with price volatility [24,25]. Although the BM shares several structural similarities with the IDM — particularly in their real-time nature and focus on short-term trading — our work centres primarily on the BM due to its finer granularity and its role in balancing supply and demand in real-time.

Despite the importance of the BM in maintaining equilibrium between supply and demand, the literature on BM forecasting is relatively underdeveloped compared to that of the DAM and IDM [26]. The BM plays a key role in balancing supply and demand with finer granularity than the IDM, particularly when responding to the fluctuations associated with renewable energy [1]. Accurate BM price forecasts are essential, especially where large-scale energy storage is not viable [27]. Studies have compared various forecasting models and their performance in BM settings, typically with limited horizons. Probabilistic forecasting methods for BM prices, such as those proposed by [28], show promise but suggest further enhancements and integration into bidding strategies. Notably, existing research often lacks critical analysis and broader context. Recent studies, including [29,30], explore diverse models for BM price forecasting, with simpler models often outperforming the more complex ML and DL models. These studies highlight the need for continued research to address current limitations and explore emerging methodologies, ensuring sharp and reliable BM price predictions. Given the challenges of renewable energy integration and market volatility, European electricity market reforms increasingly prioritise resilient mechanisms [31].

Table 1
Comparison of the literature on *Point* and *Probabilistic* EPF. The ✓ means that the paper includes that method of prediction. *CP* considers all papers using Conformal Prediction approaches. *Day-ahead* represents works focusing on DAM. Finally, *Real-time* is used to denote papers on IDM and BM. The table helps to position this work in the current literature.

Overview of the existing literature					
Reference	Point	Probabilistic	CP	Day-ahead	Real-time
[32]		✓		✓	
[33]		✓			✓
[34]		✓		✓	
[33]		✓			✓
[14]	✓			✓	
[2]		✓		✓	
[24]					✓
[19]					✓
[35]	✓			✓	
[28]		✓			✓
[22]	✓				✓
[23]		✓			✓
[36]	✓				✓
[37]		✓		✓	
[17]	✓			✓	
[38]		✓		✓	
[39]		✓	✓	✓	✓
[29]		✓			✓
[4]		✓			
[40]		✓	✓		
[41]		✓	✓		
[42]		✓	✓		
[43]		✓			
[44]		✓		✓	
[45]		✓	✓		
[46]		✓	✓		
[47]		✓		✓	
[48]		✓	✓		
[30]	✓			✓	✓
[49]		✓		✓	
[50]		✓	✓	✓	
[51]		✓	✓		
[52]		✓	✓		
[53]		✓			
Our approach		✓	✓	✓	✓

2.2. Probabilistic methods

Probabilistic forecasting methods have evolved substantially to meet the need for robust PIs in volatile electricity markets. Techniques such as Bayesian approaches, bootstrap methods, distribution-based forecasts, Monte Carlo simulations, and historical simulations each bring unique strengths but also face practical limitations. Bayesian methods effectively quantify uncertainty by combining prior knowledge with observed data, as seen in advancements like Bayesian neural networks (BNNs), stochastic volatility models, and hybrid Bayesian optimisation [35,54,55]. However, their high computational demands and reliance on strong priors limit scalability, particularly in non-stationary conditions. Bootstrap methods, such as residual-based bootstrapping, improve empirical coverage in BM settings but often produce overly wide intervals during volatile periods [56,57]. Distribution-based forecasts, including copula processes and generalised additive models, capture extreme price behaviours but are constrained by their reliance on correct distributional assumptions [58–60]. Monte Carlo simulations, enhanced through Bayesian networks and sequential Markov Chain Monte Carlo techniques, improve real-time adaptability but remain computationally prohibitive for high-frequency trading [61]. Historical simulations offer simplicity and robustness but struggle to address structural shifts or unprecedented events, particularly in markets integrating renewable energy [62,63]. Recent studies have explored the use of meta-heuristic optimisation techniques for forecasting models in short-term load forecasting [64].

These limitations have driven interest in Conformal Prediction (CP) techniques, which offer a flexible, distribution-free framework for generating reliable PIs. CP avoids the restrictive assumptions of traditional methods while ensuring predefined confidence levels, even in non-stationary markets [5,6]. Recent advancements, such as Ensemble Batch Prediction Intervals (EnbPI) and Sequential Predictive Conformal Inference (SPCI), extend CP's utility to time-series applications [8]. Despite these strengths, CP's reliance on large calibration datasets and sensitivity to structural breaks can undermine performance in dynamic electricity markets [65]. Further research should focus on enhancing CP's adaptability and robustness, particularly for handling extreme price events and evolving market complexities.

Quantile regression & conformal prediction

QR and QRA are widely used in PEPF for their simplicity, efficiency, and ability to capture uncertainty across quantiles. Recent advancements have refined these techniques, such as Lasso QRA and smoothed QRA with kernel estimation, which improve predictive performance [38,43]. DL integrations have also shown promise, with methods like QR combined with LSTNet and SHAP-based feature selection enhancing sharpness through kernel density estimation [66], or hybrid QR-LSTM frameworks optimised via multi-objective algorithms improving accuracy under diverse market conditions [67]. While effective, QR and QRA face challenges in uncertainty quantification, particularly in achieving reliable coverage, motivating the exploration of CP as a complementary framework.

CP generates distribution-free PIs with predefined confidence levels, making it a robust alternative for uncertain and volatile markets [5,6]. Extensions like Ensemble Batch Prediction Intervals (EnbPI), which combines ensemble learning with block bootstrap methods, have demonstrated success in addressing seasonality and trends in energy forecasting [7,42,68]. Similarly, Sequential Predictive Conformal Inference (SPCI) dynamically recalibrates PIs to improve reliability in non-stationary settings, with Transformer-based extensions further enhancing its adaptability [51]. Hybrid approaches such as Ensemble Conformal QR and SCP with QR, offer a balance between computational efficiency and robust coverage, making them suitable for real-time applications [40,48]. However, CP still faces practical limitations, including sensitivity to structural breaks and reliance on large calibration datasets [65]. Enhancing CP's robustness, addressing extreme price events, and integrating CP with DL architectures to improve its accuracy and adaptability in dynamic electricity markets have proven beneficial [52,53].

3. Methodology

This section outlines our study's methodology, emphasising the key models and forecasting approaches. It is structured as follows: Section 3.1 introduces probabilistic forecasting models, followed by Section 3.2, which describes the proposed uncertainty quantification approaches and our proposed ensemble, Q-Ens.

3.1. Electricity price prediction models

The probabilistic regressor models benchmarked include:

- *LASSO Estimated AR* (LEAR): An autoregressive model with LASSO regularisation for feature selection and handling high-dimensional data, providing probabilistic predictions through conditional distribution estimation.
- *K-Nearest Neighbours* (KNN): A non-parametric model predicting outputs based on the "K" nearest neighbours, with probabilistic predictions derived from the neighbours' output distribution.
- *Random Forest* (RF): An ensemble model combining multiple regression trees to reduce variance and maintain low bias, generating probabilistic predictions by aggregating tree outputs.
- *Light Gradient Boosting Method* (LGBM): An ensemble model using boosting to reduce bias and variance, providing probabilistic forecasts by summing weighted outputs of weak learners (trees). We chose LGBM over XGBoost due to its leaf-wise tree growth strategy, which typically yields deeper, more accurate trees for large, complex datasets, while also offering faster training and lower memory usage.

3.2. Uncertainty estimation

In the domain of EPF, probabilistic forecasting methods offer a significant advancement over traditional point forecasts by providing insights into the uncertainty surrounding future prices. In this section, we delve into various methods for PEPF in both the DAM and BM. We outline the approaches utilised, including QR, SCP, EnbPI, SPCI, QRA, and our proposed QR-CP Ensemble (Q-Ens). Unlike confidence intervals that quantify uncertainty over mean estimates, these probabilistic forecasting methods produce PIs that reflect uncertainty around future prices, calibrated to achieve a target coverage rate.

We frame the price prediction problem as a regression task. Given a dataset $\mathcal{Z} = \{(\mathbf{x}_t, y_t)\}_{t=1}^{N}$, where $\mathbf{x}_t$ represents the covariates and y_t the response variable, our objective at each time step t is to predict the next price y_t based on previous prices and covariates $\mathbf{x}_t$. The prediction is made using a model f, such that:

$$\hat{y}_t = f(\mathbf{x}_t) \tag{1}$$

where $\hat{y}_t$ is the model's prediction. The models are trained on the historical dataset $\mathcal{Z}$, which consists of N input–output pairs. The conditional quantile function $Q_{y_t}(\alpha \mid \mathbf{x}_t)$ estimates the value below which a certain percentage α of the future price y_t is expected to fall, given the covariates $\mathbf{x}_t$. We denote the lower quantile at level α as $\hat{q}_t^{\alpha}$ and the upper quantile at level $1-\alpha$ as $\hat{q}_t^{1-\alpha}$. In the probabilistic prediction setting, as shown in Fig. 1, the model f generates a PI defined by these quantiles:

$$\Gamma_{1-2\alpha}(\mathbf{x}_t) = [\hat{q}_t^{\alpha}, \hat{q}_t^{1-\alpha}] \tag{2}$$

This interval indicates that the future price y_t is expected to fall within the range $[\hat{q}_t^{\alpha}, \hat{q}_t^{1-\alpha}]$ with a confidence level of $1-2\alpha$. For example, if $\alpha = 0.1$, the PI corresponds to an 80% confidence level ($1-2\alpha = 0.8$), implying that there is an 80% probability that y_t will lie within the interval.

3.2.1. Quantile regression

QR, introduced in [69], extends ordinary least squares regression to estimate conditional quantiles of the response variable rather than just the mean. QR provides a more comprehensive view by estimating the conditional median or other quantiles, revealing how covariates influence the entire distribution of the response variable. Consider a linear model of the form:

$$y_t = \mathbf{x}_t^{\top}\beta + \epsilon_t \tag{3}$$

where y_t is the dependent variable for the tth observation, $\mathbf{x}_t$ is a K-dimensional vector of covariates, β is a K-dimensional vector of coefficients, and ϵ_t is the error term.

The goal of QR is to estimate the conditional quantile function, which represents the value below which a certain percentage α of the response variable y_t falls, given the predictor variable $\mathbf{x_t}$. For a given α quantile level, the QR estimator $\hat{\beta}_\alpha$ is defined as the solution to the following optimisation problem:

$$\hat{\beta}_\alpha = \arg\min_{\beta} \sum_{t=1}^{N} \rho_\alpha(y_t - \mathbf{x}_t^{\top}\beta) \tag{4}$$

where $\rho_\alpha(u)$ is the check function defined by:

$$\rho_\alpha(u) = u(\alpha - \mathbf{1}\{u < 0\}) \tag{5}$$

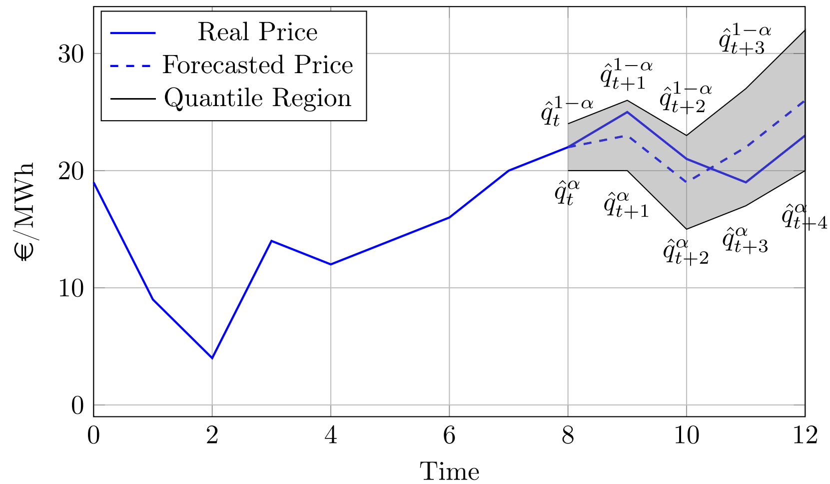


Fig. 1. Quantile forecast of electricity prices.

Here, $\mathbf{1}\{u < 0\}$ is an indicator function that equals 1 if $u < 0$ and 0 otherwise. The check function can be rewritten as:

$$\rho_\alpha(u) = \begin{cases} \alpha u & \text{if } u \geq 0 \\ (\alpha - 1)u & \text{if } u < 0 \end{cases} \tag{6}$$

This function penalises underestimation and overestimation asymmetrically, depending on the value of α. Specifically, it assigns a weight of α to positive residuals and $1 - \alpha$ to negative residuals, allowing the estimation of different quantiles.

The asymptotic properties of the QR estimator $\hat{\beta}_\alpha$ are well-established under suitable regularity conditions. The estimator $\hat{\beta}_\alpha$ is consistent, meaning it converges to the true parameter value as the sample size increases, and it is asymptotically normally distributed, approaching a normal distribution as the sample size grows. The asymptotic covariance matrix of $\hat{\beta}_\alpha$ is given by:

$$\text{Cov}(\hat{\beta}_\alpha) = \alpha(1 - \alpha) \left[f_{y_t|\mathbf{x}_t}(\hat{q}_t^\alpha)^2 \cdot (X'X)^{-1} \right] \tag{7}$$

where $\alpha(1-\alpha)$ adjusts for the variability in estimating the quantile level α, $f_{y_t|\mathbf{x}_t}(\hat{q}_t^\alpha)$ is the conditional density of y_t given $\mathbf{x}_t$ at the α quantile level, reflecting the density around this quantile, and $(X'X)^{-1}$ is the inverse of the product of the design matrix with its transpose. The design matrix X contains the values of all predictor variables for each observation $X = \{(\mathbf{x}_t)\}_{t=1}^N$, and its inverse accounts for the scaling and correlations among these predictors.

Conformal prediction

CP is a flexible, distribution-free framework for quantifying the uncertainty of predictions made by machine learning models. CP generates prediction sets with a predefined confidence level, ensuring that the true value lies within these sets with a specified probability. At the core of CP is the concept of nonconformity, which measures how unusual a new data point $(\mathbf{x}_{\text{new}}, y_{\text{new}})$ is compared to a set of previously observed examples. A nonconformity measure A is defined to assign a nonconformity score to each data point in the training set $\mathcal{Z}$. Traditional CP methods use p-values calculated from nonconformity scores to construct prediction sets. For a new example $\mathbf{x}_{\text{new}}$ with a candidate label y_{new}, the p-value is given by: $p(y_{\text{new}}) =$

$$\frac{|\{(\mathbf{x}_t, y_t) \in \mathcal{Z} : A(\mathbf{x}_t, y_t) \geq A(\mathbf{x}_{\text{new}}, y_{\text{new}})\}| + 1}{N + 1} \tag{8}$$

where N is the size of $\mathcal{Z}$. The prediction set for $\mathbf{x}_{\text{new}}$ includes all candidate labels for which the p-value exceeds a certain significance level.

3.2.2. Split conformal prediction

SCP, also known as inductive CP, is a simplified variation of standard conformal prediction that enhances computational efficiency by dividing the data into two separate sets: a training set and a calibration set. The training set is used to fit a predictive model, while the calibration set is used to generate nonconformity scores, which measure the discrepancy between the true and predicted values. The data is split into two subsets: a training set $\mathcal{Z}_{\text{train}}$ and a calibration set $\mathcal{Z}_{\text{cal}}$. A predictive model f is trained on $\mathcal{Z}_{\text{train}}$, and the model's performance is assessed on the calibration set by computing nonconformity scores for each pair $(\mathbf{x}_t, y_t) \in \mathcal{Z}_{\text{cal}}$.

The nonconformity score for an observation t is defined as:

$$A(\mathbf{x}_t, y_t) = |y_t - f(\mathbf{x}_t)| \tag{9}$$

This score measures the absolute difference between the true value y_t and the predicted value $f(\mathbf{x}_t)$, with larger scores indicating higher nonconformity between the model's prediction and the observed outcome. Once nonconformity scores are computed for all calibration examples, a quantile of the nonconformity scores is selected to construct PIs for new data points. Let λ_t be the quantile for time period t, corresponding to the desired confidence level $1 - 2\alpha$. This ensures that at least $1 - 2\alpha$ of the calibration examples have nonconformity scores smaller than or equal to λ_t, guaranteeing that the PIs will cover the true value with confidence level $1 - 2\alpha$.

For a new example $\mathbf{x}_t$, the PI is constructed as:

$$\Gamma_{1-2\alpha}(\mathbf{x}_t) = \left[f(\mathbf{x}_t) - \lambda_t, f(\mathbf{x}_t) + \lambda_t \right] \tag{10}$$

This interval represents the range of possible values for the response variable y_t that are consistent with the nonconformity scores computed from the calibration set. SCP balances computational efficiency and prediction robustness by calculating nonconformity scores on a smaller calibration set, avoiding the need to recompute predictions for each test example. This reduces the computational load but assumes data exchangeability, which limits its applicability in settings like time series data.

3.2.3. Ensemble Batch Predictive Interval (EnbPI)

EnbPI, introduced in [7], is a probabilistic method within the CP framework, specifically designed to generate PIs with strong coverage guarantees. EnbPI leverages ensemble learning techniques and bootstrap sampling to provide reliable PIs, particularly for forecasting applications. Unlike traditional CP methods that require a separate calibration set, EnbPI adopts a leave-one-out approach, which enables the method to maintain valid coverage without sacrificing data for calibration. Given a dataset $\mathcal{Z}$, EnbPI constructs PIs by first generating

B bootstrap samples from the original dataset. A predictive model $f^{(b)}$ is trained on each bootstrap sample $\mathcal{Z}^{(b)}$, allowing EnbPI to account for model variability and provide more reliable PIs. For each trained model $f^{(b)}$, residuals are computed for every observation $(\mathbf{x}_t, y_t)$ by calculating the absolute difference between the observed value y_t and the predicted value $f^{(b)}(\mathbf{x}_t)$. The residuals are given by:

$$A^{(b)}(\mathbf{x}_t, y_t) = |y_t - f^{(b)}(\mathbf{x}_t)| \tag{11}$$

These residuals reflect the model's prediction errors and are used to determine the width of the PIs. The quantile $\lambda_t^{(b)}$ for each bootstrap model is defined as:

$$\lambda_t^{(b)} = \text{Quantile}_{1-2\alpha}(A_1^{(b)}, A_2^{(b)}, \dots, A_N^{(b)}) \tag{12}$$

The PI for a new input $\mathbf{x}_t$ is constructed as $\Gamma_{1-2\alpha}(\mathbf{x}_t)$, for a confidence level of $1-2\alpha$, by averaging the predictions from all bootstrap models and incorporating the margin λ_t. The resulting PI is given by:

$$\Gamma_{1-2\alpha}(\mathbf{x}_t) = \left[\frac{1}{B}\sum_{b=1}^{B} f^{(b)}(\mathbf{x}_t) - \lambda_t, \frac{1}{B}\sum_{b=1}^{B} f^{(b)}(\mathbf{x}_t) + \lambda_t\right] \tag{13}$$

where λ_t is the average of the quantiles $\lambda_t^{(b)}$ across all bootstrap models:

$$\lambda_t = \frac{1}{B}\sum_{b=1}^{B} \lambda_t^{(b)} \tag{14}$$

By averaging across multiple models and adjusting the interval width using the quantile of residuals, EnbPI mitigates the effects of outliers and reduces the variance in PIs. Its flexibility allows use across various models and data types, making it ideal for complex EPF. The use of ensemble learning also improves the resilience of the PIs.

3.2.4. Sequential Predictive Conformal Inference (SPCI)

SPCI, introduced in [8], extends the EnbPI method to handle non-exchangeable time series data. Recognising the limitations of traditional CP methods in the context of temporal dependencies, SPCI incorporates quantile regression forests (QRF) to estimate conditional quantiles of residuals. This allows SPCI to better adapt to the temporal dependencies and heteroskedasticity often encountered in time series data. SPCI works by training a predictive model f on the dataset $\mathcal{Z}$. The model's residuals are computed using Eq. (9) on a rolling calibration set and used to estimate future residuals through QR. By applying QRF to the residuals, SPCI estimates the conditional quantile $\hat{q}_t^{\alpha}$, where $1-2\alpha$ represents the desired confidence level. This approach helps SPCI account for the serial dependence in the residuals, which is critical for accurate PIs in time series forecasting. To generate PIs for a new input $\mathbf{x}_t$, SPCI constructs the interval:

$$\Gamma_{1-2\alpha}(\mathbf{x}_t) = \left[f(\mathbf{x}_t) - \lambda_t, f(\mathbf{x}_t) + \lambda_t\right] \tag{15}$$

where λ_t is derived from the quantiles of past residuals, adjusted for the level of uncertainty in the data. One of the main advantages of SPCI is its ability to dynamically re-estimate residual quantiles as new data becomes available. This adaptive mechanism allows the method to maintain valid coverage over time, ensuring that the true outcome falls within the PI with a specified confidence level $(1-2\alpha)$. Furthermore, SPCI is applicable to a wide range of predictive models and data types, making it highly versatile for complex time series forecasting tasks. Compared to SCP and EnbPI, SPCI often yields narrower intervals without sacrificing coverage, offering a powerful tool for uncertainty quantification in time series.

3.2.5. Quantile regression averaging

QRA was introduced in [32]. QRA applies QR to a set of point forecasts from different models, allowing direct estimation of electricity spot price distribution without separating the probabilistic forecast into a point forecast and an error distribution. This method directly leverages electricity price distribution properties for improved forecasting accuracy. It considers M different point prediction models f^m and feeds their output to a probabilistic model:

$$\Gamma_{1-2\alpha}(\mathbf{x}_t, [f^1(\mathbf{x}_t), f^2(\mathbf{x}_t), \dots, f^M(\mathbf{x}_t)]) = [\hat{q}_t^{\alpha}, \hat{q}_t^{1-\alpha}] \tag{16}$$

In our implementation, two sets of point forecasts are generated:

- *QRA-R:* Includes forecasts from models trained on different data contexts (full datasets, historical, and future datasets) to capture diverse trends and improve reliability.
- *QRA-CP:* Consists of forecasts from various Conformal Prediction (CP) models like SCP, EnbPI, and SPCI, each contributing unique strengths to quantile predictions.

3.2.6. QR-CP ensemble: A robust approach

In predictive modelling, uncertainty is categorised into epistemic and aleatoric types. Epistemic uncertainty arises from limited knowledge of the underlying model and can be reduced with more data or model improvements, while aleatoric uncertainty results from inherent variability in the data and cannot be mitigated by additional data. Our Q-Ens combines QR, EnbPI, and SPCI aims to address both uncertainties, improving the reliability and sharpness for PEPF. QR captures a wide range of possible outcomes by estimating multiple conditional quantiles, thus addressing both epistemic and aleatoric uncertainties. EnbPI reduces epistemic uncertainty by averaging predictions from bootstrap samples, thereby lowering variance and reflecting aleatoric uncertainty in the constructed PIs. SPCI adapts dynamically to new data, updating PIs in real-time to account for evolving data variability, thus capturing both types of uncertainty as they change over time.

The ensemble combines the strengths of these methods, working from a similar approach as EnbPI, averaging the models outputs to both reduce epistemic uncertainty further, and improve overall forecast performance. The ensemble prediction is obtained by averaging the quantile predictions from QR, EnbPI, and SPCI models: $\Gamma_{1-2\alpha} =$

$$[\frac{1}{3}(\text{QR}^{\alpha} + \text{EnbPI}^{\alpha} + \text{SPCI}^{\alpha}), \frac{1}{3}(\text{QR}^{1-\alpha} + \text{EnbPI}^{1-\alpha} + \text{SPCI}^{1-\alpha})] \tag{17}$$

with an outline of the PIs construction shown in Fig. 2.

4. Evaluation metrics

Our evaluation of probabilistic predictions hinges on two crucial dimensions: validity and efficiency. Efficiency, gauged through metrics like sharpness and interval width, enhances precision. Simultaneously, validity, including calibration and coverage, requires the generation of precise PIs to affirm the reliability of our forecasts.

4.1. Efficiency

Efficiency in probabilistic forecasting is crucial for precision. It is evaluated using two key metrics: the Pinball Score and Interval Width.

4.1.1. Pinball score

In both the DAM and BM, sharpness plays an important role in accurate anticipation, hedging, and real-time adjustments. The Pinball Score [70], derived from the Pinball Loss function, measures the sharpness of the forecast based on the quantile prediction $\hat{q}_t^{\alpha}$ and the observed price y_t for time period t:

$$PS(\hat{q}_t^{\alpha}, y_t, \alpha) = \begin{cases} (1-\alpha)(\hat{q}_t^{\alpha} - y_t) & \text{for } y_t \le \hat{q}_t^{\alpha} \\ \alpha(y_t - \hat{q}_t^{\alpha}) & \text{for } y_t \ge \hat{q}_t^{\alpha} \end{cases} \tag{18}$$

Here, α denotes the quantile level, such as 0.1 for the 10th percentile; hence, an 80% PI ($[\hat{q}_t^{\alpha}, \hat{q}_t^{1-\alpha}]$) corresponds to quantile levels 0.1 and 0.9.

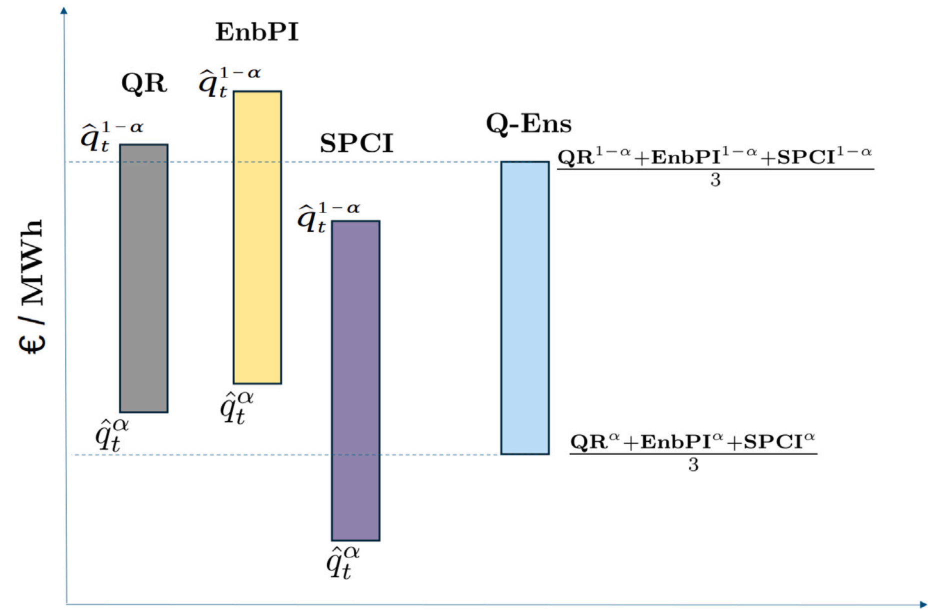


Fig. 2. Construction of the ensemble, Q-Ens.

4.1.2. Interval width

The Interval Width is another important metric for assessing the efficiency of probabilistic forecasts. While sharpness ensures reliability, efficiency is pivotal in refining precision, and it is intricately linked to the width of PIs. The Interval Width (IW_t) for a given PI can be defined as follows:

$$IW_t = \hat{q}_t^{1-\alpha} - \hat{q}_t^{\alpha} \tag{19}$$

where $\hat{q}_t^{1-\alpha}$ and $\hat{q}_t^{\alpha}$ are the upper and lower quantiles of the PI, respectively. Narrower intervals indicate higher efficiency, reflecting greater precision in the forecast.

4.2. Validity

Validity in probabilistic forecasting ensures that the predicted intervals accurately reflect the true distribution of the observed data. Two primary metrics used to evaluate validity are coverage and the Winkler Score.

4.2.1. Coverage

Both reliability and accuracy of probabilistic forecasts are crucial, with a particular focus on evaluating their validity. We examine the models' coverage to highlight their ability to capture the true distribution, focusing on precision in capturing the price y_t within a predefined PIs. The empirical coverage, indicating the agreement of predictions with the specified intervals, is denoted by the binary indicator I_t^{α}:

$$I_t^{\alpha} = \begin{cases} 1 & \text{if } y_t \in [\hat{q}_t^{\alpha}, \hat{q}_t^{1-\alpha}] \\ 0 & \text{if } y_t \notin [\hat{q}_t^{\alpha}, \hat{q}_t^{1-\alpha}] \end{cases} \tag{20}$$

The Coverage Probability (Λ) is then defined as:

$$\Lambda = \frac{1}{N} \sum_{t}^{N} I_t^{\alpha} \tag{21}$$

where N is the total number of forecasts. Ideally, the coverage probability should be close to the nominal level (e.g., 80% if $\alpha = 0.1$), indicating that the PIs are valid.

4.2.2. Winkler score

The Winkler Score (W_t) combines reliability and sharpness into a single metric, offering a comprehensive assessment of probabilistic forecasts. This score evaluates the interval width (IW_t), see Eq. (19), of PIs based on the observed values (y_t), incorporating a penalty factor (τ) for deviations from the interval bounds:

$$W_t = \begin{cases} IW_t & \text{if } y_t \in [\hat{q}_t^{\alpha}, \hat{q}_t^{1-\alpha}] \\ IW_t + \frac{2}{\tau}(\hat{q}_t^{\alpha} - y_t) & \text{if } y_t < \hat{q}_t^{\alpha} \\ IW_t + \frac{2}{\tau}(y_t - \hat{q}_t^{1-\alpha}) & \text{if } y_t > \hat{q}_t^{1-\alpha} \end{cases}$$

The Winkler Score penalises larger deviations when observed prices fall outside the PIs, offering insight into the model's balance of accuracy and interval width.

5. Case studies

We present two case studies to demonstrate the effectiveness of our QR-CP Ensemble (Q-Ens) approach for PEPF in the DAM and BM. Utilising datasets from [30] spanning 2019 to 2022, we focus on the Irish electricity market, which is characterised by high curtailment rates due to surplus renewable generation during low-demand periods [71]. To account for key time series characteristics, we performed an exploratory decomposition of the data to assess trend, seasonality, and noise. Although long-term trends were minimal, we captured structural effects and temporal dependencies using lagged prices, system forecasts, and autoregressive terms. Intraday and weekly seasonality were modelled using calendar-based features such as hour-of-day, day-of-week, and month, which allow the models to learn recurring demand–supply cycles. A full decomposition and exploratory analysis is included in the public repository to support transparency and reproducibility. This section evaluates probabilistic approaches and forecasting models in both the DAM 5.1 and the BM 5.2. We examine the effectiveness of each model, looking at QR and CP approaches, with a focus on the performance of our proposed QR-CP Ensemble. The analysis can be divided into two main aspects: efficiency and validity. Efficiency evaluates the sharpness of the forecasts by examining metrics such as the Aggregate Pinball Score (APS) and interval width. Validity assesses the reliability of the forecasts by measuring each approach's ability to achieve coverage. These two aspects are integrated using the Winkler Score, which ties together efficiency and validity. This culminates in a financial performance evaluation of these probabilistic methods in Section 5.3. To ensure reproducibility, the Python code and

Table 2
APS for the DAM. Best approach for each regressor marked in bold, worst in red.

Model	QR	SCP	EnbPI	SPCI	QRA-R	QRA-CP	Q-Ens
KNN	6.65	7.79	**5.71**	6.09	6.20	6.01	5.96
LEAR	8.35	4.81	4.18	**4.08**	8.36	8.35	4.54
LGBM	3.62	3.67	3.76	3.71	3.67	3.66	**3.48**
RF	3.63	3.85	3.81	3.84	3.60	3.61	**3.59**
Avg.	5.56	5.03	**4.37**	4.43	5.43	5.36	4.39

Table 3
Winkler scores for the DAM. Best approach for each regressor marked in bold, worst in red.

Model	QR	SCP	EnbPI	SPCI	QRA-R	QRA-CP	Q-Ens
KNN	57.02	64.16	51.47	56.16	53.70	**51.73**	54.33
LEAR	61.97	36.30	37.37	**35.06**	61.82	61.74	39.37
LGBM	29.87	30.39	32.76	31.31	31.5	31.47	**29.79**
RF	33.7	30.81	32.14	31.65	33.00	32.36	**29.99**
Avg.	42.94	40.42	38.43	38.55	45.01	44.33	**38.37**

a full decomposition and exploratory analysis of the dataset used for this section are available at: GitHub.[1]

5.1. Day-ahead market

The DAM dataset includes historical data and the system operator (TSO) predictions, revealing past market behaviours and aiding in forecasting future conditions. In this case study, we predict DAM prices for the next 24 settlement periods using historical DAM prices and TSO wind and demand forecasts from the past 168 h. This one-day-ahead forecasting horizon aligns with the market design of the DAM, where participants submit bids by 12:00 CET for delivery the following day. Forecasts at this horizon are the most actionable, as they benefit from the most up-to-date information, which is particularly important in systems with high shares of renewable generation, such as the Irish and broader European markets, where short-term volatility in wind and demand significantly affects bidding strategies.

5.1.1. Efficiency in the day-ahead market: Pinball score & interval width

Analysing the APS for DAM models in Table 2 highlights the strong performance for the Q-Ens approach, producing the two lowest APS in the DAM for the RF and LGBM models.

This includes strong performance for CP approaches EnbPI and SPCI, in particular improving LEAR and KNN models, taking advantage of CPs model agnostic feature. Time series adapted CP approaches EnbPI and SPCI also demonstrate notable improvements over SCP. QR and QRA models struggle, but this is largely as a result of their poor performance with LEAR, substantially dragging down the average.

Looking further into efficiency, we evaluate the interval width of the models for the DAM in Fig. 3, centred around a median of 50 for clarity. The analysis of IW across different models reveals key insights into model efficiency and precision. A pronounced contrasting outcome in Fig. 3 between QR and CP approaches is how CP approaches, SCP, EnbPI, and SPCI, result in a wider interval width for less accurate models KNN and LEAR, and a narrowing interval width for the more accurate RF and LGBM. SPCI, developed as a successor to EnbPI to produce narrower intervals, fails to achieve this for KNN in the DAM but succeeds slightly for LEAR, LGBM, and RF. This trend is absent entirely for QR and QRA approaches with QR and QRA producing intervals independently of models accuracy, with QRA-CP producing similarly sized Interval widths compared to QRA-QR, despite the contrasting contents of each approaches dataset. Q-Ens closely mirrors EnbPI and SPCI for LEAR and KNN models. However, for LGBM, Q-Ens produces narrower interval widths due to the model's inherently narrow intervals when using QR. Conversely, Q-Ens with RF produces wider intervals than EnbPI or SPCI due to the wider interval width of QR.

5.1.2. Validity in the day-ahead market: Coverage

Fig. 4 summarises coverage across the 0.1–0.9 quantile range in the DAM, highlighting each forecasting approaches ability to achieve coverage. In the DAM context, aiming for a coverage of 0.8 within the 0.1–0.9 quantile range, CP models consistently outperform QR counterparts. QR produces the worst results by a margin, failing to achieve coverage in a single model. Both QRA approaches begin to diverge from QR, improving coverage for most models, LGBM in particular, and achieving the desired 0.8 coverage for RF in both QRA-R and QRA-CP. Our Q-Ens approach yields slightly inferior results compared to EnbPI and SPCI. However, considering the lower APS for RF and LGBM models compared to EnbPI and SPCI, as shown in Table 2, this trade-off is justified.

Focusing on the 0.3–0.7 quantile pair with a target of 0.4 in Fig. 4, we see a similar trend to Fig. 4 with QR struggling and QRA improving QR for each model, with the exception of RF this time. All CP approaches achieve coverage for all models this time, highlighting the advantages of CP when it comes to improving a models reliability. Our Q-Ens achieves coverage for all models, with slightly lower scores than EnbPI and SPCI, but substantially better than QR, with the exception of RF (see Fig. 5).

5.1.3. Efficiency & validity in the day-ahead market: Winkler score

Table 3 provides a detailed overview of Winkler Scores, highlighting each approach's ability to tie together both efficiency and validity in our chosen forecasting models. Our approach Q-Ens excels producing the lowest average score, and the two lowest scores for all models with LGBM and RF. QR and QRA pay the price of their poor performance in coverage, producing the highest average winkler scores. CP approaches produce lower winkler scores, underscoring the benefits of their strong performance in coverage, with EnbPI strong performance just missing out on our Q-Ens approach.

5.2. Balancing market

The BM dataset include historical data and TSO predictions, revealing past market behaviours and aiding in forecasting future conditions. For the BM, we forecast prices for the next 16 open settlement periods starting at $t+2$, utilising historical BM prices, volumes, wind discrepancies, interconnector flows, and DAM prices from the past 24–48 h. We also incorporate future physical notifications, interconnector schedules, renewable forecasts, demand forecasts, and DAM prices for the next 8 h. The forecasting horizon is limited to 8 h ahead starting from $t+2$, in line with SEMO's publication schedule for forecast data and market rules, which allow participants to adjust balancing market positions only for settlement periods beyond $t+1$. Despite similar average prices with the DAM (BM: €54.772/MWh, DAM: €53.77/MWh), the BM exhibits higher price volatility (standard deviation: €64.60/MWh) compared to the DAM (€37.38/MWh). More Details surrounding both the DAM and BM datasets can be found at [30].

5.2.1. Efficiency in the balancing market: Pinball score & interval width

Analysing the APS for BM models in Table 4 follows a similar trend to Table 2 with Q-Ens showing strong performance producing the lowest average this time, except we see EnbPI and SPCI, two of the top approaches for the DAM, produce the two highest average scores. SCP produces markedly better scores, with QR and QRA approaches also performing better than EnbPI and SPCI. The two lowest scores are achieved by LGBM with QRA-R and RF with QR.

Examining efficiency in the BM, we evaluate the interval width of the models in Fig. 6. Compared to the average interval width for the

[1] https://github.com/ciaranoc123/PEPF_Conformal.

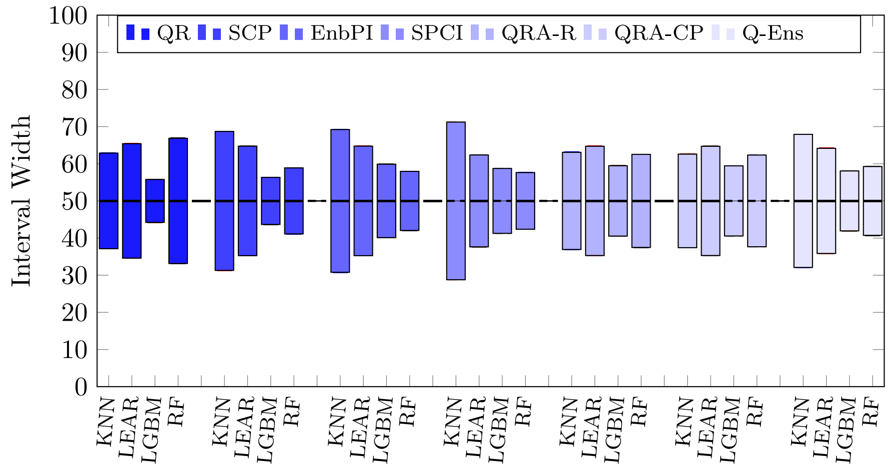


Fig. 3. Interval width for each model in the DAM.

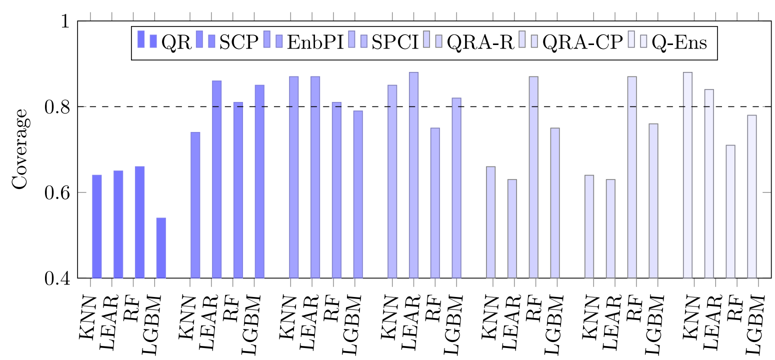


Fig. 4. DAM coverage for 0.1–0.9 quantile pair.

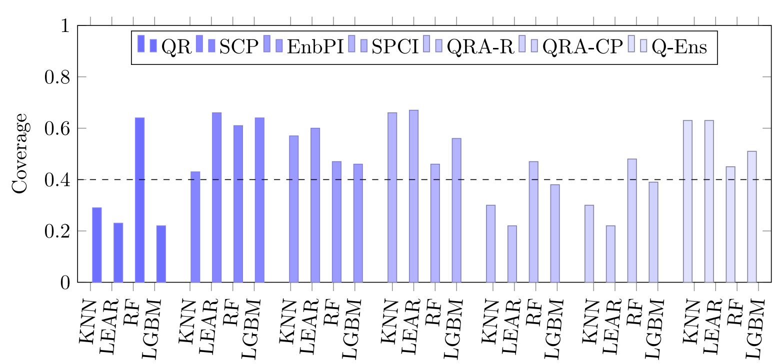


Fig. 5. DAM coverage for 0.3–0.7 quantile pair.

Table 4
APS for the BM. Best approach for each regressor marked in bold, worst in red.

Model	QR	SCP	EnbPI	SPCI	QRA-R	QRA-CP	Q-Ens
KNN	14.09	15.19	14.60	15.84	13.66	**13.49**	14.25
LEAR	15.22	**12.91**	13.96	13.79	15.33	15.22	13.21
LGBM	12.48	12.87	13.55	13.6	12.58	**12.41**	12.70
RF	**12.66**	13.01	13.34	13.48	12.79	12.89	12.88
Avg.	13.61	13.50	13.91	14.18	13.59	13.50	**13.26**

DAM in Fig. 3, all models in the BM have significantly wider interval widths, reflecting the greater inherent uncertainty in the BM. We see the same trend as the DAM with interval width being independent of model accuracy for QR and QRA approaches, but again a strong correlation between interval width and model accuracy for all CP approaches, albeit less pronounced for SPCI than in the DAM. SPCI produces narrower interval widths than EnbPI as expected, but much more pronounced than in the DAM. Q-Ens produces wider interval widths than SPCI and QR, skewed by EnbPI large interval width.

5.2.2. *Validity in the balancing market: Coverage*

Fig. 7 summarises coverage across the 0.1–0.9 quantile range in the BM. Aiming for a coverage of 0.8 within this range, all models in the BM come significantly closer to achieving the desired coverage compared to their performance in the DAM. CP approaches similarly excel, with all models achieving coverage, excluding LGBM for EnbPI and RF for SPCI, both just missing the coverage target. QRA again improves all QR models noticeably, with RF for both QRA approaches achieving coverage. Our Q-Ens fails to perform as well as any CP approach, but outperforms all QR and QRA models bar RF for QRA.

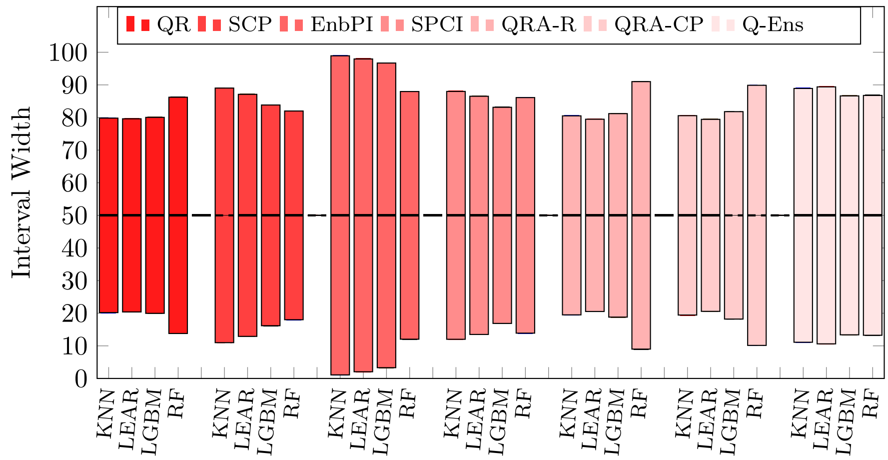


Fig. 6. Interval Width for each model in the BM.

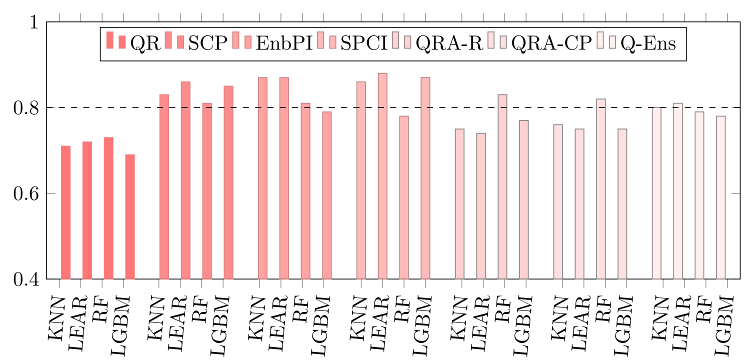


Fig. 7. BM coverage for 0.1–0.9 quantile pair.

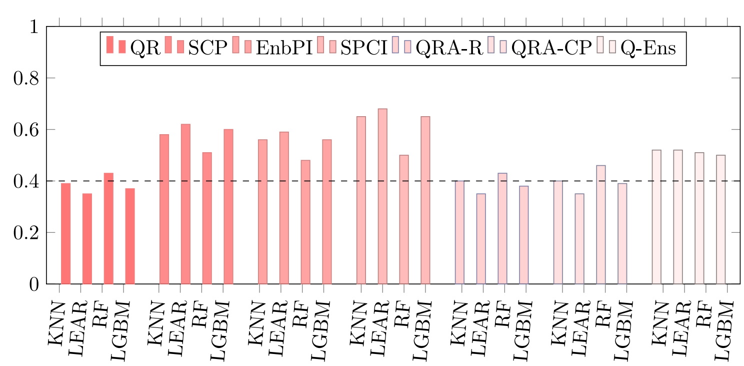


Fig. 8. BM coverage for 0.3–0.7 quantile pair.

Focusing on the 0.3–0.7 quantile pair aiming for a coverage of 0.4 in Fig. 8, again all CP approaches achieve coverage, including Q-Ens. QRA fails to demonstrate a noticeable improvement off of QR, but all models come close to achieving coverage, with RF for QR and both QRA achieving the 0.4 target.

5.2.3. Efficiency & validity in the balancing market: Winkler score

Table 5 provides a overview for Winkler Scores in the BM, offering insights into each approaches ability to tie efficiency and validity together. The analysis in Table 5 highlights the strong performance of our ensemble approach again, resulting in the second lowest winkler score, slightly underperforming QRA-CP. QR and QRA demonstrate considerably improved performance than they did for winkler scores in the DAM, with QRA-CP coming first, QR third, and QRA-R fourth. The last 3 spots are occupied by all three CP approaches with SPCI changing from just 0.07 off the top approach in the DAM, to the worst performer in the BM. QR achieves the lowest score of any model with LGBM, while Q-Ens has the second lowest with RF. These results are likely due to the closer coverage scores and narrower interval widths for the QR and QRA approaches.

5.3. Financial performance analysis

For our financial evaluation, we utilise a Battery Energy Storage System (BESS). This section outlines our BESS trading strategy, crucial for ensuring grid stability and seamlessly integrating renewable energy into the dynamic energy landscape. This consists of our Single trade strategy (TS1) and our high-frequency trading strategy (TS2):

- TS1 is a rule-based heuristic trading strategy adopted from [38, 43]. This strategy utilises quantile-based forecasts to optimise

Table 5
Winkler scores for the BM. Best approach for each regressor marked in bold, worst in red.

Model	QR	SCP	EnbPI	SPCI	QRA-R	QRA-CP	Q-Ens
KNN	124.18	134.77	131.35	132.18	**119.21**	120.89	126.35
LEAR	127.79	**115.02**	122.63	126.16	129.03	127.75	124.09
LGBM	**107.34**	115.77	111.18	122.41	109.47	108.34	111.18
RF	114.37	113.88	111.84	114.32	116.15	115.07	**111.39**
Avg.	118.42	119.86	119.25	123.77	118.47	**118.01**	118.25

Table 6
Financial performance comparison of RF & LGBM models. Best approach marked in bold, worst in red.

Model	TS1 DAM	TS1 BM	TS2 DAM	TS2 BM	Avg. DAM	Avg. BM
QR	€16,376	€3634	€21,887	€16,112	€19,131	€9873
SCP	€16,032	−€1215	€17,610	€11,119	€16,821	€4952
EnbPI	€16,806	€3214	€20,932	€16,237	€18,869	€9725
SPCI	€16,486	€3280	€21,632	€16,807	€19,059	€10,044
QRA-R	€13,868	€2768	€20,938	€14,705	€17,403	€8736
QRA-CP	€13,880	**€3923**	€20,805	€15,061	€17,342	€9492
Q-Ens	**€16,935**	€3748	**€21,926**	**€17,400**	**€19,430**	**€10,574**

trading decisions involving a hypothetical 1 MWh battery with no discharge limit, 80% discharge efficiency, and 98% charge efficiency. Over the specified time horizon (1 day for DAM or 8 h for BM), a single buy-sell pair trade is permitted, with the requirement that the buy trade occurs before the sell trade.

- TS2 leverages optimisation techniques to maximise profits in electricity markets, adopted from [72]. It incorporates flexible buy/sell timestamps within bottleneck-constrained volumes, considering realistic constraints such as battery capacity, charging speed, discharging rate, minimum charge level, and energy efficiency. The dynamic adjustment of charging and discharging operations, based on forecasted minimum and maximum price periods, is a fundamental aspect of this adaptive approach. The overarching goal of TS2 is to enhance profitability within bottleneck limitations, significantly increasing trading frequency and contributing to overall market efficiency.

For more detailed information on the implementation of both TS1 and TS2 see [72]. Table 6 summarises the financial performance of various forecasting approaches in the DAM and BM for trading strategies TS1 and TS2, excluding the KNN and LEAR models to focus on top performers. SCP consistently produces the worst results, with the lowest averages and the poorest performance in TS2 for both markets and TS1 for the BM. In contrast, EnbPI and SPCI demonstrate strong performance, with SPCI leading among the CP approaches, highlighting the benefits of time series-adapted methods. QRA does not surpass QR, with QR achieving better average scores in both markets and QRA-R ranking last for TS1 in the DAM. However, QRA-CP secures the top result for TS1 in the BM and scores closer to QR in the BM. The most notable outcome is the strong performance of our ensemble approach, which achieves the highest average result in both the DAM and BM, only missing the top position for TS1 in the BM.

5.4. Key findings and limitations

The application of CP techniques, particularly EnbPI and SPCI, improves the accuracy of PEPF in both DAM and BM, outperforming traditional QR models by providing more reliable PIs with consistent coverage. The ensemble approach, combining QR with EnbPI and SPCI, improves the balance between narrow intervals and coverage, resulting in more consistent and accurate forecasts. Financial analysis of these forecasting methods indicates that CP techniques, particularly our ensemble, Q-Ens, can increase profitability in trading algorithms; however, their dependence on large amounts of historical data and higher computational requirements may limit their practicality in real-time scenarios. These models can also face challenges when encountering extreme price events or sudden market shifts.

Table 7
Ablation analysis of Q-Ens for LGBM on the DAM and BM, evaluated using APS, Interval Width (IW), coverage (80%), and Winkler score. The full Q-Ens ensemble achieves the best balance across all metrics.

Market	Method	APS	IW	Coverage	Winkler score
DAM	**Q-Ens**	**3.48**	16.13	0.78	29.79
	EnbPI & SPCI	3.62	18.57	**0.82**	31.35
	QR & SPCI	3.48	**14.31**	0.73	**29.23**
	QR & EnbPI	3.49	15.51	0.75	29.68
BM	**Q-Ens**	**12.70**	73.24	0.78	111.18
	EnbPI & SPCI	13.13	79.83	**0.83**	114.88
	QR & SPCI	12.74	76.73	0.76	113.35
	QR & EnbPI	12.73	**73.16**	0.75	**109.47**

5.4.1. Ablation analysis of Q-Ens

Table 7 presents an ablation study of Q-Ens for LGBM on both the DAM and BM. The full Q-Ens model consistently delivers the best trade-off between sharpness and validity across all metrics.

Removing the QR module leads to the largest increase in interval width and Winkler score, indicating reduced sharpness and efficiency. Excluding EnbPI slightly narrows the PIs but decreases coverage, undermining reliability. Dropping SPCI yields marginal changes but results in lower coverage in both markets. Overall, each component contributes to performance, with QR and EnbPI being particularly important for maintaining calibration and PI quality.

5.4.2. Computational cost

To support reproducibility and assess operational feasibility, all experiments were conducted on a local machine with an Intel Core i7-6800K CPU @ 3.40 GHz and 64 GB RAM, running Ubuntu 22.04, without GPU acceleration or cloud computing services. In practice, as updated market data becomes available, the models are retrained on this new data and a new forecast is generated. Training times ranged from under two minutes for QR and SCP to 10–25 min for EnbPI and SPCI, with the Q-Ens ensemble requiring approximately 25–45 min in total. Inference times for all models were under 30 s. These computational requirements are well within the time-frame needed for operational deployment on standard desktop machines.

6. Conclusion

This study evaluated several probabilistic forecasting methods for electricity prices in the Irish DAM and BM, focusing on QR, QRA-R, QRA-CP, SCP, EnbPI, SPCI, and the proposed ensemble method Q-Ens. Our evaluation covered multiple dimensions—accuracy, reliability, interval width, coverage, and financial performance. We found that QR and QRA approaches performed well in efficiency metrics but fell short in coverage, where CP methods like SCP, EnbPI, and SPCI excelled. The ensemble approach (Q-Ens) demonstrated strong results in both efficiency and validity, and it delivered superior financial outcomes in a single trade and high-frequency trading strategy, confirming its practical value for market participants. The integration of CP techniques into EPF offers a promising way to improve forecast reliability and financial outcomes, especially as the energy market evolves and renewable energy integration increases. However, the financial evaluation, based on simulated trading, may not fully capture the complexities of real-world market conditions, including transaction costs, market liquidity, and sudden structural shifts. Therefore, further research should focus on testing these methods in diverse markets, including IDM, and exploring the practical limitations of CP models, particularly in environments with extreme price events or limited historical data. Future studies could also delve deeper into more granular 5 min BM datasets and incorporate probabilistic forecasting frameworks that address transaction costs and market conditions. Additionally, improving the robustness of CP methods, particularly with respect to outliers and non-exchangeable time series, will be fundamental. Exploring the application of CP models in renewable energy forecasting, smart grid management, and real-time data streams presents a rich avenue for future development, which will help to create more adaptable and reliable forecasting systems in increasingly complex energy markets.

CRediT authorship contribution statement

Ciaran O'Connor: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Mohamed Bahloul:** Writing – review & editing, Validation, Supervision. **Roberto Rossi:** Writing – review & editing, Validation, Supervision. **Steven Prestwich:** Supervision, Project administration, Funding acquisition. **Andrea Visentin:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT to improve the paper's readability. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Andrea Visentin reports was provided by Research Ireland. Ciaran O'Connor reports financial support was provided by Research Ireland. Mohamed Bahloul reports financial support was provided by Sustainable Energy Authority of Ireland. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was conducted with the financial support of Science Foundation Ireland, Ireland under Grant Nos. 18/CRT/6223 and 12/RC/2289-P2 which are co-funded under the European Regional Development Fund. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Data availability

The data and algorithms used are available on a GitHub repository provided.

References

[1] Ortner A, Totschnig G. The future relevance of electricity balancing markets in Europe. Energy Strat Rev 2019;24:111–20.
[2] Nowotarski J, Weron R. Recent advances in electricity price forecasting: A review of probabilistic forecasting. Renew Sustain Energy Rev 2018.
[3] Ziel F, Steinert R. Probabilistic mid-and long-term electricity price forecasting. Renew Sustain Energy Rev 2018;94:251–66.
[4] Khajeh H, Laaksonen H. Applications of probabilistic forecasting in smart grids: A review. Appl Sci 2022.
[5] Gammerman A, Vovk V, Vapnik V. Learning by transduction. Proc the 14th Conf Uncertain Artif Intell (UAI) 1998.
[6] Vovk V, Gammerman A, Shafer G. Algorithmic learning in a random world, vol. 29, Springer; 2005.
[7] Xu C, Xie Y. Conformal prediction interval for dynamic time-series. In: International conference on machine learning. PMLR; 2021, p. 11559–69.
[8] Xu C, Xie Y. Sequential predictive conformal inference for time series. In: International conference on machine learning. PMLR; 2023, p. 38707–27.
[9] Newbery D, Strbac G, Viehoff I. The benefits of integrating European electricity markets. Energy Policy 2016;94:253–63.
[10] Ilea V, Bovo C, et al. European day-ahead electricity market coupling: Discussion, modeling, and case study. Electr Power Syst Res 2018;155:80–92.
[11] Miraki A, Parviainen P, Arghandeh R. Probabilistic forecasting of renewable energy and electricity demand using graph-based denoising diffusion probabilistic model. Energy AI 2025;19:100459.
[12] Liu H, Zhang Z. Probabilistic forecasting of multiple plant day-ahead renewable power generation sequences with data privacy preserving. Energy AI 2025;19:100463.
[13] Ugurlu U, Oksuz I, Tas O. Electricity price forecasting using recurrent neural networks. Energ 2018;11(5):1255.
[14] Lago J, De Ridder F, De Schutter B. Forecasting spot electricity prices: Deep learning approaches and empirical comparison of traditional algorithms. Appl Energy 2018;221:386–405.
[15] Chen Y, Wang Y, Ma J, Jin Q. BRIM: An accurate electricity spot price prediction scheme-based bidirectional recurrent neural network and integrated market. Energ 2019;12(12):2241.
[16] Li W, Becker DM. Day-ahead electricity price prediction applying hybrid models of LSTM-based deep learning methods and feature selection algorithms under consideration of market coupling. Energy 2021;237:121543.
[17] 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 2021.
[18] ElRobrini F, Bukhari SMS, Zafar MH, Al-Tawalbeh N, Akhtar N, Sanfilippo F. Federated learning and non-federated learning based power forecasting of photovoltaic/wind power energy systems: A systematic review. Energy AI 2024;100438.
[19] Shinde P, Amelin M. A literature review of intraday electricity markets and prices. 2019 IEEE Milan PowerTech 2019;1–6.
[20] Piantadosi G, Dutto S, Galli A, De Vito S, Sansone C, Di Francia G. Photovoltaic power forecasting: A transformer based framework. Energy AI 2024;18:100444.
[21] Samarajeewa C, De Silva D, Manic M, Mills N, Moraliyage H, Alahakoon D, Jennings A. An artificial intelligence framework for explainable drift detection in energy forecasting. Energy AI 2024;17:100403.
[22] Uniejewski B, Marcjasz G, Weron R. Understanding intraday electricity markets: Variable selection and very short-term price forecasting using LASSO. Int J Forecast 2019;35(4):1533–47.
[23] Narajewski M, Ziel F. Econometric modelling and forecasting of intraday electricity prices. J Commod Mark 2020;19:100107.
[24] Koch C, Hirth L. Short-term electricity trading for system balancing: An empirical analysis of the role of intraday trading in balancing Germany's electricity system. Renew Sustain Energy Rev 2019;113:109275.
[25] Birkeland D, AlSkaif T. Research areas and methods of interest in European intraday electricity market research—A systematic literature review. Sustain Energy, Grids Netw 2024;101368.
[26] O'Connor C, Bahloul M, Prestwich S, Visentin A. A review of electricity price forecasting models in the day-ahead, intra-day, and balancing markets. Energ 2025;18(12):3097.
[27] Mazzi N, Pinson P. 10 - wind power in electricity markets and the value of forecasting. In: Kariniotakis G, editor. Renewable energy forecasting. Woodhead Publishing; 2017.
[28] Dumas J, Boukas I, de Villena MM, Mathieu S, Cornélusse B. Probabilistic forecasting of imbalance prices in the belgian context. In: 2019 16th international conference on the European energy market. EEM, IEEE; 2019, p. 1–7.

[29] Narajewski M. Probabilistic forecasting of German electricity imbalance prices. Energ 2022;15.
[30] O'Connor C, Collins J, Prestwich S, Visentin A. Electricity price forecasting in the irish balancing market. Energy Strat Rev 2024;54:101436.
[31] Zachmann G, Hirth L, Heussaff C, Schlecht I, Mühlenpfordt J, Eicke A. The design of the European electricity market–Current proposals and ways ahead. Luxembourg: European Parliament; 2023.
[32] Nowotarski J, Weron R. Computing electricity spot price prediction intervals using quantile regression and forecast averaging. Comput Statist 2015.
[33] Hagfors LI, Bunn D, Kristoffersen E, Staver TT, Westgaard S. Modeling the UK electricity price distributions using quantile regression. Energy 2016;102.
[34] Maciejowska K, Nowotarski J, Weron R. Probabilistic forecasting of electricity spot prices using factor quantile regression averaging. Int J Forecast 2016;32(3):957–65.
[35] Cheng H, Ding X, Zhou W, Ding R. A hybrid electricity price forecasting model with Bayesian optimization for German energy exchange. Int J Electr Power Energy Syst 2019;110:653–66.
[36] Lucas A, Pegios K, Kotsakis E, Clarke D. Price forecasting for the balancing energy market using machine-learning regression. Energ 2020;13(20):5420.
[37] Marcjasz G, Uniejewski B, Weron R. Probabilistic electricity price forecasting with NARX networks: Combine point or probabilistic forecasts? Int J Forecast 2020;36(2).
[38] Uniejewski B, Weron R. Regularized quantile regression averaging for probabilistic electricity price forecasting. Energy Econ 2021;95.
[39] Kath C, Ziel F. Conformal prediction interval estimation and applications to day-ahead and intraday power markets. Int J Forecast 2021.
[40] Jensen V, Bianchi FM, Anfinsen SN. Ensemble conformalized quantile regression for probabilistic time series forecasting. IEEE Trans Neural Netw Learn Syst 2022.
[41] Hu J, Luo Q, Tang J, Heng J, Deng Y. Conformalized temporal convolutional quantile regression networks for wind power interval forecasting. Energy 2022;248:123497.
[42] Zaffran M, Féron O, Goude Y, Josse J, Dieuleveut A. Adaptive conformal predictions for time series. In: International conference on machine learning. PMLR; 2022, p. 25834–66.
[43] Uniejewski B. Smoothing quantile regression averaging: A new approach to probabilistic forecasting of electricity prices. 2023, arXiv:2302.00411.
[44] Marcjasz G, Narajewski M, Weron R, Ziel F. Distributional neural networks for electricity price forecasting. Energy Econ 2023;125:106843.
[45] Ghosh S, Shi Y, Belkhouja T, Yan Y, Doppa J, Jones B. Probabilistically robust conformal prediction. In: Proceedings of the 39th conference on uncertainty in artificial intelligence. UAI, vol. 216, 2023, p. 681–90.
[46] Dewolf N, Baets BD, Waegeman W. Valid prediction intervals for regression problems. Artif Intell Rev 2023;56(1):577–613.
[47] Trebbien J, Pütz S, Schäfer B, Nygård HS, Gorjão LR, Witthaut D. Probabilistic forecasting of day-ahead electricity prices and their volatility with LSTMs. In: 2023 IEEE PES innovative smart grid technologies Europe. ISGT EUROPE, IEEE; 2023, p. 1–5.
[48] Cordier T, Blot V, Lacombe L, Morzadec T, Capitaine A, Brunel N. Flexible and systematic uncertainty estimation with conformal prediction via the MAPIE library. In: Conformal and probabilistic prediction with applications. PMLR; 2023.
[49] Monjazeb MR, Amiri H, Movahedi A. Wholesale electricity price forecasting by quantile regression and Kalman filter method. Energy 2024;290:129925.
[50] De la Torre J, Rodriguez LR, Monteagudo FE, Arredondo LR, Enriquez JB. Electricity price forecast in wholesale markets using conformal prediction: Case study in Mexico. Energy Sci Eng 2024.
[51] Lee J, Xu C, Xie Y. Transformer conformal prediction for time series. 2024, arXiv preprint arXiv:2406.05332.
[52] Xu C, Jiang H, Xie Y. Conformal prediction for multi-dimensional time series by ellipsoidal sets. 2024, arXiv preprint arXiv:2403.03850.
[53] Angelopoulos A, Candes E, Tibshirani RJ. Conformal pid control for time series prediction. Adv Neural Inf Process Syst 2024;36.
[54] Vahidinasab V, Jadid S. Bayesian neural network model to predict day-ahead electricity prices. Eur Trans Electr Power 2010;20(2):231–46.
[55] Kostrzewski M, Kostrzewska J. Probabilistic electricity price forecasting with Bayesian stochastic volatility models. Energy Econ 2019;80:610–20.
[56] Uniejewski B, Marcjasz G, Weron R. On the importance of the long-term seasonal component in day-ahead electricity price forecasting: Part II—Probabilistic forecasting. Energy Econ 2019;79:171–82.
[57] Khosravi A, Nahavandi S, Creighton D. Quantifying uncertainties of neural network-based electricity price forecasts. Appl Energy 2013;112:120–9.
[58] Klein N, Smith MS, Nott DJ. Deep distributional time series models and the probabilistic forecasting of intraday electricity prices. J Appl Econometrics 2023;38(4):493–511.
[59] Serinaldi F. Distributional modeling and short-term forecasting of electricity prices by generalized additive models for location, scale and shape. Energy Econ 2011;33(6):1216–26.
[60] Monteiro C, Ramirez-Rosado IJ, Fernandez-Jimenez LA, Ribeiro M. New probabilistic price forecasting models: Application to the Iberian electricity market. Int J Electr Power Energy Syst 2018;103:483–96.
[61] Yuanchen W, Yizhi H, Tong G, Lei D, Ying W, Wenbin X. Electricity price forecasting based on Bayesian network and Monte Carlo simulation. In: 2023 IEEE 6th international conference on automation, electronics and electrical engineering. AUTEEE, IEEE; 2023, p. 542–6.
[62] Maciejowska K, Nitka W. Multiple split approach–multidimensional probabilistic forecasting of electricity markets. 2024, arXiv preprint arXiv:2407.07795.
[63] Janczura J. Expectile regression averaging method for probabilistic forecasting of electricity prices. 2024, arXiv preprint arXiv:2402.07559.
[64] Bhatnagar M, Rozinaj G, Vargic R. Using crafted features and polar bear optimization algorithm for short-term electric load forecast system. Energy AI 2025;100470.
[65] Foygel Barber R, Candes EJ, Ramdas A, Tibshirani RJ. Conformal prediction beyond exchangeability. 2022, p. arXiv e–Prints. arXiv–2202.
[66] Liu H, Shen X, Tang X, Liu J. Day-ahead electricity price probabilistic forecasting based on SHAP feature selection and LSTNet quantile regression. Energ 2023;16(13):5152.
[67] Xu Y, Li J, Wang H, Du P. A novel probabilistic forecasting system based on quantile combination in electricity price. Comput Ind Eng 2024;187:109834.
[68] O'Connor C, Prestwich S, Visentin A. Conformal prediction techniques for electricity price forecasting. In: International workshop on advanced analytics and learning on temporal data. Springer; 2024, p. 1–17.
[69] Koenker R, Bassett Jr G. Regression quantiles. Econ: J Econ Soc 1978.
[70] Zhao C, Wan C, Song Y. An adaptive bilevel programming model for nonparametric prediction intervals of wind power generation. IEEE Trans Power Syst 2019;35(1):424–39.
[71] Yasuda Y, Bird L, Carlini EM, Eriksen PB, Estanqueiro A, Flynn D, Fraile D, Lázaro EG, Martín-Martínez S, Hayashi D, et al. CE (curtailment–Energy share) map: An objective and quantitative measure to evaluate wind and solar curtailment. Renew Sustain Energy Rev 2022.
[72] O'Connor C, Collins J, Prestwich S, Visentin A. Optimising quantile-based trading strategies in electricity arbitrage. Energy AI 2025;100476.