Energy Economics 106 (2022) 105742

Contents lists available at ScienceDirect

Energy Economics

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

Forecasting day-ahead electricity prices: A comparison of time series and neural network models taking external regressors into account

Malte Lehna [a,*], Fabian Scheller [b,c], Helmut Herwartz [d]

[a] Fraunhofer Institute for Energy Economics and Energy System Technology (IEE), Germany
[b] Energy Economics and System Analysis, Division of Sustainability, Department of Technology, Management and Economics, Technical University of Denmark (DTU), Denmark
[c] Institute for Infrastructure and Resources Management (IIRM), University Leipzig, Germany
[d] Chair of Econometrics, Georg-August-University Goettingen, Germany

ARTICLE INFO

Keywords:
Electricity price forecast
Time series forecasting
(S)ARIMA(X)
Vector autoregressive model
Long-short term memory neural network
Convolutional neural network

ABSTRACT

The amount of renewable energies in electricity production has increased significantly in the last decade, resulting in more variability of the day-ahead electricity spot price. The Electricity Price Forecast (EPF) has to adapt to the new situation by applying flexible models. However, the numerous available forecasting methods differ widely, with no distinct candidate offering the best solution. Against this background, we conduct a comparative study of four different approaches to forecasting the German day-ahead electricity spot price. In addition to the prominent Seasonal Integrated Auto-Regressive Moving Average model ((S)ARIMA(X)) and the Long-Short Term Memory (LSTM) neural network models, we employ a Convolutional Neural Network LSTM (CNN-LSTM) and an extended two-stage multivariate Vector Auto-Regressive model (VAR) approach as hybrid models. For better performance, we include common external influences such as the consumer load, fuel and CO_2 emission prices, average solar radiation and wind speed in our analysis. We analyse hourly data for twelve samples from October 2017 to September 2018. Each model is implemented to deliver price forecasts at three horizons, i.e., one day, seven days and thirty days ahead. While the LSTM model achieves the best forecasting performance on average, the two-stage VAR follows closely behind and performs exceedingly well for shorter prediction horizons. Further, we provide evidence that a combination of both forecasting methods outperforms each of the single models. This indicates that combining advanced methods could lead to further improvements in electricity spot price forecasts.

1. Introduction

1.1. Price forecasting

Ever since the liberalisation of the electricity markets, both the energy industry as well as the research community have been highly interested in Electricity Price Forecast (EPF). Different disciplines of energy economics offer a variety of methods and models to capture complex dynamics. However, the application of typical financial models for the forecast has to be executed with caution, due to the specific characteristics of the traded commodity. The electricity market differs from other financial markets, for the reason that both the demand and supply sides are exceedingly complex. Among other things, end-consumers have a highly inelastic demand primarily depending on specific time patterns. Furthermore, the demand has to be precisely balanced to avoid blackouts or an overload of the network structure.

Abbreviations: (S)ARIMA(X), Seasonal Integrated Auto-Regressive Moving Average model; AR(X), Autoregressive model; ARIMA, Integrated Auto-Regressive Moving Average model; ARMA(X), Auto-Regressive Moving Average model; BIC, Bayesian Information Criterion; CNN, Convolutional Neural Network; CNN-LSTM, Hybrid Convolutional and Long-Short Term Memory neural network; EPEX SPOT, European Power Exchange; EPF, Electricity Price Forecast; EU-ETS, European Emission Trading System; FSAPE, Frequency of Smallest Absolute Prediction Error; GARCH, Generalised AutoRegressive Conditional Heteroscedasticity; LSTM, Long-Short Term Memory; MA, Moving Average; MAE, Mean Absolute Error; mlog, Mirror-logarithmic; NARX-NN, Nonlinear Auto-regressive Neural Network; OLS, Ordinary Least Squares; PEPF, Probabilistic Electricity Price Forecast; RMSE, Root Mean Square Error; VAR, Vector Auto-Regressive model

* Corresponding author.
E-mail address: malte.lehna@iee.fraunhofer.de (M. Lehna).

https://doi.org/10.1016/j.eneco.2021.105742
Received 26 October 2020; Received in revised form 16 November 2021; Accepted 23 November 2021
Available online 20 December 2021
0140-9883/© 2021 Elsevier B.V. All rights reserved.

Concerning the supply side, both inflexible and volatile electricity plants are combined, thus increasing the complexity of the energy distribution. All these different factors have to be considered due to their influence on the price and the corresponding price prediction.

For the most important electricity markets, most research is conducted on the day-ahead spot market. Even though the trade volume on the intraday market has increased significantly in recent years (EPEX SPOT, 2021), the major amount of trades is still conducted on the day-ahead market, where the electricity prices for the next day delivery are traded and determined in an auction exchange. Regarding the German spot market, the day-ahead prices are determined at the European Power Exchange (EPEX SPOT) on the previous day at 12.00 pm through an auction under the merit order principle. Accordingly, available offers are sorted in ascending order, based on their auction prices. Thereafter, the orders are distributed to meet the electricity demand, starting with the lowest order. The resulting market-clearing price is then the marginal cost of the last supplier that obtains the order and defines the spot price for the respective product. Different electricity products, i.e., other distribution intervals, are traded on the EPEX SPOT, however, we focus in this study on the hourly products.

In terms of the model architecture, the recent EPF literature has indicated a need for more flexible models as well as the inclusion of new variables in the research design (Gürtler and Paulsen, 2018). Especially, the rise of renewable energies has induced additional uncertainty in the spot price, which corresponds to a larger variability in the price development. In this context, up to 37.8% of German electricity was produced in 2018 from renewable energy sources, which consist primarily of wind (49.5%) and solar (20.5%) energy (AGEE-Stat Umweltbundesamt, 2019). Major reasons for the increase in renewable energies have been governmental subsidiaries and regulations introduced in Germany (Cludius et al., 2014). Due to special accountability and their low marginal cost, renewable energies further impact the spot price through the merit order principle (Cludius et al., 2014; Ketterer, 2014; Paschen, 2016). Consequently, researchers not only need to incorporate the operating prices of carbon-based power plants, CO_2 compensations and consumer loads in their models but also have to account for the influence of renewable energies.

1.2. Related work

Depending on the interpretation, contributions to the field of EPF can be separated into four (Gürtler and Paulsen, 2018) or five (Weron, 2014) groups, classifying them into multi-agent, fundamental, reduced form, statistical and computer intelligence models. Research has shifted from theoretical, long-term simulation models to more statistical and computational intelligence models in the last years. In the bibliometric analysis, Weron (2014) noted that more than two-third of EPF papers either include time series models, neural network models or both. In contrast, Gürtler and Paulsen (2018) identified the autoregressive time series models as the most frequently used approaches. Within the class of time series models, one can primarily differentiate between univariate and multivariate regression settings. Prominent models from the univariate framework are, according to Gürtler and Paulsen (2018), the Autoregressive model (AR(X)), the Auto-Regressive Moving Average model (ARMA(X)) and the Seasonal Integrated Auto-Regressive Moving Average model ((S)ARIMA(X)), all augmented with exogenous variables. These models have been frequently used and extended within the literature (Crespo Cuaresma et al., 2004; Contreras et al., 2003; Weron and Misiorek, 2008; Xie et al., 2013). In addition, researchers have often often used the AR(X) and Integrated Auto-Regressive Moving Average model (ARIMA) models as baseline models in comparative studies, e.g., Chen et al. (2019), Peng et al. (2018) and Ugurlu et al. (2018). Next to the basic ARIMA models, other univariate autoregressive models have been proposed in the EPF, such as Threshold Autoregressive models (Weron and Misiorek, 2008) or Generalised AutoRegressive Conditional Heteroscedasticity (GARCH) volatility models (Liu and Shi, 2013). In terms of the multivariate time series models, especially the Vector Auto-Regressive model (VAR) seems to be of interest in the EPF community (Paschen, 2016; Ziel and Weron, 2018; Ziel et al., 2015; Haldrup et al., 2010). While (Ziel and Weron, 2018) implemented the VAR model for a comparison between multivariate and univariate time series models, Paschen (2016) addressed the structural shocks of wind and solar power on the German spot market with a Structural VAR model. Moreover, Ziel et al. (2015) analysed the performance of a VAR model in combination with a Threshold Autoregressive Conditional Heteroskedasticity model on the German spot market, and compared them with other univariate and multivariate models. Finally, Haldrup et al. (2010) used a combination of a VAR model with regime switching mechanisms to model energy price congestion on the Scandinavian Nord Pool electricity market.

As reviewed by Zhang and Fleyeh (2019), several approaches are available within the class of the neural networks. While (Zhang and Fleyeh, 2019) primarily focused on Artificial Neural Networks, they remarked that especially the deep learning networks show potential in the EPF. The defining characteristics of these models are multiple layer structures of neurons that enable complex data assessment and, in some cases, even the description of time-dependent structures. While (Sharma and Srinivasan, 2013) applied simple Recurrent Neural Networks in a multi-layer framework, Marin et al. (2018) utilised a Nonlinear Auto-regressive Neural Network (NARX-NN) that also captures the autoregressive components of the data. Furthermore, Lago et al. (2018) made predictions by means of more complex neural networks with structures such as the Long-Short Term Memory (LSTM), a Gated Recurrent Unit or a Convolutional Neural Network (CNN). The advantage of these three models is that they distinguish themselves by their scope to filter past effects with different methods. For this reason, other researchers such as Peng et al. (2018) also analysed the performance of LSTM models and came to similar conclusions as Lago et al. (2018). Finally, Zhang and Fleyeh (2019) additionally asserted in their review paper that hybrid combinations which include at least one neural network further increase the predictive power.

In a direct comparison between the time series and the neural network approaches, researchers have been unclear which class outperforms the other in terms of EPF (Gürtler and Paulsen, 2018). On the one hand, Marin et al. (2018) compared the forecastability of an ARMA(X) model with a NARX-NN and stated no significant difference. In contrast, Ugurlu et al. (2018) applied a Gated Recurrent Unit neural network which outperformed the SARIMA(X) in forecasting the electricity prices of the Turkish market. In Lago et al. (2018), a large set of statistical and neural network approaches have been tested on the Belgian electricity market and the results show a clear advantage of the neural networks. Furthermore, another problem regarding the comparability of the research results is the difference between the target market and the underlying external variables across the literature. According to Gürtler and Paulsen (2018), the most prominent markets are the US (California, PJM Power Pool) as well as the European markets (Spain, Nord Pool, Germany, UK, ...), yet all these markets incorporate different characteristics and price developments depending on their national legislative environment. Concerning explanatory variables, the most chosen variable was the load of the consumers, either as an actual variable or as a day-ahead forecast (Gürtler and Paulsen, 2018). Other factors such as demand and supply ratio (Ugurlu et al., 2018), export measurement (Lago et al., 2018), or temperature data of the country to assess the heating demand (Gürtler and Paulsen, 2018) have been utilised as well. With the increased expansion of renewable energies, more and more researchers have included the production of renewable energies (Ketterer, 2014; Paschen, 2016; Ziel, 2017). The results of Ketterer (2014) demonstrate that in intervals with high wind power generation, a decrease in the overall spot price is visible while the volatility of the price increase. Similar results have been presented by Ziel (2017), where an impact of the wind and solar forecasting

errors on the intraday spot price of Germany could be shown. Finally, according to Nowotarski and Weron (2018) recent contributions suggest to depict the EPF by means of Probabilistic Electricity Price Forecast (PEPF) models that describe both the point forecast as well as the variability of the expected electricity price. With regard to neural networks, respective advancements have been made by Dudek (2016) and Marcjasz et al. (2020) who suggest a multilayer perceptron and a NARX-NN, respectively. Regarding the time series models, Muniain and Ziel (2020) suggest ARX models with a constant conditional correlation GARCH extension, while (Uniejewski and Weron, 2021; Maciejowska and Nowotarski, 2016) propose Quantile Regression Averaging for the PEPF. Note that other methods have been proposed as well for probabilistic forecasts, especially Bayesian approaches in Kostrzewski and Kostrzewska (2019).

In summary, it needs to be noted that there is still no dominant model for EPF even though the number of comparative studies has become sizeable in the last years. While the optimal selection of the variables is not clearly defined, a shift to modelling the influence of wind and solar production is visible.

1.3. Research objective

Regarding the previously mentioned research advancements, we address the EPF model uncertainty by offering a comparative investigation of forecasting models from the time series and neural network prediction classes. Next to using two prominent EPF models, i.e., LSTM and (S)ARIMA(X), we further analyse the performance of two hybrid models. The hybrid models in question are the CNN-LSTM, and an own suggestion of a two-stage VAR model which combines an Ordinary Least Squares (OLS) estimation step with a multivariate VAR time series model. As such, the suggested multivariate model is straightforward to implement and interpret. Finally, we apply two model combinations of the different prediction classes. Ultimately, this paper answers the following research question: With the best candidate from each prediction class selected, which is the most appropriate model to predict the German day-ahead spot price for monthly, weekly, and daily periods?

In addition, we propose a more indirect approach with regard to the inclusion of renewable energies in the form of external regressors. While various external variables are taken into account to capture the electricity price variation in other recent research papers, we use weather data to proxy renewable energies. Moreover, the forecasting performance of the alternative models is assessed for each month of a year concerning longer (thirty days), medium (seven days) and shorter (one day) forecasting horizons and three different performance metrics (Root Mean Square Error (RMSE), the Mean Absolute Error (MAE), and the Frequency of Smallest Absolute Prediction Error (FSAPE)). Finally, we provide evidence that different prediction classes complement each other and that a combination of neural networks with multivariate time series frameworks leads to improvement in the forecasting performance.

The remaining parts of this work are organised as follows: in Section 2, we outline the theoretical background of all four models as well as the combination approaches. After that, in Section 3 we introduce our research methodology, consisting of the data description, spot price transformation, the model selection process and the forecast performance metrics. Subsequently, we present our research results in Section 4. Complementary to the overall results, we conduct the analysis for alternative forecast horizons and analyse the scope of model combinations for forecasting electricity prices. In Section 5 we discuss the implications of the forecast results and focus primarily on the best performing models. Section 6 summarises and concludes.

2. Forecasting modelling concepts

2.1. (S)ARIMA(X) model

The first model is the (S)ARIMA(X) model, which is an extension of the ARIMA time series model based on Box et al. (2008). However, in EPF research, e.g., Xie et al. (2013), the basic model has been extended further to include seasonal components as well as external variables. Overall, there are multiple advantages of the (S)ARIMA(X) model concerning the EPF. First, the electricity price displays both strong autoregressive as well as seasonal patterns, which both can be described by means of the (S)ARIMA(X) model. Furthermore, it is possible to include external variables, which can increase the forecast performance considering the influence of consumer loads and renewable energies. Another advantage of the model is that coefficients can be interpreted in terms of their influence and significance, which, e.g., helps in the selection process of new explanatory variables. Finally, through the repeated usage of the (S)ARIMA(X) model by other researchers, as elaborated in Section 1.2 and seen in Crespo Cuaresma et al. (2004) and Gürtler and Paulsen (2018), one can consider the model as a baseline model. Consequently, we include the (S)ARIMA(X) model in our study as a benchmark to compare the performance of three further model candidates.

According to Hyndman and Khandakar (2008), the theoretical structure of the (S)ARIMA(X) model can be described through a combination of autoregressive (AR) and moving average components (MA), where y_t denotes the respective time series, ϵ_t the independently and identically distributed residuals with mean zero and constant variance, $\epsilon_t \overset{iid}{\sim} (0, \sigma^2)$, and B the backshift operator:

$$\Phi(B^S)\phi(B)(1-B^S)^D(1-B)^d y_t = X_t\beta + \Theta(B^S)\theta(B)\epsilon_t \tag{1}$$

The autoregressive influence is modelled by means of theAR(X) component $\phi(B)$ of order p and the seasonal AR(X) component $\Phi(B)$ of order P. Moreover, $\theta(B)$ and $\Theta(B)$ characterise the MA and seasonal MA influence of order q and Q. Finally, $(1-B^S)^D(1-B)^d$ indicate the differencing of the time series with the order d and D, respectively. Next to the seasonal patterns, the (S)ARIMA(X) further includes with $X_t\beta$ the influence of the external variables in the model (Makridakis et al., 2008). Note that for all extensions of the ARIMA model stationary conditions exist (Hyndman and Khandakar, 2008).

2.2. The two-stage VAR model

For the second forecasting approach, a two-stage model has been developed by combining an Ordinary Least Squares (OLS) estimation step with a multivariate VAR time series model (Lütkepohl, 2005). With regard to the EPF, the VAR model has not been used frequently to forecast spot prices, especially to the best of our knowledge, not in combination with a first step OLS estimation. While (Ziel and Weron, 2018) proposed the VAR model for the EPF, they did not include external variables within their experiment. However, as previously discussed, it is necessary to include external variables in order to describe the spot price adequately, hence our proposition of a two-stage model.

In the first stage, we presume a linear dependency between the external variables and the corresponding time series as follows:[1]

$$y_t = X_t\beta + e_t \tag{2}$$

[1] To examine this presumption, we conducted a correlation test (for both for Pearson and Spearman correlation coefficient) between the explanatory variables and the spot price (RDocumentation, 2021; Best and Roberts, 1975). With the rejection of the H_0 hypothesis, i.e., that the correlation between two samples is zero, we can assume that some relationship exists between the variables.

The OLS estimation of the model in (2) obtained that all covariates in X_t were highly significant. However, we assume that the residuals e_t still exhibit patterns of unmodelled serial correlation that provide valuable information for predictive analysis. Thus, we propose a VAR model, as the second component to account for multivariate autoregressive structures within the error terms. In a VAR model, the variables are generally expected to occur simultaneously, the realisations of the variables have joint interaction and are to some extend correlated with each other. Based on the general concept of Lütkepohl (2005), the VAR model for residual estimates is developed in a daily framework in accordance with (Ziel and Weron, 2018). To begin with, the residuals e_t are combined to a daily variable E_τ, with dimension $k = 24$ and a length of $\tau = 1, \dots, T/k$:

$$E_1 = \begin{bmatrix} e_1 \\ e_2 \\ \dots \\ e_{24} \end{bmatrix}, \; E_2 = \begin{bmatrix} e_{25} \\ e_{26} \\ \dots \\ e_{48} \end{bmatrix}, \; \dots, \; E_\tau = \begin{bmatrix} e_{(\tau-1)k+1} \\ e_{(\tau-1)k+2} \\ \dots \\ e_{\tau k} \end{bmatrix} \tag{3}$$

With the newly defined variable E_τ the VAR(p) model with the autoregressive order p can then be formalised as

$$E_\tau = \eta + A_1 \, E_{\tau-1} + A_2 \, E_{\tau-2} + \dots + A_p \, E_{\tau-p} + \upsilon_t \tag{4}$$

Accordingly, the VAR model is based on the assumption that the observations E_τ are determined through an autoregressive influence of prior observations as well as a random component υ_t. Similar to the univariate case, the υ_t is considered a random white noise, however, υ_t is multivariate distributed. To qualify as serially uninformative white noise, υ_t aligns with the following conditions:

$$E(\upsilon_t) = 0, \;\; E(\upsilon_t \upsilon_s') = 0 \text{ for } s \neq t \;\; \text{and} \;\; E(\upsilon_t \upsilon_t') = \Sigma_\upsilon \;\; \forall t \tag{5}$$

The VAR(p) process is stationary if it holds that (Lütkepohl, 2005):

$$\det(I_k - A_1 z - \dots - A_p z^p) \neq 0 \text{ for } |z| \leq 1 \tag{6}$$

Regarding the effects of $E_{\tau-i}$, A_i are parameter $(k \times k)$ parameter matrices.[2] Hence, each variable of the previous observation does not only influence itself but also has a specific effect on the other variables of the observation τ. Next to the autoregressive component, it is also possible to incorporate a fixed intercept η which is independent of time.

With the multivariate approach, the interpretation of the electricity price forecast changes in a major way. Instead of assuming that the price is primarily influenced by previous hours, the model now implies that it is determined by the realisations of the previous day, while the 24 h of the given day are only correlated with each other. Through the combination of the two models in (2) and (4), it is, therefore, possible to include external influences while simultaneously implementing a multivariate framework to depict the daily structure of the data. In Section 5, we further justify this particular model feature.

2.3. LSTM neural network model

The third model is a LSTM neural network which is frequently used for time series forecasting and displayed on the left-hand side of Fig. 1 With regard to the EPF, the LSTM offers multiple advantages which make the neural network especially appealing (Zhang and Fleyeh, 2019). First, through its neural network structure, the model is not limited to the description of linear relationships but instead is able to depict non-linear and more complex relationships. Furthermore, the network is trained autonomously, without prior knowledge of the data. As a result, the model is able to detect which variables are important for the spot price prediction, even though some relationships between the variables might be unknown to the researcher. Finally, another major advantage of the neural network is its scope to process large data sets. Given the recent advances in deep neural networks, it is possible to adjust the architecture of the models to meet with the complexity of the underlying data.

The general LSTM model has been developed as an improvement of the Recurrent Neural Network, in order to solve the vanishing gradient problem (Hochreiter, 1991). As a solution, Hochreiter and Schmidhuber (1997) proposed the LSTM model that was capable of modelling recurrent time series data. Instead of a simple recurrent structure, the LSTM model consisted of multiple cells that filter the input, store information in a specific memory and return the output depending on both input and memory, as displayed in Fig. 2. In the following, we describe the structure of the LSTM cell and the mechanisms within.

Overall, the core of each cell is the cell state vector C_t which operates as the long-term memory of the network. For each observation, the previous cell state C_{t-1} is extracted, updated to C_t and then used for the computation of the output. In order to correctly specify this updating process, the LSTM cell consists of three gates that manage the information flow of the cell state. Within each gate, embedded activation functions transform the input into a signal. In the case of the LSTM these activation functions are generally the following *sigmoid* ($\sigma(x)$) and a *tanh* ($tanh(x)$) activation functions, however, other activation functions are available as well (Karlik and Olgac, 2011):

$$\sigma(x) = \frac{1}{1 + exp(-x)} \tag{7}$$

$$tanh(x) = 1 - \frac{exp(x) - exp(-x)}{exp(x) + exp(-x)} \tag{8}$$

In terms of the gating mechanisms, the first gate functions as a filter for the past information of C_{t-1}. Given the previous output z_{t-1} of the cell and current input X_t, a *sigmoid* activation function defines the proportion of information that should remain in the cell state:

$$f_t = \sigma(W_f [z_{t-1}, X_t] + b_F) \tag{9}$$

In order to add information to the cell state, the second gate is implemented that manages new input for the cell state. Within this gate, the new information $\tilde{C}_t$ as well as the proportion i_t of the input are both calculated based on z_{t-1} and X_t:

$$\tilde{C}_t = tanh(W_C [z_{t-1}, X_t] + b_C) \tag{10}$$

$$i_t = \sigma(W_i [z_{t-1}, X_t] + b_i) \tag{11}$$

Through a combination of the two gates the previous C_{t-1} is updated to the new cell state C_t:

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{12}$$

Lastly, the third gate transforms the information of the new cell state through a *tanh* activation function in order to model the output z_t. Similar to the input gate, a proportion measurement o_t is further constructed based on z_{t-1} and X_t:

$$o_t = \sigma(W_o [z_{t-1}, X_t] + b_o) \tag{13}$$

$$z_t = o_t * tanh(C_t) \tag{14}$$

Given this general idea, multiple LSTM cells can be combined to one large framework to model time-dependent variables. After that, the respective model is optimised through an error minimisation framework, with the trainable parameters of the model being the weights (W_i) and the bias (b_i) of the respective cells. For the optimisation, a loss function, illustrating the misspecification of the data, is defined and minimised through a gradient descent algorithm (Hecht-Nielsen, 1992). For the empirical analysis in this work, we decided to use the Adam algorithm (Kingma and Ba, 2017) as the optimisation algorithm due to its high computational efficiency.

[2] Note that the order p has to be selected with care because a higher-order increases the number of unknown model parameters considerably. Due to the size of each A_i matrix, a total of pk^2 parameters have to be estimated.

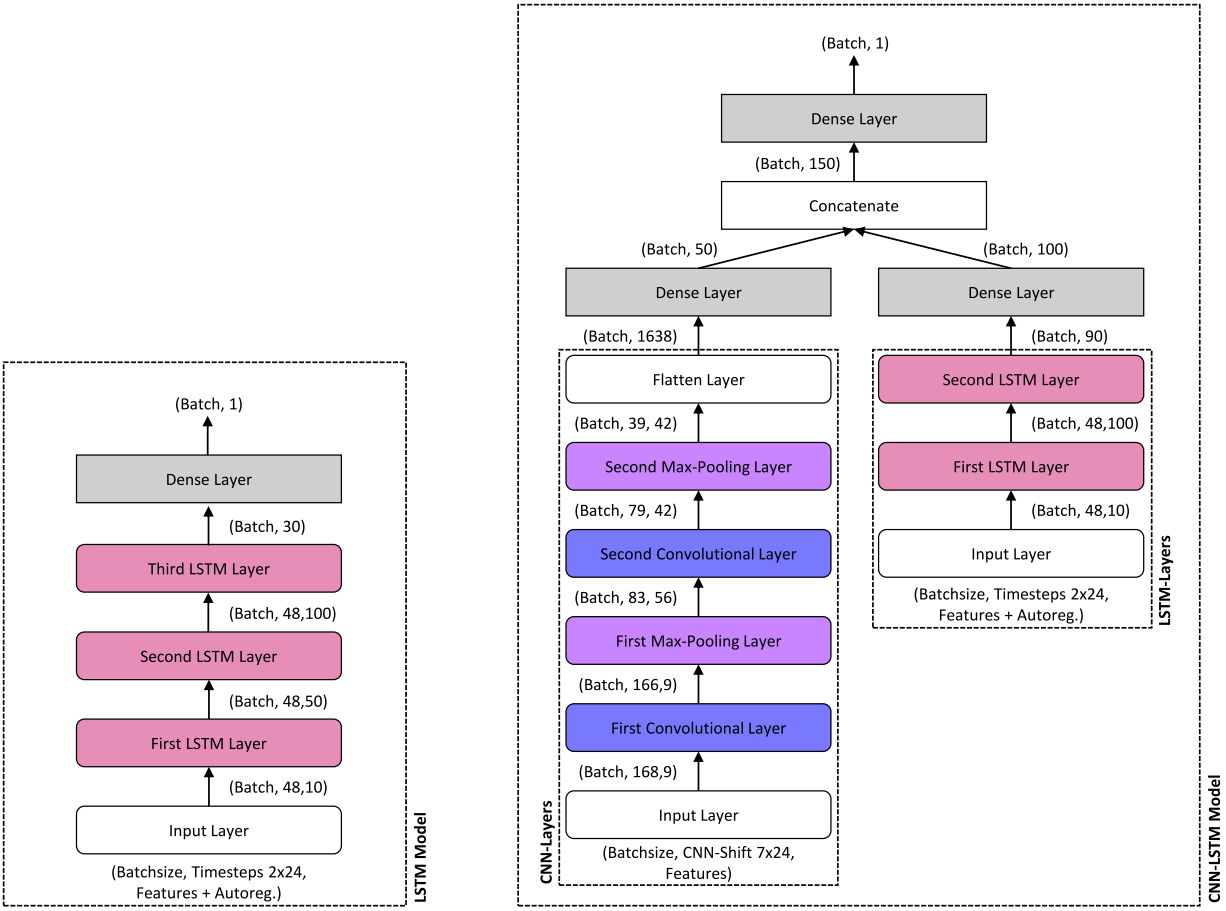


Fig. 1. Visualisation of the LSTM and CNN-LSTM model. In the LSTM model the information is fed through three LSTM layers before being aggregated in the final Dense layer. The CNN-LSTM is constructed with two input branches, the CNN branch and the LSTM branch. Note that we only include the explanatory variables in the CNN branch, while the LSTM branch also includes the lagged spot price, thus explaining the difference in the feature dimension.

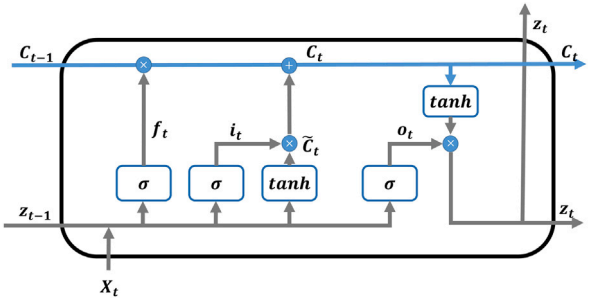


Fig. 2. Visualisation of LSTM cell, where the blue line depicts the cell state vector C_t. The figure is based on a visual explanation of Olah (2015). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

2.4. CNN-LSTM neural network model

As the fourth prediction approach, we propose an extension of the LSTM model by combining the model with a CNN to a hybrid CNN-LSTM model as displayed on the right-hand side of Fig. 1. Even though the CNN is primarily known for image classification, researchers recently adopted the model in the time series context (Kuo and Huang, 2018; Lago et al., 2018). Overall, the general advantages of the CNN model are its ability to aggregate sequences from the data and filter only the important information. Concerning the one-dimensional time series data, the network is therefore capable of identifying patterns and selecting specific seasonal structures. In the EPF context, the CNN-LSTM has already been proposed by Kuo and Huang (2018). Next to describing the strong time dependency of the German spot price data by means of the LSTM component, it is possible to distinguish patterns in the data and select distinct features. Thus, through the combination of the networks, one can analyse the spot price from multiple perspectives. The following descriptions of the one-dimensional CNN are based on Kiranyaz et al. (2021).

Generally, the idea behind the one-dimensional CNN is that a large data set is reduced to multiple feature maps that each exhibit specific characteristics of the original data set. This reduction of the data is primarily accomplished through a combination of two specific layers, which are the Convolutional layer and the Max-pooling layer, as exemplary displayed in Fig. 3. The purpose of the Convolutional layer is to aggregate the overall data, by creating multiple feature maps, i.e., filters, each different filtering characteristic of the data. In the aggregation process of the feature maps, multiple neurons with

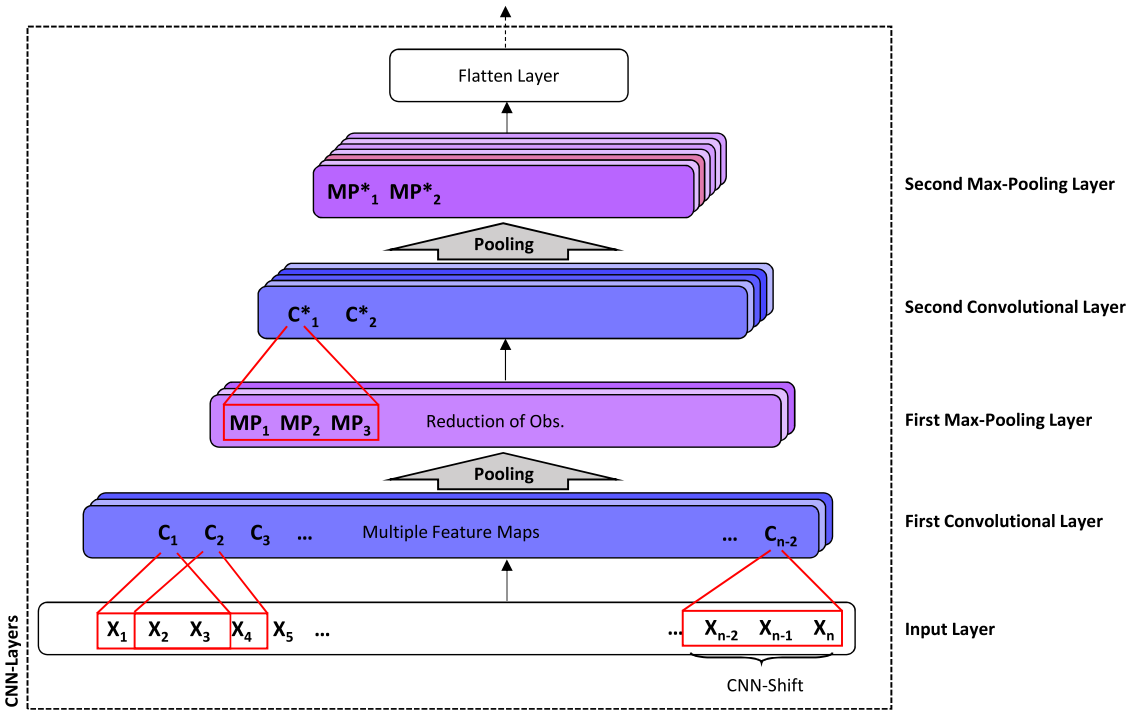


Fig. 3. Visualisation of the CNN branch with two convolutional layers and two max-pooling layers. The pooling results in a reduction of half the data size.

different weights inspect the underlying data through a window filter, where the network shifts with a specific kernel size across the data and feeds the information into the neurons. Given the different activation results, each feature map extracts individual and distinct information from the data and thus highlights specific characteristics. As a result, the Convolutional layers can identify patterns and reduce the noise within the data. However, in many cases, the data sample is quite large compared with the window size. Thus, one simple Convolutional layer is insufficient to summarise the data, and further reduction measures are necessary. Therefore, the Max-pooling layer is additionally implemented within the CNN architecture. Similar to the Convolutional layer, the purpose of the Max-pooling layer is to aggregate important information. However, instead of using a window filter that shifts over the data, the layer compares a portion of the input (C_i) and extracts the maximum value from the selected segment. Depending on the pool size, the pooling layer reduces the feature map to at least half the size. Thus only the most important information remains within the network. In the last step, the information is fed through a flatten layer to reduce the dimensions and a dense layer to return the output. By means of a suitable combination of both Convolutional layers and Max-pooling layers, it is possible to extract the most important information of the data while simultaneously shrinking the overall input for the layers significantly. Furthermore, similar to the LSTM, the CNN is also modelling the relationships of interest in a non-linear fashion.

2.5. Combined forecast models

The alternative forecasting methods described so far allow for a broad classification into time series regression models on the one hand and neural network approaches on the other hand. These model classes can be considered to differ structurally concerning the processing of available information. To take benefit from potential model complementarities between the classes, we further pursue model combinations for predictive analysis. The idea that ensemble models might outperform single prediction models in the EPF has been suggested, for instance, by Neupane et al. (2017). While, in principle, one could think of four alternative model combinations with one representative from both predictor classes, we decide to combine those two model approaches from each class that outperform their within-class counterparts in terms of overall performance criteria. Let $\hat{y}_t^{TS}$ and $\hat{z}_t^{NN}$ denote the predicted energy price of the overall best performing time series regression (TS) and neural network (NN) model, respectively. We suggest two alternative model combinations, a simple averaging method and a more advanced ensemble approach to analyse the scope of model combinations. Concerning the first approach ($Combined_{naive}$), the combined forecast is given as follows:

$$y_t^{CN} = \frac{\hat{y}_t^{TS} + \hat{z}_t^{NN}}{2} \tag{15}$$

As a further and more complex model combination, we construct an ensemble model ($Ensemble_{LSTM}$) that is based on neural network ensemble models (Ardabili et al., 2019). Conditional on a backward looking horizon h, the predictor $y_t^{Ensemble}$ is the following:

$$y_t^{Ensemble} = f(\hat{y}_t^{TS}, \hat{z}_t^{NN}, \hat{y}_{t-1}^{TS}, \hat{z}_{t-1}^{NN} \dots, \hat{y}_{t-h}^{TS}, \hat{z}_{t-h}^{NN}) \tag{16}$$

Due to the fact that we want to include a possible non-linear relation, we design the model $f()$ similar to the LSTM model of Section 2.3. However, instead of external data, the model was solely trained on the previous output of the single models. Thus, it can only be influenced by the information within the forecasted values $\hat{y}_i^{TS}$ and $\hat{z}_i^{NN}$ for $i = t, t-1, \dots, t-h$. A potential advantage of the second model is that it combines previous and current forecasts of the two single models to a refined prediction. Consequently, it might exploit possible structural differences between the forecasts more effectively than the $Combined_{naive}$ approach.

3. Methodology

3.1. Case study data

In this analysis, we chose the hourly German day-ahead spot price as our target variable, which is traded at the EPEX SPOT. To compare the model performance across different months, we selected a period between October 2017 and September 2018 and separated the interval

into twelve monthly prediction batches. The choice of the prediction interval was influenced by the fact that the German electricity market changed in October 2018 due to a market split into the German, and the Austrian Electricity market (Next Kraftwerke AT GmbH (NEXT), 2019). Further, due to the different lengths of the months, we display the exact time intervals both for training and testing in Table 6 in Appendix A. The size of each monthly prediction batch has been set to 720 observations, thus having a prediction horizon of 30 days. For additional analysis, we further selected the first 24 and 168 observations from each batch to conduct a one-day-ahead and one-week ahead forecast. For each prediction period, we chose the previous 360 days as training period, resulting in 8640 hourly observations per training batch. In addition to the spot price, a group of exogenous variables have been selected to enhance the forecasting performance, see Table 1. Following the suggestions of Gürtler and Paulsen (2018), energy consumption of the consumers, weather data, CO_2 emission prices, as well as the cost of fuel, were taken into account.

To capture the electricity demand, we decided to include the hourly logarithmised load forecast of the German consumers. The load forecast was commercially distributed by the network operator and has frequently been used in the EPF literature (Gürtler and Paulsen, 2018). As the second group of variables, we chose to include the effect of renewable energies in our analysis. However, we did not implement renewable electricity production directly but instead provided weather data as proxies to describe the production indirectly. The chosen proxies were the average wind speed and average solar radiation which corresponds to the largest two renewable energy sources of Germany (AGEE-Stat Umweltbundesamt, 2019). Given that both wind and solar parks are not uniformly distributed across Germany, we averaged the observations of three weather stations that were located within important production regions.[3] Furthermore, it is important to note that we used the actual observations instead of forecasted weather data. This decision is because we did not want to induce further uncertainty into our comparative analysis, because the primary focus of this analysis is the performance of the models. The third group of variables was long-term factors that influence the cost of production and could therefore be used to explain possible trends in the spot price. The first of these variables was the price for CO_2 emissions which we included in the form of primary auction prices from the European Emission Trading System (EU-ETS). Since 2005, energy companies have to compensate for the emitted CO_2, thus we expected the emission price to influence the cost of energy production. Furthermore, we included the fuel prices for the fossil-fuelled power plants with the historical gas and coal prices. However, both the fuel prices and CO_2 emission prices have been recorded as daily variables, thus we cannot explain sudden hourly changes in the spot price by means of these variables. Lastly, two additional exogenous variables were created to describe seasonal patterns. The first was a categorical variable that categorises the year into winter, summer and intermediate seasons. The second variable was a Fourier Frequency from the spot price of order three which was provided for the time series models to describe a daily wave pattern. With regard to the neural networks, the frequencies induced a weaker performance, thus they were excluded from these two models. Regarding the data sources, the spot price and the load forecast data, were taken from the Open Power System Data Project (Open Power System Data (OPSD), 2019). In the case of the weather data, both wind speed and solar radiation were supplied by the German Weather Service (Deutscher Wetterdienst (DWD), 2019b,a). Finally, the EU-ETS auction results were taken from European Comission (EC) (2019) and the coal and gas prices have been taken from Finanzen.net (2019b) and Finanzen.net (2019a), respectively.

3.2. Spot price transformation

Given the variability of the spot price, it is common to deflate the influence of extreme outliers in the data which in most cases has been done by logarithmising the variable (Gürtler and Paulsen, 2018). However, this approach is only suitable if the variable in question is strictly positive. In recent years, due to repeatedly negative values in the German spot price, this condition could not be met. Thus, based on Uniejewski et al. (2018) a Mirror-logarithmic (mlog) transformation has been used in this work to increase the predictability and reduce the influence of outliers. With the transformation parameter $c \in (0, 1)$ the mlog(c) is defined as

$$y_t = \text{sgn}(p_t) \left(\log \left[|p_t| + \frac{1}{c} \right] + \log(c) \right) \tag{17}$$

Note that prior to the transformation, the spot price P_t had to be standardised (p_t). Moreover, the parameter c is not only a scaling parameter but also shifts the implemented p_t values by $\frac{1}{c}$. This ensures that no values smaller than one are logarithmised which is essential for re-transformation. Based on different test trials, we decided to set $c = \frac{1}{3}$ which accords with (Uniejewski et al., 2018). To re-transform the data the following inverse is used for calculations:

$$p_t = \text{sgn}(y_t) \left(\exp \left[|y_t| - \log(c) \right] - \frac{1}{c} \right) \tag{18}$$

Next to the mlog-transformation of the spot price, we further standardised the external variables for the neural networks to ensure stable performance.

3.3. Model selection

Given the above-mentioned data settings, all four prediction models needed to be specified concerning their optimal model fit.

For the (S)ARIMA(X) model, the selection process has been based on the Bayesian Information Criterion (BIC) (Schwarz et al., 1978).[4] For each of the twelve training periods, permutations of the model order $(p, d, q)x(P, D, Q)_{24}$ were fitted and the best candidate with the lowest BIC selected. Thereafter, the most promising candidate among all periods was chosen as the global model order. The seasonal frequency was set to $S = 24$ to capture daily patterns of the spot price and the maximal order was restricted to a total order of six. With a total of nine out of twelve samples, the (S)ARIMA(X) $(1, 0, 1)x(2, 0, 0)_{24}$ model showed the lowest BIC statistics. Thus, the model in (1) can be reformulated with the given model order:

$$y_t = \mu + \beta X_t + \phi_1 y_{t-1} + \Phi_{24} y_{t-24} + \Phi_{48} y_{t-48} - \theta_1 \epsilon_{t-1} + \epsilon_t \tag{19}$$

In the case of the two-stage VAR model, the selection process has only been necessary for the $VAR(p)$ process. Similar to the (S)ARIMA(X) model, we chose the BIC as selection criterion. The results were unambiguous across all samples to favour a representation of the model in (4), i.e.,

$$E_\tau = \eta + A_1 \; E_{\tau-1} + \upsilon_t \tag{20}$$

For the LSTM neural network the selection process needed to be different in comparison with the time series models. First, we had to define the LSTM architecture and thereafter optimise the hyperparameters of the model. Regarding the architecture, we built the network with three LSTM layers plus an additional Dense layer to aggregate the final results (see Fig. 1). For the hyperparameter optimisation, each training data set was split and the last 25% were used as validation samples

[3] We decided for the respective weather stations based on the wind and solar performance of the corresponding regions, as seen in Fraunhofer-Institut für Umwelt-, Sicherheits- und Energietechnik (2015b,a). In addition, note that some observations of the selected weather stations had to be replaced by neighbouring stations due to missing values.

[4] The choice of the Bayesian Information Criterion (BIC) is in accordance with the related literature, where the BIC and the Akaike information criterion are often preferred for model selection purposes, see Gürtler and Paulsen (2018). In this analysis, we decided to use the BIC, because it penalises the number of model parameters more strongly (Schwarz et al., 1978).

Table 1
Summary of the spot price and the external variables.

Variable	Mean	St. Dev.	Min	Max	Source	Remarks
Spot Price	34.923	16.467	−130.090	163.520	Open Power System Data (OPSD) (2019)	
Forecasted consumer load	10.901	0.179	10.269	11.237	Open Power System Data (OPSD) (2019)	Logarithmised
Average wind speed	4.403	2.106	0.367	17.000	Deutscher Wetterdienst (DWD) (2019b)	
Average solar radiation	47.329	72.336	0	335	Deutscher Wetterdienst (DWD) (2019a)	
CO_2 emission prices	8.589	4.807	3.940	24.850	European Comission (EC) (2019)	Daily frequency
Coal prices	46.256	8.854	34.570	69.090	Finanzen.net (2019b)	Daily frequency
Gas prices	2.494	0.421	1.490	4.280	Finanzen.net (2019a)	Daily frequency

for the optimisation. We decided to use the Hyperband algorithm based on Li et al. (2018), because it combined the advantages of the Random Search approach, and managed the computational resources more effectively. The optimisation was conducted across all training samples separately but with the same set of parameters, thus the best hyperparameters minimise the overall group error. These parameters were, next the neurons of each layer, also stabilisation parameters such as dropouts, recurrent dropouts and kernel regularisation. The detailed hyperparameters of the model architecture can be found in Table 7 in Appendix B. Given the fact that each LSTM network differs slightly, due to the random initialisation of the weights, an additional measure was necessary to ensure stable results. For each prediction period, the LSTM model was calculated and predicted 10 times and thereafter the median value was used as the corresponding forecast. Although the architecture differed slightly, the selection process regarding the CNN-LSTM model was similar to the LSTM network. In order to describe both the hourly time features as well as weekly patterns, the CNN-LSTM consisted of two branches that were concatenated in the final layer. In the first branch, we constructed the CNN model, in which two times a Convolutional layer followed by a Max-pooling layer was implemented. After that, the results were combined in a Dense layer and fed into the final layer for prediction. The second branch of the model consisted of the LSTM structures similar to the LSTM model. However, we decided to implement only two LSTM layers to not over-inflate the model. Similar to the CNN gate, we then included a Dense layer that fed the information into the last layer of the model. Concerning the input variables, it is important to note that we include both the explanatory variables (X_t) and the lagged spot price in the LSTM branch. In the CNN branch, only the explanatory variables were incorporated. With respect to the defined architecture, we split the training data again by 25% and optimised the hyperparameter based on the Hyperband algorithm (Li et al., 2018). The detailed results can be found in Table 8 in Appendix B. Similar to the LSTM approach, we recalculated and predicted each period ten times and took the median values as predictors to ensure stable results.

We want to point out that for both neural network models, a time lag of 48 h was chosen for the input data (denoted as Time Steps in Tables 7 and 8). This decision is based on three reasons. First, the lag corresponds to the lag of the (S)ARIMA(X) model; thus, a fair comparison between the models is enabled. Second, with the lag of 48 h we include daily patterns into the model, which can be utilised by the LSTM layers of the models. Third, preliminary analysis with other lags showed a clear improvement across the neural networks, when considering a larger training horizon. For example, the networks improved their accuracy with the shift from lag 12 h to lag 48 h for both the RMSE and MAE error metric. In terms of the RMSE error metric, the reduction achieved by the LSTM model was between 11.7% up to 49.1% depending on the forecasting period. Similarly, the CNN-LSTM achieved a reduction between 2.9% up to 25.2% for the RMSE. Therefore, the larger lag is clearly advantageous for the neural networks. Lastly, we report the hyperparameter of the $Ensemble_{LSTM}$. Generally, the model was constructed identically to the LSTM with mostly similar hyperparameters in comparison to the LSTM, as provided in Table 7. However, the model was solely trained on the in-sample predictions from the two best forecasting models, i.e., predictions on the training data. Thus, taking into account that only two inputs were available for the prediction, it was necessary to conduct two minor adjustments in the hyperparameters.[5] First, we dispensed the recurrent dropout as well as the shuffling of the validation data because it showed no clear improvement. Second, only a *lag* of 6 steps backwards in time was used to ensure a short term perspective of the forecast.

3.4. Performance metric definitions

In order to compare the forecasting performance of the models, we implemented three different performance metrics in this work: the Root Mean Square Error (RMSE), the Mean Absolute Error (MAE), and the Frequency of Smallest Absolute Prediction Error (FSAPE). Given the true values y_i and the forecasted values $\hat{y}_i$ with $i = 1, \dots, N$, observations of the forecast period, we define the residuals e_i with $e_i = y_i - \hat{y}_i$. Respectively, the RMSE is defined as the total sum of the squared residuals:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} e_i^2} \tag{21}$$

However, Willmott and Matsuura (2005) argued that the RMSE overstates larger errors. Under the consideration that the spot price includes large outliers as well, one can assume that the error measurement penalises predictions with large outliers. Therefore, we further include the MAE as the second performance metric as a sum of the absolute values of the residuals:

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |e_i| \tag{22}$$

As a third performance measurement, we introduce a relative measurement that directly compares the prediction results of the forecasting models. Under the consideration of $J = 4$ models and their respective residuals $e_i j$, the FSAPE is defined for the model j as follows:

$$FSAPE_j = \frac{1}{N}\sum_{i=1}^{N} a_{ij} with\ a_{ij} = \begin{cases} 1 & if\ |e_{ij}| = \min(|e_{ij}|)\ \forall\ j \\ 0 & all\ else \end{cases} \tag{23}$$

Accordingly, the FSAPE performance metric displays the proportion of the smallest absolute forecasting residuals, where a high percentage indicates an outperforming of the other candidates. In comparison with the MAE and RMSE, the advantage of the FSAPE metric is that it examines the residuals in a binary approach. Therefore, the FSAPE is not directly influenced by any scaling or outliers of the forecast data if the other forecasting models are also not able to detect the outlier.

4. Forecasting results

4.1. Exemplary forecast periods

Before the general model comparison, we first want to give an exemplary overview of the spot price time series and the forecast performance of the models. We selected three distinct prediction periods that differ in their characterisation and outline these periods in Fig. 4. Overall, the depicted three periods allow insights into various

[5] Note that these improvements were tested on the validation data.

degrees of forecasting accuracy of the four models. While the price development and structure of the German day-ahead spot price was adequately forecasted in period *Apr18* (Fig. 4c), the periods *Oct17* (Fig. 4a) and *Jan18* (Fig. 4b) show larger deviations between the true EPEX SPOT prices and the forecasted prices across all models. In the case of the *Oct17* period, a large forecasting loss was induced by the negative outliers at the end of the month. The most plausible cause for this variation could have been the extreme storm front *Herwart* that occurred from October 28th until October 30th (Deutscher Wetterdienst (DWD), 2018). As a result, a rapid price drop led to the largest negative price in our data set with −83.06 €/MWh. As one could see in Fig. 4a, none of the forecasting models captured this price drop to its full extent. In contrast, the prediction errors during the *Jan18* period can be largely attributed to the variability of the spot price during the first half of the month. Within the first two weeks, both large negative and positive values of the spot price were observed, resulting in larger forecasting losses, especially for the time series models. Again, some explanation of the negative prices can be found in the storm front *Burglind* on January 2nd and January 3rd (Deutscher Wetterdienst (DWD), 2018); however, extreme weather events cannot fully explain the positive differences to the results of the forecasts. Note that even though the models did not perform perfectly in the *Sep17* and *Jan18* period, it should still be emphasised that we did not use a rolling window forecast approach and updated the spot price. Instead, we only provided the explanatory variables in the forecast interval. Thus, no model received any information on the spot price, and the autoregressive components could only process previously forecasted values. Given this research setting, one would expect worse results in later observations. However, none of the three figures indicates a structural deterioration in terms of predictive accuracy. Therefore, the conclusion may be drawn that the exogenous variables determine a large proportion of the spot price.

After the visual analysis, we further want to review the performance in the context of the RMSE and MAE metrics to reassure the previous deductions. The results of all performance metrics are documented in Tables 2, 3 and 4 for the RMSE, MAE and FSAPE, respectively. Compared with the other two intervals, it is clear to see that all models were able to perform more accurately within the *Apr18* interval. This is visible in terms of both RMSE and MAE results, with all models showing more favourable performance metrics in comparison with their overall mean. This holds especially for the two-stage VAR model, which showed the best performance in this period. In contrast, all models show less accuracy in terms of the *Oct17* and *Jan18* interval. It is interesting to see that in the *Oct17* interval the VAR and LSTM perform similarly, with the VAR obtaining the smaller MAE and the LSTM the smaller RMSE metric. The large outlier might induce this difference at the end of the period, which is penalised to a larger extent by the RMSE. Nevertheless, the remaining observations in *Oct17* were predicted quite well. However, with regard to the *Jan18* interval, we observe a clear disadvantage of the time series models, which both have their largest MAE in this period. This result clearly corresponds to the visual analysis, where especially the VAR results of the first weeks are constantly below the spot price.

4.2. General model performance

After the first analysis of the exemplary periods, we now focus on the results for all periods documented in Tables 2–4. Overall, there is a close call between the LSTM and the two-stage VAR model. On the one hand, the LSTM model surpasses all other candidates in all three performance metrics. Depending on the performance metric, the LSTM was the best performing model in six (RMSE) or five (MAE and FSAPE) forecasting periods. On the other hand, the two-stage VAR follows closely behind with the second-best results, i.e., it obtains the most favourable performance statistics for four periods (across all performance metrics). While one model shows the best performance in many cases, the other model has the second-best outcome. These close

Table 2
Score of the RMSE performance metric on all twelve monthly forecasting periods (720 observations), as well as the average for each model. The right column displays the best performing model for the respective period.

Period	(S)ARIMA(X)	VAR	LSTM	CNN-LSTM	Best performance
Oct17	19.034	15.072	13.146	17.153	LSTM
Nov17	13.458	9.847	11.249	13.332	VAR
Dec17	14.617	9.410	10.662	11.030	VAR
Jan18	18.031	18.213	11.937	13.447	LSTM
Feb18	9.157	9.442	6.686	8.612	LSTM
Mar18	10.888	9.907	7.574	10.232	LSTM
Apr18	6.633	6.149	6.938	7.931	VAR
Mai18	10.568	9.349	8.520	9.850	LSTM
Jun18	8.160	6.363	5.812	6.299	LSTM
Jul18	8.596	5.485	5.885	5.029	CNN-LSTM
Aug18	10.044	7.296	9.083	7.410	VAR
Sep18	10.228	11.743	11.889	11.623	(S)ARIMA(X)
Average	11.618	9.856	9.115	10.162	LSTM

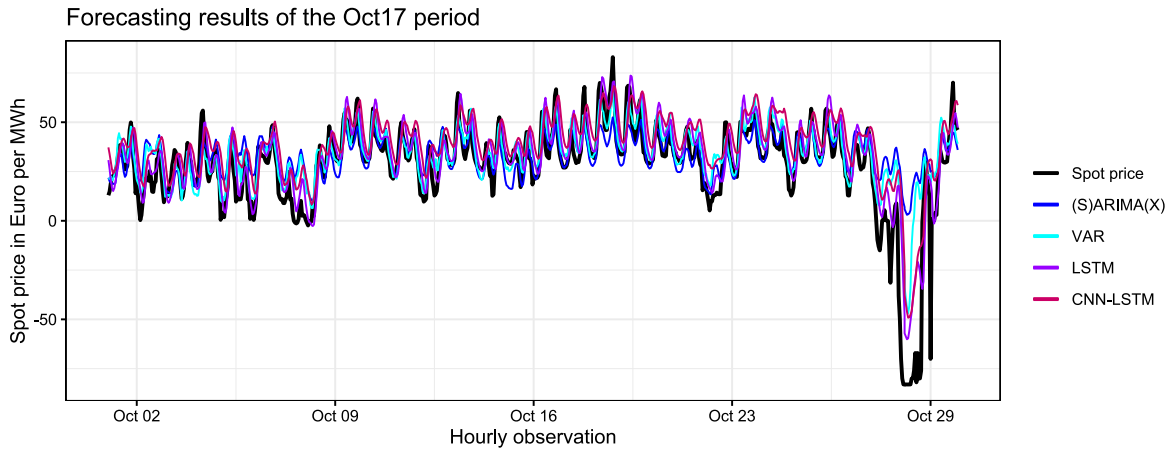
results are also visible in the averages of the performance metrics. Here, the LSTM had better performance metrics for the RMSE and MAE, while the VAR scored most favourable in terms of the FSAPE metric. This is especially interesting, considering that the FSAPE is not influenced by outliers. Instead, the FSAPE depicts the forecasting performance in relation to the other forecasting models. With similar FSAPE scores, but different RMSE and MAE results, one could argue that the two models have an equal proportion of the best individual predictions, however, the LSTM predicts with smaller deviations from the original data. In regard to specific forecast periods, the LSTM performed exceedingly well for the first half of 2018, with the exception of *Apr18*. On the other hand, the VAR showed good results for the *Oct17-Dec17* interval. As a potential explanation, some of the performance might be induced by seasonal structural differences. In general, one can say that the LSTM ranks slightly better than the VAR model, however, one could not detect a clearly superior performance of the LSTM.

Regarding the remaining two forecasting models, the CNN-LSTM and the (S)ARIMA(X) models show very similar predictive accuracy, with a slightly better performance of the CNN-LSTM. While the FSAPE average is similar for both models, the RMSE and MAE average performance metrics show a larger difference in favour of the CNN-LSTM. The average results are further supported when consulting the individual forecasting intervals. The CNN-LSTM achieves the best performance for the intervals *Jul18* and *Aug18*. Within the *Jul18* period the CNN-LSTM shows the most favourable overall RMSE and MAE scores. Finally, the (S)ARIMA(X) models did not capture the spot price to the same extent as competing approaches which is especially reflected in average performance metrics. Nevertheless, conditional on the *Sep18* period the (S)ARIMA(X) outperformed the remaining models for at least one sample. In sum, the CNN-LSTM is ranked third and the (S)ARIMA(X) model last.

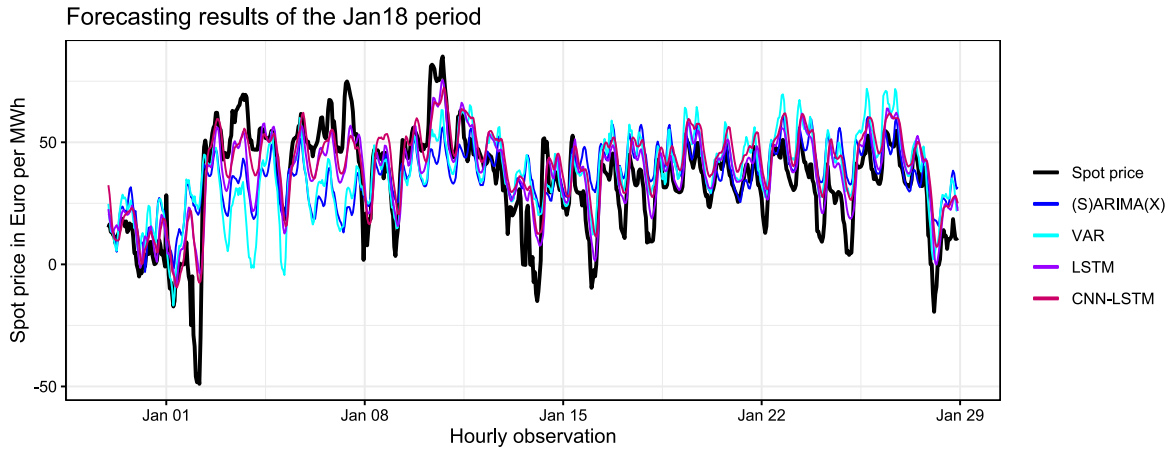
4.3. Shorter forecasting horizon

Next to the monthly forecasts (720 observations), we considered a one day-ahead forecast (24 observations) and a seven day-ahead forecast (168 observations) for each prediction period, due to the fact that shorter forecasting horizons have been considered in related studies, e.g., see Gürtler and Paulsen (2018). In Table 5 we outline the average model performance for each prediction horizon. As expected, reducing the number of observations translates into lower prediction loss averages for almost all models. However, it is worth noticing that the neural networks only partly improved their performance at shorter time horizons. Regarding the results of the LSTM model, we can observe that the MAE and FSAPE_{single} metric of the one day-ahead forecasts are worse in direct comparison with the 30 days-ahead forecasts.[6] In

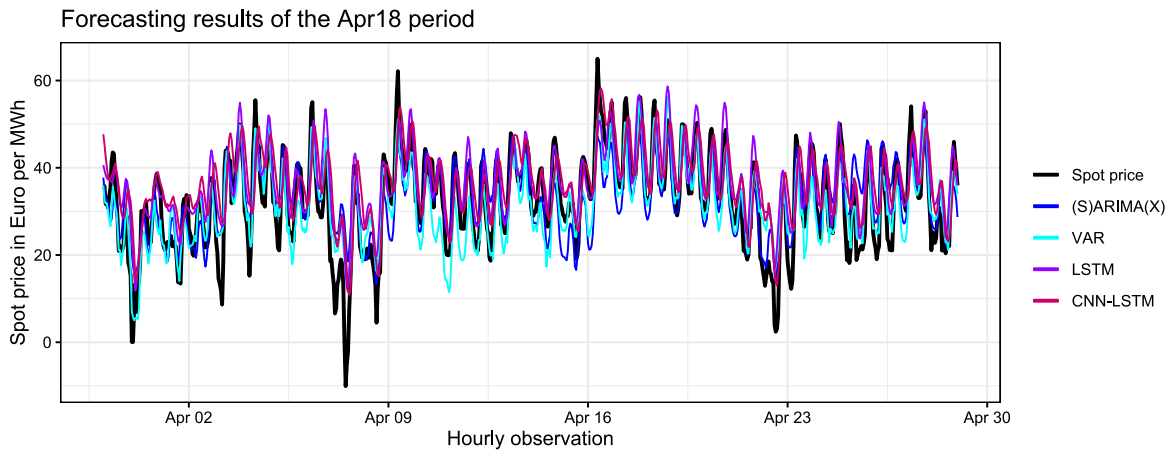
[6] Note that at this point we only compare the FSAPE_{single} metric because it does not include the combined models in the computation.



(a) Forecast Results of Oct17 period



(b) Forecast Results of Jan18 period



(c) Forecast Results of Apr18 period

Fig. 4. Forecasting results for three different monthly periods. The black line indicates the observed EPEX SPOT prices, while the coloured lines describe results of the different models. Each prediction batch consists of 720 hourly observations. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

contrast, the time series models, especially the VAR, are more precise at shorter prediction horizons. This materialises in an increase of $FSAPE_{single}$ statistics for the VAR approach for seven and one day-ahead forecasts, while respective frequencies assigned to the neural networks decrease. Furthermore, for one day-ahead forecasts the VAR approach shows most favourable RMSE and MAE results. Overall, one could draw two interpretations, based on these differences of results conditioning on alternative forecasting horizons. Firstly, we see that the prediction horizon is an important determinant of model performance. With different length choices across the literature, e.g., seen in Gürtler and Paulsen (2018), this might explain the heterogeneous results in the research community. Secondly, the dependence of predictive performance on the forecast horizon suggests a possible non-linear relationship of the spot price, given that the linear models decrease in their performance for the longer-term forecast.

Table 3
Score of the MAE performance metric on all twelve monthly forecasting periods (720 observations) as well as the average for each model. The right column displays the best performing model for the respective period.

Period	(S)ARIMA(X)	VAR	LSTM	CNN-LSTM	Best performance
Oct17	10.514	8.391	9.327	13.154	VAR
Nov17	9.477	7.216	9.332	11.119	VAR
Dec17	10.841	6.711	7.862	8.388	VAR
Jan18	13.836	14.342	9.400	11.050	LSTM
Feb18	7.204	7.678	4.380	6.554	LSTM
Mar18	7.642	7.553	5.235	8.174	LSTM
Apr18	5.088	4.977	5.558	6.442	VAR
Mai18	6.728	6.866	6.430	7.526	LSTM
Jun18	6.057	4.968	4.469	4.712	LSTM
Jul18	7.269	4.483	4.971	3.841	CNN-LSTM
Aug18	8.481	5.827	7.643	5.648	CNN-LSTM
Sep18	7.897	9.434	9.147	9.059	(S)ARIMA(X)
Average	8.419	7.370	6.980	7.972	LSTM

Table 4
Score of the FSAPE performance metric on all twelve monthly forecasting periods (720 observations) as well as the average for each model. The right column displays the best performing model for the respective period.

Period	(S)ARIMA(X)	VAR	LSTM	CNN-LSTM	Best performance
Oct17	0.274	0.342	0.281	0.104	VAR
Nov17	0.319	0.393	0.196	0.092	VAR
Dec17	0.167	0.415	0.235	0.183	VAR
Jan18	0.232	0.153	0.371	0.244	LSTM
Feb18	0.146	0.171	0.451	0.232	LSTM
Mar18	0.201	0.224	0.365	0.210	LSTM
Apr18	0.308	0.328	0.199	0.165	VAR
Mai18	0.287	0.210	0.307	0.196	LSTM
Jun18	0.181	0.226	0.300	0.293	LSTM
Jul18	0.136	0.254	0.211	0.399	CNN-LSTM
Aug18	0.156	0.321	0.138	0.386	CNN-LSTM
Sep18	0.347	0.232	0.192	0.229	(S)ARIMA(X)
Average	0.230	0.272	0.270	0.228	LSTM

4.4. Combined forecasting models

As outlined in Section 2.5, we conduct in the last step an additional analysis on the consolidation of the different forecasting approaches to identify their potential for ensemble forecasting as proposed in Neupane et al. (2017). With two neural networks and two-time series approaches, this would result in four different model combinations. However, given the distinct performance characteristics of the two single prediction models, we are primarily interested in the performance of their joint forecast, thus in this work, we only combine the VAR and LSTM models. Further, again note that we trained the $Ensemble_{LSTM}$ solely with the fitted values of the VAR and LSTM and that the $Ensemble_{LSTM}$ received only the forecasted values of the single models. Thus, we ensure that no external information influenced the ensemble predictors. The results of the combined forecasting models for all three periods are documented in Table 5. With respect to the results, one can see that the combination of the two models did improve the forecasting performance. Both the $Combined_{naive}$ and $Ensemble_{LSTM}$ outperformed the single VAR and LSTM model in terms of the MAE and RMSE.

In particular, the results of the $Ensemble_{LSTM}$ are interesting, considering that it performed better than the naive approach, even though no additional data was given. Considering the individual episodes, the combined forecasts outperformed the single models in 11 out of 12 episodes.[7] However, with regard to FSAPE results, the combined forecasting models do not outperform the constituent models, especially

[7] The $Ensemble_{LSTM}$ achieved a total of 7 (MAE) and 8 (RMSE) best performances and the $Combined_{naive}$ 4 and 3, respectively.

Table 5
Aggregated results of the long, medium and short-term forecasting horizons for the four single models as well as the combined forecasting approaches from Section 4.4. Considering that the FSAPE is not invariant with respect to the inclusion of additional models, we report two FSAPE values. In $FSAPE_{single}$ the $Combined_{naive}$ and $Ensemble_{LSTM}$ have been excluded in the computation, while in $FSAPE_{comb}$ they were included. The first sub-table corresponds to the monthly results (720 observations) as outlined in more detail in Tables 2 and 3 as well as Table 4 for the $FSAPE_{single}$. For the second and third sub-table, a seven day-ahead (168 observations) and one day-ahead (24 observations) forecast were computed for each of the twelve prediction periods, and the results have been averaged.

Thirty day-ahead forecast						
Single models					Combined forecasting models	
	(S)ARIMA(X)	VAR	LSTM	CNN-LSTM	$Combined_{naive}$	$Ensemble_{LSTM}$
RMSE	11.618	9.856	9.115	10.162	8.455	7.917
MAE	8.420	7.370	6.980	7.972	6.272	5.919
$FSAPE_{single}$	0.230	0.272	0.270	0.228	–	–
$FSAPE_{comb}$	0.178	0.196	0.185	0.184	0.095	0.162
Seven day-ahead forecast						
Single models					Combined forecasting models	
	(S)ARIMA(X)	VAR	LSTM	CNN-LSTM	$Combined_{naive}$	$Ensemble_{LSTM}$
RMSE	10.296	8.960	8.401	9.404	7.637	7.152
MAE	7.990	6.931	6.500	7.362	5.850	5.444
$FSAPE_{single}$	0.220	0.288	0.267	0.225	–	–
$FSAPE_{comb}$	0.165	0.202	0.176	0.180	0.102	0.176
One day-ahead forecast						
Single models					Combined forecasting models	
	(S)ARIMA(X)	VAR	LSTM	CNN-LSTM	$Combined_{naive}$	$Ensemble_{LSTM}$
RMSE	8.220	6.598	8.791	9.166	6.519	6.409
MAE	6.819	5.491	7.276	7.486	5.363	5.173
$FSAPE_{single}$	0.253	0.340	0.194	0.212	–	–
$FSAPE_{comb}$	0.174	0.264	0.111	0.153	0.111	0.188

for the case of 30 days-ahead forecasts (see Table 11 in the Appendix). One possible explanation for this performance difference could be the binary nature of the FSAPE metric, where the individual predictions closest to the true observations are counted. One can assume that the combined forecast values are largely located between the VAR and LSTM values. Therefore, it is unlikely that the combined models are closer to the true values if both single models over- or underestimate in the electricity price. Nevertheless, the outstanding $Ensemble_{LSTM}$ performance in terms of MAE and RMSE shows that the combination of the two forecasts improves the overall predictive performance and assigns additional value to the constituent forecasts.

5. Discussion

Regarding the overall forecasting results, one has to acknowledge that the neural network approaches outperformed the time series models in seven out of twelve prediction samples and the LSTM had the most satisfactory performance metrics on average. Nevertheless, while the LSTM showed the best forecasting results in the majority of prediction samples, the two-stage VAR turned out to provide the second-best EPF, surpassing the CNN-LSTM. Furthermore, the VAR provides better results in terms of the FSAPE metric and for issues of short-term prediction. Lastly, through the combination of the two models, we were able to show that both these models seem to have advantages for the provision of spot price forecasts at the German EPEX SPOT market.

Given the empirical evidence, the results fortify the usage of LSTM based models for the EPF and support the recent findings of other researchers, such as Lago et al. (2018), Ugurlu et al. (2018), Peng et al. (2018) and Liu et al. (2020). In this regard, we especially want to call attention to Lago et al. (2018), which provide a detailed analysis of different EPF approaches and document a dominance of neural network approaches in their study. We were able to support

these findings by showing that the LSTM is superior to the simple (S)ARIMA(X) approach. Overall, the non-linear approach of the LSTM seems to depict the relationship between the spot price and the external variables while simultaneously describing the dynamic relation with past values adequately. We also conducted other model runs without including past spot price information to observe that the exclusion of historical price information deteriorated the performance significantly.[8] Moreover, we also confirm the findings of Lago et al. (2018) that hybrid CNN-LSTM models do not perform as well as simple LSTM, which contradicts the results of Kuo and Huang (2018). In terms of the network structure, our results align with Lago et al. (2018) as deeper neural networks and a larger number of neurons increase the forecasting performance. However, with the results of the state-of-the-art hyperparameter optimisation, we recommend dropouts, recurrent dropouts, and kernel regularisation to ensure network stability. In addition, given the marked fluctuations in the predictions of the networks, we advise to train multiple LSTMs and subsequently average the results to increase the stability even further.

Concerning the time series models, we further want to highlight a possible reason for the two-stage VAR-model to outperform both the CNN-LSTM as well as the (S)ARIMA(X) model. Given the structural similarities between the (S)ARIMA(X) model and the VAR model, the multivariate framework might be an explanation for the superior performance over the (S)ARIMA(X).[9] While this conclusion might not be intuitive at first, it can be justified by the structure of the spot market. As mentioned in Section 1, all German spot prices of the next day are auctioned at 12.00 pm at the EPEX SPOT. As a result, the prices are determined simultaneously instead of consecutively. Thus, the market participants integrate their expectations and knowledge of the other prices of the day into their bidding process. The VAR approach has the potential to capture this relationship and describes the variables simultaneously with an hourly resolution at the daily frequency. Hence, one could argue that a multivariate framework such as the two-stage VAR might actually depict the market structures to a better extent and as a result, enhance the accuracy of the EPF. While other researchers (Gianfreda et al., 2020; Ziel and Weron, 2018; Ziel et al., 2015) have already applied the VAR for spot price forecasting, we have extended their approach with a first step OLS regression. We thereby include the information of the explanatory variables.

By combining best-performing forecasting models, we demonstrated that the time series VAR approach and the neural network LSTM approach complement each other. Even though our combined $Ensemble_{LSTM}$ model only consisted of two predictions and we did not employ any further fine-tuning, such as the fallback mechanism of Neupane et al. (2017), forecast combinations improved upon the performance of the single constituent models. In addition, the comparison between the $Combined_{naive}$ and $Ensemble_{LSTM}$ showed that there might be distinct differences between the single models in terms of their forecasting strategy, which the $Ensemble_{LSTM}$ was able to exploit. As a result, we recommend additional EPF research in the combination of different forecasting approaches. Given the performance of the LSTM, a combination of a multivariate framework and a neural network might offer new possibilities for the EPF.

Finally, we shortly want to address the importance of the research environment, i.e., the different spot markets. While (Lago et al., 2018; Ugurlu et al., 2018) considered the Belgian and the Turkish electricity market, respectively, we obtained similar results for the German spot market, even though we had a stronger focus on renewable energies. This shows the flexibility and applicability of neural networks, but at the same time makes comparability rather difficult. Therefore, an open-access benchmark, similar to other fields of machine learning, should be discussed and supported among the EPF-research community.[10] One very promising open-source candidate was proposed by Lago et al. (2020), which includes six years of data for five different spot price markets (Nordpool, Pennsylvania-New Jersey-Maryland, Belgium, France and Germany), and distinct baseline models in order to compare newly proposed approaches. Thus, as a future research outlook, it might be interesting to test both the LSTM and the two-stage VAR on other spot price markets against the baseline models of Lago et al. (2020).

6. Conclusion

This work provides a comparative study aiming to identify one most favourable model out of four prominent approaches to EPF. These approaches consisted of two time series models, one univariate (S)ARIMA(X) and a two-stage multivariate VAR model, and two neural networks, i.e. the LSTM and the CNN-LSTM model. We conducted the EPF on the German spot price market, with various external features added to the analysis. Thereby, the prediction interval was divided into a total of twelve monthly periods to analyse the potential of seasonal effects on the performance of the forecasting models. Based on this framework, the single models were optimised by means of state-of-the-art approaches and individually fitted for each prediction period. Finally, a combination of the neural network and the multivariate model was presented.

We could identify two models that showed promising results across the performance metrics on average in terms of the overall performance. On the one hand, the LSTM model outperformed the other approaches in terms of most of the performance metrics, which is following recent publications in the field of EPF. However, the multivariate VAR approach followed closely and indicated a similar predictive accuracy in the direct comparison with the LSTM. Further, the time series model seemed especially suited for predicting shorter day-ahead horizons such a one day. In addition, by combining our best-performing forecasting models, we demonstrated that the time series VAR approach and the neural network LSTM approach complement each other and achieved the best forecasting performance in comparison with the single models. In line with that, future research should assess further combinations of forecasting models and methods to take advantage of the strengths of different prediction classes. Given that the VAR was the only multivariate framework in this analysis, we can recommend further research, especially in this direction, both for time series models as well as neural network approaches.

CRediT authorship contribution statement

Malte Lehna: Writing – original draft, Conceptualisation, Methodology, Software, Resources, Formal analysis. **Fabian Scheller:** Writing – review & editing, Conceptualisation, Visualisation, Validation. **Helmut Herwartz:** Supervision, Conceptualisation, Writing – review & editing.

Acknowledgement

For valuable feedback, the authors wish to thank two anonymous reviewers and the Associate Editor. This work was supported by Fraunhofer Cluster of Excellence Integrated Energy Systems *CINES*.

[8] The results of the other model runs are included in the supplementary data and are not further discussed.

[9] The (S)ARIMA(X) model and the VAR model have similar structures in especially two characteristics. First, both the (S)ARIMA(X) and the two-stage VAR incorporate a linear relationship between the spot price and the explanatory variables. Furthermore, both models incorporate autoregressive structures. In terms of the (S)ARIMA(X) model, the influence is included through the seasonal lagged components. In contrast, the two-stage VAR model comprises the seasonality by stacking hourly observations with the 24 variable frameworks.

[10] A good example are the Atari 2600 games as a benchmark for reinforcement learning, e.g., see Mnih et al. (2015).

Table 6
Characterisation of the forecasting intervals as well as the training intervals. Each forecast interval had a fixed batchsize of 30 days for the forecast and the previous 360 days for training period. The hourly changes (CEST and CET) come from the day light saving in March and October.

Forecast period	Name	First observation	Last observation	Trainings period
1	Oct17	2017-10-01 00:00:00 CEST	2017-10-30 22:00:00 CET	2016-10-05 00:00:00 CEST - 2017-09-30 23:00:00 CEST
2	Nov17	2017-10-30 23:00:00 CET	2017-11-29 22:00:00 CET	2016-11-04 23:00:00 CET - 2017-10-30 22:00:00 CET
3	Dec17	2017-11-29 23:00:00 CET	2017-12-29 22:00:00 CET	2016-12-04 23:00:00 CET - 2017-11-29 22:00:00 CET
4	Jan18	2017-12-29 23:00:00 CET	2018-01-28 22:00:00 CET	2017-01-03 23:00:00 CET - 2017-12-29 22:00:00 CET
5	Feb18	2018-01-28 23:00:00 CET	2018-02-27 22:00:00 CET	2017-02-02 23:00:00 CET - 2018-01-28 22:00:00 CET
6	Mar18	2018-02-27 23:00:00 CET	2018-03-29 23:00:00 CEST	2017-03-04 23:00:00 CET - 2018-02-27 22:00:00 CET
7	Apr18	2018-03-30 00:00:00 CEST	2018-04-28 23:00:00 CEST	2017-04-04 00:00:00 CEST - 2018-03-29 23:00:00 CEST
8	Mai18	2018-04-29 00:00:00 CEST	2018-05-28 23:00:00 CEST	2017-05-04 00:00:00 CEST - 2018-04-28 23:00:00 CEST
9	Jun18	2018-05-29 00:00:00 CEST	2018-06-27 23:00:00 CEST	2017-06-03 00:00:00 CEST - 2018-05-28 23:00:00 CEST
10	Jul18	2018-06-28 00:00:00 CEST	2018-07-28 23:00:00 CEST	2017-07-03 00:00:00 CEST - 2018-06-27 23:00:00 CEST
11	Aug18	2018-07-29 00:00:00 CEST	2018-08-27 23:00:00 CEST	2017-08-02 00:00:00 CEST - 2018-07-28 23:00:00 CEST
12	Sep18	2018-08-28 00:00:00 CEST	2018-09-27 23:00:00 CEST	2017-09-01 00:00:00 CEST - 2018-08-27 23:00:00 CEST

Table 7
The table describes the hyperparameter of the LSTM model. Note that the time step variable of the LSTM was selected by the authors. The remaining hyperparameters were optimised through the Hyperband algorithm. In order to restrict over fitting of the model, we further implemented an early stopping criterion which stops the calculation if no improvement of the validation loss occurred in the last 20 observations.

Hyperparameters LSTM		
1. LSTM layer	Number of neurons	50
2. LSTM layer	Number of neurons	100
3. LSTM layer	Number of neurons	30
LSTM time steps	Look-back (fixed)	48
Dense layer	Number of neurons	1
Dropout	For all LSTM Layers	0.03
Recurrent dropout	For all LSTM Layers	0.0001
Kernel regulariser	L1 and L2	0 and 0.000001
Batch size		360
Epochs	(max)	65
Early stopping	On val_loss	20

Table 8
The table describes the hyperparameter of the CNN-LSTM. Note that the time step variable of the LSTM and CNN were selected by the authors. The remaining hyperparameters were optimised through the Hyperband algorithm. In order to restrict over fitting of the model, we further implemented an early stopping criterion which stops the calculation if no improvement of the validation loss occurred in the last 20 observations.

Hyperparameters CNN-LSTM		
1. 1D convolution layer	Filter size	64
	Kernel size	28
2. 1D Convolution layer	Filter size	32
	Kernel size	14
1.+2. 1D max pooling	Pooling size	2
CNN dense layer	Number of neurons	24
CNN time steps	Look-back (fixed)	168
1. LSTM layer	Number of neurons	55
2. LSTM layer	Number of neurons	30
LSTM dense layer	Number of neurons	24
LSTM time steps	Look-back (fixed)	48
Dropout	For all LSTM layers	0.3
Recurrent dropout	For all LSTM layers	0.3
Kernel regulariser	L1 and L2	0
Final dense layer	Number of neurons	1
Batch size		360
Epochs	(max)	65
Early stopping	On val_loss	20

Appendix A. Overview over forecast horizon

See Table 6.

Appendix B. Hyperparameter results

See Tables 7 and 8.

Table 9
Comparison of the RMSE results between the best performing single models VAR and LSTM and the ensemble models $Combined_{naive}$ and $Ensemble_{LSTM}$. The (S)ARIMA(X) and the CNN-LSTM model were excluded for display purposes.

	$Combined_{naive}$	$Ensemble_{LSTM}$	VAR	LSTM	Best
Oct17	13.277	11.475	15.072	13.146	$Ensemble_{LSTM}$
Nov17	9.564	8.760	9.847	11.249	$Ensemble_{LSTM}$
Dec17	9.204	8.433	9.410	10.662	$Ensemble_{LSTM}$
Jan18	14.519	13.644	18.213	11.937	LSTM
Feb18	7.216	6.032	9.442	6.686	$Ensemble_{LSTM}$
Mar18	8.086	6.977	9.907	7.574	$Ensemble_{LSTM}$
Apr18	5.639	6.076	6.149	6.938	$Combined_{naive}$
Mai18	7.939	8.079	9.349	8.520	$Combined_{naive}$
Jun18	5.520	5.408	6.363	5.812	$Ensemble_{LSTM}$
Jul18	4.589	5.059	5.485	5.885	$Combined_{naive}$
Aug18	7.018	6.490	7.296	9.083	$Ensemble_{LSTM}$
Sep18	8.891	8.573	11.743	11.889	$Ensemble_{LSTM}$
Average	8.455	7.917	9.856	9.115	

Table 10
Comparison of the MAE results between the best performing single models VAR and LSTM and the ensemble models $Combined_{naive}$ and $Ensemble_{LSTM}$. The (S)ARIMA(X) and the CNN-LSTM model were excluded for display purposes.

	$Combined_{naive}$	$Ensemble_{LSTM}$	VAR	LSTM	Best
Oct17	8.284	7.172	8.391	9.327	$Ensemble_{LSTM}$
Nov17	7.635	7.054	7.216	9.332	$Ensemble_{LSTM}$
Dec17	6.559	6.063	6.711	7.862	$Ensemble_{LSTM}$
Jan18	11.526	10.823	14.342	9.400	LSTM
Feb18	5.362	4.281	7.678	4.380	$Ensemble_{LSTM}$
Mar18	5.871	4.990	7.553	5.235	$Ensemble_{LSTM}$
Apr18	4.354	4.723	4.977	5.558	$Combined_{naive}$
Mai18	5.606	6.059	6.866	6.430	$Combined_{naive}$
Jun18	4.244	4.287	4.968	4.469	$Combined_{naive}$
Jul18	3.690	4.143	4.483	4.971	$Combined_{naive}$
Aug18	5.670	5.188	5.827	7.643	$Ensemble_{LSTM}$
Sep18	6.467	6.250	9.434	9.147	$Ensemble_{LSTM}$
Average	6.272	5.919	7.370	6.980	

Appendix C. Combined model results

See Tables 9–11.

Appendix D. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.eneco.2021.105742.

Table 11
FSAPE performance of all twelve forecasting periods including the combined model results. Under consideration that the error metric is a relative measure, the inclusion of the two models also influences the metric of the single models. As a consequence, the results differ from those in Table 4.

	(S)ARIMA(X)	VAR	LSTM	CNN-LSTM	$Combined_{naive}$	$Ensemble_{LSTM}$	Best Performance
Oct17	0.250	0.258	0.190	0.067	0.065	0.169	VAR
Nov17	0.296	0.286	0.143	0.064	0.068	0.143	(S)ARIMA(X)
Dec17	0.125	0.307	0.169	0.158	0.092	0.149	VAR
Jan18	0.214	0.104	0.296	0.207	0.035	0.144	LSTM
Feb18	0.089	0.112	0.304	0.174	0.086	0.235	LSTM
Mar18	0.151	0.164	0.244	0.179	0.081	0.181	LSTM
Apr18	0.225	0.243	0.117	0.121	0.138	0.157	VAR
Mai18	0.236	0.149	0.211	0.147	0.111	0.146	(S)ARIMA(X))
Jun18	0.153	0.174	0.194	0.254	0.089	0.136	CNN-LSTM
Jul18	0.078	0.204	0.131	0.322	0.125	0.140	CNN-LSTM
Aug18	0.096	0.211	0.090	0.322	0.099	0.182	CNN-LSTM
Sep18	0.225	0.138	0.125	0.193	0.153	0.167	(S)ARIMA(X)
Average	0.178	0.196	0.185	0.184	0.095	0.162	0

References

AGEE-Stat Umweltbundesamt, 2019. Erneuerbare energien in deutschland daten zur entwicklung im jahr 2018. URL https://www.umweltbundesamt.de/sites/default/files/medien/1410/publikationen/uba_hgp_eeinzahlen_2019_bf.pdf.

Ardabili, S., Mosavi, A., Várkonyi-Kóczy, A.R., 2019. Advances in machine learning modeling reviewing hybrid and ensemble methods. In: International Conference on Global Research and Education. Springer, pp. 215–227.

Best, D., Roberts, D., 1975. Algorithm AS 89: the upper tail probabilities of Spearman's rho. J. R. Stat. Soc. C 24 (3), 377–379. http://dx.doi.org/10.2307/2347111.

Box, G.E.P., Jenkins, G.M., Reinsel, G.C., 2008. Time Series Analysis: Forecasting and Control. John Wiley & Sons Inc. Publication.

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

Cludius, J., Hermann, H., Matthes, F.C., Graichen, V., 2014. The merit order effect of wind and photovoltaic electricity generation in Germany 2008–2016: Estimation and distributional implications. Energy Econ. 44, 302–313. http://dx.doi.org/10.1016/j.eneco.2014.04.020.

Contreras, J., Espinola, R., Nogales, F.J., Conejo, A.J., 2003. ARIMA models to predict next-day electricity prices. IEEE Trans. Power Syst. 18 (3), 1014–1020. http://dx.doi.org/10.1109/TPWRS.2002.804943.

Crespo Cuaresma, J., Hlouskova, J., Kossmeier, S., Obersteiner, M., 2004. Forecasting electricity spot-prices using linear univariate time-series models. Appl. Energy 77 (1), 87–106. http://dx.doi.org/10.1016/S0306-2619(03)00096-5.

Deutscher Wetterdienst (DWD), 2018. Rückblick Sturmsaison 2017/2018. URL https://www.dwd.de/DE/wetter/thema_des_tages/2018/9/22.html.

Deutscher Wetterdienst (DWD), 2019a. Open data DWD solar data. URL https://opendata.dwd.de/climate_environment/CDC/observations_germany/climate/hourly/solar/.

Deutscher Wetterdienst (DWD), 2019b. Open data DWD wind data. URL https://opendata.dwd.de/climate_environment/CDC/observations_germany/climate/hourly/wind/historical/.

Dudek, G., 2016. Multilayer perceptron for GEFCom2014 probabilistic electricity price forecasting. Int. J. Forecast. 32 (3), 1057–1060. http://dx.doi.org/10.1016/j.ijforecast.2015.11.009.

EPEX SPOT, 2021. Facts and figures 2021. URL https://www.epexspot.com/en/downloads.

European Comission (EC), 2019. EU emissions trading system (EU ETS). URL https://ec.europa.eu/clima/policies/ets_en.

Finanzen.net, 2019a. Erdgaspreis historisch. URL https://www.finanzen.net/rohstoffe/erdgas-preis-natural-gas/historisch.

Finanzen.net, 2019b. Kohlepreis historisch. URL https://www.finanzen.net/rohstoffe/kohlepreis/historisch.

Fraunhofer-Institut für Umwelt-, Sicherheits- und Energietechnik, 2015a. Maps4Use: Energieproduktion- solarenergie. URL https://maps4use.de/solarenergie/.

Fraunhofer-Institut für Umwelt-, Sicherheits- und Energietechnik, 2015b. Maps4Use: Energieproduktion-windenergie. URL https://maps4use.de/windenergie/.

Gianfreda, A., Ravazzolo, F., Rossini, L., 2020. Comparing the forecasting performances of linear models for electricity prices with high RES penetration. Int. J. Forecast. 36 (3), 974–986. http://dx.doi.org/10.1016/j.ijforecast.2019.11.002.

Gürtler, M., Paulsen, T., 2018. Forecasting performance of time series models on electricity spot markets: a quasi-meta-analysis. Int. J. Energy Sect. Manage. 12 (1), 103–129. http://dx.doi.org/10.1108/IJESM-06-2017-0004.

Haldrup, N., Nielsen, F.S., Ørregaard Nielsen, M., 2010. A vector autoregressive model for electricity prices subject to long memory and regime switching. Energy Econ. 32 (5), 1044–1058. http://dx.doi.org/10.1016/j.eneco.2010.02.012.

Hecht-Nielsen, R., 1992. III.3 - Theory of the backpropagation neural network. In: Wechsler, H. (Ed.), Neural Networks for Perception. Academic Press, pp. 65–93. http://dx.doi.org/10.1016/B978-0-12-741252-8.50010-8.

Hochreiter, S., 1991. Untersuchungen zu Dynamischen Neuronalen Netzen, Vol. 91 (Diploma Thesis). Technische Universität München, (1).

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780. http://dx.doi.org/10.1162/neco.1997.9.8.1735.

Hyndman, R.J., Khandakar, Y., 2008. Automatic time series forecasting: The forecast package for R. J. Stat. Softw. Artic. 27 (3), 1–22. http://dx.doi.org/10.18637/jss.v027.i03.

Karlik, B., Olgac, A.V., 2011. Performance analysis of various activation functions in generalized MLP architectures of neural networks. Int. J. Artif. Intell. Expert Syst. 1 (4), 111–122.

Ketterer, J.C., 2014. The impact of wind power generation on the electricity price in Germany. Energy Econ. 44, 270–280. http://dx.doi.org/10.1016/j.eneco.2014.04.003.

Kingma, D.P., Ba, J., 2017. Adam: A method for stochastic optimization. arXiv:1412.6980.

Kiranyaz, S., Avci, O., Abdeljaber, O., Ince, T., Gabbouj, M., Inman, D.J., 2021. 1D convolutional neural networks and applications: A survey. Mech. Syst. Signal Process. 151, 107398. http://dx.doi.org/10.1016/j.ymssp.2020.107398.

Kostrzewski, M., Kostrzewska, J., 2019. Probabilistic electricity price forecasting with Bayesian stochastic volatility models. Energy Econ. 80, 610–620. http://dx.doi.org/10.1016/j.eneco.2019.02.004.

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

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

Lago, J., Marcjasz, G., Schutter, B.D., Weron, R., 2020. Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark. arXiv:2008.08004.

Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., Talwalkar, A., 2018. Hyperband: A novel bandit-based approach to hyperparameter optimization. arXiv:1603.06560.

Liu, H., Shi, J., 2013. Applying ARMA–GARCH approaches to forecasting short-term electricity prices. Energy Econ. 37, 152–166. http://dx.doi.org/10.1016/j.eneco.2013.02.006.

Liu, S., Zhang, L., Zou, B., 2020. Study on electricity market price forecasting with large-scale wind power based on LSTM. In: 2019 6th International Conference on Dependable Systems and their Applications (DSA). pp. 297–303. http://dx.doi.org/10.1109/DSA.2019.00045.

Lütkepohl, H., 2005. New Introduction to Multiple Time Series Analysis. Springer Science & Business Media.

Maciejowska, K., Nowotarski, J., 2016. A hybrid model for GEFCom2014 probabilistic electricity price forecasting. Int. J. Forecast. 32 (3), 1051–1056. http://dx.doi.org/10.1016/j.ijforecast.2015.11.008.

Makridakis, S., Wheelwright, S.C., Hyndman, R.J., 2008. Forecasting Methods and Applications. John Wiley & Sons Inc. Publication.

Marcjasz, G., Uniejewski, B., Weron, R., 2020. Probabilistic electricity price forecasting with NARX networks: Combine point or probabilistic forecasts? Int. J. Forecast. 36 (2), 466–479. http://dx.doi.org/10.1016/j.ijforecast.2019.07.002.

Marin, J.B., Orozco, E.T., Velilla, E., 2018. Forecasting electricity price in Colombia: A comparison between neural network, ARMA process and hybrid models. Int. J. Energy Econ. Policy 8 (3), 97–106.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., Hassabis, D., 2015. Human-level control through deep reinforcement learning. Nature 518 (7540), 529–533. http://dx.doi.org/10.1038/nature14236.

Muniain, P., Ziel, F., 2020. Probabilistic forecasting in day-ahead electricity markets: Simulating peak and off-peak prices. Int. J. Forecast. 36 (4), 1193–1210. http://dx.doi.org/10.1016/j.ijforecast.2019.11.006.

Neupane, B., Woon, W.L., Aung, Z., 2017. Ensemble prediction model with expert selection for electricity price forecasting. Energies 10 (1), 77. http://dx.doi.org/10.3390/en10010077.

Next Kraftwerke AT GmbH (NEXT), 2019. Strompreiszonentrennung deutschland/österreich. URL https://www.next-kraftwerke.at/wissen/strommarkt/strompreiszonentrennung.

Nowotarski, J., Weron, R., 2018. Recent advances in electricity price forecasting: A review of probabilistic forecasting. Renew. Sustain. Energy Rev. 81, 1548–1568. http://dx.doi.org/10.1016/j.rser.2017.05.234.

Olah, C., 2015. Colah's blog: Understanding LSTM networks. URL https://colah.github.io/posts/2015-08-Understanding-LSTMs/.

Open Power System Data (OPSD), 2019. Time series: Load, wind and solar, prices in hourly resolution. URL https://data.open-power-system-data.org/time_series/2019-06-05.

Paschen, M., 2016. Dynamic analysis of the German day-ahead electricity spot market. Energy Econ. 59, 118–128. http://dx.doi.org/10.1016/j.eneco.2016.07.019.

Peng, L., Liu, S., Liu, R., Wang, L., 2018. Effective long short-term memory with differential evolution algorithm for electricity price prediction. Energy 162, 1301–1314. http://dx.doi.org/10.1016/j.energy.2018.05.052.

RDocumentation, 2021. Test for association/correlation between paired samples. URL https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/cor.test.

Schwarz, G., et al., 1978. Estimating the dimension of a model. Ann. Statist. 6 (2), 461–464.

Sharma, V., Srinivasan, D., 2013. A hybrid intelligent model based on recurrent neural networks and excitable dynamics for price prediction in deregulated electricity market. Eng. Appl. Artif. Intell. 26 (5), 1562–1574. http://dx.doi.org/10.1016/j.engappai.2012.12.012.

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

Uniejewski, B., Weron, R., 2021. Regularized quantile regression averaging for probabilistic electricity price forecasting. Energy Econ. 95, 105121. http://dx.doi.org/10.1016/j.eneco.2021.105121.

Uniejewski, B., Weron, R., Ziel, F., 2018. Variance stabilizing transformations for electricity spot price forecasting. IEEE Trans. Power Syst. 33 (2), 2219–2229. http://dx.doi.org/10.1109/TPWRS.2017.2734563.

Weron, R., 2014. Electricity price forecasting: A review of the state-of-the-art with a look into the future. Int. J. Forecast. 30 (4), 1030–1081. http://dx.doi.org/10.1016/j.ijforecast.2014.08.008.

Weron, R., Misiorek, A., 2008. Forecasting spot electricity prices: A comparison of parametric and semiparametric time series models. Int. J. Forecast. 24 (4), 744–763. http://dx.doi.org/10.1016/j.ijforecast.2008.08.004.

Willmott, C.J., Matsuura, K., 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim. Res. 30 (1), 79–82. http://dx.doi.org/10.3354/cr030079.

Xie, M., Sandels, C., Zhu, K., Nordström, L., 2013. A seasonal ARIMA model with exogenous variables for elspot electricity prices in Sweden. In: 2013 10th International Conference on the European Energy Market (EEM). pp. 1–4. http://dx.doi.org/10.1109/EEM.2013.6607293.

Zhang, F., Fleyeh, H., 2019. A review of single artificial neural network models for electricity spot price forecasting. In: 2019 16th International Conference on the European Energy Market (EEM). pp. 1–6. http://dx.doi.org/10.1109/EEM.2019.8916423.

Ziel, F., 2017. Modeling the impact of wind and solar power forecasting errors on intraday electricity prices. In: 2017 14th International Conference on the European Energy Market (EEM). pp. 1–5. http://dx.doi.org/10.1109/EEM.2017.7981900.

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

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