WORKing papers in Management Science

WORMS/24/03

Loss functions in regression models: Impact on profits and risk in day-ahead electricity trading

Tomasz Serafin[1]
Rafał Weron[1]

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

WORMS is a joint initiative of the Management Science departments
of the Wrocław University of Science and Technology,
Wyb. Wyspiańskiego 27, 50-370 Wrocław, Poland

Loss functions in regression models: Impact on profits and risk in day-ahead electricity trading

Tomasz Serafin[a,*], Rafał Weron[a]

[a]*Department of Operations Research and Business Intelligence, Wrocław University of Science and Technology, 50-370 Wrocław, Poland*

Abstract

We study the impact of the loss function used to estimate the parameters of a regression-type model on profits and risk in day-ahead electricity trading. To provide practical insights, we consider a strategy that incorporates battery storage and includes realistic operating costs in the calculation of revenues. Using 2021-2024 data from the German market as the testing ground, we provide evidence that minimizing a loss function that combines absolute errors with a quadratic penalty for price spread predictions of the opposite sign is the preferred option. Forecasts based on the introduced *directional loss function* repeatedly and in the majority of cases yield trading decisions that outperform those based on predictions from models estimated using squared, absolute, percentage, or asymmetric losses, as measured by the Sharpe ratio and profits per trade.

Keywords: Electricity price forecast, Day-ahead market, Loss function, Trading strategy, Battery storage, Sharpe ratio

1. Introduction

In a recent discussion, Kolassa (2020) emphasized that a "good" forecast was the one that minimized the expected value of the *loss function* $\mathcal{L}(\hat{y}, y)$. The latter (also called the *scoring function*, *cost function*, *error measure*, or *accuracy measure*; Davydenko and Fildes, 2013; Gneiting, 2011b; Hastie et al., 2015; Hyndman and Koehler, 2006) constitutes the cost or penalty if we predict $\hat{y}$ and y is observed. In a related context, Gneiting (2011a) argued that a forecaster was able to provide the optimal forecast only if the scoring function (or the functional of the predictive distribution, such as the mean, median or quantile) was disclosed *ex-ante*. This implies that the loss function minimized in the forecast generation process should be aligned with the error measure used in the subsequent forecast evaluation. However, theory does not provide insight into the selection of the optimal error measure in a practical setting.

Motivated by differing opinions of scholars and practitioners, Koutsandreas et al. (2022) investigated the importance of choosing the most appropriate loss function for training a model and then evaluating the generated forecasts. The authors considered a selection of error measures and concluded that *"forecasting methods which provide significantly better ranks are effectively identified as accurate by all accuracy measures"*. Additionally, they argued that using different criteria for model estimation and out-of-sample evaluation did not significantly impact the out-of-sample accuracy. This means that "good" forecasts, as defined by Kolassa (2020), are able to provide a similar level of accuracy as the sub-optimal ones.

Although insightful, the results of Koutsandreas et al. (2022) concern statistical, not economic, evaluation. In practice, the loss function that reflects revenues and – according to Gneiting (2011a) – should be minimized if the forecaster is to provide the optimal forecast may be complex and impossible (or at least very difficult) to implement in a predictive model. In particular, forecasts can be used in a multi-step decision-making process that takes into account the physical constraints of a system and the decision maker's expectations regarding profit maximization or risk reduction. This is a common scenario in short-term power markets (Hong et al., 2020). Forecasts of the next day's prices, load, or renewable generation are used to make informed decisions about when to buy or sell electricity and in which market (Chitsaz et al., 2018; Maciejowska et al., 2021; Marcjasz et al., 2023; Uniejewski and Weron, 2021). However, this decision-making process cannot be reduced to a simple functional, such as the mean or a quantile, of the predictive distribution.

Already in the 1990s, Murphy (1993) argued that the "goodness" of a forecast in a business forecasting setting should be evaluated by its consistency, quality, and value. *Consistency* is the agreement between a forecaster's internal judgments and actual forecasts. Although some authors adjust model-derived results using expert knowledge (Maciejowska and Nowotarski, 2016), in general, consistency cannot be measured directly because internal judgments are private. *Quality*, on the other hand, can be quantified using error metrics. Finally, *value* refers to the economic and other benefits from using predictions.

Yardley and Petropoulos (2021) emphasize that value includes not only the utility to the forecaster, but also computational (runtime, cloud computing fees, etc.) and opportunity costs (i.e., the resources wasted on implementing complex methods that decision makers ultimately do not use due to a lack of confidence in them). They also point out that the value of forecasts in markets of a financial nature can be evaluated

*Corresponding author; *email:* t.serafin@pwr.edu.pl

using different trading strategies, and claim that the literature generally reveals a disagreement between traditional error metrics and economic measures of performance. Therefore, Yardley and Petropoulos (2021) question the use of statistical error metrics to select forecasting methods.

In this study, we address this issue. We consider the most commonly used loss functions for evaluating both point and probabilistic forecasts, namely squared, absolute and percentage errors (Davydenko and Fildes, 2013; Hyndman and Koehler, 2006), and the asymmetric pinball loss (Gneiting and Raftery, 2007). In addition, we introduce a custom scoring function that severely penalizes predictions of the opposite sign, which we call the *directional loss function* (DLF). We then calibrate regression-type expert models (Ziel and Weron, 2018) to electricity prices from the German day-ahead market using the above error measures and evaluate the forecasts in economic terms, including the Sharpe ratio and profits. For the latter, we extend the trading strategy proposed by Uniejewski and Weron (2021) and later used, e.g., in Marcjasz et al. (2023), Nitka and Weron (2023) and Chęć et al. (2025), which involves the use of a *battery energy storage system* (BESS), and introduce thresholds that limit trading to only the most profitable opportunities. Then we examine the relationship between the choice of the loss function used to estimate the models and the economic value of the forecasts over a four-year test period spanning the Covid-19 pandemic and the Russian invasion of Ukraine. Furthermore, unlike other studies in the literature, we analyze how the operating costs of BESS affect the economic evaluation of the trading strategies.

The remainder of the paper is structured as follows. In Section 2 we present the models, the transformation used to preprocess regressors, loss functions and the dataset considered in this study. In Section 3 we describe the trading strategies, introduce economic measures used for the evaluation of forecasts and elaborate on the BESS operating costs. In Section 4 we describe the empirical findings and provide a detailed discussion on the performance of the proposed methods. Finally, in Section 5 we wrap up the results and provide directions for future research.

2. Methodology

2.1. Expert models

We focus on forecasting the day-ahead electricity prices using autoregressive models with exogenous variables inspired by well established literature benchmarks (Billé et al., 2023; Gaillard et al., 2016; Janczura and Wójcik, 2022; Maciejowska et al., 2021; Maciejowska and Nowotarski, 2016; Ziel and Weron, 2018). Since the trading strategy we use for the economic evaluation of forecasts in Section 3 requires selecting a pair of hours with the highest price difference, we can either (i) forecast all 24 prices and compute the implied price spreads for each pair of hours, or (ii) directly predict the spreads, like Abramova and Bunn (2020) and Maciejowska et al. (2019). In what follows, we consider both approaches.

Model 1 predicts the day-ahead price on day d and hour h using the regression:

$$\begin{aligned} DA_{d,h} = &\sum_{p=1}^{7} \beta_p DA_{d-p,h} + \beta_8 \overline{DA}_{d-1} + \beta_9 \underline{DA}_{d-1} \\ &+ \beta_{10} \hat{L}_{d,h} + \beta_{11} \hat{W}_{d,h} + \beta_{12} API2_{d-2} \\ &+ \beta_{13} TTF_{d-2} + \sum_{p=1}^{7} \beta_{p+13} D_p + \varepsilon_{d,h}, \end{aligned} \tag{1}$$

where $\hat{L}_{d,h}$ and $\hat{W}_{d,h}$ are the day-ahead load and wind generation forecasts for day d and hour h, $\overline{DA}_{d-1}$ and $\underline{DA}_{d-1}$ are the maximum and minimum prices from day $d-1$, $API2_{d-2}$ and TTF_{d-2} are the last known daily closing prices of API2 coal and TTF gas yearly futures contracts, and D_p are weekday dummies. See Section 2.3 for information on data sources.

Model 2 directly predicts the price spread between hours h_1 and h_2 using the following regression:

$$\begin{aligned} \Delta DA_{d,h_1,h_2} = &\sum_{p=1}^{7} \beta_p \Delta DA_{d-p,h_1,h_2} + \beta_8 \left(\overline{DA}_{d-1} - \underline{DA}_{d-1}\right) \\ &+ \beta_9 \Delta\hat{L}_{d,h_1,h_2} + \beta_{10} \Delta\hat{W}_{d,h_1,h_2} + \beta_{11} API2_{d-2} \\ &+ \beta_{12} TTF_{d-2} + \sum_{p=1}^{7} \beta_{p+12} D_p + \varepsilon_{d,h_1,h_2}, \end{aligned} \tag{2}$$

where $\Delta DA_{d,h_1,h_2} = 0.9 DA_{d,h_2} - \frac{1}{0.9} DA_{d,h_2}$ is the difference between day-ahead prices for hours h_1 and h_2 on day d after taking into account a 90% battery efficiency (see Section 3.1), and $\Delta\hat{L}_{d,h_1,h_2}$ and $\Delta\hat{W}_{d,h_1,h_2}$ are the differences of day-ahead load and wind generation forecasts for hours h_1 and h_2 on day d.

Moreover, like Ciarreta et al. (2022), Janczura (2025) and Uniejewski et al. (2018), we preprocess regressors with the *area hyperbolic sine*:

$$\operatorname{asinh}(z) = \log\left(z + \sqrt{z^2 + 1}\right), \tag{3}$$

where $z = \frac{1}{b}(x - a)$, a and b are respectively the mean and standard deviation of x in the calibration window C, and $x \in \{DA_{d,h}, \Delta DA_{d,h}, \overline{DA}_d, \underline{DA}_d, \hat{L}_{d,h}, \Delta\hat{L}_{d,h}, \hat{W}_{d,h}, \Delta\hat{W}_{d,h}, API2_d, TTF_d\}$ is the regressor. As Ziel and Weron (2018) remark, close to a the $\operatorname{asinh}(z)$ transform is almost linear, while large positive and negative values are pulled towards the center in a logarithmic way. To recover the price or price spread forecasts, we apply the hyperbolic sine to the generated predictions (see Narajewski and Ziel, 2020, for a more accurate inverse transformation). Since preprocessing the regressors improves the performance of the proposed methods with respect to all the metrics considered, but does not change the conclusions of the study, we decided, for the sake of parsimony, to report only the results for models using the asinh transformation.

2.2. Loss functions

The standard approach to estimating the parameters $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots)$ of a regression is *ordinary least squares* (OLS),

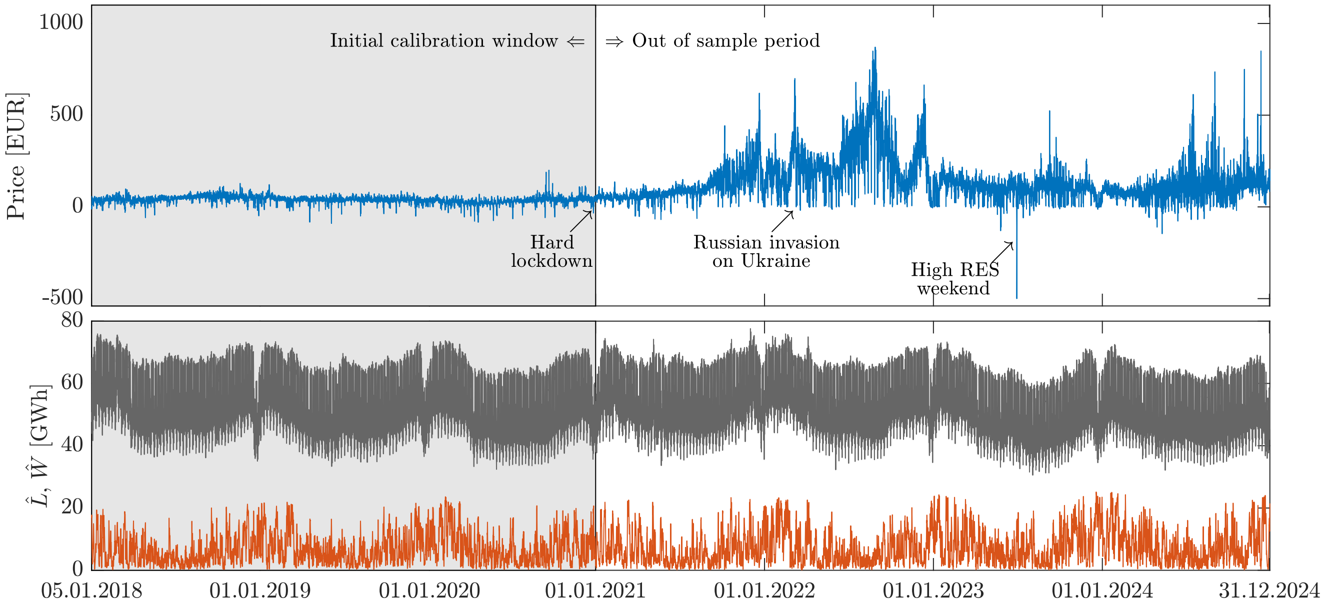

Figure 1: Day-ahead prices (blue trajectory; *top*), consumption forecasts (gray; *middle*) and wind generation forecasts (orange; *bottom*) from the German electricity market in the period from 05.01.2018 to 31.12.2024. The first 1092 days (05.01.2018–31.12.2020) are the initial calibration window C; each day the window is rolled forward by 24 hours. Three specific reference dates are marked with arrows: the first day of the hard lockdown in Germany (15.12.2020), the Russian invasion of Ukraine (24.02.2022), and the occurrence of extremely low prices due to low demand and high RES generation (on Sunday 02.07.2023 at 3 p.m. the price reached −500 EUR/MWh).

which minimizes the *mean squared error* (MSE) of the differences between the observations y_t and the predictions $\hat{y}_t$:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, \mathcal{L}(\hat{y}_t, y_t) = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left(\operatorname{mean}(y_t - \hat{y}_t)^2 \right), \tag{4}$$

where $\mathcal{L}(\cdot,\cdot)$ is the loss function and t spans the calibration window C; we are using here the simplified notation of Hyndman and Koehler (2006). We can obtain other model specifications – and consequently different forecasts – by minimizing different loss functions. In this paper, we consider the most popular loss functions used in the literature (Davydenko and Fildes, 2013; Hyndman and Koehler, 2006):

- the *mean absolute error* (MAE)

$$\mathcal{L}(\hat{y}_t, y_t) = \operatorname{mean}(|y_t - \hat{y}_t|), \tag{5}$$

- the *symmetric mean absolute percentage error* (sMAPE)

$$\mathcal{L}(\hat{y}_t, y_t) = \operatorname{mean} \frac{|y_t - \hat{y}_t|}{|y_t| + |\hat{y}_t|}. \tag{6}$$

When using the MAE, the parameter estimates are less affected by outliers than for the standard OLS procedure. On the other hand, the advantage of using the sMAPE as a percentage-based measure is that it is invariant to the scale of the underlying data values. However, despite its name, it is asymmetric in the sense that it penalizes overpredictions more than underpredictions.

Note that the so-called *relative* measures (Koutsandreas et al., 2022; Lago et al., 2021), which normalize the MAE or RMSE by the respective (out-of-sample) error of a naive forecast, e.g., RelMAE = MAE/MAE_{naive}, are not considered here. The reason is that MAE_{naive} does not depend on $\boldsymbol{\beta}$, thus the RelMAE yields the same $\hat{\boldsymbol{\beta}}$ as the MAE. For the same reason, we not consider the *mean absolute scaled error* (MASE) introduced by Hyndman and Koehler (2006), which normalizes the MAE by the in-sample MAE of a naive forecast. Furthermore, we do not report the results for the *geometric mean absolute error* (GMAE; Davydenko and Fildes, 2013), which replaces the arithmetic mean with the geometric one in Eq. (5). The reason is that the performance of the GMAE-estimated models turned out to be almost identical (but slightly inferior) to that of the MAE-estimated regressions.

Apart from loss functions associated with point forecasts, we additionally consider the *pinball loss* (PL), an asymmetric scoring function for evaluating quantile predictions (Gneiting and Raftery, 2007; Grushka-Cockayne et al., 2017):

$$\mathcal{L}(\hat{y}_t, y_t) = \operatorname{mean}\Big((\mathbb{1}_{y \le \hat{y}_t} - q)(y_t - \hat{y}_t) \Big), \tag{7}$$

where $\mathbb{1}_X$ is the indicator function of X and q is the quantile level. As Gneiting (2011b) argues, quantile forecasts can be considered as optimal point forecasts when dealing with different economic costs of underprediction and overprediction. This particular situation arises, for example, in one of the most well-known problems in operations management – the *newsvendor problem* – when there are different costs associated with having too much inventory and having too little inventory. The PL is also the function that is minimized in quantile regression, a technique used to postprocess point forecasts to derive predictive distributions (Lipiecki et al., 2024).

Finally, we introduce a custom scoring function that attempts to better reflect revenues from the trading strategies, see Section 3.1. The *directional loss function* (DLF) linearly penalizes price

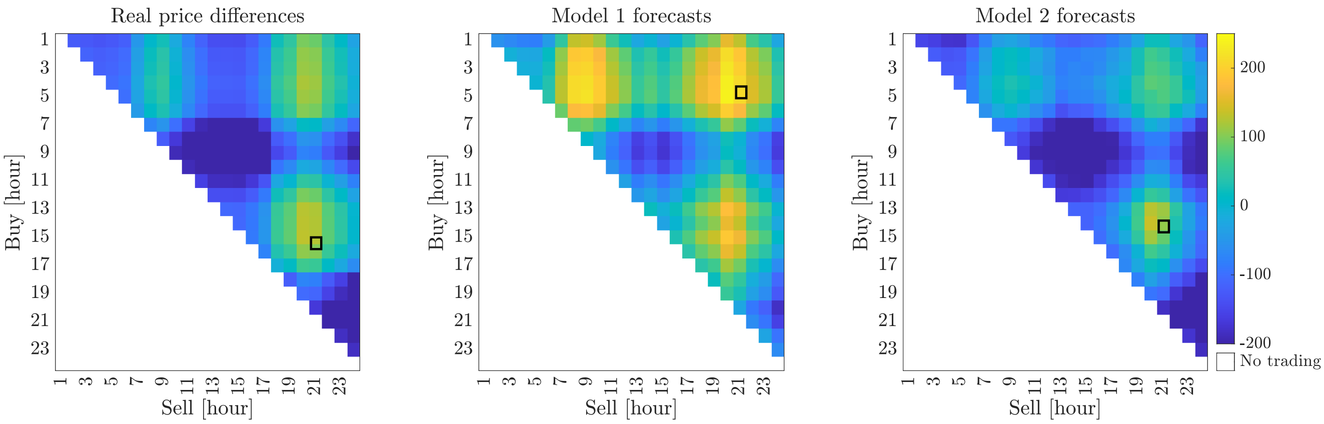

Figure 2: Real price differences (*left*), differences of predicted prices from Model 1 (*middle*) and predicted price spreads from Model 2 (*right*) on 09.12.2023. Forecasts are obtained by minimizing the mean absolute error (MAE). Black rectangles mark the pair of hours with the highest price difference (*from left to right*): (15, 20) for the actual prices which are selected by the Crystal Ball strategy, (5, 20) for Model 1, and (14, 20) for Model 2.

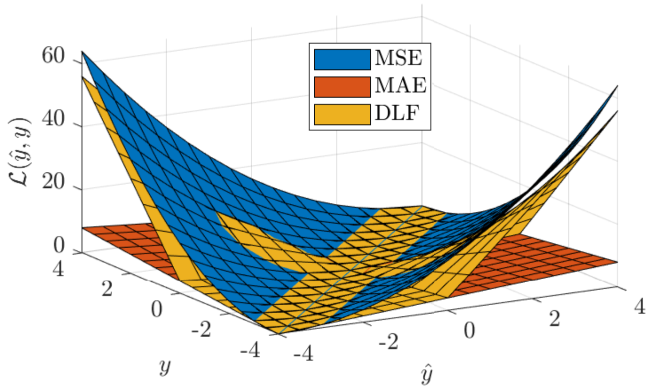

Figure 3: Three main loss functions considered in this study: MSE, Eq. (4), MAE, Eq. (5), and the directional loss function (DLF), Eq. (8). Note that the DLF is used only for price spread predictions, while the other two are also used for price forecasts themselves.

spread predictions of the same sign, but adds a quadratic cost to predictions of the opposite sign:

$$\mathcal{L}(\hat{y}_t, y_t) = \text{mean}\Big(|y_t - \hat{y}_t| + \max(0, -3y_t\hat{y}_t)\Big). \quad (8)$$

The first term is the MAE loss, see Eq. (5), while the second resembles the so-called *hinge loss* commonly used for training classifiers in machine learning (Rosasco et al., 2004). The latter penalizes when the direction of the predicted spread differs from the actual one, which would lead to buying electricity at a higher price and selling it at a lower price. Note that the coefficient of 3 in the second term is chosen arbitrarily; for this particular value, the DLF coincides with the MSE when the arguments are 2 and −2. For a comparison of the MSE, MAE and DLF see Figure 3.

2.3. Data

We consider a dataset from the German day-ahead market, which includes electricity prices ($DA_{d,h}$), as well as day-ahead load ($\hat{L}_{d,h}$) and wind generation ($\hat{W}_{d,h}$) forecasts, see Figure 1. The latter are supplemented by the last known daily closing prices of the yearly futures contracts of API2 coal ($API2_{d-2}$) and TTF gas (TTF_{d-2}). All data series are publicly available and have been downloaded from ENTSO-E (`https://transparency.entsoe.eu`; electricity prices, load forecasts, onshore + offshore wind generation forecasts) and Investing.com (`https://www.investing.com/`; closing prices of coal and natural gas futures).

The data span from 5.01.2018 to 31.12.2024, covering both the Covid-19 pandemic and the 2022 energy crisis triggered by the Russian invasion of Ukraine. Electricity prices in this period ranged from very low values and frequent negative spikes (i.e., a hallmark of the German market; Hagfors et al., 2016), to extremely volatile and high values. This dataset provides a unique testing ground to examine the performance of the forecasting methods under extreme price regimes.

2.4. Rolling window scheme

Like the majority of EPF studies, we consider a rolling scheme in which the calibration window is shifted forward by 24 hours at the end of the day. Although some authors use expanding windows (only the endpoint is shifted forward; also called the 'recursive scheme') or a combination of rolling and expanding windows (Pesaran and Pick, 2011), the latter have a major disadvantage. That is, the two most commonly used predictive ability tests (Diebold-Mariano and Giacomini-White) require that the calibration window does not grow in size (Giacomini and Rossi, 2013). This rules out expanding windows, but allows for both fixed (often used for hyperparameter selection in computationally demanding machine learning models, see Lago et al., 2021) and rolling windows.

In our setup, we use a rolling calibration window C of $3 \times 364 = 1092$ days; for shorter, one- or two-year windows, the parameter estimates of Model 2 were unstable. The first 1092 days (from 05.01.2018 to 31.12.2020, see Figure 1) of the dataset are used to calibrate the models and obtain predictions for 01.01.2021. The calibration window is then rolled forward by one day (05.01.2018 → 06.01.2018 and 31.12.2020

→ 01.01.2021) and the forecasts for 02.01.2021 are calculated. This procedure is repeated for each day in the out-of-sample period, i.e., from 01.01.2021 to 31.12.2024.

3. Trading strategies

3.1. Threshold-based strategy for predicted price spreads

We build on the strategy introduced by Uniejewski and Weron (2021) and later used, e.g., in Marcjasz et al. (2023), Nitka and Weron (2023) and Chęć et al. (2025), which mimics the day-to-day operations of electricity market participants. The strategy involves operating a 1.25 MW battery energy storage system (BESS) and using it to perform a single charge-discharge cycle during the day. For technical reasons, the battery cannot be discharged below 20% of the nominal capacity, i.e., 0.25 MWh, and its efficiency of charging as well as discharging is ca. 90% (≈ 80% for a single cycle; Sikorski et al., 2019). The approach requires selecting a pair of hours – h_1 with the lowest and $h_2 > h_1$ with the highest predicted price – for buying (→ charging the BESS) and then selling (→ discharging) the electricity in the day-ahead market, see Figure 2 for a sample illustration.

Compared to the strategy originally proposed by Uniejewski and Weron (2021) and its later variants, here we introduce a price difference *threshold* that limits trading to the most profitable opportunities. Namely, after selecting the pair of hours (h_1, h_2), we determine whether the expected profit from the trading activity, based on the generated forecasts, exceeds a predefined threshold:

$$\widehat{\Delta DA}_{d,h_1,h_2} > T. \tag{9}$$

If the above condition is met, we place price-taker buy and sell orders in the day-ahead market, otherwise we do not trade and report a profit of 0 EUR for the day. Note that while Model 2 takes into account battery efficiency, see Section 2.1, for Model 1 we have to compute: $\widehat{\Delta DA}_{d,h_1,h_2} = 0.9\widehat{DA}_{d,h_2} - \frac{1}{0.9}\widehat{DA}_{d,h_1}$.

3.2. The Crystal Ball, Persistent and Historical strategies

The trading strategy proposed in Section 3.1 has a maximum profit that can be achieved – it would only be possible if we had a crystal ball that could make a perfect prediction. For the purpose of forecast evaluation and to put the results of the considered approaches into perspective, in Section 4 we report the performance of the *Crystal Ball* strategy (abbreviated *Crystal* in the figures):

$$\widehat{\Delta DA}_{d,h_1,h_2} = \Delta DA_{d,h_1,h_2}. \tag{10}$$

Additionally, we include two strategies that do not rely on any forecasts. In the *Persistent* strategy (abbreviated *Persist* in the figures; also called 'white noise' strategy), we assume that the price spread between two hours for day d is the actual value from the previous day:

$$\widehat{\Delta DA}_{d,h_1,h_2} = \Delta DA_{d-1,h_1,h_2}. \tag{11}$$

In the *Historical* strategy (abbreviated *Hist* in the figures), we set the trading hours to those that exhibited the lowest and the highest prices on average in the calibration window C:

$$\begin{cases} h_1 = \operatorname{argmin}_h \sum_{d\in C} DA_{d,h}, \\ h_2 = \operatorname{argmax}_h \sum_{d\in C} DA_{d,h}. \end{cases} \tag{12}$$

3.3. Evaluation

We evaluate the results of the trading strategies using three economic measures: total profit (TP), profit per trade (PPT) and Sharpe ratio (SR). The *total profit* for model m and threshold T is given by:

$$\mathrm{TP}^m(T) = \sum_d P_d^m(T), \tag{13}$$

where $P_d^m(T)$ is the profit (or loss) for day d from trading based on forecasts of model m with threshold T, and the sum ranges over the whole test period from 01.01.2021 to 31.12.2024. Note that m refers to a combination of the model class (Model 1 or 2, one of the benchmarks) and the loss function.

Depending on the threshold, the total number of trading days may vary between the models and thus the TP may favor quantity over quality. Since more frequent transactions lead to higher transaction costs, this can significantly reduce revenues. Therefore, from a practical perspective, the *profit per trade*:

$$\mathrm{PPT}^m(T) = \frac{\mathrm{TP}^m(T)}{\#trades}, \tag{14}$$

where the denominator is the total number of days in which model m traded given threshold T, is a more important metric. As Maciejowska et al. (2024) and Marcjasz et al. (2023) emphasize, the PPT can be easily adjusted for transaction and BESS operating costs. In this study, we ignore transaction costs because they would be at least an order of magnitude smaller than BESS costs for most participants (see Section 3.4).

As a third metric, we use the *Sharpe ratio* (Agakishiev et al., 2025; Kath and Ziel, 2018):

$$\mathrm{SR}^m(T) = \frac{\frac{1}{\#trades}\sum_{d^*} P_{d^*}^m(T)}{\sigma}, \tag{15}$$

where d^* marks days when trading happened and σ is the standard deviation of the obtained profits.

3.4. BESS operating cost

In most studies on electricity price forecasting that examine profits from trading, the authors do not consider the costs and commissions involved. Only a handful take into account transaction or liquidity costs, mainly in the context of intraday or continuous-time markets (Narajewski and Ziel, 2022; Kuppelwieser and Wozabal, 2021). However, to the best of our knowledge, none have analyzed how the operating costs of BESS affect the profits from intraday trading in day-ahead markets.

The methods used to calculate the cost of energy storage systems vary widely between studies. Few consider only the initial cost of building the system, while others calculate the so-called *life-cycle cost* (LCC), which includes the costs of operation and maintenance, parts replacement, and eventual decommissioning. Rahman et al. (2020) discuss this issue in detail, pointing out that for new technologies such as BESS, there is a high degree of uncertainty in the cost estimates, mostly due to different assumptions of the system parameters considered in the studies.

Additionally, the perspective of the company utilizing the BESS is crucial – the cost estimate will be very different if we compare building the BESS from scratch versus operating an existing infrastructure. Since the exact pricing of BESS costs is not the primary focus of this paper, for simplicity we assume that the cost of operating BESS is 100 EUR per single charge-discharge cycle. We derive this value by considering the approximate cost of battery cells (excluding installation and construction costs) in a utility-scale lithium-ion battery system (NREL, 2024), at the level of 300 EUR/kWh. We scale it to a 1MWh battery size and assume a battery lifetime of 3000 cycles (da Silva Lima et al., 2021). According to Cole and Karmakar (2023), the cost of operating BESS is expected to decrease rapidly over the next decade. Therefore, to complement the outlook, we additionally consider a 50 EUR charge-discharge cycle cost when reporting the results of the economic evaluation. Finally, to emphasize the importance of considering the operating costs of BESS, we show how different the conclusions can be if such costs are ignored.

4. Economic results

The trading frequency, total trading profits, profits per trade and Sharpe ratios for the considered models (to be more precise: loss functions used to estimate the models) are shown in Figures 4 and 5. Due to significant changes in electricity price dynamics during the test period, we report results separately for each of the years 2021, 2022, 2023, and 2024. This allows us to evaluate the performance of the models and loss functions under different market conditions.

4.1. Trading frequency

The total number of BESS charge-discharge cycles and the corresponding trades made for a given model and loss function are shown in the top row of Figure 4 as a percentage of all possible trades, i.e., the number of days per year. As can be seen, the PL-estimated models for high quantile levels (→ dark gray shaded areas) and the sMAPE-estimated models (→ solid and dashed purple lines) trade most frequently. At the same time, the MSE loss function (→ solid and dashed blue lines) leads to more frequent trading than the MAE (→ solid and dashed orange lines) in a 'calm' year 2023, while the opposite is true in a 'volatile' year 2022. In 2021 and 2024, both models trade similarly frequently. The DLF-estimated models (→ yellow line) and the PL-estimated models for low quantile levels (→ light gray shaded areas) are consistently among the least frequently trading ones, indicating that they are more discriminating in their choice of trading opportunities than the MSE and MAE.

By construction, see Section 3.2, if the Crystal (Ball) benchmark (→ green line) traded yesterday, then the Persist(ent) strategy (→ dash-dotted red line) will trade today, and the number of trading days in a year will differ by no more than one day for them. They both traded relatively often in 2021. In the remaining years, they traded frequently for high thresholds and infrequently for low thresholds. Finally, the Hist(orical) strategy (→ dotted red line) traded rarely in the considered years.

4.2. Total trading profits

Looking first at the most realistic case given today's prices of BESS (2nd row of Figure 4; note that y-axes differ between rows and columns), we can see that in 2023-2024, and to some extent in 2021, the less frequently trading models – the DLF (→ yellow line) and pinball loss for low quantiles (→ light gray areas), followed by the MAE (→ orange lines) and MSE (→ blue lines) – were the top performers when each charge-discharge cycle costs 100 EUR. However, in 2022, when the energy crisis reached its climax and electricity prices soared, the MSE-estimated price spread model (→ solid blue line) led to the highest revenues. Regarding the performance of Model 1 (→ dashed lines) vs. Model 2 (→ solid lines), for the MSE the latter performed better, while for the MAE the evidence was mixed – the price spread model was better in 2021-2022, but slightly worse in 2023-2024. Similarly, for the worst performing sMAPE-estimated models (→ purple lines), the evidence was mixed.

Interestingly, with a BESS operating cost of 100 EUR per cycle, trading was barely profitable in 2021 and 2023 for the best approaches considered, and would be extremely unprofitable without proper model and threshold selection.

Clearly, the inclusion of costs in the evaluation reveals the need to introduce thresholds in the trading strategy. For lower thresholds, even the best performing models generate significant losses. With such high trading costs, traders need to consider only the most profitable opportunities to generate positive returns.

For the charge-discharge cost of 50 EUR (3rd row of Figure 4) there is a trade-off between the number of trades (lower values of T) and the profitability of the trading opportunities (higher values of T). Initially, the total profit increases with the value of the threshold by including only more profitable opportunities – up to $T \approx 50$ EUR – and then begins to decrease as the total number of trades begins to decrease. Like in the top row in Figure 4, also for the twice lower operating cost the DLF price spread model (→ yellow line) yielded the highest revenues in 2023 and 2024, followed by the MAE- and MSE-estimated regressions; this time the pinball loss for low quantiles (→ light gray areas) performs poorly, especially for high T's. Similarly as for the BESS cost of 100 EUR, also in this case the MSE-estimated price spread model (→ solid blue line) excels in 2022. In 2021 there is no clear winner, the DLF-, MAE- and MSE-estimated regressions perform equally well for the lowest thresholds yielding the highest revenues. Again, the evidence is mixed when comparing Model 1 with Model 2.

Finally, if we assume that the charge-discharge cycle has no cost (bottom row of Figure 4), the models that trade more frequently are the top performers. In addition, trading with a higher threshold limits the total profit in each year of the test period. This is not surprising, since in this case trading is free

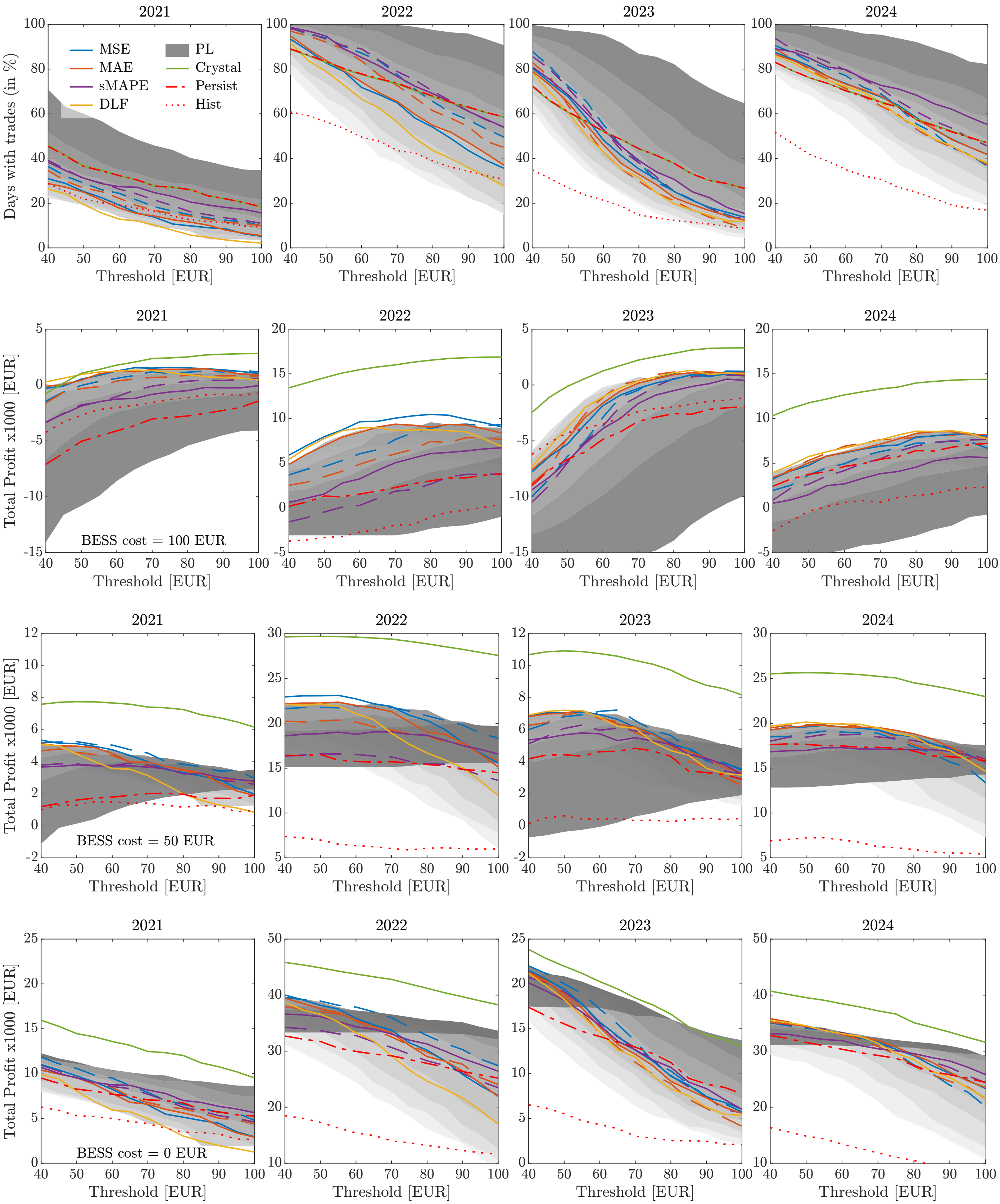

Figure 4: *Top row*: The percent of days with trades for different model classes and loss functions, independently for each of the test period years (*left to right*). *Remaining rows*: The total profits, see Eq. (13), for different model classes and loss functions, independently for each of the test period years and three levels of BESS operating cost per charge-discharge cycle: 100, 50 and 0 EUR (*top to bottom*). The solid lines represent the results of Model 2, which predicts price spreads, while the dashed lines of corresponding colors represent the results of Model 1, which predicts prices themselves. The shaded areas mark the performance of Model 2 with the pinball loss (PL) and q ranging from 10% (light gray) to 90% (dark gray). Note that y-axes differ in each row: 2021 and 2023 use different values than 2022 and 2024.

and even the potentially less profitable opportunities increase the total revenue.

4.3. Profits per trade

In the top row of Figure 5 we present the profit per trade for different model classes and loss functions, independently for each of the test period years and BESS operating cost of 100 EUR. Note that for operating costs of 50 and 0 EUR the profit per trade will be exactly 50 and 100 EUR higher, respectively, since changing the cost only shifts the results by a constant. Similarly to the total profit depicted in Figure 4, years 2022 and 2024 provided the most profitable trading opportunities with an average profit per trade of ca. 85 and 55 EUR, respectively, for the top performing models and high values of T. Unlike the total profit, $\text{PPT}^m(T)$ increases with T for all of the considered approaches.

In year 2021 the DLF- and MSE-estimated regressions performed the best, closely followed by the MAE- and PL-estimated for low quantiles. Interestingly, in 2021 the models were able to outperform the Crystal Ball benchmark. Because of the threshold and imperfect forecasts, the models traded less frequently, omitting some of the less profitable situations. For instance, for $T = 50$ and BESS operating cost of 100 EUR, if the Crystal Ball forecast was 55 EUR and the model forecast 49 EUR, the model would not trade and 'gain' over the Crystal Ball which effectively lost 45 EUR.

In the most profitable year 2022, the DLF, MSE and particularly the pinball loss for low quantiles outperformed all other approaches, while the MAE performed relatively worse compared to the previous year. Similarly to 2021, the MSE- and MAE-estimated Model 1 (→ dashed lines) performed significantly worse than its counterpart directly predicting price spreads (→ solid lines).

In 2023 and 2024, the DLF and the pinball loss for low quantiles remain the strongest performers and the MAE trades places with the MSE as the third best approach. This time, similarly to the total profit for the BESS cost of 100 EUR, Model 1 performs slightly better than Model 2 for the MAE. For the MSE there is no clear winner in 2023 and 2024, Model 2 outperforms Model 1 by a slim margin.

4.4. Sharpe ratios

In the bottom row of Figure 5 we plot the Sharpe ratios for the models and scenarios considered. Similarly to the profits per trade, the Sharpe ratios generally increase with increasing thresholds. The highest values of $\text{SR}^m(T)$ were obtained in 2021, likely due to the much lower standard deviation of the profits, see Eq. (15). Here, the price spread regressions (i.e., Model 2) estimated using DLF, MSE and MAE outperformed all other approaches, even – what may seem surprising – the Crystal Ball benchmark. Alongside PL-estimated models for low quantiles, they were again the top performers in 2022. For the calmer year that followed, the DLF-estimated model was the best. For 2023 and 2024, the MSE-estimated regressions were outperformed by their MAE-estimated counterparts. However, the DLF-estimated model performed even better. On the other hand, sMAPE-estimated regressions consistently underperformed for all threshold values. Additionally, Model 2 generated forecasts that for MSE and MAE performed better in terms of the Sharpe ratio than the forecasts of Model 1 (for MAE only for 2021 and 2022).

5. Conclusions

In this study, we calibrated regression-type expert models to electricity prices from the German day-ahead market using a range of well-known loss functions (MAE, MSE, sMAPE, PL) and a newly-introduced *directional loss function* (DLF). We evaluated the forecasts in economic terms, including the Sharpe ratio and profits (total and per trade), over an out-of-sample test period of four years (2021-2024). We extended the trading strategy of Marcjasz et al. (2023) and introduced thresholds that limit trading to only the most profitable moments. Additionally, unlike other studies, in the evaluation of the proposed methods, we considered realistic costs associated with operating a *battery energy storage system* (BESS), which currently amount to ca. 100 EUR per charge-discharge cycle.

Due to significant changes in electricity price dynamics during the test period, we reported results separately for each year. In 2021, the DLF-estimated models performed well in terms of total profits and outperformed all other approaches in terms of profits per trade and Sharpe ratios for almost all thresholds. In the most volatile, yet most profitable year 2022, the MSE-estimated regressions performed extremely well in terms of total profits, producing similar profits per trade and Sharpe ratios to DLF- and PL-estimated (for low quantiles) models. However, in the calmer year 2023 and the moderately volatile year 2024, again the DLF- and PL-estimated (for low quantiles) models were the top performers. Although MSE-estimated models outperformed their MAE-estimated counterparts in 2021 and 2022, the opposite could be observed in 2023 and 2024.

Overall, our results provide evidence that the newly introduced *directional loss function* (DLF) and the pinball loss for low quantiles exhibit the most consistent performance, being the best-performing models in terms of profits per trade (a measure independent of the trading volume) and risk-adjusted profits (as measured by the Sharpe ratio). We argue that while the total profit is commonly used in the literature, it favors frequent trading and could lead to misleading conclusions, especially when the BESS operating costs are not considered. Therefore, we recommend using profits per trade and Sharpe ratios to assess the economic value of forecasts.

By examining two similarly structured regression models, we can conclude that directly predicting price spreads between a pair of hours, as opposed to predicting prices for each hour separately, produces forecasts that yield generally higher profits as well as superior profits per trade and risk-adjusted profits for the MAE and MSE loss functions, especially in an extremely volatile price environment as observed in 2022.

In addition, we find that the introduced threshold-based strategy allows us to increase the performance with respect to all considered economic measures when BESS operating costs are

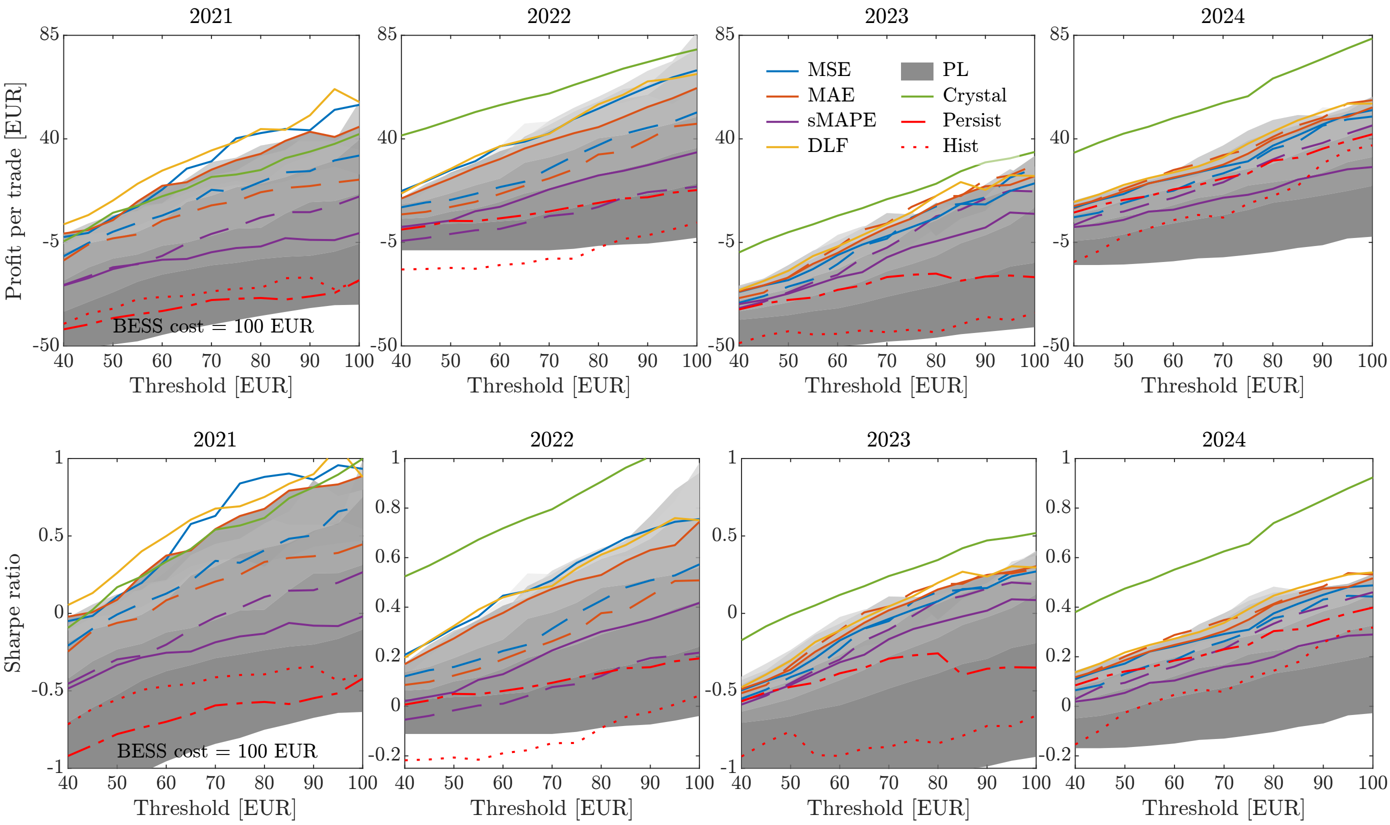

Figure 5: Profits per trade (*top*) and Sharpe ratios (*bottom*) for different model classes and loss functions, independently for each of the test period years (*left to right*) and BESS operating cost of 100 EUR. Note that with operating costs of 50 and 0 EUR, the profit per trade will be exactly 50 and 100 EUR higher, respectively. This does not hold for the Sharpe ratio, but the results are qualitatively the same for the different operating costs. The line styles are the same as in Figure 4. Note that y-axes differ in the bottom row: 2021 and 2023 use different values than 2022 and 2024.

taken into account. We recommend that these costs are not ignored in the economic evaluation of forecasts, as this may lead to a suboptimal model choice.

Our study can be further expanded in several directions. First, more advanced and accurate forecasting models can be considered, e.g., LASSO-estimated autoregressions or deep neural networks (LEAR, DNN; Lago et al., 2021; Uniejewski, 2024). Second, the proposed framework can be adapted to the joint modeling of prices for the 24 hours of the day, e.g., using vector autoregressions (VAR; Maciejowska, 2022; Ziel and Weron, 2018), machine learning models with multiple outputs (Lago et al., 2018) or functional autoregressions (FAR; Chen and Li, 2017), to minimize loss functions directly related to trading profits. Finally, since energy commodities tend to exhibit strong seasonal patterns (Lisi and Pelagatti, 2018; Maciejowska, 2020), seasonal decomposition could be used to improve the accuracy of price and price spread forecasts (Chęć et al., 2025).

Acknowledgements

This work was partially supported by the Ministry of Science and Higher Education (MNiSW, Poland) through Diamond Grant No. 0009/DIA/2020/49 (to T.S.) and the National Science Center (NCN, Poland) through grant No. 2018/30/A/HS4/00444 (to R.W.). The authors would like to thank David Obst (EDF R&D) and Tomasz Sikorski (WUST) for discussions on BESS operating costs, as well as the participants of Energy Finance Italia EFI9 and the Energy Finance Christmas EFC24 workshop for their constructive comments and suggestions.

References

Abramova, E., Bunn, D., 2020. Forecasting the intra-day spread densities of electricity prices. Energies 13, 687.

Agakishiev, I., Härdle, W.K., Kopa, M., Kozmik, K., Petukhina, A., 2025. Multivariate probabilistic forecasting of electricity prices with trading applications. Energy Economics 141, 108008.

Billé, A., Gianfreda, A., Del Grosso, F., Ravazzolo, F., 2023. Forecasting electricity prices with expert, linear, and nonlinear models. International Journal of Forecasting 39, 570–586.

Chęć, K., Uniejewski, B., Weron, R., 2025. Extrapolating the long-term seasonal component of electricity prices for forecasting in the day-ahead market. Journal of Commodity Markets 37, 100449. doi:`https://doi.org/10.1016/j.jcomm.2024.100449`.

Chen, Y., Li, B., 2017. An adaptive functional autoregressive forecast model to predict electricity price curves. Journal of Business and Economic Statistics 35, 371–388.

Chitsaz, H., Zamani-Dehkordi, P., Zareipour, H., Parikh, P., 2018. Electricity price forecasting for operational scheduling of behind-the-meter storage systems. IEEE Transactions on Smart Grid 9, 6612–6622.

Ciarreta, A., Muniain, P., Zarraga, A., 2022. Do jumps and cojumps matter for electricity price forecasting? Evidence from the German-Austrian day-ahead market. Electric Power Systems Research 212, 108144.

Cole, W., Karmakar, A., 2023. Cost projections for utility-scale battery storage: 2023 update. U.S. Department of Energy, Technical Report OSTI 1984976.
Davydenko, A., Fildes, R., 2013. Measuring forecasting accuracy: The case of judgmental adjustments to SKU-level demand forecasts. International Journal of Forecasting 29, 510–522.
Gaillard, P., Goude, Y., Nedellec, R., 2016. Additive models and robust aggregation for GEFCom2014 probabilistic electric load and electricity price forecasting. International Journal of Forecasting 32, 1038–1050.
Giacomini, R., Rossi, B., 2013. Forecasting in macroeconomics, in: Hashimzade, N., Thornton, M. (Eds.), Handbook of Research Methods and Applications on Empirical Macroeconomics. Edward Elgar, pp. 381–407.
Gneiting, T., 2011a. Making and evaluating point forecasts. Journal of the American Statistical Association 106, 746–762.
Gneiting, T., 2011b. Quantiles as optimal point forecasts. International Journal of Forecasting 27, 197–207.
Gneiting, T., Raftery, A., 2007. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association 102, 359–378.
Grushka-Cockayne, Y., Lichtendahl Jr, K.C., Jose, V.R.R., Winkler, R.L., 2017. Quantile evaluation, sensitivity to bracketing, and sharing business payoffs. Operations Research 65, 712–728.
Hagfors, L., Kamperud, H., Paraschiv, F., Prokopczuk, M., Sator, A., Westgaard, S., 2016. Prediction of extreme price occurrences in the German day-ahead electricity market. Quantitative Finance 16, 1929–1948.
Hastie, T., Tibshirani, R., Wainwright, M., 2015. Statistical Learning with Sparsity: The Lasso and Generalizations. CRC Press.
Hong, T., Pinson, P., Wang, Y., Weron, R., Yang, D., Zareipour, H., 2020. Energy forecasting: A review and outlook. IEEE Open Access Journal of Power and Energy 7, 376–388.
Hyndman, R.J., Koehler, A.B., 2006. Another look at measures of forecast accuracy. International Journal of Forecasting 22, 679–688.
Janczura, J., 2025. Expectile regression averaging method for probabilistic forecasting of electricity prices. Computational Statistics doi:10.1007/s00180-024-01508-y.
Janczura, J., Wójcik, E., 2022. 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, 106015.
Kath, C., Ziel, F., 2018. The value of forecasts: Quantifying the economic gains of accurate quarter-hourly electricity price forecasts. Energy Economics 76, 411–423.
Kolassa, S., 2020. Why the "best" point forecast depends on the error or accuracy measure. International Journal of Forecasting 36, 208–211.
Koutsandreas, D., Spiliotis, E., Petropoulos, F., Assimakopoulos, V., 2022. On the selection of forecasting accuracy measures. Journal of the Operational Research Society 73, 937–954.
Kuppelwieser, T., Wozabal, D., 2021. Liquidity costs on intraday power markets: Continuous trading versus auctions. Energy Policy 154, 112299.
Lago, J., De Ridder, F., De Schutter, B., 2018. Forecasting spot electricity prices: Deep learning approaches and empirical comparison of traditional algorithms. Applied Energy 221, 386–405.
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. Applied Energy 293, 116983.
Lipiecki, A., Uniejewski, B., Weron, R., 2024. Postprocessing of point predictions for probabilistic forecasting of day-ahead electricity prices: The benefits of using isotonic distributional regression. Energy Economics 139, 107934.
Lisi, F., Pelagatti, M.M., 2018. Component estimation for electricity market data: Deterministic or stochastic? Energy Economics 74, 13–37.
Maciejowska, K., 2020. Assessing the impact of renewable energy sources on the electricity price level and variability – a quantile regression approach. Energy Economics 85, 104532.
Maciejowska, K., 2022. Portfolio management of a small RES utility with a structural vector autoregressive model of electricity markets in Germany. Operations Research and Decisions 32, 75–90.
Maciejowska, K., Nitka, W., Weron, T., 2019. Day-ahead vs. intraday – forecasting the price spread to maximize economic benefits. Energies 12, 631.
Maciejowska, K., Nitka, W., Weron, T., 2021. Enhancing load, wind and solar generation for day-ahead forecasting of electricity prices. Energy Economics 99, 105273.
Maciejowska, K., Nowotarski, J., 2016. A hybrid model for GEFCom2014 probabilistic electricity price forecasting. International Journal of Forecasting 32, 1051–1056.
Maciejowska, K., Serafin, T., Uniejewski, B., 2024. Probabilistic forecasting with a hybrid factor-qra approach: Application to electricity trading. Electric Power Systems Research 234, 110541. doi:https://doi.org/10.1016/j.epsr.2024.110541.
Marcjasz, G., Narajewski, M., Weron, R., Ziel, F., 2023. Distributional neural networks for electricity price forecasting. Energy Economics 125, 106843.
Murphy, A., 1993. What is a good forecast? an essay on the nature of goodness in weather forecasting. Weather & Forecasting 8, 281–293.
Narajewski, M., Ziel, F., 2020. Econometric modelling and forecasting of intraday electricity prices. Journal of Commodity Markets 19, 100107.
Narajewski, M., Ziel, F., 2022. Optimal bidding in hourly and quarter-hourly electricity price auctions: Trading large volumes of power with market impact and transaction costs. Energy Economics 110, 105974.
Nitka, W., Weron, R., 2023. Combining predictive distributions of electricity prices. Does minimizing the CRPS lead to optimal decisions in day-ahead bidding? Operations Research and Decisions 33, 103–116.
NREL, 2024. NREL - Utility-Scale Battery Storage. https://atb.nrel.gov/electricity/2023/2023/utility-scale_battery_storage (Page accessed on 31.05.2024).
Pesaran, M., Pick, A., 2011. Forecast combination across estimation windows. Journal of Business and Economic Statistics 29, 307–318.
Rahman, M.M., Oni, A.O., Gemechu, E., Kumar, A., 2020. Assessment of energy storage technologies: A review. Energy Conversion and Management 223, 113295.
Rosasco, L., De Vito, E., Caponnetto, A., Piana, M., Verri, A., 2004. Are loss functions all the same? Neural Computation 16, 1063–1076.
Sikorski, T., Jasinski, M., Ropuszynska-Surma, E., Weglarz, M., Kaczorowska, D., et al., 2019. A case study on distributed energy resources and energy-storage systems in a virtual power plant concept: Economic aspects. Energies 12, 4447.
da Silva Lima, L., Quartier, M., Buchmayr, A., Sanjuan-Delmás, D., Laget, H., et al., 2021. Life cycle assessment of lithium-ion batteries and vanadium redox flow batteries-based renewable energy storage systems. Sustainable Energy Technologies and Assessments 46, 101286.
Uniejewski, B., 2024. Regularization for electricity price forecasting. Operations Research and Decisions 34, 267–286.
Uniejewski, B., Weron, R., 2021. Regularized quantile regression averaging for probabilistic electricity price forecasting. Energy Economics 95, 105121.
Uniejewski, B., Weron, R., Ziel, F., 2018. Variance stabilizing transformations for electricity spot price forecasting. IEEE Transactions on Power Systems 33, 2219–2229.
Yardley, E., Petropoulos, F., 2021. Beyond error measures to the utility and cost of the forecasts. Foresight Q4, 36–45.
Ziel, F., Weron, R., 2018. Day-ahead electricity price forecasting with high-dimensional structures: Univariate vs. multivariate modeling frameworks. Energy Economics 70, 396–420.