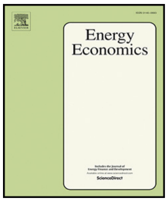
Contents lists available at ScienceDirect

Energy Economics

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



Postprocessing of point predictions for probabilistic forecasting of day-ahead electricity prices: The benefits of using isotonic distributional regression

Arkadiusz Lipiecki [a], Bartosz Uniejewski [b], Rafał Weron [b,*]

[a] Doctoral School, Faculty of Management, Wrocław University of Science and Technology, Wrocław, 50-370, Poland
[b] Department of Operations Research and Business Intelligence, Wrocław University of Science and Technology, Wrocław, 50-370, Poland

ARTICLE INFO

JEL classification:
C21
C22
C45
C53
Q47

Keywords:
Day-ahead electricity price
Quantile regression
Conformal prediction
Isotonic distributional regression
Combining probabilistic forecasts
Shapley values

ABSTRACT

Operational decisions relying on predictive distributions of electricity prices can result in significantly higher profits compared to those based solely on point forecasts. However, the majority of models developed in both academic and industrial settings provide only point predictions. To address this, we examine three postprocessing methods for converting point forecasts of day-ahead electricity prices into probabilistic ones: Quantile Regression Averaging, Conformal Prediction, and the recently introduced Isotonic Distributional Regression. We find that while the latter demonstrates the most varied behavior, it contributes the most to the ensemble of the three predictive distributions, as measured by Shapley values. Remarkably, the performance of the combination is superior to that of state-of-the-art Distributional Deep Neural Networks over two 4.5-year test periods from the German and Spanish power markets, spanning the COVID pandemic and the war in Ukraine.

1. Introduction

Recent studies demonstrate that operational decisions based on probabilistic price forecasts can lead to significantly (up to 20%) higher profits in day-ahead electricity trading than those relying solely on point predictions (Uniejewski and Weron, 2021; Marcjasz et al., 2023). However, constructing models that can yield probabilistic forecasts is a complex task. No wonder the majority of methods developed by both academics and practitioners provide only point predictions (Nowotarski and Weron, 2018; Ziel and Steinert, 2018). A workable solution is to use so-called *postprocessing* to convert point forecasts into probabilistic ones (Chen et al., 2024; Vannitsem et al., 2021), as such an approach can benefit from developments in the point forecasting literature (Liu et al., 2017).

In this study, we compare an established postprocessing method in energy forecasting – *Quantile Regression Averaging* (QRA; Liu et al., 2017; Wang et al., 2019; Kath and Ziel, 2021; Uniejewski and Weron, 2021; Nitka and Weron, 2023; Yang et al., 2023) – with *Conformal Prediction* (CP; Shafer and Vovk, 2008; Kath and Ziel, 2021), popular in the machine learning community, and the recently introduced *Isotonic Distributional Regression* (IDR; Henzi et al., 2021; Gneiting et al., 2023). Since we are not interested in developing point forecasting models for day-ahead markets, but rather in employing point predictions as inputs to postprocessing schemes, we use a variant of the well-performing *LASSO-Estimated AutoRegressive* (LEAR) model of Lago et al. (2021), and a simple similar-day 'naive' benchmark, commonly used as a reference point in *electricity price forecasting* (EPF; Weron, 2014). The obtained predictive distributions are compared to three probabilistic benchmarks built on point forecasts of the LEAR or the naive model and normally $N(0, \hat{\sigma})$ distributed errors, as well as state-of-the-art Distributional Deep Neural Networks (DDNNs; Marcjasz et al., 2023). Two major European electricity markets – Germany and Spain – serve as our testing ground.

The remainder of the paper is structured as follows. In Section 2 we present the datasets, then in Section 3 we explain how the point forecasts of day-ahead electricity prices are computed. Next, in Section 4 we describe the three postprocessing schemes: Quantile Regression Averaging, Conformal Prediction, and Isotonic Distributional Regression. In Section 5 we first briefly recall the Continuous Ranked Probability Score (CRPS), then discuss the obtained results in terms of the CRPS and the test for Conditional Predictive Ability (CPA) of Giacomini and White (2006). Next, we use Shapley values (Covert et al., 2020; Lundberg et al., 2020) to see which component contributes the most to the ensemble of the three predictive distributions. We conclude Section 5 by taking a risk management perspective and presenting

[*] Corresponding author.
E-mail address: rafal.weron@pwr.edu.pl (R. Weron).

https://doi.org/10.1016/j.eneco.2024.107934
Received 26 May 2024; Received in revised form 24 August 2024; Accepted 22 September 2024
Available online 5 October 2024

results for the tails of the predictive distribution. Finally, in Section 6 we summarize the main results.

2. Datasets

The data we use is publicly available and has been downloaded from ENTSO-E (https://transparency.entsoe.eu; day-ahead prices, day-ahead load forecasts, day-ahead onshore/offshore wind and solar generation forecasts) and Investing.com (https://www.investing.com/; carbon emission, natural gas, crude oil and coal closing prices). More precisely, the German day-ahead prices are for the BZN|DE-LU bidding zone (BZN|DE-AT-LU until 30.09.2018) and the Spanish day-ahead prices for the BZN|ES bidding zone. The day-ahead load forecasts and *renewable energy sources* (RES) generation forecasts are for the two countries — DE and ES, respectively. Since some of the ENTSO-E data has a 15-minute resolution, we have aggregated it to hourly values. The European Union Allowance (EUA) carbon emission prices, natural gas prices from the Title Transfer Facility (TTF) virtual trading point in the Netherlands, Brent crude oil prices, and API2 coal prices are the last known closing prices at the time of bidding in the day-ahead market.

Datasets for both markets span from 01.01.2015 to 31.12.2023; the 4.5-year out-of-sample test periods start on 27.06.2019, see Fig. 1. All time series were preprocessed to account for transitions to/from daylight saving time (DST). Missing values, which occur during the switch to DST, were replaced with the arithmetic average of the observations from the surrounding hours. Duplicate values, which occur during the switch back, were replaced by their arithmetic mean.

3. Computing point forecasts

3.1. The LEAR model

We use a variant of the *LASSO-Estimated AutoRegressive* (LEAR) model of Lago et al. (2021) to generate high quality point forecasts $\hat{p}_{d,h}$ of day-ahead electricity prices for day d and hour h. It is a parameter-rich autoregressive structure with exogenous variables estimated using the *Least Absolute Shrinkage and Selection Operator* (LASSO; Hastie et al., 2015). In the original formulation, the regressors include past prices $\boldsymbol{p}_{d-k} = \{p_{d-k,1}, \dots, p_{d-k,24}\}$ for lags $k = 1, 2, 3, 7$, day-ahead predictions $\boldsymbol{x}^{(i)}_{d-k} = \{x^{(i)}_{d-k,1}, \dots, x^{(i)}_{d-k,24}\}$ of two ($i = 1, 2$) fundamental variables for lags $k = 0, 1, 7$, and daily dummies to capture the weekly seasonality. The LEAR model we use has the form:

$$\begin{aligned} p_{d,h} = {} & \sum_{h=1}^{24} \beta_h p_{d-1,h} + \sum_{h=1}^{24} \beta_{h+24} p_{d-2,h} + \\ & + \sum_{h=1}^{24} \beta_{h+48} p_{d-3,h} + \sum_{h=1}^{24} \beta_{h+72} p_{d-7,h} \\ & + \sum_{h=1}^{24} \beta_{h+96} \hat{L}_{d,h} + \sum_{h=1}^{24} \beta_{h+120} \hat{L}_{d-1,h} + \sum_{h=1}^{24} \beta_{h+144} \hat{L}_{d-7,h} \\ & + \sum_{h=1}^{24} \beta_{h+168} \hat{R}_{d,h} + \sum_{h=1}^{24} \beta_{h+192} \hat{R}_{d-1,h} + \sum_{h=1}^{24} \beta_{h+216} \hat{R}_{d-7,h} \\ & + \beta_{241} \mathrm{EUA}_{d-2} + \beta_{242} \mathrm{NG}_{d-2} + \beta_{243} \mathrm{Brent}_{d-2} \\ & + \beta_{244} \mathrm{API2}_{d-2} + \sum_{i=1}^{7} \beta_{i+244} D_i + \varepsilon_{d,h}, \end{aligned} \tag{1}$$

since, following (Marcjasz et al., 2023), we:

- Use day-ahead predictions of the system-wide load $\hat{L}_{d-k,h}$ as $x^{(1)}_{d-k,1}$ and day-ahead RES (sum of onshore/offshore wind and solar) generation $\hat{R}_{d-k,h}$ as $x^{(2)}_{d-k,1}$;
- Additionally include four macroeconomic variables that have a major impact on European electricity prices: EUA_{d-2} carbon emission prices, NG_{d-2} natural gas prices, Brent_{d-2} crude oil prices, and $\mathrm{API2}_{d-2}$ coal prices; all four are the last known closing prices on day $d-2$.

Moreover, like Lago et al. (2021) and Ziel and Weron (2018) but unlike Marcjasz et al. (2023), we preprocess the electricity prices with the *area hyperbolic sine* variance stabilizing transformation:

$$\operatorname{asinh}(x) = \log\left(x + \sqrt{x^2 + 1}\right), \tag{2}$$

where x is the price standardized by subtracting the in-sample median and dividing by the median absolute deviation (MAD), adjusted by the 75% quantile of the standard normal distribution for asymptotical consistency with the standard deviation (Uniejewski et al., 2018). To recover price forecasts we apply the inverse transformation, i.e., the hyperbolic sine, to the generated predictions; see Narajewski and Ziel (2020) for a more accurate back-transformation.

Finally, unlike Lago et al. (2021) and Marcjasz et al. (2023), instead of using the faster but less accurate Least Angle Regression (LARS; Efron et al., 2004), we use the standard coordinate descent LASSO estimator (as implemented in Matlab 2024a; see Friedman et al., 2010) to estimate the model coefficients. We combine the latter with 7-fold cross-validation (CV), like (Marcjasz et al., 2023) but unlike (Lago et al., 2021) who used the Akaike Information Criterion (AIC) for initial estimation and coordinate descent for the final run. Since in our setup CV involves a random split of the training data, resulting in a slightly different forecast for each run, we compute the predictions of the LEAR model not once, but 5 times for each training window length and average the 5 individual results to obtain the final LEAR forecast. As we will see in Table 2, these changes, compared to the variant used by Marcjasz et al. (2023), result in more accurate point forecasts, leading to significantly better predictive distributions.

We consider a rolling window setup, where forecasts of all 24 h on day d are calculated in the morning of day $d-1$ and the model parameters are reestimated each day using a calibration sample of D most recent past observations. As in the original LEAR formulation, the parameters are estimated separately for each of the 4 training window lengths $D = 56, 84, 1092$ and 1456, yielding point forecasts $\hat{p}^{56}_{d,h}$, $\hat{p}^{84}_{d,h}$, $\hat{p}^{1092}_{d,h}$ and $\hat{p}^{1456}_{d,h}$ for each day and hour in the test period.

3.2. The naive benchmark

As a reference point, we use a popular in EPF implementation of the similar-day approach, often called the *naive* method (Lago et al., 2021; Weron, 2014; Ziel and Weron, 2018). It uses last week's prices to forecast the prices on Monday, Saturday and Sunday, and yesterday's prices for the remaining days:

$$\hat{p}_{d,h} = \begin{cases} p_{d-7,h}, & \text{for } d = \text{Mon, Sat or Sun}, \\ p_{d-1,h}, & \text{otherwise}. \end{cases} \tag{3}$$

4. Postprocessing point forecasts

Our goal is to obtain 99 percentiles of the predictive distribution $\hat{F}_p$ of $p_{d,h}$. Like for the LEAR model, we construct separate distributional models for each hour h and retrain them daily, using 4 different calibration windows with the most recent data: $\{(\hat{p}_{t,h}, p_{t,h})\}_{t=d-m}^{d-1}$ with $m \in \{28, 56, 91, 182\}$. Note that we use the term 'training/calibration window' to refer to the data used to estimate the point/probabilistic forecasting model. For postprocessing we use the open source Julia package *PostForecasts.jl* (Lipiecki and Weron, 2024).

Before fitting distributional models, we must first generate point forecasts for 182 days, and thus our approach requires a total of 1456 (longest training window) + 182 (longest calibration window) = 1638 days to generate the first predictive distribution. The final probabilistic forecasts are obtained via probability (or 'vertical'; Lichtendahl et al., 2013) averaging of 4 distributions obtained for different calibration window lengths m; except for the Naive-1N model which uses only the 182-day window, see Section 4.5 for details. Let us now briefly describe the postprocessing schemes.

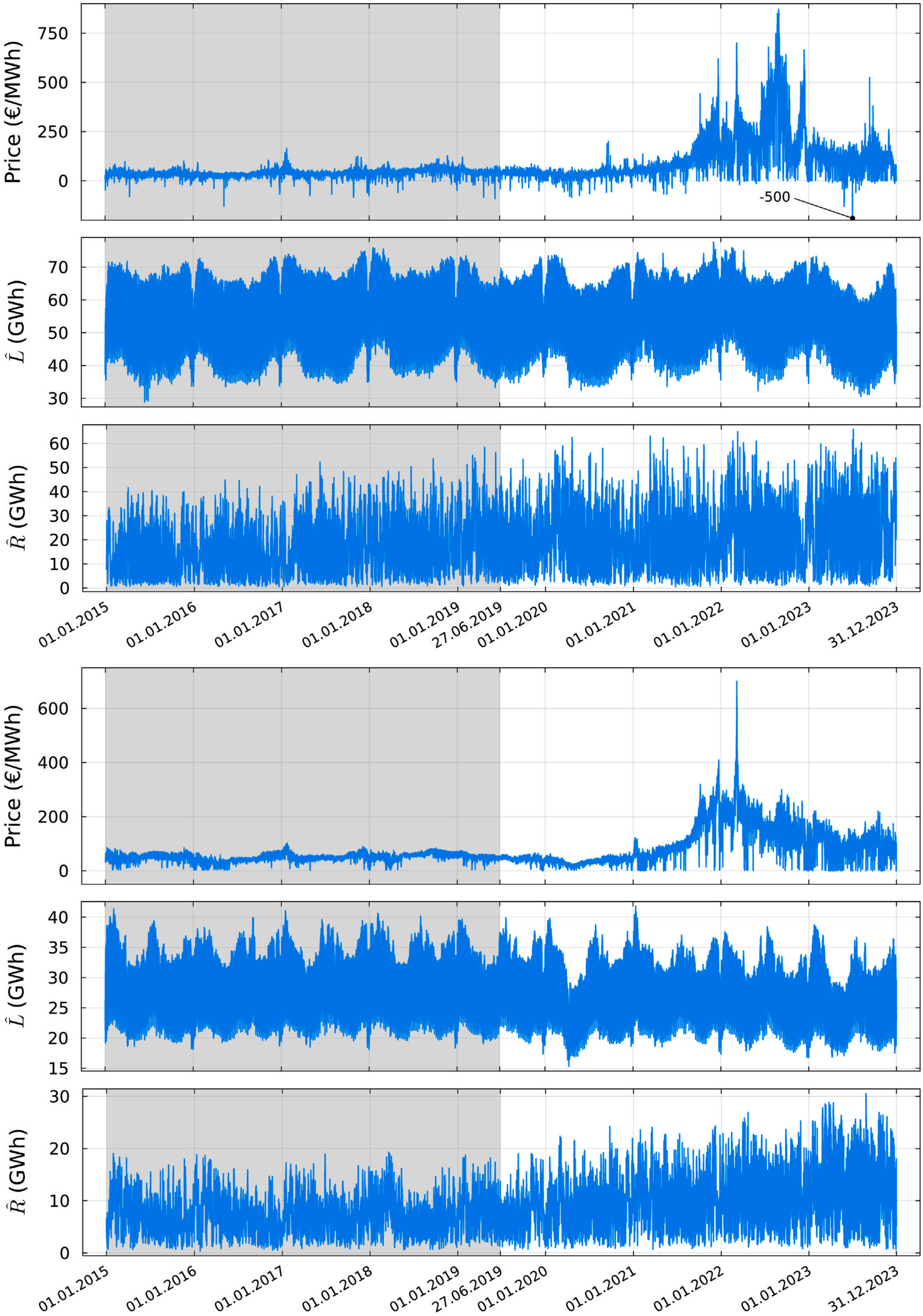


Fig. 1. *From top to bottom*: Day-ahead electricity prices $p_{d,h}$ and day-ahead predictions of load $\hat{L}_{d,h}$ and renewable generation $\hat{R}_{d,h}$ (onshore/offshore wind and solar) in Germany (*top three panels*) and Spain (*bottom three panels*). Gray background marks the initial calibration window (1.01.2015–26.06.2019), while white corresponds to the 4.5-year test period (27.06.2019–31.12.2023). Note that in Germany the prices can be negative and on 02.07.2023 the day-ahead price dropped to the minimum admissible level of −500 EUR/MWh.

4.1. Quantile Regression Averaging (QRA)

Formally introduced by Nowotarski and Weron (2015), and successfully used in the GEFCom2014 competition (Maciejowska and Nowotarski, 2016; Gaillard et al., 2016) and later energy forecasting applications (Liu et al., 2017; Wang et al., 2019; Kath and Ziel, 2021; Uniejewski and Weron, 2021; Nitka and Weron, 2023; Yang et al., 2023; Cornell et al., 2024), the method estimates conditional quantiles of the target variable as a linear combination of point predictions in a quantile regression setting:

$$\hat{q}(\alpha|\hat{\boldsymbol{p}}_{d,h}) = [1, \hat{\boldsymbol{p}}_{d,h}]\ \boldsymbol{\beta}_\alpha, \tag{4}$$

where $\hat{q}(\alpha|\cdot)$ is the conditional αth quantile, $\hat{\boldsymbol{p}}_{d,h}$ is the row vector of point predictions (see the next paragraph for details), and $\boldsymbol{\beta}_\alpha$ is the column vector of coefficients. Prediction intervals (PIs) are obtained by running QRA for two selected quantiles, e.g., the 5% and 95% quantiles yield the 90% PI. The coefficients are computed by minimizing the pinball score, so to obtain $\hat{F}_p$, a linear optimization problem must be solved independently for each quantile (Nowotarski and Weron, 2018). This makes QRA by far the most computationally intensive method we consider, yet still feasible on a consumer-grade laptop.

Given the pool of four point forecasts $\hat{p}^{56}_{d,h}, \hat{p}^{84}_{d,h}, \hat{p}^{1092}_{d,h}, \hat{p}^{1456}_{d,h}$, we initially examined three approaches:

(i) using a single model with all individual point forecasts as regressors $\hat{\boldsymbol{p}}_{d,h} = [\hat{p}^{56}_{d,h}, \hat{p}^{84}_{d,h}, \hat{p}^{1092}_{d,h}, \hat{p}^{1456}_{d,h}]$, like in the original formulation of QRA (Nowotarski and Weron, 2015);

(ii) using a single model with one regressor being the average point forecast $\hat{\boldsymbol{p}}_{d,h} = \hat{p}^{ave}_{d,h} = \frac{1}{4}(\hat{p}^{56}_{d,h} + \cdots + \hat{p}^{1456}_{d,h})$, a variant dubbed Quantile Regression Machine (QRM) in Marcjasz et al. (2020); and

(iii) averaging (over quantiles or probabilities) the predictive distributions $\hat{F}^{56}_p, \dots, \hat{F}^{1456}_p$ obtained from the individual point forecasts $\hat{p}^{56}_{d,h}, \dots, \hat{p}^{1456}_{d,h}$, respectively.

Below we present the results for approach (ii), which turned out to be the fastest and the most accurate. We call it the **LEAR-QRM** model.

4.2. Conformal Prediction (CP)

This is a framework for computing PIs based on absolute point prediction errors in a chosen calibration window (Shafer and Vovk, 2008). CP produces valid intervals for a given confidence level and requires no distributional assumptions (Zaffran et al., 2022). However, the estimated PIs are centered on the point forecast, so obtaining quantile forecasts from CP is only possible under the assumption of symmetrically distributed errors. The αth quantile is given by:

$$\hat{q}(\alpha|\hat{p}_{d,h}) = \begin{cases} \hat{p}_{d,h} - \lambda^{2\alpha} & \text{if} \quad \alpha < 1/2, \\ \hat{p}_{d,h} + \lambda^{2(1-\alpha)} & \text{otherwise,} \end{cases} \tag{5}$$

where λ^α is the so-called *nonconformity score* such that $[\hat{p}_{d,h} - \lambda^\alpha, \hat{p}_{d,h} + \lambda^\alpha]$ is a $(1-\alpha)$ PI. We use the same inductive scheme as (Kath and Ziel, 2021), but take a different approach to selecting training and calibration sets. Although the rolling windows we use are not disjoint, each point prediction error calculated in a calibration window does not belong to the training window used for generating that particular prediction. Similar to QRA, we present the results for CP computed for $\hat{p}_{d,h} = \hat{p}^{ave}_{d,h} = \frac{1}{4}(\hat{p}^{56}_{d,h} + \cdots + \hat{p}^{1456}_{d,h})$ in Eq. (5); we call it the **LEAR-CP** model. This approach outperformed combining the predictive distributions $\hat{F}^{56}_p, \dots, \hat{F}^{1456}_p$.

4.3. Isotonic Distributional Regression (IDR)

This is a nonparametric method for learning conditional distributions under the stochastic order constraint (Henzi et al., 2021; Walz et al., 2024). The output $\hat{F}_p$ minimizes the Continuous Ranked Probability Score (CRPS; see Section 5.1) under the isotonic constraint, which requires that the conditional cumulative distribution function (CDF) of the response be non-increasing (or equivalently, that the quantiles of the response be non-decreasing) with respect to the regressor. This makes postprocessing point forecasts a natural setting for IDR, since a forecast of the response variable generally satisfies the isotonic relation. For a calibration window of m data points $(p_{i,h}, \hat{p}_{i,h})_{i=d-m,\dots,d-1}$, renumbered $(p_{i,h}, \hat{p}_{i,h})_{i=1,\dots,m}$ so that $\hat{p}_{1,h} \le \cdots \le \hat{p}_{m,h}$, IDR estimates m conditional distributions $\hat{F}_i(z) \equiv \hat{F}(z|\hat{p}_{i,h})$:

$$\left(\hat{F}_1(z), \dots, \hat{F}_m(z)\right) = \operatorname*{argmin}_{(\eta_1,\dots,\eta_m)} \sum_{i=1}^{m} \left(\eta_i - \mathbb{1}_{\{p_{i,h} \le z\}}\right)^2, \tag{6}$$

with $\eta_1 \ge \cdots \ge \eta_m$ and $\eta_i \in [0,1]$. To obtain the conditional distribution for any $\hat{p}_{d,h} \in \mathbb{R}$, we adopt the interpolation method suggested in Henzi et al. (2021):

$$\hat{F}_d(z) = \frac{\hat{p}_{d,h} - \hat{p}_{i,h}}{\hat{p}_{i+1,h} - \hat{p}_{i,h}} \hat{F}_i(z) + \frac{\hat{p}_{i+1,h} - \hat{p}_{d,h}}{\hat{p}_{i+1,h} - \hat{p}_{i,h}} \hat{F}_{i+1}(z), \tag{7}$$

for any $i \in \{1, \dots, m-1\}$ such that $\hat{p}_{d,h} \in [\hat{p}_{i,h}, \hat{p}_{i+1}]$; if $\hat{p}_{d,h} < \hat{p}_{1,h}$ or $\hat{p}_{d,h} > \hat{p}_{m,h}$, we set $\hat{F}_d(z) = \hat{F}_1(z)$ or $\hat{F}_d(z) = \hat{F}_m(z)$, respectively.

To solve Eq. (6), we use the abridged pool-adjacent violators algorithm (Henzi et al., 2022). This is illustrated in Fig. 2 for a sample calibration set $(p_{i,h}, \hat{p}_{i,h})_{i=1,\dots,4}$ of $m = 4$ days for hour h. The four price forecasts are sorted to satisfy: $\hat{p}_{1,h} \le \cdots \le \hat{p}_{4,h}$. But then the respective prices $p_{1,h}, \dots, p_{4,h}$ are not, since $p_{2,h} > p_{3,h}$. This requires pooling together the two observations which violate the isotonic constraint, see the orange ellipse in panel (a), with $\frac{1}{2}$ probability mass assigned to $p_{2,h}$ and $\frac{1}{2}$ to $p_{3,h}$. The resulting conditional CDFs are plotted in panel (b); note that $\hat{F}_i(z) \equiv \hat{F}(z|\hat{p}_{i,h})$ are defined only for $z \in \{p_{1,h}, \dots, p_{4,h}\}$. Clearly, $\hat{F}_2(z)$ and $\hat{F}_3(z)$ overlap due to the pooling. On the other hand, as shown in panel (c), $\hat{F}(p_{2,h}|\hat{p})$ and $\hat{F}(p_{3,h}|\hat{p})$ as a function of $\hat{p}$ do not overlap. In this panel, the values for $\hat{p}_{d,h} \notin \{\hat{p}_{1,h}, \dots, \hat{p}_{4,h}\}$ are interpolated using Eq. (7). Finally, the predictive distribution $\hat{F}_p$ for the next day's forecast $\hat{p}_{5,h}$ is obtained from the intersections of $\hat{F}(p_{i,h}|\hat{p})$ and a vertical line at $\hat{p} = \hat{p}_{5,h}$. In panel (c) we use dashed lines to indicate three hypothetical next day's forecasts: $\hat{p}_{5',h}$, $\hat{p}_{5'',h}$ and $\hat{p}_{5''',h}$. The corresponding predictive distributions are plotted in panel (d).

We separately solve Eq. (6) for each of the four point prediction models to obtain $\hat{F}^{56}_p, \hat{F}^{84}_p, \hat{F}^{1092}_p$ and $\hat{F}^{1456}_p$. Then, we take a simple 'vertical' average to obtain the **LEAR-IDR** model:

$$\hat{F}_p(z) = \frac{1}{4}\left(\hat{F}^{56}_p(z) + \hat{F}^{84}_p(z) + \hat{F}^{1092}_p(z) + \hat{F}^{1456}_p(z)\right). \tag{8}$$

We can do this since the distributions are trained on the same set of prices $p_{1,h}, \dots, p_{m,h}$, see panel (d) in Fig. 2. We also tested a variant of IDR that used an average point forecast as a regressor, but it performed significantly worse.

4.4. Ensemble of predictive distributions

Since combining forecasts typically improves the results (Olivares et al., 2023; Nitka and Weron, 2023), not only in electricity price forecasting (Baran and Lerch, 2018; Grushka-Cockayne and Jose, 2020), we consider an ensemble of the LEAR-QRM, LEAR-CP and LEAR-IDR predictive distributions and call it **LEAR-Ave**. We compute it as an average over probabilities ('vertical'; Marcjasz et al., 2020) of the three distributions. We also tested combinations of any two predictive distributions, but their performance was worse than of the LEAR-Ave.

4.5. N(0, $\hat{\sigma}$)-based benchmarks

A common assumption underlying time series models is that the innovations are Gaussian. Under this assumption, probabilistic forecasts can be obtained by computing the standard deviation $\hat{\sigma}$ of the prediction errors $\epsilon_{d,h} = p_{d,h} - \hat{p}_{d,h}$ in the calibration sample, then taking appropriate quantiles of the N(0, $\hat{\sigma}$) distribution and adding them to the point forecast $\hat{p}_{d,h}$ for the target day and hour (Nowotarski and Weron, 2018). We use this approach to construct three benchmark models:

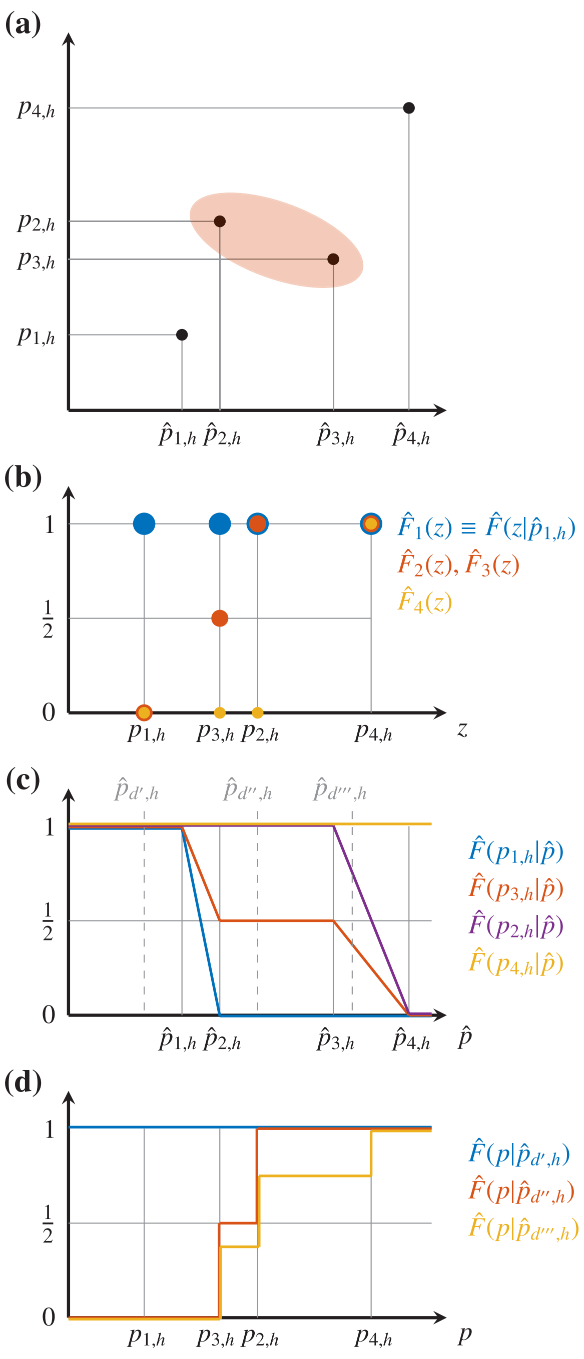

(a)
$p_{4,h}$
$p_{2,h}$
$p_{3,h}$
$p_{1,h}$
$\hat{p}_{1,h}$ $\hat{p}_{2,h}$ $\hat{p}_{3,h}$ $\hat{p}_{4,h}$
(b)
1
$\frac{1}{2}$
0
$\hat{F}_1(z) \equiv \hat{F}(z|\hat{p}_{1,h})$
$\hat{F}_2(z), \hat{F}_3(z)$
$\hat{F}_4(z)$
$p_{1,h}$ $p_{3,h}$ $p_{2,h}$ $p_{4,h}$ z
(c)
$\hat{p}_{d',h}$ $\hat{p}_{d'',h}$ $\hat{p}_{d''',h}$
1
$\frac{1}{2}$
0
$\hat{F}(p_{1,h}|\hat{p})$
$\hat{F}(p_{3,h}|\hat{p})$
$\hat{F}(p_{2,h}|\hat{p})$
$\hat{F}(p_{4,h}|\hat{p})$
$\hat{p}_{1,h}$ $\hat{p}_{2,h}$ $\hat{p}_{3,h}$ $\hat{p}_{4,h}$ $\hat{p}$
(d)
1
$\frac{1}{2}$
0
$\hat{F}(p|\hat{p}_{d',h})$
$\hat{F}(p|\hat{p}_{d'',h})$
$\hat{F}(p|\hat{p}_{d''',h})$
$p_{1,h}$ $p_{3,h}$ $p_{2,h}$ $p_{4,h}$ p

Fig. 2. Schematic representation of the IDR algorithm. Panel (a) shows the calibration set of $m = 4$ days for hour h, with price predictions $\hat{p}_{1,h} \leq \hat{p}_{2,h} \leq \hat{p}_{3,h} \leq \hat{p}_{4,h}$ and respective prices $p_{1,h}, \dots, p_{4,h}$. Panel (b) displays the conditional CDFs after pooling together the two observations which violate the isotonic constraint (orange ellipse). Panel (c) shows $\hat{F}(z|\hat{p})$ as a function of $\hat{p}$, interpolated using Eq. (7). The predictive distribution for the next day is obtained from the intersections of $\hat{F}(p_{i,h}|\hat{p})$ and a vertical line at the next day's point forecast. In panel (d) we depict $\hat{F}_p$ corresponding to three hypothetical next day's forecasts: $\hat{p}_{d',h}, \hat{p}_{d'',h}$ and $\hat{p}_{d''',h}$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

- **Naive-1N** which uses point forecasts of the naive model defined in Eq. (3) and estimates $\hat{\sigma}$ on one calibration window of $m = 182$ days;

Table 1
Computational time required for each model/component to generate forecasts for a 4.5 year test period on a server equipped with an AMD EPYC 7713 64-core processor and 256 GB of RAM. For comparison, hyperparameter optimization for DDNN-JSU would take 2–3 weeks, as estimated for 16 times smaller hyperparameter sets (128 instead of 2048 elements).

Model/component	Time
LEAR (4 training windows, 5 runs, Matlab 2024a)	2:30–3:00 h
QRA (4 calibration windows, Julia 1.10)	10–15 min
CP (4 calibration windows, Julia 1.10)	10–15 s
IDR (4 calibration windows, Julia 1.10)	15–20 s
LEAR-Ave (all components, Julia 1.10)	2:45–3:15 h
DDNN-JSU (4 networks, TensorFlow 2.9)	6:00–6:30 h

- **Naive-N** which uses point forecasts of the naive model defined in Eq. (3) and estimates $\hat{\sigma}$ on calibration windows of $m \in \{28, 56, 91, 182\}$ days;
- **LEAR-N** which uses point forecasts of the LEAR model defined in Section 3.1 and estimates $\hat{\sigma}$ on calibration windows of $m \in \{28, 56, 91, 182\}$ days.

4.6. The DDNN-JSU benchmark

The fourth benchmark is the DDNN-JSU-pEns model of Marcjasz et al. (2023); we refer to it as **DDNN-JSU**. It is based on a Distributional Deep Neural Network architecture that outputs four parameters of Johnson's SU distribution. To estimate the DDNN-JSU model and obtain day-ahead price forecasts, we use the Python codes available on GitHub: https://github.com/gmarcjasz/distributionalnn.

For the whole 4.5-year German and Spanish out-of-sample test sets, we use the hyperparameter set optimized by Marcjasz et al. (2023) for German data over the period 1.01.2015–31.12.2018; the files are available on GitHub. The rationale for this approach is provided by Marcjasz (2020), who showed that hyperparameters optimized for one electricity market can be effectively used in another one. Optimizing the hyperparameter set more frequently and for both markets could lead to better predictions, but is extremely time consuming. A single hyperparameter optimization run takes weeks even on multi-core computing servers, see Table 1 and the DDNN documentation on GitHub.

5. Empirical results

5.1. Comparison in terms of the CRPS

The Continuous Ranked Probability Score (CRPS; Gneiting and Raftery, 2007) is a proper scoring rule and the standard metric for evaluating probabilistic forecasts (Billé et al., 2023; Marcjasz et al., 2023; Nowotarski and Weron, 2018). It is defined as:

$$\mathrm{CRPS}(\hat{F}, x) = \int_{-\infty}^{\infty} \left(\hat{F}(y) - \mathbb{1}_{\{x \leq y\}} \right)^2 dy, \tag{9}$$

where $\hat{F}$ is the predictive distribution and x is the observation, e.g., electricity price $p_{d,h}$. It can be approximated by[1]:

$$\mathrm{CRPS}(\hat{F}, x) \approx \frac{2}{M} \sum_{i=1}^{M} \mathrm{PS}\left(\hat{q}, x, q_i\right), \tag{10}$$

where $\left(q_1, \dots, q_M\right)$ is an equidistant monotonically increasing dense grid of probabilities, e.g., the 99 percentiles, $\hat{q} \equiv \hat{F}^{-1}(q)$ is the quantile forecast for quantile level $q \in (0, 1)$, and

$$\mathrm{PS}(\hat{q}, x, q) = \left(\mathbb{1}_{\{x<\hat{q}\}} - q \right) (\hat{q} - x) \tag{11}$$

[1] Note that the scaling factor of 2 in Eq. (10) is usually omitted in practice (Nitka and Weron, 2023). This is also the case here.

Table 2
Continuous Ranked Probability Scores (CRPS; i.e., Aggregate Pinball Score across all 99 percentiles, compare with Table 3) for the considered models and markets. Cells are colored independently for each row and market. The test period labeled '2000†' spans from 27.06.2019 to 31.12.2020 (554 days), the remaining three span full years (365 days). Note, that Marcjasz et al. (2023) reported a CRPS of 1.662 for the LEAR-QRM model and 1.304 for the DDNN-JSU model in the first 554-day test period for Germany; see text for details and discussion.

Model	2020†	2021	2022	2023
		Germany		
Naive-1N	3.548	9.494	25.346	12.078
Naive-N	3.488	9.322	25.064	11.464
LEAR-N	1.408	4.370	10.878	4.641
LEAR-QRM	1.350	4.189	10.651	4.422
LEAR-CP	1.369	4.399	10.864	4.582
LEAR-IDR	1.422	4.389	10.926	4.336
LEAR-Ave	1.310	3.970	10.199	4.215
DDNN-JSU	1.342	5.395	13.375	5.265
		Spain		
Naive-1N	2.110	7.373	12.553	8.999
Naive-N	2.065	7.201	12.287	8.806
LEAR-N	1.018	4.166	7.412	4.735
LEAR-QRM	0.976	4.034	7.136	4.723
LEAR-CP	1.014	4.208	7.371	4.699
LEAR-IDR	0.986	4.179	7.268	4.361
LEAR-Ave	0.938	3.832	6.983	4.369
DDNN-JSU	0.989	4.627	8.299	4.379

is the so-called *pinball score*, also known as the *pinball loss*, *quantile loss* or *check function* (Berrisch and Ziel, 2023; Grushka-Cockayne et al., 2017; Maciejowska et al., 2023).

In Table 2 we report the CRPS for the considered models and markets. As can be seen, the LEAR-Ave ensemble yields the lowest CRPS across both markets and all four test subperiods, while the Naive-1N and Naive-N benchmarks are the worst. Of the latter two, Naive-N significantly outperforms Naive-1N at the 5% level for all subperiods and both markets, as measured by the CPA test of Giacomini and White (2006), see Section 5.2. This underscores the importance of estimating $\hat{\sigma}$ on calibration windows of different lengths, see Section 4.5.

The LEAR-N benchmark is much more competitive than Naive-based benchmarks due to much more accurate point forecasts. Interestingly, in some subperiods it even outperforms some of the other LEAR-based competitors. On the other hand, the DDNN-JSU model is a disappointment. It is much worse than all LEAR-based models during the energy crisis and the initial phase of the war in Ukraine (2021–2022), and performs well only in the first subperiod labeled '2000†' in Germany and in last year in Spain. In the latter case, the LEAR-IDR model is the best performer and the second best in 2023 in Germany. It seems that LEAR-IDR excels in (relatively) calm periods that follow more volatile ones, but is not an all-rounder like LEAR-QRM.

5.1.1. Temporal performance

In Fig. 3 we plot the rolling 182-day CRPS-based Skill Score (SS; see Rasp and Lerch, 2018) with respect to the LEAR-N model, i.e., the LEAR model with normally $\mathrm{N}(0, \hat{\sigma})$ distributed errors:

$$\mathrm{SS}_d^{model} = 1 - \frac{\sum_{k=0}^{181} \sum_{h=1}^{24} \mathrm{CRPS}_{d-k,h}^{model}}{\sum_{k=0}^{181} \sum_{h=1}^{24} \mathrm{CRPS}_{d-k,h}^{\text{LEAR-N}}}, \tag{12}$$

where $d = 25.12.2019, \ldots, 31.12.2023$ and $\mathrm{CRPS}_{d,h}^{model}$ is the CRPS of $model$ for day d and hour h.

Among the three postprocessing schemes, IDR shows the most uneven performance. In Germany relatively poor for the 182-day windows ending between Dec 2019 and Apr 2021, between Dec 2021 and Apr 2022, and between Aug 2022 and Apr 2023, while relatively good for the remaining periods. In Spain relatively poor for the 182-day windows ending between Nov 2020 and Apr 2021, and between Oct and Dec 2022, while relatively good for the remaining periods, especially after May 2023. For windows that span periods of moderately increasing prices after calm periods (e.g., May-Sep 2021 in Germany and Spain) or normal prices after a spiky period (e.g., Sep-Dec 2023 in Germany and May-Dec 2023 in Spain), the IDR average significantly outperforms the QRA and CP averages. The CP scheme gives a relatively stable performance, both for the individual calibration windows and the average, while the QRA approach shows an intermediate behavior. Analyzing the four distributions $\hat{F}_p^{56}$, $\hat{F}_p^{84}$, $\hat{F}_p^{1092}$ and $\hat{F}_p^{1456}$ corresponding to the individual calibration windows of $m = 28, 56, 91$ and 182 days, the IDR-generated ones are the most volatile and different from each other, while the CP-generated ones are the least; not depicted in Fig. 3.

5.1.2. Shapley values and component contribution to the ensemble

We use Shapley values to assess which component contributes the most to the ensemble of the three predictive distributions in the LEAR-Ave model. Recall, that Shapley values were originally developed to fairly distribute total wins (→ predictive power) among players (→ ensemble components) in a cooperative game based on their individual contributions. Our approach is similar in spirit to *Loss SHapley Additive exPlanations* (LossSHAP; Lundberg et al., 2020) and *Shapley Additive Global importancE* (SAGE; Covert et al., 2020), which aim to explain the contribution of features to the model's performance measured by a given loss function.

In Fig. 4 we plot Shapley values based on the CRPS loss – see Lipiecki and Weron (2024) for details – for the whole 4.5-year test sets in Germany and Spain. Clearly, across both markets, IDR contributes the most to the LEAR-Ave ensemble, while CP the least. When analyzed independently for the four subperiods ('2000†', 2021, 2022 and 2023), the contribution of IDR is by far the highest in all but the first subperiod in Germany, when all three postprocessing schemes contribute approximately equally. In 2023 the contribution of IDR exceeds 75% in both markets.

5.1.3. Comparison with the results of Marcjasz et al. (2023)

The test period labeled '2000†' spans from 27.06.2019 to 31.12.2020 (554 days) and is the same as considered for Germany by Marcjasz et al. (2023). Interestingly, the latter article reported a CRPS of 1.662 for the LEAR-QRM model and 1.304 for the DDNN-JSU model for Germany. The much better performance of the LEAR-QRM model in our study (CRPS = 1.350) is a result of the improvements discussed in Section 3.1: the use of the asinh variance stabilizing transformation (Uniejewski et al., 2018) and the more time consuming, but more accurate coordinate descent LASSO estimator (Friedman et al., 2010) combined with 7-fold cross-validation.

The differences in the results of the DDNN-JSU model – a CRPS of 1.304 vs. 1.342 in our study – are harder to explain. We use exactly the same set of hyperparameters and the same Python code to estimate the weights of the neural network and make the predictions. The difference in the CRPS of $1.342 - 1.304 = 0.038$ cannot only be attributed to the stochastic nature of the estimation process; a limited simulation study suggests that this randomness could be responsible for a ±0.01 discrepancy, but not more. The answer is surprising and lies in the dataset. Marcjasz et al. (2023) used load and RES generation forecasts that are not consistent with those currently available on ENTSO-E for the years 2015–2017, and which we use in this study. In particular, the series of RES forecasts exhibit differences of considerable magnitude in both directions — for individual hours, a median deviation of 372 MWh or ca. ±3% with respect to the median RES level in the years 2015–2017, and a maximum deviation of 6,388 MWh or ca. ±45%!

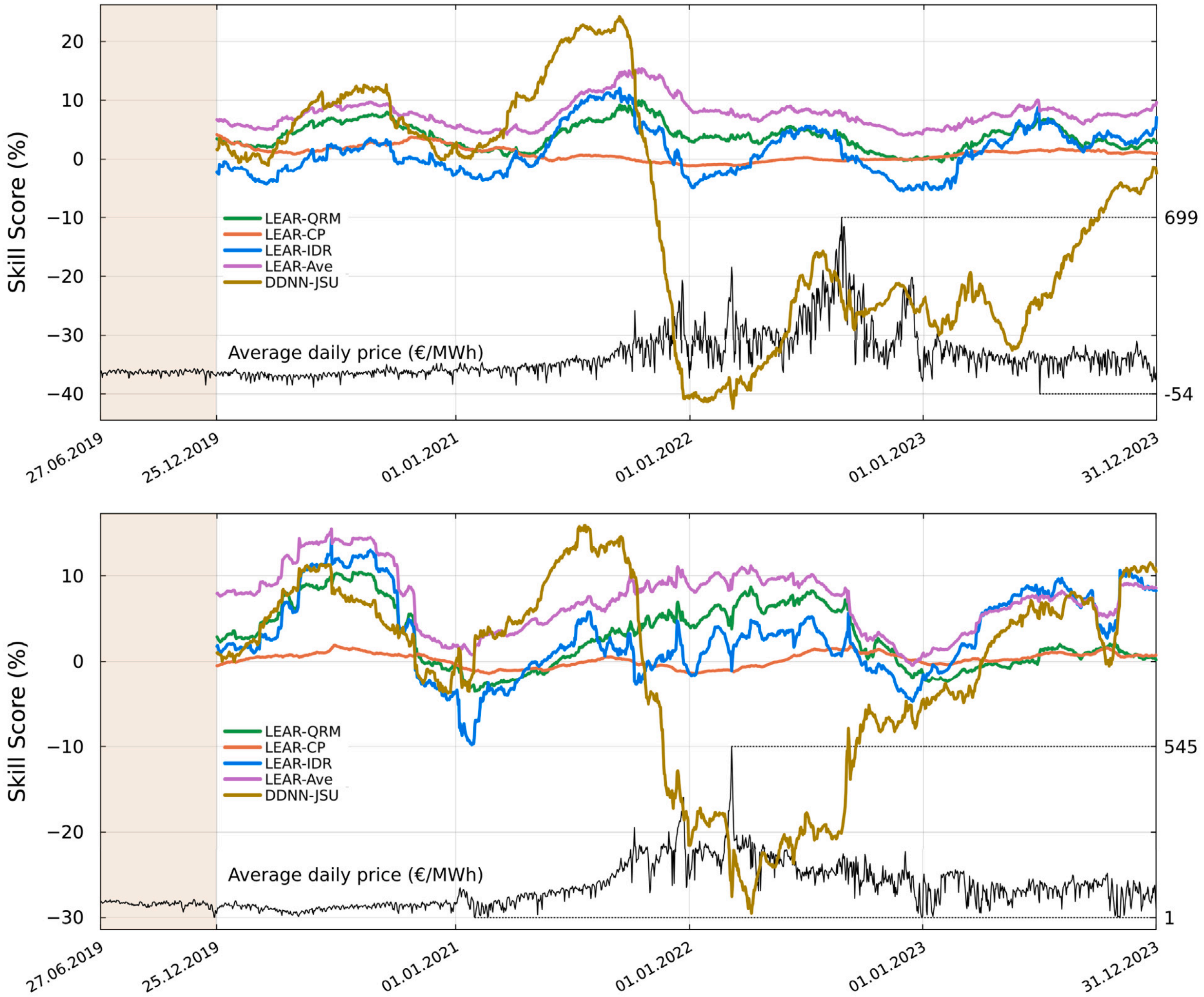


Fig. 3. Rolling 182-day Skill Scores, see Eq. (12), for Germany (*top*) and Spain (*bottom*) with respect to the LEAR-N model. Values for 25.12.2019 are the scores for 27.06–25.12.2019, i.e., the beige shaded area, values for 26.12.2019 are the scores for 28.06–26.12.2019, etc. The black curves in both panels are the mean daily electricity prices $p_d = \frac{1}{24}\sum_h p_{d,h}$ in the depicted period; the maximum and minimum values (in EUR/MWh) are shown on the right axis. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

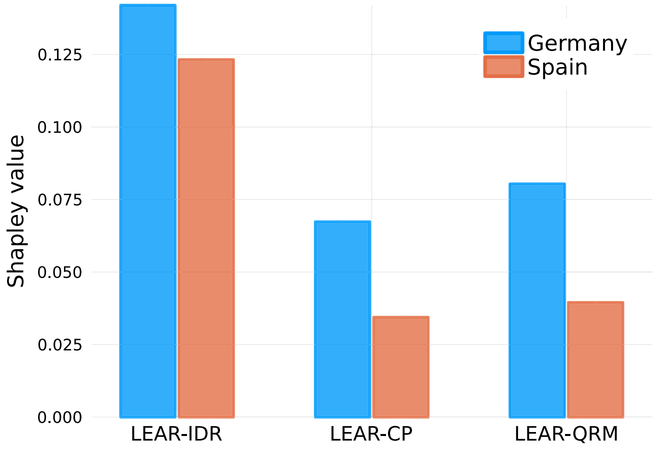


Fig. 4. Shapley values for the whole 4.5-year test sets in Germany and Spain. Across both markets, IDR contributes the most to the LEAR-Ave ensemble.

5.2. Conditional predictive ability

Following Lago et al. (2021) and Olivares et al. (2023), we run the test of *conditional predictive ability* (CPA; Giacomini and White, 2006) to formally assess the performance of the different models. Namely, we test the null $H_0 : \boldsymbol{\phi} = 0$ in the regression:

$$\Delta_d = \boldsymbol{\phi}' X_{d-1} + \epsilon_d, \tag{13}$$

where X_{d-1} contains elements from the information set on day $d-1$, i.e., a constant and lags of the loss differential series $\Delta_d = \|\epsilon_{1,d}\|_p - \|\epsilon_{2,d}\|_p$, $\epsilon_{i,d}$ is the H-dimensional vector of prediction errors of model i for day d, $\|\epsilon_{i,d}\|_p = (\sum_{h=1}^{H} |\epsilon_{i,d,h}|^p)^{1/p}$ is the pth norm of that vector, and ϵ_d is an error term. Results of the CPA test for all pairs of models (due to the very poor performance reported in Table 2, the Naive-1N and Naive-N benchmarks are not considered) and both markets are illustrated in Fig. 5. Heat maps are used to denote the range of p-values – the smaller they are (→ dark green), the more significant the difference between the two forecasts (the model on the X-axis outperforms the model on the Y-axis).

Clearly, the LEAR-Ave model significantly outperforms all models; the columns corresponding to this ensemble are dark green in both panels of Fig. 5. Remarkably, all LEAR-based models, even the LEAR-N

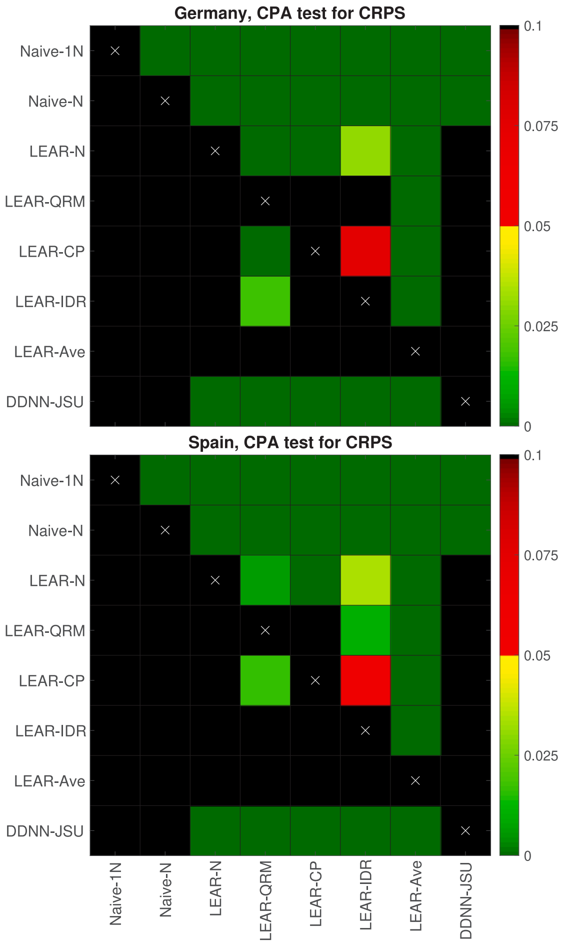


Fig. 5. Results of the CPA test of Giacomini and White (2006) for the CRPS, i.e., the aggregate pinball score for all 99 percentiles, for the whole 4.5-year German (*top*) and Spanish (*bottom*) test period. Heat maps are used to illustrate the range of *p*-values – the smaller they are (→ dark green), the more significant the difference between the two forecasts (the model on the X-axis outperforms the model on the Y-axis). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

benchmark, significantly outperform the DDNN-JSU, a model that is much more complex and computationally much more demanding; see the five dark green cells in the bottom row in both panels. This is a result of the poor performance of the neural network during the energy crisis and the war in Ukraine — Nov 2021 to Dec 2023 in Germany and Nov 2021 to Mar 2023 in Spain. Potentially, hyperparameter optimization conducted every few months could improve the model's predictive accuracy. This, however, would be a very time consuming task, see Section 4.6.

5.3. Performance in the tails of the distribution

In a risk management context we are interested in the tail behavior of the profit and loss (P&L) distribution. Hence, following (Uniejewski et al., 2019), we now consider only the percentiles that correspond to confidence levels typically used in risk management: below 10% and above 90%, i.e., the lower 10 and the upper 10 percentiles. In Table 3 we report the APS_{20} for the considered models and markets. The corresponding *p*-values of the CPA test for all pairs of models are illustrated in Fig. 6.

Table 3
Aggregate Pinball Score (APS_{20}) for 20 extreme percentiles (1, …, 10 and 90, …, 99, i.e., corresponding to confidence levels typically considered in risk management) for the considered models and markets. Like in Table 2, cells are colored independently for each row and market. The test period labeled '2000†' spans from 27.06.2019 to 31.12.2020 (554 days), the remaining three span full years (365 days).

Model	2020†	2021	2022	2023
		Germany		
Naive-1N	1.728	4.804	11.334	5.786
Naive-N	1.669	4.331	10.805	5.270
LEAR-N	0.691	2.006	4.629	2.121
LEAR-QRM	0.602	1.819	4.579	1.949
LEAR-CP	0.655	2.045	4.631	2.081
LEAR-IDR	0.648	2.176	4.985	1.914
LEAR-Ave	0.575	1.654	4.327	1.837
DDNN-JSU	0.555	2.763	6.321	2.437
		Spain		
Naive-1N	0.981	3.914	5.851	4.184
Naive-N	0.920	3.490	5.514	4.003
LEAR-N	0.447	1.870	3.290	2.117
LEAR-QRM	0.402	1.841	3.177	2.045
LEAR-CP	0.446	1.922	3.278	2.102
LEAR-IDR	0.431	2.130	3.333	1.856
LEAR-Ave	0.381	1.671	3.031	1.876
DDNN-JSU	0.401	2.259	3.696	1.922

This time, the LEAR-Ave ensemble yields the lowest CRPS across both markets and all test subperiods, except the first subperiod labeled '2000†' in Germany (where DDNN-JSU excels, but the difference is statistically insignificant) and year 2023 in Spain (where it is outperformed by LEAR-IDR). Still, the combination significantly outperforms all other models over the entire 4.5-year test period, see the CPA test results in Fig. 6.

Similarly as for the CRPS, all LEAR-based models significantly outperform the DDNN-JSU ensemble across the whole test sets; see the five dark green cells in the bottom row in both panels of Fig. 6. Yet, the high accuracy of the neural network in both markets in the first subperiod, i.e., directly after hyperparameter optimization, suggests that the DDNN-JSU has potential, especially in a risk management context.

Finally, comparing Figs. 5 and 6, we can observe that LEAR-IDR performs better overall (its forecasts are significantly more accurate that those of LEAR-N and LEAR-CP for Germany, and those of all three LEAR-based models for Spain) than for the extreme 20 percentiles (cells in the column corresponding to LEAR-IDR are black in Fig. 6). This indicates that IDR is better than its competitors for the more central quantiles.

6. Conclusions

Our study is the first to consider Isotonic Distributional Regression (IDR) and one of the first to use Conformal Prediction (CP) for electricity price forecasting. Overall, it highlights postprocessing as a relatively simple and well-performing means of deriving predictive distributions from point forecasts in such a challenging environment.

Like (Nitka and Weron, 2023), we find that introducing diversity to a pool of forecasts is highly beneficial. Combining the IDR-generated predictive distributions with those of the generally better performing QRA and CP schemes significantly improves the accuracy, as measured by Shapley values. The resulting LEAR-Ave combination outperforms state-of-the-art Distributional Deep Neural Networks of Marcjasz et al. (2023) over two 4.5-year test periods from the German and Spanish power markets, spanning the COVID pandemic and the war in Ukraine.

In the tails of the predictive distribution the situation is less straightforward. While for the whole test periods the LEAR-Ave ensemble significantly outperforms the DDNN-JSU model for both markets, as measured by the Conditional Predictive Ability (CPA) test of Giacomini and White (2006), in the first 1.5-year subperiod in Germany the

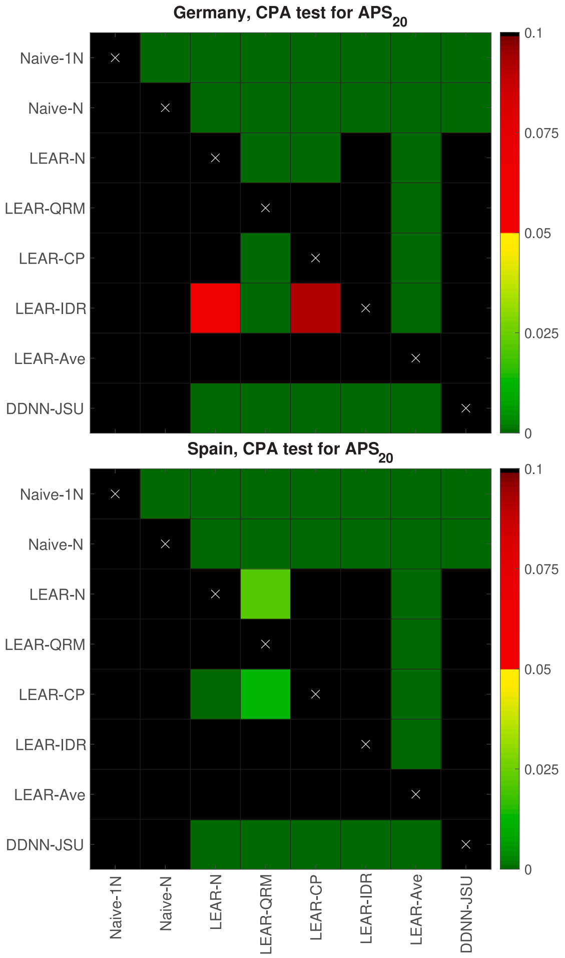


Fig. 6. Results of the CPA test of Giacomini and White (2006) for the APS$_{20}$ metric, i.e., the aggregate pinball score for the extreme top 10 and bottom 10 percentiles, for the whole 4.5-year German (*top*) and Spanish (*bottom*) test periods. The same type of a heat map is used as in Fig. 5.

DDNN-JSU network excels (the difference is statistically insignificant). Overall, we recommend the LEAR-Ave ensemble as a top performer and the LEAR-QRA model as a powerful all-rounder, second only to the combination. The DDNN-JSU network can provide accurate predictions in the tails of the distribution. However, it is beyond the scope of this study to examine whether frequent (extremely time-consuming) hyperparameter optimization would allow it to perform well during periods of extreme prices.

CRediT authorship contribution statement

Arkadiusz Lipiecki: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Conceptualization. **Bartosz Uniejewski:** Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Conceptualization. **Rafał Weron:** Writing – review & editing, Writing – original draft, Validation, Supervision, Funding acquisition, Conceptualization.

Acknowledgments

The study was partially supported by the National Science Center (NCN, Poland) through grant no. 2018/30/A/HS4/00444 (to AL), grant no. 2023/49/N/HS4/02741 (to BU) and grant no. 2021/43/I/HS4/02578 (to RW).

References

Baran, S., Lerch, S., 2018. Combining predictive distributions for the statistical post-processing of ensemble forecasts. Int. J. Forecast. 34 (3), 477–496.

Berrisch, J., Ziel, F., 2023. CRPS learning. J. Econometrics 237, 105221.

Billé, A., Gianfreda, A., Del Grosso, F., Ravazzolo, F., 2023. Forecasting electricity prices with expert, linear, and nonlinear models. Int. J. Forecast. 39 (2), 570–586.

Chen, J., Janke, T., Steinke, F., Lerch, S., 2024. Generative machine learning methods for multivariate ensemble postprocessing. Ann. Appl. Stat. 18 (1), 159–183.

Cornell, C., Dinh, N.T., Pourmousavi, S.A., 2024. A probabilistic forecast methodology for volatile electricity prices in the Australian national electricity market. Int. J. Forecast. 40 (4), 1421–1437.

Covert, I.C., Lundberg, S., Lee, S.-I., 2020. Understanding global feature contributions with additive importance measures. In: NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems. pp. 17212–17223.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression. Ann. Statist. 32 (2), 407–499.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33 (1), 1–22.

Gaillard, P., Goude, Y., Nedellec, R., 2016. Additive models and robust aggregation for GEFCom2014 probabilistic electric load and electricity price forecasting. Int. J. Forecast. 32 (3), 1038–1050.

Giacomini, R., White, H., 2006. Tests of conditional predictive ability. Econometrica 74 (6), 1545–1578.

Gneiting, T., Lerch, S., Schulz, B., 2023. Probabilistic solar forecasting: Benchmarks, post-processing, verification. Sol. Energy 252, 72–80.

Gneiting, T., Raftery, A., 2007. Strictly proper scoring rules, prediction, and estimation. J. Amer. Statist. Assoc. 102 (477), 359–378.

Grushka-Cockayne, Y., Jose, V.R.R., 2020. Combining prediction intervals in the M4 competition. Int. J. Forecast. 36 (1), 178–185.

Grushka-Cockayne, Y., Lichtendahl, K.C., Jose, V.R.R., Winkler, R.L., 2017. Quantile evaluation, sensitivity to bracketing, and sharing business payoffs. Oper. Res. 65 (3), 712–728.

Hastie, T., Tibshirani, R., Wainwright, M., 2015. Statistical Learning with Sparsity: The Lasso and Generalizations. CRC Press.

Henzi, A., Mösching, A., Dümbgen, L., 2022. Accelerating the pool-adjacent-violators algorithm for isotonic distributional regression. Methodol. Comput. Appl. Probab. 24 (4), 2633–2645.

Henzi, A., Ziegel, J.F., Gneiting, T., 2021. Isotonic distributional regression. J. R. Stat. Soc. Ser. B Stat. Methodol. 83 (5), 963–993.

Kath, C., Ziel, F., 2021. Conformal prediction interval estimation and applications to day-ahead and intraday power markets. Int. J. Forecast. 37 (2), 777–799.

Lago, J., Marcjasz, G., De Schutter, B., Weron, R., 2021. Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark. Appl. Energy 293, 116983.

Lichtendahl, K.C., Grushka-Cockayne, Y., Winkler, R.L., 2013. Is it better to average probabilities or quantiles? Manage. Sci. 59 (7), 1594–1611.

Lipiecki, A., Weron, R., 2024. PostForecasts.jl – Julia package for postprocessing forecasts. https://github.com/lipiecki/PostForecasts.jl.

Liu, B., Nowotarski, J., Hong, T., Weron, R., 2017. Probabilistic load forecasting via quantile regression averaging on sister forecasts. IEEE Trans. Smart Grid 8, 730–737.

Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.-I., 2020. From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence 2 (1), 56–67.

Maciejowska, K., Nowotarski, J., 2016. A hybrid model for GEFCom2014 probabilistic electricity price forecasting. Int. J. Forecast. 32 (3), 1051–1056.

Maciejowska, K., Uniejewski, B., Weron, R., 2023. Forecasting electricity prices. In: Oxford Research Encyclopedia of Economics and Finance. Oxford University Press, pp. 1–34.

Marcjasz, G., 2020. Forecasting electricity prices using deep neural networks: A robust hyper-parameter selection scheme. Energies 13 (18), 13184605.

Marcjasz, G., Narajewski, M., Weron, R., Ziel, F., 2023. Distributional neural networks for electricity price forecasting. Energy Econ. 125, 106843.

Marcjasz, G., Uniejewski, B., Weron, R., 2020. Probabilistic electricity price forecasting with NARX networks: Combine point or probabilistic forecasts? Int. J. Forecast. 36 (2), 466–479.

Narajewski, M., Ziel, F., 2020. Econometric modelling and forecasting of intraday electricity prices. J. Commod. Mark. 19, 100107.

Nitka, W., Weron, R., 2023. Combining predictive distributions of electricity prices. Does minimizing the CRPS lead to optimal decisions in day-ahead bidding? Oper. Res. Decis. 33, 103–116.

Nowotarski, J., Weron, R., 2015. Computing electricity spot price prediction intervals using quantile regression and forecast averaging. Comput. Statist. 30 (3), 791–803.

Nowotarski, J., Weron, R., 2018. Recent advances in electricity price forecasting: A review of probabilistic forecasting. Renew. Sustain. Energy Rev. 81, 1548–1568.

Olivares, K.G., Challu, C., Marcjasz, G., Weron, R., Dubrawski, A., 2023. Neural basis expansion analysis with exogenous variables: Forecasting electricity prices with NBEATSx. Int. J. Forecast. 39 (2), 884–900.

Rasp, S., Lerch, S., 2018. Neural networks for postprocessing ensemble weather forecasts. Mon. Weather Rev. 146 (11), 3885–3900.

Shafer, G., Vovk, V., 2008. A tutorial on conformal prediction. J. Mach. Learn. Res. 9, 371–421.

Uniejewski, B., Marcjasz, G., Weron, R., 2019. On the importance of the long-term seasonal component in day-ahead electricity price forecasting: Part II – probabilistic forecasting. Energy Econ. 79, 171–182.

Uniejewski, B., Weron, R., 2021. Regularized quantile regression averaging for probabilistic electricity price forecasting. Energy Econ. 95, 105121.

Uniejewski, B., Weron, R., Ziel, F., 2018. Variance stabilizing transformations for electricity spot price forecasting. IEEE Trans. Power Syst. 33, 2219–2229.

Vannitsem, S., Bremnes, J., Demaeyer, J., Evans, G., Flowerdew, J., Hemri, S., Lerch, S., et al., 2021. Statistical postprocessing for weather forecasts: Review, challenges, and avenues in a big data world. Bull. Am. Meteorol. Soc. 102 (3), E681–E699.

Walz, E.-M., Henzi, A., Ziegel, J., Gneiting, T., 2024. Easy uncertainty quantification (EasyUQ): Generating predictive distributions from single-valued model output. SIAM Rev. 66 (1), 91–122.

Wang, Y., Zhang, N., Tan, Y., Hong, T., Kirschen, D., Kang, C., 2019. Combining probabilistic load forecasts. IEEE Trans. Smart Grid 10 (4), 3664–3674.

Weron, R., 2014. Electricity price forecasting: A review of the state-of-the-art with a look into the future. Int. J. Forecast. 30 (4), 1030–1081.

Yang, D., Yang, G., Liu, B., 2023. Combining quantiles of calibrated solar forecasts from ensemble numerical weather prediction. Renew. Energy 215, 118993.

Zaffran, M., Féron, O., Goude, Y., Josse, J., Dieuleveut, A., 2022. Adaptive conformal predictions for time series. Proc. Mach. Learn. Res. 162, 25834–25866.

Ziel, F., Steinert, R., 2018. Probabilistic mid- and long-term electricity price forecasting. Renew. Sustain. Energy Rev. 94, 251–266.

Ziel, F., Weron, R., 2018. Day-ahead electricity price forecasting with high-dimensional structures: Univariate vs. multivariate modeling frameworks. Energy Econ. 70, 396–420.