Conformal Prediction Techniques for Electricity Price Forecasting

Ciaran O'Connor[1][0000−0002−6364−7712], Steven Prestwich[3][0000−0002−6218−9158], and Andrea Visentin[2][0000−0003−3702−4826]

[1] SFI CRT in Artificial Intelligence, School of Computer Science & IT, University College Cork, Cork, Ireland 119226305@umail.ucc.ie
[2] Insight Centre for Data Analytics, School of Computer Science & IT, University College Cork, Cork, Ireland
{andrea.visentin,s.prestwich}@insight-centre.org

Abstract. Integrating the erratic production of renewable energy into the electricity grid poses numerous challenges. One approach to stabilising market prices and reducing energy losses due to curtailments is the deployment of batteries. Efficient electricity arbitrage is crucial to make investments in storage systems financially viable; trading solutions to achieve this rely on price forecasting techniques. This study delves into the application of Conformal Prediction (CP) techniques, including Ensemble Batch Prediction Intervals (EnbPI) and Sequential Predictive Conformal Inference for Time Series (SPCI), for generating probabilistic forecasts in the Irish electricity market. Recent advancements in CP have addressed temporal considerations inherent in time series forecasting, eliminating the need for exchangeability assumptions. Our study demonstrates that despite potential efficiency trade-offs, CP methods consistently yield precise and reliable prediction intervals, ensuring comprehensive coverage. We assess the impact of CP on the financial results of a simulated trading algorithm. Monetary outcomes achieved with EnbPI and SPCI outperform those of both split CP and traditional quantile regression models, highlighting the practical superiority of CP in electricity price forecasting.

1 Introduction

Electricity price forecasting (EPF) is paramount for energy companies navigating volatile markets, sudden price shifts, and changing demand patterns. The widespread integration of renewables can introduce volatility in net power supply due to rapid and unforeseen changes in their output, potentially resulting in reliability concerns within the power system (Martinez-Anido et al. [2016]). The incorporation of energy storage technologies such as Battery Energy Storage Systems (BESS) can enhance the reliability and efficiency of the grid, improving market liquidity and reducing price volatility. With the integration of renewable energy sources and smart grids, forecasting accuracy becomes increasingly critical. Accurate predictions stabilize energy production planning and inform risk-aware strategies.

Ireland is particularly interesting in this perspective. The renewable component accounted for 39.5% of the total electricity production in 2020 (EirGrid [2022]). However, a consistent part of this power is wasted due to curtailments, e.g. when the production exceeds the demand. In Ireland, between 3% to 6% of electricity is lost every year due to this offer/demand mismatch. This barrier strongly limits the extension of the current renewable production and hinders the efforts to reach carbon neutrality. The introduction of BESS is an efficient way to tackle this issue. Precise forecasts are indispensable for market operations, notably in the Day-Ahead Market (DAM), Intra-Day Market (IDM), and Balancing Market (BM) within European electricity markets (Green and Vasilakos [2010], Martinez-Anido et al. [2016]). In the Irish single electricity market, the DAM represents a market with significant contributions to grid stability, with volume far exceeding both the IDM and BM. The DAM facilitates trading for electricity delivery the next day, with daily auctions at noon CET that establish initial market prices for electricity. The transition from deterministic to probabilistic forecasting signifies a significant shift in the need for nuanced predictions in the face of escalating uncertainties in future supply, demand, and prices. Probabilistic Electricity Price Forecasting (PEPF), in particular quantile forecasts, has emerged to address uncertainties and provide decision-makers with a comprehensive understanding of potential outcomes. While traditional point forecasting methods established the groundwork, the rise of quantile forecasting, spurred by initiatives like the Global Energy Forecasting Competition 2014 (GEFCom2014), has ushered in a new era. However, the challenge of robust uncertainty quantification persists, prompting exploration beyond conventional methodologies such as Quantile Regression (QR) and QR Averaging (QRA) (Maciejowska et al. [2016], Nowotarski and Weron [2018], Uniejewski and Weron [2021]).

Conformal Prediction (CP) emerges as a promising alternative to both QR and QRA for generating prediction intervals (PI), offering both validity and adaptability without relying on strict assumptions. Initially introduced in Gammerman et al. [1998] and subsequently extended to regression and classification domains by Vovk et al. [2005] and Shafer and Vovk [2008]. CP works by using past data to create a model that predicts future outcomes, providing a measure of confidence in the PI. A key feature of CP is its ability to deliver valid PI regardless of the underlying model, making it a flexible and reliable tool for uncertainty quantification. One of the foundational concepts in CP is data exchangeability, which assumes that the order of data points does not affect the statistical properties of the data set. While this assumption simplifies the development of CP methods, it presents significant challenges in time series applications, where the order and dependencies of data points are crucial. Traditional CP methods, which depend on exchangeability, often struggle to handle these dependencies effectively. To address these limitations, recent advancements in CP, such as Ensemble Batch Prediction Intervals (EnbPI) (Xu and Xie [2021]) and Sequential Predictive Conformal Inference for Time Series (SPCI) (Xu and Xie [2023]), have been developed. EnbPI enhances the flexibility and accuracy of PI by using ensemble methods, which combine multiple models to improve predictive

performance. SPCI, on the other hand, modifies the CP framework to better handle the sequential nature of time series data, ensuring that the PI remain valid even when data points are dependent on previous values. Focusing on these novel methods is important because they significantly improve the applicability of CP in dynamic fields like electricity markets, where accurate and reliable predictions are crucial for decision-making. By overcoming CP's traditional limitations, EnbPI and SPCI provide more reliable PI, crucial for renewable energy integration, where high uncertainties make reliable forecasts essential for market stability and efficient trading. This paper undertakes a thorough investigation into PEPF by leveraging recent advancements in CP methodologies, particularly adaptations tailored for time series data. We examine these adapted CP techniques alongside traditional QR approaches. Our study focuses on recent contributions to probabilistic forecasting, emphasizing how uncertainty can be transformed into an opportunity rather than a source of risk. In an extensive numerical analysis, the significance of coverage guarantees in enhancing trading strategies and fostering market resilience is assessed with an economic simulation.

The structure of this paper is as follows: Section 2 provides an overview of recent advancements in PEPF within the context of the DAM. Section 3 presents the dataset used in our empirical analysis. In Section 4, we detail our methodological framework, including our approach, models, and trading strategies. Section 5 presents the empirical findings, comparing the efficacy of CP with traditional methods through quantitative metrics and financial evaluations. Finally, Section 6 summarizes our findings.

2 Related Work

In this section, we present an overview of recent advancements in PEPF methodologies, focusing on modern CP techniques within the context of their application to the DAM.

2.1 Probabilistic Forecasting

Recent reviews by Khosravi and Nahavandi [2014] and Khajeh and Laaksonen [2022] have underscored the growing prominence of probabilistic forecasting in addressing uncertainties inherent in smart grids, supply-demand dynamics, and price variations. Notably, Nowotarski and Weron [2018] and Tzallas et al. [2022] have provided extensive insights into various PEPF methodologies, ranging from autoregressive models to neural networks and ensemble techniques. These reviews have critically evaluated methods like QRA, highlighting its dominance despite inherent limitations. Recent advancements in prediction interval generation, such as the methodology proposed in S. Salem et al. [2020], which leverages ensemble neural networks to generate prediction intervals alongside point estimates, contribute to the evolving landscape of regression analysis. Leverger et al. [2021] introduces a method that uses clustering to identify seasonal patterns and classification to enhance forecast accuracy. This hybrid approach improves the reliability of

probabilistic forecasts for seasonal data. Furthermore, Oesterheld et al. [2023] delves into the realm of performative predictions, where the act of making predictions can influence outcomes, shedding light on a crucial aspect of predictive modelling. In response to these limitations, recent studies by Uniejewski and Weron [2021], Uniejewski [2023] have introduced novel approaches such as Lasso QRA and smoothed QRA with kernel estimation, showcasing superior performance in trading strategies and financial outcomes. O'Connor et al. [2024] compares statistical, machine learning, and deep learning models for EPF in the Irish BM, finding that simpler statistical models like LEAR outperform more complex ones. The study provides a framework for model evaluation and offers an open-source dataset and models, which we utilize for our forecasting evaluation. Additionally, advancements in deep learning models, as explored in Lago et al. [2021, 2018], Marcjasz et al. [2020, 2022], have demonstrated notable improvements in both point and probabilistic forecasting.

2.2 Conformal Prediction

CP has emerged as a versatile framework for uncertainty quantification, as highlighted by Dewolf et al. [2023], emphasizing its significance in regression prediction intervals. Notably, recent work in Foygel Barber et al. [2022] has addressed CP's traditional challenge in handling time series data by introducing weighted residual distributions, enhancing robustness and reliability in prediction intervals, while Ghosh et al. [2023] presents an approach to improve CP's robustness to outliers. In the domain of time series forecasting, Jensen et al. [2022] introduced Ensemble Conformalized Quantile Regression (EnCQR), showcasing superior performance in handling heteroscedastic data and ensuring reliable prediction intervals. Similarly, Hu et al. [2022] employed Conformal Quantile Regression (CQR) with neural networks, surpassing traditional methods in wind power forecasting accuracy. In a PEPF context, the only paper of its kind, Kath and Ziel [2021] looks at CP as a robust framework to enhance time series forecasting and manage uncertainty. CP provides dynamic and symmetric prediction intervals, emphasizing balanced construction, effective sampling, and error-based normalization. Comparative analysis with QRA underscores CP's market sensitivity and model selection importance. Of particular relevance to our study, Kath and Ziel [2021] investigated CP as a framework to enhance time series forecasting and manage uncertainty. Recent advancements like EnbPI proposed by Xu and Xie [2021] and SPCI introduced by Xu and Xie [2023] have addressed exchangeability issues in time series data, offering promising avenues for uncertainty estimation in PEPF applications.

2.3 Trading

In the realm of energy trading, Krishnamurthy et al. [2017], l Narajewski and Ziel [2021], Uniejewski and Weron [2021], Uniejewski [2023] have made significant contributions to the development of effective trading strategies for energy storage systems. Noteworthy findings include the supremacy of stochastic models and

optimal bidding strategies. Additionally, studies by Staffell and Rustomji [2016], Tohidi and Gibescu [2019], and Abramova and Bunn [2021] have explored the economic viability of energy storage systems, addressing various aspects such as profitability, revenue evaluation, and battery pack degradation. O'Connor et al. [2024b] integrates renewable energy into markets with battery storage. The study enhances DAM and BM trading using quantile-based forecasts and increased trading frequency. It highlights the economic viability of larger batteries, precise quantile pair selection, and high-frequency trading for maximizing profits. In summary, this review highlights advancements in PEPF methodologies, encompassing DAM forecasting, CP integration, insights into energy storage systems, and effective trading strategies. However, notable gaps remain, particularly regarding the applicability of recent improvements in probabilistic approaches to EPF. Our study aims to address these gaps by presenting methodologies to enhance energy trading strategies in the DAM, focusing on modern adaptations of CP for time series applications.

3 Datasets

Data for this study were sourced from the Single Electricity Market Operator for Ireland (SEMO). Information regarding prices, network parameters and forecasts are available from SEMO[3] & SEMOpx[4]. We collected and analyzed historical data and Transmission System Operator (TSO) predictions from 2019 to 2022, revealing that electricity prices exhibit significant volatility closely linked to demand and wind forecasts during this period, as illustrated in 1.

In the Irish DAM, prices are determined with hourly granularity, established at 11 pm on the preceding day. Bids for the DAM must be submitted before midday of the previous day to facilitate efficient market operations. Focusing on predicting DAM prices for the 24 settlement periods of the subsequent day. This temporal alignment ensures timely dissemination of forecasting insights, enabling market participants to strategise and make informed bidding decisions in advance. Our analysis incorporates a comprehensive set of regressors to predict DAM prices. These include historical DAM prices spanning the previous 168 hours, alongside corresponding forecasts of demand and wind speed. Additionally, we integrate past 168-hour DAM prices. This selection of attributes offers a robust foundation for price forecasting, capturing both historical trends and relevant external factors such as demand and weather conditions. For further details on our forecasting approach, market structure, datasets, and variables for both the DAM and BM, please refer to O'Connor et al. [2024a].

4 Methodology

This section delineates our study's methodology, focusing on key models and forecasting approaches across three primary stages: probabilistic forecasting models

[3] https://www.sem-o.com/
[4] https://www.semopx.com/market-data/

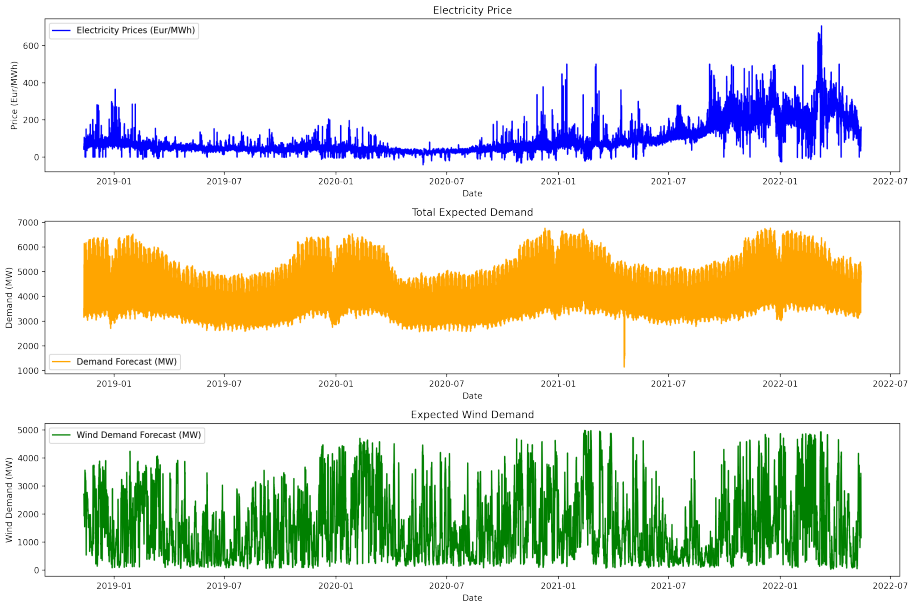

Fig. 1: Electricity Price and Demand Forecasts

(Section 4.1), corresponding approaches (Section 4.2), and the evaluation of our models (Section 5.1). Our approach, encompassing both forecasting and trading, comprises five key steps:

1. *Data Collection and Preparation:* Aggregating historical and forward-looking data from the ISEM for the DAM.
2. *Data Pre-processing and Model Optimization:* We pre-process the data and optimize each predictive model. Hyperparameter tuning is conducted using three-month data subsets to enhance model performance.
3. *Walk-Forward Model Validation:* Performing iterative walk-forward validation to ensure the reliability of our models, continuously updating the time horizon.
4. *Quantile Forecasting:* Generating quantile forecasts using optimized models based on unseen test data, spanning 24 hours.
5. *Financial Evaluation:* We compare all quantile pairs for each methodology in a single trade scenario to assess average forecasting model performance.

4.1 Quantile Regression Models

Our study employs quantile regression models, a statistical technique estimating conditional quantiles. These quantiles form a "quantile pair," defining a forecast range with lower and upper bounds, offering potential values within a specified confidence level. Specifically, we examine two quantile pairs: QP1, encompassing the 0.1-0.9 quantiles, denoted by $\alpha = 0.1$ and its complementary quantile 0.9,

and QP2, comprising the 0.3-0.7 quantiles. The probabilistic regressor models benchmarked are:

- *LASSO Estimated AR* (LEAR): A modified autoregressive time series approach incorporating LASSO regularization for improved performance and feature selection.
- *K-Nearest Neighbors* (KNN): A non-parametric instance-based learning approach that predicts an instance's output by comparing it to the "K" nearest neighbors in the training set.
- *Random Forest* (RF): An ensemble model that combines multiple regression trees.
- *Light Gradient Boosting Method* (LGBM): Similar to RF, LGBM uses multiple regression trees but follows the boosting principle.

4.2 Split Conformal Prediction

Split Conformal Prediction (SCP) (Shafer and Vovk [2008]) has limited application to time series data due to strict requirements for data exchangeability, it contributes valuable insights into constructing prediction intervals without relying on specific distribution assumptions.

Ensemble Batch Prediction Intervals (EnbPI) The EnbPI algorithm, as introduced in Xu and Xie [2021], emerges as a leading CP method tailored for dynamic time-series forecasting, effectively addressing the challenges posed by time series data without relying on data exchangeability assumptions. EnbPI's key features include its adaptability to dynamic time series, ensuring the importance of the sequence of data points is acknowledged, which is a crucial factor often overlooked by conventional CP methods. It provides PI with finite-sample, approximately valid marginal coverage, particularly for regression functions and time series with mildly mixing stochastic errors. EnbPI also demonstrates computational efficiency by avoiding overfitting without the need for data splitting or training multiple ensemble estimators.

Sequential Predictive Conformal Inference for Time Series (SPCI) The SPCI algorithm, introduced as an advancement on EnbPI in Xu and Xie [2023], presents a more versatile framework for time-series forecasting by directly leveraging the dependency of residuals when constructing PI, offering distinct advantages over previous methods. SPCI's key features include its utilization of residual dependencies, enhancing adaptability to dynamic time series. It employs a conditional quantile estimator rather than relying on empirical quantiles, resulting in more accurate PI estimates. SPCI also serves as a more general framework compared to both EnbPI and split conformal methods, capable of encompassing the functionalities of both through appropriate component selection. Furthermore, the computational efficiency of SPCI is maintained by fitting conditional quantile estimators using quantile models, ensuring effectiveness in sequential settings. For both EnbPI and SPCI, each model is run once, with bootstrap set to 15 for both the DAM and BM.

5 Experimental Results

This section analyzes probabilistic approaches and forecasting models in the DAM. Using diverse metrics, statistical tests, and financial indicators, we assess each model's efficacy in quantile regression and modern CP adaptations. Starting with Section 5.1, we analyze forecast accuracy through calibration, coverage, sharpness, and statistical testing metrics, aiming to evaluate the accuracy, reliability, and precision of probabilistic forecasts. Subsequently, in Section 5.2, we conduct a comparative analysis of outcomes across forecasting models, uncovering economic implications associated with each approach.

5.1 Evaluation

Our evaluation of probabilistic predictions focuses on two crucial dimensions: validity and efficiency. Efficiency, gauged through metrics like sharpness and interval width, enhances precision. Simultaneously, validity, including calibration and coverage, demands the generation of precise PI to affirm the reliability of our forecasts. Subsequent sections delve into each of these aspects with statistical testing. The evaluation culminates in the financial assessment of probabilistic approaches in Section 5.2. The Python code used for this section and the dataset are made available to ensure reproducibility GitHub[5].

Efficiency: Pinball Score & Interval Width In the DAM, sharpness plays an important role in accurate anticipation, hedging, and real-time adjustments. The Pinball Score, derived from the Pinball Loss function, reflects the sharpness of the forecast based on the quantile prediction $\hat{q}_{\alpha,P}$ and observed price $P_{d,h}$: $PS(\hat{q}_{\alpha,P}, P_{d,h}, \alpha) =$

$$\begin{cases} (1-\alpha)(\hat{q}_{\alpha,P} - P_{d,h}) \text{ for } P_{d,h} \leq \hat{q}_{\alpha,p} \\ (\alpha)(P_{d,h} - \hat{q}_{\alpha,P}) \quad \text{for } P_{d,h} \geq \hat{q}_{\alpha,p} \end{cases} \tag{1}$$

Analyzing the Aggregate Pinball Score (APS) for DAM models, as illustrated in Table 1, with the lowest APS for marked in bold, we observe that both SCP and QR under-perform time-series adapted EnbPI and SPCI. Despite this under-performance, the top two models in APS are QR versions of RF and LGBM. This can be attributed to the high baseline accuracy of these models, limiting the potential improvement introduced by CP. In contrast, models with lower baseline accuracy, such as KNN and LEAR, experience a substantial enhancement with CP. LEAR sees a notable reduction in split CP and further in EnbPI in SPCI. Similarly, KNN exhibits a reduction from 6.65 to 5.71 for EnbPI and 6.09 for SPCI but faces a sharp increase in CP. This indicates that CP methods effectively mitigate the limitations of less accurate models, leading to significant improvements in forecast performance.

[5] `https://github.com/ciaranoc123/PEPF_Conformal`

Model	QR	CP	EnbPI	SPCI
KNN	6.65	7.79	**5.71**	6.09
LEAR	8.35	4.81	4.18	**4.08**
LGBM	**3.62**	3.67	3.76	3.71
RF	**3.63**	3.85	3.81	3.84
Avg.	5.56	4.99	**4.37**	4.43

Table 1: Aggregrate Pinball Score Scores. in green, the best approach for each regressor.

Moving further into efficiency, the assessment extends beyond sharpness, focusing on width. While sharpness ensures reliability, efficiency, intricately linked to PI width, is pivotal in refining precision. Post-validity optimization enhances overall robustness and accuracy. An important highlight of the models utilizing CP is the interval width, showcased in Figure 2. CP for highly accurate models reduces the interval width, while for less accurate models, it widens the interval width. This trend is evident in the consistent decline for EnbPI and SPCI models, where greater accuracy corresponds to a smaller interval and vice versa. This occurs due to the coverage guarantee, where less accurate models widen their intervals to meet this guarantee, a behaviour not observed in QR models. RF, while showing a similar high accuracy compared to LGBM, has a considerably different interval width, but it does achieve better coverage compared to other QR models. This holds for SCP, which, despite the high accuracy, fails to produce a narrow interval width as EnbPI and SPCI succeed with. The variance between the average Interval Widths is minimal.

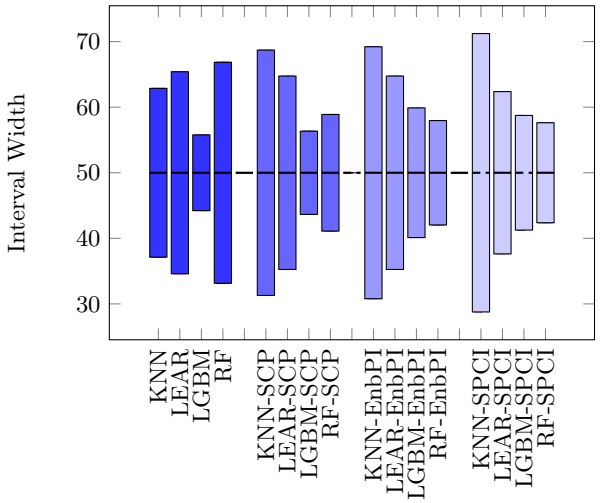

Fig. 2: Interval Width for each model in the DAM

Validity: Coverage and Kupiec Test Both reliability and accuracy of probabilistic forecasts are vital, with a particular focus on evaluating their validity. For this, we examine the model's coverage, which entails assessing precision in capturing the price $P_{d,h}$ within predefined probability levels or intervals $(\hat{L}d, h^{\alpha}, \hat{U}d, h^{\alpha})$, with close scrutiny of the nominal coverage level α. Empirical coverage, indicating the alignment of predictions with specified intervals, is expressed through the binary indicator $I^{\alpha}_{d,h}$ for a given day d and hour h:

$$I^{\alpha}_{d,h} = \begin{cases} 1 \text{ for } P_{d,h} \in [\hat{L}^{\alpha}_{d,h} \hat{U}^{\alpha}_{d,h}] \\ 0 \text{ for } P_{d,h} \notin [\hat{L}^{\alpha}_{d,h} \hat{U}^{\alpha}_{d,h}] \end{cases} \tag{2}$$

Coverage metrics highlight the models' ability to capture the true distribution. Figure 3 provides a comprehensive coverage analysis across the 0.1-0.9 quantile range in the DAM, revealing the efficacy of diverse forecasting methodologies. In the DAM context, targeting a coverage of 0.8 for the 0.1-0.9 quantile range, CP models outperform QR counterparts. SCP, EnbPI, and SPCI models achieve commendable average coverages of 0.82, 0.83, and 0.80, respectively. In contrast, QR models lag significantly with an average coverage of 0.62, highlighting the superiority of CP methodologies in attaining target coverage levels and affirming their advantage over traditional QR techniques in probabilistic forecasting.

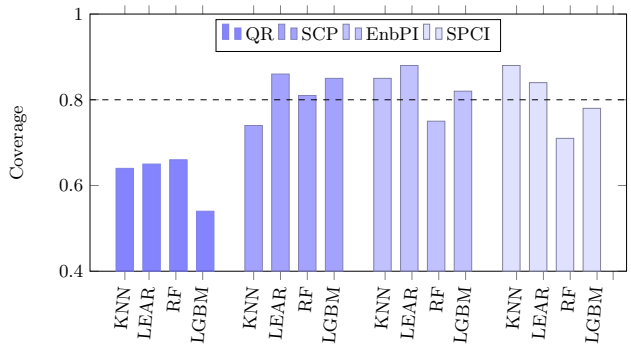

Fig. 3: Coverage for 0.1-0.9 quantile pair

In the 0.3-0.7 quantile pair analysis aiming for 0.4 coverage within the DAM framework (see Figure 4), CP methodologies demonstrate pronounced impact. All CP-based models achieve the desired coverage, contrasting sharply with only one QR model meeting the criterion (RF being the sole exception). This stark contrast underscores CP's pivotal role in forecasting, ensuring robust coverage guarantees. The coverage performance gap between QR and CP approaches is significant. The QR model's average coverage is only 0.35, notably lower than that of CP-based strategies. SCP, EnbPI, and SPCI models yield average coverages of 0.59, 0.59, and 0.56, respectively, highlighting CP methodologies'

superior performance in attaining target coverage levels and enhancing the reliability of probabilistic forecasts within the dynamic electricity market domain.

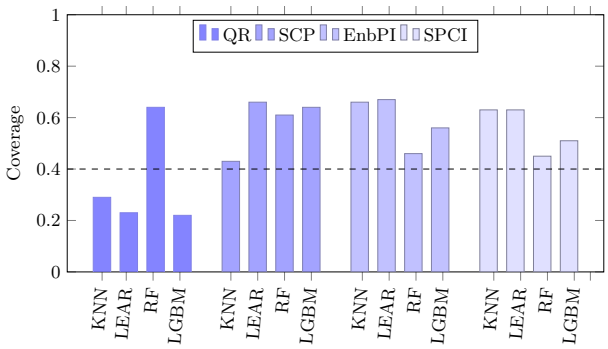

Fig. 4: Coverage for 0.3-0.7 quantile pair

The Kupiec test evaluates binary prediction model accuracy using labels (1 or 0) determined by predicted values falling within specified upper and lower bounds. For a model with T total observations, R events, predicted probability p, and significance level α, the test statistic is:

$$\chi^2 = -2\left[R\log(p) + (T-R)\log(1-p)\right]$$

The critical value χ^2_{crit} (one degree of freedom) is used to compare with X^2. If $\chi^2 > \chi^2_{crit}$, the null hypothesis of well-calibrated predictions is rejected.

To provide a comprehensive assessment, we perform the Kupiec test for unconditional coverage at 40% and 80% PI for each hour of the day, reporting the number of hours that pass the test.

	QR		SCP		EnbPI		SPCI	
Model	QP1	QP2	QP1	QP2	QP1	QP2	QP1	QP2
KNN	0	0	7	**20**	14	0	0	0
LEAR	0	0	3	0	0	0	9	0
RF	2	5	21	0	6	8	0	12
LGBM	0	0	7	0	22	0	**23**	0
Total	2	5	38	**20**	**42**	8	32	12

Table 2: Number of Hours that pass the Kupiec test. QP1 = quantile pair 0.1-0.9, QP2=quantile pair 0.3-0.7

Table 2 provides insights into model performance across quantile pairs in the DAM, focusing on the hours passing the Kupiec Test, with the best results

for each QP marked in bold. Only the RF model among QR models passes the test, totalling 7 hours out of 192. In contrast, CP approaches exhibit notable performance, exhibiting statistically significant performances 58 times. EnbPI models pass 50 hours, while SPCI models pass 44 hours. The significant difference in results between QR and CP approaches aligns with coverage outcomes in Figures 3 and 4.

Efficiency & Validity: Winkler Score The Winkler Score ($W_{t,h}$) amalgamates reliability and sharpness, providing a concise metric for the comprehensive assessment of probabilistic forecasts. It evaluates the width ($B_{t,h}$) of prediction intervals based on observed values ($y_{t,h}$), incorporating a penalty factor (α) for deviations from the interval bounds:

$$W_{t,h} = \begin{cases} B_{t,h} & \text{if } y_{t,h} \in [L_{t,h}, U_{t,h}] \\ B_{t,h} + \alpha(L_{t,h} - y_{t,h}) & \text{if } y_{t,h} < L_{t,h} \\ B_{t,h} + \alpha^2(y_{t,h} - U_{t,h}) & \text{if } y_{t,h} > U_{t,h} \end{cases}$$

Ultimately, Winkler Scores shed light on how well models ability to strike a balance between accuracy and interval width. Table 3 provides a detailed overview of Winkler Scores, offering insights into the efficiency and validity of different forecasting models in the dynamic electricity market landscape. The analysis

Model	QR	CP	EnbPI	SPCI
KNN	57.02	64.16	**51.47**	56.16
LEAR	61.97	36.30	37.37	**35.06**
LGBM	**29.87**	30.39	32.76	31.31
RF	**22.89**	30.81	32.14	31.65
Avg.	42.94	39.51	**38.44**	38.55

Table 3: Winkler Scores

in Table 3 highlight RF's dominance across all methodologies, with a Winkler Score of 22.89 in the QR framework. However, QR approaches are hindered by both KNN and LEAR's subpar results. In contrast, CP methods demonstrate consistent performance, with SPCI notably benefiting LEAR. Despite QR models, RF and LGBM, low accuracy and narrow interval width, challenges persist with models like KNN and LEAR. This underscores the need to explore alternative strategies, particularly CP approaches. Overall, accurate QR models excel in accuracy and precision in interval estimation, emphasizing their significance in probabilistic forecasting, especially in scenarios prioritizing accuracy and interval width.

Statistical Testing: Giacomini and White (2006) CPA Test To draw statistically significant conclusions, we employ the Giacomini and White (Giacomini

and White [2006]) Conditional Predictive Ability (CPA) test, utilizing a generalized Diebold-Mariano approach with a 24-dimensional vector of Pinball Scores for each day. The test statistic $\Delta_{X,Y,d}$ measures the difference between the L1 norms of APS vectors for models X and Y:

$$\Delta_{X,Y,d} = \|APS_{X,d}\| - \|APS_{Y,d}\|$$

where:

$$\|APS_{X,d}\| = \sum_{h=1}^{24} \sum_{\alpha=0.1}^{0.9} PS(\hat{q}_{\alpha,P}, P_{d,h}, \alpha)$$

for model X. P-values for the CPA test are computed for each model pair and dataset under the null hypothesis $H_0 : \phi = 0$ in the regression: $\Delta_{X,Y,d} = \phi' X_{d-1} + \epsilon_d$, where X_{d-1} contains day $d-1$ information, including a constant and lags of $\Delta_{X,Y,d}$. This test evaluates the reliability and precision of probabilistic

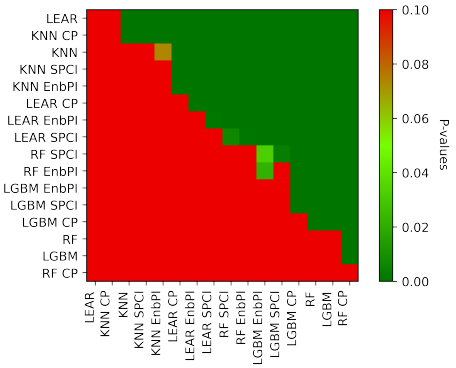

Fig. 5: Giacomini and White Test DAM

forecasts across a 24 hour horizon, offering insights into a model's capability to navigate evolving dynamics and uncertainties in electricity markets.

Figure 5 presents p-values using a chessboard representation. Dark green shades indicate the most significant differences between a model's forecast on the X-axis (better) and the forecast on the Y-axis (worse), with models arranged by APS. All CP methods, EnbPI, and SPCI, show statistically significant improvements for LEAR, suggesting enhanced forecasting accuracy. Conversely, KNN is significantly outperformed by all other models. RF, the top-performing model, demonstrates significant outperformance over all others, with SCP showing the best APS.

5.2 Financial Performance Analysis

This section outlines our BESS trading strategy, which is crucial for ensuring grid stability and seamlessly integrating renewable energy into the dynamic energy

landscape. This is evaluated with our Single trade strategy (TS1). TS1 is a rule-based heuristic trading strategy adapted from Uniejewski and Weron [2021], Uniejewski [2023]. This strategy utilizes quantile-based forecasts to optimize trading decisions involving a hypothetical 1 MWh battery with no discharge limit, 80% discharge efficiency, and 98% charge efficiency. Over the time horizon of 24 hours, a single buy-sell pair trade is permitted, with the requirement that the buy trade occurs before the sell trade.

Table 4: Financial Performance Comparison of DAM Models

Model	QR	CP	EnbPI	SPCI
KNN	€14,133	€14,342	**€14,488**	€13,374
LEAR	€2,889	€15,970	**€17,136**	€16,869
LGBM	€16,101	**€16,932**	€16,749	€16,676
RF	€16,652	**€16,883**	€16,862	€16,295
Avg.	€12,444	€16,032	**€16,309**	€15,804

In Table 4, CP adoption significantly improves LEAR and KNN models, with LEAR showing the most substantial enhancement. However, QR models' performance is notably impacted by LEAR and KNN, dragging down their average performance. Despite this, QR models perform comparably to CP approaches for LGBM and RF models, although CP models surpass QR for LGBM and two out of three for RF. Interestingly, although QR models like RF and LGBM exhibit superior APS and Winkler Scores, the inclusion of a coverage guarantee through CP appears to significantly influence financial outcomes. This highlights the nuanced interplay between model accuracy, interval width, and coverage assurance in financial performance evaluation. CP approaches demonstrate consistent performance across all models, highlighting the influence of CP's coverage guarantee in ensuring a robust forecasting framework compared to traditional quantile regression models.

Discussion

Model evaluation in the DAM reveals the intricate interplay between forecast accuracy, interval width, and financial outcomes. While QR models excel in accuracy and precision, challenges persist with models like KNN and LEAR, impacting financial metrics. In contrast, CP methodologies, including SCP, EnbPI, and SPCI, demonstrate reliability across all models, leveraging coverage guarantees for robust forecasting. Despite QR's accuracy, CP's coverage guarantee influences financial outcomes, highlighting trade-offs between accuracy, interval width, and coverage assurance. QR and CP approaches complement each other, enhancing model performance and decision-making in the DAM.

6 Conclusion

In this study, we conducted a comprehensive analysis of probabilistic forecasting models within the dynamic electricity market, focusing on the DAM. Our investigation encompassed a diverse array of metrics, statistical tests, and financial indicators to evaluate the efficacy of traditional QR techniques and modern CP adaptations.

CP emerges as a potent approach for bolstering the reliability and precision of probabilistic forecasts within the DAM. CP's adaptive methodology addresses the intrinsic uncertainties of the DAM, offering superior coverage guarantees and interval width optimization compared to traditional QR techniques. SCP, EnbPI, and SPCI CP methodologies consistently outperform QR across various quantile pairs, exhibiting commendable performance in coverage metrics. CP effectively mitigates the limitations of less accurate models, resulting in substantial forecast improvements, making it a powerful tool for decision-making in the DAM. However, traditional QR models demonstrated exceptional accuracy and precision in interval estimation, underscoring their significance in probabilistic forecasting, despite challenges with certain models such as KNN and LEAR. Regarding financial performance, CP methodologies displayed remarkable consistency and reliability, leveraging their coverage guarantees to ensure robust forecasting frameworks.

Our analysis underscores the complementary roles of QR and CP approaches, with integrating CP alongside QR promising to advance probabilistic forecasting in DAM, facilitating more informed decision-making in dynamic electricity markets.

References

Ekaterina Abramova and Derek Bunn. Optimal daily trading of battery operations using arbitrage spreads. *Energies*, 14(16):4931, 2021.

Nicolas Dewolf, Bernard De Baets, and Willem Waegeman. Valid prediction intervals for regression problems. *Artificial Intelligence Review*, 56(1):577–613, 2023.

EirGrid. Renewable energy, 2022. URL `https://www.eirgridgroup.com/how-the-grid-works/renewables/`.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. *arXiv e-prints*, pages arXiv–2202, 2022.

A Gammerman, V Vovk, and V Vapnik. Learning by transduction. *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence (UAI)*, 1998.

Subhankar Ghosh, Yuanjie Shi, Taha Belkhouja, Yan Yan, Jana Doppa, and Brian Jones. Probabilistically robust conformal prediction. *Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence (UAI)*, 216:681–690, 31 Jul–04 Aug 2023.

Raffaella Giacomini and Halbert White. Tests of conditional predictive ability. *Econometrica*, 74(6):1545–1578, 2006.

Richard Green and Nicholas Vasilakos. Market behaviour with large amounts of intermittent generation. *Energy Policy*, 38(7):3211–3220, 2010.

Jianming Hu, Qingxi Luo, Jingwei Tang, Jiani Heng, and Yuwen Deng. Conformalized temporal convolutional quantile regression networks for wind power interval forecasting. *Energy*, 248:123497, 2022.

Vilde Jensen, Filippo Maria Bianchi, and Stian Normann Anfinsen. Ensemble conformalized quantile regression for probabilistic time series forecasting. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

Christopher Kath and Florian Ziel. Conformal prediction interval estimation and applications to day-ahead and intraday power markets. *International Journal of Forecasting*, 37(2):777–799, 2021.

Hosna Khajeh and Hannu Laaksonen. Applications of probabilistic forecasting in smart grids: A review. *Applied Sciences*, 12(4):1823, 2022.

Abbas Khosravi and Saeid Nahavandi. Closure to the discussion of "prediction intervals for short-term wind farm generation forecasts" and "combined nonparametric prediction intervals for wind power generation" and the discussion of "combined nonparametric prediction intervals for wind power generation". *IEEE Transactions on Sustainable Energy*, 5(3):1022–1023, 2014.

Dheepak Krishnamurthy, Canan Uckun, Zhi Zhou, Prakash R Thimmapuram, and Audun Botterud. Energy storage arbitrage under day-ahead and real-time price uncertainty. *IEEE Transactions on Power Systems*, 33(1):84–93, 2017.

Micha l Narajewski and Florian Ziel. Optimal bidding on hourly and quarter-hourly day-ahead electricity price auctions: trading large volumes of power with market impact and transaction costs. *arXiv preprint arXiv:2104.14204*, 2021.

Jesus Lago, Fjo De Ridder, and Bart De Schutter. Forecasting spot electricity prices: Deep learning approaches and empirical comparison of traditional algorithms. *Applied Energy*, 221:386–405, 2018.

Jesus Lago, Grzegorz Marcjasz, Bart De Schutter, and Rafał Weron. Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark. *Applied Energy*, 293:116983, 2021.

Colin Leverger, Thomas Guyet, Simon Malinowski, Vincent Lemaire, Alexis Bondu, Laurence Rozé, Alexandre Termier, and Régis Marguerie. Probabilistic forecasting of seasonal time series: Combining clustering and classification for forecasting. In *International Conference on Time Series and Forecasting*, pages 47–63. Springer, 2021.

Katarzyna Maciejowska, Jakub Nowotarski, and Rafał Weron. Probabilistic forecasting of electricity spot prices using factor quantile regression averaging. *International Journal of Forecasting*, 32(3):957–965, 2016.

Grzegorz Marcjasz, Bartosz Uniejewski, and Rafał Weron. Probabilistic electricity price forecasting with narx networks: Combine point or probabilistic forecasts? *International Journal of Forecasting*, 36(2):466–479, 2020.

Grzegorz Marcjasz, Michał Narajewski, Rafał Weron, and Florian Ziel. Distributional neural networks for electricity price forecasting. *arXiv preprint arXiv:2207.02832*, 2022.

Carlo Brancucci Martinez-Anido, Greg Brinkman, and Bri-Mathias Hodge. The impact of wind power on electricity prices. *Renewable Energy*, 94:474–487, 2016.

Jakub Nowotarski and Rafał Weron. Recent advances in electricity price forecasting: A review of probabilistic forecasting. *Renewable and Sustainable Energy Reviews*, 81:1548–1568, 2018.

Ciaran O'Connor, Joseph Collins, Steven Prestwich, and Andrea Visentin. Electricity price forecasting in the irish balancing market. *arXiv preprint arXiv:2402.06714*, 2024a.

Ciaran O'Connor, Joseph Collins, Steven Prestwich, and Andrea Visentin. Optimizing quantile-based trading strategies in electricity arbitrage, 2024b.

Caspar Oesterheld, Johannes Treutlein, Emery Cooper, and Rubi Hudson. Incentivizing honest performative predictions with proper scoring rules. In *Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence*, pages 1564–1574, 2023.

Ciaran O'Connor, Joseph Collins, Steven Prestwich, and Andrea Visentin. Electricity price forecasting in the irish balancing market. *Energy Strategy Reviews*, 54:101436, 2024. ISSN 2211-467X. `https://doi.org/https://doi.org/10.1016/j.esr.2024.101436`. URL `https://www.sciencedirect.com/science/article/pii/S2211467X24001433`.

Tárik S. Salem, Helge Langseth, and Heri Ramampiaro. Prediction intervals: Split normal mixture from quality-driven deep ensembles. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 1179–1187, 2020.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

Iain Staffell and Mazda Rustomji. Maximising the value of electricity storage. *Journal of Energy Storage*, 8:212–225, 2016.

Yaser Tohidi and Madeleine Gibescu. Stochastic optimisation for investment analysis of flow battery storage systems. *IET Renewable Power Generation*, 13(4):555–562, 2019.

Petros Tzallas, Napoleon Bezas, Ioannis Moschos, Dimosthenis Ioannidis, and Dimitrios Tzovaras. Probabilistic quantile multi-step forecasting of energy market prices: A uk case study. In *Artificial Intelligence Applications and Innovations.*, pages 301–313. Springer, 2022.

Bartosz Uniejewski. Smoothing quantile regression averaging: A new approach to probabilistic forecasting of electricity prices. *arXiv preprint arXiv:2302.00411*, 2023.

Bartosz Uniejewski and Rafał Weron. Regularized quantile regression averaging for probabilistic electricity price forecasting. *Energy Economics*, 95:105121, 2021.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

Chen Xu and Yao Xie. Conformal prediction interval for dynamic time-series. In *International Conference on Machine Learning*, pages 11559–11569. PMLR, 2021.

Chen Xu and Yao Xie. Sequential predictive conformal inference for time series. In *International Conference on Machine Learning*, pages 38707–38727. PMLR, 2023.