

Vogler, Arne; Ziel, Florian

Working Paper

On the evaluation of binary event probability predictions in electricity price forecasting

HEMF Working Paper, No. 11/2019

Provided in Cooperation with:

University of Duisburg-Essen, Chair for Management Science and Energy Economics

Suggested Citation: Vogler, Arne; Ziel, Florian (2019) : On the evaluation of binary event probability predictions in electricity price forecasting, HEMF Working Paper, No. 11/2019, University of Duisburg-Essen, House of Energy Markets & Finance (HEMF), Essen

This Version is available at:

<https://hdl.handle.net/10419/234065>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.



House of
Energy Markets
& Finance

On the Evaluation of Binary Event Probability Predictions in Electricity Price Forecasting

HEMF Working Paper No. 11/2019

by

*Arne Vogler,
and
Florian Ziel*

UNIVERSITÄT
DUISBURG
ESSEN

Open-Minded

Abstract

In this paper we present an evaluation framework for predictions of binary events in probabilistic electricity price forecasting. It employs the MSE-equivalent QPS together with the DM test and allows for further insights about deficiencies of the considered models. Additionally, techniques from the field of classification are considered, which extend our framework and are particularly suited for the evaluation of predictions of rare events. We consider binary events with direct applicability to a generator's daily decision making such as profitability of a pumped-hydro storage plant and evaluate the respective forecasts statistically. We show that the task of forecast evaluation can be simplified from assessing a multivariate distribution over prices to assessing a univariate distribution over a binary outcome, fully characterized by a single probability.

Keywords : Probabilistic Forecasting, Binary Predictions, Classification, Electricity Price Forecasting

JEL-Classification : C53, C38, Q47

Arne Vogler

House of Energy Markets and Finance
University of Duisburg-Essen, Germany
arne.vogler@wiwinf.uni-due.de

Florian Ziel

House of Energy Markets and Finance
University of Duisburg-Essen, Germany
florian.ziel@uni-due.de

The authors are solely responsible for the contents, which do not necessarily represent the opinion of the House of Energy Markets and Finance.

Contents

- List of Figures IV
- List of Tables IV
- Abbreviations V
- 1 Introduction 1
- 2 Binary Events 3
 - 2.1 Pumped-hydro Storage Plant Event 3
 - 2.2 Six Hours of Negative Electricity Prices Event 4
- 3 Event Probability Forecasting 5
 - 3.1 The Econometric Approach 5
 - 3.2 The Classification Approach 6
- 4 Evaluation of Event Probability Forecasts 7
 - 4.1 Quadratic Probability Score 8
 - 4.2 Evaluation of Classification Models 10
- 5 Empirical Results and Discussion 11
- 6 Conclusion 17
- References VII

List of Figures

1	Probability Time Series for Events	12
2	p-Values of DM Test	14
3	QPS Decompositions for Events	16

List of Tables

1	Specification Overview	6
2	Contingency Table	10
3	Quadratic Probability Score	14
4	AUROC	17
5	H-Measure	17

Abbreviations

AUROC	Area under Receiver Operating Characteristic Curve
CRPS	Continuous Ranked Probability Score
DM	Diebold Mariano
EPF	Electricity Price Forecasting
ES	Energy Score
FPR	False Positive Rate
GDP	Gross Domestic Product
MAE	Mean Absolute Error
MD	Murphy Decomposition
MPDFB	Mean Percentage Deviation from Best
MSE	Mean Square Error
OLS	Ordinary Least Squares
PEPF	Probabilistic Electricity Price Forecasting
PIT	Probability Integral Transform
PS	Pinball Score
PV	Photovoltaic
QPS	Quadratic Probability Score
QR	Quantile Regression
REL	Reliability
RES	Renewable Energy Sources
RESO	Resolution
RMSE	Root Mean Squared Error
ROC	Receiver Operating Characteristic
SVR	Support Vector Regression
TPR	True Positive Rate
UNC	Uncertainty
WBC	Within-Bin Covariance

WBV Within-Bin Variance

YD Yates Decomposition

1 Introduction

Electricity Price Forecasting (EPF) has become an indispensable part of energy companies' asset scheduling and short-term trading decision making. Since the advent of the EPF literature a plethora of forecasting models rooted in various fields such as econometrics and engineering have been studied. Among the considered evaluation measures for point forecasts both the Mean Absolute Error (MAE) and the Root Mean Squared Error (RMSE), paired with the Diebold Mariano (DM) test to establish statistically significant deviations in prediction performance, have proven to be the most popular (e.g., Jonsson et al. 2013, Weron 2014, Ziel 2016, Gürtler and Paulsen 2018). With the increasing infeed of intermittent Renewable Energy Sources (RES), relevance of demand response and the associated elevation of uncertainty in electricity prices, the literature has turned to Probabilistic Electricity Price Forecasting (PEPF) (e.g., Pape et al. 2017, Marcjasz et al. 2018, Nowotarski and Weron 2018). Forecasts are considered probabilistic if they constitute probability distributions over future events or quantities, Gneiting and Katzfuss 2014. In context of PEPF one aims to either characterize the full multivariate distribution of electricity prices or other associated characteristics such as marginal densities, intervals or specific quantiles (e.g., value-at-risk). The evaluation of such probabilistic forecasts is complicated by the fact that one only observes one realization of the underlying predicted distribution. The literature has established the evaluation paradigm of maximizing sharpness subject to calibration (e.g., Gneiting et al. 2007, Gneiting and Katzfuss 2014). The latter measures the correspondence between the probabilistic forecast and the realization. Calibration can be assessed using the Probability Integral Transform (PIT). For a forecast to be calibrated the PIT values should be uniformly identically and independently distributed between zero and one. Sharpness captures the concentration of the forecast distribution and can be assessed using the average width of prediction interval or using sharpness diagrams, which are less influenced by the presence of conditional heteroskedasticity, Gneiting et al. 2007. Additionally, calibration and sharpness may be assessed simultaneously using proper scoring rules. A score is considered proper if issuing the true underlying distribution as forecast distribution minimizes the score in expectation. In PEPF the Pinball Score (PS) and the Continuous Ranked Probability Score (CRPS) have proven to be popular (e.g., Jónsson et al. 2014, Pape et al. 2017, Nowotarski and Weron 2018). Yet, both the PS and the CRPS allow only for the evaluation of marginal distribution forecasts. The Energy Score (ES) can alternatively be used to assess the validity of a multivariate distribution forecast, Weron and Ziel 2018. All aforementioned evaluation techniques are statistical in nature. Consequently, some authors have noted that they are not ideal in the sense that they may not sufficiently reflect the associated economic consequences of preferring a particular forecasting model over another (e.g. Delarue et al. 2010, Zareipour et al. 2010, Mohammadi-Ivatloo et al. 2011, Doostmohammadi et al. 2017). Additionally, the notion of an optimal forecast may differ across applications and forecast users. Thus, rather than the forecast alone, the application to which it constitutes an input should form the basis of forecast evaluation. Mohammadi-Ivatloo et al. 2011 study the self-scheduling problem of an electricity generator under perfect-foresight

as well as predicted prices and assess the economic value of improved forecast accuracy using both statistical and profit-based measures. The notion of application-based forecast evaluation has also been recently introduced to the realm of electricity trading and risk management (e.g., Bunn et al. 2018, Kath and Ziel 2018). Furthermore, focusing on the eventual application of the forecast may also be extended to its actual derivation. A related idea is what Weron 2014 terms threshold forecasting. It also constitutes an approach motivated by the application of forecasts, as it may be sufficient to establish whether future prices surpass a specific threshold or fall into a predefined interval in the set of all possible prices for a generator's decision making (e.g., Zareipour et al. 2010). It is thus closely related to the work presented here. A price falling within a specific predefined interval constitutes an event, the occurrence of which could be predicted discretely or probabilistically with a certain probability assigned to its occurrence. Cast in this way it represents a univariate event; yet, by defining a threshold for a succession of prices turns the event into a multivariate event, the prediction and evaluation of which are the subject of the present paper. Probabilistic forecasting in the realm of EPF is thus far understood as forecasting the entire distribution of a continuous variable. As such PEPF mirrors a paradigm shift across a wide area of fields, where forecasting distributions across general types of variables becomes increasingly common. Similarly, probabilistic forecasts over binary events have a long tradition in other fields such as weather forecasting and economics, (e.g., Murphy 1988, Stephenson 2000, Diebold et al. 1998, Lahiri and Wang 2013). Such probabilistic forecasts of binary events have not yet been fully considered in the realm of PEPF and no coherent evaluation framework has been established. In this paper we fill the void by presenting methods to derive probabilistic forecasts over binary events and to evaluate such probabilistic forecasts. For illustration purposes we consider well-established expert models (e.g., Ziel and Weron 2018) as well as classification models. As the whole series of prices is of primary interest, the considered events are multivariate in nature, an issue that we take up in a later section. Moreover, by considering forecasts over binary events applicable to the daily decision making of a generator and evaluating them statistically, we bridge the gap between the strand of the literature concerned with the practical applicability and the forecasting literature rooted in the realm of statistics. From a practical standpoint it may often be sufficient to define a statistical event of interest motivated by a particular business problem, to forecast the associated probability and to evaluate this probability forecast, rather than to characterize and evaluate the entire multivariate distribution. The remainder of the paper is organized as follows: In section 2 we present the illustrative binary events considered and motivate their applicability. Section 3 introduces the econometric and classification models, while the proposed evaluation framework for event forecasts is presented in section 4. The results are presented and discussed in section 5. Section 6 concludes.

2 Binary Events

A binary event constitutes an outcome that assumes a value of zero or one. Generally, the event is considered to occur when it takes a value of one. In the context of the present paper the event's occurrence depends on the underlying path of day-ahead electricity prices. Specifically, we define two illustrative events motivated by the daily decision making of a generator and predict the probability of occurrence. Yet, as the occurrence of the event depends on all 24 day-ahead electricity prices of a given day and thus the full multivariate distribution, the underlying stochastics are multivariate. Consequently, we contribute to the literature by simplifying the task of forecast evaluation from assessing a multivariate distribution over continuous outcomes to assessing a univariate distribution over a binary outcome, fully characterized by a single probability and directly linked to the eventual application of the forecast.

2.1 Pumped-hydro Storage Plant Event

The first illustrative event is concerned with the profit from time spread arbitrage of a pumped-hydro storage plant exceeding a specified threshold on a given day. We refer to it as the 10k-pump event in the remainder of the present paper. A RES-based energy system is associated with increased importance of storage and flexibility options. Pumped-hydro storage plants constitute such a flexibility option and have thus received considerable attention in the literature (e.g., Brown et al. 2008, Steffen and Weber 2016, Braun and Hoffmann 2016). Steffen and Weber 2016 maintain that the traditional *modus operandi* in thermal-dominated electricity markets has been to pump at night and to turbine around noon. However, the economic rationale for pumped-hydro storage plants has been undermined by the success of PV generation in particular as this has largely suppressed peak electricity prices around noon. It is thus of importance for operators of pumped-hydro storage plants to assess whether the asset's operation will be profitable in the day-ahead market above a specified threshold, potentially derived from considerations of fixed cost coverage, or whether the flexibility should be held for more short-term markets.¹ To forecast the probability of profitability the optimal operation program for a given price curve has to be solved. The optimization problem considered in this study closely follows Steffen and Weber

¹We should not that the natural profit threshold is zero, as the conventional logic of the merit order model implies that a power plant should run, if a positive contribution margin can be generated. Yet, given the simplified pumped-hydro storage optimization problem considered here and the focus on the presentation of the evaluation framework, we have opted to consider an arbitrary threshold of 10,000, which subsequently constitutes an event with balanced occurrence and non-occurrence.

2016 but is formulated in discrete rather than continuous time. It is described by the following equations.

$$\max_{T_t, S_t, F_t} \sum_{t=1}^H P_t(T_t \Delta t - S_t \Delta t) \quad (1)$$

$$s.t. \quad F_t - F_{t-1} = -T_t \Delta t + \eta S_t \Delta t \quad (2)$$

$$0 \leq T_t \leq K_s \quad (3)$$

$$0 \leq S_t \leq K_s \quad (4)$$

$$0 \leq F_t \leq K_F \quad (5)$$

$$F_0 = F^0 \quad (6)$$

$$F_H \geq F_0 \quad (7)$$

It is assumed that the reservoir is filled with $F_0 \leq 0$ at $t = 0$ and the profits from operation of the pumped-hydro storage plant (1) are optimized subject to the set of constraints. The equation of motion (2) ensures that the change in the fill level of the reservoir is equal to the sum of turbining T_t and pumping S_t , accounting for the efficiency factor η . Constraints 3 -5 ensure that the control variables remain within the possible ranges. Since we are considering the profitability of time spread arbitrage the fill level of the reservoir cannot fall below the fill level at the beginning of the optimization period. Following Steffen and Weber 2016 we consider a hydro-pumped storage with pumping and turbining capacity of 200 MW, a maximum storage capacity of 1000 MWh and a starting storage level of 500 MWh. η is assumed to be 0.8. After solving for the optimal schedule given a price path, we calculate the profits and compare them to the profit threshold which is assumed to be 10,000 Euros. Repeating the optimization over a variety of price paths allows us to derive a forecast of the likelihood of profitability as outlined in section 3.

2.2 Six Hours of Negative Electricity Prices Event

The occurrence of n or more consecutive hours of negative electricity prices constitutes the second considered event and we refer to it as the 6h-negative event. Increasing intermittent RES capacity in combination with conventional generation of limited flexibility has raised the likelihood of negative electricity prices (e.g., Agora Energiewende 2014). These reduce the market reference value of RES generation and subsequently increase the pay-out under the German renewable subsidy scheme. Consequently, the German Renewable Energy Sources Act (§51 EEG 2017) stipulates that subsidy payments to RES are retrospectively withheld in case of six or more hours with negative electricity prices. Energy Brainpool GmbH & Co. KG 2017 estimates the revenue shortfall associated with the so-called six-hour rule to amount to 54,000 Euros per installed MW for an onshore wind asset over a period of twenty years. Operators and

direct marketers are thus incentivized to cut infeed in these hours and do thus have an incentive to correctly forecast the occurrence of this event.

3 Event Probability Forecasting

The present study considers two approaches to derive probability predictions for the occurrence of binary events related to electricity prices. In what we term the econometric approach, day-ahead electricity price paths are simulated based on an EPF model and the probability forecast is calculated as the relative frequency of occurrence across the ensemble of simulation paths. Specifically, each simulated day-ahead price path is mapped to the event indicator variable I_m^E , which takes a value of one if prices along path m are such that event E occurs. The day-ahead probability forecast for the binary event E is given by the relative frequency, that is; $\frac{1}{M} \sum_{m=1}^M I_m^E$, across the ensemble of price paths. We set $M = 3000$. To the contrary, the second approach, termed the classification approach here, directly provides predictions for the event's occurrence. Fawcett 2006 defines a classification model as a mapping from instances to classes, where an element of the set of class labels, that is zero or one, is assigned to each instance, using information about that instance. It may output either a predicted class label directly or a predicted probability of class membership, being referred to as a discrete or probabilistic classifier, respectively. Specifically, a series of binary event indicators is modelled using electricity prices of the preceding days. To this end, similar to above, given the vector of electricity prices of day t , P_t , one can define the event indicator variable I_t^E , which takes a value of one if prices on day t are such that event E occurs. Clearly, using the available information at day $t - 1$, one may construct a classifier to predict the class label I_t^E . It should be noted that the considered models serve merely as examples in the exposition of the evaluation framework for binary event probability predictions.

3.1 The Econometric Approach

The econometric approach is based here on two well-established models from the literature. In the naive model, the electricity price of a particular hour h on day t is equal to the price of the same hour the week before, if t constitutes a Monday, Saturday or Sunday, or it is equal to the price of the same hour the day before for all other days (e.g., Conejo et al. 2005a, Conejo et al. 2005b).

$$P_{t,h} = \begin{cases} P_{t-7,h}, D_t \in \{1, 6, 7\} \\ P_{t-1,h}, D_t \notin \{1, 6, 7\} \end{cases} \quad (8)$$

The second model belongs to the class of so-called expert models and is directly taken from Ziel and Weron 2018. It characterizes the electricity price of a particular hour h on day t as a

function of autoregressive terms, non-linear terms, the price of the last hour of the preceding day and dummy variables that capture calendar information.

$$\begin{aligned}
P_{t,h} = & \beta_{h,0} + \beta_{h,1}P_{t-1,h} + \beta_{h,2}P_{t-2,h} + \beta_{h,3}P_{t-7,h} \\
& + \beta_{h,4}P_{t-1,h}^{Max} + \beta_{h,5}P_{t-1,h}^{Min} \\
& + \beta_{h,5}P_{t-1,24} + \sum_{i=1}^6 \beta_{h,6+i}D_t^i + \varepsilon_{t,h}
\end{aligned} \tag{9}$$

We estimate the parameters of the expert model using the Ordinary Least Squares (OLS) estimator (mean regression) and the Quantile Regression (QR) estimator with $\tau = 0.5$ (median regression). Additionally, a Support Vector Regression (SVR) with the same explanatory variables is considered. The hyperparameters of the SVR are selected using the analytic approach of Cherkassky and Ma 2002. The day-ahead price for each individual hour is then forecasted based on both models and random disturbances are added to generate an ensemble of simulated day-ahead price paths. The present study considers two approaches to generate said disturbances. They are either drawn from a multivariate Student's t-distribution, which has been fit to the sample of residuals, or derived using residual-based bootstrapping. It should be noted that we fit both a multivariate Student's t-distribution as well as a multivariate normal distribution, as the limiting case of the former, to the residuals. We subsequently consider whichever achieves the higher likelihood and refer to it as multivariate Student's t-distribution. The non-parametric bootstrapping algorithm is also multivariate in the sense that it returns a vector of twenty-four residuals of a particular day to preserve the intraday correlation structure. The various combinations of econometric models, estimation techniques and simulation approaches provide eight different specifications, the details of which are summarized in Table 1. The probability predictions are subsequently derived as outlined above.

N-Boot	Ex-Boot	QREx-Boot	SVREx-Boot
Naive	Expert	Expert	Expert
-	OLS	QR ($\tau = 0.5$)	SVR
Bootstrap	Bootstrap	Bootstrap	Bootstrap
N-t	Ex-t	QREx-t	SVREx-t
Naive	Expert	Expert	Expert
-	OLS	QR ($\tau = 0.5$)	SVR
Student's t	Student's t	Student's t	Student's t

Table 1: Specification Overview

3.2 The Classification Approach

We model the probability of event E occurring, $\mathbb{P}(I_t^E = |P_{t-1}, P_{t-2}, P_{t-7}, D_1, \dots, D_6)$, given the available price and calendar information, with a regularized logistic regression (RLog) and a

Naive Bayesian classifier (NBayes). It should be noted that both models constitute probabilistic classifiers. The logistic regression model is described by the following equation.

$$\begin{aligned} & \mathbb{P}(I_t^E | P_{t-1}, P_{t-2}, P_{t-7}, D_1, \dots, D_6) \\ &= \frac{1}{1 + e^{-(\beta_0 + \sum_{h=1}^{24} \sum_{k \in \{1,2,7\}} \beta_{h,k} P_{t-k,h} + \sum_{i=1}^6 D_i + \varepsilon_t)}} \end{aligned} \quad (10)$$

Its parameters are estimated by regularized maximum likelihood. Following the estimation of the parameters, the model directly provides probability forecasts for the next-day occurrence of the binary event under study. Similarly, the Naive Bayesian classifier also lends probability forecasts directly yet does not require any parameter estimation. The probability of the event's occurrence is calculated based on Bayes' theorem and a conditional independence assumption.

$$\begin{aligned} & \mathbb{P}(I_t^E | P_{t-1}, P_{t-2}, P_{t-7}, D_1, \dots, D_6) \\ &= \frac{\mathbb{P}(I_t^E) \mathbb{P}(P_{t-1} | I_t^E) \dots \mathbb{P}(D_6 | I_t^E)}{\mathbb{P}(P_{t-1}, P_{t-2}, P_{t-7}, D_1, \dots, D_6)} \end{aligned} \quad (11)$$

The logistic regression and Naive Bayesian classifier constitute the ninth and tenth specification considered in the present study.

4 Evaluation of Event Probability Forecasts

We predict the day-ahead probability of the event's occurrence over the out-of-sample test period and thus observe a series of probability forecasts f_t for each specification. Additionally, the corresponding event indicator series x_t is observed. It should be noted that x_t constitutes the out-of-sample equivalent to I_t^E defined above with the event superscript E suppressed for notational convenience. To evaluate forecasting accuracy, one may compare the predicted probabilities with the realizations of the event. The average of the squared deviations over the out-of-sample period lends the Quadratic Probability Score (QPS), which constitutes the equivalent to the Mean Square Error (MSE) for probability forecasts. However, the precedign approach is suboptimal for the evaluation of probability predictions for rare events. Murphy 1991 defines a rare event as an event that occurs on less than five per cent of forecasting occasions. To suitably evaluate probability predictions of rare events, machine learning techniques developed for the evaluation of classifiers are additionally considered. These techniques also provide further tools for the analysis of more frequent events.

4.1 Quadratic Probability Score

The QPS averages the squared deviation over the out-of-sample period.

$$QPS(f, x) = \frac{1}{T} \sum_{t=1}^T (f_t - x_t)^2 \quad (12)$$

The QPS constitutes a proper, negatively oriented score that takes values between zero and one, where zero denotes perfect forecast accuracy. Since it evaluates accuracy over the entire range of probabilities, the QPS is a global measure of forecast accuracy. To establish statistically significant conclusions on deviations in forecasting accuracy between any two model, as indicated by differences in their QPS, the DM test is applied (e.g., Diebold and Mariano 2002, Ziel and Weron 2018). One can also obtain an understanding of the deficiencies of the considered forecasting models, using decompositions of the QPS. The Murphy Decomposition (MD) decomposes the QPS into the following sum of five terms.

$$\begin{aligned} QPS(f, x) = & \underbrace{\bar{x}(1 - \bar{x})}_{UNC} + \underbrace{\frac{1}{T} \sum_{j=1}^J T_j (\bar{f}_j - \bar{x}_j)^2}_{REL} \\ & - \underbrace{\frac{1}{T} \sum_{j=1}^J T_j (\bar{f}_j - \bar{x}_j)^2}_{RESO} + \underbrace{\frac{1}{T} \sum_{j=1}^J T_j \sum_{t=1}^{T_j} (f_{tj} - \bar{f}_j)^2}_{WBV} \\ & - \underbrace{\frac{1}{T} \sum_{j=1}^J T_j \sum_{t=1}^{T_j} (x_{tj} - \bar{x}_j)(f_{tj} - \bar{f}_j)^2}_{WBC} \end{aligned} \quad (13)$$

It should be noted that Murphy 1972 proposes a decomposition into the first three terms, while the formulation above is due to Stephenson et al. 2008. The original MD requires the evaluation of conditional means of the event indicator series given the forecasts. To this end, one can either condition on the individual probability forecasts directly or assign them to predefined bins of probability. Stephenson et al. 2008 maintains that the binning approach is more common in the literature but requires the effect of binning to be considered in the derivation of the MD. We apply the binning approach by considering J bins of probability and adjust for the effect via the fourth (Within-Bin Variance (WBV)) and fifth term (Within-Bin Covariance (WBC)) as proposed by Stephenson et al. 2008. The first term (Uncertainty (UNC)) represents the uncertainty a forecaster faces when issuing the forecast. It is given by the variance of the event indicator series x_t , which is unobserved at the time of forecast issuance. Note that the uncertainty term is also equivalent to the QPS of a constant probability forecast given by the unconditional mean of the event indicator series over the hold-out sample. The notion of reliability, given by the second term (Reliability (REL)), captures the correspondence between conditional mean observation and conditioning forecast; that is, the correspondence between the mean of the event

indicator series and forecasts within a particular bin. Ideally, for the forecast to be reliable, the probability attached to the realization of the event in a bin should equal its average occurrence within it. Any deviation from perfect reliability increases the QPS above uncertainty. To the contrary, resolution, the third term (Resolution (RESO)), reduces the QPS. It represents the relation between the conditional mean observation and the unconditional mean observation; that is, how well a particular forecasting model distinguishes a particular probability case from relative frequency and attaches different probabilities to different realizations. Thus, the notion of resolution is what makes forecasts useful. Clearly, a forecaster is faced with a trade-off between resolution and reliability in decreasing uncertainty and the issued forecast is useful if it can reduce the QPS below this uncertainty. To adjust the decomposition of the QPS for the effect of binning we follow Stephenson et al. 2008 and calculate the WBV and WBC, which will both be zero if only one forecast value is issued per bin. Stephenson et al. 2008 propose to combine the two within-bin terms with the resolution term to form a generalized resolution term that is less sensitive to the binning of probability forecasts.

The present study considers two approaches to binning, where both of them lend a series of partitions of the unit interval with the number of subintervals ranging from one to ten. The first approach simply divides the unit interval into the specified number of subintervals of equal size. The second approach utilizes a slightly altered version of the constrained k-means algorithm of Bradley et al. 2000. It clusters the probability predictions of all models for a particular event but the constraint set is such that at least five observations of each model fall within each cluster. The bin boundaries are derived from the respective midpoints between the cluster centroids. We find that, when using the binning-robust form of the MD, the differences between the decompositions under the two binning approaches are negligible. Some gains in accuracy are uncovered for constrained k-means binning, when the non-robust decomposition is used but our results are unaffected.

Another, yet more crude, decomposition of the QPS is the Yates Decomposition (YD) (e.g., Yates 1982). It is written as

$$QPS(f, x) = \bar{x}(1 - \bar{x}) + s_f^2 + (\bar{f} - \bar{x})^2 - 2s_{f,x} \quad (14)$$

As in the MD the first term captures the faced uncertainty when issuing the forecast. The third term reflects the notion of bias; that is, how well the forecasting model under consideration performs on average by comparing the unconditional mean forecast to the unconditional mean observation. Clearly, any deviation between these two increases the QPS. The second and fourth term constitute the unconditional variance of the predicted probabilities and the covariance with the observations, respectively. While the variance increases the QPS, the covariance decreases the QPS. One does, however, face a trade-off between the two terms. It should be noted that the variance of the predicted probabilities could be minimized to zero by issuing a constant forecast, which would, however, reduce the covariance to zero as well, implying that one has to strike a balance. Since the covariance between observations and forecasts can be written as $s_{f,x} = (\bar{f}_{x=1} - \bar{f}_{x=0})\bar{x}(1 - \bar{x})$ (Lahiri and Wang 2013) one can consider it as fixing a given

difference between the mean of the forecasts conditional on the value of the instances. It can be shown that required variance, the minimum so to say, to support a given wedge is equal to $s_{f,min}^2 = (\bar{f}_{x=1} - \bar{f}_{x=0})^2$. To better reflect the effect of variance, one can reformulate the YD to

$$QPS(f, x) = \bar{x}(1 - \bar{x}) + \Delta s_f^2 + s_{f,min}^2 + (\bar{f} - \bar{x})^2 - 2s_{f,x} \quad (15)$$

where the third term reflects the minimum variance, while the second term reflects the exceedance of this minimum variance by the variance of the predicted probabilities. Any nonzero excess variance constitutes variability in the predicted probabilities that is unnecessary to support a given wedge of the conditional means and thus the skill of the forecaster is reflected in the ability to minimize this excess variance. Thus, while the covariance can be interpreted as the forecaster's responsiveness to information related to the occurrence of the event, the excess variance represents her responsiveness to information unrelated to the occurrence of the event. Naturally, this should be minimized.

4.2 Evaluation of Classification Models

Consider a discrete classifier for a binary outcome; that is, a model that directly predicts occurrence or non-occurrence rather than probabilities of occurrence. The accuracy of said classifier over the out-of-sample test set can be summarized in a so-called contingency table, which illustrates the correspondence between forecasts and realizations. To analyze the performance of the

	$x_t = 1$	$x_t = 0$
$f_t = 1$	True Positives	False Positives
$f_t = 0$	False Negatives	True Negatives
	P	N

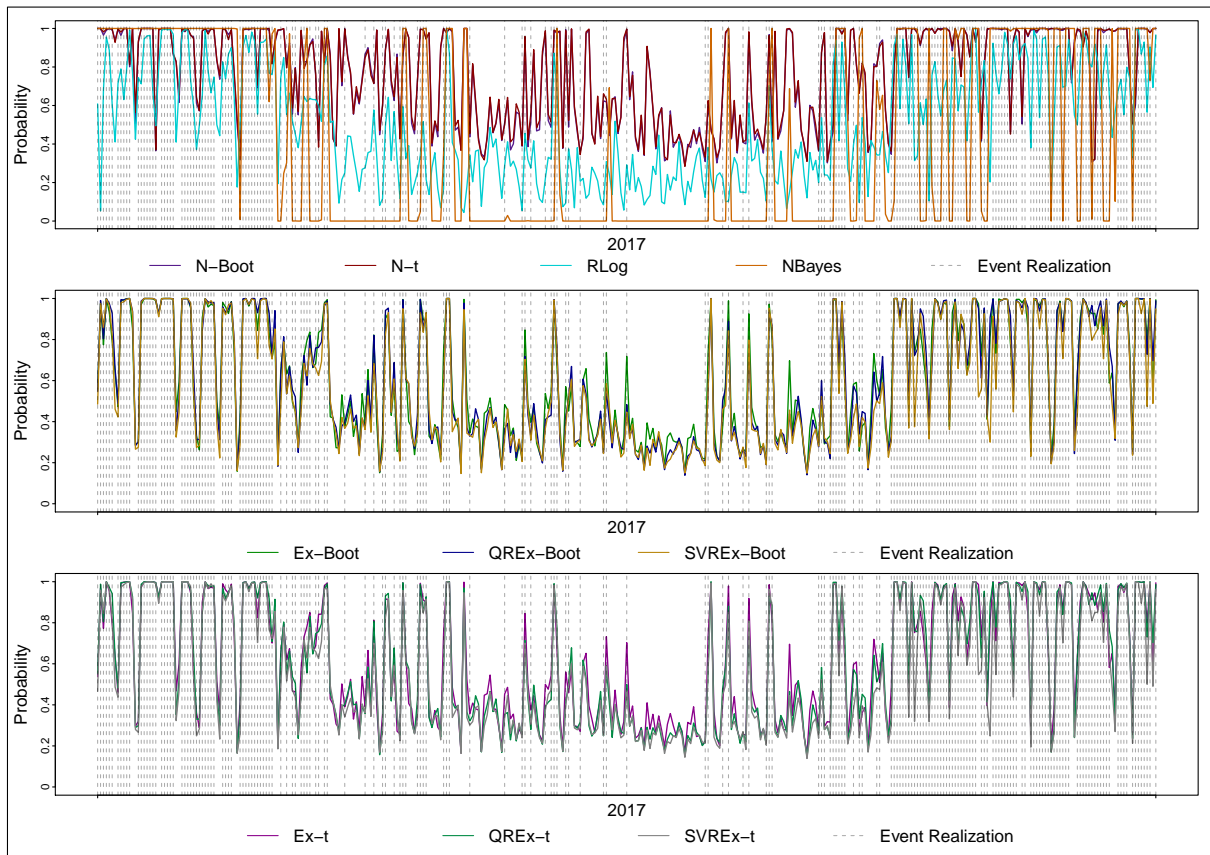
Table 2: Contingency Table

discrete classifier, define the True Positive Rate (TPR) and False Positive Rate (FPR), which denote the proportion of observations where the event was predicted and did occur ($TPR = \frac{TP}{P}$) and the proportion of observation where it was predicted but did not occur ($FPR = \frac{FP}{N}$), respectively. One can subsequently plot the FPR against the TPR in a two-dimensional graph. Since both measures lie strictly between 0 and 1, the potential space is given by the unit square and referred to as Receiver Operating Characteristic (ROC) space. A discrete classifier is represented by a single point in the ROC space with the point of optimality given by $(0, 1)$, where a discrete classifier exhibits a TPR of 1 and an FPR of 0. Thus, over a set of binary classifiers the one closest to $(0, 1)$ achieves the highest forecasting accuracy. By focusing solely on the cases, where the event was forecast to realize, it constitutes a better approach for the evaluation of a rare event's probability predictions, especially when its occurrence is of primary concern to the

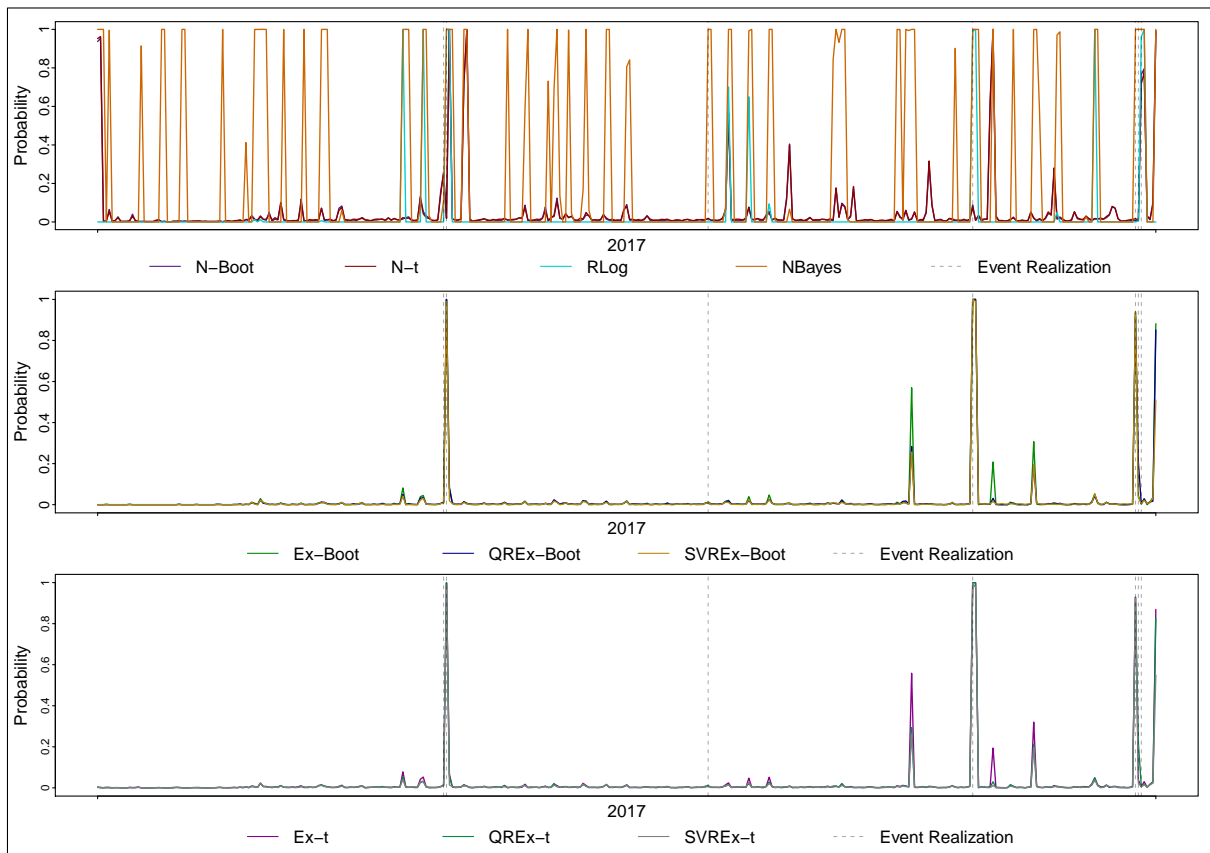
forecast user. Yet, the approach requires the forecasts from a probabilistic classifier to be transformed to discrete forecasts, taking values of zero or one. Said transformation can be achieved by specifying a probability threshold, where the event is predicted when a probability lies above it. By varying the threshold one can trace out a number of points in the ROC space, which lend the ROC curve of a probabilistic classifier. ROC curves themselves constitute a tool of classifier evaluation and exhibit the nice property of being invariant to class distribution. Thus, if we were to consider two out-of-sample test sets with different number of occurrences of the rare event, the ROC curve of a particular classifier would not change. Nevertheless, although it is possible to compare prediction models on the basis of their corresponding ROC curves, it is more common to derive a scalar measure of aggregate performance, which is the Area under Receiver Operating Characteristic Curve (AUROC). Since the AUROC always constitutes a subarea of the unit square, it lies strictly between 0 and 1. One established shortcoming is that ROC curves may cross, implying that one curve and hence one model may exhibit a larger AUROC although the alternative model may exhibit a better performance, as indicated by a higher ROC curve, over the majority of the range of classification thresholds. Hand 2009 derives another fundamental deficiency of the AUROC as measure of forecasting performance. He shows that a comparison of AUROC values amounts to comparing the forecasting models using metrics that themselves depend on the models, essentially meaning that the comparison uses a different metric per model. To address said problem of evaluation, Hand 2009 proposes the so-called H-Measure, which the present study reports alongside the AUROC to evaluate forecasting accuracy for rare events.

5 Empirical Results and Discussion

To illustrate the applicability of the proposed evaluation framework for binary event probability predictions, we conduct an out-of-sample forecasting study on German day-ahead electricity prices using a rolling window approach. The in-sample period covers the last 730 days and we predict the probability of the illustrative events for the following day over the out-of-sample test set, ranging from 1st January 2016 to 31st December 2017, comprising $T = 731$ days. We display the time series of probability forecasts across events and models for 2017 in Figure 1. The colored lines constitute the probability forecasts while the dashed black vertical lines indicate the occurrence of the respective event. Clearly, the considered 10k-pump event constitutes a rather common event in 2017. In contrast the 6h-negative event rarely realized. In fact, with 14 occurrences over the out-of-sample period it falls within the rare event definition of Murphy 1991. Figure 1 indicates that the predicted probabilities vary both across time and models. For example, the exceedance of the predefined profit threshold for the pumped-hydro storage plant is ex ante more likely during winter when less peak shaving due to Photovoltaic (PV) generation occurs. Also, the specifications based on the naive electricity price model structurally assign higher probabilities to the event, while the naive Bayesian classifier mainly provides extreme



(a) 10k-Pump Event



(b) 6h-Negative Event

Figure 1: Probability Time Series for Events

predictions of zero or one. Similarly, the specifications based on the expert model assign much lower probabilities to six consecutive hours of negative prices than the naive model with the majority of days with larger probability in the second half of 2017. Additionally, it seems that the models predict the realization of the rare event rather well, which is however misleading. Close inspection reveals that the realization of the rare event is predicted for the day after its occurrence; the reason being that the prices, which are such, that the event occurs subsequently form the basis for the day-ahead prediction and thus assign a high probability to the event's occurrence. QPS values for the considered forecasting models are reported in Table 3. For the 10k-pump event one can observe from Table 3 that the expert-based specifications and the regularized logistic regression outperform the naive approaches, while the SVR models outperform both the mean- and median-regression models. SVREx-t constitutes the best overall model and has a slightly lower QPS value than SVREx-Boot and RLog. In Figure 2 we summarize the results of the corresponding DM tests. Each square displays the p-value of a pairwise test of equal predictive performance against the alternative hypothesis that the model in the row predicts significantly less accurately than the model in the corresponding column. White squares indicate that no significant difference in forecasting performance can be uncovered, whereas green squares indicate significant deviations in forecasting performance at the ten, five and one per cent level of significance with lighter green implying a more significant difference. The results of the DM tests underscore the preceding discussion based on Table 3, as the deviations in forecasting performance between the expert-based models as well as the regularized logistic regression and the naive models are found to be significant at the one per cent level. Furthermore, the superiority in predictive ability of the support-vector-regression-based specifications is confirmed. The overall best model (SVREx-t) significantly outperforms the second-best model (SVREx-Boot). Yet, the null of equal predictive performance between SVREx-t and the third-best model (RLog) cannot be rejected, despite a higher deviation in the QPS value. This somewhat surprising result is due to the difference in the standard deviation of the respective loss differential time series, affecting the test statistic of the DM test. Interestingly, within each subclass of expert-based specifications, the model using the t-distribution outperforms the bootstrap approach, suggesting that the assumed distribution of the innovations is critical to the achieved performance. With the exception of the naive Bayesian classifier, we find the overall level of the QPS to be much lower for the 6h-negative event than for the 10k-pump event, which illustrates the influence of the frequent realization of the rare event on the evaluation measure. Since the models generally assign low probabilities to the day-ahead occurrence of the event, their respective scores are low. In contrast, the naive Bayesian classifier predicts the event's occurrence rather frequently and thus exhibits a large QPS value. Consequently, we find it to be significantly outperformed by all other models. Similarly, the expert-based specifications significantly outperform the naive and logistic specification. QREx-t constitutes the best overall model and has just a slightly lower QPS value than QREx-Boot. In fact, the median regressions perform significantly better than the mean regression. Yet, among them no significant difference in predictive performance can be uncovered, suggesting that it is the median-regression approach rather than the assumed dis-

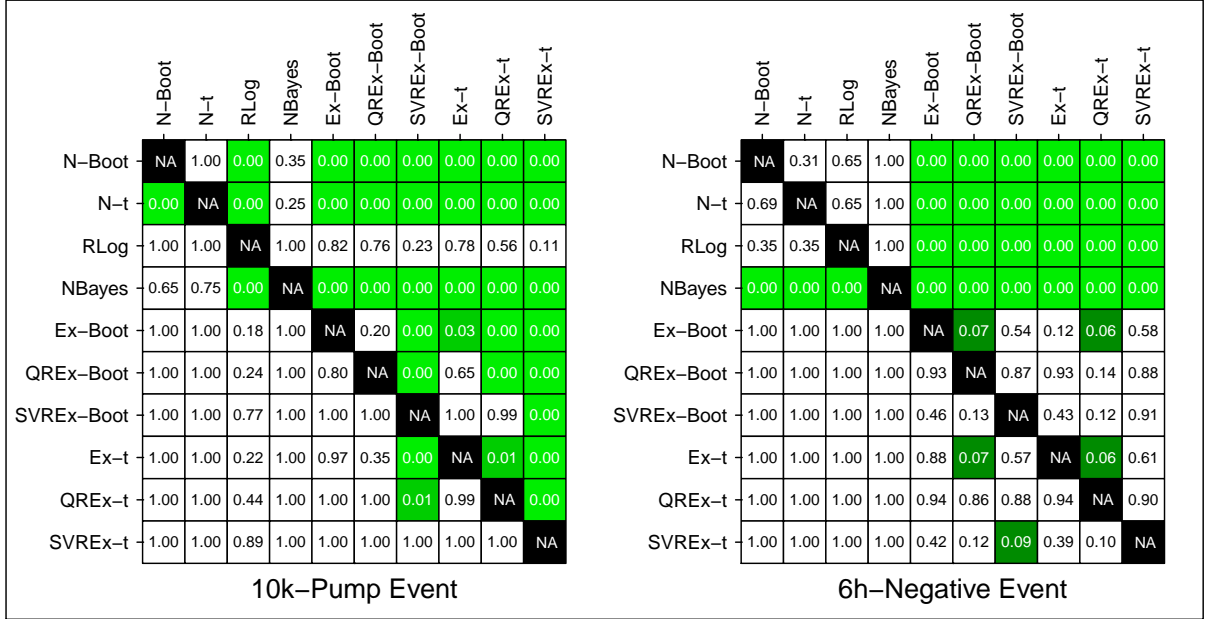


Figure 2: p-Values of DM Test

tribution of innovations that drives the result. Nevertheless, the median-regression-based model fail to significantly outperform the SVR models.

Event	N-Boot	N-t	RLog	NBayes	Ex-Boot
10k-Pump	0.2746	0.2797	0.1798	0.2675	0.1860
6h-Negative	0.0315	0.0314	0.0339	0.2201	0.0155
Event	QREx-Boot	SVREx-Boot	Ex-t	QREx-t	SVREx-t
10k-Pump	0.1844	0.1756	0.1852	0.1808	0.1728
6h-Negative	0.0141	0.0157	0.0154	0.0140	0.159

Table 3: Quadratic Probability Score

The MD provides further insights into the deficiencies of the considered models. Plots of the MD and its respective components are presented in the two left panels of Figure 3. It should be noted that the uncertainty component, being derived from the event indicator series over the out-of-sample test set, is the same across all models for a given event. We find that the naive specifications increase the QPS above said uncertainty for the 10k-pump event, as they are highly miscalibrated and provide little resolution. In contrast, the expert-based specifications and regularized logistic regression succeed in reducing the uncertainty. The RLog model provides slightly lower resolution than the expert-based models, meaning it is less able to distinguish between the respective cases of the event, but does so at a lower level of miscalibration, thus explaining its lower QPS value. Additionally, the MD provides an explanation why the specifications based on SVR perform better than the other models in a respective class. For similar levels of resolution achieved, the SVR models are the least miscalibrated. For the 6h-negative event we find that the naive specifications and the regularized logistics regression increase the QPS

above uncertainty, due to substantial miscalibration, which is particularly acute for the naive Bayesian classifier. It should be noted that both the QPS and reliability bar have been capped at 0.035. In contrast, the expert-based specifications succeed in reducing the uncertainty. Within that class the median-regression-based models are the least miscalibrated with slightly higher resolution, implying that overall the issued forecasts correspond well with the realization of the event and that the models are most effective in using the provided information to distinguish cases of occurrence and non-occurrence of the event.

The YD provides additional insights regarding the deficiencies of the considered forecasting models. Its components are shown in the two right panels of Figure 3. It should be noted again that the uncertainty component is the same across all models for a given event. For the 10k-pump event the expert-based specifications all achieve similar levels of covariance between forecasts and observations as well as similar levels of excess variance. Yet, the SVRs do so at a lower level of bias, explaining their overall best performance. Similarly, the RLog model exhibits lower covariance and higher excess variance, but it achieves the lowest level of bias among all models. Interestingly, the naive Bayesian classifier achieves the highest overall covariance between forecasts and observation, yet it also has the highest excess variance. The percentage of excess variance of overall variance is found to be 79%, 80%, 71%, 80%, 65%, 65%, 65%, 65%, 64% and 64%, respectively. Thus, the considered models have problems to sufficiently incorporate information related to the event’s occurrence and the subjective forecasts are scattered unnecessarily around the conditional means of the forecasts. We establish similar results for the 6h-negative event. The naive specifications together with the regularized logistics regression again increase the QPS above uncertainty. It should be noted that the values for naive Bayesian classifier have again been capped. As before it achieves the highest covariance between forecasts and observations but at the cost of excessive bias. Among the expert-based specifications the SVR-based specifications achieve the lowest excess variance but also the lowest covariance. The trade-off between variance and covariance is best achieved by the median-regression-based models. Overall, the percentage of excess variance of the overall variance is rather elevated. It is found to amount to 98%, 98%, 93%, 96%, 75%, 72%, 80%, 75%, 72% and 81%, respectively. Thus, for the rare event the problem of sufficiently incorporating information related to the event’s occurrence is even more acute.

In contrast to the QPS both the AUROC and the H-Measure are positively oriented measures of forecasting accuracy, focusing on the event’s occurrence. The respective values per model are provided in Table 4 and Table 5. It should be noted that all AUROC values are larger than 0.5 implying that all models perform better than random class guessing. For the 10k-pump event the naive specifications exhibit the lowest AUROC among all considered models, underscoring the results derived from the QPS comparisons. Additionally, the expert-based specifications exhibit the highest AUROC values and within each subclass the model using the t-distribution outperforms the model using the bootstrap, which confirms our previous findings. Yet, the model with the highest AUROC (QREx-t) is not equivalent with the model with the lowest QPS (SVREx-t), despite the DM test suggesting a significant difference in forecasting performance. Interestingly,

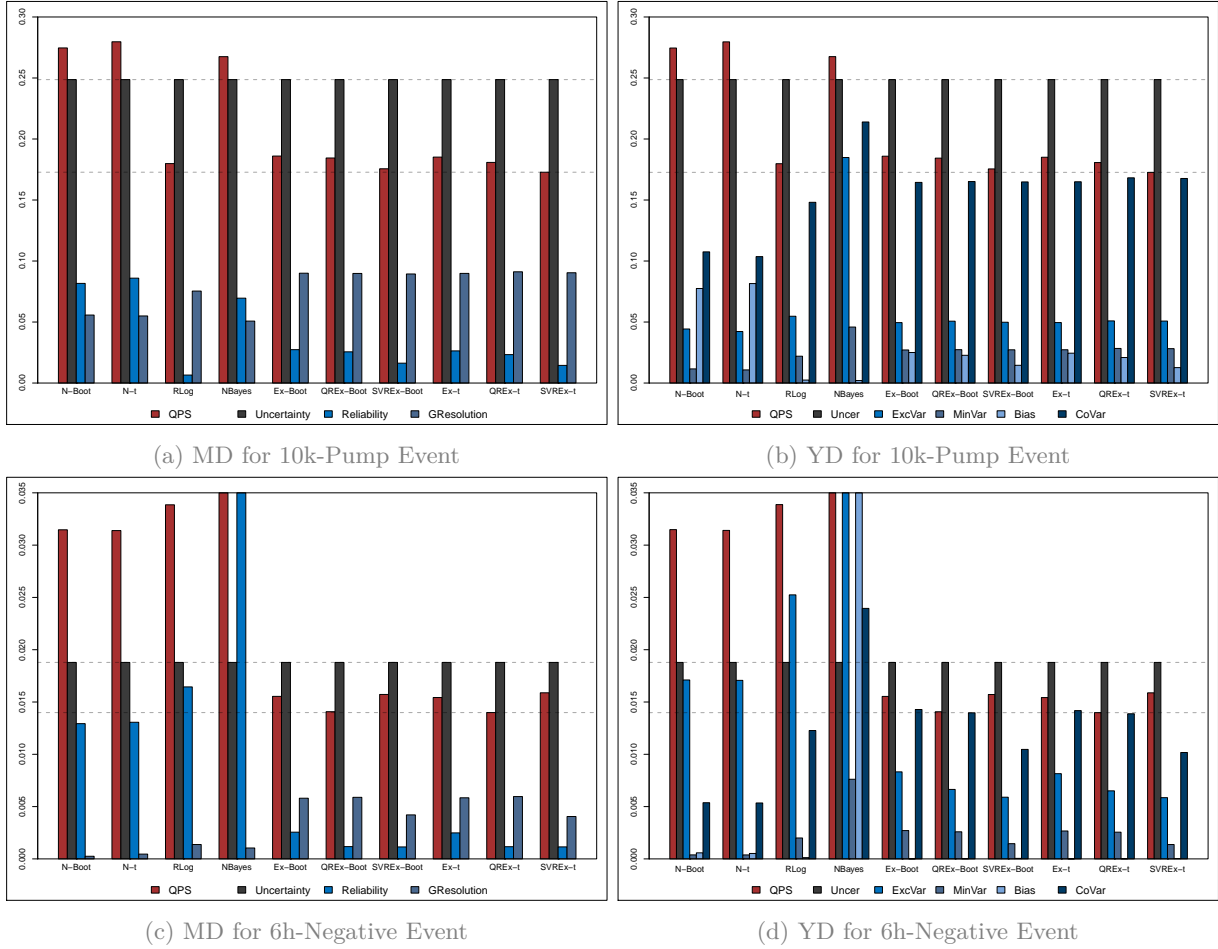


Figure 3: QPS Decompositions for Events

the RLog model, which constitutes the third-best model according to QPS is outperformed by all expert-based specifications. The reason may be that the RLog model fails to correctly predict numerous occurrences of the event during the summer months (Figure 1). For the 6h-negative event the naive power price specifications (N-Boot and N-t) exhibit the lowest AUROC values among all considered models. Among the remaining specifications the QREx-t model achieves the highest AUROC value and it is also the model that achieves the lowest QPS value. Yet, both RLog and the naive Bayesian classifier achieve higher AUROC values than some specifications in the expert-based class. This finding is in stark contrast to the results presented above, where the two models are significantly outperformed by all expert-based specifications. The result seems to suggest that an evaluation that does not account for the frequent realization provides misleading conclusions when forecasting probabilities for rare events. In particular, the finding suggests that the RLog and especially the naive Bayesian model exhibit very high QPS values overall, as they forecast the occurrence of the rare event too frequently but they seem to exhibit a higher hit rate, when the rare event actually occurs.

Yet, as Hand 2009 derives some series deficiencies of the AUROC as a measure of forecasting accuracy, we additionally evaluate our models with his proposed H-Measure. We find the Ex-

Event	N-Boot	N-t	RLog	NBayes	Ex-Boot
10k-Pump	0.776	0.774	0.805	0.762	0.839
6h-Negative	0.610	0.652	0.885	0.894	0.855
Event	QREx-Boot	SVREx-Boot	Ex-t	QREx-t	SVREx-t
10k-Pump	0.839	0.837	0.840	0.842	0.838
6h-Negative	0.843	0.856	0.902	0.903	0.898

Table 4: AUROC

Boot model to outperform all other models for the 10k-pump event. Additionally, the bootstrap approach outperforms the t-distribution approach in each subclass of the expert-based specifications, which is contrary to our result established using the QPS. Yet, the all expert-based specifications achieve a higher H-Measure than the RLog model. This result mirrors our findings for the AUROC measure but is at odds with our conclusion reached for the QPS. Thus, it suggests that, focusing on its occurrence for model evaluation, the 10k-pump event is best forecast using the expert-based specifications. For the 6h-negative event the results using the H-Measure and the AUROC are very similar. The QREx-t model constitutes the overall best model and again the RLog model as well the naive Bayesian classifier are less deficient according to the H-Measure than according to the QPS, suggesting that they forecast the realization of the rare event rather well. However, the expert-based specifications all achieve a higher H-Measure, which resuscitates our findings based on the QPS, where they clearly outperformed the remaining models.

Event	N-Boot	N-t	RLog	NBayes	Ex-Boot
10k-Pump	0.290	0.281	0.363	0.275	0.423
6h-Negative	0.204	0.220	0.527	0.569	0.582
Event	QREx-Boot	SVREx-Boot	Ex-t	QREx-t	SVREx-t
10k-Pump	0.420	0.414	0.417	0.419	0.413
6h-Negative	0.575	0.609	0.578	0.626	0.585

Table 5: H-Measure

6 Conclusion

Probabilistic forecasts over binary events have a long tradition in fields such as weather forecasting and economics. Despite a paradigm shift from point to probabilistic forecasting in the realm of EPF, such forecasts of binary events have not yet been fully considered and no coherent evaluation framework has been established. The present study fills the void by proposing an evaluation framework that ties in and extends the existing EPF framework. It employs the MSE-equivalent QPS together with the DM test and allows for further insights about deficiencies of the considered models. Additionally, we consider techniques from the field of classification,

which extend our framework and are particularly suited for the evaluation of predictions of rare events. We demonstrate the applicability of our framework with two illustrative examples motivated by energy companies' daily asset scheduling. Overall, we find that the well-established expert models also form a reliable basis for probability forecasts of binary events. Concerning the evaluation of such forecasts, we establish that our proposed framework provides valuable insights about the considered specifications and that care needs to be taken, when evaluating forecasts with just the traditional QPS, especially for events that rarely realize. Decompositions of the QPS and additional evaluation techniques are worthwhile considering for the identification of the overall best specification. Furthermore, we reconcile the strand of the literature concerned with the practical applicability of forecasts and the forecasting evaluation literature rooted in the realm of statistics. By considering binary events with direct applicability to a generator's daily decision making and evaluating the respective forecast statistically, we show that the task of forecast evaluation can be simplified from assessing a multivariate distribution over continuous outcomes to assessing a univariate distribution over a binary outcome, fully characterized by a single probability. Whether a forecaster utilizes the former or the latter approach depends on her preferences and the specific forecasting problem at hand, which we do not address. Yet, we find that our simplified evaluation approach is sufficient from the perspective of the eventual application of the forecast, provides benefits in its own right and provides an interesting path for future research.

References

- Agora Energiewende (2014). *Negative Strompreise: Ursachen und Wirkungen: Eine Analyse der aktuellen Entwicklungen und ein Vorschlag für ein Flexibilitätsgesetz* (cit. on p. 4).
- Bradley, P. S., Bennett, K. P., and Demiriz, A. (2000). *Constrained K-Means Clustering: Microsoft Research, Redmond 20.0* (cit. on p. 9).
- Braun, S. and Hoffmann, R. (2016). “Intraday Optimization of Pumped Hydro Power Plants in the German Electricity Market”. In: *Energy Procedia* 87, pp. 45–52 (cit. on p. 3).
- Brown, P. D., Peas Lopes, J. A., and Matos, M. A. (2008). “Optimization of Pumped Storage Capacity in an Isolated Power System With Large Renewable Penetration”. In: *IEEE Transactions on Power Systems* 23.2, pp. 523–531 (cit. on p. 3).
- Bunn, D. W., Gianfreda, A., and Kermer, S. (2018). “A Trading-Based Evaluation of Density Forecasts in a Real-Time Electricity Market”. In: *Energies* 11.10, p. 2658 (cit. on p. 2).
- Cherkassky, V. and Ma, Y. (2002). “Selection of Meta-parameters for Support Vector Regression”. In: *Artificial Neural Networks — ICANN 2002*. Ed. by G. Goos, J. Hartmanis, J. van Leeuwen, and J. R. Dorronsoro. Vol. 2415. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 687–693 (cit. on p. 6).
- Conejo, A. J., Plazas, M. A., Espinola, R., and Molina, A. B. (2005a). “Day-Ahead Electricity Price Forecasting Using the Wavelet Transform and ARIMA Models”. In: *IEEE Transactions on Power Systems* 20.2, pp. 1035–1042 (cit. on p. 5).
- Conejo, A. J., Contreras, J., Espínola, R., and Plazas, M. A. (2005b). “Forecasting electricity prices for a day-ahead pool-based electric energy market”. In: *International Journal of Forecasting* 21.3, pp. 435–462 (cit. on p. 5).
- Delarue, E., van den Bosch, P., and D’haeseleer, W. (2010). “Effect of the accuracy of price forecasting on profit in a Price Based Unit Commitment”. In: *Electric Power Systems Research* 80.10, pp. 1306–1313 (cit. on p. 1).
- Diebold, F. X., Gunther, T. A., and Tay, A. S. (1998). “Evaluating Density Forecasts with Applications to Financial Risk Management”. In: *International Economic Review* 39.4, p. 863 (cit. on p. 2).
- Diebold, F. X. and Mariano, R. S. (2002). “Comparing Predictive Accuracy”. In: *Journal of Business & Economic Statistics* 20.1, pp. 134–144 (cit. on p. 8).
- Doostmohammadi, A., Amjady, N., and Zareipour, H. (2017). “Day-Ahead Financial Loss/Gain Modeling and Prediction for a Generation Company”. In: *IEEE Transactions on Power Systems* 32.5, pp. 3360–3372 (cit. on p. 1).
- Energy Brainpool GmbH & Co. KG (2017). *Einfluss der Sechs-Stunden-Regel auf die Erlöse einer Wind und PV Anlage: Energy Brainpool White Paper* (cit. on p. 4).
- Fawcett, T. (2006). “An introduction to ROC analysis”. In: *Pattern Recognition Letters* 27.8, pp. 861–874 (cit. on p. 5).
- Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). “Probabilistic Forecasts, Calibration and Sharpness”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69.2, pp. 243–268 (cit. on p. 1).

- Gneiting, T. and Katzfuss, M. (2014). “Probabilistic Forecasting”. In: *Annual Review of Statistics and Its Application* 1.1, pp. 125–151 (cit. on p. 1).
- Gürtler, M. and Paulsen, T. (2018). “Forecasting performance of time series models on electricity spot markets: a quasi-meta-analysis”. In: *International Journal of Energy Sector Management* 12.1, pp. 103–129 (cit. on p. 1).
- Hand, D. J. (2009). “Measuring classifier performance: a coherent alternative to the area under the ROC curve”. In: *Machine Learning* 77.1, pp. 103–123 (cit. on pp. 11, 16).
- Jónsson, T., Pinson, P., Madsen, H., and Nielsen, H. (2014). “Predictive Densities for Day-Ahead Electricity Prices Using Time-Adaptive Quantile Regression”. In: *Energies* 7.9, pp. 5523–5547 (cit. on p. 1).
- Jonsson, T., Pinson, P., Nielsen, H. A., Madsen, H., and Nielsen, T. S. (2013). “Forecasting Electricity Spot Prices Accounting for Wind Power Predictions”. In: *IEEE Transactions on Sustainable Energy* 4.1, pp. 210–218 (cit. on p. 1).
- Kath, C. and Ziel, F. (2018). “The value of forecasts: Quantifying the economic gains of accurate quarter-hourly electricity price forecasts”. In: *Energy Economics* 76, pp. 411–423 (cit. on p. 2).
- Lahiri, K. and Wang, J. G. (2013). “Evaluating probability forecasts for GDP declines using alternative methodologies”. In: *International Journal of Forecasting* 29.1, pp. 175–190 (cit. on pp. 2, 9).
- Marcjasz, G., Uniejewski, B., and Weron, R. (2018). *Probabilistic electricity price forecasting with NARX networks: Combine point or probabilistic forecasts? HSC Research Report HSC/18/05* (cit. on p. 1).
- Mohammadi-Ivatloo, B., Zareipour, H., Ehsan, M., and Amjady, N. (2011). “Economic impact of price forecasting inaccuracies on self-scheduling of generation companies”. In: *Electric Power Systems Research* 81.2, pp. 617–624 (cit. on p. 1).
- Murphy, A. H. (1972). “Scalar and Vector Partitions of the Probability Score: Part I. Two-state Situation”. In: *Journal of Applied Meteorology* 11.2, pp. 273–282 (cit. on p. 8).
- Murphy, A. H. (1988). “Skill Scores Based on the Mean Square Error and Their Relationships to the Correlation Coefficient”. In: *Monthly Weather Review* 116.12, pp. 2417–2424 (cit. on p. 2).
- Murphy, A. H. (1991). “Probabilities, Odds and Forecasts of Rare Events”. In: *Weather and Forecasting* 6.2, pp. 302–307 (cit. on pp. 7, 11).
- Nowotarski, J. and Weron, R. (2018). “Recent advances in electricity price forecasting: A review of probabilistic forecasting”. In: *Renewable and Sustainable Energy Reviews* 81, pp. 1548–1568 (cit. on p. 1).
- Pape, C., Woll, O., and Weber, C. (2017). *Forecasting the Distribution of Hourly Electricity Spot Prices: Accounting for Serial Correlation Patterns and Non-Normality of Price Distributions: HEMF Working Paper 05/2017* (cit. on p. 1).
- Steffen, B. and Weber, C. (2016). “Optimal operation of pumped-hydro storage plants with continuous time-varying power prices”. In: *European Journal of Operational Research* 252.1, pp. 308–321 (cit. on pp. 3 sq.).

- Stephenson, D. B. (2000). "Use of the "Odds Ratio" for Diagnosing Forecast Skill". In: *Weather and Forecasting* 15.2, pp. 221–232 (cit. on p. 2).
- Stephenson, D. B., Coelho, C. A. S., and Jolliffe, I. T. (2008). "Two Extra Components in the Brier Score Decomposition". In: *Weather and Forecasting* 23.4, pp. 752–757 (cit. on pp. 8 sq.).
- Weron, R. (2014). "Electricity price forecasting: A review of the state-of-the-art with a look into the future". In: *International Journal of Forecasting* 30.4, pp. 1030–1081 (cit. on pp. 1 sq.).
- Weron, R. and Ziel, F. (2018). *Electricity price forecasting: HSC Research Report HSC/18/08* (cit. on p. 1).
- Yates, J. F. (1982). "External Correspondence: Decompositions of the Mean Probability Score". In: *Organizational Behavior and Human Performance* 30.1, pp. 132–156 (cit. on p. 9).
- Zareipour, H., Canizares, C. A., and Bhattacharya, K. (2010). "Economic Impact of Electricity Market Price Forecasting Errors: A Demand-Side Analysis". In: *IEEE Transactions on Power Systems* 25.1, pp. 254–262 (cit. on pp. 1 sq.).
- Ziel, F. (2016). "Forecasting Electricity Spot Prices Using Lasso: On Capturing the Autoregressive Intraday Structure". In: *IEEE Transactions on Power Systems* 31.6, pp. 4977–4987 (cit. on p. 1).
- Ziel, F. and Weron, R. (2018). "Day-ahead electricity price forecasting with high-dimensional structures: Univariate vs. multivariate modeling frameworks". In: *Energy Economics* 70, pp. 396–420 (cit. on pp. 2, 5, 8).

Correspondence

Arne Vogler

(Corresponding Author)

Research Associate

House of Energy Markets and Finance

University of Duisburg-Essen

Universitätsstr. 12

45117 Essen

Germany

E-Mail arne.vogler@wiwinf.uni-due.de

Prof. Dr. Florian Ziel

Professor

House of Energy Markets and Finance

University of Duisburg-Essen, Germany

Universitätsstr. 12

45117 Essen

Germany

E-Mail florian.ziel@uni-due.de