

Investigating Calibration Challenges in Probabilistic Electricity Price Forecasting

Jan Niklas Lettner*

Institute for Automation and Applied Informatics

Karlsruhe Institute of Technology
Karlsruhe, Baden-Württemberg
Germany

jan.lettner@kit.edu

Hadeer El Ashhab*

Institute for Automation and Applied Informatics

Karlsruhe Institute of Technology
Karlsruhe, Baden-Württemberg
Germany

hadeer.elashhab@kit.edu

Benjamin Schäfer

Institute for Automation and Applied Informatics

Karlsruhe Institute of Technology
Karlsruhe, Baden-Württemberg
Germany

benjamin.schaefer@kit.edu

Abstract

As renewable energy integration increases market volatility, probabilistic electricity price forecasting has become essential for effective risk management. However, current proper-scoring rules often prioritize forecast sharpness at the expense of calibration, leading to overconfident and statistically unreliable uncertainty estimates. This work highlights the critical gap between theoretical scoring and practical calibration, demonstrating that models can become mere proxies for deterministic forecasts when reliability is neglected. We conclude that future research must shift toward calibration-aware objectives and architectures to ensure the distributional integrity of energy market forecasts.

CCS Concepts

• Applied computing → Physical sciences and engineering; • Computing methodologies → Machine learning.

Keywords

Probabilistic Forecasting, Calibration, Quantile Regression, Time Series, Time Series Forecasting, Energy, Electricity Prices, Electricity Price Forecasting, Cross-border

ACM Reference Format:

Jan Niklas Lettner, Hadeer El Ashhab, and Benjamin Schäfer. 2026. Investigating Calibration Challenges in Probabilistic Electricity Price Forecasting. In *ACM Sustainability Week 2026 (ACM Sustainability Week Companion '26)*, June 22–25, 2026, Banff, AB, Canada. ACM, New York, NY, USA, 2 pages. <https://doi.org/10.1145/3765611.3815392>

1 Introduction

In recent years, forecasting research has increasingly shifted from point forecasts toward probabilistic forecasts [4] as it outperforms point estimates by quantifying predictive uncertainty, providing better decision support under volatility. This shift necessitates the use of proper scoring rules that account for both sharpness and calibration—the fundamental paradigm being to maximize sharpness subject to calibration [2]. Calibration refers to the statistical

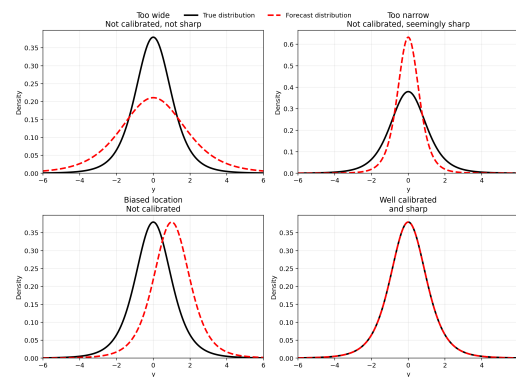


Figure 1: Illustrative forecast examples showing underdispersion, overdispersion, bias, and ideal calibration. Note that calibration is a collective property requiring statistical diagnostics (e.g., PIT histograms)

consistency between the predicted probabilities and the actual observed frequencies; essentially, if an event is predicted with a 90% probability, it should occur 90% of the time in the long run. Sharpness, on the other hand, refers to the concentration of the predictive distributions; a "sharper" forecast provides narrower prediction intervals, offering more precise information for the decision-maker. If a model focuses exclusively on sharpness while neglecting calibration, it risks becoming a mere proxy for deterministic forecasting—providing precise-looking estimates that lack statistical reliability and fail to represent the true underlying risks. Figure 1 provides a schematic illustration of these concepts by contrasting ideal, underdispersed, overdispersed, and biased forecast distributions.

Quantile Regression Averaging (QRA) [5] generates probabilistic forecasts by treating point forecasts and explanatory variables as inputs for separate quantile regressions. These models are optimized via pinball loss, a proper scoring rule minimized only when the forecast matches the true underlying distribution. Despite the theoretical appeal of proper scoring rules, empirical evidence shows they can still result in miscalibration. Notably, Si et al. observed this in normalizing flows trained with negative log-likelihood (NLL) [7], while Chung et al. found that pinball loss can prioritize sharpness over calibration [1]. This leads to overconfident forecasts, violating the principle of maximizing sharpness subject to calibration. This work aims to highlight the critical gap between existing proper scoring rules and the requirement for metrics that more effectively

*Jan Niklas Lettner and Hadeer El Ashhab contributed equally to this work.



penalize miscalibration. By addressing this gap, we hope that forecasts remain statistically sound and well-calibrated rather than merely superficially precise. Building on climate change mitigation initiatives that advocate for the integration of highly volatile renewable energy sources [6], probabilistic electricity price forecasting (PEPF) has become essential for both market participants and grid operators alike. To this end, the data selection, feature engineering, and model architectures employed in this work are all inspired by prior work in PEPF [3].

2 Methods

Our analysis uses the same hourly EU electricity prices as in earlier work [3]. To predict day-ahead prices, the model uses the previous 168 hours as input and outputs forecasts for the next 24 hours. Our pipeline uses a two-stage design: first, NHITS generates stochastic point forecasts via Monte Carlo dropout. These forecasts, along with calendar features, are then passed to a QRA head to produce final probabilistic predictions, following the architecture of Lettner et al. [3]. We compare two objectives for quantile estimation: the standard pinball loss and the calibration-oriented loss proposed by Chung et al. [1]. The latter includes a calibration component that checks whether the empirical coverage of the predicted quantiles matches their nominal coverage, by shifting quantile predictions upward or downward when needed, as well as a sharpness term to avoid excessively wide prediction intervals. Under the standard pinball loss, the QRA head consists of a separate linear regression model for each horizon and quantile. When applying the loss of Chung et al., we instead aim to more closely follow their original architecture by using a neural network with two hidden layers of 64 units each and a learning rate of 10^{-3} . In this setting, the model is trained separately for each horizon. The weighting parameter controlling the trade-off between sharpness and calibration is tuned in the main paper [1]. In our experiments, we set it to 0.5, thereby assigning equal importance to both objectives. We investigate group calibration to improve predictive reliability. While average calibration only requires quantile accuracy across the entire dataset, group calibration demands this accuracy within specific subgroups. Following Chung et al., we enforce this by alternating between standard and group-based training batches. We set the alternating frequency to 2, meaning the model switches batching schemes every other step. We also train a Normalizing Flow model using masked autoregressive layers, conditioned on latent representations from a Transformer encoder-decoder. While such models usually minimize NLL, Si et al. suggest that NLL alone may be insufficient and report better performance using alternative loss functions [7].

3 Results and Discussion

As shown in Table 1, the pinball loss of NHITS+QRA remains a strong baseline. The loss by Chung et al. performed substantially worse in continuous ranked probability score (CRPS) and expected calibration error (ECE), with group batching yielding no improvement. This discrepancy may stem from the non-i.i.d. nature of electricity prices; unlike standard regression datasets, temporal dependencies and seasonal shifts may undermine the effectiveness of Chung et al.'s calibration objective. Alternatively, our use of separate models for each horizon and quantile under the pinball setup

Used Loss	CRPS	ECE	Used Loss	CRPS	ECE
Pinball	16.87	0.107	NLL	31.51	0.219
Chung	21.74	0.273	CRPS	24.87	0.145
Chung + G.	21.74	0.274	ES	25.08	0.156

G. denotes group batching by weekday/hour.

ES denotes Energy Score.

(a) Results for NHITS+QRA

(b) Results for Normalizing flows

Table 1: Comparison of probabilistic forecasting models in terms of CRPS and ECE on the test dataset.

may offer greater flexibility than a single, more expressive model. For the normalizing flow model, training with CRPS yielded the best CRPS and ECE results. These findings demonstrate that, while various losses are theoretically motivated, the choice of scoring rule significantly impacts empirical calibration in practice. Probability integral transform (PIT) histograms further reveal the nature of these errors, with results for both models showing clear signs of bias or overconfidence.

4 Conclusion and Outlook

The integration of renewable energy sources necessitates PEPF to manage increasing market volatility. However, generating probabilistic forecasts that are both sharp and well-calibrated remains a challenge. While sharpness can be captured through standard deterministic metrics, the true value of a probabilistic model lies in its calibration—the extent to which predictive distributions reflect actual outcomes. We show that current "proper" scoring rules often fail to ensure this calibration, frequently favoring sharpness at the expense of reliability. Ultimately, because the primary goal of probabilistic forecasting is to provide a dependable measure of uncertainty, future research should shift focus from maximizing sharpness to developing calibration-aware objectives and architectures that prioritize the distributional integrity of the forecasts.

References

- [1] Youngseog Chung, Willie Neiswanger, Ian Char, and Jeff Schneider. 2021. Beyond Pinball Loss: Quantile Methods for Calibrated Uncertainty Quantification. In *Advances in Neural Information Processing Systems* (2021), Vol. 34. Curran Associates, Inc., 10971–10984. https://proceedings.neurips.cc/paper_files/paper/2021/hash/5b168fdb5ee5ea262cc2d4c0b457697-Abstract.html
- [2] Tilmann Gneiting and Matthias Katzfuss. 2014. Probabilistic Forecasting. 1, 1 (2014), 125–151. doi:10.1146/annurev-statistics-062713-085831
- [3] Jan Niklas Lettner, Hadeer El Ashhab, Veit Hagenmeyer, and Benjamin Schäfer. 2026. Assessing the Performance-Efficiency Trade-off of Foundation Models in Probabilistic Electricity Price Forecasting. arXiv:2604.14739 [cs.LG] <https://arxiv.org/abs/2604.14739>
- [4] Katarzyna Maciejowska, Bartosz Uniejewski, and Rafał Weron. 2022. *Forecasting Electricity Prices*. arXiv:2204.11735 [q-fin] doi:10.48550/arXiv.2204.11735
- [5] Jakub Nowotarski and Rafał Weron. 2015. Computing Electricity Spot Price Prediction Intervals Using Quantile Regression and Forecast Averaging. 30, 3 (2015), 791–803. doi:10.1007/s00180-014-0523-0
- [6] Sebastian Pütz, Hadeer El Ashhab, Matthias Hertel, Ralf Mikut, Markus Götz, Veit Hagenmeyer, and Benjamin Schäfer. 2024. Feasibility of Forecasting Highly Resolved Power Grid Frequency Utilizing Temporal Fusion Transformers. In *Proceedings of the 15th ACM International Conference on Future and Sustainable Energy Systems (E-Energy '24)*. Association for Computing Machinery, New York, NY, USA, 447–453. doi:10.1145/3632775.3661963
- [7] Phillip Si, Zeyi Chen, Subham Sekhar Sahoo, Yair Schiff, and Volodymyr Kuleshov. 2023. Semi-Autoregressive Energy Flows: Exploring Likelihood-Free Training of Normalizing Flows. In *Proceedings of the 40th International Conference on Machine Learning* (2023-07-03). PMLR, 31732–31753. <https://proceedings.mlr.press/v202/si23a.html>