Vol. 33, No. 3 (2023) | DOI: 10.37190/ord230307

OPEN ACCESS

Operations Research and Decisions

www.ord.pwr.edu.pl

OPERATIONS RESEARCH AND DECISIONS QUARTERLY

Combining predictive distributions of electricity prices. Does minimizing the CRPS lead to optimal decisions in day-ahead bidding?

Weronika Nitka[1*] Rafał Weron[1]

[1] *Department of Operations Research and Business Intelligence, Wrocław University of Science and Technology, Wrocław, Poland*

[*] *Corresponding author, email address: weronika.nitka@pwr.edu.pl*

Abstract

Probabilistic price forecasting has recently gained attention in power trading because decisions based on such predictions can yield significantly higher profits than those made with point forecasts alone. At the same time, methods are being developed to combine predictive distributions, since no model is perfect and averaging generally improves forecasting performance. In this article, we address the question of whether using CRPS learning, a novel weighting technique minimizing the continuous ranked probability score (CRPS), leads to optimal decisions in day-ahead bidding. To this end, we conduct an empirical study using hourly day-ahead electricity prices from the German EPEX market. We find that increasing the diversity of an ensemble can have a positive impact on accuracy. At the same time, the higher computational cost of using CRPS learning compared to an equal-weighted aggregation of distributions is not offset by higher profits, despite significantly more accurate predictions.

Keywords: *decision support, day-ahead electricity bidding, predictive distribution, combining forecasts, CRPS learning*

1. Introduction

To mitigate risks or increase profits from trading in day-ahead power markets, market participants use data-driven decision support techniques [12, 16, 17, 28]. For years, these have relied on point forecasts of the major variables of interest: loads (or demand for electricity), generation from renewable energy sources (RES), and electricity prices [10, 30]. However, as recently shown by Uniejewski and Weron [27], decisions based on probabilistic price forecasts, i.e., quantiles, prediction intervals, or whole predictive distributions, can yield significantly higher profits. For the quantile-based bidding strategies considered in the Polish day-ahead power market, the profit obtained was from 5% to 19% higher than for the strategy based on point forecasts alone.

Point forecasts are far more popular in the electricity price forecasting (EPF) literature, not only in a decision support context. As reported by Maciejowska et al. [18], probabilistic EPF was not a part of

Received 27 August 2023, accepted 6 September 2023, published online 21 September 2023
ISSN 2391-6060 (Online)/

The costs of publishing this issue have been cofinansed by the program *Development of Academic Journals* of the Polish Ministry of Education and Science under agreement No. RCN/SP/0241/2021/1

the mainstream literature until the Global Energy Forecasting Competition in 2014 [9], and even now, no more than 15% of the Scopus-indexed articles concern it. Although business analysts have begun to recognize their importance in the planning and operation of energy systems (see, e.g., [13]), it is not easy to generate accurate probabilistic predictions. Combining forecasts obtained from different model specifications [19] or calibration sample lengths [11, 26] one can significantly increase accuracy without sacrificing computational complexity or interpretability. Compared to selecting a single best-performing forecast, combining forecasts from multiple models offers several advantages such as increased resilience against model uncertainty or misspecification, and better adaptability in the event of structural breaks [29].

Forecast combinations (also called ensemble forecasts) involve assigning weights to the individual predictions (or experts). While naive, i.e., equal, weighting is a straightforward – and surprisingly robust – way of averaging point forecasts, in the case of predictive distributions, a choice must be made about what to combine. Two natural approaches are vertical averaging of probabilities and horizontal averaging of quantiles [15, 20] but the authors do not agree on which is better. Berrisch and Ziel [1] have recently proposed a cutting-edge weighting technique, called CRPS learning that accounts for variations in predictive performance over time and across quantiles of the distribution. It optimizes weights with respect to the continuous ranked probability score (CRPS), the standard error metric for probabilistic forecasts [8, 18].

In this article, we address the question of whether forecast combinations obtained by minimizing the CRPS lead to optimal decisions in day-ahead bidding. To this end, we conduct a comprehensive empirical study involving:

- six years of hourly day-ahead electricity prices from the German EPEX market,
- state-of-the-art probabilistic forecasts generated by Marcjasz et al. [20] using distributional deep neural networks (DDNN), as well as deep neural networks (DNN) and LASSO-estimated autoregressive (LEAR) models combined with quantile regression (QR),
- two approaches to combining predictive distributions – horizontal averaging of quantiles [15] and CRPS learning [1].

Since statistical measures of forecast accuracy do not assess the utility of a forecast to its potential end users [10, 13, 18, 31], we calculate the profits of a day-ahead bidding strategy [20, 25]. The latter aims to find the most financially beneficial hours of the next day to buy electricity and charge a battery, then discharge it and sell electricity. To minimize the risk of losses, limit orders are submitted to the power exchange with the limits determined by selected quantiles of the predictive distributions.

The remainder of the article is organized as follows. In Section 2, the dataset and assumptions for the forecasting problem are introduced. Section 3 describes the details of ensemble construction. The results are presented in Section 4, with the forecast accuracy being the focus of Section 4.1, the trading simulation described in 4.2, and its financial results in Section 4.3. Finally, Section 5 wraps up the results and concludes.

2. Preliminaries and data sources

We assume a standard short-term forecast horizon of 1 day, performed in a rolling window scheme [30]. More precisely, forecasts of all 24 hourly prices on the day d are calculated at the same time in the

morning of day $d-1$, i.e., before the day-ahead market for day d closes, and that the model parameters are estimated using a calibration sample of D most recent past observations. In our case, the underlying data are hourly day-ahead electricity prices from the German EPEX market spanning the period from 1 January 2015 to 31 December 2020. The prices, day-ahead predictions of the loads, and RES generation are publicly available from the ENTSO-E Transparency platform (https://transparency.entsoe.eu). The full dataset, including emission allowances and fuel prices, is also available from https://github.com/gmarcjasz/distributionalnn, a GitHub repository that accompanies [20].

The first $D = 1456$ days of the dataset are used as the initial calibration sample for all models, and the additional 182 days are needed to calculate the quantile regression forecasts. The remaining 554 days from 27 June 2019 until 31 December 2020 constitute the out-of-sample test period. Note that the latter includes a major drop in the level of prices associated with a decrease in the demand for electricity during the initial stage of the COVID-19 pandemic.

Since this study focuses on the evaluation of combination schemes for probabilistic forecasts and not on the computation of predictive distributions themselves, we work directly with a pool of readily available state-of-the-art forecasts generated by Marcjasz et al. [20]. The latter takes the form of 99 predicted percentiles for each day and hour, which approximate the predictive distribution quite well. They are generated by twelve different models, eight of which are distributional deep neural networks (DDNN) with the output layer returning fitted parameters of the normal or Johnson's SU (JSU) distributions. Since the quantile functions have no closed-form representations, the percentile forecasts are obtained as empirical quantiles of a 10000-element random sample generated from the output normal or JSU distribution. These forecasts are denoted further in the text as `DDNN_N_{1-4}` and `DDNN_JSU_{1-4}`, respectively, with the numbers representing the hyperparameter set used for tuning the DDNNs.

The remaining models directly predict 99 percentiles with the use of quantile regression averaging [QRA; 22] or quantile regression machine [QRM; 21] methods, applied to point forecasts of two well-performing benchmarks – LASSO-estimated autoregressive models (LEAR) and deep neural networks (DNN) [14]. The combinations of these techniques make up the final four forecasts used in the ensembles: `LEAR_QRA`, `LEAR_QRM`, `DNN_QRA` and `DNN_QRM`. Note that in the LEAR models, the prices for each of the 24 hourly load periods are treated as separate time series and estimated independently, whereas in the DNN and DDNN neural networks, the 24 prices or 24 distributions are estimated jointly.

3. Methods

Combining forecasts has become a well-established method to increase predictive accuracy. The advantages of using ensembles of experts in place of individual models include diversification of used information and increasing robustness against model misspecification and structural breaks in the data [24]. While the literature generally recommends combining forecasts, many questions still remain open regarding the construction of ensembles. Across a multitude of possible specifications, the forecaster must decide on how many predictions to combine, how to perform forecast selection, and how to choose weights for each expert. Combining probabilistic forecasts is even more tricky, as the assigned weights may change not just across experts and time, but also across quantile levels [29].

3.1. Equal weighting

In point forecasting, the use of naive, i.e., equal, weights is often found to outperform more sophisticated weighting schemes because the latter introduce excessive estimation bias [4]. In the case of predictive distributions, however, a choice must be made about what to combine. Two natural approaches are vertical averaging of probabilities, which boils down to computing a mixture distribution, and horizontal averaging of quantiles, where each quantile of the ensemble forecast is a weighted average of the corresponding quantiles of all individual experts [15]. While the literature does not agree on which approach is better, Marcjasz et al. [20] emphasize that horizontal averaging is more robust and results in a sharper, i.e., more concentrated, unimodal distribution. On the other hand, vertical averaging may lead to increased variance and multimodality. For this reason, as well as potential information loss due to interpolation needed to perform vertical averaging, only horizontal averaging of quantiles is considered in this paper. For consistency with other EPF studies, we denote it in the text by `qEns`.

3.2. CRPS learning

Berrisch and Ziel [1, 2] have recently proposed a cutting-edge weighting technique that accounts for variations in predictive performance over time and across quantiles; it is freely available in the *profoc* package for R ([3], https://cran.r-project.org/web/packages/profoc). The authors called it CRPS learning since it optimizes weights with respect to the continuous ranked probability score (CRPS). The latter is a proper scoring rule and the standard error metric for probabilistic forecasts [7, 8]. It is defined as:

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

where F is the cumulative distribution function of the evaluated probabilistic forecast. It can equivalently be represented as a scaled integral of the quantile loss, which for an equidistant grid can be approximated by:

$$\mathrm{CRPS}(F,\, x) \approx \frac{2}{M} \sum_{i=1}^{M} \mathrm{QL}_{p_i}\left(F^{-1}(p_i),\, x\right) \tag{2}$$

where $(p_1, \ldots, p_M)$ is an equidistant monotonically increasing dense grid of probabilities and $\mathrm{QL}_p(q, x) = \left(\mathbb{1}_{\{x<q\}} - p\right)(q - x)$ is the quantile loss for a quantile forecast q of true value x for probability $p \in (0,\ 1)$, also known as the pinball score [1, 18]. In practice, the scaling factor of 2 in eq. (2) is typically omitted; this is also the case here.

The CRPS learning algorithm aims to combine probabilistic forecasts by selecting optimal weights for averaging across quantiles to minimize the CRPS of the resulting ensemble. The weight functions are subject to online updating throughout the forecasting period and are chosen pointwise, i.e., for each quantile of the distribution separately, depending on each expert's performance. The framework additionally includes smoothing procedures that reduce estimation noise of the weights [2].

In this study, the CRPS learning framework was applied once per ensemble, with the following arbitrarily chosen set of parameters: Bernstein online aggregation (BOA) for updating weights, penalized probabilistic smoothing with $\lambda = 2^{(-5, \ldots, 5)}$ updated based on past performance, and no forgetting past

regret. The remaining options were set to the *profoc* package defaults. Such an approach is denoted in the text by `CRPS`. Finally, note that the time required to compute forecasts of a single CRPS learning ensemble for the entire test period is ca. 500 times longer than that for the naive `qEns` weighting. However, it does not exceed 20 s on a laptop equipped with a 9th-generation Intel Core i7-9750H processor.

3.3. Comparison of the two weighting schemes

The general idea of averaging across quantiles, as well as differences between the two weighting schemes, are shown in Figure 1. The illustration shows a toy example of a two-forecast ensemble. Among the two experts, the `DDNN_JSU_1` forecast (teal color) is sharper, i.e., more concentrated, predicting prices between 26 and 37 €, and has a smoother cumulative distribution function (CDF), while the `LEAR_QRA` predictive distribution (red color) is less sharp (with prices between 15 and 44 €) and more rugged. Medians of both experts are relatively close to the actual observed price (31.89 €, vertical line), albeit leaving room for improvement, i.e., with absolute errors of 0.52 and 1.24 €, respectively.

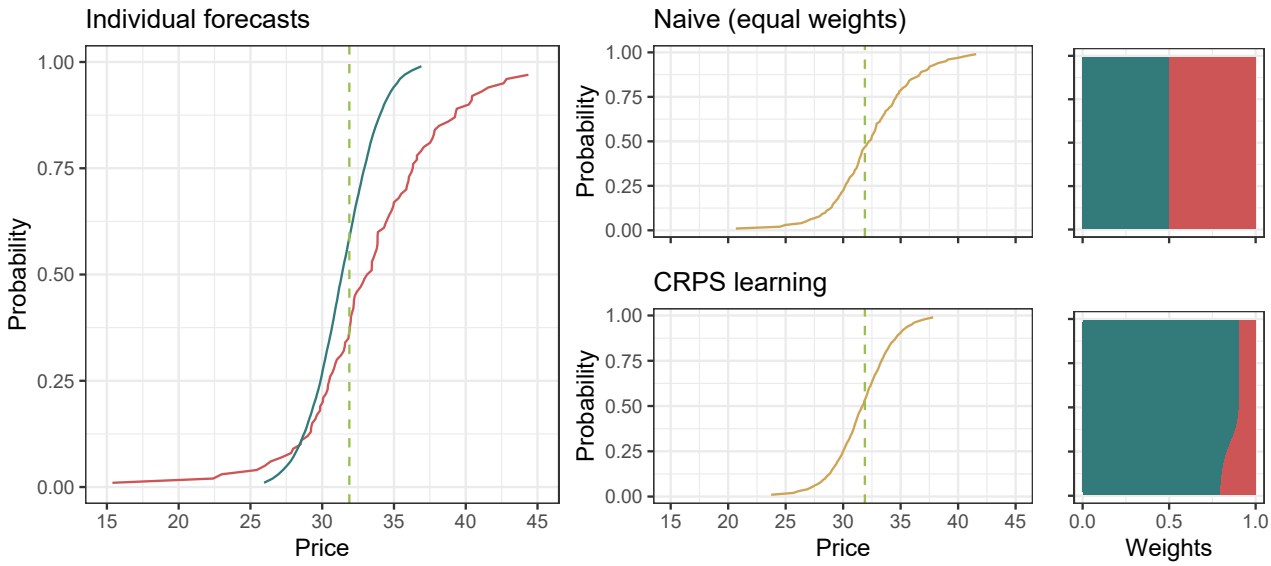

Figure 1. Illustration of the two weighting schemes. The left panel shows predictive distributions obtained from the `DDNN_JSU_1` (teal color) and `LEAR_QRA` (red color) models for a selected hour and day. The center panels present the resulting ensemble forecasts obtained by estimating weights with naive (top) and CRPS learning (bottom) methods. The right panels illustrate the relative weights for each quantile; these are horizontally stacked bar plots with the length of the bar representing the weight of the forecast in the corresponding color and all weights summing up to 1. The dashed vertical line marks the actual price

The two individual forecasts are combined using the two weighting approaches, with the resulting CDFs and the assigned weights shown in the panels to the right. It can be seen that while the `qEns` approach, by definition, assigns equal weights to all models and quantiles, CRPS learning assigns larger weights to the `DDNN_JSU_1` forecast, based on its better past performance (not shown in the plot). The share of the `DDNN_JSU_1` forecast is smaller for the lowest 25 percentiles, but nevertheless, it still dominates the `CRPS` ensemble, leading to its higher sharpness (price range of [24, 38] €) and smoothness compared to the equally weighted ensemble (with values in the range [21, 42] €). However, both forecast combinations provide a more accurate median forecast than the individual experts, with absolute errors of 0.36 € for `qEns` ensemble and 0.22 € for CRPS learning.

3.4. Selection of experts

The forecaster's second decision is the selection of experts that are aggregated in the ensemble. Following [20], each ensemble we consider in this study contains a set of four DDNN forecasts, either `DDNN_N_{1-4}` or `DDNN_JSU_{1-4}`. Furthermore, to diversify the pool of experts we additionally include benchmark quantile regression-based forecasts – either `LEAR_QRA` and `LEAR_QRM` or `DNN_QRA` and `DNN_QRM`. While they have been demonstrated to perform significantly worse on their own, using them can lead to a higher prediction accuracy of the ensembles by avoiding overfitting [29]. Thus, the resulting naming convention for ensembles is:

`DDNN_{distribution}_{averaging}_{experts}`

with `distribution={N,JSU}` denoting whether normal or JSU forecasts were used, `averaging={qEns,CRPS}` indicating the use of equal or CRPS learning-derived weights, and `experts={LEAR,DNN}` added when additional experts – respectively `LEAR_QRA` and `LEAR_QRM` or `DNN_QRA` and `DNN_QRM` – were included in the ensemble. A graphical illustration of the steps performed in order to construct the ensemble forecasts is shown in Figure 2.

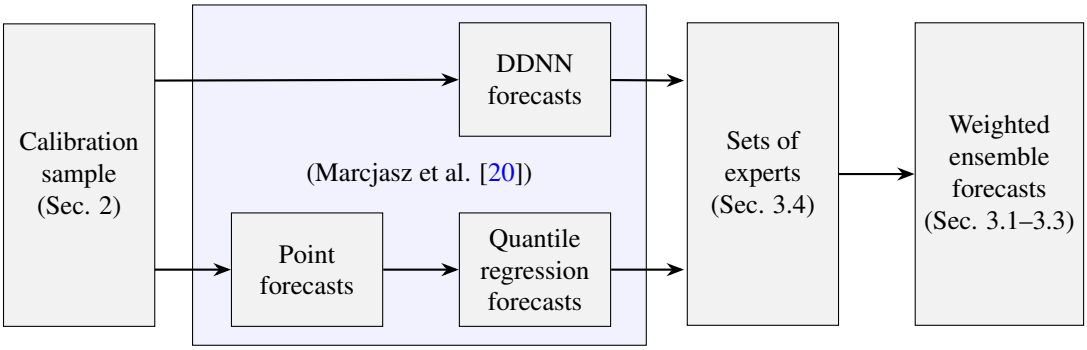

Figure 2. Schematic illustration of the process of generating ensemble forecasts

It should be noted that additional forecast combinations were explored during the course of this research. The complete list included smaller ensembles (four DDNN forecasts and a single QR-based forecast; best 5 performing models), other combinations of quantile regression forecasts (e.g., four DDNN forecasts, `LEAR_QRA`, `DNN_QRA`) and larger ensembles (four DDNN forecasts and four QR-based forecasts; eight DDNN forecasts as in [2]; all available forecasts). They offered comparable or, in the case of the largest ensembles, significantly inferior performance to the combinations listed above, and have been omitted from the presentation of results for the sake of clarity.

In an online learning setting, the forecaster may decide to discard the initial part of the test sample as a burn-in period, which is beneficial for the stability of weights and hyperparameters, see, e.g., [2], which uses a burn-in period of 182 days for combinations of DDNN forecasts. However, the quantile regression forecasts are only available within the 554-day out-of-sample test period, the entirety of which has been used in Berrisch et al. [2] and Marcjasz et al. [20] for evaluation. Since the majority of ensembles we consider in this study include these quantile regression forecasts, for the sake of consistency no burn-in period has been applied.

4. Forecast evaluation

In this section, the generated ensemble forecasts are compared to each other and individual expert models. The evaluation is divided into two parts. First, in Section 4.1, we measure the predictive accuracy in terms of statistical error metrics:

- the mean absolute error (MAE) and the root mean squared error (RMSE) for median and mean forecasts, respectively [6],
- the continuous ranked probability score (CRPS) for probabilistic forecasts [8, 23].

Then, in Sections 4.2–4.3, we measure the predictive accuracy in terms of profits – total and per trade – from a day-ahead bidding strategy that utilizes probabilistic forecasts [20, 25]. Note that the CRPS is approximated by a sum of pinball scores on a grid of 99 percentiles, see eq. (2). The statistical significance of differences in CRPS scores is assessed using the Diebold–Mariano test [5].

4.1. Evaluation in terms of statistical error measures

As Gneiting et al. [7, 8] argue, the goal of probabilistic forecasting is to *maximize the sharpness of the predictive distributions subject to calibration*. Here, calibration (also called reliability or unbiasedness) refers to the statistical consistency between the probabilistic forecasts and the observations, e.g., whether the 50% prediction interval (PI) covers 50% of the actual observations. Sharpness, on the other hand, refers to the concentration of the predictive distributions. For instance, given two reliable 50% PIs, the sharper or more narrow one is better. The CRPS introduced in Section 3 assesses calibration and sharpness simultaneously [7]. Moreover, for a point forecast, it is equal to the MAE [23].

The CRPS values for all models ordered from the lowest/best to the highest/worst are shown in the left panel of Figure 3; the corresponding MAE and RMSE errors in the right panel. Clearly, the two `LEAR` forecasts perform the worst, while the `DDNN_JSU_4` and two `DNN` forecasts the best out of the individual models. Moreover, the individual experts are outclassed by all `DDNN_N` ensembles, which are further outperformed by the `DDNN_JSU` combinations. The `DDNN_JSU_CRPS_LEAR` ensemble achieves the lowest CRPS score. For all ensembles, both weighting schemes typically result in very similar forecasts and thus accuracy. Nevertheless, CRPS learning yields slightly better predictions on average. A similar, though not identical, ordering can be observed for the point forecasting error metrics. The most significant differences are obtained for the `DDNN_JSU_{1,3,4}` experts in terms of the RMSE. As in [20], we calculate the MAE for the median (i.e., the 50th percentile) and the RMSE for the expected value of each distribution; the errors presented for the individual forecasts as well as the `DDNN_N_qEns` and `DDNN_JSU_qEns` ensembles are consistent with the results reported in [20].

To assess the statistical significance of differences in CRPS scores, we perform the Diebold–Mariano (DM) test [5]. In order to correct for daily seasonality in CRPS values, following [14] and [20], we consider a multivariate loss differential series defined for a pair of models A and B as:

$$\Delta_d^{A,B} = \|L_d^A\|_1 - \|L_d^B\|_1 \tag{3}$$

where $L_d^X = \{L_{d,1}^X, \ldots, L_{d,24}^X\}$ is the 24-dimensional vector of hourly CRPS values for model X on day d and $\|L_d^X\|_1$ is its L_1 norm. For each pair of models we apply two one-sided DM tests.

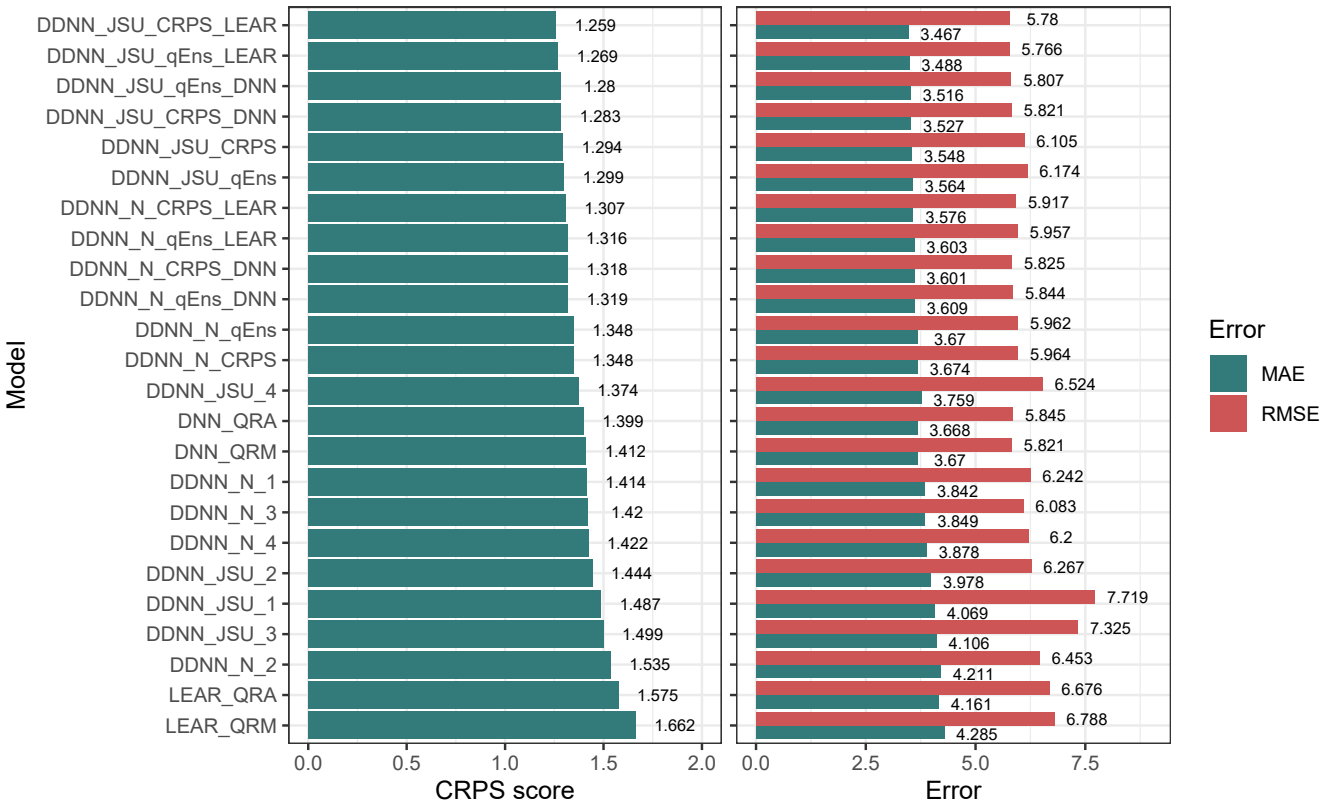

Figure 3. CRPS scores (left panel) and MAE and RMSE errors (right panel) for all models and ensembles ordered from the lowest to the highest CRPS. Compare with a CRPS of 1.284 of the best performing model of Berrisch and Ziel [2]

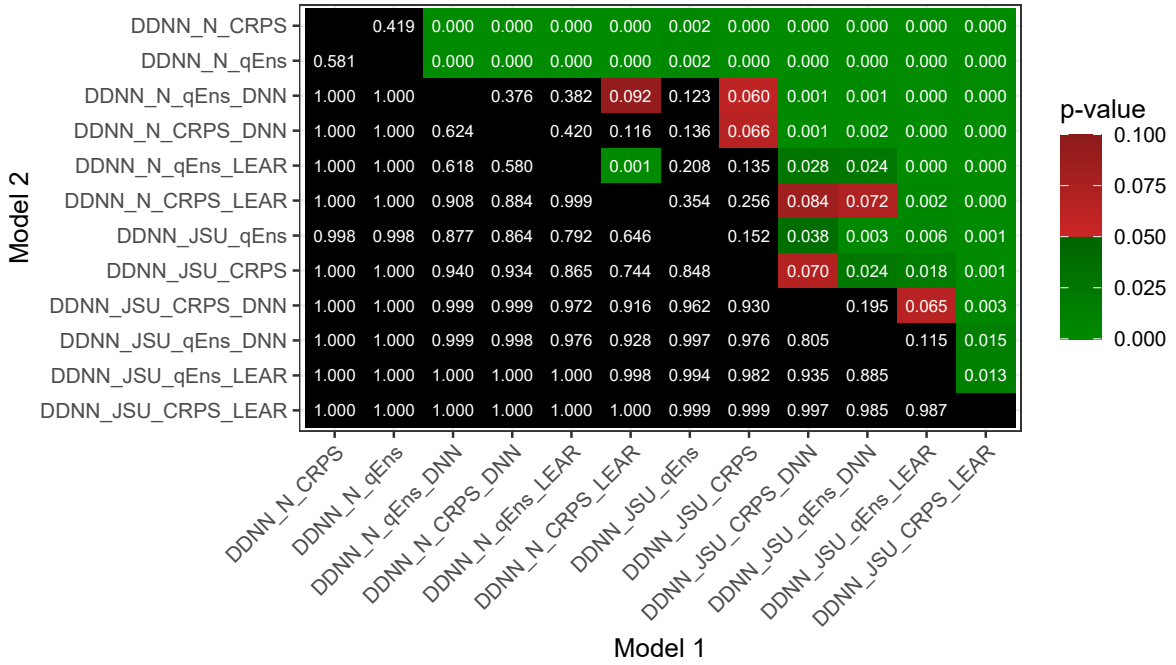

Figure 4. Results (p-values) of the Diebold–Mariano test for the CRPS loss; the lower it is the more significant is the difference between the forecasts of a model on the X-axis (better) and the forecasts of a model on the Y-axis (worse). We use a coloring scheme to highlight the differences

A heatmap of the respective p-values is presented in Figure 4. The results indicate that the predictions of the DDNN_JSU_CRPS_LEAR ensemble are significantly better than those of all competing models. The predictions of the remaining ensembles within the top 4 do not significantly differ from each other. Another ensemble whose forecasts are significantly better than those ranked lower in terms of the CRPS is DDNN_JSU_CRPS_DNN, while most other ensembles do not yield significantly better predictions than ensembles similarly ranked in terms of the CRPS.

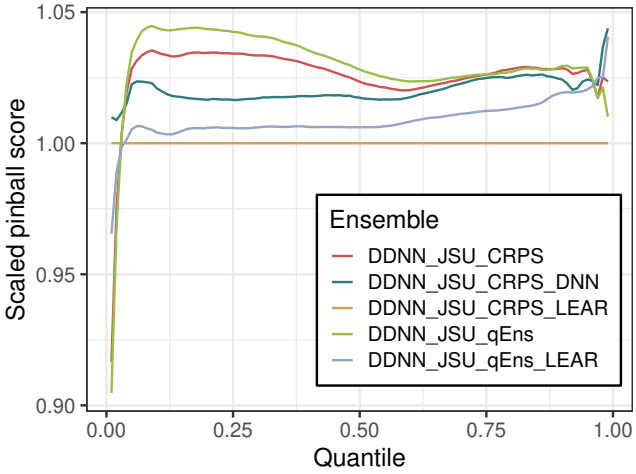

Figure 5. Pinball scores of selected best performing ensembles across quantiles, relative to the DDNN_JSU_CRPS_LEAR ensemble. A lower score corresponds to better performance

The CRPS score provides a single number for all quantiles (and each time point in the test period). To see how the pinball scores for individual percentiles contribute to the CRPS, in Figure 5 we plot them for selected best-performing ensembles. To enhance readability, all values are plotted with respect to the pinball scores of the best-performing ensemble, i.e., DDNN_JSU_CRPS_LEAR. Clearly, the relative performance of the ensembles is not uniform across the entire distributions. The largest disparity can be seen below the median, with the relative ranking of the ensembles changing for the three lowest percentiles. Above the median, the ensembles perform similarly.

4.2. Day-ahead bidding

Following Uniejewski [25], we consider a realistic trading strategy that utilizes battery storage and day-ahead bidding based on probabilistic price forecasts. The goal is to buy electricity cheaply at hour $h1$ and charge the battery, then discharge it and sell the electricity expensively at hour $h2 > h1$. To minimize the risk of losses, limit orders are submitted to the power exchange with the limits determined by selected – based on the trader's risk appetite – quantiles of the predictive distributions.

We assume that the efficiency of charging as well as discharging the battery is 90%. Hence, $1/0.9 \approx 1.1$ MWh is needed to charge the battery by 1 MWh. Similarly, discharging 1 MWh generates only 0.9 MWh. Further, we assume that the total usable capacity of the battery is $B = 2$ MWh and that at the beginning of the simulation period, the battery starts halfway charged ($B = 1$). If both orders are executed on the next day, this state persists. If $B = 0$ at the beginning of a day, an unlimited bid to buy 1 MWh is placed at hour $h* < h2$, and if $B = 2$, an unlimited offer to sell 1 MWh is placed at hour $h* < h1$.

For each day in the out-of-sample test period, the following two steps are performed. First, based on median price forecasts $Y^{0.5}_{d,h}$ for day d and hours $h = 1, 2, ..., 24$ computed on day $d-1$, hours $h1$, $h2$ and $h*$ are selected to maximize the profit:

$$\Pi_d = -\frac{1}{0.9}\hat{Y}^{0.5}_{d,h1} + 0.9\hat{Y}^{0.5}_{d,h2} - \mathbb{1}_{\{B=0\}}\frac{1}{0.9}\hat{Y}^{0.5}_{d,h*} + \mathbb{1}_{\{B=2\}}0.9\hat{Y}^{0.5}_{d,h*} \tag{4}$$

When the battery is halfway charged ($B = 1$), this optimization problem reduces to selecting hours with the lowest and the highest predicted median price. In other cases, linear programming is used to optimize the selection of $h1$, $h2$, and $h*$.

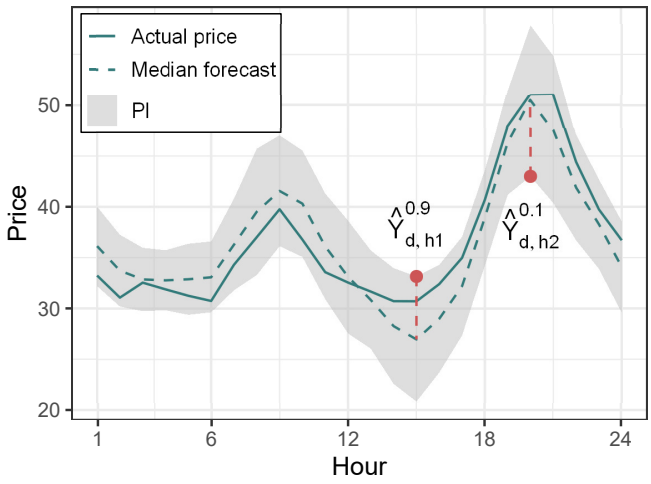

Figure 6. Illustration of the trading strategy with limit orders defined by the 80% PIs, corresponding to a risk appetite of 0.8. Red dots indicate the price limits for the selected hours

Next, following Marcjasz et al. [20], a profitability condition is checked. If the transaction is expected to be profitable, i.e., the sum of the first two terms in eq. (4) is greater than zero, a buy order with price limit $\hat{Y}^{1-q}_{d,h1}$ and a sell order with price limit $\hat{Y}^{q}_{d,h2}$ are placed. Here $q = (1-\alpha)/2$ and α is trader's risk appetite, i.e., the PI level, set only once for the whole test period. This is illustrated in Figure 6 for a sample day and forecasts generated by the `DDNN_N_qEns` ensemble. In this example, both orders would be accepted since the actual price falls within the 80% PI, corresponding to a risk appetite of $\alpha = 0.8$. However, hour h_2 is predicted suboptimally, a slightly higher price was observed for hour 21.

4.3. Evaluation in terms of trading profits

The total profits are presented in Table 1 for five values of risk appetite α ranging from 0.5 to 0.9; the minor differences between the reported values and those in [20] for the `DDNN_N_qEns` and `DDNN_JSU_qEns` ensembles are a result of correcting a bug in the original software. The profitability results somewhat correspond to the CRPS results, although with a few notable exceptions. On average, the `DDNN_JSU` ensembles achieve higher total profits than the `DDNN_N` ensembles, mirroring their better performance in terms of the CRPS. The detailed ranking of those ensembles is where the outcomes start to vary. While the CRPS weighting scheme outperforms equal weights in terms of forecast accuracy, this trend is mostly reversed in the financial results. This is especially true for lower values of the risk appetite.

Table 1. Total profits from the quantile-based trading strategy in the whole test period for risk appetite ranging from 0.5 to 0.9. The highest values in each column are in bold. Cells are colored independently in each column from the best (→ green) to the worst (→ red)

Model	Risk appetite 0.5	0.6	0.7	0.8	0.9
DDNN_N_CRPS	11603	11669	11418	10522	7283
DDNN_N_qEns	11772	11710	11322	10120	6544
DDNN_N_qEns_DNN	11719	11850	11609	11196	8794
DDNN_N_CRPS_DNN	11587	11790	11798	11635	9792
DDNN_N_qEns_LEAR	11871	11903	11818	11163	8359
DDNN_N_CRPS_LEAR	11524	11764	11830	11304	9040
DDNN_JSU_qEns	12122	12186	12106	11715	10189
DDNN_JSU_CRPS	11806	12044	12033	11719	10500
DDNN_JSU_CRPS_DNN	11514	11852	11978	**11960**	10868
DDNN_JSU_qEns_DNN	11601	11844	11994	11820	10374
DDNN_JSU_qEns_LEAR	**12210**	**12225**	**12115**	11922	10409
DDNN_JSU_CRPS_LEAR	11924	12138	11995	11931	**10911**

Figure 7. Average daily (dark green) and the minimum and maximum hourly (light green) prices in Germany from 27 June 2019 to 31 December 2020 (top panel). 30-day moving average of cumulative profit for the best-performing strategies, shown as a difference between the cumulative profit of each ensemble and `DDNN_JSU_CRPS_LEAR` for risk appetite $\alpha = 0.6$ (center panel) and $\alpha = 0.8$ (bottom panel)

For instance, the most accurate in terms of the CRPS ensemble, i.e., `DDNN_JSU_CRPS_LEAR`, yields lower profits than its `qEns` counterpart. This is likely due to the fact that while the CRPS weighting is more accurate on average, it is significantly outperformed by the naive weighting for the few lowest percentiles, see Figure 5, giving the latter an advantage during the initial stage of the COVID-19 pandemic (middle part of the test period), see Figure 7. A similar behavior can be observed for the

DDNN_JSU_CRPS_DNN ensemble, which performs worse than its competitors for the extreme quantiles, being at a significant disadvantage in the second half of the evaluation period, when the daily price spread is higher than in the beginning.

It can be expected that an optimal trading strategy would result in executing exactly two trades per day, buying on the low and selling on the high. With such a "crystal ball" strategy, the trader would earn 13,587 € throughout the whole evaluation period. Conversely, taking the worst possible decisions would lead to a total loss of $-21,425$ €. On this scale of possible profits, all evaluated ensembles rank relatively well. The lowest profit presented in Table 1 reaches 80% of the maximum, while the best of all forecasts as much as 96%. For comparison, a naive strategy of placing bids at fixed hours selected *ex-post* as having the highest price spread on average – buying at hour 3 and selling at hour 19 – would lead to total profits of 8048 €, or 84% of the maximum, see [20].

Table 2. Profits per trade from the quantile-based trading strategy in the whole test period for risk appetite ranging from 0.5 to 0.9. The highest values in each column are in bold. Cells are colored independently in each column from the best (→ green) to the worst (→ red).

Model	Risk appetite 0.5	0.6	0.7	0.8	0.9
DDNN_N_CRPS	11.20	11.29	11.56	11.77	12.78
DDNN_N_qEns	11.32	11.46	11.65	11.85	13.25
DDNN_N_qEns_DNN	11.23	11.33	11.61	12.12	**14.42**
DDNN_N_CRPS_DNN	11.16	11.36	11.52	12.12	14.07
DDNN_N_qEns_LEAR	11.60	11.60	11.77	12.38	13.88
DDNN_N_CRPS_LEAR	11.40	11.44	11.60	11.92	13.03
DDNN_JSU_qEns	11.57	11.81	11.91	12.33	13.96
DDNN_JSU_CRPS	11.33	11.56	11.75	12.06	13.57
DDNN_JSU_CRPS_DNN	11.16	11.42	11.70	12.25	13.55
DDNN_JSU_qEns_DNN	11.22	11.48	11.71	12.34	14.25
DDNN_JSU_qEns_LEAR	**11.72**	**11.89**	**11.95**	**12.47**	13.99
DDNN_JSU_CRPS_LEAR	11.67	11.76	11.83	12.08	13.40

The profit per trade results reported in Table 2 are less clear-cut, with the DDNN_N ensembles no longer being completely outclassed by the DDNN_JSU ensembles, especially when there are fewer total trades. As profits per trade can be seen as an indicator of the trader's risk, this disparity is consistent with literature findings [12]. It is worth noticing that, perhaps unintuitively, higher values of the risk appetite correspond to higher risk aversion. This is an effect of the final step of the strategy, which checks the income of the worst case scenario. With higher values of the risk appetite, this predicted profit tends to be lower, leading the trader to act more cautiously. This seemingly leads to a dominance of models which are more accurate across the entire distribution rather than only in the extreme quantiles, compare the center (risk appetite $\alpha = 0.6$) and bottom (risk appetite $\alpha = 0.8$) panels of Figure 7 with Figure 5. However, this is only a conjecture, as the relationship between daily price levels, price spreads, and PIs is not direct or linear.

5. Conclusions and discussion

In this article, we address the question of whether minimizing the continuous ranked probability score (CRPS) – the standard error metric for probabilistic forecasts – leads to optimal decisions in day-ahead bidding. Conducting an extensive empirical study, we find that introducing diversity to a pool of fore-

casts is highly beneficial, both in terms of forecast accuracy measured by the CRPS and profits from a trading strategy implemented in the German day-ahead power market. Also optimizing combination weights with CRPS learning positively impacts forecast accuracy. This is likely caused by the uneven performance of experts across time and quantiles, which is an outcome consistent with the literature.

While trading profits generally follow forecast accuracy, the benefits of using CRPS learning are not as pronounced in the trading scenario, especially considering the ca. 500 times higher computational burden. The precise cause-and-effect relationships between the predictive accuracy and profits are difficult to disentangle. However, the performance for the extreme quantiles of the distribution seems to be related to some of the observed patterns. In general, using any of the considered ensembles leads to achieving satisfactory profits, especially when compared to the best-case and worst-case scenarios.

For the sake of clarity, only selected forecast sets were considered. Extending the pool of experts and ensembles could lead to a more comprehensive evaluation. Other possible extensions of this study include comparisons with other weighting schemes, as well as automated methods for expert selection.

Acknowledgement

This research was partially supported by the Ministry of Science and Higher Education (MNiSW, Poland) through Diamond Grant No. 0027/DIA/2020/49 (to W.N.) and the National Science Center (NCN, Poland) through grant No. 2018/30/A /HS4/00444 (to R.W.).

References

[1] Berrisch, J., and Ziel, F. CRPS learning. *Journal of Econometrics* (2021), 105221. DOI: 10.1016/j.jeconom.2021.11.008.

[2] Berrisch, J., and Ziel, F. Multivariate probabilistic CRPS learning with an application to day-ahead electricity prices, 2023. DOI: 10.48550/arXiv.2303.10019. Working paper version available from arXiv: https://arxiv.org/abs/2303.10019.

[3] Berrisch, J., and Ziel, F. *The profoc Package: An R package for probabilistic forecast combination using CRPS Learning*, 2023. R package version 1.2.0.

[4] Blanc, S. M., and Setzer, T. Bias–variance trade-off and shrinkage of weights in forecast combination. *Management Science 66*, 12 (2020), 5720–5737.

[5] Diebold, F. X., and Mariano, R. S. Comparing predictive accuracy. *Journal of Business & Economic Statistics 13*, 3 (1995), 253–263.

[6] Gneiting, T. Making and evaluating point forecasts. *Journal of the American Statistical Association 106*, 494 (2011), 746–762.

[7] Gneiting, T., and Katzfuss, M. Probabilistic forecasting. *The Annual Review of Statistics and Its Application 1* (2014), 125–151.

[8] Gneiting, T., and Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association 102*, 477 (2007), 359–378.

[9] Hong, T., Pinson, P., Fan, S., Zareipour, H., Troccoli, A., and Hyndman, R. J. Probabilistic energy forecasting: Global Energy Forecasting Competition 2014 and beyond. *International Journal of Forecasting 32*, 3 (2016), 896–913.

[10] Hong, T., Pinson, P., Wang, Y., Weron, R., Yang, D., and Zareipour, H. Energy forecasting: A review and outlook. *IEEE Open Access Journal of Power and Energy 7* (2020), 376–388.

[11] Hubicka, K., Marcjasz, G., and Weron, R. A note on averaging day-ahead electricity price forecasts across calibration windows. *IEEE Transactions on Sustainable Energy 10*, 1 (2019), 321–323.

[12] Janczura, J., and Puć, A. ARX-GARCH probabilistic price forecasts for diversification of trade in electricity markets—variance stabilizing transformation and financial risk-minimizing portfolio allocation. *Energies 16*, 2 (2023).

[13] Janczura, J., and Wójcik, E. Dynamic short-term risk management strategies for the choice of electricity market based on probabilistic forecasts of profit and risk measures. The German and the Polish market case study. *Energy Economics 110* (2022), 106015.

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

[15] Lichtendahl, K. C., Grushka-Cockayne, Y., and Winkler, R. L. Is it better to average probabilities or quantiles? *Management Science 59*, 7 (2013), 1594–1611.

[16] Maciejowska, K. Portfolio management of a small RES utility with a structural vector autoregressive model of electricity markets in Germany. *Operations Research and Decisions 32*, 4 (2022), 75-90.

[17] Maciejowska, K., Nitka, W., and Weron, T. Enhancing load, wind and solar generation for day-ahead forecasting of electricity prices. *Energy Economics 99* (2021), 105273.

[18] Maciejowska, K., Uniejewski, B., and Weron, R. Forecasting electricity prices. In *Oxford Research Encyclopedia of Economics and Finance*. Oxford University Press, 2023. DOI: 10.1093/acrefore/9780190625979.013.667. Working paper version available from arXiv: https://doi.org/10.48550/arXiv.2204.11735.

[19] Makridakis, S., Spiliotis, E., and Assimakopoulos, V. The M4 Competition: Results, findings, conclusion and way forward. *International Journal of Forecasting 34*, 4 (2018), 802–808.

[20] Marcjasz, G., Narajewski, M., Weron, R., and Ziel, F. Distributional neural networks for electricity price forecasting. *Energy Economics 125* (2023), 106843.

[21] Marcjasz, G., Uniejewski, B., and Weron, R. Probabilistic electricity price forecasting with NARX networks: Combine point or probabilistic forecasts? *International Journal of Forecasting 36*, 2 (2020), 466–479.

[22] Nowotarski, J., and Weron, R. Computing electricity spot price prediction intervals using quantile regression and forecast averaging. *Computational Statistics 30*, 3 (2015), 791–803.

[23] Nowotarski, J., and Weron, R. Recent advances in electricity price forecasting: A review of probabilistic forecasting. *Renewable and Sustainable Energy Reviews 81* (2018), 1548–1568.

[24] Timmermann, A. Forecast Combinations. In *Handbook of Economic Forecasting*, G. Elliott, C. W. J. Granger, and A. Timmermann, Eds., vol. 1. Elsevier, 2006, pp. 135–196.

[25] Uniejewski, B. Smoothing Quantile Regression Averaging: A new approach to probabilistic forecasting of electricity prices, 2023. DOI: 10.48550/arXiv.2302.00411. Working paper version available from arXiv: https://arxiv.org/abs/2302.00411.

[26] Uniejewski, B., and Maciejowska, K. Lasso principal component averaging: A fully automated approach for point forecast pooling. *International Journal of Forecasting* (2022). DOI: 10.1016/j.ijforecast.2022.09.004, (in press).

[27] Uniejewski, B., and Weron, R. Regularized quantile regression averaging for probabilistic electricity price forecasting. *Energy Economics 95* (2021), 105121.

[28] Vogler, A., and Ziel, F. Event-based evaluation of electricity price ensemble forecasts. *Forecasting 4*, 1 (2022), 51–71.

[29] Wang, X., Hyndman, R. J., Li, F., and Kang, Y. Forecast combinations: An over 50-year review. *International Journal of Forecasting* (2022). DOI: 10.1016/j.ijforecast.2022.11.005, (in press).

[30] Weron, R. Electricity price forecasting: A review of the state-of-the-art with a look into the future. *International Journal of Forecasting 30*, 4 (2014), 1030–1081.

[31] Yardley, E., and Petropoulos, F. Beyond error measures to the utility and cost of the forecasts. *Foresight: The International Journal of Applied Forecasting 63* (2021), 36–45.