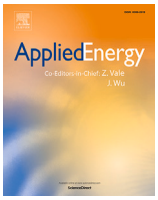
Applied Energy 398 (2025) 126412

Contents lists available at ScienceDirect

Applied Energy

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



On-line conformalized neural networks ensembles for probabilistic forecasting of day-ahead electricity prices

Alessandro Brusaferri [a,*], Andrea Ballarino [a], Luigi Grossi [b], Fabrizio Laurini [c]

[a] CNR, Institute of Intelligent Industrial Technologies and Systems for Advanced Manufacturing, via A. Corti 12, Milan, Italy
[b] University of Parma, Department of Engineering for Industrial Systems and Technologies, Viale delle Scienze, Parma, Italy
[c] University of Parma, Department of Economics and Management, Via J.F. Kennedy 6, Parma, Italy

HIGHLIGHTS

- Novel probabilistic electricity price forecasting approach.
- Extending neural network based methods through conformal inference.
- Experiments on German market and Italian regional bidding zones.
- Improved hourly reliability while providing stable probabilistic scores.
- Open GitHub repository provided to fully reproduce paper experiments.

ARTICLE INFO

Keywords:
Conformal prediction
Day-ahead market
Electricity price
Ensembles
Neural networks
Probabilistic forecasting
Time series

ABSTRACT

Probabilistic electricity price forecasting (PEPF) is subject of increasing interest, following the demand for proper quantification of prediction uncertainty, to support the operation in complex power markets with increasing share of renewable generation. Distributional neural network ensembles have been recently shown to outperform state of the art PEPF benchmarks. Still, they require critical reliability enhancements, as they fail to pass the coverage tests at various steps on the prediction horizon. In this work, we propose a novel approach to PEPF, extending the state-of-the-art methods based on neural network ensembles through conformal-inference techniques, deployed within an on-line recalibration procedure. Experiments have been conducted on multiple market regions, achieving day-ahead forecasts with improved hourly coverage and stable probabilistic scores.

1. Introduction

Trustworthy electricity price forecasting (EPF) systems represent fundamental strategic tools for utilities, retailers, aggregators, and large consumers to achieve effective participation in liberalized energy markets. Aside from being exploited to anticipate price movements and perform bidding strategies, EPF is a key enabler of further crucial decision-making stages [1], such as optimal generation and asset management [2] and energy-aware planning and scheduling [3]. In particular, hourly day-ahead EPF - i.e., the prediction of the hourly prices for the next day - is a challenging task and a subject of continuous interest in both academia and industry [4]. Compared to other traded commodities, electricity is still not economically storable on a large scale; thus, a constant balance between demand and supply is essential to achieve overall system stability. Hence, the price profiles in liberalized power markets often exhibit quite peculiar characteristics [5]. Complex relationships with conditioning variables (such as load demand, fuel costs, electricity production and weather conditions) are typically involved, encompassing meshed short- and long-term seasonalities, as well as considerable volatility, usually orders of magnitude larger than in other utility-trading contexts [6–9].

Critical deviations have been injected by the repercussions of the steep gas price variations on the power plants. The increasing penetration of renewable sources in the generation mix, fundamental to counter global warming, is introducing further short-term price fluctuations to be promptly tackled. Additional variables, such as the CO_2 certificate prices, interact with the power price features [10]. Additional complex market dynamics arise from the increasing competition among a

* Corresponding author.
Email address: alessandro.brusaferri@stiima.cnr.it (A. Brusaferri).

https://doi.org/10.1016/j.apenergy.2025.126412
Received 7 June 2024; Received in revised form 23 April 2025; Accepted 29 June 2025
Available online 5 July 2025
0306-2619/© 2025 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

growing number of highly active participants, each pursuing innovative business models [11,12].

Overall, forecasting is becoming harder to perform than ever before, and advances in this field are strongly requested to cope with such an evolving context [4,13].

Undoubtedly, the EPF research community is devoting considerable effort to addressing these key challenges, which is evident from the substantial number of scientific studies developed in the literature, as detailed in the following subsection.

1.1. Literature review

Over the years, a broad set of approaches have been investigated to perform EPF. A non-exhaustive list includes auto-regressive models and related extensions, exponential smoothing, generalized additive models, Gaussian processes, gradient boosting, neural networks, support vector machines, random forest, fuzzy logic, wavelet ensembles, and hybrid models. More detailed descriptions and comparisons of these techniques can be found in [1,14,15], and references therein. According to the aforementioned reviews, the EPF literature was primarily dedicated to simple statistical models and small neural network architectures until the early 2010s, mainly due to computational constraints of the technologies available. Besides, they proved effective in addressing the more stable price series occurring during that period [16].

Following the demand for more enhanced modeling capabilities to properly tackle modern energy market dynamics, the EPF community has been investigated more flexible but computationally intensive machine learning approaches [12,17–19].

Within this rapidly developing and fascinating field of research, much interest is being devoted to modern neural network architectures [20,21]. Such research momentum is mainly propelled by the significant results achieved across a broad set of computer science applications, including natural language processing [22] and computer vision [23]. Several research studies have contributed to the exploration and development of neural networks (NN) in the EPF field (see e.g., [4,24,25] for recent reviews). To date, the most comprehensive experimental investigation of NN based approaches to EPF is provided by Lago et al. [26], including feed-forward, recurrent and convolutional models, as well as several more traditional statistical techniques (such as ARX, ARIMA-GARCH, etc.), primarily adopted in the literature. When tested on the EPEX (Belgium day-ahead market benchmark), a deep feed-forward architecture (labeled DNN hereafter) achieved the highest predictive accuracy, outperforming both the traditional methods and the best recurrent neural networks with Long Short-Term Memory (LSTM) cells and Gated Recurrent Unit (GRU) cells.

Building upon the previous works, in [4] and [27] the authors have conducted a comparison between the DNN architecture and the recently proposed LASSO Estimated AutoRegressive model, which is arguably the most accurate approach within the broader statistical family. By deploying an extensive automatic hyperparameter optimization and evaluating a set of benchmark EPF tasks covering five different day-ahead markets, they confirmed the performance gains provided by the more flexible DNN. Indeed, the latter paper is titled “The dawn of machine learning” for EPF. Following these results, several NN-based techniques have been explored in recent years with the aim of reducing prediction error (see e.g., Olivares et al. [28] and references therein). In economic terms, even single percentage gains can lead to annual savings of up to millions of dollars for companies [13].

Although most of the available studies address point predictions (e.g., mean or median values) of day-ahead hourly prices, there is an increasing research momentum in the field of NN-based probabilistic EPF (PEPF). Specifically, PEPF is aimed at characterizing the underlying uncertainty of the model in the target space, e.g., via prediction intervals, quantiles or full predictive distributions. This is essential for companies operating within power markets showing increasing volatility, e.g., to support extremely relevant business tasks, such as effective risk optimization, stochastic optimization and what-if scenario analysis before trading. Several recent studies investigate this topic from different perspectives, reporting both additional problems and opportunities insisting on the EPF approach (see e.g., [29–32] and references therein for details). A thorough empirical comparison of deep-learning architectures for multivariate probabilistic energy forecasting is performed in [24], including DeepAR, DeepTCN, DSANet and LSTNet models.

According to several research works (see e.g., [1,33,34] and references therein), Quantile Regression Average (QRA) represents the reference benchmark to turn point NNs into a PEPF setup, following the ranking in the GEFCom competition.[1]

Distributional NNs [24] and Bayesian Deep Learning [25] techniques have been proposed to approximate the conditional distribution of the hourly prices given the input features, through parametric forms. The Gaussian distribution is commonly employed for this purpose. However, it has been observed that the prices tend to exhibit complex patterns, including heteroskedasticity, significant skewness and fat tails [33,36]. Deep quantile regression (DQR) based methods have been introduced to support non-parametric uncertainty characterization (see e.g., Zhou et al. [37] and references therein). Despite their asymptotic properties, they inherit potential quantile overfitting and overconfidence on the extreme quantiles in finite samples. Hence, ensembles of distributional NNs (i.e., Deep Ensembles) parameterizing flexible Johnson’s SU distributions have been recently proposed for this purpose, outperforming the state of the art PEPF benchmarks including both conventional Gaussian forms and the widely applied quantile regression on NN ensembles [33]. Despite the significant improvements, it has been observed that such models lack the required calibration capabilities—i.e., the statistical consistency between the probabilistic forecasts and the observations—to pass the coverage tests at various hours on the prediction horizon.

Conformal Prediction (CP) provides a principled framework to attain distribution-free finite sample marginal calibration guarantees. Pioneered in the early 2000 s, CP has recently become subject of a renewed research interest within the statistical and machine learning communities (see e.g., Balasubramanian et al. [38] and references therein for a detailed review). However, CP has so far attracted only minor attention in the EPF field. To date, only two papers have investigated CP-based approaches for this purpose. A first empirical study has been performed by Kath and Ziel [34] where the conventional split CP is compared to a normalized absolute deviation on LASSO, KNN, and SVM models, achieving promising results. Later, Zaffran et al. [39] implemented an adaptive conformal inference procedure on Random Forest based point predictors. As this presents a relatively novel and still understudied research direction, further extensions are expected from the community in the near future.

By looking beyond the EPF literature, the latest developments in the Conformal Prediction research domain provide a lot of compelling ideas to be further explored for achieving sharp and calibrated probabilistic forecasts (see e.g., Angelopoulos and Bates [40] and references therein). Enhancements that deserve particular attention include techniques to improve local adaptivity and to address the lack of exchangeability in practical time series applications [38]. To the best of our knowledge, such CP extensions have not yet been deployed on the state-of-the-art neural network-based approaches for PEPF to assess potential probabilistic performance gains.

[1] The Global Energy Forecasting Competition (GEFCom) began in August 2014, with a specific emphasis on probabilistic energy forecasting, including load, price, wind, and solar predictions. The price category drew interest from 287 participants globally, with the top-performing teams subsequently invited to contribute papers to a special 2016 edition of the International Journal of Forecasting [35].

1.2. Contributions and organization of the paper

Leveraging the available state of the art and moving from reported open issues, the major scope of the present work is to contribute to the further investigation of the CP framework within the PEPF field. In particular, we aim to explore whether it can provide a valuable support to address the limited hourly calibration of the state-of-the-art quantile regression-based and distributional neural networks (NN). Specifically, we target the approximation of the feature-conditioned day-ahead price distribution through a discrete set of quantiles (e.g., deciles), coupled in prediction intervals (PIs) of increasing coverage degree around the median.

Then, we propose a novel approach to PEPF, extending the Deep Ensembles based methods through conformal-inference-based techniques deployed within an on-line recalibration procedure. The developed method consists of the following major ingredients. First, we leverage the asymmetric Conformalized Quantile formulation introduced by Romano et al. [41] for regression tasks, here developed through a daily-recalibrated multi-horizon time series setup to support flexible step-wise compensations of each upper/lower band. Overall, this enhances local PIs' reliability (i.e., sample-wise efficiency) in conjunction with the conventional CP marginal coverage. Then, we deploy the adversarial conformal inference setup proposed in [42] for uncertainty quantification under non-exchangeable conditions, such as distribution shifts. Hence, the target prediction bands over the different coverage levels are dynamically adjusted by tracking the related quantiles of the conformity score sequences, incorporating the running sum of the miscoverage events to support a stable long-run calibration. To estimate the quantiles to be calibrated, we explore both a DQR setup and samples from different parameterized distributional NNs to assess their potential impact on both sample efficiency and reliability. A uniform vincentization technique is exploited for ensembles aggregation, leading to sharper bounds than the alternative probability aggregation, aside from marginalizing the different local minimizers reached by the training algorithm. Moreover, a post-hoc sorting operator is included before combination to achieve conditional quantile non-crossing.

We structured an open GitHub repository with the datasets and code to reproduce the experiments, as well as to support the integration and comparison of further datasets and PEPF techniques.[2]

The experiments are performed following the established best practice guidelines stated in [4] for EPF research. As case studies, we focused on the German market dataset (i.e., the benchmark application in the original Distributional NN paper [33]), as well as the different regional bidding zones constituting the Italian day-ahead markets made available by Golia et al. [43], providing a compelling setup for testing under heterogeneous conditions. A comparison against state of the art benchmarks is performed, including QRA, DQR, Distributional NNs, as well as conventional absolute-score and normalized-score-based CP settings.

The rest of the paper is structured as follows: Section II elaborates on the developed PEPF approach, while articulating the techniques in mathematical form; Section III provides a detailed description of the experimental setups and reports the results achieved; Section IV summarizes the conclusions and the envisioned future work.

2. Methods

In this section, we deepen the developed PEPF approach. A schematic representation of the main stages is shown in Fig. 1. We start by providing a brief introduction to the general conformal inference framework. Then, we address the local adaptivity issue by deepening the conformalized quantile regression technique. Subsequently, we describe the methods deployed to address the lack of robustness of CP under non-exchangeable conditions. Finally, we detail the prediction quantile

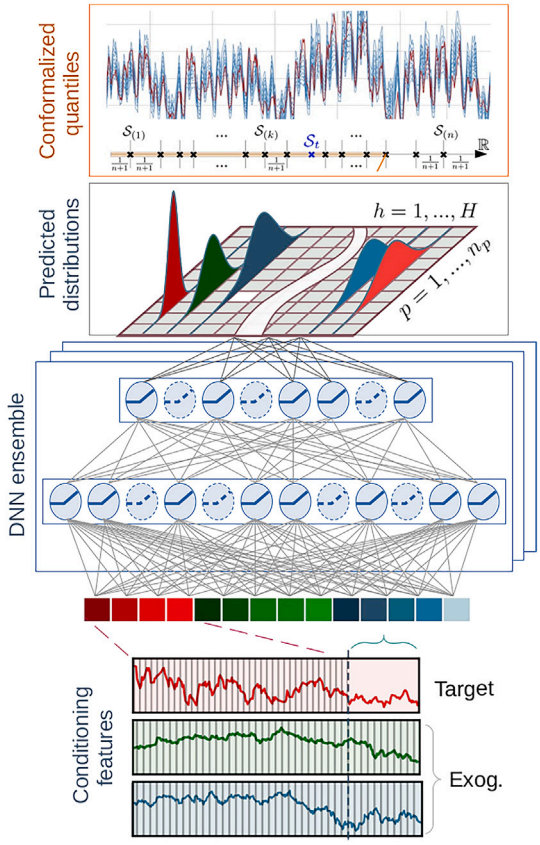


Fig. 1. Schematic representation of the main stages in the overall approach.

estimation procedures, the network architectures employed, as well as the deep ensemble combination techniques.

2.1. Conformal prediction

Conformal Prediction (CP) provides a general framework to obtain prediction intervals $C(D_n, \alpha, x_t) \equiv C(x_t)$, at any test point t, from black-box models with marginal coverage guarantees under finite sample settings. Formally, this is stated as: $\mathbb{P}(y_t \in C(x_t)) \geq 1 - \alpha$, with α-error probability, given a dataset $D_n \equiv \{(x_i, y_i)\}_{i=1}^{n}$, $y_i \in \mathbb{R}$ response variable and features vector $x_i = [x_i(1), \ldots, x_i(d)]$ involving continuous or discrete components. Moreover, CP is fully distribution-free, as opposed to distributional neural network techniques (parameterizing e.g. Gaussian, Student's t, mixtures, etc.), thus retaining validity for arbitrary latent distributions. Clearly, this is trivial without efficiency requirement (e.g., $C(x_t) = \mathbb{R} \Rightarrow \mathbb{P}(.) = 1$). Hence, the goal is to achieve a sharp interval close to the equality $\mathbb{P}(y_t \in C(x_t)) \approx 1 - \alpha$.

The core concepts behind CP are the conformity scores, which are exploited to assess the degree of "conformity" (thus the name of the approach) of the trained model's prediction with reference to a held-out calibration bag (i.e., split CP). For continuous target values, the absolute score $S(x_i, y_i) = |y_i - f(x_i)|$ is conventionally employed, where $f(x_i)$ represents the model's sample prediction.

Then, by computing the empirical quantiles of the order statistic obtained by ranking the conformity scores (depicted in Fig. 2), it is easy to show that for any t-th test sample:

$$\mathbb{P}(|y_t - f(x_t)| \leq S_{(\lceil (n+1)(1-\alpha) \rceil)}) =$$
$$\mathbb{P}(y_t \in \underbrace{f(x_t) \pm S_{(\lceil (n+1)(1-\alpha) \rceil)}}_{C_{1-\alpha}(x_t)}) = \frac{\lceil (n+1)(1-\alpha) \rceil}{n} \quad (1)$$

[2] https://github.com/bruale/PefCodeBench

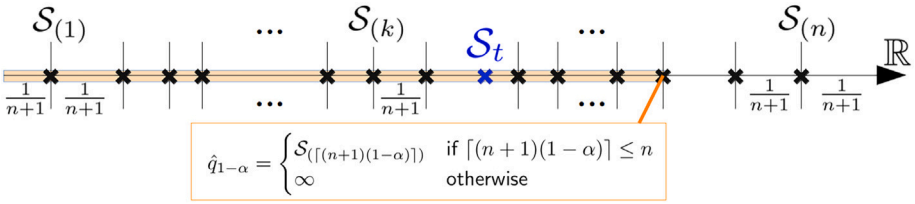


Fig. 2. Empirical quantiles of the conformity scores.

Besides, using a random tie-breaking rule (to avoid ties in the absolute residuals ranking), it has been shown that: $\mathbb{P}(y_t \in C_{1-\alpha}(x_t)) \leq 1-\alpha+1/(n+1)$. Further details on CP and related proofs are reported in [38] and references therein. As apparent from the Eq. (1), the conventional CP based on absolute scores leads to symmetric average bands of weakly varying extent across the input space, which has been shown to result in overly conservative predictions [44]. Since full conditional coverage (i.e., $\mathbb{P}(y_t \in C(x_t)|x_t = \chi) \geq 1-\alpha$, for almost all χ) is not achievable in finite samples under weak distribution-free assumptions, the goal is to obtain sample-wise efficiency (local PIs adaptivity) with proven marginal validity.

2.2. Improved local adaptivity by conformalized quantiles

Two major classes of methods have been recently proposed in the CP literature to enhance the capability to adapt the PI's width on simpler/harder test conditions (e.g., heteroscedasticity), namely normalized conformal inference-based, and conformalized quantile regression (CQR)-based approaches (see e.g., Balasubramanian et al. [38] and references therein). The former typically rescales the CP scores depending on the conditional spread of the observations, e.g., by parameterizing the target variance via dedicated network outputs as in distributional NNs. Still, this procedure leads to symmetric bounds around the point predictions despite the underlying distribution shape. Furthermore, it has been observed to suffer from critical PI inflation and systematic underestimation issues [40].

In consideration of these results in the literature, we deploy a CQR-based approach in the PEPF framework. Specifically, to compute the scores and the conformalized quantiles, we leverage the asymmetric form introduced by Romano et al. [41], which can be easily adapted to the multi-horizon forecasting setup proposed in [45] as follows:

$$C^h_{1-\alpha}(x_t) = [q^h_{\alpha/2}(x_t) - l^h_{1-\alpha/2}(\mathcal{I}_{c_t}), q^h_{1-\alpha/2}(x_t) + u^h_{1-\alpha/2}(\mathcal{I}_{c_t})] \tag{2}$$

where $C^h_{1-\alpha}(x_t)$ represents (1-α)-level PI at prediction step $h = [1, \dots, H]$ given the input features. $l^h_{1-\alpha/2}$ and $u^h_{1-\alpha/2}$ depict the (1-α/2)-th empirical quantiles of $q^h_{\alpha/2}(x_i) - y^h_i : i \in \mathcal{I}_{c_t}$ and $y^h_i - q^h_{1-\alpha/2}(x_i) : i \in \mathcal{I}_{c_t}$ respectively, computed on the calibration subset $\mathcal{I}_{c_t}$ for test time t. $\{q^h_\gamma(x_j)\}_{\gamma\in\Gamma}$, $q^h_\gamma(x_j) \leq q^h_{\gamma'}(x_j)\ \forall \gamma < \gamma'$ represent the discrete set Γ of the quantiles predicted by the NN ensemble (see Section 2.4), while y^h_t are the observed values. The PI bounds at specific coverage levels are straightforwardly derived as the predicted quantiles in the set Γ are defined by balanced pairs (e.g., distribution deciles, percentiles, etc).

Through such formulation, upper/lower quantiles can be specifically adjusted at each stage over the prediction horizon, considering the related coverage on the calibration sets. Overall, this yields stronger coverage capabilities by enabling more efficient input features conditioned approximate PI estimation beyond the CP marginal coverage. In fact, more flexibility in compensating for the predicted bands is deemed useful for addressing the complex shapes typically involved in PEPF applications, such as heteroscedasticity, skewness and fat tails [33]. This complements the asymptotic validity supported by distributional and quantile regression NNs (i.e., DQR) by introducing finite sample coverage guarantees, aside from compensating for potential quantile overfitting and overconfidence.

2.3. Addressing non-exchangeable conditions

Despite the aforementioned enhancements, the CQR-based formulation still inherits the lack of robustness of CP to practical settings where the backbone exchangeability assumption is not properly satisfied [44]. This simply states that the target (arbitrary) joint distribution is invariant to observation permutations, thus representing a weaker assumption than the i.i.d. commonly employed in machine learning applications (see e.g., Fontana et al. [46] for further details). In particular, covariates and concept shifts in the target series are recognized as critical issues impacting CP validity in practical time-series applications, as is the case for EPF.

Three main classes of approaches have been recently investigated in the CP literature devoted to non-exchangeable time series tasks, namely online sequential split CP (see e.g. Kath and Ziel [34]), bootstrapped estimators without underlying model refitting (as e.g., in [47,48]) and adversarial CP settings (see e.g., [39,42,49]).

In this work, we integrate the multi-step CQR-based compensations (reported in Section 2.2) within an on-line NN ensemble (i.e., Deep Ensemble) recalibration procedure, including the quantile tracking and coverage error integration stages proposed in [42] (i.e., Conformal PI control). Specifically, the quantiles predicted by the Deep Ensemble (DE) are incrementally conformalized over the test set following a daily retraining (i.e., recalibration). Details about how prediction model recalibration is commonly employed in the EPF context can be found in [4]. Formally, the predicted quantiles are corrected as follows at each stage t (i.e., day) in the test set:

$$\bar{q}^h_{t+1} = \bar{q}^h_t + \eta \nabla \rho_{1-\alpha/2}\left(s^h_t - \bar{q}^h_t\right) + r_t\left[\sum_{j=1}^{t}(\mathbb{1}\{\bar{q}^h_j \geq s^h_j\} - \alpha/2)\right] \tag{3}$$

$$r_t(x) = K_I \tan\left(\frac{x\log(t)}{tC_{sat}}\right) \tag{4}$$

where s^h_t summarizes the asymmetric CQR scores, $\mathbb{1}\{.\}$ is the indicator function, and $\nabla\rho_{1-\alpha/2}(.)$ denotes the subgradient of the quantile loss, with:

$$\nabla\rho_{1-\alpha/2}(s^h_t - \bar{q}^h_t) = \begin{cases} 1-\alpha/2 & \text{if: } s^h_t > \bar{q}^h_t \\ -\alpha/2 & \text{otherwise} \end{cases} \tag{5}$$

The operation is performed on both the (1-α/2)-th empirical quantiles of the asymmetric upper/lower scores $q^h_{\alpha/2}(x_t) - y^h_t$ and $y^h_t - q^h_{1-\alpha/2}(x_t)$. We keep $r_t(x)$ as the tangent integrator, leaving the investigation of alternative admissible saturation functions to future extensions. The constant C_{sat} bounds the asymptotic guarantee of the integral action at a level of at least $1-\alpha-\delta$:

$$C_{sat} = 2/\pi\left[\lceil\log(T)\delta\rceil - 1/\log(T)\right] \tag{6}$$

with $\delta > 0$ representing a small constant (e.g., 1e-2). η, $K_I > 0$ are further tunable hyperparameters controlling the proportional and integral actions respectively(see Section 3).

The whole procedure is first executed across a sub-sequence of samples close to the first test date to acquire the initial bag of calibration scores. It then progresses in a rolling window fashion.

The motivations behind such methodological choices are summarized hereafter. First of all, while in principle CP yields marginal coverage despite the accuracy of the underlying model, the latter impacts the local reliability of the prediction intervals [40]. Hence, updating the DE by exploiting the last observations can support probabilistic performance besides point accuracy. Besides, the ensemble combination (detailed in Section 2.4) provides a simple but efficient technique to get rid of poor local minimizers in the network parameter space (see section), often leading to improved accuracy beyond single networks (see e.g., Lago et al. [4]). The investigation of conformalized quantile regression methods without DE recalibration (as e.g. in [48]) and further on-line adaptive CP wrappers (as e.g., Zaffran et al. [39]) is left to future works.

We note that despite being more computationally demanding than the implementation of a single model training for the whole out-of-sample test set, online retraining is commonly employed in the EPF context, e.g., within the widely adopted Quantile Regression Averaging (QRA) technique as well as for the distributional neural networks (see e.g., Marcjasz et al. [33]). In fact, all the methods deployed in Section 3 follow the same online retraining procedure (i.e., daily recalibration) for the ensemble components, including the point prediction settings for CP/QRA, quantile regression NNs, and the distributional NNs with varied parameterizations. Hence, they exhibit comparable computational times, as discussed in Section 3. In addition, both CP and CQR computations are executed ex-post on coherent hold-out samples, with a negligible increase in computational time.

The quantiles tracking and integration provides an additional mechanism to compensate for coverage degradation and systematic errors occurring over the recalibration window, e.g., due to sensible short-term drifts. This supports a principled long-run reliability, i.e., $\lim_{T\to+\infty} 1/T \sum_{t=1}^{T} \mathbb{1}\{y_t^h \notin C_{1-\alpha}^h(x_t)\} = \alpha$, for each target coverage degree $1-\alpha$ (see [42] for further details on long-run coverage and related proofs). The exploration of the derivative action (i.e., the scorecasting component) is left to future works devoted to specific robust design for addressing the potential degradation in stability.

2.4. Prediction quantiles estimation and DE combination

To first estimate the conditional quantiles to be calibrated, we exploit both a quantile regression setup and samples from distributional NNs. In principle, the former enables flexible non-parametric approximations, but suffers from potential quantile overfitting and overconfidence in the extremes under finite data regimes [1]. The latter can mitigate such issues by leveraging the parametric form, but requires a careful selection of the density for each task at hand. Therefore, a dedicated experimental comparison is deemed useful to assess their capabilities in practical conditions.

The output layers of the NNs and the loss functions are specialized accordingly, i.e. providing the predicted quantiles averaged by the pinball score in the former, and parameterizing the target distribution being evaluated through a negative log-likelihood in the latter.

The next step in designing the PEPF framework regards the specification of the NN architectural form employed to compute the day-ahead predictions. Either single-step (i.e., using 24 hour specific models) and multi-step mapping forms can be considered for this purpose (see e.g., Lago et al. [26]). Besides, as introduced in Section 1 a broad range of architectures have been investigated in the EPF literature, including feed-forward, recurrent NNs and Transformers to cite a few (see e.g., Mashlakov et al. [24] and references therein for further details).

As the state-of-the-art Distributional NN for PEPF is based on multi-step feed-forward maps, we follow the same structural design in this work. This enables a proper comparison of the proposed PEPF technique under coherent settings (see Section 3). The investigation of further NN architectural forms is left to future studies. Such NN configuration is labeled DNN hereafter (i.e., Deep Neural Network), following the naming convention in [4,33].

Formally, each DE component is defined as a parameterized function providing day-ahead price predictions over the whole horizon in a unique pass, given the input conditioning set x_i at time i. Considering two hidden layers of $n_{u_1}, n_{u_2} \in \mathbb{Z}^+$ units to lighten the notation, the feed-forward map is mathematically expressed as:

$$\begin{aligned} \ell_1 &= \mathrm{g}(x_i W_1 + b_1) \\ \ell_2 &= \mathrm{g}(\ell_1 W_2 + b_2) W_3 + b_3 \end{aligned} \tag{7}$$

where $W_1 \in \mathbb{R}^{n_x \times n_{u_1}}$, $W_2 \in \mathbb{R}^{n_{u_1} \times n_{u_2}}$, $W_3 \in \mathbb{R}^{n_{u_2} \times H \cdot n_p}$, $b_1 \in \mathbb{R}^{n_{u_1}}$, $b_2 \in \mathbb{R}^{n_{u_2}}$, $b_3 \in \mathbb{R}^{H \cdot n_p}$ represent the weight and bias parameters in each hidden layer and $g(.)$ the nonlinear activation function (e.g., ReLU, LeakyReLU, GELU, etc.).

The input feature x_i can involve both past values of the target series and a set of exogenous variables (see Section 3 for the sets involved in each benchmark application). Hence, the learning pipeline is aimed at identifying the latent relationships and patterns within the information sources involved in the conditioning set [26].

The parameter $n_p \in \mathbb{Z}^+$ is defined depending on the subsequent network setup. Specifically, it is set as:

- $n_p = 1$: to achieve point predictions at each hour, as employed e.g., in the QRA-based PEPF setup (see Section 3);
- $n_p = \#\Gamma$: for quantile regression settings, with $\#\Gamma$ representing the number of estimated quantiles (e.g., $\#\Gamma = 10$ for decile approximation);
- $n_p = p$: for distributional NN settings, where p represents the number of parameters of the approximated conditional distribution, e.g., $p = 4$ for the Johnson's SU form (see Section 3).

For the Distributional NN configurations, the output of the last layer in Eq. 7 (i.e., ℓ_2) is further processed before being passed to the succeeding distributional layer in order to achieve valid parameterizations. For instance, in the case of the Johnson's SU form:

$$f^h(\chi) = \frac{\tau_i^h}{\sigma_i^h \sqrt{2\pi}} \frac{1}{\sqrt{1+\left(\frac{\chi-\lambda_i^h}{\sigma_i^h}\right)^2}} e^{-\frac{1}{2}\left[\zeta_i^h + \tau_i^h \sinh^{-1}\left(\frac{\chi-\lambda_i^h}{\sigma_i^h}\right)\right]^2} \tag{8}$$

the output of the last hidden layer is processed as:

$$\lambda_i^h = \ell_2^{[h]} \tag{9}$$

$$\sigma_i^h = \epsilon + \gamma \ \mathrm{Softplus}\left(\ell_2^{[H+h]}\right) \tag{10}$$

$$\tau_i^h = 1 + \gamma \ \mathrm{Softplus}\left(\ell_2^{[2\cdot H+h]}\right) \tag{11}$$

$$\zeta_i^h = \ell_2^{[3\cdot H+h]} \tag{12}$$

$$\mathrm{Softplus}(x) = \log(1+e^x) \tag{13}$$

with $\lambda_i^h, \sigma_i^h, \tau_i^h, \zeta_i^h$ defining the density location, scale, tailweight and skewness given by the feed-forward map, conditioned on the input feature x_i. $\ell_2^{[j]}$ defines the element in the output vector ℓ_2 at index j. $\epsilon = 1e^{-3}$ and $\gamma = 3$ are correction factors commonly introduced for computational purposes. As is apparent from the formulations above, a specific distribution parameterization is computed at each stage over the prediction horizon. Similar transformations are employed for the other density forms, e.g., to constrain positive values for standard deviation and degrees of freedom in Normal and Student-t (see Section 3).

For the quantile regression setup, the predicted conditional quantiles (e.g., deciles) at each stage $\hat{q}_\gamma^h(x_i)$ are simply extracted by indexing the last layer ℓ_2 in the feed-forward map by stage h and level γ. Then, the average Pinball loss across the discrete set $\gamma \in \Gamma$ of target quantiles is

computed as:

$$\sum_i \sum_h \sum_\gamma (y_i^h - \hat{q}_\gamma^h(x_i))\gamma \mathbb{1}\{y_i^h > \hat{q}_\gamma^h(x_i)\} + (\hat{q}_\gamma^h(x_i) - y_i^h)(1-\gamma)\mathbb{1}\{y_t^h \leq \hat{q}_\gamma^h(x_i)\} \tag{14}$$

Still, the approximated conditional quantiles can be subject of critical crossing issues. To address this problem, the introduction of non-crossing constraints within the quantile loss function has been proposed in the literature. Despite providing a valuable solution, it results in a sensible increase in the computational effort during learning. Besides, it does not provide non-crossing guarantees for the whole conditioning space. Therefore, in this work we exploit a post-hoc sorting operator, formally defined as:

$$\{\tilde{q}_\gamma^h(x_i)\}_{\gamma\in\Gamma} = \text{Sort}(\{\hat{q}_\gamma^h(x_i)\}_{\gamma\in\Gamma}),\ \tilde{q}_\gamma^h(x_i) \leq \tilde{q}_{\gamma'}^h(x_i)\ \forall \gamma < \gamma' \tag{15}$$

Aside from providing a computationally cheap solution to achieve non-crossing $\forall x$ during prediction by construction, it has been recently shown that such operation can only improve the pinball loss in the post-processed forecasting model [50].

The last design stage regards the definition of the approach to combine the neural network components within the DE. To this end, a wide set of methods have been studied in the time-series forecasting literature, from uniform marginal aggregations, to complex conditional weighting (see [51] for a recent detailed review). Still, the uniform aggregation (i.e., ensemble combination considering equal weights) has been shown to provide a robust alternative difficult to beat in practical settings. Similar observations have been reported under both point [4] and probabilistic [33] DNN-based settings in EPF studies. Moreover, it constitutes a widely adopted approach in the broader Deep Learning literature devoted to DE (see e.g., Lakshminarayanan et al. [52]). Therefore, we exploit a uniform ensemble combination for the present study, leaving the investigation of alternative techniques to future extensions. For point predictors (e.g., in the case of QRA and base CP in the next section), this results in a simple average between the DNN components. For Distributional NNs, both probability aggregation and vincentization techniques can be considered. We resort to the latter since it typically leads to sharper bounds, aside from marginalizing the different local minimizers reached by the training algorithm [51]. Formally, this is computed for each prediction stage and quantile as:

$$q_\gamma^h(x_i) = \sum_{j=1}^{n_e} \tilde{q}_\gamma^{h(\mathbf{j})}(x_i),\ \forall \gamma \in \Gamma,\ \forall h \in [1,\dots,H] \tag{16}$$

where n_e represents the number of neural networks constituting the ensemble, indexed by (j). The same approach is employed for forecasting combination in the quantile regression configuration.

The overall DE components are trained end-to-end in parallel, starting from different random initializations, by passing evenly spaced batches of data including the whole set of conditioning features and the target price values. To form the batches from the input time series during daily recalibration, we employ a common sliding window approach, as detailed in Section 3.

3. Applications and results

3.1. Case studies

As case studies, we focused on both the German market (GE) from Marcjasz et al. [33] and the different bidding zones constituting the Italian day-ahead markets (namely NOR, CNOR, CSOU, SOU, SARD, SICI) made available by Golia et al. [43], providing a compelling setup for comparative evaluations under heterogeneous conditions. The exogenous set for the GE market includes the day-ahead load and renewable generation forecasts as well as the most recent gas closing prices. The Italian datasets comprise the hourly load and wind generation predictions.

The Italian datasets cover the period from 10/1/2015 to 31/8/2019. For each region, the test set starts from 31/8/2018 onwards, while the validation subset covers the last year of observations before the first test recalibration.

Descriptive statistics of the Italian regions are reported in Table 1. Prices across the zones of continental Italy (NOR, CNOR, CSOU, SOU) exhibit a limited range (51–52), while in the islands, they show more variability, with the highest prices in SARD (59.63) and the lowest in SICI (49.96). Median values follow a similar trend and are closely aligned with the average prices, suggesting that differences are due to the fat-tailed distribution of prices. This is confirmed by kurtosis values, which generally exceed 5, reflecting occasional extreme price values. Price fluctuations are most pronounced in SARD, which has the highest standard deviation (23.02) and relative variability (SD/Mean*100 = 38.6 %). The other zones have standard deviations around 15–16, with relative variability close to 30 %, indicating more stability compared to SARD. Skewness values are close to 1 for most regions, indicating mildly right-skewed distributions. SICI has the lowest skewness (0.5), indicating a more symmetric distribution. SARD and SICI also show slightly lower kurtosis values (5.11 and 5.24), suggesting fewer extreme outliers compared to other regions.

Average electricity demand is highest in NOR and significantly lower in other regions, with CNOR and CSOU also showing relatively high demand levels. These figures reflect the distribution of electricity demand in Italy, with the North accounting for around 50 % of national demand. Median values closely align with means, indicating symmetrical distributions, as confirmed by skewness indices. NOR shows the highest standard deviation, reflecting greater variability in demand. However, when normalized by the mean (SD/Mean*100), NOR's relative variability is moderate at 25.94 %. SARD has the lowest relative variability at 15.99 %, indicating more stable demand patterns. Kurtosis values range from 1.92 to 2.77, suggesting relatively normal distributions without extreme outliers, consistent with stable and predictable demand patterns across all regions.

Wind generation varies significantly between regions, with SOU exhibiting the highest mean, followed by SICI and CSOU. NOR has the lowest average. Medians are considerably lower than means, particularly in regions with high wind generation (SOU, CSOU, SICI), indicating that most values are below the mean and a few high outliers drive the averages. Relative variability is exceptionally high across all regions (e.g., SD/Mean*100 ranging from 90 %–102 %), reflecting substantial fluctuations in wind generation. This variability is highest in SARD and lowest in SOU, though all regions show high levels of fluctuation, characteristic of renewable energy sources. Skewness is highest in NOR, indicating a strong right-skew with occasional high wind generation values. Other regions also exhibit positive skewness (around 1.15–1.35), suggesting that high wind generation events, though infrequent, are significant. Kurtosis values range from moderate to high (3.58–6.73), particularly in NOR, highlighting the presence of extreme outliers in wind generation data across all regions.

Considering the detailed analysis reported in [43], we have included in the input features the price settlements across the different hours of the last 7 days (hence 168 values), each of the 24 values for both day-ahead load and wind generation forecasts, as well as the weekday.

The German dataset spans observations from 1/1/2015 to 31/12/2020, with out-of-sample tests starting from 27/6/2019. The last 364 days before the beginning of the recalibration warm-up (which starts on 27/12/2018) are employed as validation data for hyperparameter tuning.

As input features, we included the subset selected with frequency 100 % during the experiments performed in [33]. We refer interested readers to this paper for a detailed analysis of the dataset. To summarize, we adopted: the price values over the previous 2 days t-1, t-2, i.e., 48 lags; the day-ahead load forecasts available at time t; the renewable energy sources predictions for the target date t as well as the previous day t-1; the most recent closing gas price available, i.e. at t-2; the weekday

Table 1
Descriptive statistics for prices, day-ahead demand, and day-ahead wind generation across zones of the Italian electricity market.

	NOR	CNOR	CSOU	SOU	SARD	SICI
Prices						
Mean	52.27	51.45	52.34	51.32	59.63	49.96
Median	50.04	49.92	50.16	49.88	57.01	49
1st Qu.	41.7	41.5	41.36	41.37	44.37	40.55
3rd Qu.	60.81	59.95	61	59.97	68.7	58.38
SD	15.99	15.2	16.36	15.49	23.02	14.62
SD/Mean*100	30.59	29.54	31.26	30.18	38.6	29.26
Skewness	1.02	0.88	1.11	0.79	1	0.5
Kurtosis	6.05	5.92	6.43	5.95	5.11	5.24
Day-ahead Demand						
Mean	18603.25	3584.38	5220.47	2906.16	981.21	2031.15
Median	17,948	3493	5287	2878	977	2024
1st Qu.	14575.75	2844	4206	2421	863	1718
3rd Qu.	22857.25	4300	6087	3323	1088	2309.25
SD	4825.83	866.59	1169.59	621.18	156.85	404.42
SD/Mean*100	25.94	24.18	22.4	21.37	15.99	19.91
Skewness	0.18	0.24	0.12	0.37	0.21	0.23
Kurtosis	1.92	2.06	2.19	2.73	2.77	2.46
Day-ahead wind Generation						
Mean	4.69	18.1	298.16	870.14	192.77	324.58
Median	3.38	12.63	199.45	664.99	117.93	220.58
1st Qu.	1.35	5.48	71.6	311.68	45.74	95.53
3rd Qu.	7.05	26.69	454.17	1243.15	278.13	469.38
SD	4.23	16.08	283.81	709	197.07	299.62
SD/Mean*100	90.2	88.88	95.19	81.48	102.23	92.31
Skewness	1.52	1.15	1.16	1.16	1.35	1.25
Kurtosis	6.73	3.58	3.65	3.92	4.04	3.93

encoding. The motivation behind such input feature selection is twofold. On the one hand, we aim to explore a group consistent with the one employed in the previous work implementing the baseline distributional neural network approach [33]. On the other hand, we target the assessment of the different PEPF models under consistent input variables. In [33], the authors reported sensible variations both in terms of the hyperparameters chosen (including NNs input feature selection via indicator variables) and the consequent test performances. Still, a deeper dataset analysis and feature selection beyond the one performed may lead to average performance improvements. We leave such investigation to future extensions.

Following [33], we included weekday dummies as input, which have been represented by means of the cyclical feature encoded in sine-cosine form, computing the components in the vector:

$$c_t = \left[\sin\left(2\pi d_t/7\right),\ \cos\left(2\pi d_t/7\right)\right] \tag{17}$$

where $d_t \in [0, \ldots, 6]$ indexes the day of the week for the predicted sample at time t. The investigation of alternative techniques (e.g., one-hot encoding) is left for future studies. The targets include the price values for each of the day-ahead 24 hours at stage t.

The samples are generated from the time series through a moving window. A Z-Score normalization is applied within each recalibration run, fitted by involving only the past values. Besides, a batch normalization is inserted after the DNN input layer.

3.2. Benchmarks

The proposed approach is compared to several state of the art DNN based PEPF methods, including QRA, Deep Quantile Regression (QR) and Distributional NNs. In addition to the Johnson's SU (Jsu) proposed in [33], we deploy a Student's t (Stu) to investigate the impact of the different distributional setups on the case studies.

Moreover, we implement CP techniques on both normal distributions and conventional absolute residuals (see [34]) to assess the benefits of the more flexible quantile-level corrections introduced. The former has been implemented following the same setup as the other distributional NN models to achieve a fair comparison. Hence, each NN in the ensemble parameterizes the mean and scale of the Gaussian distribution, followed by quantile vincentization. Although normalized conformity scores may alternatively be computed through the mean and variance of a mixture with uniform weights, probabilistic aggregation has been shown to perform worse than quantile ensembling (see [33,53]).[3] The CP on absolute residuals is obtained from the average predictions of NNs trained by minimizing the Mean Absolute Error.

To shorten notation, in the subsequent sections we employ the following labels for referring to the different PEPF methods: CQN for the parameterized Normal form (i.e.,Conformalized Quantile Normal); CQR for Conformalized Quantile Regression; CQJ for Conformalized Quantile Johnson's SU; CQS for Conformalized Quantile Student's t. The application of the quantiles tracking and coverage error integration stages is represented by stacking the label OCQ*, e.g., OCQJ for on-line conformal PI control on Jsu distributional NNs samples quantiles.

3.3. Experimental setup

The experiments have been performed on a cluster comprising 73 nodes with IBM POWER9 processors and NVIDIA Tesla V100 GPUs. The neural networks have been developed by leveraging `Tensorflow`, including the `Tensorflow Probability` library which provides several utilities to implement distributional NNs.[4]

In particular, we adopt a set of 4 DNNs with 2 hidden layers using softplus as their activation function. Training is performed by means of Adam [54] with early-stopping, starting from different random initializations. The number of units in each layer and the learning rate are

[3] The exploration of conformal inference on further forecasting combination techniques represents an interesting direction for future research.

[4] To setup the backbone DE framework, the description of the configurations follows the one used in the analysis reported in [33] for the baseline distributional and quantile regression based methods.

Table 2
Hyperparameters selected by the grid-search procedure for each zone.

GE	CP/QRA	Norm/CQN/OCQN	Jsu/CQJ/OCQJ	Stu/CQS/OCQS	QR/CQR/OCQR
n_h	768	896	896	640	640
l_r	1e-3	1e-3	1e-3	1e-3	1e-4
NOR	CP/QRA	Norm/CQN/OCQN	Jsu/CQJ/OCQJ	Stu/CQS/OCQS	QR/CQR/OCQR
n_h	512	768	640	640	896
l_r	1e-3	1e-3	1e-3	1e-3	1e-4
CNOR	CP/QRA	Norm/CQN/OCQN	Jsu/CQJ/OCQJ	Stu/CQS/OCQS	QR/CQR/OCQR
n_h	768	960	896	896	896
l_r	1e-3	1e-3	1e-4	1e-3	1e-4
CSOU	CP/QRA	Norm/CQN/OCQN	Jsu/CQJ/OCQJ	Stu/CQS/OCQS	QR/CQR/OCQR
n_h	640	960	896	768	960
l_r	1e-3	1e-3	1e-4	1e-3	1e-4
SOU	CP/QRA	Norm/CQN/OCQN	Jsu/CQJ/OCQJ	Stu/CQS/OCQS	QR/CQR/OCQR
n_h	768	896	896	960	960
l_r	1e-4	1e-3	1e-4	1e-4	1e-4
SARD	CP/QRA	Norm/CQN/OCQN	Jsu/CQJ/OCQJ	Stu/CQS/OCQS	QR/CQR/OCQR
n_h	896	960	960	640	896
l_r	1e-3	1e-3	1e-3	1e-3	1e-3
SICI	CP/QRA	Norm/CQN/OCQN	Jsu/CQJ/OCQJ	Stu/CQS/OCQS	QR/CQR/OCQR
n_h	768	768	512	960	896
l_r	1e-4	1e-5	1e-3	1e-3	1e-5

tuned by cross-validation. The CP calibration subset involves 182 samples (i.e., the preceding 6 months) as in the QRA benchmark to achieve comparable results.

In principle, a wide range of alternative configurations could be considered during the setup of the learning framework and the hyperparameter experiments. Since our goal is not to obtain the best NN architecture but to evaluate the introduction of the conformal inference framework under coherent backbone DE setups, we performed a restricted grid search based cross-validation procedure. The following arrangements have been adopted.[5] The batch size has been set to 64. As the different output layer's parameterizations and loss functions may require specific complexities and rates, the hidden unit size and the learning tuning are searched in the discrete sets $n_h \in$ [64, 128, 512, 640, 768, 896, 960] and $l_r \in$ [1e-5, 1e-4, 1e-3, 1e-2] respectively. Clearly the selections remain equal in the subsequent conformal inference stages (e.g. in Jsu/CQJ/OCQJ). NN training is performed by a maximum number of 800 epochs, including an early stopping callback on the validation loss with a patience of 50 epochs. For the hyperparameters of the OCQ* methods we have investigated a common set across the different datasets and coverage levels, defined as $K_I = 10$, $\eta =$ 1e-2, $C_{sat} = 1.2$ given $T =$ 1e9 and a burn-in of 7 steps.

During the daily recalibration of out-of-sample test experiments, the oldest sample in the moving window is discarded, while leaving a 20 % subset to evaluate the loss for early stopping. The output quantiles of the distributional NNs are estimated by generating 10,000 samples for each test prediction.

Within the broad spectrum of approaches that can be considered to approximate the conditional price distribution given the input features in PEPF, including e.g., single PIs, quartiles, deciles, percentiles, to the full continuous density regression (see [1] and references therein), we selected the deciles as robust synthetic indicators. A finer-grained quantile regression NNs for all the percentiles along the day-ahead prediction horizon would require an ensemble of NNs each with 2400 outputs. While distributional NNs are more scalable in this regard—since the discretizations can be easily derived following sampling—we chose to conduct a proper comparison of both quantile regression and distributional NN architectures under consistent discretizations.[6]

Hence, the coupled deciles lead to a set of 4 different PIs—beyond the median—with coverage levels $1-\alpha \in$ [0.2, 0.4, 0.6, 0.8].[7] The results across all the aforementioned coverage levels $1-\alpha$ are reported in the following sections, to depict the performance among the approximated PIs, thus complementing the pinball score analysis which averages all the deciles. Besides, these provide suitable insights on the central parts of the approximate distributions, as indicated in [33] by assessing the distributional NNs coverage for $1-\alpha = 0.5$.

Overall, we implemented a quite extensive experimental setup, covering 7 different datasets, each involving a test set recalibration of at least 365 days (plus 182 days for the initial CP), 5 different network architectures (i.e., point, QR, Norm, Jsu, Stu) in ensembles of 4 components, which has led to about 76,580 different neural network trainings, beyond hyperparameter tuning.

Table 2 reports the hidden units and learning rate employed for the test set experiments, as selected by the grid-search procedure for each of the bidding zones. Large n_h have been chosen in most cases. We did not observe sensible differences in validation losses across close configurations (e.g., 640 vs 896). Still, the specific selections may be influenced by local minimizers reached by the training algorithm. These are marginalized by the Deep Ensemble during testing.[8]

3.4. Evaluation metrics and tests

To evaluate the probabilistic forecasting performances achieved on the test sets, we follow the common practices in the PEPF field (see e.g. Nowotarski and Weron [1] and references therein for further details). Calibration is first analyzed by means of the Kupiec test (with significance level of 0.05) for unconditional coverage, on both extreme and central PIs, for each day-ahead hour. Then, the average Pinball loss across the deciles (see Section 2) and the Winkler's score for the related miscoverage levels are employed as proper scoring rules. It is worth noting that the average Pinball score, computed over the distribution percentiles, is commonly employed in PEPF as a cheap discrete approximation to the continuous ranked probability score (CRPS) (see e.g. Nowotarski and Weron [1]). Statistical significance is evaluated via the multivariate Diebold and Mariano (DM) test on the differences in the loss norms between competing models. Point prediction errors are

[5] Dropout and $\ell 1/\ell 2$ regularizations have not been included in the layers since they are almost never chosen by the hyper-parameter tuner in [33].

[6] The open codebase in the GitHub repository enables future investigation of alternative distribution discretizations, e.g., non-uniform quantile discretization involving more quantiles on the tails of the approximated distribution, etc.

[7] e.g., the quantiles 0.1/0.9 lead to the PI of coverage $1-\alpha = 0.8$ with $\alpha = 0.2$, the quantiles 0.2/0.8 to the PI 1-$\alpha = 0.6$ with $\alpha = 0.4$, etc.

[8] Prediction with different model ensembling techniques is left for future extensions.

evaluated beyond the probabilistic forecasting performances by means of the Mean Absolute Error (MAE):

$$\mathrm{MAE} = \frac{1}{N}\sum_{n=1}^{N}|\mathrm{y}_n - \hat{y}_n| \tag{18}$$

3.4.1. Kupiec test

The Kupiec test is aimed at assessing the unconditional coverage of the PIs to the nominal rate $1-\alpha$. Defining the sequence of indicators: $I_n = \mathbb{1}\{\mathrm{y}_n \in [\hat{L}_n, \hat{U}_n]\}$, with $\hat{L}_n, \hat{U}_n$ as the lower and upper bounds respectively, it checks whether $\mathbb{P}(I_n = 1) = 1-\alpha$, assuming independent violations.

The test is performed on the likelihood ratio (LR) statistic for unconditional coverage, $\chi^2(1)$ asymptotically distributed, defined as:

$$\mathrm{LR}_{\mathrm{UC}} = \frac{\alpha^{n_0}(1-\alpha)^{n_1}}{(1-\pi)^{n_0}\pi^{n_1}}, \text{ with: } \pi = n_1/(n_1+n_0) \tag{19}$$

where n_1,n_0 represent the number of hits and violations in the indicator series. We refer to previous reviews in the literature (see, e.g., Weron [14]) for further details on the common exploitation of the Kupiec test within the PEPF field.

3.4.2. Winkler's score

The Winkler's score (or interval score) is a proper scoring rule to assess probabilistic forecasts formed as Prediction Intervals (PI) at discrete coverage levels $1-\alpha$ (see [55, section 5.9]). Formally, it is defined as:

$$Winkler_n = \begin{cases} \delta_n, & \text{if: } \mathrm{y}_n \in [\hat{L}_n, \hat{U}_n] \\ \delta_n + \frac{2}{\alpha}(\hat{L}_n - \mathrm{y}_n), & \text{if: } \mathrm{y}_n < \hat{L}_n \\ \delta_n + \frac{2}{\alpha}(\mathrm{y}_n - \hat{U}_n), & \text{if: } \mathrm{y}_n > \hat{U}_n \end{cases} \tag{20}$$

where $\delta_n = \hat{U}_n - \hat{L}_n$ is the width of the $1-\alpha$-PI. The first case in (20) rewards narrow PIs (i.e., sharpness), while the others penalize the occurrence of test observations outside the predicted interval.

3.4.3. Diebold–Mariano (DM) test

The model-free Diebold–Mariano (DM) test is widely adopted in the EPF literature to evaluate the statistical significance of differences between the model's predictions (see e.g., [12,14] and references therein for further details). Although we acknowledge the valid points regarding the misuse of this test reported in [56] (e.g., it was not originally intended for model comparison), we include the analysis of the DM test to maintain consistency with previous works on distributional neural networks.

Rather than averaging the prediction scores over the dataset - as in conventional test set metrics analysis - the DM test is based on the computation of pairwise score differentials between the forecasts $\hat{y}_d^{Mi}$ and $\hat{y}_d^{Mj}$ provided by competing models M_i and M_j over the test days $d = [1, \ldots, N_d]$, followed by an asymptotic z-test to assess the null hypothesis that the expected value of the differential series is zero.

Two versions of the DM test exist, namely the univariate (executed hour-wise), and multivariate (performed jointly across all day-ahead hours). Following the aforementioned research studies, we exploit the latter for the present work. Notably, aside from being more fitted to multi-hour prediction frameworks and related cross-hour combinations in the overall training loss functions, it enables a more user-friendly summary representation of the test results through graphical heat maps.

Formally, the forecast-score differentials are computed as follows:

$$\Delta_d^{M_i,M_j} = \mathcal{L}\left[\varepsilon_d^{M_i}\right] - \mathcal{L}\left[\varepsilon_d^{M_j}\right], \text{ with: } \mathcal{L}\left[\varepsilon_d^{M_i}\right] = \left(\sum_{h=1}^{24}|\varepsilon_{d,h}^{M_i}|^n\right)^{1/n} \tag{21}$$

where $\varepsilon_{d,h}^{M_i} = \left(y_{d,h} - \hat{y}_{d,h}^{M_i}\right) \in \mathbb{R}$, and the arbitrary loss function $\mathcal{L}$ is stated in common n-norm form. Then, the DM statistic is derived as:

$$\mathrm{DM}^{M_i,M_j} = \sqrt{N_d}\frac{\hat{\mu}^{M_i,M_j}}{\hat{\sigma}^{M_i,M_j}} \tag{22}$$

with $\hat{\mu}^{M_i,M_j}$ and $\hat{\sigma}^{M_i,M_j}$ depicting the sample mean and standard deviation of the differential series $\Delta_d^{M_i,M_j}$ computed over an out-of-sample test period of length N_d. In practice, the p-values of two one-sided tests are often investigated, framed on the null hypothesis $\mathrm{H}_0 : \mathbb{E}\left[\Delta_d^{M_i,M_j}\right] \leq 0$ and alternative $\mathrm{H}_1 : \mathbb{E}\left[\Delta_d^{M_i,M_j}\right] > 0$. Hence, the H_0 rejection suggests a statistically significant performance improvement in the predictions provided by model M_i with reference to the one of model M_j. To this end, a p-value of 5 % is conventionally adopted as the test threshold.

3.5. Results analysis

The calibration tests of the benchmark methods, shown in Figs. 3–9, appear consistent with the results reported in previous studies. We observe worse coverage of Jsu and Stu on GE than in other cases, which could be due to the reduced heterogeneity in the DE components impacting the quantification of forecast uncertainty [57]. It can be noted that the behaviour of the benchmarks differs between the regional markets, as shown e.g. in the Kupiec plots of NOR vs SICI. This could be related to task specific characteristics (e.g., volatility extent) and requires further investigation, e.g. by exploring alternative parameterized densities and datasets. The introduction of conformal inference has led to improved hourly reliability in both distributional and QR settings. While the conventional CP on absolute residuals already provides adequate coverage, the effect of a more flexible approach is observed by assessing sharpness beyond calibration through the scoring rules (reported below). In fact, the former results in symmetric average bands around the model predictions, thus lacking the capability to adapt the PIs' width on simpler/harder test conditions as required for sample-wise efficiency.

Further insights are provided by the hourly Prediction Interval Coverage Probability (PICP) values, reported in Figs. 10 and 20 for $\alpha = 0.2$ and $\alpha = 0.6$ respectively, and the average PICPs for each of the coupled deciles in Table 3. To support graphical readability, we have structured three subplots aggregating: the benchmark methods in the left column; the different conformalized quantile regression setups (CQ*) in the middle column and the on-line conformal PI control integration (OCQ*) in the right column. The QRA benchmark coverage is included in all subplots to provide a coherent base within the common y axis scale. The red dashed lines represent the target coverage value $1-\alpha$.[9]

The baseline Jsu and Stu have obtained hourly coverage below the targets on average, which could be due to different uncertainties quantified on the samples observed during NNs training. Conversely, the Norm distributional form has resulted in being under-confident in several cases (see e.g., NORD and CNOR in Fig. 10). The hourly coverage of the distribution free CP approach on absolute residuals agrees with the results of the Kupiec test reported above. The QR benchmark exhibits excessively narrow PIs, which can be attributed to both underlying quantile over-fitting phenomena during learning and data/distribution shifts. The combination in the ensemble seems to provide a limited contribution to fixing such issues. Despite the marginal coverage degree, the QRA application on the DE shows sensible prediction step-specific coverage fluctuations. This can be motivated by the averaging effect of the backbone quantile regression performed across the hours. These observations deserve further investigation, which is planned for future works.

The introduction of the conformalized quantiles techniques has led to improved PICPs throughout the different DE configurations. The comparison of the related columns in Figs. 10, 20, and Table 3 displays the specific contributions to the compensation of the limited calibration in the baseline models (see, e.g., QR w.r.t. CQR/OCQR in GE). Still, the final calibration achieved is influenced by their initial miscoverage degrees. This results in slight fluctuations in PICP in Table 3 and PF scores in Table 4 across different datasets. Overall, the OCQ* settings show slightly better results than the CQ*. Still, we observe performance variations in OCQ* between different markets and coverage degrees (see, e.g.,

[9] e.g., $1-\alpha = 0.8$ for $\alpha = 0.2$.

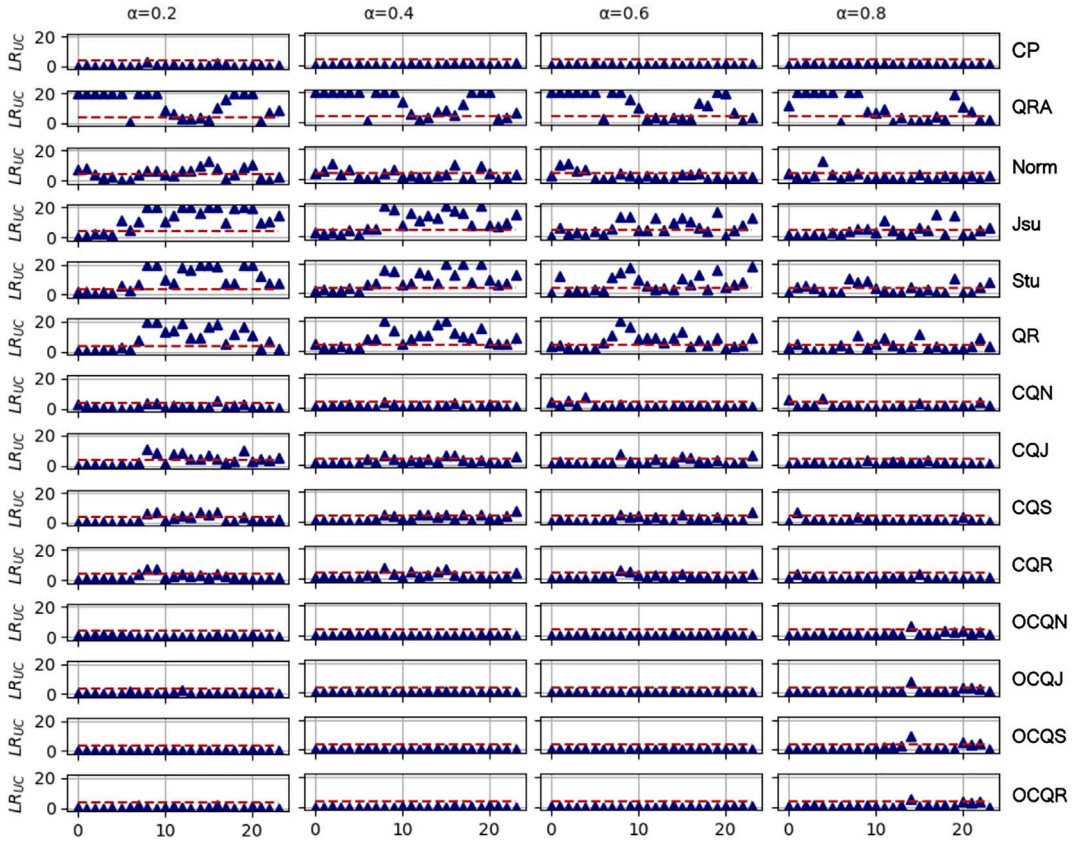


Fig. 3. Hourly Kupiec test on GE market test set.

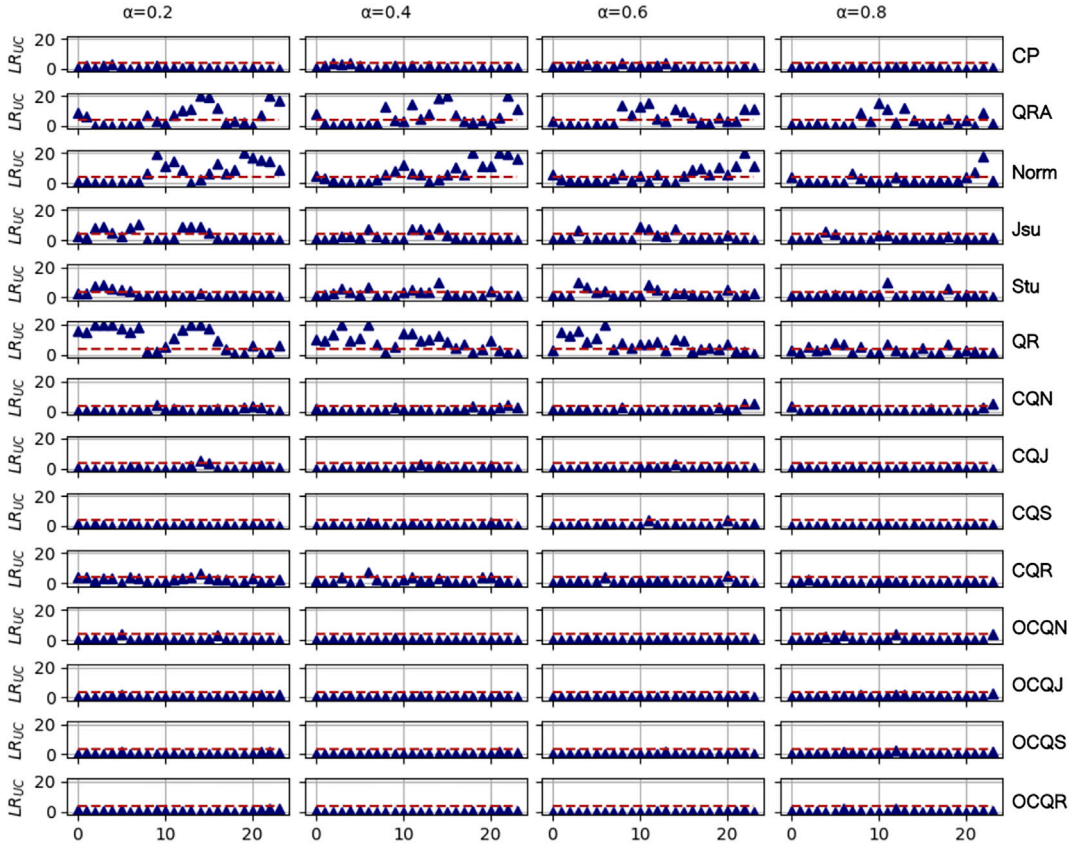


Fig. 4. Hourly Kupiec test on NOR market test set.

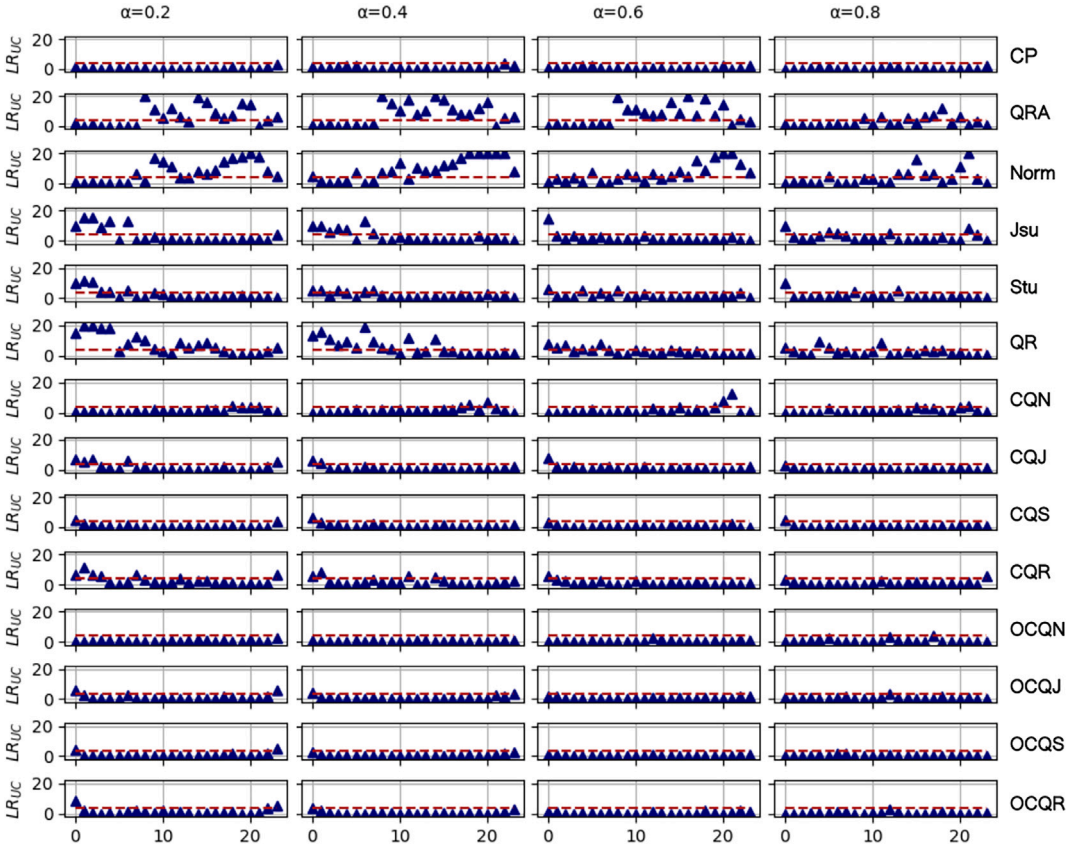


Fig. 5. Hourly Kupiec test on CNOR market test set.

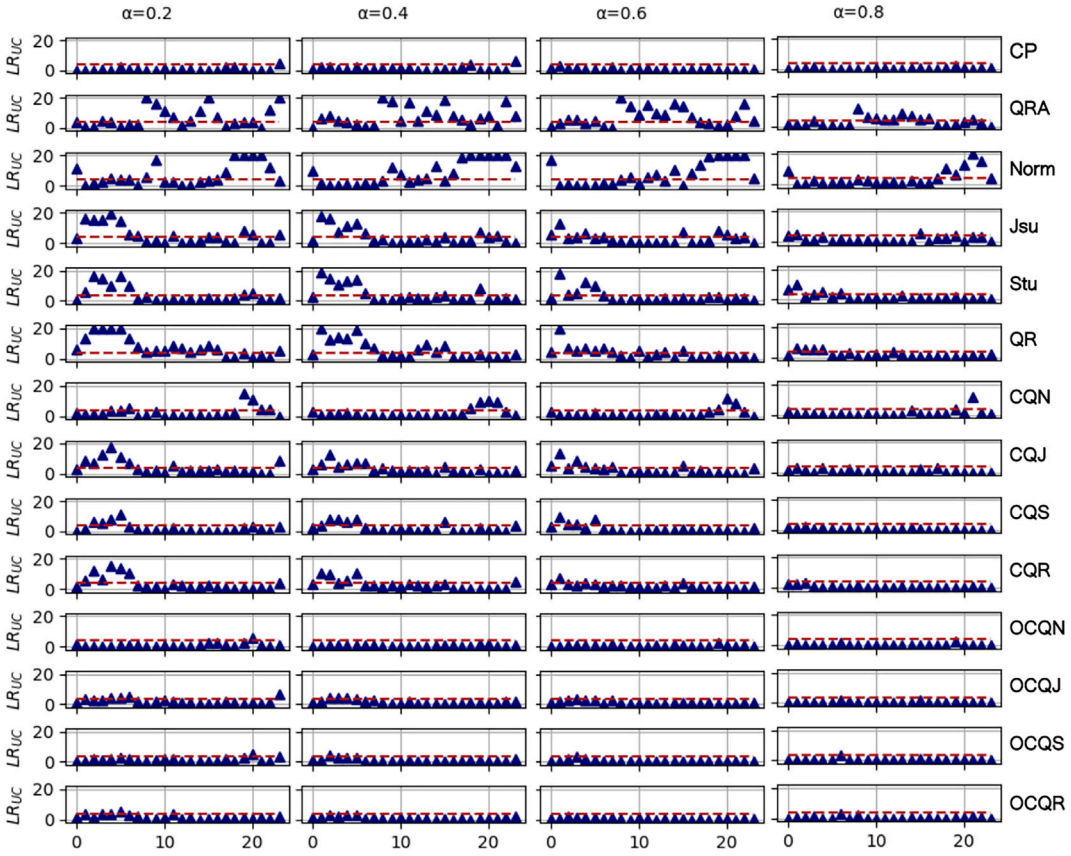


Fig. 6. Hourly Kupiec test on CSOU market test set.

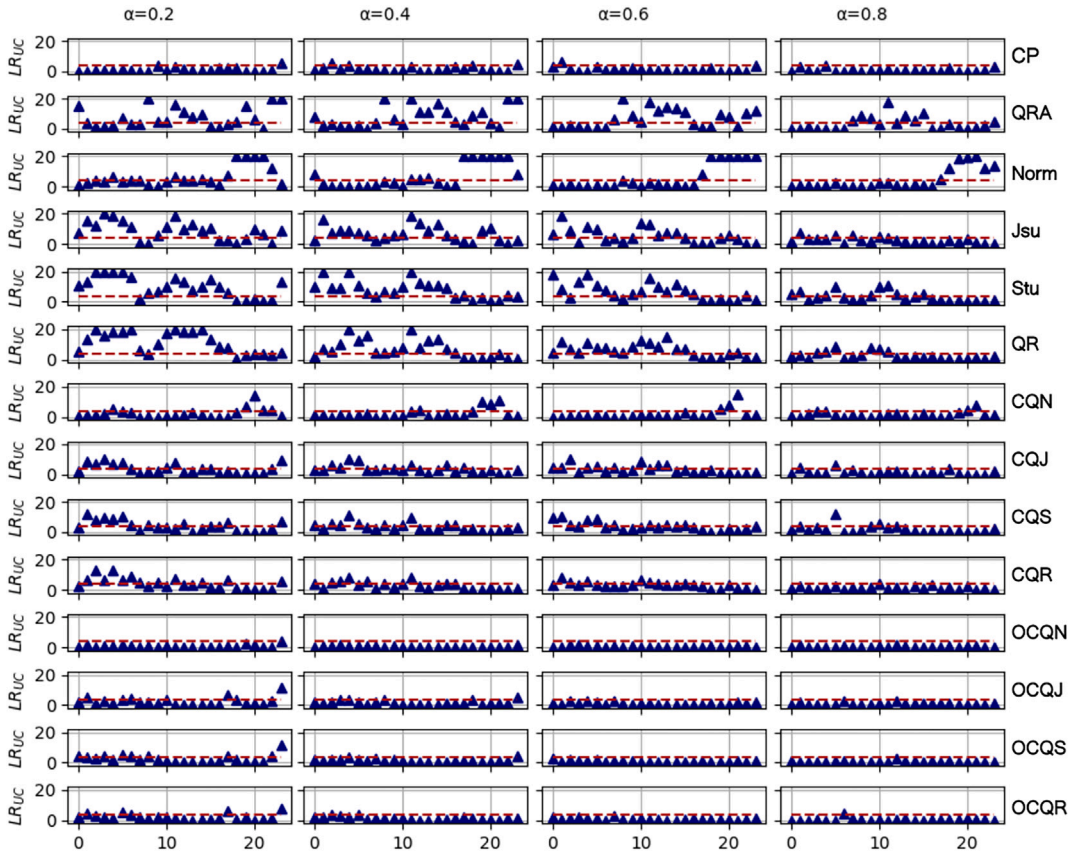


Fig. 7. Hourly Kupiec test on SOU market test set.

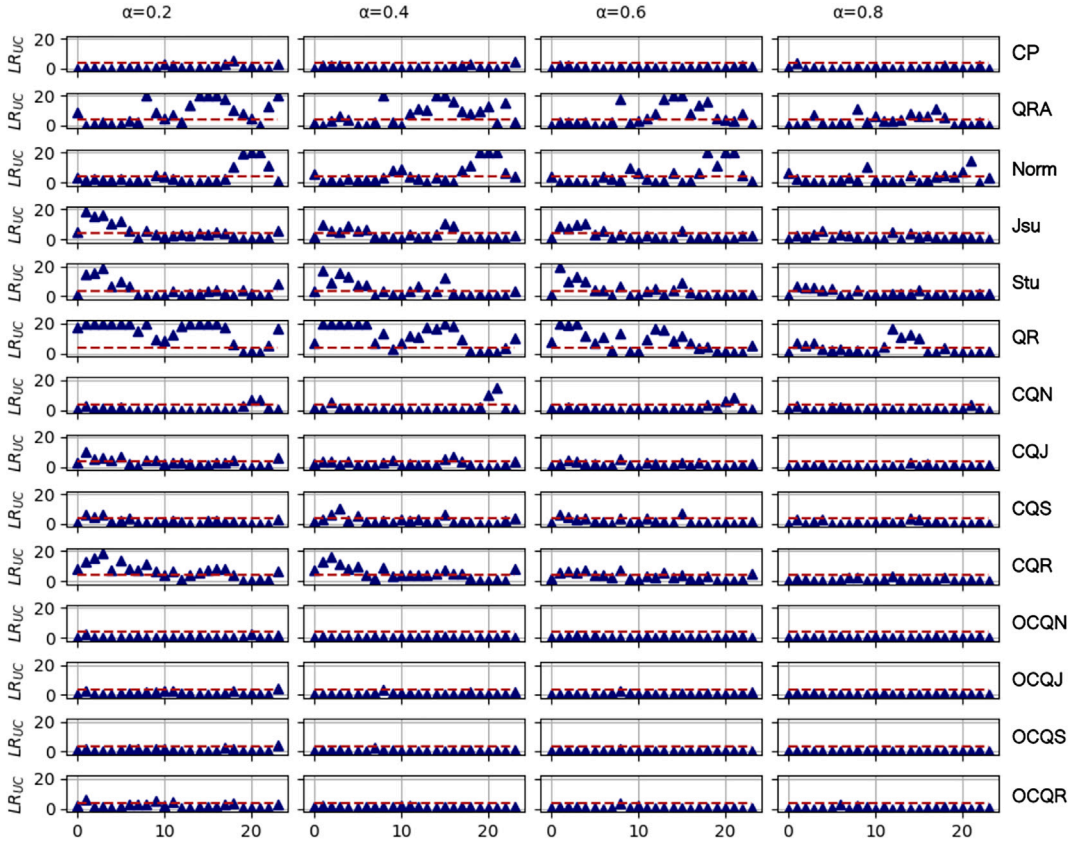


Fig. 8. Hourly Kupiec test on SARD market test set.

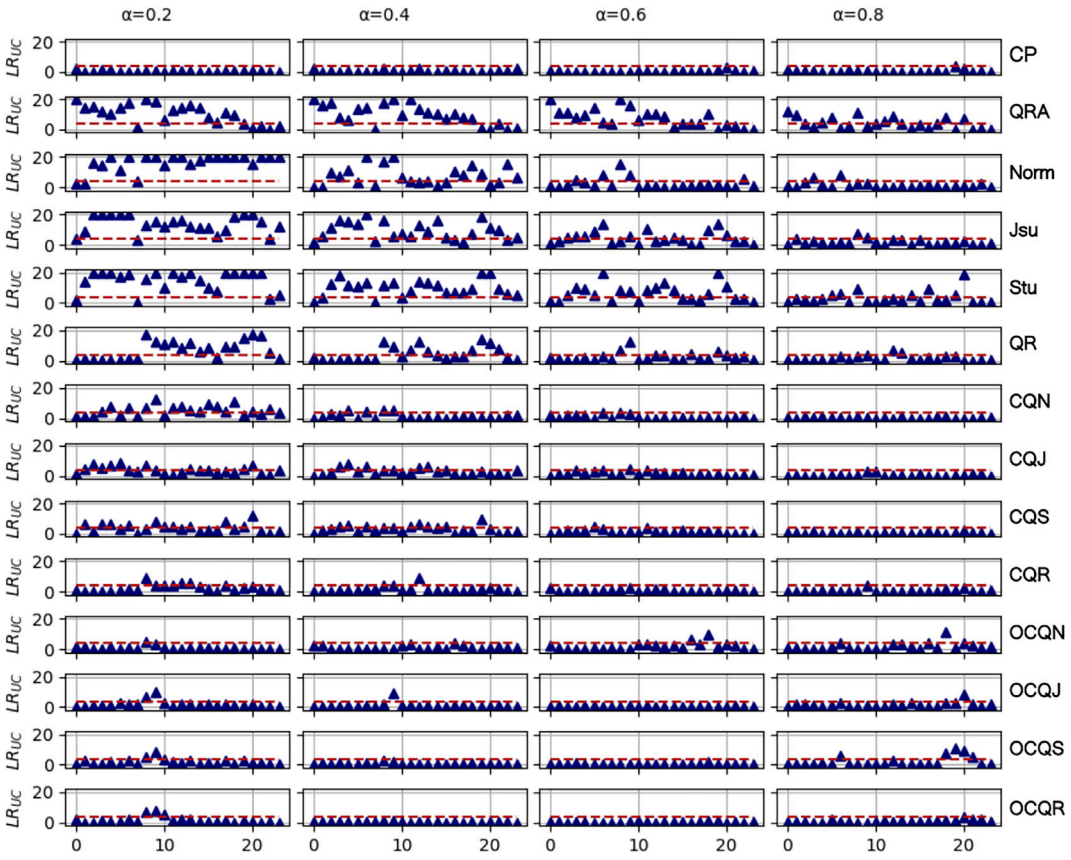


Fig. 9. Hourly Kupiec test on SICI market test set.

Table 3
Prediction interval coverage probability computed for each 1-α coverage degree.

		CP	QRA	Norm	Jsu	Stu	QR	CQN	CQJ	CQS	CQR	OCQN	OCQJ	OCQS	OCQR
GE	$PICP_{\alpha=0.2}$	0.79	0.78	0.78	0.74	0.75	0.75	0.79	0.77	0.78	0.78	0.81	0.80	0.80	0.80
	$PICP_{\alpha=0.4}$	0.59	0.59	0.60	0.54	0.54	0.54	0.60	0.57	0.57	0.57	0.61	0.60	0.61	0.60
	$PICP_{\alpha=0.6}$	0.40	0.39	0.40	0.35	0.35	0.35	0.40	0.38	0.38	0.38	0.41	0.40	0.40	0.40
	$PICP_{\alpha=0.8}$	0.20	0.20	0.20	0.17	0.18	0.18	0.20	0.19	0.19	0.19	0.22	0.21	0.21	0.21
NOR	$PICP_{\alpha=0.2}$	0.81	0.78	0.84	0.77	0.79	0.73	0.82	0.79	0.79	0.77	0.81	0.80	0.80	0.80
	$PICP_{\alpha=0.4}$	0.62	0.57	0.66	0.57	0.57	0.53	0.63	0.59	0.59	0.57	0.61	0.60	0.60	0.60
	$PICP_{\alpha=0.6}$	0.43	0.38	0.45	0.38	0.37	0.34	0.42	0.39	0.39	0.38	0.41	0.40	0.40	0.40
	$PICP_{\alpha=0.8}$	0.22	0.18	0.22	0.19	0.18	0.17	0.21	0.20	0.19	0.19	0.21	0.21	0.21	0.21
CNOR	$PICP_{\alpha=0.2}$	0.79	0.76	0.85	0.78	0.79	0.75	0.82	0.78	0.79	0.77	0.81	0.79	0.79	0.78
	$PICP_{\alpha=0.4}$	0.60	0.55	0.67	0.58	0.58	0.55	0.63	0.58	0.59	0.57	0.60	0.59	0.59	0.59
	$PICP_{\alpha=0.6}$	0.41	0.35	0.47	0.38	0.38	0.36	0.42	0.38	0.39	0.38	0.40	0.39	0.39	0.39
	$PICP_{\alpha=0.8}$	0.20	0.18	0.23	0.19	0.19	0.18	0.21	0.19	0.19	0.19	0.21	0.20	0.20	0.20
CSOU	$PICP_{\alpha=0.2}$	0.79	0.77	0.83	0.77	0.78	0.75	0.81	0.76	0.78	0.77	0.81	0.78	0.79	0.78
	$PICP_{\alpha=0.4}$	0.58	0.55	0.67	0.58	0.57	0.55	0.61	0.57	0.57	0.56	0.60	0.58	0.58	0.58
	$PICP_{\alpha=0.6}$	0.39	0.36	0.46	0.38	0.38	0.37	0.42	0.37	0.37	0.37	0.40	0.38	0.39	0.39
	$PICP_{\alpha=0.8}$	0.20	0.18	0.24	0.19	0.19	0.18	0.21	0.19	0.19	0.19	0.21	0.20	0.20	0.20
SOU	$PICP_{\alpha=0.2}$	0.78	0.79	0.81	0.75	0.74	0.74	0.80	0.76	0.76	0.76	0.79	0.78	0.77	0.78
	$PICP_{\alpha=0.4}$	0.57	0.58	0.64	0.56	0.54	0.55	0.60	0.56	0.56	0.57	0.59	0.57	0.58	0.58
	$PICP_{\alpha=0.6}$	0.38	0.37	0.44	0.36	0.34	0.36	0.40	0.36	0.35	0.36	0.39	0.38	0.38	0.38
	$PICP_{\alpha=0.8}$	0.19	0.18	0.22	0.18	0.17	0.18	0.20	0.18	0.18	0.18	0.21	0.21	0.20	0.20
SARD	$PICP_{\alpha=0.2}$	0.78	0.76	0.83	0.76	0.77	0.71	0.80	0.77	0.78	0.75	0.80	0.78	0.78	0.78
	$PICP_{\alpha=0.4}$	0.58	0.55	0.64	0.56	0.56	0.51	0.61	0.56	0.56	0.55	0.60	0.58	0.58	0.58
	$PICP_{\alpha=0.6}$	0.38	0.35	0.45	0.37	0.36	0.33	0.41	0.37	0.37	0.36	0.40	0.39	0.39	0.38
	$PICP_{\alpha=0.8}$	0.19	0.17	0.23	0.18	0.18	0.17	0.20	0.19	0.18	0.19	0.20	0.20	0.20	0.20
SICI	$PICP_{\alpha=0.2}$	0.79	0.79	0.70	0.72	0.71	0.75	0.75	0.76	0.76	0.77	0.79	0.78	0.77	0.78
	$PICP_{\alpha=0.4}$	0.59	0.58	0.54	0.52	0.52	0.56	0.58	0.56	0.56	0.57	0.62	0.59	0.59	0.60
	$PICP_{\alpha=0.6}$	0.40	0.39	0.37	0.35	0.34	0.37	0.39	0.37	0.37	0.39	0.43	0.40	0.40	0.40
	$PICP_{\alpha=0.8}$	0.20	0.20	0.19	0.18	0.17	0.19	0.20	0.20	0.19	0.20	0.23	0.22	0.22	0.22

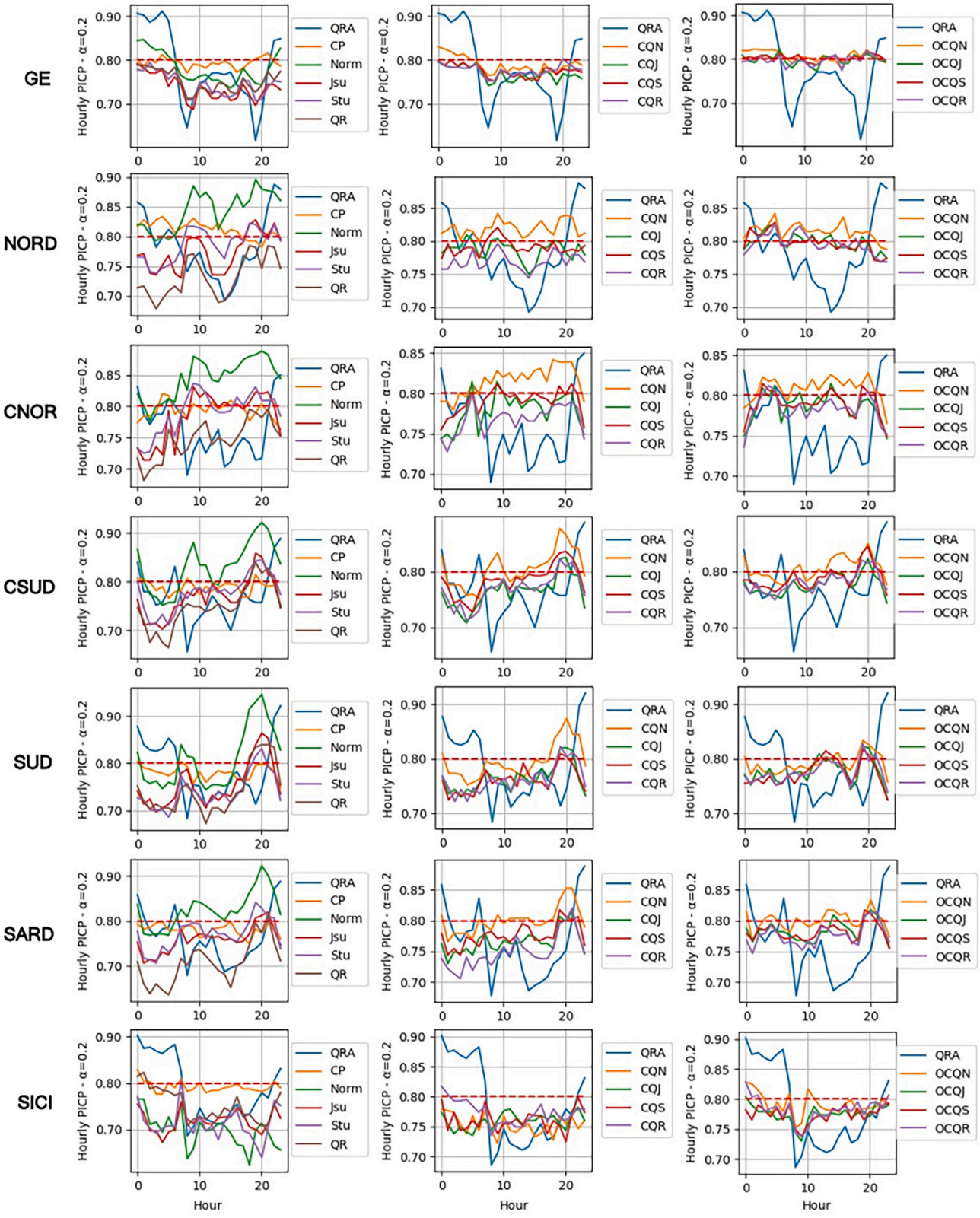


Fig. 10. Hourly PICP of the $PI_1 - \alpha = 0.8$ predicted by the different models on the test sets.

Table 4
Test set scores for each market and PEPF technique (best scores in bold).

		CP	QRA	Norm	Jsu	Stu	QR	CQN	CQJ	CQS	CQR	OCQN	OCQJ	OCQS	OCQR
GE	Pinball	1.549	1.557	1.489	1.467	1.473	1.474	1.483	1.457	1.463	1.462	1.480	**1.450**	1.456	1.459
	$\text{Winkler}_{\alpha=0.2}$	20.27	20.62	18.07	17.52	17.77	17.71	17.99	17.30	17.53	17.50	17.98	**17.18**	17.42	17.45
	$\text{Winkler}_{\alpha=0.4}$	14.43	14.57	13.68	13.45	13.51	13.52	13.62	13.36	13.40	13.41	13.58	**13.30**	13.34	13.40
	$\text{Winkler}_{\alpha=0.6}$	11.39	11.45	11.08	10.94	10.97	11.00	11.03	10.88	10.91	10.91	11.02	**10.83**	10.86	10.89
	$\text{Winkler}_{\alpha=0.8}$	9.27	9.28	9.13	9.04	9.05	9.06	9.09	8.97	9.00	9.00	9.06	**8.92**	8.96	8.97
	MAE	3.799	3.794	3.758	3.724	3.725	3.726	3.755	3.720	3.718	3.717	3.745	**3.700**	3.704	3.704
NOR	Pinball	1.847	1.861	1.855	1.823	1.809	1.806	1.852	1.827	1.810	1.804	1.850	1.822	1.804	**1.803**
	$\text{Winkler}_{\alpha=0.2}$	22.64	23.07	22.36	22.08	21.94	22.07	22.25	22.16	21.84	21.89	22.21	22.07	21.80	**21.79**
	$\text{Winkler}_{\alpha=0.4}$	16.95	17.19	17.02	16.73	16.56	16.58	16.96	16.78	16.60	16.54	16.94	16.73	**16.52**	16.53
	$\text{Winkler}_{\alpha=0.6}$	13.71	13.79	13.80	13.56	13.45	**13.43**	13.78	13.60	13.48	13.45	13.78	13.57	13.44	13.46
	$\text{Winkler}_{\alpha=0.8}$	11.31	11.34	11.41	11.19	11.11	**11.06**	11.41	11.22	11.13	11.08	11.39	11.19	11.10	11.08
	MAE	4.655	4.659	4.705	4.609	4.581	4.547	4.705	4.608	4.580	4.547	4.700	4.601	4.565	**4.543**
CNOR	Pinball	1.963	1.979	2.007	1.925	1.952	1.927	1.989	1.920	1.943	1.922	1.981	**1.913**	1.935	1.918
	$\text{Winkler}_{\alpha=0.2}$	24.30	24.64	24.65	23.60	24.18	23.87	24.30	23.59	23.94	23.70	24.20	**23.48**	23.82	23.57
	$\text{Winkler}_{\alpha=0.4}$	18.13	18.32	18.50	17.71	17.97	17.79	18.25	17.68	17.88	17.74	18.22	**17.62**	17.82	17.69
	$\text{Winkler}_{\alpha=0.6}$	14.56	14.66	14.90	14.29	14.46	14.29	14.75	14.23	14.39	14.26	14.71	**14.20**	14.34	14.25
	$\text{Winkler}_{\alpha=0.8}$	11.96	12.03	12.27	11.78	11.91	11.75	12.18	11.72	11.87	11.72	12.11	**11.69**	11.82	11.70
	MAE	4.918	4.952	5.047	4.851	4.908	4.830	5.046	4.848	4.906	4.828	5.025	4.825	4.886	**4.815**
CSOU	Pinball	1.994	1.996	2.050	1.977	1.990	1.976	2.033	1.964	1.968	1.963	2.028	**1.962**	1.970	1.964
	$\text{Winkler}_{\alpha=0.2}$	24.88	25.07	25.42	24.60	25.04	24.66	24.92	24.36	24.37	24.14	24.88	24.24	24.36	**24.09**
	$\text{Winkler}_{\alpha=0.4}$	18.38	18.45	18.96	18.27	18.40	18.29	18.73	18.17	18.13	18.15	18.71	**18.12**	18.14	18.14
	$\text{Winkler}_{\alpha=0.6}$	14.78	14.77	15.21	14.64	14.73	14.62	15.10	14.55	14.59	14.57	15.04	**14.54**	14.61	14.59
	$\text{Winkler}_{\alpha=0.8}$	12.14	12.12	12.47	12.03	12.07	12.00	12.42	11.93	11.99	11.97	12.38	11.96	12.03	11.99
	MAE	4.992	4.982	5.127	4.949	4.966	4.934	5.124	4.942	4.958	**4.927**	5.111	4.933	4.949	**4.927**
SOU	Pinball	2.176	2.170	2.209	2.164	2.178	2.130	2.184	2.138	2.152	2.110	2.185	2.136	2.150	**2.109**
	$\text{Winkler}_{\alpha=0.2}$	27.48	27.21	27.37	26.62	27.18	26.00	26.86	26.18	26.60	25.45	26.76	26.07	26.45	**25.26**
	$\text{Winkler}_{\alpha=0.4}$	20.18	20.12	20.45	19.93	20.12	19.60	20.09	19.66	19.79	19.35	20.09	19.60	19.76	**19.32**
	$\text{Winkler}_{\alpha=0.6}$	16.09	16.07	16.39	16.07	16.13	15.80	16.19	15.85	15.93	**15.68**	16.24	15.87	15.94	15.69
	$\text{Winkler}_{\alpha=0.8}$	13.17	13.16	13.44	13.22	13.24	13.05	13.33	13.06	13.12	**12.98**	13.37	13.07	13.14	12.99
	MAE	5.414	5.407	5.528	5.440	5.447	5.384	5.525	5.426	5.444	5.373	5.511	5.409	5.431	**5.368**
SARD	Pinball	2.258	2.261	2.282	2.262	2.234	2.275	2.266	2.252	**2.218**	2.255	2.267	2.254	2.221	2.251
	$\text{Winkler}_{\alpha=0.2}$	28.14	28.53	27.92	28.46	28.09	28.58	27.56	28.26	27.61	27.89	27.55	28.21	**27.54**	27.66
	$\text{Winkler}_{\alpha=0.4}$	20.84	20.93	21.01	20.97	20.59	21.00	20.82	20.86	**20.42**	20.79	20.83	20.88	20.45	20.71
	$\text{Winkler}_{\alpha=0.6}$	16.73	16.73	16.95	16.73	16.52	16.85	16.84	16.66	**16.41**	16.74	16.86	16.70	16.45	16.75
	$\text{Winkler}_{\alpha=0.8}$	13.75	13.70	13.96	13.71	13.57	13.81	13.88	13.63	**13.51**	13.74	13.90	13.67	13.55	13.76
	MAE	5.653	5.624	5.745	5.632	5.584	5.672	5.741	5.627	5.580	5.654	5.727	5.617	**5.575**	5.647
SICI	Pinball	4.433	4.440	4.646	4.396	4.428	4.341	4.581	4.346	4.362	**4.295**	4.584	4.348	4.360	4.302
	$\text{Winkler}_{\alpha=0.2}$	56.78	57.86	60.44	54.65	56.10	53.78	58.37	53.63	54.33	**52.71**	58.14	53.55	54.23	52.74
	$\text{Winkler}_{\alpha=0.4}$	41.45	41.58	43.35	40.66	41.13	40.09	42.85	40.11	40.39	**39.55**	43.02	40.25	40.39	39.66
	$\text{Winkler}_{\alpha=0.6}$	32.75	32.70	34.22	32.61	32.74	32.23	33.83	32.28	32.30	**31.91**	33.91	32.33	32.36	32.01
	$\text{Winkler}_{\alpha=0.8}$	26.62	26.54	27.83	26.72	26.75	26.43	27.43	26.42	26.43	**26.20**	27.47	26.44	26.47	26.27
	MAE	10.91	10.86	11.41	10.98	10.99	10.86	11.40	10.96	10.96	10.84	11.36	10.90	10.89	**10.81**

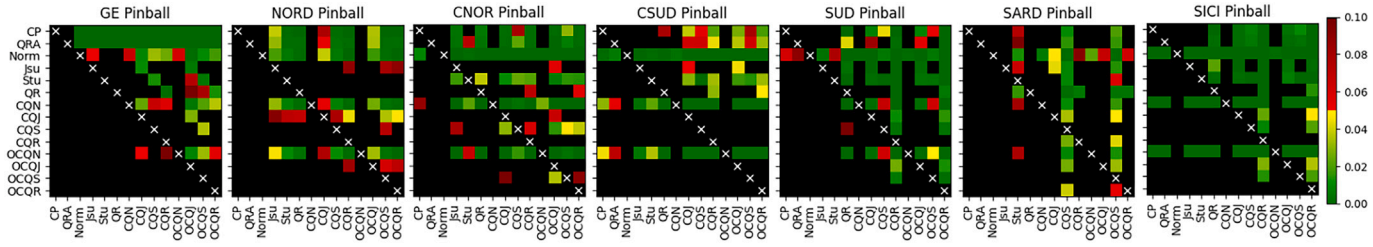


Fig. 11. DM test on average Pinball scores.

Table 5
% difference in Pinball losses, computed w.r.t. to CP, for each market and PEPF technique (lowest in bold).

	CP	QRA	Norm	Jsu	Stu	QR	CQN	CQJ	CQS	CQR	OCQN	OCQJ	OCQS	OCQR
DE	-	0.52	−3.87	−5.29	−4.91	−4.84	−4.26	−5.94	−5.55	−5.62	−4.45	**−6.39**	−6	−5.81
NOR	-	0.76	0.43	−1.3	−2.06	−2.22	0.27	−1.08	−2	−2.33	0.16	−1.35	−2.33	**−2.38**
CNOR	-	0.82	2.24	−1.94	−0.56	−1.83	1.32	−2.19	−1.02	−2.09	0.92	**−2.55**	−1.43	−2.29
CSOU	-	0.1	2.81	−0.85	−0.2	−0.9	1.96	−1.5	−1.3	−1.55	1.71	**−1.6**	−1.2	−1.5
SOU	-	−0.28	1.52	−0.55	0.09	−2.11	0.37	−1.75	−1.1	−3.03	0.41	−1.84	−1.19	**−3.08**
SARD	-	0.13	1.06	0.18	−1.06	0.75	0.35	−0.27	**−1.77**	−0.13	0.4	−0.18	−1.64	−0.31
SICI	-	0.16	4.8	−0.83	−0.11	−2.08	3.34	−1.96	−1.6	**−3.11**	3.41	−1.92	−1.65	−2.96

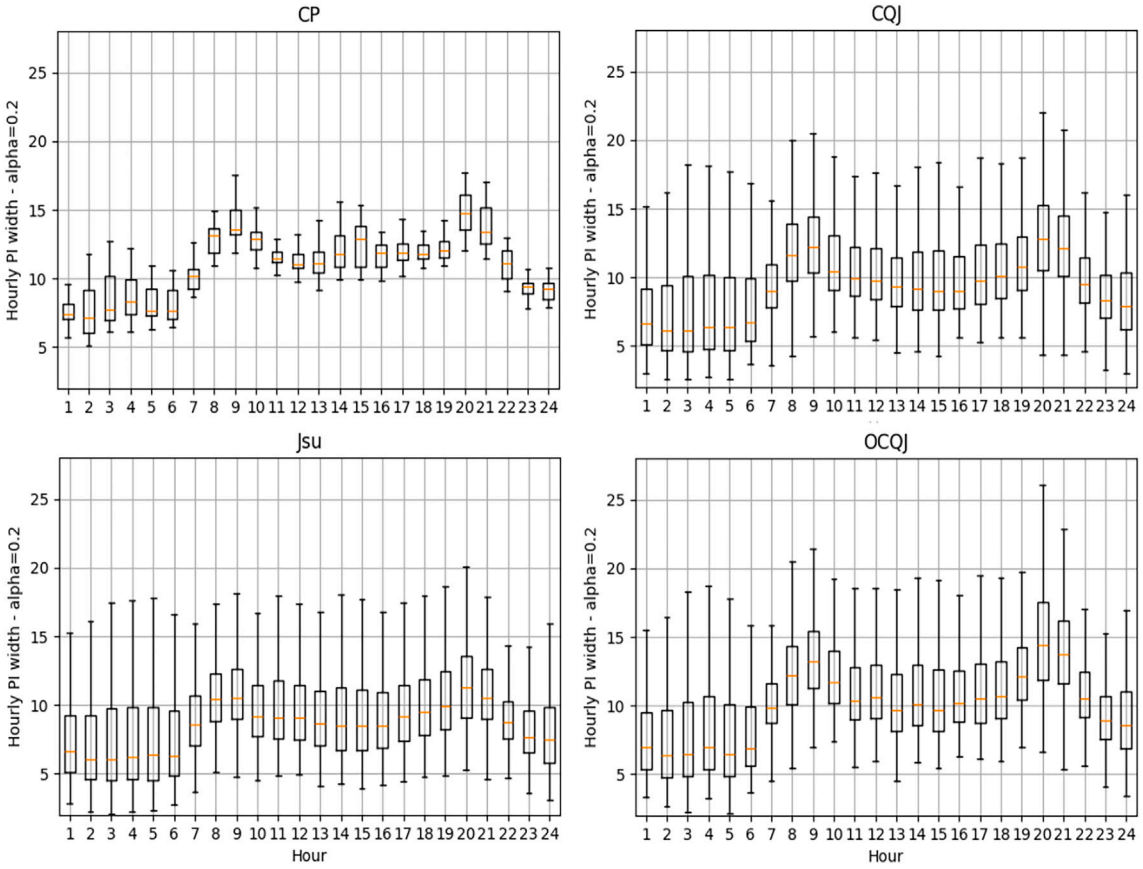


Fig. 12. Box plot of the distribution of the width of predicted PIs for 1-α = 0.8 on the GE test set.

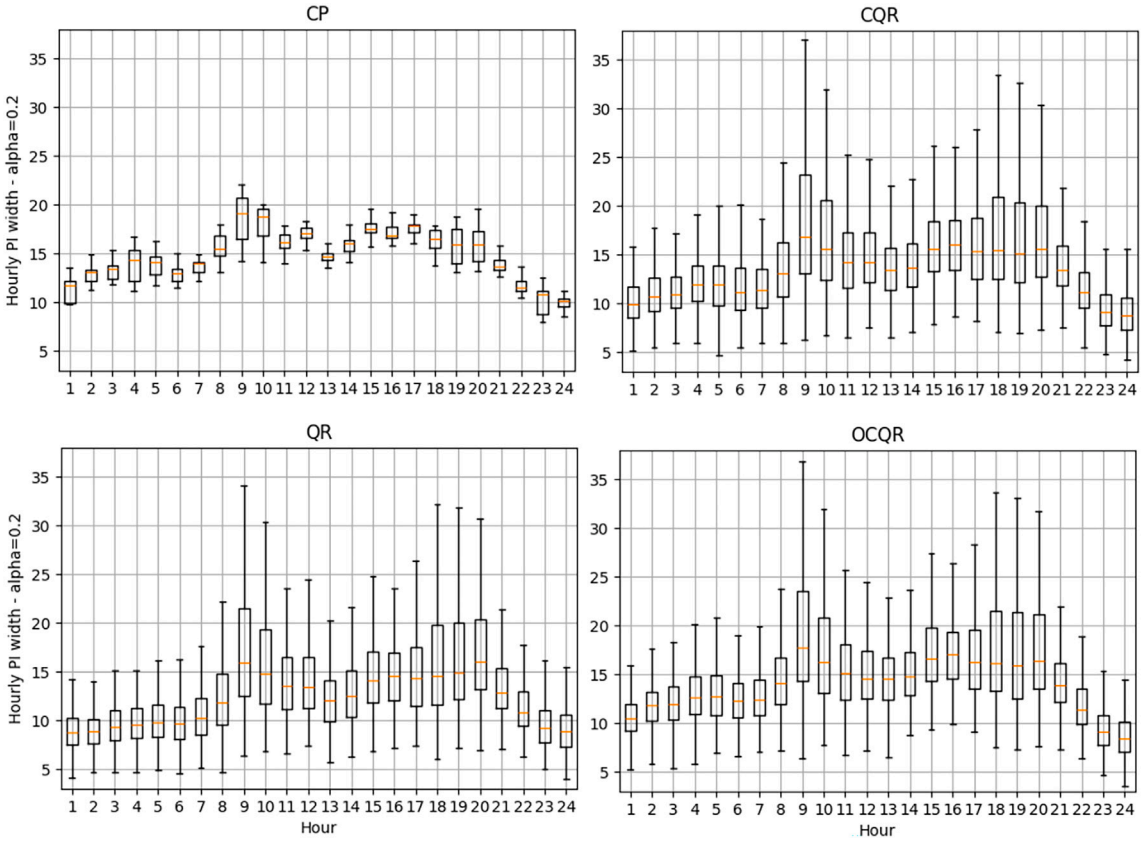


Fig. 13. Box plot of the distribution of the width of predicted hourly PIs for 1-α = 0.8 on the NORD test set.

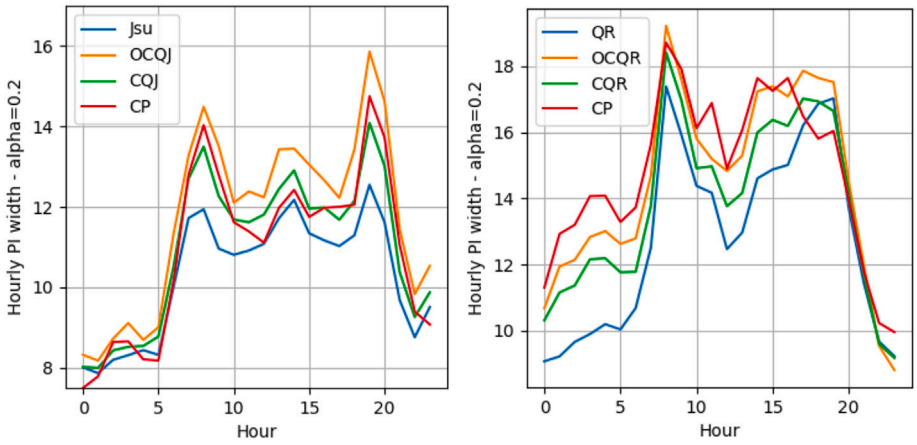


Fig. 14. Average width of predicted hourly PIs for 1-α = 0.8 on the GE (l) and NORD (r) test sets.

Table 6
Test set average PIs width for coverage 1-α = 0.8 on GE.

	CP	QRA	Norm	Jsu	Stu	QR	CQN	CQJ	CQS	CQR	OCQN	OCQJ	OCQS	OCQR
0	7.497	11.34	9.143	8.011	8.181	8.506	8.667	8.021	8.232	8.459	8.589	8.323	8.41	8.433
1	7.787	11.589	8.84	7.867	7.939	8.118	8.354	7.999	8.045	8.189	8.488	8.178	8.304	8.447
2	8.639	11.766	9.217	8.198	8.318	8.596	8.931	8.433	8.323	8.689	9.286	8.727	8.472	8.954
3	8.657	11.844	9.279	8.314	8.405	8.615	9.025	8.521	8.637	8.763	9.543	9.113	9.036	9.184
4	8.209	11.859	9.489	8.439	8.504	8.818	9.176	8.543	8.681	8.979	9.431	8.695	8.969	9.181
5	8.181	11.601	9.353	8.321	8.431	8.954	9.304	8.778	8.737	9.099	9.522	9.01	9.005	9.297
6	10.148	10.849	10.955	9.993	10.241	10.634	11.163	10.494	10.72	10.871	11.484	11.313	11.084	11.233
7	12.799	10.132	12.512	11.722	11.743	11.964	13.142	12.699	12.588	12.969	14.155	13.274	13.092	13.328
8	14.031	9.964	12.575	11.942	12.296	12.313	13.654	13.497	13.599	13.672	14.891	14.486	14.819	15.002
9	12.793	10.322	11.441	10.963	11.148	11.26	12.074	12.268	12.345	12.562	13.115	13.508	13.715	13.922
10	11.62	10.726	11.486	10.808	11.299	11.589	11.982	11.685	12.105	12.256	13.02	12.105	12.894	13.074
11	11.389	10.92	11.776	10.915	11.497	11.653	12.039	11.618	12.126	12.347	12.626	12.384	12.773	12.855
12	11.111	11.234	11.765	11.081	11.663	11.623	12.199	11.81	12.356	12.453	12.67	12.238	12.607	13.107
13	11.979	11.6	11.979	11.719	12.287	11.669	12.151	12.447	12.784	12.279	12.659	13.43	13.56	12.981
14	12.424	11.743	12.212	12.176	12.494	11.787	12.434	12.909	13.223	12.297	13.01	13.447	14.155	13.085
15	11.752	11.539	11.674	11.343	11.708	11.345	11.896	11.958	12.209	11.803	12.495	13.051	13.385	12.649
16	11.98	11.26	11.619	11.167	11.648	11.565	11.914	11.986	12.328	12.24	12.402	12.662	13.148	12.591
17	12.004	10.504	11.476	11.026	11.572	11.579	11.578	11.68	12.091	11.987	12.337	12.226	12.511	12.627
18	12.054	9.853	12.024	11.294	11.518	11.562	12.425	12.15	12.141	12.492	14.321	13.425	13.119	13.498
19	14.751	9.432	12.746	12.55	12.557	13.155	14.105	14.087	14.335	14.629	15.937	15.861	15.615	15.599
20	13.722	9.626	11.49	11.617	11.634	12.415	12.839	13.02	13.137	13.555	14.501	14.62	14.444	14.782
21	11.055	10.108	10.289	9.688	9.866	10.419	10.639	10.38	10.49	11.012	11.087	11.449	11.219	11.373
22	9.397	10.372	9.783	8.76	8.952	9.329	9.793	9.264	9.39	9.688	9.843	9.83	9.78	9.957
23	9.078	10.928	11.079	9.511	9.205	9.663	10.823	9.874	9.49	9.942	10.842	10.535	9.848	10.182

GE vs SICI). A possible explanation for this behaviour could lie in the specific tuning of the proportional/integral actions hyper-parameters, through a flat grid-search procedure. This has been deployed according to the common approach in adaptive CI frameworks in the literature (see e.g., Angelopoulos et al. [42]). Despite already yielding valuable results, the development of more enhanced fine-grained auto-tuning techniques - specifically tailored for each prediction interval (PI) coverage level - may lead to further performance improvements and more stable results across the testing conditions. This constitutes yet another compelling direction for future research.

Aside from being more reliable, the conformalized models have preserved stable and in some cases improved scores, as shown in Table 4. Furthermore, the multivariate DM tests performed on the Pinball and Winkler's scores are displayed in Figs. 11 and 18 respectively, where the coloured cells highlight the p-value for the difference between models' predictions (see Section 3.4.3). The test set Winkler's scores are computed for each of the $PI_{1-\alpha}$ obtained from the approximated distribution deciles. Moreover, Table 5 reports the percentage difference in the Pinball loss for the different models across the datasets.

Among the alternative configurations, the parameterized Normal form has obtained the worst scores, while CQR shows slightly better capabilities than CQJ and CQS. Similar results can be observed for the respective OCQ* settings. Such observations may be motivated by the major adaptivity in distribution-free methods, beyond the predetermined density forms, in characterizing complex uncertainty patterns following the specific application's needs. Besides, from the inspection of the Winkler's score, it appears that CP tends to contribute more to the correction of the PIs closed to the tails, which could be related to the behaviour of the semi-parametric quantile estimation stage in regions with limited observations. This is visible e.g., in the DM plots of the Winkler's PI with $\alpha = 0.2$ vs $\alpha = 0.6$ regarding NOR and CSOU regions. However, these are just initial explanations and are worth further empirical analysis.

Figs. 12–14 provide further insights into the differences in the PIs obtained by the models on the test sets, focusing on the coverage level 1-α = 0.8. The detailed values for the different models and regions are included in Tables 6–12. In particular, we report the plots of the hourly distributions of the PI widths from the CP-DNN, Jsu-DNN, QR-DNN and

Table 7
Test set average PIs width for coverage 1-$\alpha = 0.8$ on NOR.

	CP	QRA	Norm	Jsu	Stu	QR	CQN	CQJ	CQS	CQR	OCQN	OCQJ	OCQS	OCQR
0	2.232	2.392	2.306	1.932	1.917	1.662	2.095	2.059	2.039	1.738	2.037	2.33	2.267	1.968
1	2.164	2.352	2.389	2.033	1.996	1.749	2.268	2.083	2.139	1.974	2.154	1.942	2.161	2.099
2	2.467	2.328	2.393	2.041	2.01	1.726	2.426	2.195	2.049	2.174	2.414	2.168	2.066	2.284
3	2.494	2.31	2.447	2.077	2.023	1.853	2.424	2.32	2.276	2.058	2.682	2.566	2.407	2.376
4	3.061	2.298	2.489	2.092	2.054	1.962	2.589	2.339	2.352	2.281	2.901	2.908	2.891	2.822
5	2.705	2.33	2.492	2.12	2.046	1.809	2.498	2.301	2.483	2.251	2.594	2.479	2.667	2.455
6	2.8	2.4	2.653	2.284	2.252	2.037	2.418	2.398	2.325	2.377	2.76	2.639	2.933	2.826
7	2.89	2.463	3.087	2.547	2.525	2.428	2.579	2.445	2.486	2.567	2.123	2.693	2.693	2.718
8	4.092	2.507	4.017	3.347	3.245	3.112	3.642	3.381	3.434	3.373	3.414	3.574	3.84	3.55
9	3.228	2.503	3.873	3.149	3.035	2.777	3.503	3.146	3.099	2.829	3.76	3.219	3.175	3.007
10	3.412	2.474	3.561	2.872	2.804	2.69	3.64	3.133	3.167	2.977	3.777	3.488	3.379	3.058
11	3.802	2.46	3.619	2.834	2.783	2.619	3.734	3.151	3.3	2.992	3.558	3.435	3.975	3.388
12	3.397	2.419	3.287	2.538	2.553	2.429	3.074	2.717	2.784	2.756	3.59	2.967	3.117	3.302
13	2.873	2.405	3.204	2.67	2.64	2.45	2.995	2.859	2.727	2.661	3.018	3.4	2.946	2.809
14	3.432	2.421	3.729	2.896	2.843	2.812	3.584	2.999	3.004	3.001	3.714	3.299	3.199	3.573
15	3.411	2.446	3.89	2.989	2.875	2.794	3.79	3.073	3.069	3.069	4.122	3.245	3.276	3.534
16	3.247	2.49	4.059	3.021	2.8	2.711	3.769	3.138	2.86	2.763	3.38	3.287	3.159	2.769
17	2.832	2.527	3.88	3.152	2.979	2.763	3.564	3.112	3.099	3.01	3.599	3.557	3.213	3.244
18	3.539	2.54	4.003	3.205	3.09	2.948	3.714	3.443	3.341	3.292	3.886	3.584	3.254	3.627
19	3.463	2.535	4.179	3.292	3.1	2.954	3.648	3.371	3.115	3.153	3.951	3.514	3.456	3.551
20	3.127	2.535	3.567	2.784	2.581	2.412	3.224	3.117	2.788	2.707	3.155	2.963	3.036	3.067
21	2.391	2.511	3.05	2.299	2.159	2.168	2.561	2.491	2.17	2.21	2.383	2.519	2.387	2.431
22	1.858	2.47	2.571	2.044	2	1.762	2.118	1.924	1.937	1.794	1.781	1.823	1.764	1.711
23	1.645	2.417	2.373	1.929	1.912	1.691	2.029	1.837	1.741	1.615	2.12	2.072	1.866	1.662

Table 8
Test set average PIs width for coverage 1-$\alpha = 0.8$ on CNOR.

	CP	QRA	Norm	Jsu	Stu	QR	CQN	CQJ	CQS	CQR	OCQN	OCQJ	OCQS	OCQR
0	2.437	2.454	2.647	1.882	1.993	1.81	2.452	2.096	2.313	2.089	2.471	2.433	2.551	2.488
1	2.112	2.402	2.726	2.035	2.092	1.86	2.382	2.188	2.266	2.1	2.393	2.41	2.647	2.369
2	1.97	2.377	2.668	2.057	2.125	1.966	2.389	2.172	2.124	2.137	2.367	2.467	2.347	2.407
3	2.176	2.356	2.717	2.068	2.136	2.081	2.573	2.106	2.234	2.189	2.755	2.396	2.409	2.337
4	2.431	2.365	2.679	2.136	2.177	2.074	2.511	2.175	2.316	2.281	2.989	2.574	2.758	2.832
5	2.481	2.386	2.653	2.15	2.18	1.951	2.443	2.315	2.201	2.231	2.376	2.825	2.482	2.602
6	2.816	2.456	2.786	2.291	2.345	2.086	2.658	2.506	2.445	2.422	2.904	2.872	2.745	2.663
7	2.84	2.532	3.201	2.621	2.655	2.575	2.889	2.637	2.734	2.533	2.64	3.159	3.117	2.64
8	3.319	2.556	4.151	3.584	3.387	3.328	3.937	3.583	3.737	3.685	3.977	3.994	4.088	3.47
9	3.336	2.539	4.199	3.456	3.252	3.149	3.855	3.22	3.239	3.063	3.789	3.122	3.163	2.68
10	3.149	2.514	3.859	3.119	2.991	2.864	3.632	3.078	3.181	3.02	3.505	2.951	3.35	3.105
11	3.117	2.497	3.799	3.046	3.031	2.837	3.591	3.132	3.102	2.958	3.449	3.078	3.709	3.28
12	3.046	2.456	3.439	2.73	2.759	2.455	2.936	2.768	2.74	2.472	3.222	2.979	2.637	2.876
13	3.122	2.458	3.507	2.78	2.749	2.518	3.104	2.829	2.732	2.659	3.047	3.057	3.047	2.989
14	3.221	2.462	4.098	3.122	3.05	2.872	3.731	3.329	3.194	3.299	3.559	3.991	3.817	3.861
15	3.06	2.485	4.179	3.244	3.066	2.969	3.498	3.072	2.849	3.013	2.991	2.985	3.004	3.382
16	3.453	2.539	4.249	3.23	3.042	2.985	3.642	3.355	3.133	3.24	3.335	3.749	2.982	3.463
17	3.128	2.58	4.238	3.302	3.19	2.976	3.552	3.17	2.979	3.079	3.401	3.506	3.017	3.204
18	3.634	2.616	4.369	3.4	3.358	3.183	3.99	3.158	3.347	3.293	3.913	3.097	3.329	3.534
19	3.367	2.634	4.719	3.705	3.531	3.306	4.299	3.708	3.567	3.422	4.301	4.03	3.455	3.736
20	3.505	2.627	4.614	3.442	3.131	3.151	3.984	3.59	3.367	3.441	3.087	3.774	3.958	3.512
21	2.636	2.598	3.955	2.7	2.541	2.587	3.132	2.49	2.562	2.588	2.534	2.611	2.511	2.559
22	1.889	2.536	2.92	2.16	2.199	2.016	2.451	1.914	2.078	2.047	2.23	2.041	2.102	2.095
23	1.79	2.478	2.509	1.853	1.967	1.818	2.03	1.722	1.868	1.84	2.056	2.082	2.129	2.076

Table 9
Test set average PIs width for coverage 1-$\alpha = 0.8$ on CSOU.

	CP	QRA	Norm	Jsu	Stu	QR	CQN	CQJ	CQS	CQR	OCQN	OCQJ	OCQS	OCQR
0	2.256	2.412	2.725	1.98	2.136	2.077	2.401	2.114	2.276	2.206	2.279	2.54	2.446	2.469
1	2.797	2.374	2.739	2.06	2.138	2.083	2.557	2.252	2.556	2.513	2.781	2.691	3.205	3.077
2	2.601	2.366	2.702	2.112	2.201	2.06	2.486	2.1	2.27	2.184	2.683	2.622	2.63	2.54
3	2.496	2.366	2.763	2.136	2.199	2.131	2.581	2.212	2.386	2.378	2.848	2.657	2.506	2.675
4	2.725	2.359	2.706	2.134	2.171	2.1	2.509	2.241	2.291	2.408	2.751	2.66	2.801	2.984
5	2.172	2.399	2.63	2.143	2.161	2.128	2.599	2.315	2.401	2.492	2.533	2.398	2.685	2.736
6	2.194	2.46	2.673	2.194	2.241	2.053	2.384	2.36	2.302	2.41	2.862	2.972	2.955	2.815
7	2.115	2.509	3.175	2.643	2.745	2.581	3.078	2.635	2.914	2.728	3	2.687	3.092	2.676
8	3.936	2.541	4.364	3.626	3.479	3.448	3.948	3.673	3.538	3.627	3.375	3.46	3.432	3.296
9	3.359	2.534	4.296	3.223	3.095	3.04	3.917	3.321	3.173	3.173	3.383	3.404	3.373	3.292
10	2.915	2.498	3.857	2.987	2.879	2.792	3.32	3.1	3.221	3.054	2.981	3.123	3.528	3.258
11	3.171	2.485	3.771	2.852	2.822	2.835	3.386	2.988	2.993	2.947	2.946	3.3	3.124	3.026
12	2.716	2.477	3.42	2.681	2.646	2.619	2.97	2.895	2.679	2.893	3.09	3.154	2.721	3.021
13	3.299	2.466	3.47	2.812	2.78	2.65	3.225	2.745	3.062	2.899	3.413	3.078	3.326	3.238
14	3.483	2.482	4.114	2.969	2.905	2.982	3.849	3.119	2.955	3.07	3.6	3.176	2.683	3.238
15	3.299	2.505	3.909	2.961	2.878	2.958	3.755	3.377	3.116	3.243	3.647	3.402	3.231	3.443
16	2.872	2.532	4.028	2.955	2.905	2.932	3.608	3.141	2.98	3.084	3.681	3.136	2.9	2.963
17	2.772	2.598	4.134	3.134	3.023	2.899	3.535	3.083	2.9	3.107	3.644	3.512	2.935	3.05
18	2.794	2.594	4.395	3.311	3.248	3.018	3.542	3.163	3.057	2.998	3.529	3.306	3.311	2.984
19	3.158	2.561	4.856	3.74	3.529	3.407	4.062	3.457	3.316	3.436	3.976	3.579	3.729	3.43
20	3.063	2.506	4.61	3.353	3.049	3.08	4.046	3.285	3.127	3.073	3.757	3.452	3.031	3.231
21	2.118	2.463	3.83	2.605	2.451	2.596	2.893	2.444	2.387	2.405	2.389	2.433	2.596	2.181
22	1.822	2.441	2.778	2.019	2.08	1.963	2.23	1.84	2.111	1.818	1.828	1.92	2.188	1.816
23	1.707	2.429	2.409	1.713	1.841	1.761	2.049	1.622	1.742	1.741	2.049	1.792	1.97	1.874

Table 10
Test set average PIs width for coverage 1-$\alpha = 0.8$ on SOU.

	CP	QRA	Norm	Jsu	Stu	QR	CQN	CQJ	CQS	CQR	OCQN	OCQJ	OCQS	OCQR
0	2.267	2.656	2.959	2.16	1.989	2.288	2.795	2.358	2.225	2.623	2.798	2.631	2.44	2.769
1	2.29	2.751	2.923	2.271	2.125	2.384	2.545	2.418	2.382	2.588	2.955	3.122	3.051	2.98
2	2.529	2.834	2.893	2.373	2.195	2.398	2.569	2.422	2.337	2.551	2.849	2.878	2.652	2.808
3	2.783	2.9	2.945	2.438	2.259	2.363	2.5	2.48	2.47	2.54	2.699	2.695	2.836	2.71
4	2.69	2.899	2.871	2.422	2.219	2.435	2.912	2.644	2.495	2.76	3.27	2.99	2.911	3.017
5	2.83	2.833	2.778	2.272	2.113	2.304	2.733	2.55	2.394	2.641	3.019	2.781	2.855	3.066
6	2.371	2.663	2.825	2.341	2.178	2.256	2.594	2.384	2.233	2.404	2.973	2.91	2.651	2.874
7	2.851	2.499	3.447	2.79	2.65	2.822	3.154	2.804	2.523	2.9	3.15	3.407	2.679	2.701
8	3.436	2.446	4.135	3.311	3.119	3.292	3.815	3.549	3.498	3.573	4.233	3.6	3.685	4.11
9	3.157	2.538	3.76	2.937	2.785	2.828	3.496	3.252	3.109	3.105	3.824	3.69	3.285	3.647
10	3.2	2.69	3.627	2.971	2.865	3.022	3.426	3.338	3.361	3.637	3.475	4.04	3.69	3.931
11	3.685	2.811	3.62	3.339	3.284	3.424	3.887	3.999	4.073	3.96	4.355	4.796	4.865	4.3
12	3.934	2.911	3.637	3.526	3.548	3.481	3.806	3.803	4.107	3.751	4.861	4.915	5.002	4.729
13	3.434	2.982	3.75	3.706	3.699	3.481	3.69	3.837	4.093	3.703	4.013	4.442	4.932	4.586
14	3.549	2.938	3.836	3.624	3.563	3.558	3.538	3.865	3.773	3.89	4.785	4.101	4.489	4.534
15	3.604	2.816	3.721	3.158	3.051	3.241	3.783	3.31	3.538	3.499	4.484	3.985	4.253	4.211
16	2.734	2.66	3.624	2.821	2.711	2.815	3.227	2.886	2.734	2.926	3.106	3.418	3.316	3.286
17	2.547	2.528	4.013	2.801	2.602	2.976	3.196	2.56	2.516	2.934	2.976	2.976	2.942	3.451
18	2.649	2.376	4.406	3.022	2.705	3.069	3.374	2.804	2.569	3.069	2.904	3.223	2.769	3.27
19	2.807	2.21	4.766	3.527	3.021	3.183	3.671	2.974	2.727	3.192	3.03	2.986	2.927	3.251
20	2.331	2.187	4.677	3.177	2.644	3.023	3.664	2.852	2.657	2.883	3.331	2.742	2.845	2.94
21	2.259	2.358	3.981	2.545	2.207	2.79	2.993	2.317	2.1	2.641	2.349	2.398	2.184	2.524
22	1.85	2.525	2.979	2.026	1.82	2.073	2.469	1.865	1.909	1.966	2.149	1.844	2.202	2.094
23	1.893	2.7	2.655	1.802	1.665	1.929	2.079	1.631	1.673	1.844	2.166	1.829	2.149	2.021

Table 11
Test set average PIs width for coverage 1-$\alpha = 0.8$ on SARD.

	CP	QRA	Norm	Jsu	Stu	QR	CQN	CQJ	CQS	CQR	OCQN	OCQJ	OCQS	OCQR
0	2.229	2.588	2.839	2.248	2.196	2.064	2.537	2.281	2.264	2.307	2.23	2.246	2.187	2.53
1	2.607	2.555	2.857	2.276	2.229	2.071	2.89	2.373	2.47	2.482	3.195	2.689	2.861	2.724
2	2.538	2.526	2.86	2.343	2.327	2.206	2.847	2.506	2.564	2.741	2.793	3.122	3.261	3.289
3	2.716	2.515	2.944	2.396	2.335	2.237	2.914	2.553	2.509	2.698	2.835	3.03	3.151	3.202
4	2.658	2.524	2.895	2.363	2.317	2.12	2.579	2.592	2.486	2.455	2.981	3.211	2.833	2.86
5	2.323	2.561	2.827	2.605	2.22	2.12	2.78	2.57	2.415	2.521	2.907	2.709	2.646	2.684
6	2.725	2.66	2.835	2.37	2.333	2.279	2.794	2.736	2.858	2.918	3.091	3.555	3.753	3.465
7	3.139	2.744	3.383	2.865	2.93	2.743	3.026	2.898	2.918	3.07	2.666	2.791	2.89	3.257
8	3.848	2.818	4.493	3.803	3.653	3.459	4.1	3.789	3.998	3.869	4.012	3.97	4.217	4.111
9	2.911	2.799	4.42	3.41	3.238	3.281	3.637	3.221	3.185	3.305	3.091	3.379	3.171	2.911
10	3.068	2.745	4.097	3.226	3.043	2.924	3.424	3.29	3.196	3.071	2.901	3.24	3.413	3.255
11	3.024	2.7	4.083	3.206	3.041	2.93	3.618	3.24	3.036	3.162	2.983	3.16	2.982	3.173
12	3.547	2.689	3.819	3.05	2.929	2.841	3.632	3.233	3.242	3.349	3.58	3.249	3.606	4.081
13	3.646	2.689	4.082	3.368	3.185	3.099	3.995	3.568	3.403	3.547	4.56	4.048	3.473	4.099
14	4.34	2.713	4.753	3.697	3.328	3.339	4.27	3.929	3.654	3.87	4.558	4.59	4.062	4.373
15	3.669	2.758	4.505	3.475	3.224	3.084	4.568	3.679	3.745	3.749	4.717	4.45	4.379	4.366
16	3.129	2.8	4.401	3.385	3.188	3.189	3.942	3.618	3.336	3.451	4.123	3.782	3.526	3.667
17	3.777	2.896	4.426	3.433	3.367	3.246	3.652	3.482	3.551	3.252	3.904	3.707	3.662	3.598
18	3.264	2.899	4.565	3.664	3.546	3.265	3.73	3.512	3.478	3.33	3.504	3.406	3.768	3.488
19	3.163	2.866	4.992	3.844	3.653	3.561	4.31	3.619	3.706	3.685	4.197	3.641	3.694	4.312
20	2.915	2.733	4.817	3.408	3.218	3.354	4.268	3.545	3.388	3.459	3.904	3.778	3.563	3.474
21	2.498	2.672	4.161	2.761	2.625	2.786	3.354	2.861	2.554	2.876	2.837	3.073	2.649	3.162
22	1.961	2.646	2.998	2.281	2.256	2.258	2.772	2.364	2.254	2.245	2.804	2.421	2.487	2.473
23	2.07	2.611	2.591	2.034	2.029	2.068	2.317	1.996	2.052	2.209	2.539	2.188	2.455	2.392

Table 12
Test set average PI widths for coverage 1-$\alpha = 0.8$ on SICI.

	CP	QRA	Norm	Jsu	Stu	QR	CQN	CQJ	CQS	CQR	OCQN	OCQJ	OCQS	OCQR
0	4.137	5.123	4.26	3.966	4.018	4.569	4.156	4.123	4.064	4.341	3.951	4.521	4.158	4.321
1	4.441	4.868	4.37	3.745	3.677	4.284	4.535	4.083	3.926	4.72	4.849	4.631	4.483	4.973
2	3.348	4.735	4.18	3.562	3.388	4.183	4.373	3.552	3.283	4.306	4.841	3.977	3.927	4.706
3	3.89	4.651	4.242	3.552	3.394	4.335	4.364	3.354	3.518	4.169	4.978	4.053	3.701	4.032
4	3.698	4.664	4.109	3.559	3.378	4.185	4.18	3.813	3.393	4.158	4.573	4.247	3.779	4.351
5	4.095	4.711	4.033	3.435	3.318	3.999	3.866	3.482	3.457	4.112	4.646	3.828	3.699	4.679
6	4.082	4.838	4.024	3.73	3.647	4.309	4.171	4.252	4.137	4.592	5.202	5.461	4.807	5.01
7	5.969	5.212	5.763	5.527	5.536	5.895	5.779	5.281	5.469	5.571	6.548	5.32	5.353	5.332
8	8.063	5.562	6.769	7.182	7.048	7.15	7.912	7.434	7.77	8.042	9.576	8.356	9.877	8.677
9	6.825	5.403	6.323	6.401	6.114	6.614	7.253	6.914	6.234	7.505	8.294	7.63	7.608	8.653
10	6.056	5.092	6.319	5.796	5.342	6.015	6.405	5.753	5.483	6.166	7.093	6.364	5.775	7.094
11	5.985	4.875	6.153	5.341	4.926	5.825	6.144	5.659	5.318	5.778	6.733	5.995	6.15	6.53
12	5.885	4.77	6.065	5.228	4.722	5.677	5.867	5.864	5.193	6.015	6.98	7.425	7.11	6.62
13	5.912	4.64	6.129	5.381	4.76	5.725	6.063	5.907	5.616	6.244	7.08	6.912	6.385	7.176
14	5.727	4.706	6.222	5.438	5.043	5.749	6.364	6.084	5.834	6.426	7.357	6.4	6.165	7.225
15	5.413	4.829	6.053	5.255	4.978	5.821	5.645	5.422	5.323	5.962	6.385	5.476	5.501	6.823
16	5.554	5.105	5.974	5.898	5.665	6.248	6.388	5.814	6.154	6.566	8.633	6.325	7.365	7.506
17	6.64	5.52	6.298	6.759	6.242	7.079	6.46	6.366	6.054	7.337	7.643	6.902	7.191	8.283
18	8.544	5.976	6.84	6.894	6.789	7.429	7.15	7.603	7.616	7.912	9.875	9.665	9.878	9.586
19	7.043	6.433	6.939	6.186	5.602	6.785	7.762	8.072	7.915	7.795	9.795	10.073	10.034	9.276
20	6.016	6.632	7.514	5.934	5.165	6.374	7.588	7.779	7.529	7.185	9.618	10.472	9.395	8.301
21	6.873	6.425	6.91	6.216	6.077	6.96	7.238	6.981	7.191	7.353	8.36	8.087	8.5	8.922
22	5.386	5.983	5.677	6.016	6.045	6.253	6.249	5.888	6.394	6.167	7.123	5.36	6.026	6.559
23	4.842	5.479	5.347	5.165	4.927	5.497	5.355	4.823	4.845	5.066	6.139	5.281	5.454	5.267

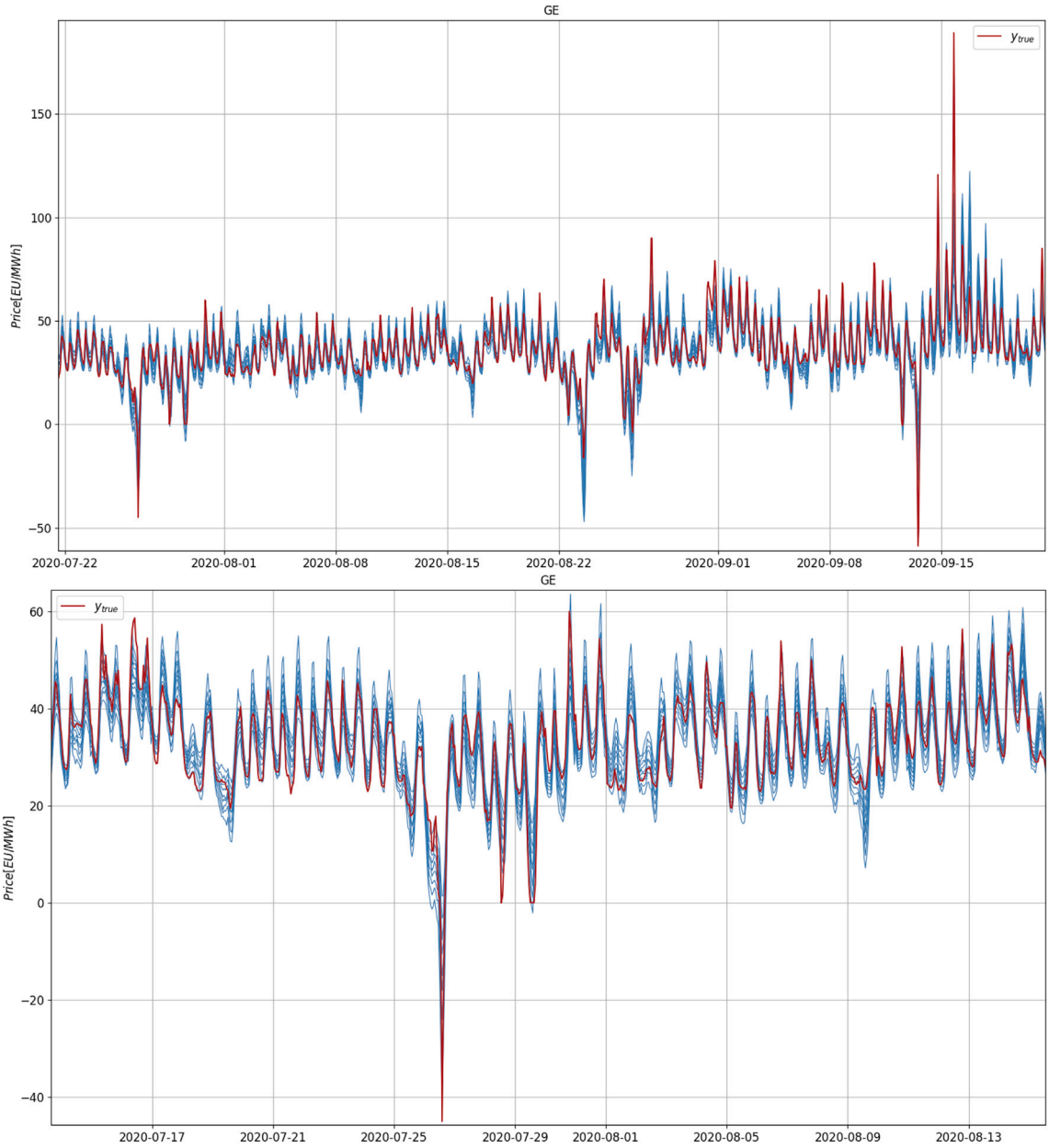


Fig. 15. Probabilistic forecasts on a subset of the GE test samples.

the related CQ* OCQ* integrations on GE and NORD bidding zones as a representative subset from the broad range of deployed configurations. Apparently, CP has led to weakly varying widths across the test samples, which aligns with the observations of previous works regarding its typical overly conservative behaviour, i.e., limited capability to adapt the PIs' width on simpler/harder test conditions [38]. Both the quantile regression and the distributional neural network forms show more adaptive patterns, which are related to their more flexible conditional mapping capabilities. Still, these models result in marginally overconfident predictions, as observed in the calibration analysis above. Notably, the conformalized quantile methods address this issue by specifically adjusting (e.g., increasing) the hourly prediction intervals (PIs) towards the target coverage level. This effect can also be observed, for example, by inspecting the widths around hour 8:00 of OCQJ with reference to the Jsu baseline, or OCQR with reference to the QR baseline. In fact, these observations align with the stable pinball scores and improved hourly coverage achieved.

Furthermore, we note that the introduction of conformal inference techniques — whether CP, CQ*, or OCQ* — has a negligible effect (i.e., a few seconds) on the computational time required for daily retraining of the backbone neural network ensembles. In the experimental setups employed, we observed that the different DNNs, whether in point form for CP/QRA, quantile regression-based, or for distributional parameterization, required similar average training times. Specifically, they took approximately $88_{\pm 26}$ seconds for a single day of recalibration, with variations depending on factors such as the selected layer dimensions and the number of epochs before early stopping intervention.

Aside from the probabilistic forecasting performances, Table 4 reports the point prediction Mean Absolute Errors, while the outputs of the related DM tests are depicted in Fig. 19. We observe that conformal

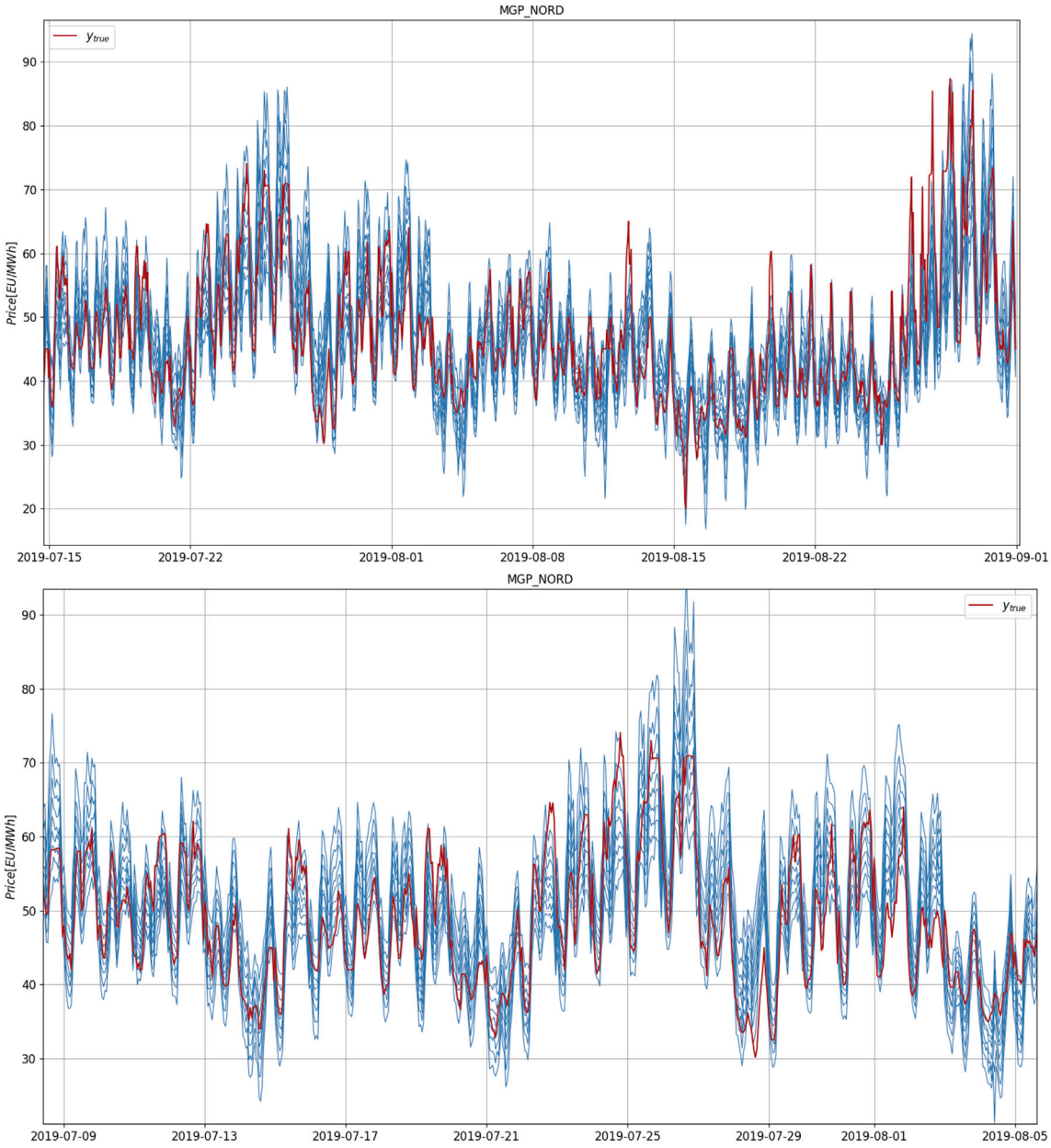


Fig. 16. Probabilistic forecasts on a subset of the NOR test samples.

inference has led to a slight improvement of the MAE in some cases (see e.g. CNOR), as the median prediction can be updated after quantile correction by the sorting procedure. The impact on point prediction accuracy of more flexible distribution parameterizations has also been reported in [33]. The gap in the MAE on GE could be related to the reduced input feature set.

To provide further insights into the different price behaviours addressed by the models among the regional markets, we plot in Figs. 15–17 extracts from the GE, NOR and SICI datasets. The deciles predicted by the OCQR model are reported along with the target series. Clearly, the price in each zone assumes specific shapes and values, which depend on the generation mix (e.g., share of wind power generation in the southern part of Italy) and the different power demands. We refer to [43] and references therein for further details on the specific characteristics of the different Italian bidding areas, and to [10] for a recent detailed analysis of the German power market.

4. Conclusions and next developments

The goal of PEPF, as probabilistic forecasting in general, is to maximize sharpness while respecting calibration requirements. In fact, efficient but reliable uncertainty quantifications are decisive in supporting decision making under stochastic conditions, as is the case in modern day-ahead EPF applications. In this work, we addressed the limited hourly calibration of the state of the art neural network (NN) ensemble based methods for PEPF by exploiting a flexible Conformal Prediction

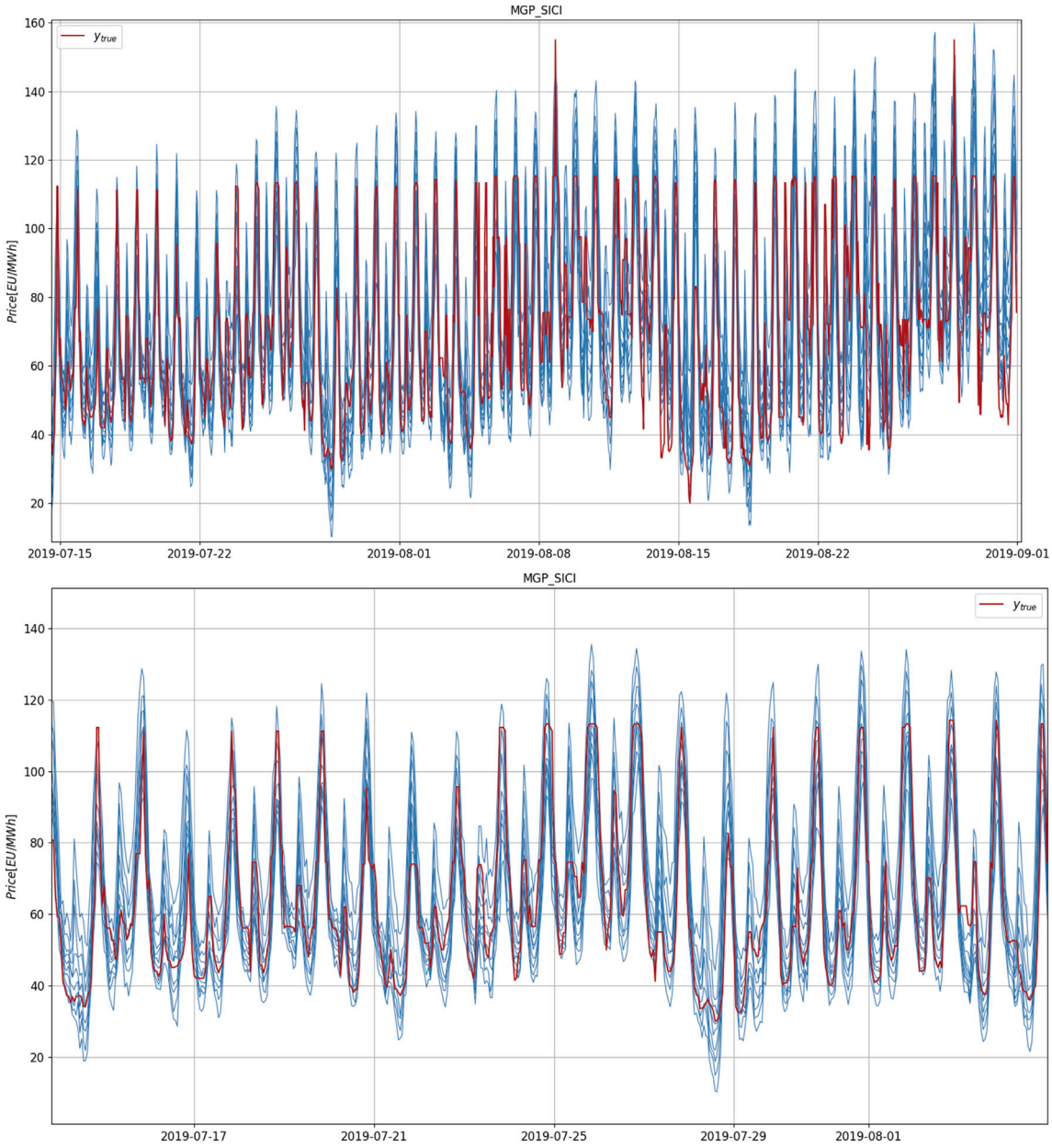


Fig. 17. Probabilistic forecasts on a subset of the SICI test samples.

(CP) framework. To this end, the quantiles approximated by the NNs at each prediction step have been incrementally conformalized over the test set following a daily retraining (i.e., recalibration). Both quantile regression and distributional NNs approaches have been exploited for estimating the conditional quantiles to be calibrated. A uniform vincentization technique has been employed to aggregate the quantiles produced by the underlying ensemble components, providing a simple but robust forecasting combination in practice. Moreover, a post-hoc sorting operator has been included to achieve conditional quantile non-crossing. The asymmetric Conformalized Quantile Regression formulation, adjusted in a multi-horizon forecasting setup, has been deployed to compute the scores and the conformalized quantiles. This supports a proper correction of both upper and lower bounds at each stage over the prediction horizon, leading to a more efficient local coverage in the features conditioned PIs estimation. Such capability is particularly crucial to tackle the complex distributional patterns typically observed in PEPF applications, such as sensible heteroskedasticity, skewness and fat tails. Besides, it compensates potential quantile overfitting of the conventional pinball loss based neural network training, enhancing the asymptotic validity with finite sample coverage guarantees. To address the lack of robustness to failures of the sample exchangeability assumptions, we have exploited an adversarial setup. Quantile tracking and coverage error integration stages have been included, aimed at compensating potential coverage degradation and systematic errors occurring over the recalibration window, e.g., due to sensible short-term drifts. The whole procedure has been started on a sample subset close to the first test date to acquire the initial bag of calibration scores, then proceeding in a rolling window fashion. Indeed, updating the ensemble components

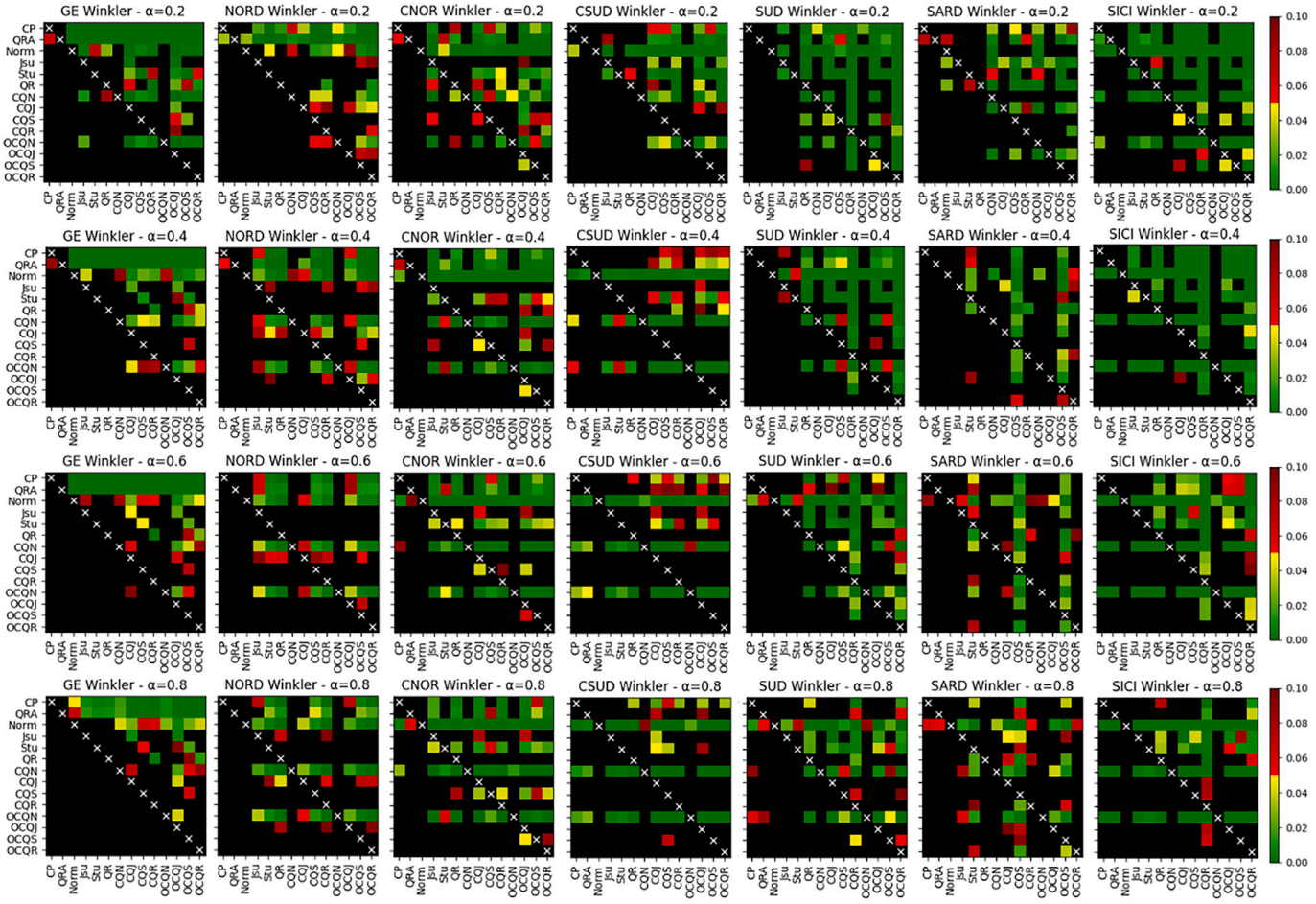


Fig. 18. DM test on Winkler's scores.

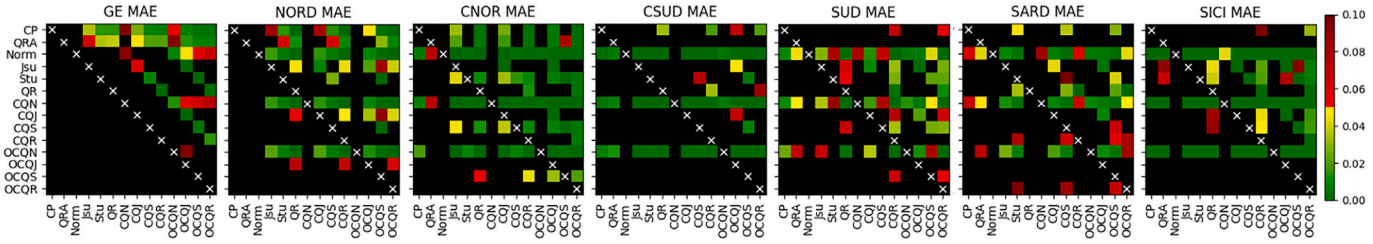


Fig. 19. DM test on MAE.

by including the most recent observations supports probabilistic performances besides point accuracy, as the underlying model quality impacts the local reliability of the prediction intervals despite the marginal coverage guarantees provided by the CP framework. Besides, the ensemble aggregation provides a further mechanism for this purpose, by getting rid of poor local minimizers in the network parameter space reached by single trainings.

The experiments have been executed on the German market dataset as well as on the different regional bidding zones constituting the Italian day-ahead markets, providing a compelling setup for testing under heterogeneous conditions. A comparison against state of the art benchmarks has been performed, including QRA, quantile regression NNs, distributional NNs, as well as conventional absolute score and normalized score based CP settings, showing the capability of the proposed approach to achieve day-ahead probabilistic forecasts with improved hourly coverage. Aside from being more reliable, the conformalized models have been shown to preserve stable and in some cases improved probabilistic scores.

Still, we envision several avenues of future research, including the integration of further adaptive CP wrappers to maintain coverage under data and concept drifts, mechanisms for fine-tuning the coverage levels, CP on early-stopping for improving sample-efficiency, as well as further baseline models, ensemble architectures and datasets. The deployment of computationally cheaper deep ensemble approaches represents an important issue to be addressed, to reduce the resource consumption of multiple network training during recalibration. Furthermore, we plan

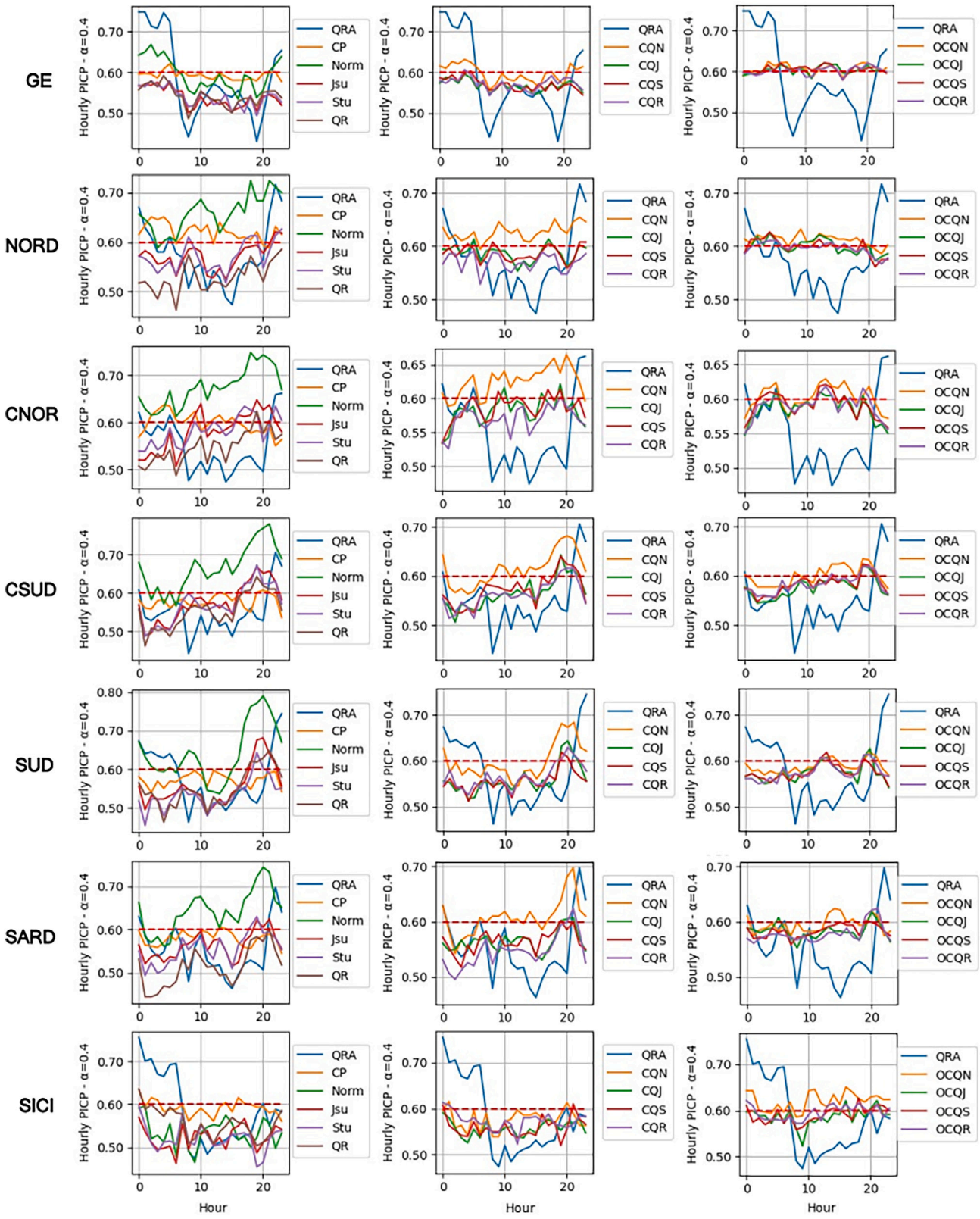


Fig. 20. Hourly PICP of miscoverege degree $\alpha = 0.4$ on the test sets.

to implement eXplainable AI techniques (e.g., SHAP [58]) to investigate the major features affecting the probabilistic predictions. Finally, despite the specific energy forecasting tasks addressed in this study, the developed approach can also be extended to other applications, such as day-ahead electricity load forecasting.

CRediT authorship contribution statement

Alessandro Brusaferri: Writing – review & editing, Writing – original draft, Supervision, Software, Methodology, Investigation, Conceptualization. **Andrea Ballarino:** Writing – review & editing, Visualization, Software, Funding acquisition, Data curation. **Luigi Grossi:** Writing – review & editing, Writing – original draft, Validation, Data curation. **Fabrizio Laurini:** Writing – review & editing, Writing – original draft, Validation, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Both the data and a structured codebase to fully reproduce the paper's experiments are available in an open GitHub repository (link also in the paper): https://github.com/bruale/PefCodeBench

References

[1] Nowotarski J, Weron R. Recent advances in electricity price forecasting: a review of probabilistic forecasting. Renew Sustain Energy Rev 2018;81:1548–68. https://doi.org/10.1016/j.rser.2017.05.234.

[2] Mitra S, Sun L, Grossmann IE. Optimal scheduling of industrial combined heat and power plants under time-sensitive electricity prices. Energy 2013;54:194–211. https://doi.org/10.1016/j.energy.2013.02.030.

[3] Ramin D, Spinelli S, Brusaferri A. Demand-side management via optimal production scheduling in power-intensive industries: the case of metal casting process. Appl Energy 2018;225:622–36. https://doi.org/10.1016/j.apenergy.2018.03.084.

[4] 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;293:116983. https://doi.org/10.1016/j.apenergy.2021.116983.

[5] Ciarreta A, Martinez B, Nasirov S. Forecasting electricity prices using bid data. Int J Forecast 2022. https://doi.org/10.1016/j.ijforecast.2022.05.011.

[6] Loutfi AA, Sun M, Loutfi I, Solibakke PB. Empirical study of day-ahead electricity spot-price forecasting: insights into a novel loss function for training neural networks. Appl Energy 2022;319:119182. https://doi.org/10.1016/j.apenergy.2022.119182.

[7] Wagner A, Ramentol E, Schirra F, Michaeli H. Short-and long-term forecasting of electricity prices using embedding of calendar information in neural networks. J Commod Mark 2022:100246. https://doi.org/10.1016/j.jcomm.2022.100246.

[8] Lehna M, Scheller F, Herwartz H. Forecasting day-ahead electricity prices: a comparison of time series and neural network models taking external regressors into account. Energy Econ 2022;106:105742. https://doi.org/10.1016/j.eneco.2021.105742.

[9] Sgarlato R, Ziel F. The role of weather predictions in electricity price forecasting beyond the day-ahead horizon. IEEE Trans Power Syst 2022:1. https://doi.org/10.1109/TPWRS.2022.3180119

[10] Madadkhani S, Ikonnikova S. Toward high-resolution projection of electricity prices: a machine learning approach to quantifying the effects of high fuel and co2 prices. Energy Econ 2024;129:107241. https://doi.org/10.1016/j.eneco.2023.107241.

[11] Attar M, Repo S, Mann P. Congestion management market design-approach for the nordics and central europe. Appl Energy 2022;313:118905. https://doi.org/10.1016/j.apenergy.2022.118905.

[12] Tschora L, Pierre E, Plantevit M, Robardet C. Electricity price forecasting on the day-ahead market using machine learning. Appl Energy 2022;313:118752. https://doi.org/10.1016/j.apenergy.2022.118752.

[13] Kezunovic M, Pinson P, Obradovic Z, Grijalva S, Hong T, Bessa R. Big data analytics for future electricity grids. Electr Power Syst Res 2020;189:106788. https://doi.org/10.1016/j.epsr.2020.106788.

[14] Weron R. Electricity price forecasting: a review of the state-of-the-art with a look into the future. Int J Forecast 2014;30(4):1030–81. https://doi.org/10.1016/j.ijforecast.2014.08.008.

[15] Lu H, Ma X, Ma M, Zhu S. Energy price prediction using data-driven models: a decade review. Comput Sci Rev 2021;39:100356. https://doi.org/10.1016/j.cosrev.2020.100356.

[16] Gianfreda A, Grossi L. Forecasting Italian electricity zonal prices with exogenous variables. Energy Econ 2012;34(6):2228–39. https://doi.org/10.1016/j.eneco.2012.06.024.

[17] Hong T, Pinson P, Wang Y, Weron R, Yang D, Zareipour H. Energy forecasting: a review and outlook. IEEE Open Access J Power Energy 2020;7:376–88. https://doi.org/10.1109/OAJPE.2020.3029979

[18] Lu X, Qiu J, Lei G, Zhu J. Scenarios modelling for forecasting day-ahead electricity prices: case studies in Australia. Appl Energy 2022;308:118296. https://doi.org/10.1016/j.apenergy.2021.118296.

[19] Fraunholz C, Kraft E, Keles D, Fichtner W. Advanced price forecasting in agent-based electricity market simulation. Appl Energy 2021;290:116688. https://doi.org/10.1016/j.apenergy.2021.116688.

[20] Keles D, Scelle J, Paraschiv F, Fichtner W. Extended forecast methods for day-ahead electricity spot prices applying artificial neural networks. Appl Energy 2016;162:218–30. https://doi.org/10.1016/j.apenergy.2015.09.087.

[21] Goodfellow IJ, Bengio Y, Courville A. Deep learning. Cambridge, MA, USA: MIT Press; 2016. http://www.deeplearningbook.org

[22] Lauriola I, Lavelli A, Aiolli F. An introduction to deep learning in natural language processing: models, techniques, and tools. Neurocomputing 2022;470:443–56. https://doi.org/10.1016/j.neucom.2021.05.103.

[23] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: transformers for image recognition at scale, In: International conference on learning representations; 2021.

[24] Mashlakov A, Kuronen T, Lensu L, Kaarna A, Honkapuro S. Assessing the performance of deep learning models for multivariate probabilistic energy forecasting. Appl Energy 2021;285:116405. https://doi.org/10.1016/j.apenergy.2020.116405.

[25] Brusaferri A, Matteucci M, Portolani P, Vitali A. Bayesian deep learning based method for probabilistic forecast of day-ahead electricity prices. Appl Energy 2019;250:1158–75. https://doi.org/10.1016/j.apenergy.2019.05.068.

[26] 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. https://doi.org/10.1016/j.apenergy.2018.02.069.

[27] Jedrzejewski A, Lago J, Marcjasz G, Weron R. Electricity price forecasting: the dawn of machine learning. IEEE Power Energy Mag 2022;20(3):24–31. https://doi.org/10.1109/MPE.2022.3150809

[28] Olivares KG, Challu C, Marcjasz G, Weron R, Dubrawski A. Neural basis expansion analysis with exogenous variables: forecasting electricity prices with N-BEATSx. Int J Forecast 2023;39(2):884–900. https://doi.org/10.1016/j.ijforecast.2022.03.001.

[29] Dumas J, Wehenkel A, Lanaspeze D, Cornélusse B, Sutera A. A deep generative model for probabilistic energy forecasting in power systems: normalizing flows. Appl Energy 2022;305:117871. https://doi.org/10.1016/j.apenergy.2021.117871.

[30] Grothe O, Kächele F, Krüger F. From point forecasts to multivariate probabilistic forecasts: the schaake shuffle for day-ahead electricity price forecasting. Energy Econ 2023;120:106602. https://doi.org/10.1016/j.eneco.2023.106602.

[31] Zhang C, Fu Y. Probabilistic electricity price forecast with optimal prediction interval. IEEE Trans Power Syst 2024;39(1):442–52. https://doi.org/10.1109/TPWRS.2023.3235193

[32] Cramer E, Witthaut D, Mitsos A, Dahmen M. Multivariate probabilistic forecasting of intraday electricity prices using normalizing flows. Appl Energy 2023;346:121370. https://doi.org/10.1016/j.apenergy.2023.121370.

[33] Marcjasz G, Narajewski M, Weron R, Nov F. Distributional neural networks for electricity price forecasting. Energy Econ 2023;125:106843. https://doi.org/10.1016/j.eneco.2023.106843.

[34] Kath C, Ziel F. Conformal prediction interval estimation and applications to day-ahead and intraday power markets. Int J Forecast 2021;37(2):777–99. https://doi.org/10.1016/j.ijforecast.2020.09.006.

[35] Hong T, Pinson P, Fan S, Zareipour H, Troccoli A, Hyndman RJ. Probabilistic energy forecasting: global energy forecasting competition 2014 and beyond. Int J Forecast 2016;32(3):896–913. https://doi.org/10.1016/j.ijforecast.2016.02.001.

[36] Brusaferri A, Matteucci M, Ramin D, Spinelli S, Vitali A. Probabilistic day-ahead energy price forecast by a mixture density recurrent neural network. In: 2020 7th international conference on control, decision and information technologies (CoDIT); vol. 1. 2020. p. 523–28. https://doi.org/10.1109/CoDIT49905.2020.9263898

[37] Zhou X, Wang J, Wang H, Lin J. Panel semiparametric quantile regression neural network for electricity consumption forecasting. Ecol Inf 2022;67:101489. https://doi.org/10.1016/j.ecoinf.2021.101489.

[38] Balasubramanian V, Ho S-S, Vovk V. Conformal prediction for reliable machine learning: theory, adaptations and applications. 1st ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 2014.

[39] Zaffran M, Feron O, Goude Y, Josse J, Dieuleveut A. Adaptive conformal predictions for time series. In: Chaudhuri K, Jegelka S, Song L, Szepesvari C, Niu G, Sabato S, editors. Proceedings of the 39th international conference on machine learning; vol. 162. PMLR; 2022. p. 25834–66 of Proceedings of Machine Learning Research. URL:https://proceedings.mlr.press/v162/zaffran22a.html

[40] Angelopoulos AN, Bates S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv:2107.07511 [Preprint]. 2022.

[41] Romano Y, Patterson E, Candes E. Conformalized quantile regression, In: Wallach H, Larochelle H, Beygelzimer A, d'Alché F, Fox E, Garnett R, editors. Advances in neural information processing systems; vol. 32. Curran Associates, Inc.; 2019. https://proceedings.neurips.cc/paper_files/paper/2019/file/5103c3584b063c431bd1268e9b5e76fb-Paper.pdf

[42] Angelopoulos AN, Candes EJ, Tibshirani RJ. Conformal pid control for time series prediction. arXiv:2307.16895 [Preprint]. 2023.

[43] Golia S, Grossi L, Pelagatti M. Machine learning models and intra-daily market information for the prediction of Italian electricity prices. Forecasting 2023;5(1):81–101. https://doi.org/10.3390/forecast5010003. https://www.mdpi.com/2571-9394/5/1/3

[44] Angelopoulos AN, Barber RF, Bates S. Theoretical foundations of conformal prediction. arXiv:2411.11824 [Preprint]. 2024. https://arxiv.org/abs/2411.11824

[45] Stankeviciute K, Alaa AM, van der Schaar M. Conformal time-series forecasting. In: Ranzato M; Beygelzimer A; Dauphin Y; Liang PS, Vaughan JW, editors. Advances in neural information processing systems; vol. 34. Curran Associates, Inc.; 2021. p. 6216–28. https://proceedings.neurips.cc/paper_files/paper/2021/file/312f1ba2a72318edaaa995a67835fad5-Paper.pdf
[46] Fontana M, Zeni G, Vantini S. Conformal prediction: a unified review of theory and new challenges. Bernoulli 2023;29(1):1–23. https://doi.org/10.3150/21-BEJ1447
[47] Xu C, Xie Y. Conformal prediction for time series. IEEE Trans Pattern Anal Mach Intell 2023;45(10):11575–87. https://doi.org/10.1109/TPAMI.2023.3272339
[48] Jensen V, Bianchi FM, Anfinsen SN. Ensemble conformalized quantile regression for probabilistic time series forecasting. IEEE Trans Neural Netw Learn Syst 2022:1–12. https://doi.org/10.1109/TNNLS.2022.3217694
[49] Gibbs I, Candes E. Adaptive conformal inference under distribution shift. In: Ranzato M, Beygelzimer A, Dauphin Y, Liang PS, Vaughan JW, editors. Advances in neural information processing systems; vol. 34. Curran Associates, Inc.; 2021. p. 1660–72. https://proceedings.neurips.cc/paper_files/paper/2021/file/0d441de75945e5acbc865406fc9a2559-Paper.pdf
[50] Fakoor R, Kim T, Mueller J, Smola AJ, Tibshirani RJ. Flexible model aggregation for quantile regression. arXiv:2103.00083 [Preprint]. 2023.
[51] Wang X, Hyndman RJ, Li F, Kang Y. Forecast combinations: an over 50-year review. Int J Forecast 2023;39(4):1518–47. https://doi.org/10.1016/j.ijforecast.2022.11.005.
[52] Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. In: Guyon I, Von Luxburg U, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, editors. Advances in neural information processing systems; vol. 30. Curran Associates, Inc.; 2017. https://proceedings.neurips.cc/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf
[53] Lichtendahl KC, Grushka-Cockayne Y, Winkler RL. Is it better to average probabilities or quantiles? Manag Sci 2013;59(7):1594–611. https://doi.org/10.1287/mnsc.1120.1667
[54] Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv:1412.6980 [Preprint]. 2017.
[55] Hyndman R, Athanasopoulos G. Forecasting: principles and practice. 3rd ed. Australia: OTexts; 2021.
[56] Diebold FX. Comparing predictive accuracy, twenty years later: a personal perspective on the use and abuse of diebold-mariano tests. J Bus Econ Stat 2015;33(1):1. https://doi.org/10.1080/07350015.2014.983236
[57] Brusaferri A, Matteucci M, Spinelli S, Vitali A. Probabilistic electric load forecasting through bayesian mixture density networks. Appl Energy 2022;309:118341. https://doi.org/10.1016/j.apenergy.2021.118341.
[58] Lundberg S, Lee S-I. A unified approach to interpreting model predictions. arXiv:1705.07874 [Preprint]. 2017.