Renewable and Sustainable Energy Reviews 81 (2018) 1548–1568

Contents lists available at ScienceDirect

Renewable and Sustainable Energy Reviews

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

Recent advances in electricity price forecasting: A review of probabilistic forecasting

Jakub Nowotarski, Rafał Weron*

Department of Operations Research, Wrocław University of Science and Technology, 50-370 Wrocław, Poland

ARTICLE INFO

Keywords:
Electricity price forecasting
Probabilistic forecast
Reliability
Sharpness
Day-ahead market
Autoregression
Neural network

ABSTRACT

Since the inception of competitive power markets two decades ago, *electricity price forecasting* (EPF) has gradually become a fundamental process for energy companies' decision making mechanisms. Over the years, the bulk of research has concerned point predictions. However, the recent introduction of smart grids and renewable integration requirements has had the effect of increasing the uncertainty of future supply, demand and prices. Academics and practitioners alike have come to understand that probabilistic electricity price (and load) forecasting is now more important for energy systems planning and operations than ever before. With this paper we offer a tutorial review of probabilistic EPF and present much needed guidelines for the rigorous use of methods, measures and tests, in line with the paradigm of 'maximizing sharpness subject to reliability'. The paper can be treated as an update and a further extension of the otherwise comprehensive EPF review of Weron [1] or as a standalone treatment of a fascinating and underdeveloped topic, that has a much broader reach than EPF itself.

1. Introduction

In their analysis of research in time series forecasting, covering the period 1982–2005 and summarizing over 940 papers, De Gooijer and Hyndman [2] conclude that the use of prediction intervals and densities, or *probabilistic forecasting*, has become much more common over the years, as 'practitioners have come to understand the limitations of point forecasts'. Nevertheless, back in 2013, when Weron started writing his review [1], this did not seem to be the case for *electricity price forecasting* (EPF).[1] The article speculated, however, that probabilistic forecasting was one of five directions that should and would develop over the next decade or so. Somewhat surprisingly, this 'prophecy' has already come true. After a decade of limited interest, probabilistic EPF gained momentum with the Global Energy Forecasting Competition (GEFCom2014), which commenced in August 2014 and focused solely on probabilistic energy (load, price, wind and solar) forecasting [3]. The price track attracted 287 contestants worldwide and the best ranking teams were later invited to submit a paper to the 2016 special issue of the *International Journal of Forecasting*, see Section 2. Altogether, seven price forecasting articles appeared in the issue, marking the beginning of the era of probabilistic EPF.

Naturally, the GEFCom2014 competition was not the reason, rather the effect of increased interest in probabilistic energy forecasting. The energy industry has been going through a significant modernization process. In the last decade, the increased market competition, aging infrastructure, introduction of smart grids and renewable integration requirements have had the effect of probabilistic load and price forecasting becoming more and more important to energy systems planning and operations [4–7]. And probabilistic forecasting has a lot to offer, in particular, improved assessment of future uncertainty, ability to plan different strategies for the range of possible outcomes, increased effectiveness of submitted bids and possibility of more thorough forecast comparisons [1,8,9].

However, probabilistic EPF is an underdeveloped topic, with both academics and practitioners not using the correct evaluation or testing procedures (as discussed below). With this paper we offer a much needed tutorial review that explains the complexity of the available

* Corresponding author.
E-mail addresses: jakub.nowotarski@pwr.edu.pl (J. Nowotarski), rafal.weron@pwr.edu.pl (R. Weron).

[1] To avoid ambiguous and verbose presentation – unless stated otherwise – we use the term *price forecasting* to refer to *electricity price forecasting*. We also use EPF as the abbreviation for both *electricity price forecasting* and *electricity price forecast*, while PEPF for *probabilistic* EPF. The plural form, i.e., *forecasts*, is abbreviated EPFs and PEPFs, respectively.

http://dx.doi.org/10.1016/j.rser.2017.05.234
Received 19 September 2016; Received in revised form 15 March 2017; Accepted 25 May 2017
Available online 16 June 2017
1364-0321/

solutions, including notable techniques, statistically sound and less formal evaluation methods and common misunderstandings. The paper can be treated as an update and a further extension of the otherwise comprehensive EPF review of Weron [1] or as a standalone treatment of a fascinating and underdeveloped topic, that has a much broader reach than EPF itself. In particular, as Raza and Khosravi [10] argue, the electricity price is one of the influential explanatory variables in load forecasting and PEPFs could be considered as input in probabilistic load forecasting models for smart grids and buildings.

We start with a top-down literature review in Section 2. We first conduct an extensive bibliometric study of the Web of Science and Scopus databases. Then, acknowledging the importance of the GEFCom2014 competition, in particular its competitiveness and unified forecast evaluation, we summarize the methods used by the top four winning teams in the price track (note that we will utilize two of these approaches in the empirical study in Section 5). Finally, in Section 2.3 we review other important PEPF publications.

In Section 3 we first formulate the probabilistic forecasting problem, then discuss four approaches to constructing probabilistic forecasts: (i) historical simulation (or empirical/sample prediction intervals,[2] PIs), (ii) distribution-based probabilistic forecasts, (iii) bootstrapped PIs and (iv) Quantile Regression Averaging (QRA). Next, in Section 4, following the paradigm of 'maximizing sharpness subject to reliability' [15–17], we first present the numerical tools and statistical tests to assess *reliability* (i.e., the statistical consistency between the distributional forecasts and the observations; also called *calibration* or *unbiasedness*), then discuss the techniques for measuring and analyzing the *sharpness* (i.e., the concentration of the predictive distributions).

In the empirical study of Section 5 we employ most of the methods detailed in the preceding two Sections. To provide transparency and replicability, we use a dataset that comes from the price track of the GEFCom2014 competition, which is available as supplementary material accompanying Ref. [3]. In the closing paragraphs, in Section 5.4, we put forward recommendations for the evaluation of probabilistic forecasts. Finally, in Section 6 we conclude.

2. Literature review

Compared to probabilistic wind power forecasting [18–21], the literature on probabilistic EPF is relatively scarce, even taking into account the 2016 special issue on the GEFCom2014 competition [3]. This 'maturity' of wind power forecasting is likely due to its close relationship to meteorological forecasting, where probabilistic predictions are well-established and commonly accepted. On the other hand, EPF has not picked up before the deregulation of the 1990s and the establishment of power markets for trading electricity [22]. As Hong et al. [3] argue, electricity prices, and especially price spikes, are influenced heavily by a wide range of factors other than the electricity demand, such as transmission congestion, generation outages, market participant behaviors, etc. These factors, and the uncertainties associated with them, are hard to incorporate into EPF models. So the first wave of models focused on point forecasting, which is generally less demanding and easier to comprehend and implement than PEPF [1,7]. To put probabilistic EPF in perspective we start with a bibliometric study of EPF itself.

2.1. Bibliometric survey

In this section, we report on the bibliometric analysis we performed on 15 March 2017, nearly three years after a similar study of Weron [1]. We use two well-established, constantly expanding and generally acknowledged databases: Web of Science (WoS) and Scopus. We will first present general results for both databases, then more specialized queries for Scopus only (its search engine is more user-friendly and allows for more refined queries). Since the collections of publications indexed by WoS and Scopus are not the same, the results do differ quantitatively but the overall picture is similar.

In Fig. 1 we plot the number of WoS- and Scopus-indexed EPF publications in the years 1992–2016.[3] The overall number of publications is 559 for WoS and 664 for Scopus. Respectively 285 (51%) and 328 (49%) of these are journal articles. Both databases are constantly being expanded to cover more journals and proceedings volumes, but still the indexed publications are not representative of the true number of conference papers. Since the latter are typically also of lower quality than journal articles, like Weron [1], we mostly concentrate on articles. Note, however, that because we have modified the queries to better filter out relevant EPF publications, the results are not fully comparable between the two bibliometric studies.

Except for a handful of papers, EPF publications have not appeared in the literature before year 2000. The next major breakthrough were the years 2005 and 2006 when the number of publications first doubled, then tripled with respect to 2002–2004 figures, mainly due to conference papers; journal articles followed with a delay. The publication rate rapidly increased until 2009, then dropped to pre-2009 levels, to pick up again in 2012. As of 2016 the topic seems to have regained interest, with the figures for 2015–2016 being significantly higher than the numbers for 2009. This is also visible in the constantly increasing numbers of citations, see the left panel in Fig. 2.

As far as probabilistic forecasting[4] is concerned, the topic was not present in the EPF literature until Zhang et al. published two conference papers in 2002 (not visible in Fig. 2) and the first probabilistic EPF article in 2003 in *IEEE Transactions on Power Systems* [23], see the narrow white bars in the right panel of Fig. 2. Between 2005 and 2009 further eight articles were published, including two papers in a special issue of the *International Journal of Forecasting* on 'Energy Forecasting' [24,25]. The topic picked up again in 2011 and averaged four articles per year in the period 2012–2015. A big change came in 2016 with the special issue on the GEFCom2014 competition [3], which included 7 papers on probabilistic EPF. As of 15 March 2017, 38 articles (and 9 conference papers) have been indexed by Scopus for the period 2002–2016, with another 4 articles published in 2017.

Regarding the methods used, there is no clear temporal pattern, see the right panel in Fig. 2. Overall, the share of 'neural network'-type (including support vector machines and fuzzy logic) methods exceeds

[2] Some authors have erroneously used the term *confidence interval* (CI) instead of *prediction interval* (PI) [11–13]. However, in most EPF applications we are interested in PIs associated with electricity prices yet to be observed, i.e., intervals which contain the true values of future prices with specified probability, not in CIs quantifying the uncertainty of a parameter estimate. See Hyndman [14] for a discussion.

[3] To search publication titles, abstracts and keywords for EPF-related phrases we have used the following Scopus query: `(TITLE(((((''electric*'' OR ''energy market'' OR ''power price'' OR ''power market'' OR ''power system'' OR pool OR ''market clearing'' OR ''energy clearing'') AND (price OR prices OR pricing)) OR lmp OR ''locational marginal price'') AND (forecast OR forecasts OR forecasting OR prediction OR predicting OR predictability OR ''predictive densit*'')) OR (''price forecasting'' AND ''smart grid*'')) OR TITLE-ABS(''electricity price forecasting'' OR ''forecasting electricity price'' OR ''day-ahead price forecasting'' OR ''day-ahead mar* price forecasting'' OR (gefcom2014 AND price) OR ((''electricity market'' OR ''electric energy market'') AND ''price forecasting'') OR (''electricity price'' AND (''prediction interval'' OR ''interval forecast'' OR ''density forecast'' OR ''probabilistic forecast'')))) AND NOT TITLE (''unit commitment'')) AND (EXCLUDE(AU-ID,"[No Author ID found]'' undefined))` and the equivalent WoS query. All look-ups have been further refined to exclude non-English language texts or include only specific document types.

[4] To search for probabilistic EPF papers the Scopus query given in footnote 3 was appended in front by: `(TITLE((''probabilistic'' AND ''forecasting'') OR interval OR density) OR TITLE-ABS-KEY(''probabilistic forecast*'' OR ''interval forecast*'' OR ''density forecast*'' OR ''prediction interval*'')) AND.`

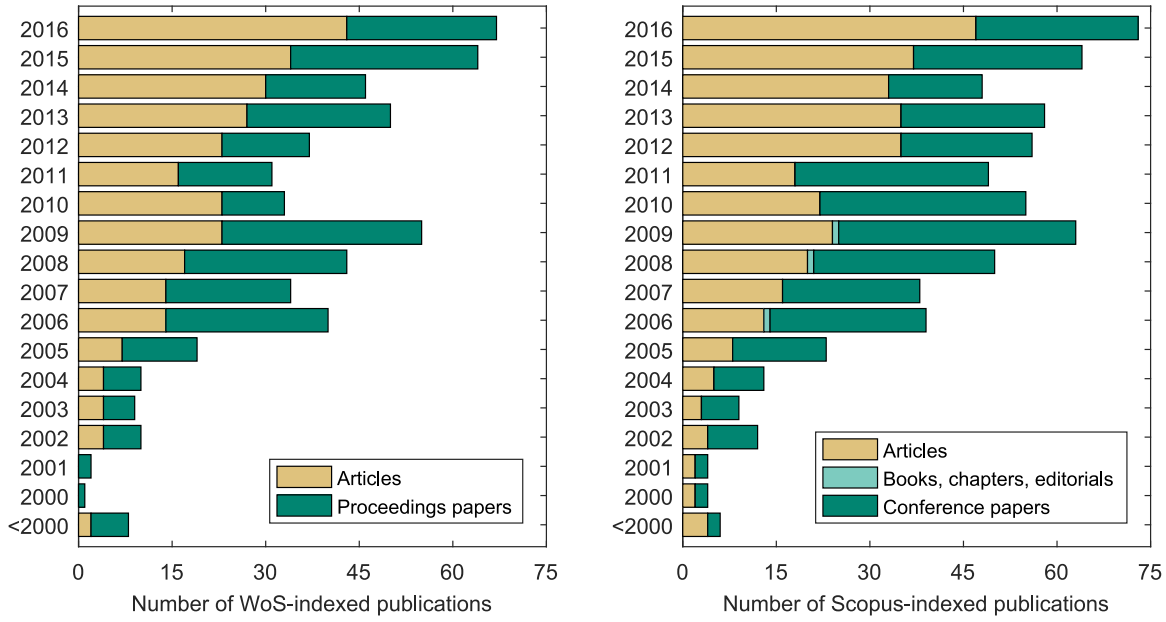

Fig. 1. The number of WoS- (*left panel*) and Scopus-indexed (*right panel*) electricity price forecasting (EPF) publications in the years 1992–2016. All publications prior to year 2000 (8 for WoS, 6 for Scopus) have been aggregated into one category '<2000'.

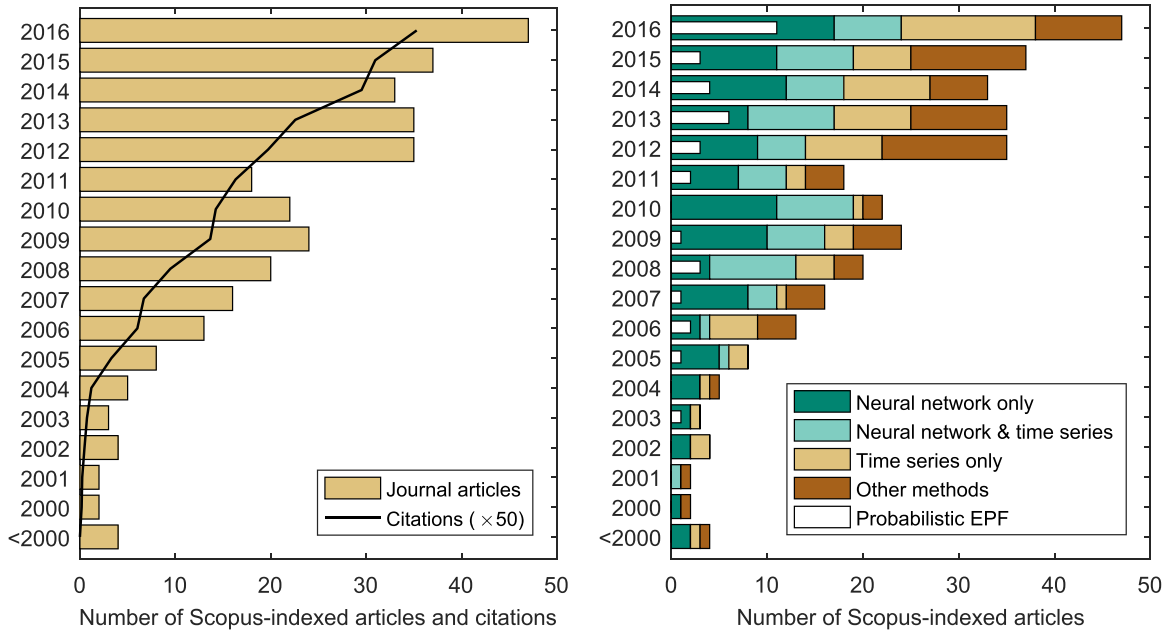

Fig. 2. *Left panel*: The number of Scopus-indexed EPF journal articles and citations to those articles in the years 1994–2016. Four articles prior to year 2000 have been aggregated into one category '<2000'. *Right panel*: The number of Scopus-indexed EPF journal articles in each of four 'method' classes (see text for details). Additionally, the number of probabilistic EPF papers in each year is indicated by a narrow white bar.

that of 'statistical time series' models. It should be noted, however, that the classification was automatic and possibly includes some errors. In particular, the look-up for 'statistical time series' methods is more complicated as there are many commonly used keywords and phrases.[5] Out of the 328 articles indexed by Scopus, the search yielded 184 'neural network'-type papers and 137 'statistical time series' papers. However, in some articles both types of methods are used, in other none of the tools automatically classified as coming from one of the two groups. If we consider four disjoint sets: (i) 'neural network' papers, (ii) papers where both 'neural network' and 'statistical time series' models are used, (iii) 'statistical time series' papers and (iv) papers where neither 'neural network' nor 'statistical time series' methods are used, then the overall count is 115, 69, 68 and 76, respectively. Apparently 'neural network'-type methods are nearly twice as popular as 'statistical time series' techniques.

Let us now see which are the most popular outlets for EPF articles, see the top panel in Fig. 3. Clearly the number one journal is *IEEE Transactions on Power Systems* with 35 publications (out of 328 indexed by Scopus). Like for other engineering journals, the share of

[5] To search for 'neural network'-type papers the Scopus query given in footnote 3 was appended in front by: `TITLE-ABS-KEY(''neural network'' OR ''support vector machine'' OR fuzzy) AND`, while for 'statistical time series' methods by: `(TITLE-ABS-KEY (''AR'' OR ''ARMA'' OR ''ARIMA'' OR ''GARCH'' OR ''VaR'' OR ''regression'' OR ''autoregressive'' OR ''autoregression'') OR ABS(''time series model'')) AND`.

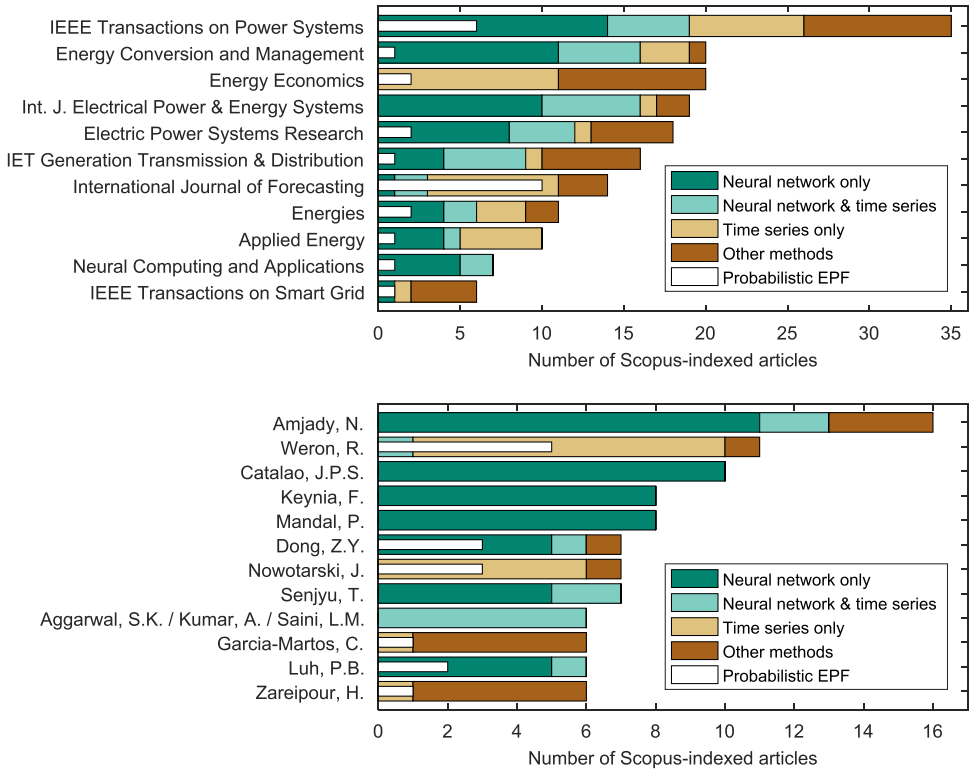

Fig. 3. *Top panel*: The number of Scopus-indexed EPF articles published in the years 1994–2016 in the 11 most popular outlets. *Bottom panel*: The number of Scopus-indexed EPF articles published in the years 1994–2016 by the 14 most prolific authors (S.K. Aggarwal, A. Kumar and L.M. Saini have jointly written all six of their EPF papers, hence are listed as 'one' author). In both panels the papers are subdivided into four 'method' classes and additional narrow white bars indicate the number of probabilistic EPF papers in each journal (*top panel*) and by each author (*bottom panel*).

'neural network'-type methods exceeds that of 'statistical time series' models. On the other hand, the latter methods are mainly published in non-engineering journals: *Energy Economics* and *International Journal of Forecasting*. In particular, not a single article published in *Energy Economics* involved neural networks, support vector machines or fuzzy logic. As argued by Weron [1], a likely reason for the latter situation is the difference in educational training of electrical engineers (focused on computational intelligence) and econometricians/statisticians (focused on regression and time series models), who constitute the two main groups of authors submitting papers to those two journal classes. Unfortunately, these differences in educational training have their consequences in the quality of research. Typically 'electrical engineering' papers consider sophisticated computational intelligence tools and relatively simple (or not properly applied) statistical models, while 'econometric' or 'statistical' papers usually show that (advanced) statistical models outperform (simple) computational intelligence techniques. There is definitely room for improvement and closer cooperation between the two communities.

In the bottom panel of Fig. 3 we summarize the research output of the 14 most prolific authors. The list is headed by Nima Amjady (Semnan University, Iran), who has (co-)authored 16 EPF articles, including the influential *Day-ahead price forecasting of electricity markets by a new fuzzy neural network* [26] (175 citations in Scopus since 2006; w/o self citations). The second on the list is Rafał Weron (Wrocław University of Technology, Poland) with 11 Scopus-indexed articles, including the most comprehensive EPF review to date – *Electricity price forecasting: A review of the state-of-the-art with a look into the future* [1] (103 citations since 2014); Weron has also authored the first monograph devoted to EPF [22] (344 citations in Scopus since 2006; included in Fig. 1, but not in Figs. 2–3). The third on the list is João P.S. Catalão (University of Beira Interior, Portugal), who has published nine EPF articles, including the highly cited *Short-term electricity prices forecasting in a competitive market: A neural network approach* [27] (154 citations since 2007). Generally, the top publishing authors are not very diversified in their use of forecasting tools – most specialize in 'neural network'-type techniques (Amjady, Catalão, Keynia, Mandal, Dong, Senjyu), some in 'statistical time series' models (Weron, Nowotarski) and a few in data-mining and dimension reduction procedures (Garcia-Martos, Zareipour). However, three of the listed authors have published only 'multi-method'/review-type papers (Aggarwal, Kumar and Saini have jointly written all six of their EPF papers, hence are listed as 'one' author in Fig. 3).

Regarding probabilistic forecasting, only two journals can boast a sizable amount of probabilistic EPF studies – *IEEE Transactions on Power Systems* (6 articles) and *International Journal of Forecasting* (10 articles; largely a result of the 2016 special issue on the GEFCom2014 competition), see the top panel in Fig. 3. Only three of the 14 most prolific EPF authors – Rafał Weron (Wrocław University of Technology, Poland) [1,12,24,28,29], Z.Y. (Joe) Dong (University of Sydney, Australia) [30–32] and Jakub Nowotarski (Wrocław University of Technology, Poland) [28,29,33] – have (co-)authored at least three probabilistic EPF articles, see the bottom panel in Fig. 3.

2.2. Winners of the GEFCom2014 price track

The dataset available to participants of the price track consisted of three time series at an hourly resolution: locational marginal prices and day-ahead predictions of zonal and system loads, see Section 5.1 for

details. During the competition the information set was being extended on a weekly basis. For the first of the 12 competition Tasks (or weeks; there were also three 'trial' pre-competition Tasks) almost 2.5 years of historical data was available. The objective was to forecast 99 quantiles[6] (as an approximation of the predictive distribution) of the next day's 24 hourly prices, i.e., arrays of 99×24 values.

Given the high participation rate (the price track attracted 287 contestants worldwide) and the unified forecast evaluation scheme (entries of all participants were ranked using the pinball loss function, see Section 4.2.2), the GEFCom2014 competition provided a unique, large scale test ground for PEPF, something that was missing in EPF studies thus far [1]. A total of 14 teams beat the benchmark and submitted final reports [3]. The top four teams submitted papers to the special issue; their methodology is discussed in this Section. Interestingly, three of them used quantile regression [34] as the main tool for obtaining quantiles of the predictive distribution. However, what is particularly worth emphasizing, the best performing models beat dozens of competitors in a fair 'battle' and, as such, are recommended for benchmarks in future probabilistic EPF research.

Team TOLOLO were the winners of both the load and price tracks. Gaillard, Goude and Nedellec [35] used three methods for the more challenging price track. The best on average approach, dubbed *quantGAM*, utilizes general additive models (GAM) introduced by Hastie and Tibshirani [36] and quantile regression. The former can be viewed as an extension of linear regression – the dependent variable is explained by a sum of smooth functions of the different covariates. The second best approach, dubbed *quantMixt*, is an extension of Quantile Regression Averaging (QRA) introduced by Nowotarski and Weron [28], see Section 3.5 for details, with up to 13 individual point forecasting models (including variants of autoregression, regression, GAM, random forests and gradient boosting) combined using a version of the ML-Poly forecaster. Finally, the third approach, dubbed *quantGLM*, is a kernel-based quantile regression with a lasso penalty [37]. Team TOLOLO started out by using several versions of quantMixt for Tasks 2–8, then different versions of quantGAM for Tasks 9–12 with particularly spiky prices and, finally, for Tasks 13–15 they used quantGLM, which is designed specifically for Winter. Like that of other teams, their methodology evolved over the course of the competition.

TEAM POLAND, ranked 2nd in the price track, proposed a hybrid approach which consists of four major blocks: point forecasting, pre-filtering, quantile regression modeling and post-processing. Maciejowska and Nowotarski [33] argue that their approach maintains a proper balance between flexibility and accuracy, and allows the blocks to be developed independently. Two autoregressive models with the same structure (but different calibration samples) are used to compute point EPFs. They are later used, together with other explanatory variables (hourly, mean daily and ratios of load forecasts, average daily price forecasts and their squares), in a quantile regression setting [34]. As such, that TEAM POLAND's approach can be viewed as yet another extension of QRA [28]. In the post-processing step, the 99 quantiles are sorted and the quantile curves smoothed. This last step is important, since the neighboring quantiles may be overlapping due to numerical inefficiency, a problem that is also known as *quantile crossing* [34,38].

Team GMD, ranked 3rd, used a relatively simple neural network for computing EPFs. Dudek's [39] model is based on a multilayer perceptron (MLP) with five sigmoid neurons in the hidden layer and one linear neuron in the output layer. To facilitate and accelerate the MLP learning, the input and output variables are preprocessed by mapping them to the interval $[-0.9, 0.9]$. In the first step, the MLP with system and zonal loads (original and squared) as the only input variables is used to obtain point EPFs. The parameters are estimated once for all 24 h of the next day using 312 hourly loads (i.e., data from 13 previous days). In the second step, the residuals are computed. Since they are assumed to be $N(0, \sigma^2)$, the computation of quantile forecasts is straightforward, see Section 3.2.

The C3 GREEN TEAM, ranked 4th, used machine learning techniques.[7] Juban et al. [40] explain that the core part of their model used quantile regression with a regularization term and included a variable selection procedure based on leave-one-out cross validation for linear regression (i.e., for point forecasting). The latter considers predicted loads, historical prices, maximum variation and standard deviation of the previous day's prices and calendar effects as potential explanatory variables. In the next step, the input variables are transformed using a radial basis function to incorporate non-linear dependencies. Finally, the quantile regression minimization problem is solved with the help of alternating direction method of multipliers (ADMM) [41].

2.3. Other notable probabilistic EPF papers

2.3.1. The first years

In the first journal article on probabilistic EPF, Zhang et al. [23] propose an algorithm for obtaining the PIs (which they erroneously call 'confidence intervals'; see footnote 2) from a cascaded neural network model. In a follow-up paper, Zhang and Luh [11] develop a modified U-D factorization method within the decoupled extended Kalman filter framework. The computational speed and numerical stability of this method are improved significantly relative to the earlier method. The new method also provides smaller PIs, though their quality is not formally assessed.

Another popular computational intelligence tool, the Support Vector Machine (SVM), has been used for the first time in probabilistic EFP by Zhao et al. [30]. They propose a data mining-based approach in order to achieve two major objectives: to forecast electricity spot prices (using the SVM) and to compute the PIs (by introducing a heteroskedastic variance equation to the SVM). Zhao et al. conclude that their method is highly effective relative to existing techniques such as GARCH models.

In the 'statistical time series' stream of EPF literature, Misiorek et al. [12] and Weron [22] were the first to consider probabilistic forecasts. For three expert[8] autoregressive models studied, Misiorek et al. compute the PIs (erroneously called 'confidence intervals') by taking the quantiles of a standard normal random variable rescaled by the standard deviation of the residuals in the calibration period (see Section 3.2 for details on the distribution-based PIs). For the Markov Regime-Switching (MRS; see [1,44]) model they use Monte Carlo simulations to obtain potential future values and, consequently, the empirical PIs. Misiorek et al. evaluate the quality of the PIs only by comparing the nominal coverage of the models to the true coverage, see Section 4.1.1. In the first monograph devoted to EPF, Weron [22] does not perform empirical analyses of probabilistic forecasts. However, he does discuss the construction of interval forecasts (historical simulation and distribution-based), which are implemented in the Matlab toolbox accompanying the book. In an article published in the same year, Zhou et al. [45] compute the PIs for SARIMA models fitted to California power market prices, but they use them only as a trigger to stop an iterative SARIMA estimation scheme and do not evaluate nor analyze the PIs.

In a study that complements Ref. [12], Weron and Misiorek [24] compare the accuracies of 12 expert time series models, and evaluate

[6] The qth *quantile* of random variable X is the value below which a fraction q of observations of this random variable fall, i.e., x_q satisfies $F_X(x_q) = q$, where F_X is the cummulative distribution function of X. A *sample quantile* refers to a value that splits the sample into subsamples of q and $(1 - q)$ observations. For example, the $q = 0.1$ or 10% quantile is the value below which 10% of the observations may be found. Quantiles $q = 0.01, 0.02, \ldots, 0.99$ are also called *percentiles*.

[7] Originally ranked 5th, but the team ranked 4th did not submit a valid report.

[8] We adopt the terminology of Uniejewski et al. [42] and Ziel [43] who refer to such parsimonious structures as *expert models*, since they are usually built on some prior knowledge of experts.

their performances in terms of one-step-ahead point and interval forecasts. Two types of PIs are computed: distribution-based and empirical (see Section 3.2). The former are computed as quantiles of the error term density: Gaussian for AR-type models and kernel estimator-implied for the semiparametric models. The reliability of the PIs is assessed with the Christoffersen test [46] for unconditional and conditional coverage, see Sections 4.1.1–4.1.2, which is an innovation in the EPF literature. Weron and Misiorek find that the semiparametric models, and SNARX in particular, generally lead to better PIs than their competitors, and also, more importantly, have the potential to perform well under diverse market conditions.

2.3.2. Density forecasts

In the first EPF paper that considers density forecasts, Panagiotelis and Smith [25] develop a first order vector autoregressive (VAR) model with exogenous effects and skew t distributed innovations within a Bayesian framework. They estimate the model using Markov Chain Monte Carlo and judge the effectiveness of their model by computing the Continuous Ranked Probability Score (CRPS; see Section 4.2.4) obtained from a 30 day forecasting trial. This is probably the first PEPF paper where the reliability and sharpness of the predictive densities was jointly evaluated by computing the CRPS, one of the measures recommended in Section 4.2.

Serinaldi [13] introduces the class of Generalized Additive Models for Location, Scale and Shape (GAMLSS) and computes the PIs (called 'confidence intervals') as the time-varying quantiles of the density forecasts. The accuracy of the PIs is checked by comparing the nominal coverage with the actual one. Surprisingly, the density forecasts themselves are not analyzed.

Huurman et al. [47] consider GARCH-type time-varying volatility models and compute density forecasts. To assess their reliability they use the probability integral transform (PIT) and the Berkowitz [48] test, see Section 4.1.4. Huurman et al. also measure the relative predictive accuracy by applying the Kullback-Leibler Information Criterion (KLIC) [49].

In a more recent paper, Jonsson et al. [50] develop a semi-parametric methodology for generating prediction densities by combining a time-adaptive quantile regression [34] model for the 5–95% quantiles with an exponential distribution for the tails. They jointly evaluate the reliability and sharpness of the predictive densities by computing the average CRPS (see Section 4.2.4) and the related Continuous Ranked Probability Skill Score (CRPSS).

2.3.3. Bootstrapped PIs

The method of constructing PIs via the bootstrap (see Section 3.4 for details) is very popular in the 'neural network' PEPF literature (though, it has also been used in 'statistical time series' papers [51]). For instance, Chen et al. [31] combine the extreme learning machine (ELM) with a wild (or external) bootstrap approach, and use them to compute point and interval forecasts of half-hourly spot prices in the Australian electricity market. The uncertainty of data noise is not considered in the construction of the PIs, and the accuracy of the PIs is only checked by comparing the nominal coverage with the actual one. In a follow-up paper, Wan et al. [32] first use the ELM to obtain point forecasts of half-hourly Australian spot prices, then use a bootstrap-based 'neural network' procedure (involving $N + 1$ additional neural networks) to compute the PIs. They use the Winkler score (see Section 4.2.3) to evaluate the PIs.

Khosravi et al. [52] use a neural network for point forecasts and estimate it with k-fold cross-validation (to determine the number of neurons in each of two hidden layers). In the second step, they apply the 'delta method' or the bootstrap to construct 90% PIs. The interval forecasts are evaluated with coverage, interval width and the flawed Coverage Width-based Criterion (CWC; see [53,54] and the discussion in Section 4.3). In a related article, Khosravi et al. [55] propose a hybrid method for the construction of PIs, which uses moving block bootstrapped neural networks and GARCH models for forecasting electricity prices. Rather than employing the traditional maximum likelihood estimation, the parameters of the GARCH model are adjusted via the minimization of a PI-based cost function. The authors claim that the proposed method generates narrow PIs with a large coverage probability, however, they again use the CWC to evaluate the intervals.

More recently, Rafiei et al. [56] consider a two-layer neural network with the clonal selection algorithm and extreme learning machine. Wavelets are used for pre-processing, to split the original time series into one approximation and three details series. The neural network is fitted to each of them and their model uncertainty is computed with the bootstrap. The data uncertainty is computed afterwards on the aggregated series, again via the bootstrap. The forecasts are evaluated only with descriptive statistics, coverage and mean width.

2.3.4. Factor models and medium-term forecasts

In a multivariate context, Garcia-Martos et al. [57] construct PIs based on one-day-ahead forecasts of the common volatility factors in the proposed GARCH-SeaDFA (Seasonal Dynamic Factor Analysis) model, but do not evaluate them. In a related study by the same research team, Alonso et al. [51] construct the PIs via the bootstrap (see Section 3.4), however, the evaluation is limited to just one week and assessed only with the coverage rate. Yet, the authors claim that the SeaDFA model allows to capture seasonality and forecast prices up to one year ahead.

Wu et al. [58] propose a recursive dynamic factor analysis (RDFA) algorithm, where the principal components (PC) are tracked recursively using a subspace tracking algorithm, while the PC scores are tracked further and predicted recursively via the Kalman filter. The PIs are obtained from the latter and their reliability is checked by comparing the nominal coverage with the actual one (called 'calibration bias') and their sharpness by computing the 'interval score' (i.e., the Winkler score, see Section 4.2.3).

In a recent paper, Bello et al. [59] use scenario generation and a market-equilibrium framework to compute the forecasts. A large number of scenarios is analyzed and transformed into 250 by spatial interpolation techniques. PEPFs are computed directly from them and evaluated with the pinball loss function (see Section 4.2.2) and coverage rate.

2.3.5. Spike occurrence and threshold forecasting

Price spikes are a characteristic feature of electricity markets. They may also play a special role in EPF. They can be treated as any other price (as in most of the reviewed above papers), treated as outliers [60] and the input prices pre-filtered to minimize or eliminate spikes [22,24,61–63] or they can be treated as the main object of study, as in spike occurrence and threshold forecasting.

Spike occurrence forecasting is similar to predicting individual quantiles that separate two regimes – normal and spiky prices. For instance, Christensen et al. [64] treat the time series of spikes as a discrete-time point process and represent it as a nonlinear variant of the autoregressive conditional hazard (ACH) model. They conclude that the ACH model performs better than the benchmark logit model in terms of MAE, RMSE and the log-probability score error (LPSE). Bello et al. [65] compare a number of methods: logistic regression, decision trees, multilayer perceptons as well as a hybrid approach that merges logistic regression with a fundamental market equilibrium model, and evaluate the forecasts with the Brier score [16,66].

Threshold forecasting is a generalization of spike occurrence forecasting, where the number of regimes is more than two. It could be also considered as a special case of interval forecasting where, instead of constructing a PI around a point forecast, a future price is allocated to one of a few prespecified price intervals spanning the entire range of attainable prices [1,7]. A nice example of threshold forecasting is the paper by Zareipour et al. [67], who use two SVM-based models to

classify future electricity prices in the Ontario and Alberta markets into three price groups with respect to prespecified price thresholds. They evaluate the forecasts using the mean percentage classification error (MPCE), i.e., a percentage of misclassifications.

3. Constructing probabilistic forecasts

3.1. Problem statement

To define the probabilistic forecasting problem let us start with a point forecast of the electricity spot price (i.e., the 'best guess' or expected value of the spot price[9]). Note that the actual price at time t, i.e. P_t, can be expressed as:

$$P_t = \widehat{P}_t + \varepsilon_t, \tag{1}$$

where $\widehat{P}_t$ is the point forecast of the spot price at time t made at an earlier point in time and ε_t is the corresponding error. In a vast majority of EPF papers the analysis ends at this point, since the authors focus only on point predictions, see [1] and [7] for reviews.

The most common extension from point to probabilistic forecasts is to construct *prediction intervals* (PIs). A number of methods can be used for this purpose, the most popular take into account both the point forecast and the corresponding error [24,29]: the center of the PI at the $(1-\alpha)$ confidence level is set equal to $\widehat{P}_t$ and its bounds are defined by the $\frac{\alpha}{2}$ th and $(1-\frac{\alpha}{2})$ th quantiles of the cumulative distribution function (CDF) of ε_t. For instance, for the commonly used 90% PIs, the 5% and 95% quantiles of the error term are required. We later denote such a PI of the spot price at time t by $[\widehat{L}_t, \widehat{U}_t]$, where $\widehat{L}_t$ and $\widehat{U}_t$ are the lower and upper bounds, respectively. We skip the nominal rate $(1-\alpha)$ for simplicity.

A forecaster may extend their study further and construct multiple PIs. The final outcome may be a set of quantiles on many levels, e.g., all 99 percentiles as in the GEFCom2014 competition. Such a set of 99 quantiles ($q = 1\%, 2\%, \ldots, 99\%$) is also a reasonable discretization of the price distribution. In general, a density forecast corresponding to Eq. (1) can be defined as a set of PIs for all $\alpha \in (0, 1)$. In other words, computing a probabilistic forecast requires estimation of $\widehat{P}_t$ and the distribution of ε_t. Equivalently, the problem can be formulated in terms of the inverse of the CDF of P_t and of ε_t:

$$F^{-1}_{P_t}(q) = \widehat{P}_t + F^{-1}_{\varepsilon_t}(q). \tag{2}$$

Note that splitting the probabilistic forecast into a point forecast and the distribution of the error term is not the only possible approach. The problem may be stated in a more general way. In particular, Gneiting and Katzfuss [17] define the probabilistic forecast as 'a forecast in the form of a probability distribution over future quantities or events' and associate it with a random variable. In our case this means finding the distribution of the electricity spot price itself, i.e., $\widehat{F}_{P_t}$. The latter approach is utilized in Quantile Regression Averaging (QRA; see [28] and Section 3.5 below).

Finally, two important aspects of the problem have to be mentioned at this point. First, in the above discussion we do not mention the probability density function (PDF) of ε_t. While there are papers in which probabilistic price forecasts are expressed in those terms, see Section 2, other authors argue that such a statement should be avoided. For instance, Ziel and Steinert [68] analyze aggregated supply and demand curves in the EPEX market and find that the price distribution is not continuous as it has additional point masses at certain prices. This implies the same conditions for the predicted distribution. Indeed, the authors predicted point probability masses up to around 20% (especially at 0 EUR/MWh).

[9] Note that although commonly used, the term *point forecast* is not precisely defined. While most often it refers to the mean of a future value, it may as well refer to the median.

The second point relates to the statistical nature of predicting day-ahead prices. Since 24 hourly predictive distributions have to be constructed at once, their cross-dependencies should be taken into account. However, most studies simplify the framework and predict 24 marginal distributions, and do not discuss their joint distribution. Such a 'simplistic' approach is also taken in Section 5. It should be noted, though, that in other areas of energy probabilistic forecasting this problem has already been addressed, see eg. [69]. Two main solutions are available. The first one is to model and predict the correlation between the marginal distributions. This, however, has a major drawback – it allows to capture only linear relationships between the hours and a question of proper evaluation arises. The second solution requires simulating 24-h paths of the day-ahead prices, which then can be treated as vectors from the joint 24-dimensional distribution.

3.2. Historical simulation

The method of calculating empirical (or sample) PIs is extremely simple and in the Value-at-Risk literature is known as *historical simulation* [70]. It is a model-independent approach which consists of computing sample quantiles of the empirical distribution of ε_t [1]. Later in the text we use the suffix **-H** to denote probabilistic forecasts obtained via historical simulation. EPF studies where PIs are obtained using this approach include [12,22,24,28] among others.

3.3. Distribution-based probabilistic forecasts

For time series models driven by Gaussian noise (AR, ARIMA, etc.), the density forecasts can be set equal to the Gaussian distribution approximating the error density and the PIs can be computed analytically as quantiles of this distribution [12]. Later in the text we use the suffix **-G** to denote such probabilistic forecasts. This approach differs from historical simulation in that first the standard deviation of the error density, $\hat{\sigma}$, is computed and then the lower and upper bounds of the PI are set equal to selected quantiles of the $N(0, \hat{\sigma}^2)$ distribution. The same approach can be used as long as the distribution of the noise term is parametric, for instance, student-t as in [25]. For time series models driven by non-parametric noise, like the IHMAR and SNAR models in [24,28,71], the distribution-based lower and upper bounds of the PI can be computed as quantiles of the kernel estimator of the PDF of ε_t. EPF studies where distribution-based PIs are computed include [12,24,25,29,30,39,72] among others.

3.4. Bootstrapped PIs

The third approach, commonly used in neural network EPF studies, is the *bootstrap*. For one step-ahead forecasts, the method consists of the following steps [73,74]:

1. Estimate the set of model parameters, $\widehat{\Theta}$, obtain a fit and the corresponding residuals, $\hat{\varepsilon}_t$.
2. Generate pseudo-data recursively using $\widehat{\Theta}$ and sampled normalized residuals ε_t^*.
   - For a model with no autoregression on P_t (like the neural network model of Dudek [39]; see also Section 5.2.5) simply set $P_t^* = \widehat{f}(X_t) + \varepsilon_t^*$, where ε_t^* is the sampled residual and $\widehat{f}(X_t)$ is an estimated function of exogenous variables X_t.
   - For an AR(1) model first set $P_1^* = P_1$ and then recursively put $P_t^* = \widehat{\beta} P_{t-1}^* + \varepsilon_t^*$ for all $t \in \{2, 3, \ldots, T\}$, where T is the time index of the last observation in the calibration window.
   - For a more general case of an autoregressive model of order r with exogenous variables first set $P_1^* = P_1, \ldots, P_r^* = P_r$ and then recursively put $P_t^* = \widehat{\beta}_1 P_{t-1}^* + \cdots + \widehat{\beta}_r P_{t-r}^* + \widehat{f}(X_t) + \varepsilon_t^*$ for all $t \in \{r+1, \ldots, T\}$.
3. Estimate the model again and compute the bootstrap-implied one step-ahead (point) forecast for time $t = T + 1$.

4. Repeat steps 2 and 3 B times and obtain the bootstrap sample of the forecasted price, $\{\widehat{P}^{i}_{T+1}\}^{B}_{i=1}$.
5. Compute desired quantiles of $\{\widehat{P}^{i}_{T+1}\}^{B}_{i=1}$ to obtain PIs.

The advantage of the bootstrap over historical simulation or distribution-based PIs is that it takes into account not only historical forecast errors but also parameter uncertainty. The disadvantage is the significantly increased computational burden. Later in the text we use the suffix **-B** to denote probabilistic forecasts obtained via the bootstrap. EPF studies where this approach is used to compute PIs include [31,32,51,52,56,65] among others.

3.5. Quantile Regression Averaging

The fourth method we discuss is *Quantile Regression Averaging* (QRA), proposed by Nowotarski and Weron [28]. Its very good forecasting performance has been verified by a number of authors [29,72,75], not only in the area of EPF [6,76]. However, its most spectacular success came during the GEFCom2014 competition – the top two winning teams in the price track used variants of QRA [33,35], see Section 2.2.

The method involves applying quantile regression, see [34], to a pool of point forecasts of individual (i.e., not combined) forecasting models. As such, it directly works with the distribution of the electricity spot price, $\widehat{F}_{P_t}$, without the need to split the probabilistic forecast into a point forecast and the distribution of the error term (see Section 3.1 for a discussion). The quantile regression problem can be written as follows:

$$Q_{P_t}(q|X_t) = X_t\beta_q, \tag{3}$$

where $Q_{P_t}(q|\cdot)$ is the conditional q-th quantile of the electricity price distribution, X_t are the explanatory variables (or regressors) and β_q is a vector of parameters for quantile q. The parameters are estimated by minimizing the loss function for a particular q-th quantile:

$$\min_{\beta_q}\left[\sum_{\{t:P_t\geq X_t\beta_q\}} q|P_t - X_t\beta_q| + \sum_{\{t:P_t<X_t\beta_q\}} (1-q)|P_t - X_t\beta_q|\right]$$
$$= \min_{\beta_q}\left[\sum_{t}(q - \mathbb{1}_{P_t<X_t\beta_q})(P_t - X_t\beta_q)\right]. \tag{4}$$

In the first papers on QRA [28,75], the regressors were the point forecasts of m individual models: $X_t = [1, \widehat{P}_{1,t}, \ldots, \widehat{P}_{m,t}]$. The choice of the number of individual models can be made arbitrarily (the best three models, all models, etc.; see the empirical study in Section 5) or, in case of dozens of competing models, using dimension reduction techniques [29]. During the GEFCom2014 competition the vector of explanatory variables was further expanded to include important exogenous variables (hourly, mean daily and ratios of load forecasts, average daily price forecasts and their squares) [33]. There are no limits as to the components of X_t, as long as it includes forecasts of individual models the method can still be regarded as QRA. The Matlab function `qra.m` that allows to run QRA on a pool of point forecasts is available from the HSC RePEc repository (https://ideas.repec.org/s/wuu/hscode.html).

4. Evaluation metrics

When evaluating a probabilistic forecast, the main challenge is that we never observe the true distribution of the underlying process. In other words, we cannot compare the predictive distribution, $\widehat{F}_{P_t}$, or the prediction interval, $[\widehat{L}_t, \widehat{U}_t]$, with the actual distribution of the electricity spot price, F_{P_t}, only with observed past prices, P_τ, $\tau < t$.

Over the years, a number of ways have been developed to evaluate probabilistic forecasts. The approach depends on the forecasting target – a quantile forecast requires a different evaluation than a predictive distribution, but sometimes it may also depend on the preference of a forecaster. Some methods admit formal statistical tests, while other result in a single number which has a clear interpretation and is easy to compare. We summarize the more popular evaluation metrics in Table 1. Note, however, that the Table does not include measures that can be found in the EPF literature but are not recommended, we discuss some of them in Section 4.3.

In a series of papers on probabilistic forecasting, Gneiting et al. [15–17] argue that 'probabilistic forecasting aims to maximize the sharpness of the predictive distributions, subject to reliability'. *Reliability* (also called *calibration* or *unbiasedness*) refers to the statistical consistency between the distributional forecasts and the observations. For instance, if a 90% PI covers 90% of the observed prices, then this PI is said to be reliable[10] [18,96], well calibrated[11] [15–17] or unbiased [97]. *Sharpness*, on the other hand, refers to how tightly the predicted distribution covers the actual one, i.e., to the concentration of the predictive distributions. This definition derives from the idea that reliable predictive distributions of null width would correspond to perfect point predictions [16,18]. Unlike reliability, which is a joint property of the predictions and the observations, sharpness is a property of the forecasts only. In Section 4.1 we discuss methods for the evaluation of reliability for different types of probabilistic forecasts, then do the same for sharpness in Section 4.2.

Note that there is one more commonly used attribute for probabilistic forecast evaluation, especially in the meteorological or wind power forecasting literature [18,19]. *Resolution* refers to how much the predicted density varies over time, stated differently, to the ability of providing probabilistic forecasts (e.g., wind power) conditional to the forecast conditions (e.g., wind direction). As Pinson et al. [18] note, sharpness and resolution are equivalent when probabilistic forecasts have perfect reliability. In view of the 'maximizing sharpness subject to reliability' paradigm we advocate, the evaluation of resolution is not critical. Hence, we do not discuss it any further.

4.1. Reliability

4.1.1. Unconditional coverage and the Kupiec test

Let us start with prediction intervals. The simplest and the most common approach for assessing the quality PIs is the *unconditional coverage* (UC). By definition, the empirical coverage should match the nominal rate: $\mathbb{P}(P_t \in [\widehat{L}_t, \widehat{U}_t]) = (1-\alpha)$. For instance, the 90% PI (i.e., with $\alpha = 10\%$) should yield the nominal coverage of 90%. To obtain the empirical coverage we typically focus on the indicator I_t series of 'hits and misses':

$$I_t = \begin{cases} 1 & \text{if } P_t \in [\widehat{L}_t, \widehat{U}_t] \rightarrow \text{'hit'}, \\ 0 & \text{if } P_t \notin [\widehat{L}_t, \widehat{U}_t] \rightarrow \text{' miss' (or 'violation')}. \end{cases} \tag{5}$$

Note that I_t may be also considered for individual quantiles, as is common in the risk management (Value-at-Risk) literature [70,82]. In such a case, the forecaster predicts only the lower or the upper quantile, i.e., in Eq. (5) either $\widehat{U}_t$ is replaced by ∞ or $\widehat{L}_t$ by $-\infty$. Some authors simply report the empirical coverage itself (sometimes called the *PI coverage probability*, PICP), while others subtract it from the nominal level (sometimes called the *PI nominal coverage*, PINC) to obtain the *average coverage error* (ACE = PICP - PINC), see e.g. [32,56]. Either way, the conclusions from the comparison will be the same.

[10] Note that in the electric power industry the term 'reliability' is often used to describe the ability of power systems to perform the required functions under stated conditions. For this reason, in their review on probabilistic load forecasting, Hong and Fan [5] use the term 'unconditional coverage' as a substitute for 'reliability'. However, our perspective is that 'reliability' is a broader concept than 'unconditional coverage' and covers 'independence' and 'conditional coverage' as well. Since we are not concerned here with power system performance, using the term 'reliability' does not lead to ambiguity.

[11] To avoid ambiguity, in this paper we use the term 'calibration' only as a substitute for 'estimation', when we refer to the process of estimating the parameters of a model.

Table 1
A comparison of evaluation metrics for probabilistic forecasting. Statistics and tests in *italics* are discussed in the text, but not illustrated in the empirical study in Section 5.

Interval forecasts		**Density forecasts**	
Statistics	**Tests**	**Statistics**	**Tests**
Reliability/calibration/unbiasedness			
Unconditional coverage [46,77]	Kupiec [77]	Probability Integral Transform (PIT) [15,78]	Visual 'tests' [15,17] *Tests for uniformity* [79,80]
Conditional coverage [46] (CC = UC + Independence)	Christoffersen [46] (*Lagged* [81]) *Ljung-Box Christoffersen* [82] *Duration-based tests* [83,84] *Dynamic Quantile (DQ)* [85] *VQR* [86]	Berkowitz CC statistic [48]	Berkowitz [48]
Sharpness (and reliability)			
Pinball loss [87,88] Winkler (interval) score [89]	Diebold-Mariano [90,91] *Model confidence set* [92] *Forecast encompassing* [93]	Continuous Ranked Probability Score (CRPS) [16,94] *Logarithmic score* [95]	Diebold-Mariano [90,91] *Model confidence set* [92] *Forecast encompassing* [93]

Generally, the closer is the empirical coverage to the nominal rate the better. However, if we want to know if 'close is close enough' we have to run a formal statistical test. The Kupiec [77] test checks whether $\mathbb{P}(I_t = 1) = (1 - \alpha)$ under the assumption that the violations are independent, which is equivalent to testing that the sequence I_t is identically and independently distributed (i.i.d.) Bernoulli with mean $(1 - \alpha)$. The test rejects the null hypothesis of an accurate PI if the actual fraction of PI violations is statistically different than α. The Kupiec test is carried out in the likelihood ratio (LR) framework. The LR statistics for unconditional coverage:

$$LR_{UC} = -2\log\left\{\frac{(1-c)^{n_0}c^{n_1}}{(1-\pi)^{n_0}\pi^{n_1}}\right\} \tag{6}$$

is distributed asymptotically as $\chi^2(1)$ [46,77]. Here $c = (1 - \alpha)$ is the nominal coverage rate, $\pi = n_1/(n_0 + n_1)$ is the percentage of 'hits' and n_0 and n_1 are respectively the number of zeros and ones in the indicator I_t series.

4.1.2. Independence, conditional coverage and the Christoffersen test

As noted by Christoffersen [46], the Kupiec [77] test evaluates the coverage of the PI but it does not have any power against the alternative that the ones and zeros come clustered together in the indicator I_t series. In other words, in the Kupiec test the order of the PI violations does not matter, only the total number of violations plays a role. To make up for this deficiency, Christoffersen introduced the *independence* and *conditional coverage* (CC) tests; the latter is simply a joint test for independence and UC. Note that some authors use the term 'Christoffersen test' to refer to all three tests (UC, independence, CC), see [1,7].

Both tests are carried out in the LR framework. Independence is tested against an explicit first-order Markov alternative. Hence, the LR statistics for independence is given by [46]:

$$LR_{Ind} = -2\log\left\{\frac{(1-\pi_2)^{n_{00}+n_{10}}\pi_2^{n_{01}+n_{11}}}{(1-\pi_{01})^{n_{00}}\pi_{01}^{n_{01}}(1-\pi_{11})^{n_{10}}\pi_{11}^{n_{11}}}\right\}, \tag{7}$$

where $\pi_2 = (n_{01} + n_{11})/(n_{00} + n_{10} + n_{01} + n_{11})$, n_{ij} is the number of observations with value i followed by j and $\pi_{ij} = \mathbb{P}(I_t = j|I_{t-1} = i)$. Like LR_{UC}, also LR_{Ind} is distributed asymptotically as $\chi^2(1)$. Furthermore, if we condition on the first observation, then the conditional coverage LR test statistics is the sum of the other two, i.e., $LR_{CC} = LR_{UC} + LR_{Ind}$, and is distributed asymptotically as $\chi^2(2)$.

The Matlab function `christof.m` that allows to run all three tests (i.e., UC, independence and CC) is available from the HSC RePEc repository (https://ideas.repec.org/s/wuu/hscode.html). Note, however, that since the day-ahead electricity price forecasts typically use the same information set for predicting the next day's prices and hence are correlated by construction, the tests are usually conducted separately for each of the 24 h [1,7,24,28,29,72].

4.1.3. Extensions and alternatives to the Christoffersen test

As Clements and Taylor [81] note, we can conduct the independence test (and consequently the CC test) for any time lag h, in order to capture more than just the first-order dependency. The idea of the independence test is based on the Markov chain framework, and relies on investigating transition probabilities $\pi_{ij}^h = \mathbb{P}(I_t = j|I_{t-h} = i)$ for $h = 1$. However, the latter restriction is not crucial. We can relax it and test independence of PI violations for any time lag h. Maciejowska et al. [29] argue that testing independence makes particular sense for $h = 1, 2$ and 7 days, as these lags are typically the most significant when modeling and forecasting electricity spot prices. Note that the mentioned in Section 4.1.2 Matlab function `christof.m` allows to run the lagged Christoffersen test as well.

Berkowitz et al. [82] go a step further and suggest to use the Ljung-Box statistics for a joint test of independence for the first h lags. Finally, Wallis [79] recasts Christoffersen's tests in the framework of χ^2 statistics, and considers their extension to density forecasts. The use of contingency tables allows for the incorporation of a more informative decomposition of the χ^2 goodness-of-fit statistic and the calculation of exact small-sample distributions.

The popular in the risk management literature Dynamic Quantile (DQ) test of Engle and Manganelli [85] goes in a different direction. It is based on a linear regression model of the violations variable on a set of explanatory variables including a constant, the lagged values of the violations variable and any function of the past information set suspected of being informative (for instance, the lower $\widehat{L}_t$ and upper $\widehat{U}_t$ quantiles themselves). The DQ test rejects the PIs if the intercept is significantly different from $(1 - \alpha)$ or the remaining coefficients are significantly different from zero. There are also duration-based tests, which check if the duration (i.e., the time interval) between violations of the PI is unpredictable [83]. However, as shown in [82], the DQ test has more power against misspecified PIs than the duration-based tests and is the preferred option. Gaglianone et al. [86] argue that using only binary variables, such as whether or not there was a violation, sacrifices too much information. They propose the VQR ('Value-at-Risk model based on quantile regressions') test, which uses more information to reject a misspecified model and, hence, has more power in finite samples than the Christoffersen or the DQ tests.

4.1.4. Probability Integral Transform (PIT) and the Berkowitz test

Testing for the goodness-of-fit of a predictive distribution is, in general, more challenging than evaluating the reliability of a PI. Dawid's [78] so-called *prequential principle* states that the predictive distributions need to be assessed on the basis of the forecast-observation pairs $(\widehat{F}_{P_t}, P_t)$ only, regardless of their origins. Indeed, the true distributions, F_{P_t}, are unknown, hence standard goodness-of-fit tests cannot be utilized. In this context, Dawid [78] proposed the use of the *Probability Integral Transform*:

$$PIT_t = \widehat{F}_{P_t}(P_t), \tag{8}$$

which can be traced back at least to the works of Karl Pearson in 1930s, see [15]. If the distributional forecast, $\widehat{F}_{P_t}$, is perfect (i.e., is the same as the true distribution of the spot price process, F_{P_t}), then PIT_t is independent and uniformly distributed [98]. Although this problem formulation enables us to utilize statistical tests, see e.g. [79,80], the common approach is to assess the uniformity and independence graphically [17]. The tools to examine it are the histogram (if the forecast is constructed properly, the histogram of PIT_t shows a uniform distribution) and the plot of the autocorrelation function, respectively. Non-uniformity may lead to quick conclusions how to improve the model. For instance, a histogram with too much probability mass in the center (inverse U-shape) indicates that the predictive distribution has too fat tails. Conversely, a U-shape suggests that the tails of the predictive distribution are not heavy enough.

In the risk management literature the following transformation of PIT has been popularized by Berkowitz [48]:

$$\nu_t = \Phi^{-1}(PIT_t) = \Phi^{-1}(\widehat{F}_{P_t}(P_t)), \tag{9}$$

where $\Phi^{-1}(\cdot)$ is the inverse of the standard normal distribution function. The argument behind it is that in finite-samples tests based on the Gaussian likelihood are more convenient and flexible than tests of uniformity. Given the transformed sequence ν_t, we can test the null hypothesis of independence and normality against a first-order autoregressive alternative with mean and variance possibly different from 0 and 1, respectively. Writing down the first-order autoregression:

$$\nu_t - \mu = \rho(\nu_{t-1} - \mu) + \varepsilon_t, \tag{10}$$

the null hypothesis becomes equivalent to $\mu = 0$, $\sigma^2 = \mathrm{Var}(\varepsilon_t) = 1$ and $\rho = 0$.

Like the Kupiec [77] and Christoffersen [46] tests, the Berkowitz test is carried out in the likelihood ratio (LR) framework. The LR statistics for independence:

$$LR_{Ind}^{Ber} = -2\{L(\hat{\mu}, \hat{\sigma}^2, 0) - L(\hat{\mu}, \hat{\sigma}^2, \hat{\rho})\}, \tag{11}$$

where $L(\cdot, \cdot, \cdot)$ is the standard normal log-likelihood function and the hats denote estimated values, is distributed as $\chi^2(1)$. Moreover, the LR statistics for a joint test of independence and normality (or conditional coverage):

$$LR_{CC}^{Ber} = -2\{L(0, 1, 0) - L(\hat{\mu}, \hat{\sigma}^2, \hat{\rho})\} \tag{12}$$

is distributed as $\chi^2(3)$ [48]. The implementation of the test is straightforward. The Matlab function `berkowitz.m` that allows to run the joint test is available as part of the MFE Toolbox from Kevin Sheppard's webpage (http://www.kevinsheppard.com/MFE_Toolbox).

4.2. Sharpness

4.2.1. Proper scoring rules

Sharpness, a measure of concentration of the predictive distribution, is closely related to the concept of the so-called proper scoring rules. Recall, that *scoring rules* provide summary measures for the evaluation of probabilistic forecasts, by assigning a numerical score, $S(\widehat{F}_{P_t}, P_t)$, based on the predictive distribution, $\widehat{F}_{P_t}$, and on the actually observed price, P_t [16,94]. In fact, scoring rules assess reliability and sharpness simultaneously [17]. A *proper scoring rule* is designed in such a way that quoting the true distribution as the forecast distribution is an optimal strategy in expectation, i.e., it minimizes the score. More formally, denote by $\overline{S}(\widehat{F}_{P_t}, F_{P_t})$ the expected value of $S(\widehat{F}_{P_t}, P_t)$ under the true price distribution of P_t. A scoring rule S is proper if $\overline{S}(F_{P_t}, F_{P_t}) \leq \overline{S}(\widehat{F}_{P_t}, F_{P_t})$ for any probabilistic forecast $\widehat{F}_{P_t}$ and any true distribution F_{P_t}. The term *proper* was coined by Winkler and Murphy [99], but the idea dates back at least to Brier [66].

In Sections 4.2.2–4.2.4, we present three proper scoring rules that have seen limited use in probabilistic energy forecasting [3,5,6,19,29,32,50,58,72,100,101] and definitely deserve to be recommended. As Gneiting and Raftery [16] emphasize, *score propriety* is essential in forecast evaluation. They also discuss potential issues that result from the use of intuitively appealing but improper scoring rules. Unfortunately, as the case of the relatively popular, but improper CWC score [52,102] shows, score propriety has not received enough attention in the probabilistic energy forecasting literature, see Section 4.3.

4.2.2. Pinball loss

The pinball loss gained popularity during the GEFCom2014 competition, where it was used as the scoring function of the contestants' entries [3]. It was chosen over the more popular in probabilistic forecasting, but conceptually more complex Continuous Ranked Probability Score (CRPS; see Section 4.2.4). The *pinball loss* is a special case of an *asymmetric piecewise linear loss* function [87,88,103]:

$$Pinball(\widehat{Q}_{P_t}(q), P_t, q) = \begin{cases} (1-q)(\widehat{Q}_{P_t}(q) - P_t), & \text{for } P_t < \widehat{Q}_{P_t}(q), \\ q(P_t - \widehat{Q}_{P_t}(q)), & \text{for } P_t \geq \widehat{Q}_{P_t}(q), \end{cases} \tag{13}$$

where $\widehat{Q}_{P_t}(q)$ is the price forecast at the q-th quantile and P_t is the actually observed price. This proper scoring rule is also known in the literature as the *linlin*, *bilinear* or *newsboy* loss [103,104]; the latter name refers to a newsboy who must order papers when he is uncertain about the demand and unsold papers are worthless to him [94]. Note that pinball loss is the function to be minimized in quantile regression [6,34] and is similar to Eq. (4), the loss function minimized in Quantile Regression Averaging [28]. Secondly, the loss function in Eq. (13) is also the loss function for a regression problem with asymmetric Laplace density assumption for the residuals (instead of Gaussian as in the standard OLS). The target quantile is the asymmetry parameter of the density [105].

The pinball loss, as defined by Eq. (13), is a measure of fit for one quantile only. It can be averaged across different quantiles to provide an aggregate score. Note that in the GEFCom2014 competition it was averaged not only across 99 quantiles ($q = 1\%, 2\%, \ldots, 99\%$; i.e., percentiles), but also across the 24 h of the target day [3]. A lower score indicates a better probabilistic forecast.

4.2.3. Winkler score

When faced by multiple PIs with similarly accurate levels of coverage, our preference is to choose the narrowest intervals. Interestingly, reliability and interval width can be assessed jointly using the score function that was proposed by Winkler [89] and is now known as the *Winkler* or *interval score* [16]. For a central $(1-\alpha) \times 100\%$ prediction interval it is defined as:

$$Winkler_t = \begin{cases} \delta_t, & \text{for } P_t \in [\widehat{L}_t, \widehat{U}_t], \\ \delta_t + \frac{2}{\alpha}(\widehat{L}_t - P_t), & \text{for } P_t < \widehat{L}_t, \\ \delta_t + \frac{2}{\alpha}(P_t - \widehat{U}_t), & \text{for } P_t > \widehat{U}_t, \end{cases} \tag{14}$$

where $\widehat{L}_t$ and $\widehat{U}_t$ are respectively the lower and upper bounds of the PI, $\delta_t = \widehat{U}_t - \widehat{L}_t$ is the interval width and P_t is the actual price. The Winkler score gives a penalty if an observation (the actual price) lies outside the constructed interval and rewards a forecaster for a narrow PI; naturally the lower the score the better the PI. Note that the Winkler score, like the pinball score, is a proper scoring rule, which makes it an appealing measure for PI evaluation.

4.2.4. Continuous Ranked Probability Score (CRPS)

The *logarithmic score* [95], also known as *predictive deviance* or the *ignorance score*, is a popular proper scoring rule that has many desirable properties, but lacks robustness [16]. It is calculated as the negative of the logarithm of the predictive density evaluated at the observed electricity price, P_t. This restriction to density forecasts can be impractical, however, and the *Continuous Ranked Probability Score* (CRPS) is defined directly in terms of the predictive CDF, $\widehat{F}_{P_t}$:

$$CRPS(\widehat{F}_{P_t}, P_t) = \int_{-\infty}^{\infty} (\widehat{F}_{P_t}(x) - \mathbb{1}_{\{P_t \le x\}})^2 dx, \tag{15}$$

where $\mathbb{1}$ is the indicator function. The idea behind the CRPS can be traced back to the article of Matheson and Winkler [94], but the name itself was probably used for the first time by Unger [106]. The CRPS has several appealing properties [107]: (i) its definition does not require the introduction of a number of predefined classes (e.g., quantiles in the pinball score) on which results may depend, (ii) for a deterministic forecast, it is equal to the well known Mean Absolute Error, and (iii) it can be interpreted as an integral over all possible Brier scores [66]. However, from a practical perspective, the integral in Eq. (15) poses numerical difficulty [16,106].

Interestingly, the CRPS can be defined equivalently as follows:

$$CRPS(\widehat{F}_{P_t}, P_t) = \int_0^1 \{Pinball(\widehat{Q}_{P_t}(q), P_t, q)\} dq = \tag{16}$$

$$= \mathbb{E}_{\widehat{F}_{P_t}} |Y_t - P_t| - \frac{1}{2} \mathbb{E}_{\widehat{F}_{P_t}} |Y_t - Y'_t|, \tag{17}$$

where $\widehat{Q}_{P_t}(q)$ is a q-quantile forecast of the electricity price, and random variables Y_t and Y'_t are two independent copies distributed as $\widehat{F}_{P_t}$. Formula (16) creates a direct link to the pinball loss function (13). Its discretization, e.g., replacing the integral by a sum over quantiles $q = 0.01, \ldots, 0.99$, enables us to avoid the complications with the direct use of Eq. (15); we use this approach in the empirical study in Section 5. Formula (17), on the other hand, is a decomposition of the CRPS into absolute differences (first component; which reduces to the absolute error if $\widehat{F}_{P_t}$ is a point forecast) and spread (second component; which measures the lack of sharpness); it has been utilized recently by Taieb et al. [108] in the context of forecasting uncertainty in smart meter data.

4.2.5. Equal predictive performance and the Diebold-Mariano test

Quite often we are faced with a situation when we have two (or more) competing forecasting methods and we wish to find the best one. We may rank them by their average score over a test set:

$$\widehat{S} = \frac{1}{T} \sum_{t=1}^{T} S(\widehat{F}_{P_t}, P_t), \tag{18}$$

where $S(\cdot, \cdot)$ is a score function and T is the length of the out-of-sample test period. However, we may wish to test if one method significantly outperforms the other, more formally, to test the hypothesis that these two methods have equal predictive performance.

The extremely simple Diebold-Mariano (DM) [90] test can be used for exactly this purpose; see Diebold [91] for a recent discussion of its uses and abuses. Although the DM test is much more popular in the point forecasting literature, in particular on EPF [71,109–114], it is readily applicable to probabilistic forecasts. Indeed, Tastu et al. [115] and Baran and Lerch [116], among others, conduct DM tests for probabilistic wind forecasts. However, to the best of our knowledge, our paper is the first where the DM test is used for evaluating probabilistic EPFs.

The DM test is simply an asymptotic z-test of the hypothesis that the mean of the *loss differential* series:

$$d_t = S_1(\widehat{F}_{P_t}, P_t) - S_2(\widehat{F}_{P_t}, P_t) \tag{19}$$

is zero [1,91], where $S_i(\cdot, \cdot)$ is the score (or loss) of model i. Note that in the point forecasting context this may simply be the squared loss, $S_i^2(\widehat{P}_t, P_t) = \varepsilon_t^2 = (\widehat{P}_t - P_t)^2$, or the absolute loss, $S_i^1(\widehat{P}_t, P_t) = |\varepsilon_t| = |\widehat{P}_t - P_t|$. In the probabilistic forecasting context the score may be any proper scoring rule, in particular the discussed above Pinball loss, Winkler score or CRPS. Given the loss differential series, we compute the statistic:

$$\mathrm{DM} = \sqrt{T} \frac{\widehat{\mu}_{d_t}}{\widehat{\sigma}_{d_t}}, \tag{20}$$

where $\widehat{\mu}_{d_t}$ and $\widehat{\sigma}_{d_t}$ are the sample mean and standard deviation of d_t, respectively, and T is the length of the out-of-sample test period. The key hypothesis of equal predictive accuracy (i.e., equal expected loss) corresponds to $\mathbb{E}(d_t) = 0$, in which case, under the assumption of covariance stationarity of d_t, the DM statistic is asymptotically standard normal, and one- or two-sided asymptotic tail probabilities are readily calculated. Many statistical computing environments, like Matlab or R, nowadays include the DM test in the standard releases or as add-ins.

In practice, typically two one-sided DM tests at the 5% significance level are conducted: (i) a standard test with the null hypothesis H_0: $\mathbb{E}(d_t) \le 0$, i.e. the outperformance of the forecasts of model 2 by those of model 1, and (ii) the complementary test with the reverse null H_0^R: $\mathbb{E}(d_t) \ge 0$, i.e., the outperformance of the forecasts of model 1 by those of model 2. To avoid a common mistake, we should remember that the DM test compares forecasts of two models, not the models themselves [91].

In day-ahead electricity markets the predictions for all 24 h of the next day are usually made at the same time using the same information set and hence forecast errors for a particular day will typically exhibit high serial correlation. Therefore, it is advisable to conduct the DM tests for each load period (e.g., each hour of the day) separately [42,71,112,117]. Even then, we should formally check that the forecasts for consecutive days, hence loss differentials, are not serially correlated. As reported by Uniejewski et al. [42], this is a generally valid assumption for well performing EPF models.

4.2.6. Alternatives to the Diebold-Mariano test

Alternative forecast comparison test procedures to the Diebold-Mariano [90] test include the *model confidence set* (MCS) approach of Hansen et al. [92] and a test of *forecast encompassing* (FE) [93]. For two models, the MCS approach is similar to the DM test but estimates the distribution of the test statistic by a bootstrap procedure. In the test of *forecast encompassing*, on the other hand, the null hypothesis is that model 2 encompasses model 1, i.e., that predictions of model 1 do not contain additional information with respect to those of model 2. In one of the few applications in EPF, Bordignon et al. [112] perform both tests to evaluate combined point forecasts.

4.3. Other measures

A number of other evaluation metrics can be found in the probabilistic forecasting literature. One example is the PI width, sometimes normalized by the range of prices. If a PI is constructed properly, with correct coverage rate, it is a good way to assess the concentration of the predictive distribution. However, these two are combined in the Winkler score (see Section 4.2.3) and hence the latter is preferred.

Another evaluation metric that has seen widespread use in energy forecasting is the so-called *Coverage Width-based Criterion* [52,102]:

$$\mathrm{CWC} = \overline{\delta}_t \{1 + \mathbb{1}(\Delta b > 0) e^{\eta \Delta b}\}, \tag{21}$$

where $\mathbb{1}$ is the indicator function, Δb is the difference between nominal and empirical coverage rates, $\overline{\delta}_t$ is the average width of the PIs and $\eta > 0$ is a free parameter that can be set to any positive value. The metric was justly criticized by Pinson and Tastu [53] and Wan et al. [54]. In particular, Pinson and Tastu show that CWC is not a proper scoring rule (see Section 4.2.1) and argue that when using it 'one can never conclude on the respective quality of the interval forecasts being evaluated'. This critique was rebutted by Khosravi and Nahavandi [102], who proposed a slightly modified version of the measure:

$$\mathrm{CWC}_{mod} = \overline{\delta}_t + \mathbb{1}(\Delta b > 0) e^{\eta \Delta b} \tag{22}$$

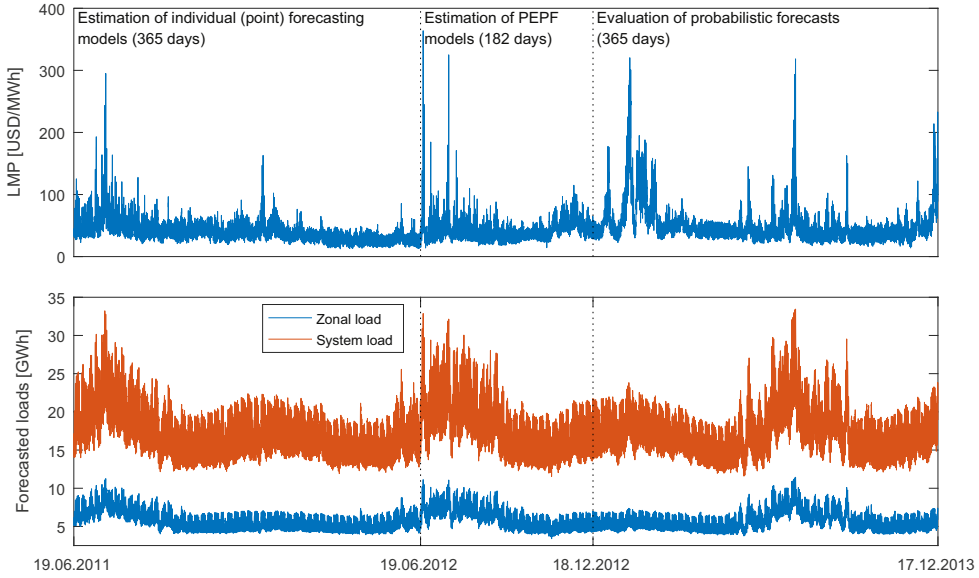

Fig. 4. GEFCom2014 hourly locational marginal prices (LMP; *top*) and hourly day-ahead predictions of the zonal and system loads (*bottom*) for the period 19 June 2011 to 17 December 2013. The first 365 days, 19 June 2011 to 18 June 2012, constitute the first window for calibrating the individual (point) forecasting models (see Section 5.2). The vertical dotted lines mark the beginning of the first 182-day long window for estimating the probabilistic EPF models (see Section 5.3.1) and the beginning of the 365-day long out-of-sample forecast evaluation period (see Sections 5.3.2–5.3.3).

Surprisingly enough, the example given by Pinson and Tastu [53] shows – contrary to what Khosravi and Nahavandi [102] write – that CWC_{mod} is not a proper scoring rule as well. As such, both the original and the modified CWC measures should be avoided in evaluation of probabilistic forecasts.

5. Empirical study

5.1. The data

The dataset used in this empirical study comes from the price track of the Global Energy Forecasting Competition 2014 (GEFCom2014), the largest energy forecasting competition to date [3]. It comprises three time series at an hourly resolution – locational marginal prices (LMP, i.e. zonal prices) and day-ahead predictions of zonal and system loads, see Fig. 4. The dataset is now available as supplementary material accompanying Ref. [3], however, during the competition the information set was being extended on a weekly basis to prevent 'peeking' into the future. The origin of the data has never been revealed by the organizers.

To illustrate the probabilistic EPF and evaluation techniques discussed in this paper, we consider a 2.5-year period from 19 June 2011 to 17 December 2013. The first 365 days, 19 June 2011 to 18 June 2012, constitute the first window for calibrating the individual autoregressive models (**ARX** and **mARX** defined in Sections 5.2.3–5.2.4). The neural network model (**NN** defined in Section 5.2.5) uses a much shorter calibration window of only 312 hourly observations (i.e. 13 past days). When the day-ahead forecasts are made for the 24 h of 19 June 2012, the 365- and 13-day windows are rolled forward by one day. This procedure is repeated until the predictions of the individual models for the last day in the sample – 17 December 2013 – are made.

The second period, initially from 19 June to 17 December 2012 (i.e. 182 days), is utilized for computing the probabilistic forecasts of the naïve, autoregressive and QRA models for 18 December 2012; like in [39], the neural network models use only a 13-day window, initially 5–17 December 2012. Then the 182- and 13-days windows are rolled forward by one day, the models are recalibrated and the probabilistic forecasts are computed for 19 December 2012. This procedure is repeated until probabilistic predictions for all 365 days in the out-of-sample test period (18 December 2012 – 17 December 2013) are obtained.

5.2. Individual (point) forecasting models

On one hand, our choice of the individual (point) forecasting models is guided by the existing literature on short-term EPF and the results of the GEFCom2014 competition, on the other, by the desire to illustrate the probabilistic EPF and evaluation techniques without the need to resort to extremely sophisticated and fine-tuned models. Overall, we consider a simple naïve benchmark and three parsimonious expert (see footnote 8) models: one commonly used in EPF since the study of Misiorek et al. [12] and two that formed the backbone of probabilistic EPF approaches ranked 2nd [33] and 3rd [39] in the price track of GEFCom2014.

In the first two autoregressive structures (**ARX** and **mARX**) the modeling is implemented separately across the hours, leading to 24 sets of parameters for each day, an approach commonly taken in EPF studies [29,35,42,43,72,117,118]. The third expert model is a neural network (**NN**) which takes into account all observations in the calibration window and leads to one set of parameters for all hours; the forecast for hour 1 is then used as model input to compute the forecast for hour 2, etc. As Ziel [43] interestingly notes, when we compare the forecasting performance of relatively simple models implemented separately across the hours and jointly for all hours, the latter generally perform better for the first half of the day, whereas the former are better in the second half of the day. This is probably the reason why the neural network models perform relatively well for the late night and early morning hours and rather poorly for the remainder of the day.

5.2.1. Data preprocessing

Like many studies in the EPF literature [1,7,98], we use transformations to make the data more symmetric and stabilize the variance. Following Uniejewski et al. [42], the autoregressive models (**ARX** and **mARX**) work on centered log-prices, $p_{d,h} = \log(P_{d,h}) - \frac{1}{365}\sum_{t=1}^{365}\log(P_{t,h})$, with the centering performed independently for each hour $h = 1, \ldots, 24$. Note that from this point onwards we use the more

natural for day-ahead markets notation and denote by $P_{d,h}$ the electricity price for day d and hour h. Clearly, the previously used single time index can be obtained through the relation $t = 24d + h$.

We can apply the logarithmic transformation since the GEFCom2014 price series is positive-valued. If datasets with zero or negative values were considered, we could work with non-transformed prices or apply a different transformation (like the area hyperbolic sine, see [119], or the probability integral transform, see [120] and Section 4.1). In the neural network model (**NN**) we follow Dudek [39] and map the prices ($P_{d,h} \to \widetilde{p}_{d,h}$) and loads ($Z_{d,h}^{zonal} \to \widetilde{z}_{d,h}$, $Z_{d,h}^{system} \to \widetilde{\zeta}_{d,h}$) to the interval $[-0.9,\ 0.9]$ to facilitate and accelerate the learning process, see also Section 5.2.5.

5.2.2. The naïve benchmark

The benchmark, most likely introduced by Nogales et al. [121] and dubbed the *naïve method*, belongs to the class of similar-day techniques (for a taxonomy of EPF approaches see e.g. [1]). It proceeds as follows: the electricity price forecast for hour h on Tuesday, Wednesday, Thursday or Friday is set equal to the price for the same hour on the previous day, i.e., $\widehat{P}_{d,h} = P_{d-1,h}$; the forecast for hour h on Saturday, Sunday or Monday is set equal to the price for the same hour a week ago, i.e., $\widehat{P}_{d,h} = P_{d-7,h}$. We denote this benchmark by **Naïve**.

5.2.3. The ARX model

The first expert model that we consider was originally proposed by Misiorek et al. [12] in one of the first probabilistic EPF studies and later used in multiple papers [13,22,24,29,35,42,43,71,72,117,122,123]. Within this model the centered log-price on day d and hour h is given by the following formula:

$$\begin{aligned} p_{d,h} = {} & \beta_{h,1}p_{d-1,h} + \beta_{h,2}p_{d-2,h} + \beta_{h,3}p_{d-7,h} + \beta_{h,4}p_{d-1}^{min} + \beta_{h,5}z_{d,h} \\ & + \beta_{h,6}D_{Sat} + \beta_{h,7}D_{Sun} + \beta_{h,8}D_{Mon} + \varepsilon_{d,h}, \end{aligned} \tag{23}$$

where the lagged log-prices $p_{d-1,h}$, $p_{d-2,h}$ and $p_{d-7,h}$ account for the autoregressive effects of the previous days (the same hour yesterday, two days ago and one week ago), $p_{d-1}^{min} \equiv \min_{h=1,\ldots,24}\{p_{d-1,h}\}$ is the minimum of the previous day's 24 hourly log-prices, $z_{d,h}$ is the logarithm of the zonal load forecast for day d and hour h, D_{Sat}, D_{Sun} and D_{Mon} are dummy variables that account for the weekly seasonality and the $\varepsilon_{d,h}$'s are assumed to be independent and identically distributed (i.i.d.) normal variables. We denote this autoregressive benchmark by **ARX** to reflect the fact that the (zonal) load forecast is used as the exogenous variable in Eq. (23).

5.2.4. The mARX model

The second expert model is an extension of **ARX**, which evolved from it during the successful participation of TEAM POLAND in the GEFCom2014 competition [33]. The rationale for the modifications stems from the observation that it may be beneficial to use different model structures for different days of the week, not only different parameter sets [117]. The so-called multi-day ARX model or **mARX** is given by the following formula:

$$\begin{aligned} p_{d,h} = {} & \left(\sum_{i\in I} \beta_{h,1,i}D_i\right)p_{d-1,h} + \beta_{h,2}p_{d-2,h} + \beta_{h,3}p_{d-7,h} + \beta_{h,4}p_{d-1}^{min} + \beta_{h,5}z_{d,h} \\ & + \beta_{h,6}D_{Sat} + \beta_{h,7}D_{Sun} + \beta_{h,8}D_{Mon} + \beta_{h,11}D_{Mon}p_{d-3,h} + \varepsilon_{d,h}, \end{aligned} \tag{24}$$

where $I \equiv \{0, Sat, Sun, Mon\}$, $D_0 \equiv 1$ and the term $D_{Mon}p_{d-3,h}$ accounts for the autoregressive effect of Friday's prices on the prices for the same hour on Monday. Note that, to some extent, this structure resembles periodic autoregressive models (i.e. PAR, PARMA), which have seen limited use in EPF [42]. Both autoregressive models (**ARX** and **mARX**) are estimated with Least Squares (LS), using Matlab's `regress.m` function.

5.2.5. The NN model

The third expert model (denoted by **NN**) is a relatively parsimonious neural network that was used by Dudek [39] in the GEFCom2014 competition. Not only does it use a different methodology, more popular among electrical engineers [7,124], but also does not contain any autoregressive terms. The rescaled price (to lie within the interval $[-0.9,\ 0.9]$; see Section 5.2.1) is modeled as:

$$\widetilde{p}_{d,h} = f(\widetilde{z}_{d,h}, \widetilde{\zeta}_{d,h}, \widetilde{z}_{d,h}^2, \widetilde{\zeta}_{d,h}^2) + \varepsilon_{d,h}, \tag{25}$$

where $\widetilde{z}_{d,h}$ and $\widetilde{\zeta}_{d,h}$ are the rescaled zonal and system load forecasts, respectively, and $f(x)$ represents a multilayer perceptron with five sigmoid neurons in the hidden layer and one linear neuron in the output layer. Unlike **ARX** and **mARX**, the **NN** model uses only the past (312 = 24 h×13 days) observations for parameter estimation. Like in [39], the **NN** model is calibrated using Matlab's Neural Network Toolbox: first, the network is set up using the `feedforwardnet.m` function, then trained with the `train.m` function for a maximum of 100 epochs using Bayesian regularization backpropagation (`trainFcn` = '`trainbr`') and the sum squared error performance function (`performFcn` = '`sse`').

5.3. Empirical results

We are now in a position to illustrate some of the techniques discussed in Sections 3 and 4. We put to work four approaches of calculating probabilistic forecasts: historical simulation, Gaussian PIs, bootstrapping and Quantile Regression Averaging (QRA). Overall we consider nine probabilistic forecasting models built on the point forecasts, $\widehat{P}_t$, of the four individual models defined in Sections 5.2.2-5.2.5 above. Naturally, unless stated otherwise, all evaluation measures and statistics presented in Sections 5.3.2–5.3.3 are averages over all days in the 365-day test period, see Fig. 4 and Section 5.1.

5.3.1. Constructing probabilistic forecasts from point predictions

Recall from Section 3.2 that historical simulation is a model-independent approach which consists of computing sample quantiles of the empirical distribution of the residuals: $\varepsilon_t = P_t - \widehat{P}_t$. We apply this technique to the naïve benchmark and both autoregressive models; as a result we obtain three PEPF models: **Naïve-H**, **ARX-H** and **mARX-H**. All three use a 182-day rolling window of residuals for constructing the probabilistic forecasts, independently for each hour of the day.

We illustrate distribution-based PIs using the neural network model. As Dudek [39], we use a 312-h (i.e., 13-day) rolling window to compute the standard deviation of the error density, $\hat{\sigma}$, then use it to retrieve quantiles of the Gaussian distribution approximating the error density, see Section 3.3 for details. The resulting probabilistic EPF model is denoted by **NN-G**.

The bootstrap is a more complex and computationally intensive approach, which is based on generating pseudo-data using bootstrapped (i.e., resampled) residuals, see Section 3.4. We apply this technique to the point forecasts of all three expert models and obtain three probabilistic models: **ARX-B**, **mARX-B** and **NN-B**. The former two use a 182-day rolling window of residuals for constructing the probabilistic forecasts (independently for each hour of the day, like **ARX-H** and **mARX-H**), the latter uses a 312-h (i.e., 13-day) rolling window of residuals (like **NN-G**).

Finally, we apply QRA either to point forecasts of the two expert autoregressive models, resulting in model **QRA(2)**, or to all three expert models, resulting in model **QRA(3)**. Recall from Section 3.5 that a rolling window of the previous 182 days is used to calibrate QRA and obtain the PIs for the next day. Note that the **Naïve** benchmark is not used for computing bootstrapped PIs or in QRA.

5.3.2. Evaluating reliability

The unconditional coverage (UC) and the related average coverage

Table 2
Unconditional coverage (UC) and average coverage error (ACE, i.e., UC minus nominal coverage; see also Section 4.1.1) of the 50% and 90% two-sided day-ahead PIs by the actual spot price for all nine models. The best results in each row are emphasized in bold, the worst are underlined. Note that the results are averages over all 24 hourly load periods.

	Naïve-H	ARX-H	ARX-B	mARX-H	mARX-B	NN-G	NN-B	QRA (3)	QRA (2)
50% prediction intervals									
UC	46.94%	48.32%	49.38%	47.55%	**50.08%**	45.50%	40.68%	47.23%	47.66%
ACE	−3.06%	−1.68%	−0.62%	−2.45%	**0.08%**	−4.50%	− 9.32%	−2.77%	−2.34%
90% prediction intervals									
UC	85.84%	86.59%	87.26%	85.96%	**87.44%**	80.29%	79.86%	85.14%	85.51%
ACE	−4.16%	−3.41%	−2.74%	−4.04%	**−2.56%**	−9.71%	− 10.14%	−4.86%	−4.49%

error (ACE; see Section 4.1.1) of the 50% and 90% PIs for the nine models are shown in Table 2. The results are averages over all 24 hourly load periods. Nearly all models yield a smaller coverage than nominal, i.e., all but one ACE errors are negative. We can observe the worst performance for the two neural network based forecasts. On the other hand, models with the best unconditional coverage (i.e., the closest to nominal) are **ARX-B** and **mARX-B**, with the remaining two autoregressive structures, both QRA models and the naïve benchmark trailing closely behind. Surprisingly, the naïve benchmark yields a much better unconditional coverage than the neural networks.

Such aggregate measures as the UC and ACE in Table 2, often reported in the probabilistic EPF literature, do not disclose the relevant details. Namely, the coverage is not uniform across the hours. In particular, for the 50% PIs, the neural network models tend to provide a little too wide (or misplaced; see the discussion on sharpness in Section 5.3.3) PIs for late night/early morning hours, with as high as 62% and 54% coverage for hour 3 for **NN-G** and **NN-B**, respectively. On the other hand, for the afternoon and evening hours they largely underestimate the variability and yield much too narrow (or significantly misplaced) PIs, with as low as 27–28% coverage for hours 19–20. The remaining models provide a more stable performance across the hours.

To formally assess coverage we run the Kupiec [77] and Christoffersen [46] tests for the 50% and 90% PIs. We conduct the tests separately for each of the 24 h since the probabilistic forecasts for consecutive hours are correlated by construction. The naïve, autoregressive and QRA models are estimated 24 times, separately for each hour of the day, using the same information set (prices and load forecasts up to midnight on the day the prediction is made). The neural networks model all hours jointly, but use the same parameter estimates to compute the probabilistic forecasts for all 24 h of the next day (i.e., using the same information set); it is the load forecasts that make the difference and diversify the price forecasts for each hour.

The results are presented in Fig. 5. For the 50% PIs, the four autoregressive models and **QRA(2)** yield the best and nearly identical unconditional coverage, see the red circles in the top nine panels. At the 5% level of significance between 22 and 24 h pass the test, at the 1% level all hours pass the test. The **QRA(3)** model and, as already visible for the aggregate UC results in Table 2, the naïve model follow closely by. The neural network models are definitely the worst, with acceptable coverage only for 6–10 h; the problems generally arise for the daytime hours (8 am-11 pm). However, the situation changes when we look at the 90% PIs, see the blue triangles in the top nine panels of Fig. 5. Models **ARX-B** and **mARX-B** are now clearly better than the competitors, with as many as 16 h passing the Kupiec test at the 5% level and 20–23 h at the 1% level. Next in line is the **ARX-H** model, then **mARX-H** and **QRA(2)**. For the 90% PIs, the **QRA(3)** model is the worst, with only three hours passing the Kupiec test at the 5% significance level and five hours at the 1% level.

Now let us look at the results of the Christoffersen test for CC. Recall from Section 4.1.1 that $LR_{CC} = LR_{UC} + LR_{Ind}$, if we condition on the first observation. Thus the differences between the UC and CC tests can be explained in terms of dependence or clustering of PI violations. Note that in Fig. 5 we are only looking at first-order dependence, however, at almost no cost we can check dependence for any lag, see Clements and Taylor [81] for a general discussion and Maciejowska et al. [29] for an application in EPF. The results for the CC test are definitely less optimistic than those for the UC test. For the 50% PIs, the bootstrapped **mARX-B** model performs slightly better than QRA(2) with 4 vs. 2 h passing the test at 5% significance and 9 vs. 7 at 1% significance. The remaining models perform much worse, with at most two hours passing the test at 1% significance.

Finally, let us investigate the reliability of the whole predictive density, not just two selected PIs. Since the methods we consider in this empirical study do not yield predictive densities, only quantile forecasts of any level, we use a set of 99 percentiles, i.e. $q = 1\%, 2\%, \ldots, 99\%$, to approximate the PDFs. In Fig. 6 we plot the histograms of the probability integral transform (PIT) histograms and sample autocorrelation functions (ACFs) of the PIT values for two selected hours: 4 (late night trough) and 19 (evening peak). We present results for four selected models representing four methods of constructing probabilistic forecasts: **Naïve-H**, **ARX-B**, **NN-G** and **QRA(2)**. Recall from Section 4.1.4, that a histogram with too much probability mass in the center (inverse U-shape; like for **NN-G** and hour 4) indicates that the predictive distribution has too fat tails, while a U-shape (like for **NN-G** and hour 19) suggests that the tails of the predictive distribution are not heavy enough. Some histograms (**Naïve-H** for hour 19, **ARX-B** and **QRA(2)** for hour 4) exhibit skewness, with more mass below the mean than above, and a little too much probability mass at the right end, suggesting that the right tail of the predictive distribution should be heavier. Only **QRA(2)** for hour 19 and to some extent **ARX-B** for the same hour yield what seems to be a uniform distribution. The ACF plots generally confirm what can be seen in the PIT histograms: **QRA(2)** and **ARX-B** perform better than **Naïve-H**, which in turn is better than **NN-G**.

To formally evaluate the predictive distributions we now conduct the Berkowitz [48] test of independence and normality (i.e., of conditional coverage). The results for all nine probabilistic models and all 24 hourly load periods are presented in Fig. 7. Clearly, none of the models is perfect. All nine have problems with passing the test for the afternoon hours, particularly hours 13 through 17. Yet, if ranked, then both bootstrapped autoregressive models lead the pack with 14–15 h passing the test at the 5% level and 17–18 h at the 1% level, and **QRA(2)** follows closely by, respectively with 12 and 15 h. At the other end are both neural networks and the naïve model, which do not provide an acceptable conditional coverage for a single hour.

Comparing the results of the Christoffersen [46] and Berkowitz [48] tests we note that while both generally rank the models in the same order, the latter provides a more holistic picture and points to problems that may not be visible when we test the CC of arbitrarily chosen PIs. As such, the Berkowitz test is the preferred option for an 'all-in-one' evaluation procedure. On the other hand, running the Christoffersen test for all PIs (from 1% to 99% nominal coverage) may indicate which quantiles are problematic and require attention or a revision of the model. For instance, the CC of the 50% PI for the **mARX-B** model is acceptable for hours 13 through 17, even at the more restrictive 5%

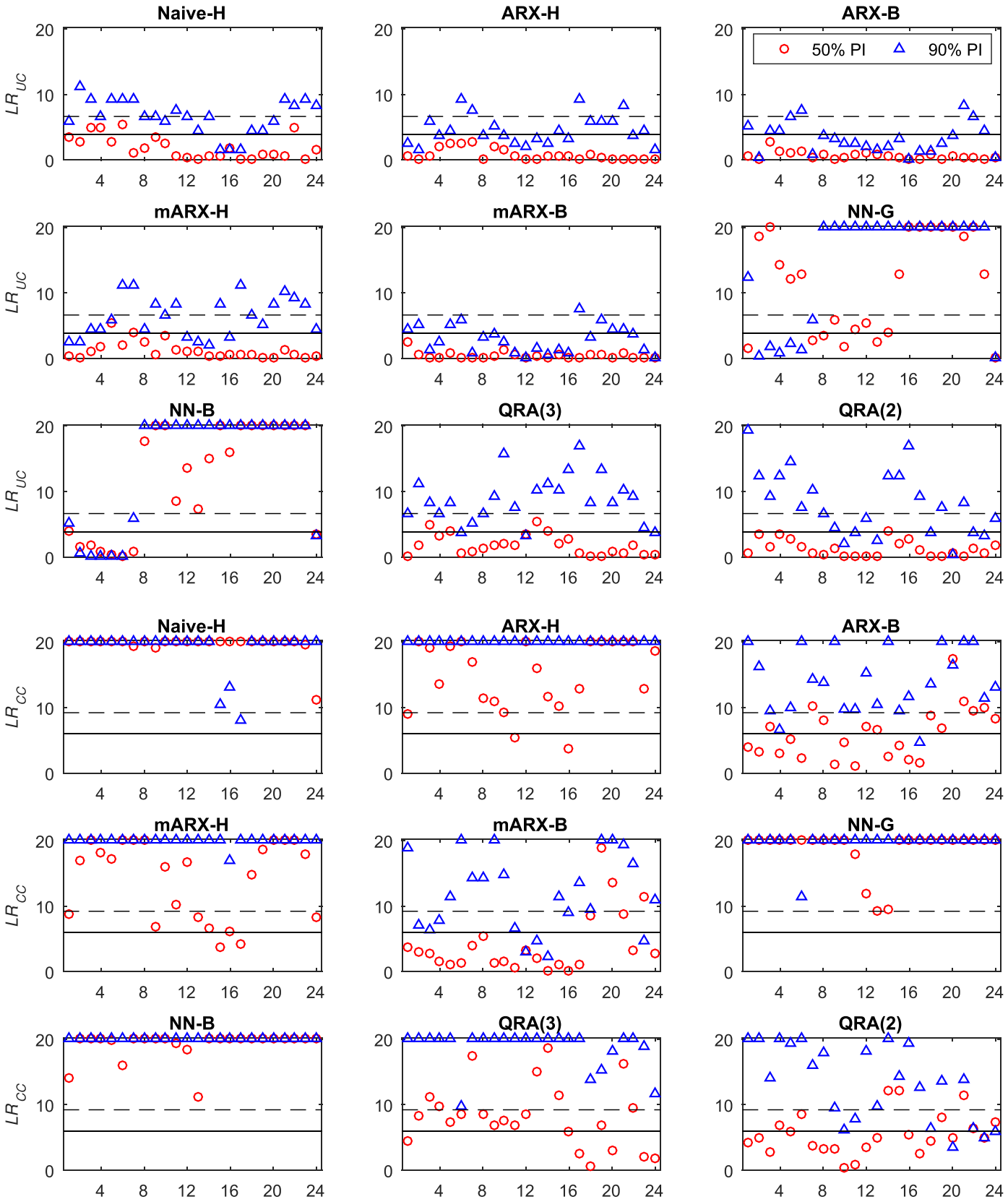

Fig. 5. The unconditional (LR_{UC}; *top three rows*) and conditional coverage (LR_{CC}; *bottom three rows*) likelihood ratio statistics for the 50% (○) and 90% (△) PIs obtained from the nine probabilistic forecasting models considered. Recall from Section 4.1.1 that $LR_{CC} = LR_{UC} + LR_{Ind}$, if we condition on the first observation. The solid (dashed) horizontal lines represent the 5% (1%) significance level of the appropriate χ^2 distribution. All test values exceeding 20 are set to 20.

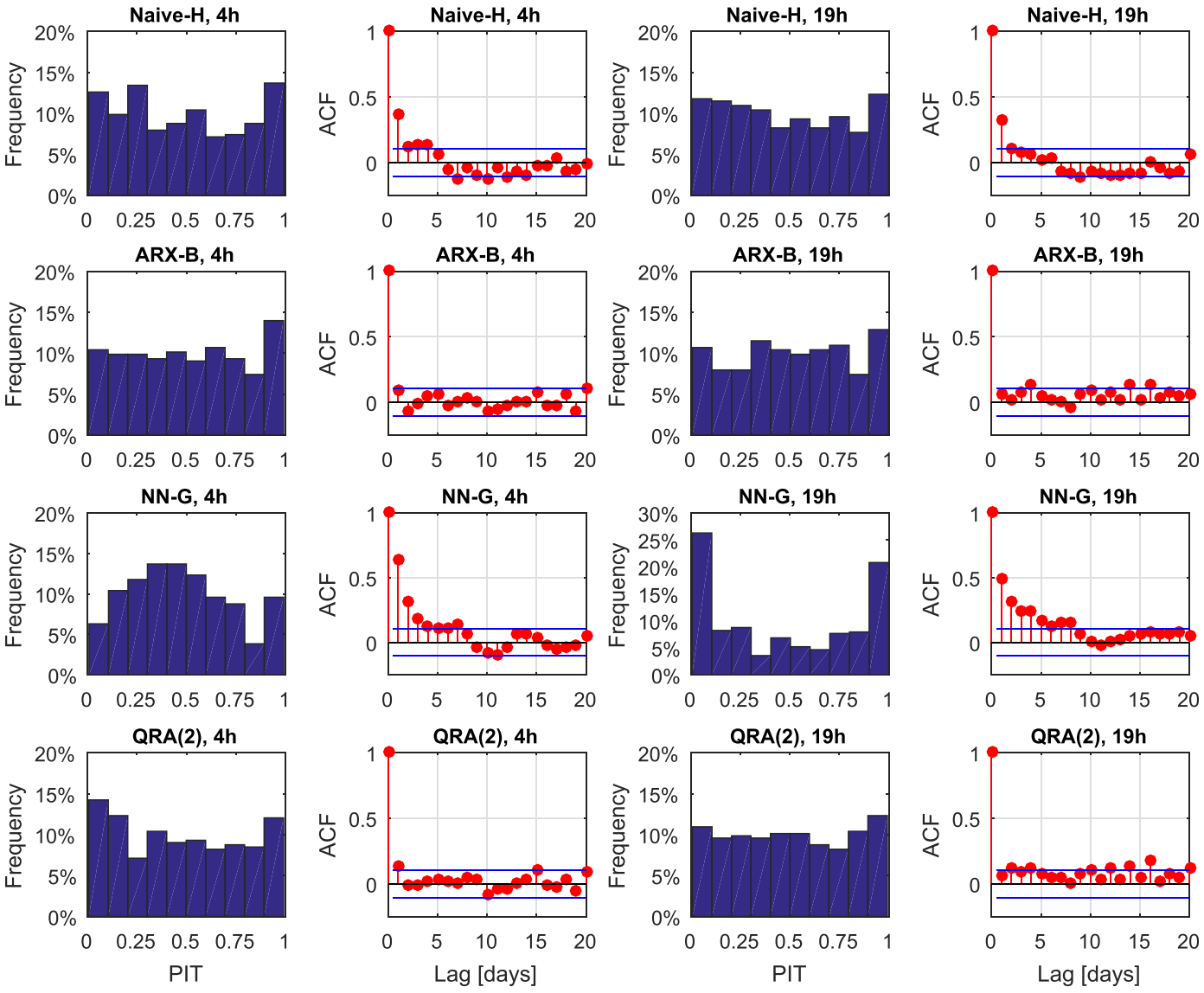

Fig. 6. Probability integral transform (PIT) histograms and sample autocorrelation functions (ACFs) of the PIT values for hours 4 (*left*) and 19 (*right*) and four probabilistic models (*top to bottom*): **Naïve-H**, **ARX-B**, **NN-G** and **QRA(2)**. Note that for the **NN-G** model and hour 19 the scale on the Y-axis is compressed due to very high bins at both ends of the PIT histogram.

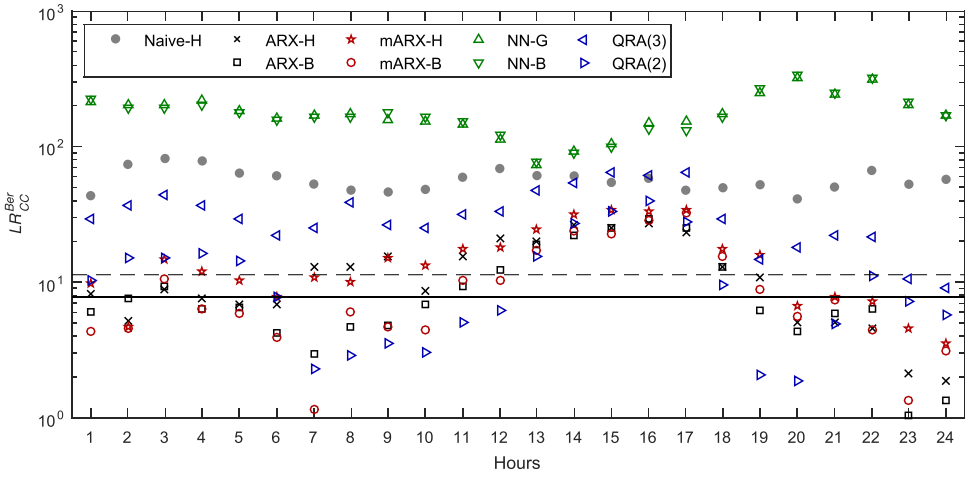

Fig. 7. The conditional coverage likelihood ratio statistics (LR_{CC}^{Ber}) for the Berkowitz [48] test of independence and normality of the predictive distributions obtained from all nine probabilistic forecasting models. The solid (dashed) horizontal lines represent the 5% (1%) significance level of the $\chi^2(3)$ distribution. Note the logarithmic scale on the Y-axis.

Table 3
Top: The pinball loss, as defined by Eqn. 13, averaged over all 24 hourly load periods and across 99 percentiles; for an analysis of selected quantiles, see Fig. 8. *Middle and bottom*: The Winkler (or 'interval') score, as defined by Eqn. 14, for the 50% and 90% two-sided day-ahead PIs averaged over all 24 hourly load periods; for an analysis of selected load periods see Fig. 9. The Winkler score is decomposed into PI width, i.e., δ_t, and a penalty for PI violations, i.e., $\frac{2}{\alpha}(\widehat{L}_t - P_t)$ or $\frac{2}{\alpha}(P_t - \widehat{U}_t)$. The best results in each row are emphasized in bold, the worst are underlined.

	Naïve-H	ARX-H	ARX-B	mARX-H	mARX-B	NN-G	NN-B	QRA (3)	QRA (2)
Average score over 99 percentiles									
Pinball loss	3.927	2.943	2.774	2.971	2.788	3.364	3.305	**2.634**	2.791
50% prediction intervals									
Winkler score	34.141	25.505	24.434	25.741	24.556	29.385	29.208	**23.108**	24.726
PI width	7.636	6.592	8.159	**6.556**	8.519	10.989	9.638	9.751	10.310
Penalty	26.505	18.914	16.275	19.185	16.037	18.396	19.570	**13.357**	14.416
90% prediction intervals									
Winkler score	98.599	74.642	55.575	75.317	55.549	68.975	64.008	**50.657**	51.017
PI width	39.662	30.723	**25.760**	30.562	26.287	26.799	27.722	26.076	28.570
Penalty	58.937	43.918	29.815	44.755	29.262	42.177	36.286	24.581	**22.447**

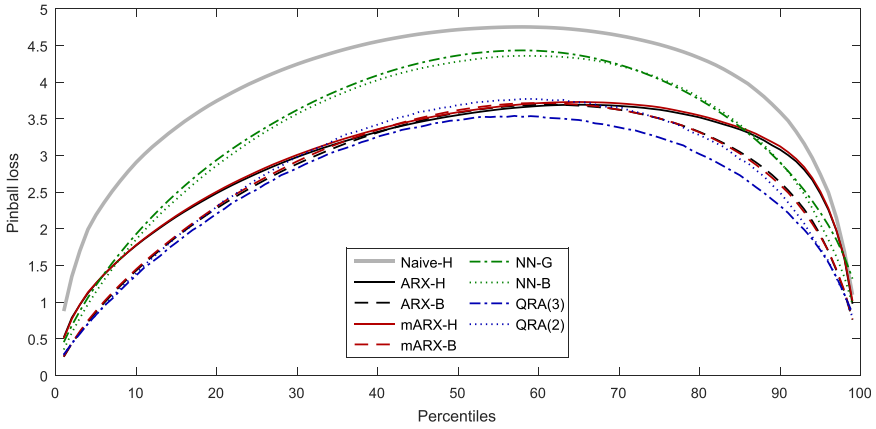

Fig. 8. The pinball loss for 99 percentiles, as defined by Eq. (13), averaged over all 24 hourly load periods. Note that the central percentiles contribute more to the aggregate pinball loss presented in Table 3 than the very low and very high percentiles.

level (see the middle panel in the second row from the bottom in Fig. 5), so it is likely that PIs with higher nominal coverage are responsible for a poor fit of the whole predictive density. Moreover, the Christoffersen test can be easily modified to evaluate more than first order dependence, see Section 4.1.3 and Refs. [29,81], while the Berkowitz test cannot.

5.3.3. Evaluating sharpness

Following the paradigm of 'maximizing sharpness subject to reliability' [15–17] and given that some of the considered models yield reliable forecasts,[12] we are now in a position to evaluate their sharpness. In Section 4 we have presented in detail three measures: the pinball loss, the Winkler score and the continuous ranked probability score (CRPS). Note, however, that for discretized predictive densities – like the ones considered in this empirical study – the CRPS boils down to the pinball loss, see formula (16). Hence, in what follows, we do not discuss the CRPS itself.

Let us first look at the results for the pinball loss, the measure used in the GEFCom2014 competition. The aggregate pinball loss presented in Table 3, i.e., Eq. (13) averaged over all 24 hourly load periods and across 99 percentiles, indicates that the predictions of the **QRA(3)** model are the sharpest, with two bootstrapped autoregressive models and **QRA(2)** following closely by. Again the neural networks and the naïve model perform the worst. There are, however, noticeable changes compared to the reliability rankings in Section 5.3.2. Namely, in terms of sharpness measured by the pinball loss, **QRA(3)** outperforms the autoregressive models, while its predictions have been found to be generally less reliable than those of the autoregressive models, especially **ARX-B** and **mARX-B**. On the other hand, while the forecasts of the **Naïve** benchmark can be regarded as more reliable than those of the neural networks, the latter provide sharper predictions (though still much worse than those of the QRA or autoregressive models).

If we decompose the pinball loss and present the contribution of each quantile to the aggregate measure, as in Fig. 8, we can draw two important conclusions. Firstly, the central percentiles contribute more to the aggregate pinball loss than the very low and very high percentiles. In other words, for sharp probabilistic forecasts it is absolutely crucial to get the point forecasts right, as they are an estimate of the mean or median of the predictive distribution. Errors made in the tails of the distribution play a lesser role, though should not be ignored completely. Secondly, the contribution is not symmetric across the percentiles. Pinball loss plots in Fig. 8 are more symmetric for some models (e.g., **QRA(3)** and **Naïve**) than for others (particularly **ARX-H** and **mARX-H**), but all penalize more for the upper quantiles. Most likely, this is due to the spikiness of electricity spot prices and the inability of the predictive distributions obtained from simple expert models to adequately describe it.

Now, let us analyze the Winkler scores for all nine probabilistic models. Recall from Section 4.2.3, that the Winkler score can be decomposed into PI width, i.e., δ_t, and a penalty for PI violations, i.e., $\frac{2}{\alpha}(\widehat{L}_t - P_t)$ or $\frac{2}{\alpha}(P_t - \widehat{U}_t)$. The latter component is similar to the pinball loss, but the former provides additional information on the sharpness

[12] At least for some of the hourly load periods. Note that we have built the empirical study around a simple naïve benchmark and three parsimonious expert models, as the focus of the paper is not on developing a very well performing model, rather on illustrating the construction and evaluation techniques for probabilistic forecasts.

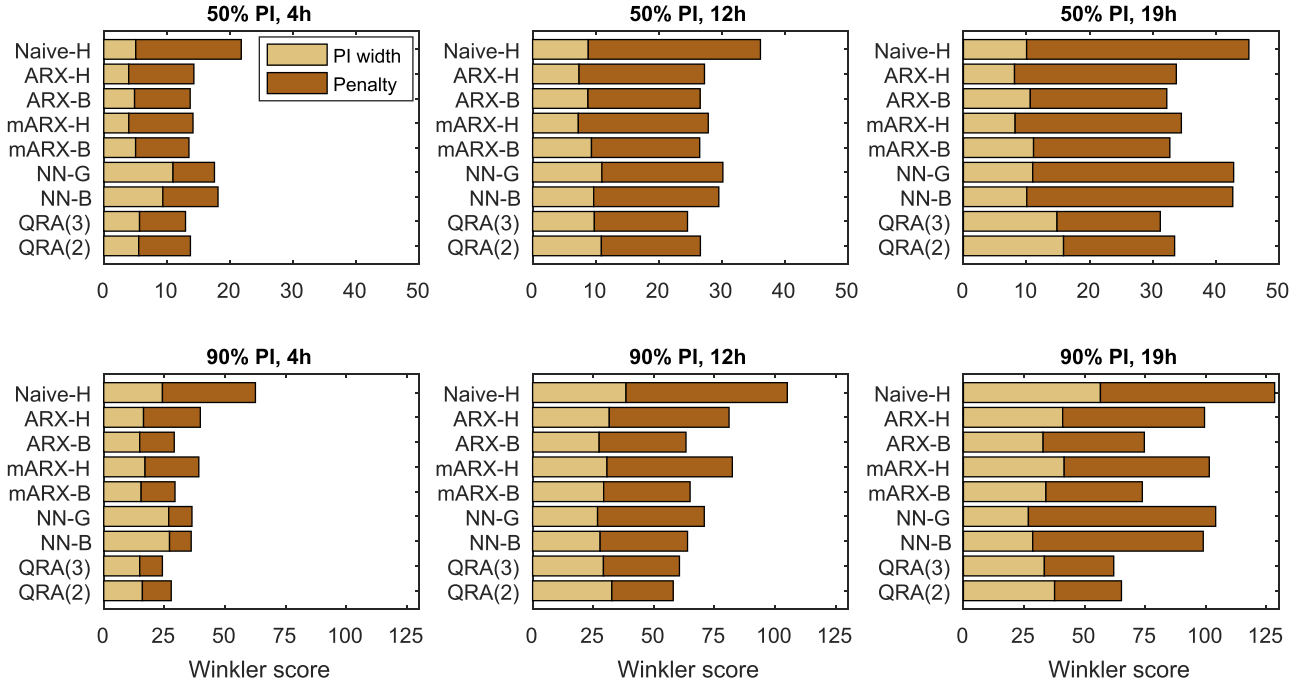

Fig. 9. The Winkler (or 'interval') score, as defined by Eq. (14), for the 50% and 90% two-sided day-ahead PIs and three selected load periods: hours 4, 12 and 19. Like in Table 3, the Winkler score is decomposed into PI width, i.e., δ_t, and a penalty for PI violations, i.e., $\frac{2}{\alpha}(\widehat{L}_t - P_t)$ or $\frac{2}{\alpha}(P_t - \widehat{U}_t)$.

of the predictive distributions. In Table 3 we present the Winkler scores for the 50% and 90% two-sided day-ahead PIs averaged over all 24 hourly load periods. Generally, both the pinball loss and the Winkler score lead to similar conclusions. According to both measures, the **QRA(3)** model yields the sharpest probabilistic forecasts, while the **Naïve** benchmark the least sharp. The **QRA(2)** model is sharper than the two bootstrapped autoregressive models for the 90% PIs, but a little less sharp for the 50% PIs. Interestingly, the sharpness of the probabilistic forecasts of **QRA(3)** and **QRA(2)** is not due to narrow PIs. Quite the opposite, the QRA models yield relatively wide intervals, especially for the nominal coverage of 50%. They excel, however, in terms of the penalty for PI violations. This is somewhat unexpected since the penalty is partly related to reliability, which is not a strong point of **QRA(3)**.

It is also interesting to see the Winkler scores for individual hours, as plotted in Fig. 9. We have selected three load periods for the comparison: hours 4 (late night trough), 12 (midday) and 19 (evening peak). While the general picture is similar to the one from Table 3, in Fig. 9 we can see that for some models the PI widths vary a lot across the hours (particularly for QRA models; i.e., they have a higher resolution across the hourly load periods, see the discussion in the first paragraphs of Section 4), while for other they do not (e.g., for neural networks). The latter may be explained to some extent by the differences in calibration windows – the neural networks use the same 312-h (i.e., 13-day) rolling window to compute the probabilistic forecasts for all hours (the differences are a result of different load forecasts for the individual hours), while the remaining models use a 182-day rolling window of residuals (independently for each hour of the day).

After a descriptive analysis of sharpness, let us now formally evaluate the differences in predictive performance with the Diebold-Mariano (DM) test [90,91]. Since predictions for all 24 h of the next day are made at the same time using the same information set, forecast errors for a particular day will typically exhibit high serial correlation. Therefore, like in [42,71,112,117], we conduct the DM tests for each of the 24 load periods separately.

In Fig. 10 we summarize the DM results for all three score functions: average Pinball loss over all 99 percentiles, Winkler score for the 50% PIs and Winkler score for the 90% PIs. Like Uniejewski et al. [42], we sum the number of significant differences in forecasting performance across the 24 h and use a heatmap to indicate the number of hours for which the forecasts of a model on the X-axis are

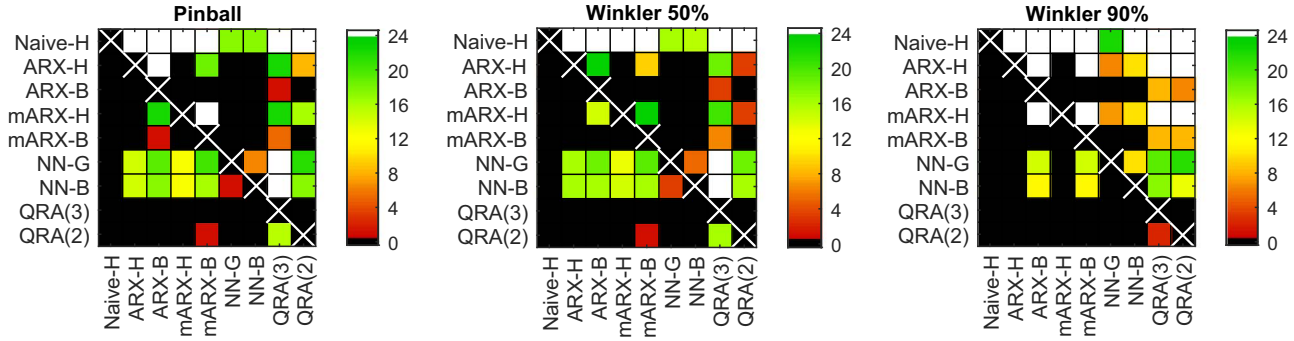

Fig. 10. Results for conducted one-sided Diebold-Mariano [90] tests at the 5% level for all nine probabilistic forecasting models and three score functions (*from left to right*): average Pinball score over all 99 percentiles, Winkler score for the 50% PIs and Winkler score for the 90% PIs. We sum the number of significant differences in forecasting performance across the 24 h and use a heat map to indicate the number of hours for which the forecasts of a model on the X-axis are significantly better than those of a model on the Y-axis. A white square indicates that forecasts of a model on the X-axis are better for all 24 h, while a black square that they are not better for a single hour.

significantly better than those of a model on the Y-axis. Two extreme cases – (i) the forecasts of a model on the X-axis are significantly better for all 24 h of the day and (ii) the forecasts of a model on the X-axis are not significantly better for any hour – are indicated by white and black squares, respectively. Naturally, the diagonal (white crosses on black squares) should be ignored as it concerns the same model on both axes. Columns with many non-black squares (the more green or white the better; this is the case for the **QRA(3)** model) indicate that the forecasts of a model on the X-axis are significantly better than the forecasts of many of its competitors. Conversely, rows with many non-black squares mean that the forecasts of a model on the Y-axis are significantly worse than the forecasts of many of its competitors (this is the case for the naïve benchmark and the neural networks to a lesser extent).

Generally, the results of the DM tests support earlier formulated conclusions. Firstly, the predictions of the **Naïve** benchmark are significantly outperformed by those of all other models, for (nearly) all hours. The forecasts of the neural networks are outperformed by the predictions of QRA and autoregressive models for a vast majority of hours. This is especially true for the pinball loss and the Winkler score for the 50% PIs. For the 90% PIs the neural networks are relatively better, yet still perform worse than other methods. On the other hand, the best method overall is **QRA(3)**. The forecasts of this model are never significantly worse than the predictions of any other model and for some hours they outperform the predictions of the other models, regardless of the evaluation metric used. **QRA(2)** and bootstrapped autoregressive models perform slightly worse, yet still yield significantly better predictions than the remaining models.

5.4. Final thoughts and recommendations

This extensive empirical study reflects the complexity of the construction and evaluation of probabilistic forecasts, which goes well beyond that of point predictions. In line with the 'maximizing sharpness subject to reliability' paradigm [15–17] we advocate, the first crucial step is a thorough evaluation and formal testing of reliability, i.e., the statistical consistency between the distributional forecasts and the observations. A number of techniques have been put to work in Section 5.3.2, several more are available in the literature. Although we strongly suggest to check both the unconditional and conditional coverage, we refrain from recommending one particular procedure. The Kupiec [77] and Christoffersen [46] tests can be replaced by the Berkowitz [48] test if the constructed PIs span the whole range of quantiles and constitute a dense grid well approximating the predictive density. However, if more than first-order dependence needs to be checked – and this may be the case for electricity price forecasts as argued by Maciejowska et al. [29] – the Christoffersen test lends itself readily to an extension which checks independence of PI violations at an arbitrary time lag [81] or all lags at once [82]. The autocorrelation plots of the PITs may be also useful in this context.

Once we are done with reliability, and can conclude that the obtained probabilistic forecasts are reliable enough, we should look at sharpness, i.e. the concentration of the predictive distributions. The latter notion is closely related to the so-called proper scoring rules, which simultaneously assess reliability and sharpness, and are minimized when the true distribution is quoted as the forecast distribution. We advocate the use of the pinball loss, the Winkler (or 'interval') score and/or the continuous ranked probability score (CRPS) to rank the forecasts. Yet, like for evaluating reliability, several more options are available in the literature. Although similar in spirit, the three scoring rules offer a slightly different view on forecast sharpness. The pinball loss lends itself readily to averaging over all considered quantiles, while the Winkler score additionally penalizes for too wide PIs. Moreover, for discretized predictive densities – like the ones considered in this empirical study – the CRPS boils down to the pinball loss, but for continuous predictive densities it is the preferred option. After ranking forecasts using one or more of the mentioned scoring rules, we suggest to formally test for statistically significant differences in the forecasting performance, e.g., using the Diebold-Mariano [90,91] test. As Weron [1] remarks, this issue has apparently been downplayed in the EPF literature, although it is a standard procedure in econometrics. Sadly, three years later not much has improved.

Before we conclude this Section let us briefly comment on the strikingly poor performance of the neural network models. According to most evaluation measures and tests they yield either the worst predictions or only better than those of the naïve benchmark. How is this possible given that one of them (**NN-G**) was ranked third in the GEFCom2014 competition? The surprising answer is that they perform very well for the 12 competition days (see Table 7 in [3]), but not in general. Namely, the aggregate pinball loss over those 12 days for the **NN-G** and **NN-B** models is 3.302 and 3.200, respectively, while for the autoregressive expert models and the naïve benchmark it is in the range 4.454–5.100. The QRA models fare better, with aggregate pinball loss of 4.209 for **QRA(2)** and only 3.089 for **QRA(3)**. Apparently, the latter model benefits from including point forecasts of the neutral network. On the other hand, this inclusion leads to a generally worse reliability over the whole 365-day out-of-sample test period, when compared to **QRA(2)**, see Section 5.3.2. This qualitatively different performance of some models on an arbitrarily, though ex-ante, selected 12-day sample (GEFCom2014) and on a large, 365-day sample (this study) emphasizes the need for long out-of-sample test periods, an issue raised by Weron [1] in his review of EPF and more recently by Hong and Fan [5] in the context of probabilistic load forecasting.

6. Conclusions

We have presented guidelines for the rigorous use of methods, measures and tests in probabilistic electricity price forecasting. However, the article has a much broader reach. None of the methods for constructing probabilistic forecasts discussed in Section 3 or the evaluation metrics reviewed in Section 4 is restricted to electricity prices. They all are general enough to be used for probabilistic energy forecasting, is it very short-term load forecasting for smart grid applications or wind and solar power forecasting. With the increasing role of probabilistic predictions in general, we truly hope that this review paper will encourage energy forecasters to develop more efficient, but at the same time statistically sound approaches. We also hope that it will propel those working in other areas of forecasting to move into the exciting and still largely unexplored world of wholesale electricity markets.

Acknowledgments

This paper has benefited from the discussions with Grzegorz Dudek, Tao Hong, Bartosz Uniejewski and Florian Ziel, as well as conversations with the participants of the IEEE PES General Meeting (PESGM2015 Denver), the Energy Finance Christmas Workshops (EFC15 Paris, EFC16 Essen), the International Summer School on Risk Measurement and Control (ISS2016 Rome), the International Symposium on Forecasting (ISF2016 Santander) and the Conference on Energy Finance (EF2016 Paris). The study was partially supported by the National Science Center (NCN, Poland) through grants 2013/11/N/HS4/03649 (to JN) and 2015/17/B/HS4/00334 (to RW).

References

[1] Weron R. Electricity price forecasting: a review of the state-of-the-art with a look into the future. Int J Forecast 2014;30(4):1030–81.
[2] De Gooijer JG, Hyndman RJ. 25 years of time series forecasting. Int J Forecast 2006;22(3):443–73.
[3] Hong T, Pinson P, Fan S, Zareipour H, Troccoli A, Hyndman RJ. Probabilistic energy forecasting: Global Energy Forecasting Competition 2014 and beyond. Int J Forecast 2016;32(3):896–913.

[4] Moghimi Ghadikolaei H, Ahmadi A, Aghaei J, Najafi M. Risk constrained self-scheduling of hydro/wind units for short term electricity markets considering intermittency and uncertainty. Renew Sustain Energy Rev 2012;16(7):4734–43.
[5] Hong T, Fan S. Probabilistic electric load forecasting: a tutorial review. Int J Forecast 2016;32:914–38.
[6] Liu B, Nowotarski J, Hong T, Weron R. Probabilistic load forecasting via Quantile Regression Averaging on sister forecasts. IEEE Trans Smart Grid 2017;8(2):730–7.
[7] Weron R, Ziel F. Forecasting electricity prices: a guide to robust modeling. CRC Press; 2018, [forthcoming].
[8] Chatfield C. Time-series forecasting. Boca Raton, Florida: Chapman & Hall/CRC; 2000.
[9] Amjady N, Hemmati M. Energy price forecasting. IEEE Power Energy Mag 2006(March/April):20–9.
[10] Raza M, Khosravi A. A review on artificial intelligence based load demand forecasting techniques for smart grid and buildings. Renew Sustain Energy Rev 2015;50:1352–72.
[11] Zhang L, Luh P. Neural network-based market clearing price prediction and confidence interval estimation with an improved extended Kalman filter method. IEEE Trans Power Syst 2005;20(1):59–66.
[12] Misiorek A, Trück S, Weron R. Point and interval forecasting of spot electricity prices: linear vs. non-linear time series models. Stud Nonlinear Dyn Econ 2006;10(3), (Article 2).
[13] Serinaldi F. Distributional modeling and short-term forecasting of electricity prices by Generalized Additive Models for location, scale and shape. Energy Econ 2011;33:1216–26.
[14] Hyndman R. The difference between prediction intervals and confidence intervals. Hyndsight Blog (13 March 2013), ⟨http://robjhyndman.com/hyndsight/intervals⟩; 2011.
[15] Gneiting T, Balabdaoui F, Raftery A. Probabilistic forecasts, calibration and sharpness. J R Stat Soc B 2007;69:243–68.
[16] Gneiting T, Raftery A. Strictly proper scoring rules, prediction, and estimation. J Am Stat Assoc 2007;102(477):359–78.
[17] Gneiting T, Katzfuss M. Probabilistic forecasting. Annu Rev Stat Appl 2014;1:125–51.
[18] Pinson P, Nielsen H, Moller J, Madsen H, Kariniotakis G. Non-parametric probabilistic forecasts of wind power: required properties and evaluation. Wind Energy 2007;10(6):497–516.
[19] Zhang Y, Wang J, Wang X. Review on probabilistic forecasting of wind power generation. Renew Sustain Energy Rev 2003;32:255–70.
[20] Jung J, Broadwater RP. Current status and future advances for wind speed and power forecasting. Renew Sustain Energy Rev 2014;31:762–77.
[21] Yan J, Liu Y, Han S, Wang Y, Feng S. Reviews on uncertainty analysis of wind power forecasting. Renew Sustain Energy Rev 2015;52:1322–30.
[22] Weron R. Modeling and forecasting electricity loads and prices: a statistical approach. Chichester: John Wiley & Sons; 2006.
[23] Zhang L, Luh P, Kasiviswanathan K. Energy clearing price prediction and confidence interval estimation with cascaded neural networks. IEEE Trans Power Syst 2003;18(1):99–105.
[24] Weron R, Misiorek A. Forecasting spot electricity prices: a comparison of parametric and semiparametric time series models. Int J Forecast 2008;24:744–63.
[25] Panagiotelis A, Smith M. Bayesian forecasting of intraday electricity prices using multivariate skew-elliptical distributions. Int J Forecast 2008;24:710–27.
[26] Amjady N. Day-ahead price forecasting of electricity markets by a new fuzzy neural network. IEEE Trans Power Syst 2006;21(2):887–96.
[27] Catalão J, Mariano S, Mendes V, Ferreira L. Short-term electricity prices forecasting in a competitive market: a neural network approach. Electr Power Syst Res 2007;77(10):1297–304.
[28] Nowotarski J, Weron R. Computing electricity spot price prediction intervals using quantile regression and forecast averaging. Comput Stat 2015;30(3):791–803.
[29] Maciejowska K, Nowotarski J, Weron R. Probabilistic forecasting of electricity spot prices using Factor Quantile Regression Averaging. Int J Forecast 2016;32(3):957–65.
[30] Zhao J, Dong Z, Xu Z, Wong K. A statistical approach for interval forecasting of the electricity price. IEEE Trans Power Syst 2008;23(2):267–76.
[31] Chen X, Dong Z, Meng K, Xu Y, Wong K, Ngan H. Electricity price forecasting with extreme learning machine and bootstrapping. IEEE Trans Power Syst 2012;27(4):2055–62.
[32] Wan C, Xu Z, Wang Y, Dong Z, Wong K. A hybrid approach for probabilistic forecasting of electricity price. IEEE Trans Smart Grid 2014;5(1):463–70.
[33] Maciejowska K, Nowotarski J. A hybrid model for GEFCom2014 probabilistic electricity price forecasting. Int J Forecast 2016;32(3):1051–6.
[34] Koenker RW. Quantile regression. Cambridge University Press; 2005.
[35] Gaillard P, Goude Y, Nedellec R. Additive models and robust aggregation for GEFCom2014 probabilistic electric load and electricity price forecasting. Int J Forecast 2016;32(3):1038–50.
[36] Hastie TJ, Tibshirani RJ. Generalized additive models. CRC Press; 1990.
[37] Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc B 1996;58:267–88.
[38] Chernozhukov V, Fernandez-Val I, Galichon A. Quantile and probability curves without crossing. Econometrica 2010;73(3):1093–125.
[39] Dudek G. Multilayer perceptron for GEFCom2014 probabilistic electricity price forecasting. Int J Forecast 2016;32:1057–60.
[40] Juban R, Ohlsson H, Maasoumy M, Poirier L, Kolter JZ. A multiple quantile regression approach to the wind, solar, and price tracks of GEFCom2014. Int J Forecast 2016;32(3):1094–102.
[41] Boyd S, Parikh N, Chu E, Peleato B, Eckstein J. Distributed optimization and statistical learning via the alternating direction method of multipliers. Found Trends Mach Learn 2011;3(1):1–122.
[42] Uniejewski B, Nowotarski J, Weron R. Automated variable selection and shrinkage for day-ahead electricity price forecasting. Energies 2016;9:621.
[43] Ziel F. Forecasting electricity spot prices using LASSO: on capturing the autoregressive intraday structure. IEEE Trans Power Syst 2016;31(6):4977–87.
[44] Janczura J, Weron R. An empirical comparison of alternate regime-switching models for electricity spot prices. Energy Econ 2010;32(5):1059–73.
[45] Zhou M, Yan Z, Ni Y, Li G, Nie Y. Electricity price forecasting with confidence-interval estimation through an extended ARIMA approach. IEEE Proc, Gener Transm Distrib 2006;153(2):187–95.
[46] Christoffersen P. Evaluating interval forecasts. Int Econ Rev 1998;39(4):841–62.
[47] Huurman C, Ravazzolo F, Zhou C. The power of weather. Comput Stat Data Anal 2012;56(11):3793–807.
[48] Berkowitz J. Testing density forecasts, with applications to risk management. J Bus Econ Stat 2001;19(4):465–74.
[49] Bao Y, Lee T-H, Saltoğlu B. Comparing density forecast models. J Forecast 2007;26(3):203–25.
[50] Jonsson T, Pinson P, Madsen H, Nielsen H. Predictive densities for day-ahead electricity prices using time-adaptive quantile regression. Energies 2014;7(9):5523–47.
[51] Alonso A, Garcia-Martos C, Rodriguez J, Sanchez M. Seasonal dynamic factor analysis and bootstrap inference: application to electricity market forecasting. Technometrics 2011;53:137–51.
[52] Khosravi A, Nahavandi S, Creighton D. Quantifying uncertainties of neural network-based electricity price forecasts. Appl Energy 2013;112:120–9.
[53] Pinson P, Tastu J. Discussion of "Prediction intervals for short-term wind farm generation forecasts" and "Combined nonparametric prediction intervals for wind power generation". IEEE Trans Sustain Energy 2014;5(3):1019–20.
[54] Wan C, Xu Z, Østergaard J, Dong ZY, Wong KP. Discussion of "Combined nonparametric prediction intervals for wind power generation". IEEE Trans Sustain Energy 2014;5(3):1021.
[55] Khosravi A, Nahavandi S, Creighton D. A neural network-garch-based method for construction of prediction intervals. Electr Power Syst Res 2013;96:185–93.
[56] Rafiei M, Niknam T, Khooban M. Probabilistic electricity price forecasting by improved clonal selection algorithm and wavelet preprocessing, Neural Computing and Applications; 2016 DOI: http://dx.doi.org/10.1007/s00521-016-2279-7.
[57] Garcia-Martos C, Rodriguez J, Sanchez M. Forecasting electricity prices and their volatilities using Unobserved Components. Energy Econ 2011;33(6):1227–39.
[58] Wu HC, Chan SC, Tsui KM, Hou Y. A new recursive dynamic factor analysis for point and interval forecast of electricity price. IEEE Trans Power Syst 2013;28(3):2352–65.
[59] Bello A, Reneses J, Muñoz A, Delgadillo A. Probabilistic forecasting of hourly electricity prices in the medium-term using spatial interpolation techniques. Int J Forecast 2016;32(3):966–80.
[60] Janczura J, Trück S, Weron R, Wolff R. Identifying spikes and seasonal components in electricity spot price data: a guide to robust modeling. Energy Econ 2013;38:96–110.
[61] Shahidehpour M, Yamin H, Li Z. Market Operations in Electric Power Systems: Forecasting, Scheduling, and Risk Management. Wiley; 2002.
[62] Conejo AJ, Contreras J, Espínola R, Plazas MA. Forecasting electricity prices for a day-ahead pool-based electric energy market. Int J Forecast 2005;21:435–62.
[63] Nowotarski J, Tomczyk J, Weron R. Robust estimation and forecasting of the long-term seasonal component of electricity spot prices. Energy Econ 2013;39:13–27.
[64] Christensen T, Hurn A, Lindsay K. Forecasting spikes in electricity prices. Int J Forecast 2012;28(2):400–11.
[65] Bello A, Reneses J, Muñoz A. Medium-term probabilistic forecasting of extremely low prices in electricity markets: application to the Spanish case. Energies 2016;9(3):193.
[66] Brier GW. Verification of forecasts expressed in terms of probability. Mon Weather Rev 1950;78(1):1–3.
[67] Zareipour H, Janjani A, Leung H, Motamedi A, Schellenberg A. Classification of future electricity market prices. IEEE Trans Power Syst 2011;26(1):165–73.
[68] Ziel F, Steinert R. Electricity price forecasting using sale and purchase curves: the X-model. Energy Econ 2016;59:435–54.
[69] Pinson P, Madsen H, Nielsen HA, Papaefthymiou G, Klöckl B. From probabilistic forecasts to statistical scenarios of short-term wind power production. Wind Energy 2009;12(1):51–62.
[70] Alexander C. Market Risk Analysis IV:Value at Risk Models. Wiley; 2008.
[71] Nowotarski J, Raviv E, Trück S, Weron R. An empirical comparison of alternate schemes for combining electricity spot price forecasts. Energy Econ 2014;46:395–412.
[72] Nowotarski J, Weron R. To combine or not to combine? Recent trends in electricity price forecasting. ARGO 2016;9:7–14.
[73] Efron B, Tibshirani RJ. An Introduction to the Bootstrap. New York: Chapman & Hall; 1993.
[74] Clements MP, Kim JH. Bootstrap prediction intervals for autoregressive time series. Comput Stat Data Anal 2007;51:3580–94.
[75] Nowotarski J, Weron R. Merging quantile regression with forecast averaging to obtain more accurate interval forecasts of Nord Pool spot prices. In: IEEE Conference Proceedings - EEM14, DOI: http://dx.doi.org/10.1109/EEM.2014.6861285; 2014.
[76] Zhang Y, Liu K, Qin L, An X. Deterministic and probabilistic interval prediction for

short-term wind power generation based on variational mode decomposition and machine learning methods. Energy Convers Manag 2016;112:208–19.
[77] Kupiec PH. Techniques for verifying the accuracy of risk measurement models. J Deriv 1995;3(2):73–84.
[78] Dawid AP. Statistical theory: the prequential approach. J R Stat Soc A 1984;147:278–92.
[79] Wallis KF. Chi-squared tests of interval and density forecasts and the Bank of England fan charts. Int J Forecast 2003;19:165–75.
[80] Corradi V, Swanson NR. Predictive density evaluation. In: Handbook of Economic Forecasting (Elliott G, Granger CWJ, Timmermann A. eds.), Elsevier, 2006 pp. 197–284.
[81] Clements MP, Taylor T. Evaluating interval forecasts of high-frequency financial data. J Appl Econ 2003;18(4):445–56.
[82] Berkowitz J, Christoffersen P, Pelletier D. Evaluating value-at-risk models with desk-level data. Manag Sci 2011;57(12):2213–27.
[83] Christoffersen P, Pelletier D. Backtesting value-at-risk: a duration-based approach. J Financ Econ 2004;2(1):84–108.
[84] Santos APP, Alves MF. A new class of independence tests for interval forecast evaluation. Comput Stat Data Anal 2012;56(11):3366–80.
[85] Engle RF, Manganelli S. CAViaR: conditional autoregressive value at risk by regression quantiles. J Bus Econ Stat 2004;22(4):367–81.
[86] Gaglianone W, Lima L, Linton O, Smith D. Evaluating Value-at-Risk models via quantile regression. J Bus Econ Stat 2011;29(1):150–60.
[87] Granger CWJ. Prediction with a generalized cost of error function. Oper Res Q 1969;20:199–207.
[88] Gneiting T. Quantiles as optimal point forecasts. Int J Forecast 2011;27(2):197–207.
[89] Winkler RL. A decision-theoretic approach to interval estimation. J Am Stat Assoc 1972;67(337):187–91.
[90] Diebold FX, Mariano RS. Comparing predictive accuracy. J Bus Econ Stat 1995;13:253–63.
[91] Diebold FX. Comparing predictive accuracy, twenty years later: a personal perspective on the use and abuse of Diebold-Mariano tests. J Bus Econ Stat 2015;33(1):1–9.
[92] Hansen PR, Lunde A, Nason JM. The model confidence set. Econometrica 2011;79:453–97.
[93] Harvey D, Leybourne S, Newbold P. Tests for forecast encompassing. J Bus Econ Stat 1998;16:254–9.
[94] Matheson JE, Winkler RL. Scoring rules for continuous probability distributions. Manag Sci 1976;22:1087–96.
[95] Good IJ. Rational decisions. J R Stat Soc B 1952;14:107–14.
[96] Pinson P, Kariniotakis G. Conditional prediction intervals of wind power generation. IEEE Trans Power Syst 2010;25(4):1845–56.
[97] Taylor JW. Evaluating volatility and interval forecasts. J Forecast 1999;18:111–28.
[98] Uniejewski B, Weron R, Ziel F. Variance stabilizing transformations for electricity spot price forecasting. IEEE Trans Power Syst 2017, [Submitted for publication].
[99] Winkler RL, Murphy AH. Good probability assessors. J Appl Meteorol 1968;7:751–8.
[100] Morales JM, Conejo AJ, Madsen H, Pinson P, Zugno M. Integrating Renewables in Electricity Markets: Operational Problems. Springer; 2014.
[101] Wan C, Xu Z, Pinson P, Dong Z, Wong K. Probabilistic forecasting of wind power generation using extreme learning machine. IEEE Trans Power Syst 2014;29(3):1033–44.
[102] Khosravi A, Nahavandi S. Closure to the discussion of "Prediction intervals for short-term wind farm generation forecasts" and "Combined nonparametric prediction intervals for wind power generation" and the discussion of "Combined nonparametric prediction intervals for wind power generation". IEEE Trans Sustain Energy 2014;5(3):1022–3.
[103] Elliott G, Timmermann A. Economic forecasting. Princeton University Press; 2016.
[104] Morgan MG, Henrion M, Small M. Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis. Cambridge University Press; 1990.
[105] Paraschiv F, Bunn D.W, Westgaard S. Estimation and application of fully parametric multifactor quantile regression with dynamic coefficients, 2016. University of St. Gallen, School of Finance Research Paper No. 2016/07. Available at SSRN: ⟨https://ssrn.com/abstract=2741692⟩.
[106] Unger D.A. A method to estimate the continuous ranked probability score. In: Proceedings of the 9th Conference on Probability and Statistics in Atmospheric Sciences, Virginia Beach, VA, 206–213; 1985.
[107] Hersbach H. Decomposition of the continuous ranked probability score for ensemble prediction systems. Weather Forecast 2000;15:559–70.
[108] Taieb SB, Huser R, Hyndman RJ, Genton MG. Forecasting uncertainty in electricity smart meter data by boosting additive quantile regression. IEEE Trans Smart Grid 2016;7(5):2448–55.
[109] Cuaresma JC, Hlouskova J, Kossmeier S, Obersteiner M. Forecasting electricity spot-prices using linear univariate time-series models. Appl Energy 2004;77(1):87–106.
[110] Gianfreda A, Grossi L. Forecasting Italian electricity zonal prices with exogenous variables. Energy Econ 2012;34(6):2228–39.
[111] Hong Y-Y, Wu C-P. Day-ahead electricity price forecasting using a hybrid principal component analysis network. Energies 2012;5(11):4711–25.
[112] Bordignon S, Bunn DW, Lisi F, Nan F. Combining day-ahead forecasts for British electricity prices. Energy Econ 2013;35:88–103.
[113] Maciejowska K, Weron R. Forecasting of daily electricity prices with factor models: utilizing intra-day and inter-zone relationships. Comput Stat 2015;30(3):805–19.
[114] Maciejowska K, Weron R. Short- and mid-term forecasting of baseload electricity prices in the UK: the impact of intra-day price relationships and market fundamentals. IEEE Trans Power Syst 2016;31(2):994–1005.
[115] Tastu J, Pinson P, Trombe P-J, Madsen H. Probabilistic forecasts of wind power generation accounting for geographically dispersed information. IEEE Trans Smart Grid 2014;5(1):480–9.
[116] Baran S, Lerch S. Mixture EMOS model for calibrating ensemble forecasts of wind speed. Environmetrics 2016;27(2):116–30.
[117] Nowotarski J, Weron R. On the importance of the long-term seasonal component in day-ahead electricity price forecasting. Energy Econ 2016;57:228–35.
[118] Karakatsani N, Bunn D. Forecasting electricity prices: the impact of fundamentals and time-varying coefficients. Int J Forecast 2008;24:764–85.
[119] Schneider S. Power spot price models with negative prices. J Energy Mark 2011;4(4):77–102.
[120] Diaz G, Planas E. A note on the normalization of Spanish electricity spot prices. IEEE Trans Power Syst 2016;31(3):2499–500.
[121] Nogales FJ, Contreras J, Conejo AJ, Espinola R. Forecasting next-day electricity prices by time series models. IEEE Trans Power Syst 2002;17:342–8.
[122] Misiorek A. Short-term forecasting of electricity prices: do we need a different model for each hour?. Medium Econom Toepass 2008;16(2):8–13.
[123] Kristiansen T. Forecasting Nord Pool day-ahead prices with an autoregressive model. Energy Policy 2012;49:328–32.
[124] Keles D, Scelle J, Paraschiv F, Fichtner W. Extended forecast methods for day-ahead electricity spot prices applying artificial neural networks. Appl Energy 2016;162:218–30.