Applied Energy 293 (2021) 116983

Contents lists available at ScienceDirect

Applied Energy

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

Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark

Jesus Lago [a,*], Grzegorz Marcjasz [b], Bart De Schutter [a], Rafał Weron [b]

[a] Delft Center for Systems and Control, Delft University of Technology, Delft, The Netherlands
[b] Department of Operations Research and Business Intelligence, Wrocław University of Science and Technology, Wrocław, Poland

ARTICLE INFO

Keywords:
Electricity price forecasting
Regression model
Deep learning
Open-access benchmark
Forecast evaluation
Best practices

ABSTRACT

While the field of electricity price forecasting has benefited from plenty of contributions in the last two decades, it arguably lacks a rigorous approach to evaluating new predictive algorithms. The latter are often compared using unique, not publicly available datasets and across too short and limited to one market test samples. The proposed new methods are rarely benchmarked against well established and well performing simpler models, the accuracy metrics are sometimes inadequate and testing the significance of differences in predictive performance is seldom conducted. Consequently, it is not clear which methods perform well nor what are the best practices when forecasting electricity prices. In this paper, we tackle these issues by comparing state-of-the-art statistical and deep learning methods across multiple years and markets, and by putting forward a set of best practices. In addition, we make available the considered datasets, forecasts of the state-of-the-art models, and a specifically designed `python` toolbox, so that new algorithms can be rigorously evaluated in future studies.

1. Introduction

The increasing penetration of renewable energy sources in today's power systems makes electricity generation more volatile and the resulting electricity prices harder to predict than ever before [1–4]. On the other hand, advances in *electricity price forecasting* (EPF) constantly provide new tools with the ultimate objective of narrowing the gap between predictions and actual prices. The progress in this field, however, is not steady and easy to follow. In particular, as concluded by all major review publications, comparisons between EPF methods are very difficult since studies use different datasets, different software implementations, and different error measures; the lack of statistical rigor complicates these analyses even further [5–8]. In particular:

- There are several studies comparing *machine learning* (ML) and statistical methods but the conclusions of these studies are contradictory. Typically, studies considering advanced statistical techniques only compare them with simple ML methods [9–11] and show that statistical methods are obviously better. Conversely, studies proposing new ML methods only compare them with simple statistical methods [12–16] and show that ML models are more accurate.
- In many of the existing studies [17–23] the testing periods are too short to yield conclusive results. In some cases, the test datasets are limited to one-week periods [22,24–30]; this ignores the problem of special days, e.g. holidays, and is not representative for the performance of the proposed algorithms across a whole year. As argued in [5], to have meaningful conclusions, the test dataset should span at least a year.
- Some of the existing papers do not provide enough details to reproduce the research. The three most common issues are: (i) not specifying the exact split between the training and test dataset [31–37], (ii) not indicating the inputs used for the prediction model [35,36,38–40], and (iii) not specifying the dataset employed [21,33,41,42]. This obviously prevents other researchers from validating the research results.

These three problems have aggravated over the last years with the increase in popularity of *deep learning* (DL). While new published papers on DL for EPF appear almost every month, and most claim to develop models that obtain state-of-the-art accuracy, the comparisons performed in those papers are very limited. Particularly, the new DL methods are usually compared with simpler ML methods [28,30,43–47]. This is obviously problematic as such comparisons are not fair. Moreover, as the proposed methods are not compared with other DL algorithms, new DL methods are continuously being proposed but it is unclear how the different models perform relative to each other.

---

[*] Corresponding author.
E-mail address: j.lagogarcia@tudelft.nl (J. Lago).

https://doi.org/10.1016/j.apenergy.2021.116983
Received 29 December 2020; Received in revised form 15 April 2021; Accepted 18 April 2021
Available online 26 April 2021
0306-2619/© 2021 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Similar problems arise in the context of *hybrid methods*. In recent years, very complex hybrid methods have been proposed. Typically, these hybrid models are based on combining a decomposition technique, a feature selection method, an ML regression model, and sometimes a meta-heuristic algorithm for optimization purposes. As with DL algorithms, these studies usually avoid comparisons with well-established methods [21,25,34,42,48–50] or resort to comparisons using outdated methodologies [22,24,26,37,41,51,52]. In addition, while a specific genetic algorithm or decomposition technique is considered, most of the studies do not analyze the effect of selecting a variant of these techniques [21,24,50–52]. Thus, the relative importance of each of the different components of the hybrid methods it is not even clear.

1.1. Motivation and contributions

The above mentioned problems call for three actions. Firstly, implementing in a popular programming environment (e.g. `python`) and making available *a set of simple but powerful open-source forecasting methods*, which can potentially obtain state-of-the-art performance, and that researchers can easily use to evaluate any new forecasting model.

Secondly, collecting and making freely available to the EPF community *a set of representative benchmark datasets* that researchers can use to evaluate and compare their methods using long testing periods. Although, some datasets are available for download without restrictions, e.g. as supplements to published articles [53] or sample transaction data [54], they are typically limited in scope (one market, a 2–3 year timespan or price series only). Hence, conclusions from such datasets are limited, results can hardly be extrapolated to other markets, and the relevance of the studies using such data are not entirely clear.

Thirdly, *putting forward a set of best practices* so that the conclusions of EPF studies become more meaningful and fair comparisons can be made.

In this paper, we try to tackle the above via three distinct contributions:

1. We analyze the existing literature and select what could arguably be considered as state-of-the-art among statistical and machine learning methods: the *Lasso Estimated AutoRegressive* (LEAR) model[1] [55] and the *Deep Neural Network* (DNN) [57], a relatively simple and automated DL method that optimizes hyperparameters and features using Bayesian optimization. Then, we make our models available to other researchers as part of an open-source `python` library (https://github.com/jeslago/epftoolbox) specially designed to provide a common research framework for EPF research [58]. Besides the models, we also provide extensive documentation [59] for the library.
2. We propose a set of five open-access benchmark datasets spanning six years each, that represent a range of well-established day-ahead, auction type power markets from around the globe. The datasets contain day-ahead electricity prices at an hourly resolution and two relevant exogenous variables each. They can be accessed from the mentioned `python` library [58]. Together with the datasets, the library also includes the forecasts of the open-access methods across the five benchmark datasets so that researchers can quickly make further comparisons without having to re-train or re-estimate the models.
3. We provide a set of best practice guidelines to conduct research in EPF so that new studies are more sound, reproducible, and the obtained conclusions are stronger. In addition, we include some of the guidelines, e.g. adequate evaluation metrics or statistical tests, in the mentioned `python` library [58] to provide a common research framework for EPF research.

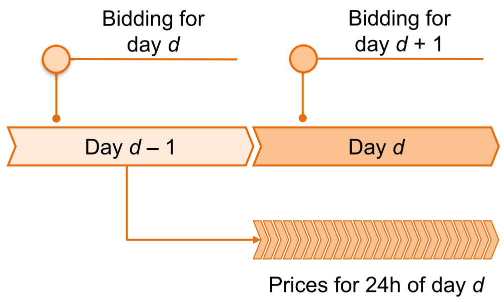

Fig. 1. Illustration of the *day-ahead* auction market, where wholesale sellers and buyers submit their bids before gate closure on day $d-1$ for the delivery of electricity during day d; the 24 hourly prices for day d are set simultaneously, typically around midday.

1.2. Paper structure

The remainder of the paper is organized a follows. Section 2 performs a literature review of the current state of EPF. Sections 3 and 4 respectively present the open-access benchmark datasets and the open-source benchmark models. Section 5 describes the set of guidelines and best practices when performing research in EPF. Section 6 discusses the forecasting results for all five datasets. Finally, Section 7 provides a summary and a checklist of the requirements for meaningful EPF research.

2. Literature review

The field of EPF aims at predicting the spot and forward prices in wholesale markets, either in a point or probabilistic setting. However, given the diversity of trading regulations available across the globe, EPF always has to be tailored to the specific market. For instance, the workhorse of European short-term power trading is the *day-ahead* market with its once-per-day uniform-price auction, see Fig. 1. On the other hand, the Australian National Electricity Market operates as a real-time power pool, where a dispatch price is determined every five minutes and six dispatch prices are averaged every half hour as pool prices [60], while electricity forward markets share many aspects with those of other energy commodities (oil, gas, coal), and quite often are only financially settled [61].

As the field of EPF is very diverse, a complete literature review is out of the scope of this paper. Instead, this section is intended to provide an overview of the three families of methods, i.e. statistical, ML, and hybrid methods, proposed for point forecasting in day-ahead markets since 2014, i.e. since the last comprehensive literature review of Weron [5]. The more recent reviews either focused on short-term [6] and medium-/long-term [7] probabilistic EPF, were not that comprehensive in scope [62,63], or concerned electricity derivatives [61]. Furthermore, our survey puts a special emphasis on DL and hybrid methods as this is the area of EPF characterized by the most rapid development and, at the same time, troubled by non-rigorous empirical studies which motivated us to write this paper in the first place.

2.1. Statistical methods

Most models in this class rely on linear regression and represent the dependent (or output) variable, i.e. the price $p_{d,h}$ for day d and hour h, by a linear combination of independent (or predictor, explanatory) variables, also called regressors, inputs, or features:

$$p_{d,h} = \boldsymbol{\theta}_h \mathbf{X}_{d,h} + \varepsilon_{d,h}, \tag{1}$$

where $\boldsymbol{\theta}_h = [\theta_{h,0}, \theta_{h,1}, \dots, \theta_{h,n}]$ is a row vector of coefficients specific to hour h, $\mathbf{X}_{d,h} = [1, X^1_{d,h}, \dots, X^n_{d,h}]^\top$ is a column vector of inputs and $\varepsilon_{d,h}$ is an error term; the intercept $\theta_{h,0}$ can be set to zero if the data is demeaned beforehand. Note that here we are using a notation common in day-ahead forecasting, which emphasizes the vector structure of these price series, see Fig. 1. Alternatively we could use single indexing: p_t with $t = 24d + h$. Although the multivariate modeling framework has been shown to be marginally more accurate than the univariate counterpart, both approaches have their pros and cons [64,65].

---

[1] Originally introduced in [55] under the name *LassoX* and based on the *fARX* model, a parameter-rich autoregressive specification with exogenous variables. The name refers to the *least absolute shrinkage and selection operator* (LASSO) [56] used to jointly select features and estimate their parameters.

In the last few years, there have been several key contributions in the field of statistical methods for EPF. Arguably, the most relevant of them has been the appearance of linear regression models with a large number of input features that utilize regularization techniques [56,66]. Classically, the regression model in (1) is estimated using *ordinary least squares* (OLS) by minimizing the *residual sum of squares* (RSS), i.e. squared differences between the predicted and actual values. However, if the number of regressors is large, using the *least absolute shrinkage and selection operator* (LASSO) [56] or its generalization the *elastic net* [66] as implicit feature selection methods have been shown to improve the forecasting results [55,57,64,67–69], also in intraday [70,71] and probabilistic [72,73] EPF. In particular, by jointly minimizing the RSS and a penalty factor of the model parameters (see Section 4.2 for details), these two implicit regularization techniques set some of the parameters to zero and thus effectively eliminate redundant regressors. As shown in the cited studies, these parameter-rich[2] regularized regression models exhibit superior performance. It is important to note that such an approach, called here *Lasso Estimated AutoRegressive* (LEAR), is in fact hybrid since LASSO (and electric nets) are considered ML techniques by some authors. However, we classify it as statistical because the underlying model is *autoregressive* (AR).

Aside from proposing parameter-rich models and advanced estimators, researchers have also improved the field by considering a variety of additional preprocessing techniques. Most notably, models using so-called variance stabilizing transformations [9,64,74,75] and long-term seasonal components [76–79] have been proposed and shown to result in statistically significant improvements. However, the applicability of these two techniques varies greatly: due to very common occurrence of price spikes, variance stabilizing transformations have become a standard and replaced the commonly used logarithmic transformation (no longer applicable due to zeros and negative values[3]) to normalize electricity prices. By contrast, the applicability of long-term seasonal components has been more limited and it is unknown whether their beneficial effect is limited to relatively parsimonious regression models or also holds for parameter-rich models.

A third innovation in the field is an ensemble (i.e. a method that combines individual forecasting models) that combines multiple forecasts of the same model calibrated on different windows. In this context, two different studies [80,81] showed that the best results are obtained with a combination of a few short (spanning 1–4 months) and a few long calibration windows (of approximately two years). Said ensembles were able to significantly outperform predictions obtained for the best ex-post selected calibration window [80–82]. But again, it has not been shown to date whether this effect is limited to relatively parsimonious regression models or also holds for LEAR models.

Interestingly, as [83] argue in an econometric context, in the presence of structural breaks it may be advisable to combine forecasts obtained for calibration windows of different lengths. Longer windows allow for a better fit, while shorter faster adapt to changes. Hence, if a structural break appears, like the COVID-19 pandemic, using models calibrated to shorter windows may better capture changes in the price dynamics. A different, but a potentially also appealing approach has been recently suggested in [84,85]. The authors assume that fundamental and price time series exhibit recurrent regimes with similar dynamics and employ cluster analysis – k-means [84] or k-nearest neighbors [85] – to identify such periods in the past. Then they calibrate models only on data segments which resemble current conditions. As such, they are able to eliminate subperiods that include structural breaks from the calibration sample.

Finally, note that in contrast to financial econometrics, where heteroskedasticity is a basic building block of many state-of-the-art approaches [86], models with *generalized autoregressive conditional heteroskedastic* (GARCH) residuals have been tried for EPF without much success, for a review and discussion see [5]. For instance, [57] compare 27 different models, among them an ARIMA–GARCH model, and find that it performs comparable to a much simpler AR model and ca. 1.5 times worse than the DNN model defined in Section 4.3. As [87] argue, GARCH effects diminish when fundamental and behavioral drivers of the electricity price volatility are taken into account and allowing for the time-varying responses of prices to fundamentals can yield more precise volatility estimates than an explicit GARCH specification.

2.2. Deep learning

In the last five years, a total of 28 deep learning papers in the context of EPF have been published.[4] Moreover, this number has been steadily increasing: while in 2017 there was only one paper, in 2018 there were 11, and in 2019 there were 16. Despite this trend, most of the published studies are very limited: the comparisons are too simplistic, e.g. avoid state-of-the-art statistical methods, and their results cannot be generalized.

The first published DL paper [12] proposes a deep learning network using stacked denoising autoencoders. The paper, despite being the first, provides a better evaluation than most studies: the new method is compared not only against machine learning techniques but also against two statistical methods. Yet, the evaluation is limited as it only considers three months of test data and simple benchmark models. In the second published DL article [57], a DNN for modeling market integration is proposed. While the method is evaluated over a year of data, the proposed model is not compared against other machine learning or statistical methods.

In the third published paper [57], four DL models (a DNNs, two *recurrent neural networks (RNNs)*, and a *convolutional network (CNN)*) are proposed. This study is, to the best of our knowledge, the most complete study up to date. In particular, the proposed DL models are compared using a whole year of data against a benchmark of 23 different models, including 7 machine learning models, 15 statistical methods, and a commercial software. Moreover, among the statistical methods, the comparison includes the fARX-Lasso and fARX-EN, i.e. the state-of-the-art statistical methods. While the study shows the superiority of the DL algorithms, very strong conclusions are not possible as the study only considers a single market.

The studies that followed in 2018 focused on one of three topics: (1) evaluating the performance of different deep recurrent networks [13,23,37,88]; (2) proposing new hybrid methods based on CNNs and LSTMs [14,44,89,90]; or (3) employing regular DNN models [23]. Independently of the focus, they were all more limited than the first and the third studies [12,57] as they failed to compare the new DL models with state-of-the-art statistical methods and/or to employ long enough datasets to derive strong conclusions.

---

[2] We define a parameter-rich linear model as a model with multiple regressors (dozens, hundreds).

[3] The logarithm of 0 or of a negative value is undefined.

[4] This data is primarily based on a Scopus search in the title, abstract, and keywords: `TITLE-ABS-KEY((((‘‘forecasting electricity’’) OR (‘‘predicting electricity’’)) AND ((‘‘electricity spot’’) OR (‘‘electricity day-ahead’’) OR (‘‘electricity price’’))) OR (((‘‘price forecasting’’) OR (‘‘price prediction’’) OR (‘‘forecasting price’’) OR (‘‘predicting price’’) OR (‘‘forecasting spikes’’) OR (‘‘forecasting VAR’’)) AND ((‘‘electricity spot price’’) OR (‘‘electricity price’’) OR (‘‘electricity market’’) OR (‘‘day-ahead market’’) OR (‘‘power market’’))) AND (‘‘deep’’) AND (‘‘learning’’))`. We have also run a second, more general query replacing (“deep”) AND (“learning”) by (“neural”) AND (“network”), however, only a few additional papers have been identified.

In detail, [13] studies the use of RNNs for forecasting electricity prices but the comparison is done in a single market and against simple statistical methods: a seasonal *auto regressive integrated moving average* (ARIMA) model, a Markov regime-switching model, and a self exciting threshold model. Moreover, while the comparison includes other DL methods, it avoids comparison with simpler ML techniques. Ref. [44] proposes a hybrid DL method composed of a CNN and a *long short-term memory* (LSTM) neural network (a type of recurrent network) for forecasting balancing prices. However, the new model is only compared against simple ML benchmarks and the evaluation is done using different periods comprising three months for training and 1 month for testing. Similarly, [14] proposes another hybrid model combining a CNN and an LSTM, but the model is only compared against two simple statistical methods: an *auto regressive moving average* (ARMA) and a GARCH model.

In [23] a regular DNN model is proposed but the model is only evaluated on a test dataset comprising a single day and compared against a simple *multilayer perceptron* (MLP). In [29], the use of an LSTM model for EPF is evaluated, but the method is only compared with three neural networks and a simple statistical method, and the evaluation is done using only 4 weeks of data. Likewise, [88] proposes a model based on an LSTM but a comparison against other methods is not performed and the test dataset only comprises 2 weeks of data. In [37], another LSTM model is proposed but, as in other studies, the test dataset comprises a few months of data and the method is only compared against a simple decision tree and a support vector regressor; moreover, the exact split between the training and test dataset is not specified and it is unclear what is exactly the performance of the model. An exception to these studies is [91] which proposes a series of DL models and compares them for a year of data against several advanced statistical methods such as LASSO and a simpler ML method. The main drawbacks of the study are that it is based on a single market and that it only considers a simple ML method as a benchmark. In addition, the study focuses on intraday electricity prices, while most of the literature (including the current paper) considers forecasting day-ahead electricity prices.

In 2019, the main focus of the papers was the same as in 2018: (1) evaluating the performance of different deep recurrent networks (mostly LSTMs) [16,30,45,47,92–94], (2) proposing new hybrid deep learning methods usually based on LSTMs and CNNs [17,28,36,92,95–97], or (3) employing regular DNN models [15,46,98]. Similarly, as with most studies in 2018, the new studies were more limited than [12, 57] as no comparisons with state-of-the-art statistical methods were made and long test datasets were seldom used. In this context, even though some studies [16,98] tried to compare the proposed methods with existing DL models [57], they either failed to re-estimate the benchmark models for the new case study [16] or they overfitted the DL benchmark models [98].

In detail, [30] proposes different LSTM models but the new models are only compared against 5 other ML techniques and using a test period of 4 weeks. In [28], a CNN model is proposed but the new model is just compared against three simple ML methods and using a test dataset that comprises a week. In [45], a model based on an LSTM is proposed but it is only compared against three simple ML methods and for a period of 12 weeks. In [46], the performance of a DNN is compared to that of an SVR model and, as the comparison only includes these two models, it is obviously very limited. In [15], a DNN is used as part of a two-step forecasting method; as in many other studies, the comparison is performed for one month of data and limited to two simple ML models (a SVR and an MLP) and a standard linear model. In [47], two DL models are proposed but the models are only compared to very simple ML methods (extreme learning machines and standard MLPs) and using a test dataset spanning eight months. In [16], a bidirectional LSTM to forecast prices in the French market is proposed; however, the study only considers historical prices as input features and the proposed method is only compared against DL models and a simple autoregressive model. In addition, the benchmark DL models are copied from [57] (a completely different case study that considers exogenous inputs and a different market) without re-tuning the hyperparameters to the new case study.

In [98], a neural network that uses data from order books is proposed and compared against DL methods from the literature, e.g. the ones proposed in [57]. While the new model outperforms existing DL methods, the DL methods from the literature are trained to overfit the training dataset.[5] Therefore, the comparison is not meaningful (the DL benchmark models will necessarily perform poorly in the test dataset) and it cannot be assessed how the new model performs. In [95], a hybrid DL forecasting method is proposed based on stacked denoising autoencoders for pre-training, regular autoencodes for feature selection, and a rough DNN as a forecasting method. As in other studies, the method is only compared against simple ML models. Moreover, the importance of each of the four modules of the hybrid method is not studied and the authors do not re-calibrate the models with new data: the models are trained once and evaluated over a whole year. Similarly, [96] proposes a CNN hybrid model that uses mutual information, random forests, gray correlation analysis, and recursive feature elimination for feature selection. Unlike most models, the algorithm is trained to classify prices instead of predicting their scalar values; however, details of how this process is done are not provided. In addition, the method is only compared against simple ML methods and evaluated for less than a year of data (the study uses one year for testing and training but the split is not specified). Likewise, [36] proposes a hybrid model based on CNNs and RNNs in the context of microgrids; as in other studies, the method is evaluated in a small dataset, it is not compared against state-of-the-art statistical methods, and the exact split between training and test datasets is not specified.

2.3. Hybrid methods

Within the field of EPF, the research area that has received the most attention in the last 5 years has been hybrid forecasting methods. In this time frame, more than 100 articles proposing new hybrid methods have been published,[6] i.e. approximately 5 times more than articles based on DL. Hybrid models are very complex forecasting frameworks that are composed of several algorithms. Usually, they comprise at least two of the following five modules:

- An algorithm for decomposing data.
- An algorithm for feature selection.
- An algorithm to cluster data.
- One or more forecasting models whose predictions are combined.
- Some type of heuristic optimization algorithm to either estimate the models or their hyperparameters.

---

[5] In the training dataset, the proposed model and some naive ML benchmark models yield a *root mean square error* (RMSE) of ca. 6. For the test dataset, for the same models, the RMSE is between 9 and 12. By contrast, the training error of the benchmark DL model is 2, and the test error is 20. Having a training error that is 1/3 of the error of other models but a test error that is 10 times larger than the training error is a clear sign for overfitting (especially when for the rest of the models the test error is just 1.5 larger than the training error).

[6] This data is based on two searches in Scopus looking for keywords in the title, abstract, and keywords. The first search is based on the following query `TITLE-ABS-KEY(((forecast*) OR (predict*)) AND (electricity) AND (price*) AND (hybrid))`. The second search is very similar but replacing the keyword `hybrid` by `neural AND network`. Note that, while this search is not as complete as the one for DL, it provides enough material for building an overview of the state of the field.

In terms of decomposition methods, the most widely used technique is the wavelet transform [17,19,22,24,34,41,49,51,52,99]. Alternative methods include empirical mode decomposition (EMD) [32, 100], the Hilbert–Huang transform which uses EMD to decompose a signal and then applies Hilbert spectral analysis [101], variational mode decomposition [27,48], and singular spectrum analysis [102, 103].

For feature selection, the most commonly utilized algorithms are correlation analysis [32,41,42,104,105] and the mutual information technique [18,42,52,106–108]. Other algorithms include classification and regression trees with recursive feature elimination [50] or Relief-F [50].

For clustering data, the algorithms are usually based on one of the following four: k-means [26,109], self-organizing maps [19,26,110], enhanced game theoretic clustering [26], or fuzzy clustering [52,111]

The selection of forecasting models is much more diverse. The most widely used method is the standard MLP [19,20,32,41,42,51,102,103, 105,107,108], followed by the *adaptive network-based fuzzy inference system* (ANFIS) [19,100,106], radial basis function network [20,24, 111], and autoregressive models like ARMA or ARIMA [20,22,24,100]. Other models include LSTM [17], linear regression [50], extreme learning machine [22,50], CNN [50], Bayesian neural network [26,110], exponential GARCH [100], echo state neural network [27], Elman neural networks [18], and support vector regressors [20]. It is important to note that in many of the approaches, the hybrid method does not consider a single forecasting model but combines several of them [19,20,24,50,100,108].

Just as for the forecasting model, the diversity of the heuristic optimization algorithms is also large. While the most often utilized algorithm is particle swarm optimization [22,48,51,106,107,111], many other approaches are also used: differential evolution [27], genetic algorithm [106], backtracking search [106], deterministic annealing [111], bat algorithm [41], vaporization precipitation-based water cycle algorithm [104], cuckoo search [103,105], or honey bee mating optimization [24].

In spite of the large number of published works, the research in hybrid methods suffers from the same problems as discussed earlier. First, most of the studies either avoid comparison with well-established methods [18–21,25,27,34,42,48–50,100,104,106,111] or resort to comparisons using outdated methodologies [22,24,26,41,51, 52,102,103]. Hence, the accuracy of the new proposed methods cannot be accurately established.

Second, the considered studies usually employ very small datasets consisting either of a few days [17–22] or a few weeks [18,19,22,24–27,41,42,49,51,102–104,106,111]. Thus, drawing conclusions is nearly impossible and it is unclear whether the accuracy results are just the outcome of selecting a convenient test period.

Besides these two problems, for many hybrid methods the effect of selecting variants of the different hybrid components is not analyzed [20,21,24,25,27,41,42,50–52,102,103]. Thus, it is not clear how relevant or useful the individual components are.

2.4. State-of-the-art models

Because of the described problems when comparing EPF models, it is very hard to establish what are the state-of-the-art methods. Nevertheless, considering the studies performed in the last years, it can be argued that the LEAR is a very accurate (if not the most accurate) linear model. Moreover, it can also be argued that the accuracy of this model can be further improved by transforming the prices using variance stabilizing transformations, combining forecasts obtained for different calibration windows, and/or using long-term seasonal decomposition.

For the case of ML models, the selection is harder as the existing comparisons are of worse quality. Considering the most complete benchmark study in terms of forecasting models [57], it seems that a simple DNN with two layers is one of the best ML models. In particular, while more complex models, e.g. LSTMs, could potentially be more accurate, there is at the moment no sound evidence to validate this claim.

In the case of hybrid models, establishing what is the best model is an impossible task. Firstly, while many hybrid methods have been proposed, they have not been compared with each other nor with the LEAR or DNN models. Secondly, as most studies do not evaluate the individual influence of each hybrid component, it is also impossible to establish the best algorithms for each hybrid component, e.g. it is unclear what are the best clustering, feature selection method, or data decomposition methods.

With that in mind, we will consider the LEAR and the DNN for the proposed open-access benchmark. In particular, not only are these two methods highly accurate, but they are also relatively simple. As such, we think that they are the best benchmarks to compare new complex EPF forecasting methods with.

3. Open-access benchmark dataset

The first contribution of the paper is to provide a large open-access benchmark dataset on which new methods can be tested, together with the day-ahead forecasts of the proposed open-access methods. In this section, we introduce this dataset, which can be accessed[7] using the `python` library built for this study.

3.1. General characteristics

For a benchmark dataset in EPF to be fair it should satisfy three conditions:

1. comprise several electricity markets so that the capabilities of new models can be tested under different conditions,
2. be long enough so that algorithms can be analyzed using out-of-sample datasets that span 1–2 years, and
3. be recent enough to include the effects of integrating renewable energy sources on wholesale prices.

Based on these conditions, we propose five datasets representing five different day-ahead electricity markets, each of them comprising 6 years of data. The prices of each market have very distinct dynamics, i.e. they all have differences in terms of the frequency and existence of negative prices, zeros, and price spikes. In addition, as electricity prices depend on exogenous variables, each dataset comprises two additional time series: day-ahead forecasts of two influential exogenous factors that differ for each market. The length of each dataset equals 2184 days, which translates to six 364-day "years" or 312 weeks.[8] All available time series are reported using the local time, and the daylight savings are treated by either arithmetically averaging two values for the extra hour or interpolating the neighboring values for the missing observation.

3.2. Nord Pool

The first dataset represents the Nord Pool (NP), i.e. the European power market of the Nordic countries, and spans from 01.01.2013 to 24.12.2018. The dataset contains hourly observations of day-ahead prices, the day-ahead load forecast, and the day-ahead wind generation forecast. The dataset was constructed using data freely available on the webpage of the Nordic power exchange Nord Pool [54]. Fig. 2(b) (top) displays the electricity price time series of the dataset; as can be seen, the prices are always positives, zero prices are rare, and prices spikes seldom occur.

---

[7] Note that we do not own the data in the dataset. However, it can be freely accessed from different websites, e.g. the ENTSO-E transparency platform [112]. In this context, the proposed `python` library [58,59] provides an interface to easily access the data.

[8] Electricity prices exhibit weekly seasonality. Thus, by approximating a year by 52 weeks because we ensure that the metrics are not impacted by a certain day, e.g. Monday, being harder to predict than the others.

3.3. PJM

The second dataset is obtained from the *Pennsylvania–New Jersey–Maryland* (PJM) market in the United States. It covers the same time period as Nord Pool, i.e. from 01.01.2013 to 24.12.2018. The three time series are: the zonal prices in the *Commonwealth Edison* (COMED) (a zone located in the state of Illinois) and two day-ahead load forecast series, one describing the system load and the second one the COMED zonal load. The data is freely available on the PJM's website [113]. Fig. 2(b) (bottom) depicts the electricity price time series of the dataset; as with the NP market, the prices are always positive and zero prices are rare; however, unlike the NP market, spikes appear frequently.

3.4. EPEX-BE

The third dataset represents the EPEX-BE market, the day-ahead electricity market in Belgium, which is operated by EPEX SPOT. The dataset spans from 09.01.2011 to 31.12.2016. The two exogenous data series represent the day-ahead load forecast and the day-ahead generation forecast in France. While this selection might be surprising, it has been shown [57] that these two are the best predictors of Belgian prices. The price data is freely available in the ENTSO-E transparency platform [112] and the ELIA website [114], and the load and generation day-ahead forecasts are freely available in [115]. It is important to note that this dataset is particularly interesting because it is harder to predict. Fig. 3 (top) shows the electricity price time series of the dataset; unlike the prices in the PJM and NP markets, negative prices and zero prices appear more frequently, and price spikes are very common.

3.5. EPEX-FR

The fourth dataset represents the EPEX-FR market, the day-ahead electricity market in France, which is also operated by EPEX SPOT. The dataset spans the same period as the EPEX-BE dataset, i.e. from 09.01.2011 to 31.12.2016. Besides the electricity prices, the dataset comprises the day-ahead load forecast and the day-ahead generation forecast. As before, the price data is freely obtained from the ENTSO-E transparency platform [112], and the load and generation day-ahead forecasts are freely available on the webpage of RTE [115], i.e. the *transmission system operator* (TSO) in France. Fig. 3 (middle) displays the electricity price time series of the dataset; as in the EPEX-BE market, negative prices, zero prices, and spikes are very common.

3.6. EPEX-DE

The last dataset describes the EPEX-DE market, the German electricity market, which is also operated by EPEX SPOT. The dataset spans from 09.01.2012 to 31.12.2017. Besides the prices, the dataset comprises the day-ahead zonal load forecast in the TSO Amprion zone and the aggregated day-ahead wind and solar generation forecasts in the zones of the 3 largest[9] TSOs (Amprion, TenneT, and 50Hertz). The price data is freely obtained from the ENTSO-E transparency platform [112], the zonal load day-ahead forecasts is freely available in the website of Amprion [116], and the wind and solar forecasts in the websites of Amprion [116], 50Hertz [117], and TenneT [118]. Fig. 3 (bottom) displays the electricity price time series of the dataset; as can be seen, while negative and zero prices occur more often than in the other four markets, price spikes are more rare.

Table 1
Start and end dates of the testing (out-of-sample) datasets for each electricity market.

Market	Test period
Nord pool	27.12.2016–24.12.2018
PJM	27.12.2016–24.12.2018
EPEX-FR	04.01.2015–31.12.2016
EPEX-BE	04.01.2015–31.12.2016
EPEX-DE	04.01.2016–31.12.2017

3.7. Training and testing periods

For each dataset, the testing period is defined as the last 104 weeks, i.e. the last two years, of the dataset. The exact dates of the testing datasets are defined in Table 1. It is important to note that, as we will argue in Section 5, selecting two years as the testing period is paramount to ensure good research practices in EPF.

Unlike the testing dataset, the training dataset cannot be defined as it will vary between different models. In general, the training dataset will comprise any data that is known prior to the target day. However, the exact data will change depending on two concepts, i.e. calibration window and recalibration:

- While there are four years of data available for estimating the model, it might be desirable to employ only recent data, e.g. to avoid estimating effects that no longer play a role. The amount of past data employed for estimation defines the calibration window.
- The model can be estimated once and then evaluated for the full test dataset, or it can be continuously recalibrated on a daily basis to incorporate the input of recent data.

For example, let us consider predicting the NP prices on 15.02.2017. A model using a calibration window of 52 weeks and no recalibration would employ a training dataset comprising the data between 29.12.2016 and 26.12.2016, i.e. one year prior to the start of the test period. By contrast, a model using a calibration window of 104 weeks and daily recalibration would employ the data between 18.02.2015 and 14.02.2017.

4. Open-access benchmark models

The second contribution of the paper is to provide a set of state-of-the-art forecasting methods as an open-source `python` toolbox. As explained in Section 2.4, the LEAR [55] and the DNN [57] models are not only highly accurate but also relatively simple. Therefore, we implement these two methods and provide their code freely available as part of the proposed toolbox [58,59]. It is important to note that the use of the proposed open-access methods is fully documented and automated so researchers can test and use them without expert knowledge.

For the sake of simplicity, the description provided here is limited to the bare minimum. For further details on the two models we refer to the original papers [55,57].

4.1. Input features

Before describing each model, let us define the input features that are considered. Independently of the model, the available input features to forecast the 24 day-ahead prices of day d, i.e. $\mathbf{p}_d = [p_{d,1}, \dots, p_{d,24}]^\top$, are the same:

- Historical day-ahead prices of the previous three days and one week ago, i.e. $\mathbf{p}_{d-1}$, $\mathbf{p}_{d-2}$, $\mathbf{p}_{d-3}$, $\mathbf{p}_{d-7}$.
- The day-ahead forecasts of the two variables of interest (see Section 3 for details) for day d available on day $d-1$, i.e. $\mathbf{x}_d^1 = [x_{d,1}^1, \dots, x_{d,24}^1]^\top$ and $\mathbf{x}_d^2 = [x_{d,1}^2, \dots, x_{d,24}^2]^\top$; note that the variables of interest are different for each market.

---

[9] There are 4 TSOs in Germany.

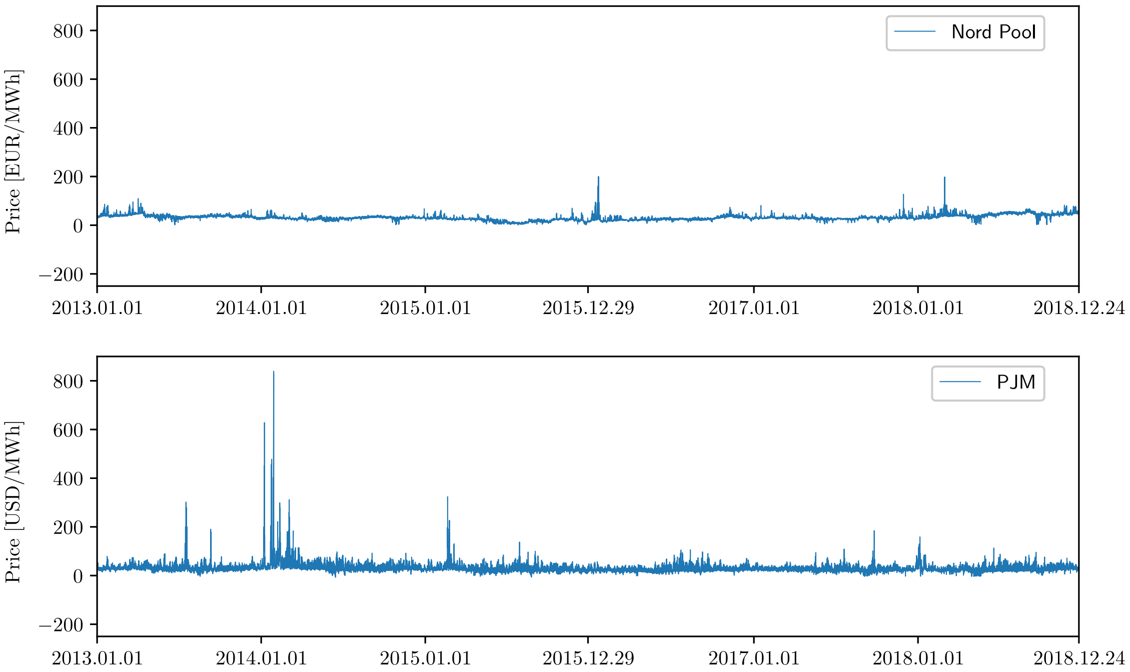

Fig. 2. Electricity price time series for two of the five datasets, i.e. Nord Pool and PJM, considered in the open-access benchmark dataset. Note that each dataset also includes two exogenous time series that are not plotted here.

- Historical day-ahead forecasts of the variables of interest the previous day and one week ago, i.e. $\mathbf{x}^1_{d-1}$, $\mathbf{x}^1_{d-7}$, $\mathbf{x}^2_{d-1}$, $\mathbf{x}^2_{d-7}$.
- A dummy variable $\mathbf{z}_d$ that represents the day of the week. In the case of the linear model, following the standard practice in the literature [55,69,81], this is a binary vector $\mathbf{z} = [z_{d,1}, \ldots, z_{d,7}]^\top$ that encodes every day of the week by setting all elements to zero except the element that identifies the day of the week, e.g. $[1,0,0,0,0,0,0]$ represents Monday and $[0,1,0,0,0,0,0]$ Tuesday. In the case of the neural network, for the sake of simplicity, the day of the week is modeled with a multi-value input $z_d \in \{1,\ldots,7\}$.

Overall, we consider a total of 247 available input features for each LEAR model and 241 input features for each DNN model. It is important to note that, while the available input features are the same, the LEAR and DNN models utilize a different feature selection procedure. Namely, each of the LEAR models finds the optimal set of features using LASSO as an embedded feature selection, i.e. each model uses L1-regularization to select among the 247 features. On the other hand, in the DNN model, as in the original study [57], the input features are optimized together with the hyperparameters using the tree Parzen estimator [119] (see Section 4.3 for details). Finally, it should be emphasized that for both types of models the feature selection is fully automated and does not require expert intervention.

4.2. *The LEAR model*

The *Lasso Estimated AutoRegressive* (LEAR) model is a parameter-rich ARX structure estimated using L1-regularization, i.e. the LASSO [56]. It was originally introduced in [55] under the name *LassoX*. The LEAR is based on the so-called *full ARX* or *fARX* model, a parameter-rich autoregressive specification with exogenous variables, which in turn is inspired by the general autoregressive model defined by Equation (2) in [68], with some important differences. While fARX includes fundamentals and a much richer seasonal structure, it does not look too far into the past and concentrates only on the last week of data. Note, that very similar models to the LEAR were used in [64] under the name *24lasso*$_{DoW,nl}$ and in [69] under the name *24Lasso*$_1$.

To enhance the model, as empirically tested and recommended in [9,64,69], the data is preprocessed with the *area* (or *inverse*) *hyperbolic sine* variance stabilizing transformation:

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

where x is the price standardized by subtracting the in-sample median and dividing by the median absolute deviation adjusted by a factor for asymptotically normal consistency to the standard deviation, for details see [9]. Long-term seasonal decomposition is not considered for the sake of simplicity; particularly, while it has been shown to further improve the performance of the LEAR, we leave it out for future research.

As in [81], to further enhance the model, we recalibrate it daily over different calibration window lengths: 8 weeks, 12 weeks, 3 years, and 4 years. We consider short windows (8–12 weeks) in combination with long windows (3–4 years) because it has been empirically shown to lead to better results [81]. In this context, we consider a minimum of 8 weeks as lower windows might not have enough information to correctly estimate parameter-rich models [81].

The LEAR model to predict price $p_{d,h}$ on day d and hour h is defined by:

$$\begin{aligned}
p_{d,h} =& f(\mathbf{p}_{d-1}, \mathbf{p}_{d-2}, \mathbf{p}_{d-3}, \mathbf{p}_{d-7}, \mathbf{x}^i_d, \mathbf{x}^i_{d-1}, \mathbf{x}^i_{d-7}, \boldsymbol{\theta}_h) + \varepsilon_{d,h} \\
=& \sum_{i=1}^{24} \theta_{h,i} \cdot p_{d-1,i} + \sum_{i=1}^{24} \theta_{h,24+i} \cdot p_{d-2,i} \\
&+ \sum_{i=1}^{24} \theta_{h,48+i} \cdot p_{d-3,i} + \sum_{i=1}^{24} \theta_{h,72+i} \cdot p_{d-7,i} \\
&+ \sum_{i=1}^{24} \theta_{h,96+i} \cdot x^1_{d,i} + \sum_{i=1}^{24} \theta_{h,120+i} \cdot x^2_{d,i} \\
&+ \sum_{i=1}^{24} \theta_{h,144+i} \cdot x^1_{d-1,i} + \sum_{i=1}^{24} \theta_{h,168+i} \cdot x^2_{d-1,i} \\
&+ \sum_{i=1}^{24} \theta_{h,192+i} \cdot x^1_{d-7,i} + \sum_{i=1}^{24} \theta_{h,216+i} \cdot x^2_{d-7,i} \\
&+ \sum_{i=1}^{7} \theta_{h,240+i} \cdot z_{d,i} + \varepsilon_{d,h}
\end{aligned} \tag{3}$$

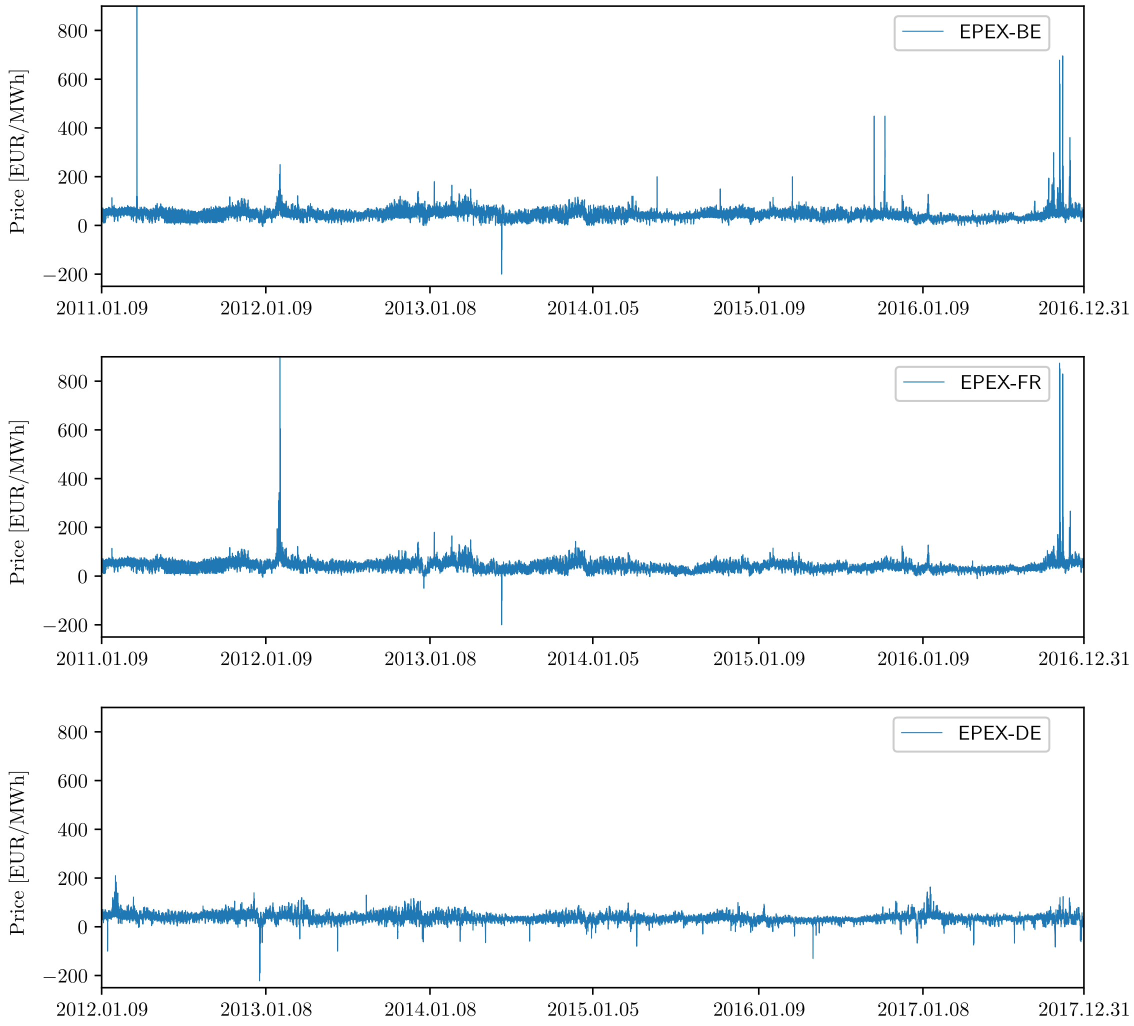

Fig. 3. Electricity price time series for three of the five datasets, i.e. EPEX-BE, EPEX-FR, and EPEX-DE, considered in the open-access benchmark dataset. Note that each dataset also includes two exogenous time series that are not plotted here. The EPEX-BE and EPEX-FR time series are similar because the EPEX-FR and EPEX-BE are highly coupled markets [57]. To keep the plots readable, the upper limit of the y-axis is below the maximum price; this only affects one spike in EPEX-FR and another one in EPEX-BE.

where $\theta_h = [\theta_{h,1}, \dots, \theta_{h,247}]^\top$ are the 247 parameters of the LEAR model for hour h. Many of these parameters become zero when (3) is estimated using LASSO:

$$\hat{\theta}_h = \underset{\theta_h}{\operatorname{argmin}}\ \text{RSS} + \lambda \left\|\theta_h\right\|_1 = \underset{\theta_h}{\operatorname{argmin}}\ \text{RSS} + \lambda \sum_{i=1}^{247} \left|\theta_{h,i}\right|, \qquad (4)$$

where $\text{RSS} = \sum_{d=8}^{N_d} (p_{d,h} - \hat{p}_{d,h})^2$ is the sum of squared residuals, $\hat{p}_{d,h}$ the price forecast, N_d is the number of days in the training dataset, and $\lambda \geq 0$ is the *tuning* (or *regularization*) hyperparameter of LASSO. Due to the computational speed of estimating with LASSO, during every daily recalibration, the hyperparameter λ that regulates the L_1 penalty is optimized. This can be done using an *ex-ante* cross-validation procedure [120]. In this study, to further reduce the computational cost, we propose an efficient hybrid approach to perform the optimal selection of λ. See Section 4.2.2 for details.

4.2.1. Regularization hyperparameter

The hyperparameter λ of LASSO can be optimized in multiple ways, each with different advantages and disadvantages. A first approach is to optimize λ once and then keep it fixed for the whole test period. Although it requires very low computation costs, the limitation of this approach is that it assumes that the hyperparameter λ does not change over time. This assumption might hinder the performance of the estimator as the regularization parameter does not change even when the market might do.

A second approach is to recalibrate the hyperparameter on a periodic basis using a validation dataset. Although this method yields good results, tuning the recalibration frequency and calibration window is complicated, the computational cost is large, and the results may vary between datasets [69].

A third option is to recalibrate the hyperparameter periodically, but using *cross-validation* (CV): splitting the data into disjoint partitions, using each possible partition once as a test dataset with the remaining data as the training dataset, and selecting the hyperparameter that performs the best across all partitions [120]. Although this approach is highly accurate, its computation costs are very large.

A fourth option is to periodically update the hyperparameter but using information criteria, e.g. the *Akaike information criterion* (AIC) or the Bayesian information criterion [64,68,121]. As before, this involves training multiple LASSO models to compute the information criteria for each possible hyperparameter value, which in turn leads to a high computational cost.

Lastly, one can use the *least angle regression* (LARS) LASSO [122] for estimating the model instead of the coordinate descent implementation. This estimation procedure has the advantage of computing the whole LASSO solution path, which in turn allows to compute the information criteria or perform CV much faster.

4.2.2. Selecting the regularization hyperparameter

To select λ we propose a hybrid approach. On a daily basis, we estimate the hyperparameter using the LARS method with the in-sample

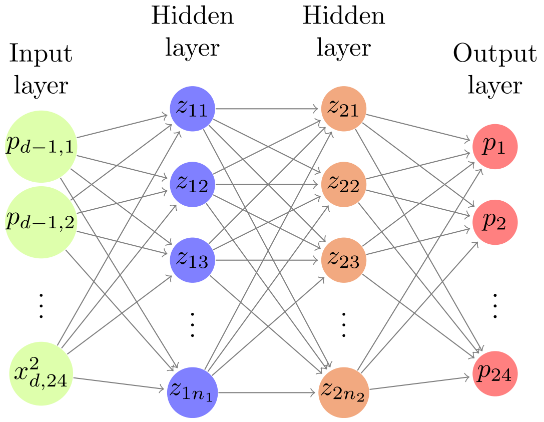

Fig. 4. Visualization of a sample DNN model.

AIC. Then, using the optimal λ obtained from the LARS method, we recalibrate the LEAR using the traditional coordinate descent implementation.

The reason for proposing this hybrid approach is that it provides a good trade-off between computational complexity and accuracy. In particular, it leverages the computational efficiency of LARS for ex-ante λ selection with the predictive performance on short calibration windows of the coordinate descent LASSO.

It is important to note that we have studied multiple approaches to select λ: (i) daily recalibration, CV, with coordinate descent; (ii) daily recalibration, CV, with LARS; (iii) daily recalibration with LARS and AIC. However, the computational cost of the first method was too high (in the same order of magnitude as the cost of the DNN model), and the accuracy of the other two was not good. By contrast, the proposed approach had a performance on par with coordinate descent LASSO using CV, but with a computational cost that was an order of magnitude lower.

4.3. *The DNN model*

The second model is the DNN [57], one of the simplest DL models whose input features and hyperparameters can be optimized and tailored for each case study without the need for expert knowledge. The DNN is a straightforward extension of the traditional multilayer perceptron (MLP) with two hidden layers.

4.3.1. *Structure*

The DNN is a deep feedforward neural network that contains 4 layers, employs the multivariate framework (single model with 24 outputs), is estimated using Adam [123], and its hyperparameters and input features are optimized using the tree Parzen estimator [119], i.e. a Bayesian optimization algorithm. Its structure is visualized in Fig. 4.

4.3.2. *Training dataset*

For estimating the hyperparameters, the training dataset is fixed and comprises the four years prior to the testing period. For evaluating the testing dataset, the DNN is recalibrated on a daily basis using a calibration window of four years.

In all cases, the training dataset is split into a training and a validation dataset, with the latter being used for two purposes: performing early stopping [124] to avoid overfitting and optimizing hyperparameters/features. While the validation dataset always comprises 42 weeks, the split between the training and validation datasets depends on whether the validation dataset is used for hyperparameter/feature selection or for the recalibration step:

- For estimating the hyperparameters, as the validation dataset is used to guide the optimization process, the validation dataset is selected as the last 42 weeks of the training dataset. This is done to keep the training and validation datasets completely independent and to avoid overfitting.[10]
- For the testing phase, as the validation dataset is only used for early stopping, it is defined by randomly selecting 42 weeks out of the total 208 weeks employed for training. This is done to ensure that the dataset used for optimizing the DNN parameters includes up-to-date data.[11]

As example, let us consider the training and evaluation of a DNN in the Nord Pool market. Before evaluating the DNN, the hyperparameter and features of the DNN are optimized. For that, the employed dataset comprises the data between 01.01.2013 and 26.12.2016, of which the training dataset represents the first 166 weeks, i.e. 01.01.2013 to 07.03.2016, and the validation dataset the last 42, i.e. 08.03.2016 to 26.12.2016. During the evaluation of the model, i.e. after the hyperparameter and feature selection, the training and validation datasets comprise the last four years of data but are randomly shuffled. For example, to evaluate the DNN during 15.02.2017, the training and validation datasets would represent the data between 20.02.2013 and 14.02.2017, of which 166 randomly selected weeks would define the training dataset and the remaining 42 the validation dataset.

4.3.3. *Hyperparameter and feature selection*

As in the original DNN paper [57], the hyperparameters and input features are optimized together using the tree-structured Parzen estimator [119], a Bayesian optimization algorithm based on sequential model-based optimization. To do so, the features are modeled as hyperparameters, with each hyperparameter representing a binary variable that selects whether or not a specific feature is included in the model (as explained in [57]). In more detail, to select which of the 241 available input features are relevant, the method employs 11 decision variables, i.e. 11 hyperparameters:

- Four binary hyperparameters (1–4) that indicate whether or not to include the historical day ahead prices $\mathbf{p}_{d-1}$, $\mathbf{p}_{d-2}$, $\mathbf{p}_{d-3}$, $\mathbf{p}_{d-7}$. The selection is done per day,[12] e.g. the algorithm either selects all 24 hourly prices $\mathbf{p}_{d-j}$ of j days ago or does not select any price from day $d-j$, hence the four hyperparameters.
- Two binary hyperparameters (5–6) that indicate whether or not to include each of the day-ahead forecasts $\mathbf{x}^1_d$ and $\mathbf{x}^2_d$. As with the past prices, this is done for the whole day, i.e. a hyperparameter either selects all the elements in $\mathbf{x}^j_d$ or none.
- Four binary hyperparameters (7–10) that indicate whether or not to include the historical day-ahead forecasts $\mathbf{x}^1_{d-1}$, $\mathbf{x}^2_{d-1}$, $\mathbf{x}^1_{d-7}$, and $\mathbf{x}^2_{d-7}$. This selection is also done per day.
- One binary hyperparameter (11) that indicates whether or not to include the variable z_d representing the day of the week.

In short, 10 binary hyperparameters indicating whether or not to include 24 inputs each and another binary hyperparameter indicating whether or not to include a dummy variable.

---

[10] Similar as it is done when splitting the dataset between the training and the test dataset.

[11] For hyperparameter optimization, as the validation dataset represents the most recent weeks of data, the neural network is trained with data that is almost one year old. While this is not a big problem when deciding on the DNN structure, it should be avoided during testing to ensure that the DNN captures new market effects.

[12] This is done for the sake of simplicity to speed up the optimization procedure of the feature selection. In particular, an alternative could be to use a binary hyperparameter for each individual historical prices; however, in most markets, that would mean using 24 as many hyperparameters as there are 24 different prices per day.

Besides selecting the features, the algorithm also optimizes eight additional hyperparameters: (1) the number of neurons per layer, (2) the activation function, (3) the dropout rate, (4) the learning rate, (5) whether or not to use batch normalization, (6) the type of data preprocessing technique, (7) the initialization of the DNN weights, and (8) the coefficient for L1 regularization that is applied to each layer's kernel.

Unlike the weights of the DNN that are recalibrated on a daily basis, the hyperparameter and features are optimized only once using the four years of data prior to the testing period. It is important to note that the algorithm runs for a number T of iterations, where at every iteration the algorithm infers a potential optimal subset of hyperparameters/features and evaluates this subset in the validation dataset. For the proposed open-access benchmark models, T is selected as 1500 iterations to obtain a trade-off between accuracy and computational requirements.[13]

4.4. Ensembles

For the open-access benchmark, in order to have benchmark predictions when evaluating ensemble techniques, we also propose ensembles of LEAR and DNNs as open-access benchmarks of ensembles methods. For the LEAR, the ensemble is built as the arithmetic average of forecasts across four calibration window lengths: 8 weeks, 12 weeks, 3 years, and 4 years. For the DNN, the ensemble is built as the arithmetic average of four different DNNs that are estimated by running the hyperparameter/feature selection procedure four times. In particular, the hyperparameter optimization is asymptotically deterministic, i.e. the global optimum is found for an infinite number of iterations. However, for a finite number of iterations and using a different initial random seed, the algorithm is non-deterministic and every run provides a different set of hyperparameters and features. Although each of these hyperparameter/feature subsets represent a local minimum, it is impossible to establish which of the subsets is better as their relative performance on the validation dataset is nearly identical. This effect can be explained by the fact that the DNN is a very flexible model and thus different network architectures are able to obtain equally good results.

4.5. Software implementation

The proposed open-access models are developed in `python`: the LEAR is implemented using the `scikit-learn` library [125] and the DNN model using the `Keras` library [126]. The reason for selecting `python` is that it is one of the most widely used programming languages, especially in the context of ML and statistical inference.

5. Guidelines and best practices in EPF

As motivated in the introduction, the field of EPF suffers from several problems that prevent having reproducible research and establishing strong conclusions on what methods work best. In this section, we outline some of these issues and provide some guidelines on how to address them.

5.1. Length of the test period

A common practice in EPF is to evaluate new methods on very short test periods. The typical approach is to evaluate the method on 4 weeks of data [18,19,22,24–26,29,30,41,42,49,51,97,102–107,110], with each week representing one of the four seasons in the year. This is problematic for three reasons:

- Selecting four weeks can lead to cherry-picking the weeks where a given method excels, e.g. a method that performs bad with spikes could be evaluated in a week with fewer spikes, leading in turn to biased estimations of the forecasting accuracy. While this is an ethical issue that most researchers would avoid, establishing four week testing periods as the standard does facilitate malpractice and should be avoided.
- Assuming that the four weeks are randomly selected and no bias is introduced in the selection, it is still not possible to guarantee that these four weeks are representative of the price behavior over a whole year. Particularly, even within a given season, the price dynamics can change dramatically, e.g. during winter there are weeks with a lot of sun and wind but there are also weeks without them. Therefore, picking only a week per season rarely represents the average performance of a forecaster in a given dataset.
- There are situations in the electrical grid that do not occur very often but that can have a very large effect on electricity prices, e.g. when several power plants are under maintenance at the same time. Forecasting methods need to be evaluated under those conditions to ensure that they are also accurate under extreme events. By selecting four weeks most of these effects are neglected.

To avoid this problem, we recommend using a minimum of one year as a testing period. This ensures that forecasting methods are evaluated considering the complete set of effects that take place during the year. To guarantee that all researchers have access to this type of data, the open-access benchmark dataset that we propose contains data from several markets and employs a testing period of two years. In addition, the open-access benchmark can be directly accessed using the proposed `epftoolbox` library [58,59].

5.2. Benchmark models

A second issue with many EPF publications is that new methods are not compared with well-established methods [14,16,18–21,23,25,27,34,36,42,46,48–50,88,100,104,106,111] or resort to comparisons using either outdated methodologies or simplified methods [13,15,22,24,26,28–30,37,41,44,45,47,51,52,95,96,102,103].

This poses a problem since it becomes very hard to establish which algorithms work best and which ones do not. To address this issue, we recommend using well-established state-of-the-art open-source methods and a common benchmark dataset. With that in mind, we have provided and made freely available an open-access benchmark dataset comprising 5 markets (as described in Section 3), and we have implemented, thoroughly tested, and made freely available two state-of-the-art forecasting methods (as described in Section 4) and their day-ahead predictions for all 5 datasets over a period of two years (as described in Section 6). Additionally, we have implemented all these resources in an easy-to-use toolbox [58] and adequately documentated it [59].

5.3. Open-access

A third issue in the field of EPF is that datasets are usually not made publicly available and the code of the proposed methods is not shared. This poses four obvious problems:

---

[13] It can be empirically observed that the performance of the models barely improves after 1000 iterations. Moreover, performing 1500 iterations takes approximately just one day on a regular quadcore laptop like the i7-6920HQ, a computation cost very acceptable when the algorithm has to run only once.

- Research cannot be reproduced as data is not available. This goes against one of the main principles of science as all research should be reproducible.
- The progress of EPF research is hindered since it is hard to establish which methodologies work well. Consequently, researchers spend unnecessary time re-evaluating methodologies that have been evaluated already.
- Comparing new methods with published ones becomes very challenging because researchers have to re-implement methods from the literature. As a result, comparisons with state-of-the-art methods are often avoided, and new methods are usually compared with simple and easy-to-implement methods.
- When new methods are proposed, they cannot be compared with published methods under the same circumstances. This leads to comparisons under different conditions and opens up the door to wrong implementations of the original methods, which in turn leads to results that are not correct.

As these problems are critical, we directly try to address them by providing an open-access benchmark/toolbox comprising five datasets, two state-of-the-art methods, and a set of day-ahead forecasts of the latter two methods. In addition, we encourage researchers in EPF to share the developed codes and to either share their datasets or use an open-access benchmark dataset.

5.4. Evaluation metrics for point forecasts

In the field of EPF, the most widely used metrics to measure the accuracy of point forecasts are the *mean absolute error* (MAE), the *root mean square error* (RMSE), and the *mean absolute percentage error* (MAPE):

$$\text{MAE} = \frac{1}{24\,N_{\text{d}}}\sum_{d=1}^{N_{\text{d}}}\sum_{h=1}^{24}|p_{d,h} - \hat{p}_{d,h}|, \tag{5}$$

$$\text{RMSE} = \sqrt{\frac{1}{24\,N_{\text{d}}}\sum_{d=1}^{N_{\text{d}}}\sum_{h=1}^{24}(p_{d,h} - \hat{p}_{d,h})^2}, \tag{6}$$

$$\text{MAPE} = \frac{1}{24\,N_{\text{d}}}\sum_{d=1}^{N_{\text{d}}}\sum_{h=1}^{24}\frac{|p_{d,h} - \hat{p}_{d,h}|}{|p_{d,h}|}, \tag{7}$$

where $p_{d,h}$ and $\hat{p}_{d,h}$ respectively represent the real and forecasted price on day d and hour h, and N_{d} is the number of days in the out-of-sample test period, i.e. in the test dataset.

Since absolute errors are hard to compare between different datasets, the MAE and RMSE are not always very informative. Moreover, since electricity costs and profits are often linearly dependent on the electricity prices, metrics based on quadratic errors, e.g. RMSE, are hard to interpret and do not accurately represent the underlying problem of most forecasting users. In particular, in most electricity trade applications, the underlying risk, profits, and costs depend linearly on the price and on the forecasting errors. Hence, linear metrics represent better than quadratic metrics the underlying risks of forecasting errors.

Similarly, since MAPE values become very large with prices close to zero (regardless of the actual absolute errors), the MAPE is usually dominated by the periods of low prices and is also not very informative. While the *symmetric mean absolute percentage error* (sMAPE) defined[14] as:

$$\text{sMAPE} = \frac{1}{24\,N_{\text{d}}}\sum_{d=1}^{N_{\text{d}}}\sum_{h=1}^{24}2\frac{|p_{d,h} - \hat{p}_{d,h}|}{|p_{d,h}| + |\hat{p}_{d,h}|} \tag{8}$$

solves some of these issues, it has (as any metric based on percentage errors) a statistical distribution with undefined mean and infinite variance [128].

5.4.1. Scaled errors

In this context, several studies advocate for the use of scaled errors [5,128,129], where a scaled error is simply the MAE scaled by the in-sample MAE of a naive forecast. A scaled error has the nice interpretation of being lower/larger than one if it is better/worse than the average naive forecast evaluated in-sample.

A metric based on this concept is the *mean absolute scaled error* (MASE), and in the context of one-step ahead forecasting is defined as [128]:

$$\text{MASE} = \frac{1}{N}\sum_{k=1}^{N}\frac{|p_k - \hat{p}_k|}{\frac{1}{n-1}\sum_{i=2}^{n}|p_i^{\text{in}} - p_{i-1}^{\text{in}}|}, \tag{9}$$

where p_i^{in} is the i^{th} price in the in-sample, i.e. training, dataset (note that in EPF $i = 24d + h$), p_{i-1}^{in} is the one-step ahead naive forecast of p_i^{in}, i.e. $\hat{p}_i^{\text{in}}$, N is the number of out-of-sample (test) datapoints, and n the number of in-sample (training) datapoints. For seasonal time series, the MASE may be defined using the MAE of a seasonal naive model in the denominator [5,129].

5.4.2. Relative measures

While scaled errors do indeed solve the issues of more traditional metrics, they have other associated problems that make them unsuitable in the context of EPF:

1. As MASE depends on the in-sample dataset, forecasting methods with different calibration windows will naturally have to consider different in-sample datasets. As a result, the MASE of each model will be based on a different scaling factor and comparisons between models cannot be drawn.
2. The same argument applies to models with and without rolling windows. The latter will use a different in-sample dataset at every time point while the former will keep the in-sample dataset constant.
3. In ensembles of models with different calibration windows, the MASE cannot be defined as the calibration window of the ensemble is undefined.
4. Drawing comparisons across different time series is problematic as electricity prices are not stationary. For example, an in-sample dataset with spikes and an out-of-sample dataset without spikes will lead to a smaller MASE than if we consider the same market but with the in-sample/out-sample datasets reversed.

To solve these issues, we argue that a better metric is the *relative* MAE (rMAE) [128,130]. Similar to MASE, it normalizes the error by the MAE of a naive forecast. However, instead of considering the in-sample dataset, the naive forecast is built based on the out-of-sample dataset. For day-ahead electricity prices of hourly frequency, rMAE is defined as:

$$\text{rMAE} = \frac{\frac{1}{24N_{\text{d}}}\sum_{d=1}^{N_{\text{d}}}\sum_{h=1}^{24}|p_{d,h} - \hat{p}_{d,h}|}{\frac{1}{24N_{\text{d}}}\sum_{d=1}^{N_{\text{d}}}\sum_{h=1}^{24}|p_{d,h} - \hat{p}_{d,h}^{\text{naive}}|}, \tag{10}$$

where the $\frac{1}{24N_d}$ factor cancels out in the numerator and the denominator. There are three natural choices for the naive forecasts:

- $\hat{p}_{d,h}^{\text{naive},1} = p_{d-1,h}$,
- $\hat{p}_{d,h}^{\text{naive},2} = p_{d-7,h}$,
- $\hat{p}_{d,h}^{\text{naive},3} = \begin{cases} p_{d-1,h}, & \text{if } d \text{ is Tue, Wed, Thu, or Fri,} \\ p_{d-7,h}, & \text{if } d \text{ is Sat, Sun, or Mon.} \end{cases}$

In the context of EPF, rMAE using $\hat{p}_{d,h}^{\text{naive},2} = p_{d-7,h}$ is arguably the best choice for two reasons: (i) it is easier to compute than the one based on $\hat{p}_{d,h}^{\text{naive},3}$ and, unlike the rMAE based on $\hat{p}_{d,h}^{\text{naive},1}$, it captures weekly effects; (ii) given a set of forecasting models, the relative ranking of the accuracy of the models is independent from the naive benchmark used (see the last paragraph of this subsection for an explanation). Hence, in the remainder of the article we will use rMAE to explicitly refer to the rMAE based on $\hat{p}^{\text{naive},2}_{d,h}$. It is important to note that, similar to rMAE, one could also define the *relative* RMSE (rRMSE) by dividing the RMSE of each forecast by the RMSE of a naive forecast.

---

[14] Note, that there are multiple versions of sMAPE, here we consider the most sensible one according to [127].

Since the dependence on the in-sample dataset is removed, using a rolling window is no longer a problem as the out-of-sample dataset stays the same. Similarly, models with different calibration windows can be compared and the rMAE of ensembles is properly defined. Moreover, as the metric is normalized by the MAE of a naive forecast for the same sample, the problem with drawing conclusions in non-stationary time series is mitigated.

Due to its better properties, rMAE should always be used to evaluate new methods in EPF. In particular, while it can be used in conjunction with other metrics, it is important to include and employ rMAE to obtain more fair evaluations and comparisons.

With that in mind, the accuracy of the open-access models in the open-access benchmark dataset is computed considering rMAE, sMAPE, MAPE, MAE, and RMSE. Then, an analysis of the different metrics is provided (see Section 6.4.2). Finally, the forecasts themselves are provided as csv files so that the accuracy results can be updated in case more adequate metrics are developed in the future.

As a final remark, let us note that, given a set of forecasting models, the relative ranking of the accuracy of the models is independent from the naive benchmark used for the rMAE or MASE. Changing it simply changes the denominator but preserves the numerator, and since the change in the denominator is the same across all methods, the relative ranking is preserved. Furthermore, as the numerator is the MAE, it follows that the ranking based on the rMAE or MASE will be the same as that based on the MAE.

5.5. Statistical testing

While using adequate metrics to compare the accuracy of the forecasts is important, it is also necessary to analyze whether any difference in accuracy is statistically significant. This is paramount to conclude whether the difference in accuracy does really exist and is not simply due to random differences between the forecasts. Despite its importance, the use of statistical testing has been downplayed in the EPF literature [5]. In particular, most publications only compare the accuracy in terms of an error metric and do not analyze the statistical significance of the accuracy differences. This trend needs to be corrected in order to compare forecasting approaches with statistical rigor. Particularly, new studies need to ensure that:

- Any new method is compared against well-established methods using a statistical test.
- The forecasts of the proposed methods are provided as open-access datasets. This ensures that, when new models are proposed, the difference in accuracy with the published methods can be analyzed in terms of statistical testing.

To facilitate statistical testing, we include in the proposed open-source `epftoolbox` library [58,59] the two most widely used statistical tests in EPF, i.e. the Diebold–Mariano and the Giacomini–White tests.

5.5.1. The Diebold–Mariano test

The Diebold–Mariano (DM) test [131] is probably the most commonly used tool to evaluate the significance of differences in forecasting accuracy. It is an asymptotic z-test of the hypothesis that the mean of the *loss differential* series:

$$\Delta^{\text{A,B}}_{d,h} = L(\varepsilon^{\text{A}}_{d,h}) - L(\varepsilon^{\text{B}}_{d,h}) \tag{11}$$

is zero, where $\varepsilon^{\text{Z}}_{d,h} = p_{d,h} - \hat{p}_{d,h}$ is the prediction error of model Z for day d and hour h, and $L(\cdot)$ is the loss function. For point forecasts, we usually take $L(\varepsilon^{\text{Z}}_{d,h}) = |\varepsilon^{\text{Z}}_{d,h}|^p$ with $p = 1$ or 2, which corresponds to the absolute and squared losses, respectively; for probabilistic forecasts, $L(\cdot)$ may be any strictly proper scoring rule, in particular the pinball loss, the *continuous ranked probability score* (CRPS), or the energy score [6,63,65]. Given the loss differential series, we compute the statistic:

$$\text{DM} = \sqrt{N}\frac{\hat{\mu}}{\hat{\sigma}}, \tag{12}$$

where $\hat{\mu}$ and $\hat{\sigma}$ are the sample mean and standard deviation of $\Delta^{\text{A,B}}_{d,h}$, respectively, and N is the length of the out-of-sample test period. Under the assumption of covariance stationarity of $\Delta^{\text{A,B}}_{d,h}$, the DM statistic is asymptotically standard normal, and one- or two-sided asymptotic tail probabilities can be easily computed.

It is important to note three things. Firstly, the DM test is model-free, i.e. it compares forecasts (of models), not models themselves. Secondly, although in the standard formulation [131] the DM test compares forecasts via the null hypothesis of the expected loss differential being zero, it is more informative to compute the p-values of two one-sided tests:

1. with the null hypothesis $H_0 : E(\Delta^{\text{A,B}}_{d,h}) \leq 0$,
2. with the alternative hypothesis null $H_1 : E(\Delta^{\text{A,B}}_{d,h}) \geq 0$.

The lower the p-value,[15] i.e. the closer it is to zero, the more the observed data is inconsistent with the null hypothesis. If the p-value is less than the commonly accepted level of 5%, the null hypothesis is typically rejected. In the DM test, this means that the forecasts of model B are significantly more accurate than those of model A.

Thirdly, the DM test requires (only) that the loss differential be covariance stationary.[16] This may not be satisfied by forecasts in day-ahead markets, since the predictions for all 24 h of the next day are computed at the same time, using the same information set. Hence, following [63], we recommend two variants of the DM test in the context of day-ahead EPF:

- a *univariate* variant with 24 independent tests performed,[17] one for each hour of the day, and comparisons based on the number of hours for which the predictions of one model are significantly better than those of another, i.e. the number of hours for which the null hypothesis is rejected,
- a *multivariate* variant with the test performed jointly for all hours using the 'daily' or multivariate loss differential series:

  $$\Delta^{\text{A,B}}_{d} = \|\varepsilon^{\text{A}}_{d}\|_p - \|\varepsilon^{\text{B}}_{d}\|_p, \tag{13}$$

  where $\varepsilon^{\text{Z}}_{d}$ is the 24-dimensional vector of prediction errors of model Z for day d, $\|\varepsilon^{\text{Z}}_{d}\|_p = (\sum_{h=1}^{24} |\varepsilon^{\text{Z}}_{d,h}|^p)^{1/p}$ is the pth norm of that vector with $p = 1$ or 2.

The univariate version of the test has the advantage of providing a deeper analysis as it indicates which forecast is significantly better for which hour of the day [6,55,57,65,133,134]. The multivariate version, introduced in [64], enables a better representation of the results as it summarizes the comparison in a single p-value, which can be conveniently visualized using heat maps arranged as chessboards [9,10,69,80], see Fig. 5.

---

[15] Recall, that the p-value is the probability of obtaining results (in our case — loss differentials) at least as large as the ones actually observed, assuming that the null hypothesis is correct.

[16] Actually covariance stationarity is sufficient but may not be strictly necessary [132].

[17] We assume that a day-ahead market has 24 prices. For markets with prices every half hour, the univariate variant comprises 48 independent tests.

5.5.2. The Giacomini–White test

In some of the more recent EPF studies [81,135,136], the DM test has been replaced by the Giacomini–White (GW) test [137] for *conditional predictive ability*. The latter is preferred because it can be regarded as a generalization of the DM test for *unconditional predictive ability*: while both tests can be used for nested and non-nested models,[18] only the GW test accounts for parameter estimation uncertainty through 'conditioning' [63].

Like the DM test, also the GW test has two variants in day-ahead EPF — the univariate and the multivariate. Without loss of generality, let us focus on the latter. It starts by building a multivariate loss differential series, see (13), for a pair of forecasts (of models A and B). Next, the test considers the following regression:

$$\Delta_d^{\mathrm{A,B}} = \boldsymbol{\phi}' X_{d-1} + \epsilon_d, \tag{14}$$

where X_{d-1} contains elements from the information set on day $d-1$, i.e. a constant and lags of $\Delta_d^{\mathrm{A,B}}$. Note that $\epsilon_d \neq \varepsilon_d^{\mathrm{Z}}$, i.e. ϵ_d is not the 24-dimensional vector of prediction errors for day d and model Z but simply an error term in the regression. Also note that using this notation the DM test can be written as [138]:

$$\Delta_d^{\mathrm{A,B}} = \mu + \epsilon_d, \tag{15}$$

i.e. with X_{d-1} containing just a constant. Finally, like for the DM test, to check the significance of differences in forecasting accuracy, the p-values of two one-sided tests can be computed. The interpretation and possible visualization (see Fig. 5) are analogous to that of the DM test.

5.6. Recalibration

An issue with many EPF studies is that forecasting models are not recalibrated. Instead, they are often estimated once using the training dataset and directly evaluated over the whole test dataset. This is problematic as it does not represent real-life conditions where forecasting models are retrained (often on a daily basis) to account for the latest market information.

To have models that are evaluated in realistic conditions, they need to be retrained considering the new incoming flow of market information. As an example, for the day-ahead market, a forecasting model should be retrained on a daily basis as new information is available. Considering a testing period of a year, this means that a realistic evaluation requires estimating the forecasting model 365 times.

5.7. Ex-ante hyperparameter optimization

A common issue in the current EPF literature is that the hyperparameter selection is often either done ex-post [49,51,139–142] or its details are not sufficiently explained [13,21,37,48,89,92,102–104, 107,110]. As an example, when models based on neural networks are proposed, the details on how the number of neurons are selected are usually not provided. In other cases, while the approach is provided, it is often based on analyzing different configurations of neurons using the test dataset and selecting the one that works best, i.e. ex-post hyperparameter selection.

Not providing enough details on how hyperparameters are selected is an obvious problem as it prevents reproducing research. Similarly, performing hyperparameter optimization ex-post leads to overfitting the test dataset, i.e. the model is partially optimized using the same dataset used for evaluating the model, and it grants the model an unfair and non-existent advantage over other models.

To prevent this, the selection of hyperparameters should be explicitly explained and always performed ex-ante using a validation dataset. With that motivation, for the open-access methods proposed, not only do we explain how the hyperparameters are obtained, but we also provide within the toolbox [58,59] a module for hyperparameter selection and the files containing the results of the hyperparameter optimization of the current study.

5.8. Computation time

An even more common problem is the fact that new models are very rarely compared in terms of their computational requirements [19,20, 22,24,32,37,41,42,51,100,102,103,105–108,111]. Although a model might be marginally better than another, it might not be worthwhile to deploy it in a practical application if its computational requirements are much larger. Particularly, higher computational requirements might pose two problems:

1. As mentioned before, forecasting models should ideally be recalibrated on a daily basis. Hence, a forecasting method is only suitable if its computational time allows this recalibration to take place. In this context, the maximum available time for estimating a model will depend on each electricity market but, as a rule of thumb, it can be argued that any model that requires more than 30 min or 1 h will unlikely be suitable for forecasting prices in the spot markets.
2. Besides recalibration, the second issue with computation time is its cost. If the computational capabilities are too large, the benefits of using a marginally better forecast might be lower than the cost of running the forecasting model on a much more expensive computer.

Hence, when new forecasting models are proposed, we argue that it is very important to provide their computation times. Moreover, we also argue that for a model to be better than the existing methods, it does not necessarily have to be the most accurate one. Instead:

1. If its computational time is large, i.e. in the order of minutes, the model should indeed be more accurate than all state-of-the-art models, e.g. DNNs.
2. If its computational time is small, i.e. in the order of seconds, the model should be more accurate than the state-of-the-art models with low computational requirements, e.g. LEAR.

In this article, we provide an analysis of the computational requirements of the proposed open-access models so other researchers can easily make such comparisons.

5.9. Reproducibility

Another related issue is that some studies lack enough details to replicate the research. Missing details vary from study to study but the four most common are:

1. the dataset used for testing and evaluation is not defined [31–37];
2. the dataset used for training is not defined [21,33,35,41,42];
3. the inputs of the model are unclear [35,36,38–40];
4. the selection of hyperparameters is unclear [13,21,37,48,89,92, 102–104,107,110].

To correct this, future EPF papers should provide enough details to allow replication and reviewers should verify that all necessary details of the employed datasets are always provided.

---

[18] This holds as long as the calibration window does not grow with the sample size [138]. This is satisfied for rolling windows, but not for extended calibration windows.

5.10. Data contamination

Another recurrent issue in the EPF literature is data contamination, which appears when a part of the training dataset is used for testing. Particularly, when working with time series data the test dataset should always comprise the last part of the dataset to avoid data contamination. If this is not done, the models can overfit the testing dataset and their accuracy can be overestimated.

Despite the importance of correctly separating the training/validation dataset from the testing dataset, some studies in EPF:

1. Do not specify the split between the training, validation, and test datasets [21,31–37,41,42]. If the datasets are not specified, it is not possible to know whether data contamination occurs.
2. Randomly sample the test dataset from the full dataset [143–146], e.g. in a dataset comprising a year of data randomly selecting 4 weeks for testing and the remaining data for training.
3. Have a partial or total overlap between the training/validation dataset and the testing dataset [51,139,140,147], e.g. by performing hyperparameter optimization ex-post.

To correct this issue, it is important that any future research in EPF ensures that: (1) the split between the datasets is correctly described; (2) the test dataset does never overlap with the training or validation datasets; (3) the test dataset is always selected as the last segment of the full dataset.

5.11. Software toolboxes

A less pressing yet relevant issue is the use of state-of-the-art software toolboxes. When comparing new methods with methods from the literature, the latter should be modeled using adequate toolboxes. Particularly, it is important to use toolboxes that are continuously updated as implementing methods using outdated libraries leads to unfair evaluations.

For example, in the context of neural networks, there are several open-source state-of-the-art toolboxes [126] that are continuously updated and that grant access to the latest development in the field of DL. Yet, in the context of EPF, new methods are often compared with neural networks that are modeled using the `MATLAB` toolbox [32,38,41,42,49,102,103,105,144], a toolbox that for many years was outdated and did not include many of the neural network developments that are critical in EPF, e.g. state-of-the-art activation functions or stochastic gradient descent algorithms [57]. As a result, many of the existing comparisons in EPF are based on evaluations where the accuracy of neural networks might be underestimated.

Besides using state-of-the-art software toolboxes, e.g. the `python` library `keras` for deep learning, it is also important to employ (whenever possible) free-to-access libraries so that research can be replicated by anyone.

5.12. Combining forecasts

As a final guideline, it is important to indicate the importance of ensembles in the context of EPF. In general, although exceptions exist [148], combining different models leads to a higher accuracy [81,134] and it is thus a good idea to build forecasts based on multiple models. However, as even the arithmetic average improves the accuracy of individual models, new ensemble techniques should be studied in comparison with other ensemble techniques, i.e. as done in [134], and not simply w.r.t. the individual models.

To maximize the forecasting accuracy, it is important to employ diverse forecasts [148], e.g. forecasts generated using different data or different models. For EPF, the former can be achieved by considering models trained using different calibration window lengths [80,135] and the latter using different modeling techniques or different sets of hyperparameters. To further maximize the performance, the number of models used in the ensemble should be limited [148], e.g. 4–10, especially in the case of heavy-tailed data for which large ensembles tend to contain outliers more often, resulting in less accurate forecasts.

With that in mind, as part of the open-access benchmark and toolbox [58,59], we also propose a series of simple ensemble techniques. Particularly, as explained in Section 4, we provide an ensemble of four LEAR models that are estimated over different calibration windows and combined using a simple arithmetic average and another ensemble using four DNNs that are estimated for different hyperparameters and combined using the arithmetic average.

6. Evaluation of state-of-the-art methods

In this section, we present the results of the open-source benchmark methods for all five datasets. For the sake of clarity, we divide the section into two parts respectively comprising the results for the error metrics and the results for statistical testing.

6.1. Accuracy metrics

We first start by presenting the results of the open-access benchmark models in terms of accuracy metrics.

6.1.1. Individual models

Table 2 compares the performance of the two individual models and their 4 variations in terms of rMAE, MAE, MAPE, SMAPE, and RMSE. The LEAR model is displayed for 4 different calibration windows representing 56, 84, 1092, and 1456 days, i.e. 8 weeks, 12 weeks, 3 years, and 4 years. The four DNNs are obtained by performing the hyperparameter/feature optimization process four times and using the best hyperparameter/feature selection of every run (see Sections 4.4 and 4.3.3 for further details).[19] Several observations can be made:

- The MAPE seems an unreliable metric as it completely disagrees with the other three linear metrics and the quadratic metric. In particular, while the rMAE, MAE, and sMAPE agree on what the best model is in all the cases, the MAPE almost never does so. This unreliability can be further seen in the German market: while the MAPE and sMAPE metrics usually have similar orders of magnitude, in the case of the German market the MAPE is approximately 10 times larger. This effect is due to prices in Germany being negative and very close to 0, leading in turn to very large absolute percentage errors that bias the MAPE.
- The DNN models seem to be more accurate than the LEAR models. Particularly, in terms of linear metrics, the best model across the five marketplaces is a DNN. Moreover, the majority of the DNN models perform better than the four LEAR models.
- Although the RMSE displays slightly different results, this is expected as the metric is based on quadratic errors and not linear ones. Nonetheless, it still shows the superiority of the DNN model: even though the DNN is estimated by minimizing absolute errors (unlike LEAR), the DNN is better in 3 of the 5 datasets. Moreover, even though the DNN seems to be worse in two markets, the RMSE metric does not correctly represent the underlying problem (see Sections 5.4 and 6.4.1) and it can be argued that it is not the best metric to assess the performance of EPF models.

---

[19] Note that, for the sake of simplicity, the features and hyperparameter selection for each model are not provided. However, they can be obtained from the website [58] accompanying this study.

Table 2

Comparison between the two individual state-of-the-art open-source methods in terms of rMAE, MAE, MAPE, sMAPE, and RMSE. Each of the two methods is listed for four different configurations. The gray cells represent the best model for a given metric.

		DNN_1	DNN_2	DNN_3	DNN_4	LEAR_{56}	LEAR_{84}	LEAR_{1092}	LEAR_{1456}
NP	rMAE	0.435	0.512	0.414	0.455	0.475	0.472	0.482	0.481
	MAE	1.797	2.118	1.712	1.883	1.964	1.952	1.993	1.990
	MAPE [%]	5.738	6.527	5.584	5.814	6.336	6.357	6.099	6.144
	sMAPE [%]	5.167	5.982	4.970	5.367	5.656	5.619	5.641	5.658
	RMSE	3.474	3.859	3.360	3.489	3.671	3.664	3.605	3.604
PJM	rMAE	0.511	0.499	0.491	0.486	0.550	0.548	0.490	0.489
	MAE	3.234	3.157	3.105	3.075	3.477	3.467	3.098	3.095
	MAPE [%]	30.622	29.345	27.554	27.975	32.520	32.341	30.279	30.239
	sMAPE [%]	12.518	12.212	12.271	12.004	13.677	13.576	12.331	12.538
	RMSE	6.231	6.773	5.200	5.498	5.718	5.709	5.264	5.142
EPEX BE	rMAE	0.620	0.621	0.620	0.597	0.682	0.669	0.649	0.653
	MAE	6.299	6.308	6.297	6.068	6.924	6.798	6.594	6.634
	MAPE [%]	24.650	27.710	27.578	25.466	32.878	32.343	26.256	22.645
	sMAPE [%]	14.543	14.723	14.980	14.106	16.197	15.954	16.867	17.293
	RMSE	16.360	16.666	16.115	15.950	16.371	16.291	16.458	16.420
EPEX FR	rMAE	0.575	0.573	0.554	0.591	0.638	0.624	0.580	0.597
	MAE	4.218	4.198	4.063	4.334	4.681	4.575	4.250	4.378
	MAPE [%]	14.284	13.757	15.160	15.513	19.031	18.087	14.955	14.896
	sMAPE [%]	12.124	11.698	11.488	12.176	13.427	13.281	13.250	14.054
	RMSE	11.772	12.345	11.880	12.354	11.732	10.759	11.337	11.462
EPEX DE	rMAE	0.407	0.422	0.406	0.394	0.506	0.499	0.450	0.451
	MAE	3.716	3.850	3.706	3.592	4.619	4.555	4.108	4.118
	MAPE [%]	77.145	137.449	100.214	90.578	129.763	133.580	128.295	124.191
	sMAPE [%]	14.970	15.356	15.508	14.680	17.600	17.491	16.984	17.054
	RMSE	6.796	7.304	6.271	6.080	8.122	7.923	6.996	6.987

Table 3

Comparison between the ensembles of the state-of-the-art open-source methods in terms of rMAE, MAE, MAPE, sMAPE, and RMSE. The comparison also includes, for each market, the best individual performing DNN and LEAR model in terms of rMAE and MAE, i.e. the two most reliable metrics. The gray cells represent the best model for a given metric.

		DNN Ensemble	LEAR Ensemble	Best[a] DNN	Best LEAR
NP	rMAE	0.407	0.420	0.414	0.472
	MAE	1.683	1.738	1.717	1.952
	MAPE [%]	5.384	5.533	5.584	6.357
	sMAPE [%]	4.880	5.009	4.970	5.619
	RMSE	3.319	3.362	3.360	3.664
PJM	rMAE	0.452	0.476	0.486	0.489
	MAE	2.862	3.013	3.075	3.095
	MAPE [%]	27.478	30.134	27.975	30.239
	sMAPE [%]	11.331	11.980	12.004	12.538
	RMSE	5.040	5.127	5.498	5.142
EPEX BE	rMAE	0.578	0.604	0.597	0.649
	MAE	5.870	6.140	6.068	6.594
	MAPE [%]	24.892	20.720	25.466	26.256
	sMAPE [%]	13.446	14.546	14.106	16.867
	RMSE	15.966	15.974	15.950	16.458
EPEX FR	rMAE	0.527	0.543	0.554	0.580
	MAE	3.866	3.980	4.063	4.250
	MAPE [%]	13.601	14.680	15.160	14.955
	sMAPE [%]	10.812	11.566	11.488	13.250
	RMSE	11.867	10.676	11.880	11.337
EPEX DE	rMAE	0.374	0.433	0.394	0.450
	MAE	3.413	3.955	3.592	4.108
	MAPE [%]	94.434	122.412	90.578	128.295
	sMAPE [%]	14.078	15.747	14.680	16.984
	RMSE	5.927	7.079	6.080	6.996

[a] Best in terms of rMAE/MAE.

6.1.2. Ensembles

The results for the ensemble methods are listed in Table 3, which compares the performance of the two ensemble models and the best DNN and LEAR models in terms of the rMAE metric, i.e. arguably the most reliable metric. From the table, several observations can be made:

- As already argued in Section 5.12, combining models usually improves the accuracy. Particularly, the ensemble of DNNs is better than the best individual DNN model for all five markets and for all reliable metrics. Similarly, the ensemble of LEAR models is better than the best individual LEAR model for all markets and reliable metrics. The exception to this observation are the MAPE and RMSE metrics but, as already noted, MAPE is an unreliable metric and RMSE does not correctly represent the underlying problem of EPF.
- As before, in terms of rMAE, the ensemble of DNNs is the most accurate model across all markets, which again seems to suggest that the DNN models are more accurate than the LEAR models.

6.2. Statistical testing

In this section, we present the results of the open-access benchmark models in terms of the statistical tests. For the sake of simplicity, we

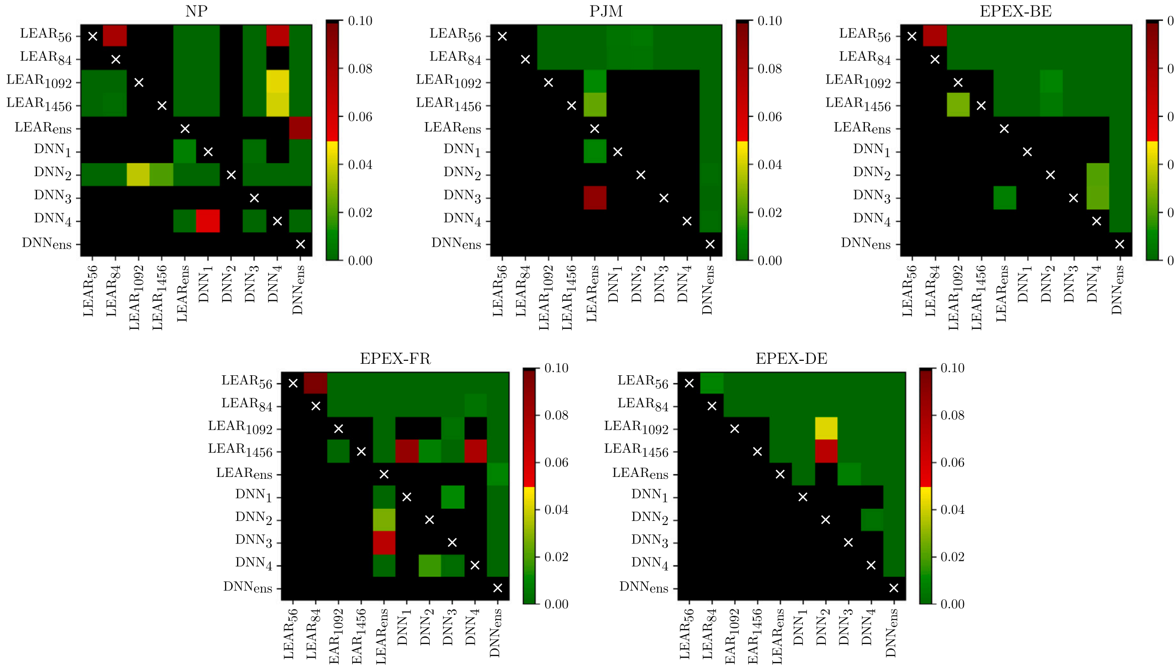

Fig. 5. Results of the GW test with the multivariate loss differential series (16) for the eight individual models and the two ensembles. A heat map is used to indicate the range of the obtained p-values for each of the five markets. The closer the p-values are to zero (dark green), the more significant the difference is between the forecasts of a model on the X-axis (better) and the forecasts of a model on the Y-axis (worse). Black color indicates p-values above the color map limit, i.e. p-values larger or equal than 0.10.

present together the results for individual methods and ensembles. The results are based on the multivariate GW test using the $L1$ norm in (13), i.e. with the following loss differential series:

$$\Delta_d^{\mathrm{A,B}} = \sum_{h=1}^{24} |\varepsilon_{d,h}^{\mathrm{A}}| - \sum_{h=1}^{24} |\varepsilon_{d,h}^{\mathrm{B}}|. \tag{16}$$

While squared losses could also be used, we do not consider them here because absolute errors better represent the underlying problem in EPF, see Section 6.4.1 for a discussion.

In Fig. 5 we display the results for the five markets. More precisely, we use heat maps arranged as chessboards to indicate the range of the obtained p-values. The closer they are to zero (dark green) the more significant is the difference between the forecasts of a model on the X-axis (better) and the forecasts of a model on the Y-axis (worse). For instance, for the EPEX-DE market the first row is green indicating that the forecasts of LEAR$_{56}$ are significantly outperformed by those of all other models. We can observe that:

- For all markets the last column is green indicating that the forecasts of the ensemble of DNNs are statistically significantly better than the predictions of all the other models for all 5 datasets. The only exception is the LEAR ensemble and the NP market, a scenario in which the difference in forecasting accuracy is not statistically significant.
- The forecasts of LEAR$_{ens}$ are statistically significantly better than those of all individual LEAR models. Together with the previous observation, i.e. the superiority of the DNN ensemble, it shows that the predictions of ensemble models usually improve upon the forecasting accuracy of individual methods.
- In two datasets (EPEX-BE and EPEX-DE), the forecasts of all the individual DNN methods are statistically significantly better than those of the individual LEAR models. In the EPEX-FR dataset, the forecasts of all the individual DNN methods are statistically significantly better than 3 out of the 4 individual LEAR models. For PJM, there are 2 DNN models whose forecasts are statistically significantly better than those of all LEAR models.
- Aside from the poor-performing DNN$_2$ model for the NP dataset, the forecasts of the individual LEAR models are never significantly better than those of the individual DNN models. Overall, it seems that forecasts based on DNNs are more likely to obtain significantly better results; this is particularly true for the DNN ensemble.

6.3. *Computation time*

As described in Section 5.8, besides comparing the predictive accuracy, it is also necessary to analyze the computation time of the forecasting methods. Table 4 lists a comparison of the computation time required for estimating the models considered, i.e. the time required to recalibrate each model on a daily basis. As the computation time is non-deterministic, its value is given as a range. These data were obtained using a regular laptop quadcore CPU, i.e. the i7-6920HQ.

As can be observed, although the LEAR model performs slightly worse than the DNN model, its computation time is 30 to 100 times lower; particularly, when considering the maximum computation time of both methods, the LEAR model is 50 times faster.

6.4. *Discussion and remarks*

In this section, we provide some final remarks behind the motivation of the metrics employed, we briefly analyze the influence of the different metrics considered, and provide a discussion on comparing new models.

6.4.1. *Absolute vs. squared errors*

Throughout the text, we have mostly considered accuracy metrics based on absolute/linear errors, i.e. metrics that evaluate the accuracy of predicting the median of the distribution. Since the LEAR model is estimated by minimizing squared errors, thus leading to forecasts of the mean [129], one could argue that a metric/test based on squared errors should be preferred. While the argument has some merits, we focused on absolute metrics for three reasons:

Table 4
Computation time that each benchmark model requires to perform a daily recalibration.

	Time
LEAR	1–10 s
LEAR ensemble	20–25 s
DNN	2–5 min
DNN ensemble	8–20 min

1. The metric used to evaluate the accuracy should be the one that better represents the underlying problem. In the case of EPF, since the cost of purchasing electricity is linear, linear metrics are arguably the best to quantify the risk associated with forecasting errors.
2. While we provided the RMSE results, they are qualitatively the same as for MAE/rMAE. Hence, as absolute errors better represent the underlying problem of EPF and the results are similar, the RMSE results are not analyzed here in detail due to space limitations.
3. While the LEAR model is indeed estimated using squared errors, this is partly done because the techniques to efficiently estimate the LASSO, e.g. coordinate descent, are based on square errors. This gives the LEAR model a computational advantage over the DNN. An alternative would be to use regularized quantile regression [149] leading, however, to an increased computational burden with little benefits on the accuracy in terms of MAE/rMAE.

6.4.2. Metrics

The obtained results validate the general guidelines proposed in Section 5.4 regarding accuracy metrics: research in EPF should avoid MAPE and only use metrics like sMAPE or RMSE in conjunction with any version of rMAE. Particularly, the results validate the following four claims:

1. MAE is as reliable as rMAE. However, as the errors are not relative, comparison between datasets is not possible and rMAE is preferred.
2. sMAPE is more reliable than MAPE and it agrees with MAE/rMAE. Yet, it has the problem of an undefined mean and an infinite variance. Thus, it is less reliable than rMAE.
3. MAPE is not a reliable metric as it gives more importance to datapoints close to zero. As such, using MAPE can lead to misleading results and wrong conclusions.
4. RMSE is more reliable than MAPE but it does not represent correctly the underlying risks of EPF. Hence, it should not be used alone to evaluate forecasting models.

6.4.3. Performance of open-access models

Based on the extensive comparison of Sections 6.1–6.3, it can be concluded that the models based on DL are more likely to outperform those based on statistical methods. This is especially true in the context of DL ensemble models as the ensemble of DNNs obtains results that are statistically significantly better than any other model.

However, while DNNs outperformed the LEAR models, the latter are still the state-of-the-art in terms of low complexity and computational cost. In particular, their performance is very close to that of DNNs, but with the advantage of having computational costs that are up to 100 times lower. As such, they are the best available option when decision making has to be done within seconds.

In short, new models for EPF should either be compared against LEAR models or DNNs depending on the decision time that is available. For a method to be considered more accurate than state-of-the-art methods, it should either be more accurate than the DNN model, or more accurate than LEAR but with similar or lower computational requirements.

7. Checklist to ensure adequate EPF research

As a final contribution, and with the goal of facilitating the work of reviewers of future EPF publications, we provide a short checklist to evaluate whether any new research in EPF satisfies the requirements to be reproducible and lead to meaningful conclusions:

1. The test dataset comprises at least a year of data.
2. Any new model is tested against state-of-the-art open-access models, e.g. the ones provided here.
3. The computational cost of new methods is evaluated and compared against the computational cost of existing methods.
4. The employed datasets are open-access.
5. The study is based on multiple markets.
6. rMAE is employed as one of the accuracy metrics to evaluate forecasting accuracy.
7. Statistical testing is used to assess whether differences in performance are significant.
8. Forecasting models are recalibrated on a daily basis and not simply estimated once and evaluated in the full out-of-sample dataset.
9. Hyperparameters are estimated using a validation dataset that is different from the test dataset.
10. The split and dates of the dataset are explicitly stated.
11. All the inputs of the model are explicitly defined.
12. The test dataset is selected as the last section of the full dataset and does not contain any overlapping data with the training or validation datasets.
13. State-of-the-art and free toolboxes are used for modeling the benchmark models.

While this is just a very short summary of the guidelines described in Section 5, we think it is very useful to have them summarized together for quick evaluations of new research.

8. Conclusion

In this paper, we have derived a set of best practices for performing research in *electricity price forecasting* (EPF). Particularly, as the field of EPF lacks a rigorous approach to compare and to evaluate new forecasting models, we have analyzed different factors affecting the quality of the research, e.g. dataset size or accuracy metrics, and we have proposed solutions to ensure that new research is adequate, reproducible, and useful.

In addition, as comparisons in EPF are often done using unique datasets that no other researchers have access to, we have proposed an extensive open-access benchmark dataset comprising 6 years of recent data in 5 different markets. The aim of the benchmark dataset is to provide a common framework for future research so that new methods can be validated under the same conditions and meaningful comparisons can be obtained. To facilitate future research, we have developed an open-source `python` library named `epftoolbox` [58,59] that provides easy access to these datasets.

Similarly, as new methods in EPF are often not compared with well-established methods, we have proposed several state-of-the-art open-source models based on statistical methods and deep learning. The methods are tuned automatically and require no expert knowledge in order to be used. These methods are provided as open-source within the proposed `epftoolbox` library [58,59] so that other researchers can employ them as benchmarks in their own studies. Although the proposed methods are currently developed in `python`, we would like to extend the support to other languages; in that spirit, we encourage other researchers to help us do so.

Finally, to have a complete open-access benchmark, we have evaluated the two proposed open-access methods in the open-access dataset and we have provided the results in terms of accuracy metrics and statistical testing. Using these results, we have shown that deep neural networks are more likely to outperform LEAR methods but that the latter are the best model for applications with short decision time-frames. Moreover, we have also shown that ensemble methods often obtain significantly better results than their individual counterparts. Based on the same results, we have also showed the importance of the guidelines as to what constitutes good practices for the rigorous use of models, metrics, and statistical tests in EPF research. The most notable guidelines were that MAPE is an unreliable metric that should be avoided, that statistical testing is mandatory to obtain meaningful conclusions, and that the length of the test dataset should be at least one year.

CRediT authorship contribution statement

Jesus Lago: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing - original draft, Writing - review & editing, Visualization, Project administration. **Grzegorz Marcjasz:** Methodology, Resources, Validation, Formal analysis, Investigation, Resources, Data curation, Writing - review & editing, Visualization. **Bart De Schutter:** Supervision, Writing - review & editing, Funding acquisition. **Rafał Weron:** Conceptualization, Investigation, Resources, Supervision, Writing - review & editing, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 675318 (INCITE), the Ministry of Science and Higher Education (MNiSW, Poland) through grant No. 0219/DIA/2019/48 and the National Science Center (NCN, Poland) through grant No. 2018/30/A/HS4/00444.

References

[1] Brancucci Martinez-Anido C, Brinkman G, Hodge B-M. The impact of wind power on electricity prices. Renew Energy 2016;94:474–87. http://dx.doi.org/10.1016/j.renene.2016.03.053.

[2] Gianfreda A, Parisio L, Pelagatti M. The impact of RES in the Italian day-ahead and balancing markets. Energy J 2016;37:161–84. http://dx.doi.org/10.5547/01956574.37.si2.agia.

[3] Grossi L, Nan F. Robust forecasting of electricity prices: Simulations, models and the impact of renewable sources. Technol Forecast Soc Change 2019;141:305–18. http://dx.doi.org/10.1016/j.techfore.2019.01.006.

[4] Maciejowska K. Assessing the impact of renewable energy sources on the electricity price level and variability – A quantile regression approach. Energy Econ 2020;85:104532. http://dx.doi.org/10.1016/j.eneco.2019.104532.

[5] 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. http://dx.doi.org/10.1016/j.ijforecast.2014.08.008.

[6] Nowotarski J, Weron R. Recent advances in electricity price forecasting: A review of probabilistic forecasting. Renew Sustain Energy Rev 2018;81(1):1548–68. http://dx.doi.org/10.1016/j.rser.2017.05.234.

[7] Ziel F, Steinert R. Probabilistic mid- and long-term electricity price forecasting. Renew Sustain Energy Rev 2018;94:251–66. http://dx.doi.org/10.1016/j.rser.2018.05.038.

[8] Hong T, Pinson P, Wang Y, Weron R, Yang D, Zareipour H. Energy forecasting: A review and outlook. IEEE Open Access J Power Energy 2020;7:376–88.

[9] Uniejewski B, Weron R, Ziel F. Variance stabilizing transformations for electricity spot price forecasting. IEEE Trans Power Syst 2018;33(2):2219–29. http://dx.doi.org/10.1109/tpwrs.2017.2734563.

[10] Marcjasz G, Uniejewski B, Weron R. On the importance of the long-term seasonal component in day-ahead electricity price forecasting with NARX neural networks. Int J Forecast 2019;35(4):1520–32. http://dx.doi.org/10.1016/j.ijforecast.2017.11.009.

[11] Cruz A, Muñoz A, Zamora J, Espínola R. The effect of wind generation and weekday on Spanish electricity spot price forecasting. Electr Power Syst Res 2011;81(10):1924–35. http://dx.doi.org/10.1016/j.epsr.2011.06.002.

[12] Wang L, Zhang Z, Chen J. Short-term electricity price forecasting with stacked denoising autoencoders. IEEE Trans Power Syst 2017;32(4):2673–81. http://dx.doi.org/10.1109/TPWRS.2016.2628873.

[13] Ugurlu U, Oksuz I, Tas O. Electricity price forecasting using recurrent neural networks. Energies 2018;11(5):1255. http://dx.doi.org/10.3390/en11051255.

[14] Zhang W, Cheema F, Srinivasan D. Forecasting of electricity prices using deep learning networks. In: Proceedings of the 2018 IEEE PES Asia-Pacific Power and Energy Engineering Conference. 2018, p. 451–6. http://dx.doi.org/10.1109/APPEEC.2018.8566313.

[15] Luo S, Weng Y. A two-stage supervised learning approach for electricity price forecasting by leveraging different data sources. Appl Energy 2019;242:1497–512. http://dx.doi.org/10.1016/j.apenergy.2019.03.129.

[16] Chen Y, Wang Y, Ma J, Jin Q. BRIM: An accurate electricity spot price prediction scheme-based bidirectional recurrent neural network and integrated market. Energies 2019;12(12):2241. http://dx.doi.org/10.3390/en12122241.

[17] Chang Z, Zhang Y, Chen W. Electricity price prediction based on hybrid model of Adam optimized LSTM neural network and wavelet transform. Energy 2019;187:115804. http://dx.doi.org/10.1016/j.energy.2019.07.134.

[18] Gao W, Darvishan A, Toghani M, Mohammadi M, Abedinia O, Ghadimi N. Different states of multi-block based forecast engine for price and load prediction. Int J Electr Power Energy Syst 2019;104:423–35. http://dx.doi.org/10.1016/j.ijepes.2018.07.014.

[19] Nazar MS, Fard AE, Heidari A, Shafie-khah M, Catalão JP. Hybrid model using three-stage algorithm for simultaneous load and price forecasting. Electr Power Syst Res 2018;165:214–28. http://dx.doi.org/10.1016/j.epsr.2018.09.004.

[20] Zhou L, Wang B, Wang Z, Wang F, Yang M. Seasonal classification and RBF adaptive weight based parallel combined method for day-ahead electricity price forecasting. In: Proceedings of the 2018 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference. 2018, p. 1–5. http://dx.doi.org/10.1109/isgt.2018.8403372.

[21] Singh N, Hussain S, Tiwari S. A PSO-based ANN model for short-term electricity price forecasting. In: Advances in Intelligent Systems and Computing. 2018, p. 553–63. http://dx.doi.org/10.1007/978-981-10-7386-1_47.

[22] Yang Z, Ce L, Lian L. Electricity price forecasting by a hybrid model, combining wavelet transform, ARMA and kernel-based extreme learning machine methods. Appl Energy 2017;190:291–305. http://dx.doi.org/10.1016/j.apenergy.2016.12.130.

[23] Chinnathambi RA, Plathottam SJ, Hossen T, Nair AS, Ranganathan P. Deep neural networks (DNN) for day-ahead electricity price markets. In: Proceedings of the 2018 IEEE Electrical Power and Energy Conference. 2018, p. 1–6. http://dx.doi.org/10.1109/epec.2018.8598327.

[24] Olamaee J, Mohammadi M, Noruzi A, Hosseini SMH. Day-ahead price forecasting based on hybrid prediction model. Complexity 2016;21(S2):156–64. http://dx.doi.org/10.1002/cplx.21792.

[25] Darudi A, Javidi MH, Bashari M. Electricity price forecasting using a new data fusion algorithm. IET Gener Transm Distrib 2015;9(12):1382–90. http://dx.doi.org/10.1049/iet-gtd.2014.0653.

[26] Ghayekhloo M, Azimi R, Ghofrani M, Menhaj M, Shekari E. A combination approach based on a novel data clustering method and Bayesian recurrent neural network for day-ahead price forecasting of electricity markets. Electr Power Syst Res 2019;168:184–99. http://dx.doi.org/10.1016/j.epsr.2018.11.021.

[27] Victoire AA, Gobu B, Jaikumar S, Arulmozhi N, Kanimozhi P, Victoire A. Two-stage machine learning framework for simultaneous forecasting of price-load in the smart grid. In: Proceedings of the 2018 IEEE International Conference on Machine Learning and Applications. 2018, p. 1081–6. http://dx.doi.org/10.1109/icmla.2018.00176.

[28] Zahid M, Ahmed F, Javaid N, Abbasi R, Zainab Kazmi H, Javaid A, et al. Electricity price and load forecasting using enhanced convolutional neural network and enhanced support vector regression in smart grids. Electronics 2019;8(2):122. http://dx.doi.org/10.3390/electronics8020122.

[29] Jiang L, Hu G. Day-ahead price forecasting for electricity market using long-short term memory recurrent neural network. In: Proceedings of the 2018 International Conference on Control, Automation, Robotics and Vision. 2018, p. 949–54. http://dx.doi.org/10.1109/icarcv.2018.8581235.

[30] Zhou S, Zhou L, Mao M, Tai H, Wan Y. An optimized heterogeneous structure LSTM network for electricity price forecasting. IEEE Access 2019;7:108161–73. http://dx.doi.org/10.1109/ACCESS.2019.2932999.

[31] Aggarwal A, Tripathi MM. A novel hybrid approach using wavelet transform, time series time delay neural network, and error predicting algorithm for day-ahead electricity price forecasting. In: Proceedings of the International Conference on Computer Applications in Electrical Engineering-Recent Advances. 2017, p. 199–204. http://dx.doi.org/10.1109/cera.2017.8343326.

[32] Hong Y-Y, Liu C-Y, Chen S-J, Huang W-C, Yu T-H. Short-term LMP forecasting using an artificial neural network incorporating empirical mode decomposition. Int Trans Electr Energy Syst 2014;25(9):1952–64. http://dx.doi.org/10.1002/etep.1949.

[33] Talari S, Shafie-khah M, Osório G, Wang F, Heidari A, Catalão J. Price forecasting of electricity markets in the presence of a high penetration of wind power generators. Sustainability 2017;9(11):2065. http://dx.doi.org/10.3390/su9112065.

[34] Singh N, Mohanty SR, Shukla RD. Short term electricity price forecast based on environmentally adapted generalized neuron. Energy 2017;125:127–39. http://dx.doi.org/10.1016/j.energy.2017.02.094.

[35] Khan GM, Arshad R, Khan NM. Efficient prediction of dynamic tariff in smart grid using CGP evolved artificial neural networks. In: Proceedings of the 2017 IEEE International Conference on Machine Learning and Applications. 2017, p. 493–8. http://dx.doi.org/10.1109/icmla.2017.0-113.

[36] Afrasiabi M, Mohammadi M, Rastegar M, Kargarian A. Multi-agent microgrid energy management based on deep learning forecaster. Energy 2019;186:115873. http://dx.doi.org/10.1016/j.energy.2019.115873.

[37] Zhu Y, Dai R, Liu G, Wang Z, Lu S. Power market price forecasting via deep learning. In: Proceedings of the 44th Annual Conference of the IEEE Industrial Electronics Society. 2018, http://dx.doi.org/10.1109/iecon.2018.8591581.

[38] Wang D, Luo H, Grunder O, Lin Y, Guo H. Multi-step ahead electricity price forecasting using a hybrid model based on two-layer decomposition technique and BP neural network optimized by firefly algorithm. Appl Energy 2017;190:390–407. http://dx.doi.org/10.1016/j.apenergy.2016.12.134.

[39] Shrivastava NA, Panigrahi BK, Lim M-H. Electricity price classification using extreme learning machines. Neural Comput Appl 2014;27(1):9–18. http://dx.doi.org/10.1007/s00521-013-1537-1.

[40] Jiang P, Ma X, Liu F. A new hybrid model based on data preprocessing and an intelligent optimization algorithm for electrical power system forecasting. Math Probl Eng 2015;2015:1–17. http://dx.doi.org/10.1155/2015/815253.

[41] Bento P, Pombo J, Calado M, Mariano S. A bat optimized neural network and wavelet transform approach for short-term price forecasting. Appl Energy 2018;210:88–97. http://dx.doi.org/10.1016/j.apenergy.2017.10.058.

[42] Khajeh MG, Maleki A, Rosen MA, Ahmadi MH. Electricity price forecasting using neural networks with an improved iterative training algorithm. Int J Ambient Energy 2017;39(2):147–58. http://dx.doi.org/10.1080/01430750.2016.1269674.

[43] Lago J, De Ridder F, Vrancx P, De Schutter B. Forecasting day-ahead electricity prices in Europe: The importance of considering market integration. Appl Energy 2018;211:890–903. http://dx.doi.org/10.1016/j.apenergy.2017.11.098.

[44] Kuo P-H, Huang C-J. An electricity price forecasting model by hybrid structured deep neural networks. Sustainability 2018;10(4):1280. http://dx.doi.org/10.3390/su10041280.

[45] Mujeeb S, Javaid N, Ilahi M, Wadud Z, Ishmanov F, Afzal M. Deep long short-term memory: A new price and load forecasting scheme for big data in smart cities. Sustainability 2019;11(4):987. http://dx.doi.org/10.3390/su11040987.

[46] Atef S, Eltawil AB. A comparative study using deep learning and support vector regression for electricity price forecasting in smart grids. In: Proceedings of the 2019 IEEE International Conference on Industrial Engineering and Applications. 2019, p. 603–7. http://dx.doi.org/10.1109/IEA.2019.8715213.

[47] Mujeeb S, Javaid N. ESAENARX and DE-RELM: Novel schemes for big data predictive analytics of electricity load and price. Sustainable Cities Soc 2019;51:101642. http://dx.doi.org/10.1016/j.scs.2019.101642.

[48] Lahmiri S. Comparing variational and empirical mode decomposition in forecasting day-ahead energy prices. IEEE Syst J 2017;11(3):1907–10. http://dx.doi.org/10.1109/jsyst.2015.2487339.

[49] Peter SE, Raglend IJ. Sequential wavelet-ANN with embedded ANN-PSO hybrid electricity price forecasting model for Indian energy exchange. Neural Comput Appl 2016;28(8):2277–92. http://dx.doi.org/10.1007/s00521-015-2141-3.

[50] Naz A, Javed M, Javaid N, Saba T, Alhussein M, Aurangzeb K. Short-term electric load and price forecasting using enhanced extreme learning machine optimization in smart grids. Energies 2019;12(5):866. http://dx.doi.org/10.3390/en12050866.

[51] Anamika, Peesapati R, Kumar N. Electricity price forecasting and classification through wavelet–dynamic weighted PSO–FFNN approach. IEEE Syst J 2018;12(4):3075–84. http://dx.doi.org/10.1109/jsyst.2017.2717446.

[52] Gao W, Sarlak V, Parsaei MR, Ferdosi M. Combination of fuzzy based on a meta-heuristic algorithm to predict electricity price in an electricity markets. Chem Eng Res Des 2018;131:333–45. http://dx.doi.org/10.1016/j.cherd.2017.09.021.

[53] 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. http://dx.doi.org/10.1016/j.ijforecast.2016.02.001.

[54] Nord Pool website. URL www.nordpoolspot.com.

[55] Uniejewski B, Nowotarski J, Weron R. Automated variable selection and shrinkage for day-ahead electricity price forecasting. Energies 2016;9(8):621. http://dx.doi.org/10.3390/en9080621.

[56] Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B Stat Methodol 1996;267–88. http://dx.doi.org/10.1111/j.2517-6161.1996.tb02080.x.

[57] Lago J, De Ridder F, De Schutter B. Forecasting spot electricity prices: deep learning approaches and empirical comparison of traditional algorithms. Appl Energy 2018;221:386–405. http://dx.doi.org/10.1016/j.apenergy.2018.02.069.

[58] Epftoolbox library. URL https://github.com/jeslago/epftoolbox.

[59] Epftoolbox documentation. URL https://epftoolbox.readthedocs.io.

[60] Mayer K, Trück S. Electricity markets around the world. J Commodity Mark 2018;9:77–100. http://dx.doi.org/10.1016/j.jcomm.2018.02.001.

[61] Aïd R. Electricity derivatives. Springer; 2015, http://dx.doi.org/10.1007/978-3-319-08395-7.

[62] Maciejowska K, Weron R. Electricity price forecasting. In: Wiley StatsRef: Statistics reference online. Wiley; 2019, p. 1–9. http://dx.doi.org/10.1002/9781118445112.stat08215.

[63] Weron R, Ziel F. Electricity price forecasting. In: Soytas U, Sari R, editors. Routledge handbook of energy economics. Routledge; 2018, p. 506–21. http://dx.doi.org/10.4324/9781315459653-36.

[64] Ziel F, Weron R. Day-ahead electricity price forecasting with high-dimensional structures: Univariate vs. multivariate modeling frameworks. Energy Econ 2018;70:396–420. http://dx.doi.org/10.1016/j.eneco.2017.12.016.

[65] Gianfreda A, Ravazzolo F, Rossini L. Comparing the forecasting performances of linear models for electricity prices with high RES penetration. Int J Forecast 2020;36:974–86. http://dx.doi.org/10.1016/j.ijforecast.2019.11.002.

[66] Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B Stat Methodol 2005;67(2):301–20. http://dx.doi.org/10.1111/j.1467-9868.2005.00503.x.

[67] Ziel F, Steinert R, Husmann S. Forecasting day ahead electricity spot prices: The impact of the EXAA to other European electricity markets. Energy Econ 2015;51:430–44. http://dx.doi.org/10.1016/j.eneco.2015.08.005.

[68] Ziel F. Forecasting electricity spot prices using lasso: On capturing the autoregressive intraday structure. IEEE Trans Power Syst 2016;31(6):4977–87. http://dx.doi.org/10.1109/tpwrs.2016.2521545.

[69] Uniejewski B, Weron R. Efficient forecasting of electricity spot prices with expert and LASSO models. Energies 2018;11(8):2039. http://dx.doi.org/10.3390/en11082039.

[70] Uniejewski B, Marcjasz G, Weron R. Understanding intraday electricity markets: Variable selection and very short-term price forecasting using LASSO. Int J Forecast 2019;35(4):1533–47.

[71] Narajewski M, Ziel F. Ensemble forecasting for intraday electricity prices: Simulating trajectories. Appl Energy 2020;279:115801.

[72] Muniain P, Ziel F. Probabilistic forecasting in day-ahead electricity markets: Simulating peak and off-peak prices. Int J Forecast 2020;36(4):1193–210.

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

[74] Schneider S. Power spot price models with negative prices. J Energy Mark 2011;4(4):77–102. http://dx.doi.org/10.21314/jem.2011.079.

[75] Diaz G, Planas E. A note on the normalization of Spanish electricity spot prices. IEEE Trans Power Syst 2016;31(3):2499–500. http://dx.doi.org/10.1109/tpwrs.2015.2449757.

[76] 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. http://dx.doi.org/10.1016/j.eneco.2013.04.004.

[77] 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. http://dx.doi.org/10.1016/j.eneco.2016.05.009.

[78] Lisi F, Pelagatti M. Component estimation for electricity market data: Deterministic or stochastic? Energy Econ 2018;74:13–37. http://dx.doi.org/10.1016/j.eneco.2018.05.027.

[79] Marcjasz G, Uniejewski B, Weron R. On the importance of the long-term seasonal component in day-ahead electricity price forecasting with NARX neural networks. Int J Forecast 2019;35(4):1520–32. http://dx.doi.org/10.1016/j.ijforecast.2017.11.009.

[80] Hubicka K, Marcjasz G, Weron R. A note on averaging day-ahead electricity price forecasts across calibration windows. IEEE Trans Sustain Energy 2019;10(1):321–3. http://dx.doi.org/10.1109/tste.2018.2869557.

[81] Marcjasz G, Serafin T, Weron R. Selection of calibration windows for day-ahead electricity price forecasting. Energies 2018;11(9):2364. http://dx.doi.org/10.3390/en11092364.

[82] Maciejowska K, Uniejewski B, Serafin T. PCA forecast averaging – Predicting day-ahead and intraday electricity prices. Energies 2020;13(14):3530.

[83] Pesaran M, Timmermann A. Selection of estimation window in the presence of breaks. J Econometrics 2007;137(1):134–61.

[84] De Marcos R, Bunn D, Bello A, Reneses J. Short-term electricity price forecasting with recurrent regimes and structural breaks. Energies 2020;13(20):5452.

[85] Nitka W, Serafin T, Sotiros D. Forecasting electricity prices: Autoregressive hybrid nearest neighbors (ARHNN) method. In: ICCS 2021. In: Lecture Notes in Computer Science. 2021 [forthcoming].

[86] Westerlund J, Narayan P. Testing for predictability in conditionally heteroskedastic stock returns. J Financ Econ 2015;13(2):342–75.

[87] Karakatsani N, Bunn D. Fundamental and behavioural drivers of electricity price volatility. Stud Nonlinear Dyn Econom 2010;14(4):4.

[88] Mujeeb S, Javaid N, Akbar M, Khalid R, Nazeer O, Khan M. Big data analytics for price and load forecasting in smart grids. Lecture Notes on Data Engineering and Communications Technologies, Springer; 2018, p. 77–87. http://dx.doi.org/10.1007/978-3-030-02613-4_7.

[89] Xie X, Xu W, Tan H. The day-ahead electricity price forecasting based on stacked CNN and LSTM. Lecture Notes in Computer Science, Springer International Publishing; 2018, p. 216–30. http://dx.doi.org/10.1007/978-3-030-02698-1_19.

[90] Ugurlu U, Tas O, Kaya A, Oksuz I. The financial effect of the electricity price forecasts' inaccuracy on a hydro-based generation company. Energies 2018;11(8):2093. http://dx.doi.org/10.3390/en11082093.

[91] Kolberg JK, Waage K. Artificial Intelligence and Nord Pool's intraday electricity market Elbas: a demonstration and pragmatic evaluation of employing deep learning for price prediction: using extensive market data and spatio-temporal weather forecasts [Master's thesis], Norwegian School of Economics; 2018.

[92] Xu J, Baldick R. Day-ahead price forecasting in ERCOT market using neural network approaches. In: Proceedings of the tenth ACM International Conference on Future Energy Systems. 2019, p. 486–91. http://dx.doi.org/10.1145/3307772.3331024.

[93] Meier J-H, Schneider S, Schmidt I, Schüller P, Schönfeldt T, Wanke B. ANN-based electricity price forecasting under special consideration of time series properties. In: Information and Communication Technologies in Education, Research, and Industrial Applications. Springer International Publishing; 2019, p. 262–75. http://dx.doi.org/10.1007/978-3-030-13929-2_13.

[94] Chang Z, Zhang Y, Chen W. Effective Adam-optimized LSTM neural network for electricity price forecasting. In: Proceedings of the 2018 IEEE International Conference on Software Engineering and Service Science. 2018, p. 245–8. http://dx.doi.org/10.1109/icsess.2018.8663710.

[95] Jahangir H, Tayarani H, Baghali S, Ahmadian A, Elkamel A, Aliakbar Golkar M, et al. A novel electricity price forecasting approach based on dimension reduction strategy and rough artificial neural networks. IEEE Trans Ind Inf 2019;16(4):2369–81. http://dx.doi.org/10.1109/TII.2019.2933009.

[96] Ahmad W, Javaid N, Chand A, Shah SYR, Yasin U, Khan M, et al. Electricity price forecasting in smart grid: A novel E-CNN model. In: Web, Artificial Intelligence and Network Applications. Springer International Publishing; 2019, p. 1132–44. http://dx.doi.org/10.1007/978-3-030-15035-8_109.

[97] Aineto D, Iranzo-Sánchez J, Lemus-Zúñiga LG, Onaindia E, Urchueguía JF. On the influence of renewable energy sources in electricity price forecasting in the Iberian market. Energies 2019;12(11):2082. http://dx.doi.org/10.3390/en12112082.

[98] Schnürch S, Wagner A. Machine learning on EPEX order books: Insights and forecasts. 2019, arXiv preprint arXiv:1906.06248.

[99] Zhang J, Tan Z, Li C. A novel hybrid forecasting method using GRNN combined with wavelet transform and a GARCH model. Energy Sources B: Econ Plann Policy 2015;10(4):418–26. http://dx.doi.org/10.1080/15567249.2011.557685.

[100] Zhang J-L, Zhang Y-J, Li D-Z, Tan Z-F, Ji J-F. Forecasting day-ahead electricity prices using a new integrated model. Int J Electr Power Energy Syst 2019;105:541–8. http://dx.doi.org/10.1016/j.ijepes.2018.08.025.

[101] Kurbatsky V, Sidorov D, Spiryaev V, Tomin N. Forecasting nonstationary time series based on Hilbert–Huang transform and machine learning. Autom Remote Control 2014;75(5):922–34.

[102] Varshney H, Sharma A, Kumar R. A hybrid approach to price forecasting incorporating exogenous variables for a day ahead electricity market. In: Proceedings of the 2016 IEEE International Conference on Power Electronics, Intelligent Control and Energy Systems. 2016, p. 1–6. http://dx.doi.org/10.1109/icpeices.2016.7853355.

[103] Xiao L, Shao W, Yu M, Ma J, Jin C. Research and application of a hybrid wavelet neural network model with the improved cuckoo search algorithm for electrical power system forecasting. Appl Energy 2017;198:203–22. http://dx.doi.org/10.1016/j.apenergy.2017.04.039.

[104] Bisoi R, Dash PK, Das PP. Short-term electricity price forecasting and classification in smart grids using optimized multikernel extreme learning machine. Neural Comput Appl 2018;32:1457–80. http://dx.doi.org/10.1007/s00521-018-3652-5.

[105] Kim MK. Short-term price forecasting of Nordic power market by combination Levenberg–Marquardt and Cuckoo search algorithms. IET Gener Transm Distrib 2015;9(13):1553–63. http://dx.doi.org/10.1049/iet-gtd.2014.0957.

[106] Pourdaryaei A, Mokhlis H, Illias HA, Kaboli SHA, Ahmad S. Short-term electricity price forecasting via hybrid backtracking search algorithm and ANFIS approach. IEEE Access 2019;7:77674–91. http://dx.doi.org/10.1109/access.2019.2922420.

[107] Ebrahimian H, Barmayoon S, Mohammadi M, Ghadimi N. The price prediction for the energy market based on a new method. Econ Res-Ekonomska Istraživanja 2018;31(1):313–37. http://dx.doi.org/10.1080/1331677x.2018.1429291.

[108] Abedinia O, Amjady N, Shafie-khah M, Catalão J. Electricity price forecast using Combinatorial Neural Network trained by a new stochastic search method. Energy Convers Manage 2015;105:642–54. http://dx.doi.org/10.1016/j.enconman.2015.08.025.

[109] Itaba S, Mori H. An electricity price forecasting model with fuzzy clustering preconditioned ANN. Electr Eng Japan 2018;204(3):10–20. http://dx.doi.org/10.1002/eej.23094.

[110] Ghofrani M, Azimi R, Najafabadi FM, Myers N. A new day-ahead hourly electricity price forecasting framework. In: Proceedings of the 2017 North American Power Symposium. 2017, p. 1–6. http://dx.doi.org/10.1109/naps.2017.8107269.

[111] Itaba S, Mori H. A fuzzy-preconditioned GRBFN model for electricity price forecasting. Procedia Comput Sci 2017;114:441–8. http://dx.doi.org/10.1016/j.procs.2017.09.010.

[112] ENTSO-E transparency platform. URL https://transparency.entsoe.eu/.

[113] PJM website. URL www.pjm.com.

[114] Elia. Grid data. URL http://www.elia.be/en/grid-data/dashboard.

[115] RTE. Grid data. URL https://data.rte-france.com/.

[116] Amprion website. URL https://www.amprion.net/.

[117] 50 Hertz website. URL https://www.50hertz.com/.

[118] TenneT website. URL https://www.tennet.eu/.

[119] Bergstra J, Bardenet R, Bengio Y, Kégl B. Algorithms for hyper-parameter optimization. In: Advances in Neural Information Processing Systems. 2011, p. 2546–54.

[120] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. Springer Series in Statistics, New York, NY, USA: Springer New York Inc.; 2001, http://dx.doi.org/10.1007/978-0-387-21606-5.

[121] Ziel F, Steinert R, Husmann S. Efficient modeling and forecasting of electricity spot prices. Energy Econ 2015;47:98–111. http://dx.doi.org/10.1016/j.eneco.2014.10.012.

[122] Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. Ann Statist 2004;32(2):407–99. http://dx.doi.org/10.1214/009053604000000067.

[123] Kingma DP, Ba J. Adam: A method for stochastic optimization. In: 3rd International Conference on Learning Representations, ICLR. 2015, http://arxiv.org/abs/1412.6980.

[124] Yao Y, Rosasco L, Caponnetto A. On early stopping in gradient descent learning. Constr Approx 2007;26(2):289–315. http://dx.doi.org/10.1007/s00365-006-0663-2.

[125] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res 2011;12:2825–30.

[126] Chollet F. Keras. In: GitHub repository. GitHub; 2015, URL https://github.com/fchollet/keras.

[127] Hyndman RJ. Errors on percentage errors. 2014, URL https://robjhyndman.com/hyndsight/smape/.

[128] Hyndman RJ, Koehler AB. Another look at measures of forecast accuracy. Int J Forecast 2006;22(4):679–88. http://dx.doi.org/10.1016/j.ijforecast.2006.03.001.

[129] Hyndman RJ, Athanasopoulos G. Forecasting: Principles and Practice. OTexts; 2018.

[130] Narayan PK, Bannigidadmath D. Are Indian stock returns predictable? J Bank Financ 2015;58:506–31.

[131] Diebold FX, Mariano RS. Comparing predictive accuracy. J Bus Econ Stat 1995;13(3):253–63. http://dx.doi.org/10.1080/07350015.1995.10524599.

[132] Diebold FX. Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of Diebold–Mariano tests. J Bus Econom Statist 2015;33(1):1–9. http://dx.doi.org/10.1080/07350015.2014.983236.

[133] Bordignon S, Bunn DW, Lisi F, Nan F. Combining day-ahead forecasts for British electricity prices. Energy Econ 2013;35:88–103. http://dx.doi.org/10.1016/j.eneco.2011.12.001.

[134] Nowotarski J, Raviv E, Trück S, Weron R. An empirical comparison of alternative schemes for combining electricity spot price forecasts. Energy Econ 2014;46:395–412. http://dx.doi.org/10.1016/j.eneco.2014.07.014.

[135] Serafin T, Uniejewski B, Weron R. Averaging predictive distributions across calibration windows for day-ahead electricity price forecasting. Energies 2019;12(13):2561. http://dx.doi.org/10.3390/en12132561.

[136] Marcjasz G, Lago J, Weron R, Schutter BD. 2020. Neural networks in day-ahead electricity price forecasting: single vs. multiple outputs,

[137] Giacomini R, White H. Tests of conditional predictive ability. Econometrica 2006;74(6):1545–78. http://dx.doi.org/10.1111/j.1468-0262.2006.00718.x.

[138] Giacomini R, Rossi B. Forecasting in macroeconomics. In: Handbook of Research Methods and Applications in Empirical Macroeconomics. Edward Elgar Publishing; 2013, p. 381–408. http://dx.doi.org/10.4337/9780857931023.00024.

[139] Ibrahim NNAN, Razak IAWA, Sidin SSM, Bohari ZH. Electricity price forecasting using neural network with parameter selection. In: Intelligent and Interactive Computing. Springer Singapore; 2019, p. 141–8. http://dx.doi.org/10.1007/978-981-13-6031-2_33.

[140] Panapakidis IP, Dagoumas AS. Day-ahead electricity price forecasting via the application of artificial neural network based models. Appl Energy 2016;172:132–51. http://dx.doi.org/10.1016/j.apenergy.2016.03.089.

[141] Singh NK, Singh AK, Tripathy M. Short-term load/price forecasting in deregulated electric environment using ELMAN neural network. In: Proceedings of the 2015 International Conference on Energy Economics and Environment. 2015, p. 1–6. http://dx.doi.org/10.1109/energyeconomics.2015.7235086.

[142] Reddy SS, Jung C-M, Seog KJ. Day-ahead electricity price forecasting using back propagation neural networks and weighted least square technique. Front Energy 2016;10(1):105–13. http://dx.doi.org/10.1007/s11708-016-0393-y.

[143] Nascimento J, Pinto T, Vale Z. Day-ahead electricity market price forecasting using artificial neural network with spearman data correlation. In: Proceedings of the 2019 IEEE PowerTech Conference. 2019, p. 1–6. http://dx.doi.org/10.1109/ptc.2019.8810618.

[144] Kotur D, Zarkovic M. Neural network models for electricity prices and loads short and long-term prediction. In: Proceedings of the 2016 International Symposium on Environmental Friendly Energies and Applications. 2016, p. 1–5. http://dx.doi.org/10.1109/efea.2016.7748787.

[145] Monteiro C, Ramirez-Rosado I, Fernandez-Jimenez L, Conde P. Short-term price forecasting models based on artificial neural networks for intraday sessions in the Iberian electricity market. Energies 2016;9(9):721. http://dx.doi.org/10.3390/en9090721.

[146] Monteiro C, Fernandez-Jimenez L, Ramirez-Rosado I. Explanatory information analysis for day-ahead price forecasting in the Iberian electricity market. Energies 2015;8(9):10464–86. http://dx.doi.org/10.3390/en80910464.

[147] Anamika, Kumar N. Market-clearing price forecasting for Indian electricity markets. In: Proceeding of International Conference on Intelligent Communication, Control and Devices. Springer Singapore; 2016, p. 633–42. http://dx.doi.org/10.1007/978-981-10-1708-7_72.

[148] Atiya AF. Why does forecast combination work so well? Int J Forecast 2020;36(1):197–200. http://dx.doi.org/10.1016/j.ijforecast.2019.03.010.

[149] Li Y, Zhu J. L1-norm quantile regression. J Comput Graph Statist 2008;17:163–85. http://dx.doi.org/10.1198/106186008x289155.