Beran, Philip; Vogler, Arne; Weber, Christoph

Working Paper

Multi-day-ahead electricity price forecasting: A comparison of fundamental, econometric and hybrid models

HEMF Working Paper, No. 02/2021

Provided in Cooperation with:
University of Duisburg-Essen, Chair for Management Science and Energy Economics

Suggested Citation: Beran, Philip; Vogler, Arne; Weber, Christoph (2021) : Multi-day-ahead electricity price forecasting: A comparison of fundamental, econometric and hybrid models, HEMF Working Paper, No. 02/2021, University of Duisburg-Essen, House of Energy Markets & Finance (HEMF), Essen

This Version is available at:
https://hdl.handle.net/10419/268017

House of Energy Markets & Finance

Multi-day-ahead Electricity Price Forecasting: A Comparison of fundamental, econometric and hybrid Models

HEMF Working Paper No. 02/2021

by

Philip Beran,

Arne Vogler

And

Christoph Weber

October 2021

UNIVERSITÄT DUISBURG ESSEN

Open-Minded

Partly funded by:

EFRE.NRW Investitionen in Wachstum und Beschäftigung

Ministerium für Wirtschaft, Energie, Industrie, Mittelstand und Handwerk des Landes Nordrhein-Westfalen

EUROPÄISCHE UNION Investition in unsere Zukunft Europäischer Fonds für regionale Entwicklung

MULTI-DAY-AHEAD ELECTRICITY PRICE FORECASTING: A COMPARISON OF FUNDAMENTAL, ECONOMETRIC AND HYBRID MODELS[+]

by Philip Beran, Arne Vogler* and Christoph Weber

Abstract

Forecasting hourly electricity prices and their characteristic properties is a core challenge for energy generation companies and trading houses. The short-term marketing and purchase of electricity is usually managed with standardized products traded on different markets and with specific temporal resolution and maturity. The size and scope of the electricity price forecasting literature has grown significantly in recent years, with the majority of studies focused on short-term (intraday and day-ahead) or long-term (investment decisions) periods. However, the literature for forecasting the period beyond the day-ahead horizon, which is relevant for trading the aforementioned products or for managing assets over several days, is rather scarce. Our paper fills this gap by developing individual forecasting models covering horizons from the day ahead up to a week ahead. We introduce hybrids of a parsimonious fundamental model and various popular econometric models. In a case study for the German day-ahead market in 2016 we test and compare the different model settings by carefully considering realistic available data and limiting the calculation time to fit typical trading time constraints. We find that the best models across the individual horizons and across all horizons jointly are hybrid model approaches. They combine the strengths of autoregressive models in terms of capturing daily - even non-linear - structures with the immediate reactions of fundamental models to short-term events or fundamental changes in the market.

Keywords: Electricity markets, Electricity Price Forecasting, Hybrid Modeling, Fundamental Modeling, Econometric Modeling, German Day-Ahead Market

JEL-Classification: C13, C22, C51, Q41, Q47

[+]) The research presented in this paper has been partially carried out within the research project "StoOpt.NRW" finance by the Ministry of Economics of the state of North-Rhine-Westphalia (MWEIMH NRW) and the European Regional Development Fund (EFRE), allocated by the European Union.

PHILIP BERAN
House of Energy Markets and Finance,
University of Duisburg-Essen, Germany
Universitätsstr. 12, 45117 Essen
philip.beran@uni-due.de
www.hemf.net

ARNE VOGLER
(*Corresponding Author)
House of Energy Markets and Finance
University of Duisburg-Essen, Germany
Universitätsstr. 12, 45117 Essen
arne.vogler@uni-due.de
www.hemf.net

CHRISTOPH WEBER
House of Energy Markets and Finance
University of Duisburg-Essen, Germany
Universitätsstr. 12, 45117 Essen
+49-(0)201 / 183-2966
christoph.weber@uni-due.de
www.hemf.net

Declarations of Interest: None

Content

1 Introduction

The optimization, trading and subsequent operation of energy companies' generation assets requires taking a view on future electricity prices. To adequately account for the technical restrictions of generation units and to serve the various electricity markets, the decision-making problems have to be considered over and beyond the day-ahead horizon. Consequently, electricity price forecasts have to be provided over multi-day-ahead horizons as well.

Since the advent of the electricity price forecasting (EPF) literature, a plethora of forecasting approaches has been studied (cf. Weron (2014), Nowotarski and Weron (2018)). The prevailing approaches depend critically on the eventual application and thus the forecasting horizon as well as on data availability. Fundamental models, for example, have been primarily considered for medium- and long-term EPF. They aim to capture the underlying economic as well as physical relationships of electricity markets. In contrast, approaches based on econometric models dominate the short-term EPF literature and such models mainly characterize electricity prices as functions of previous prices and potentially additional exogenous variables.

Short-term fundamental EPF models have recently started to be proposed in the literature. In addition, such models have also been considered in studies on performance-improving combinations with existing econometric approaches, lending so-called hybrid models. In contrast to the majority of the short-term EPF literature, which is primarily focused on forecasting over the day-ahead horizon, they have been applied to multi-day-ahead forecasting. Yet, no study has investigated the effect of the individual forecasting horizon on their predictive performance. Also, it has not been generally addressed in the literature whether EPF models for multi-day-ahead horizons should be considered in recursive or direct form.

The present paper considers forecasts of German day-ahead electricity prices over multi-day-ahead horizons using hybrids of a parsimonious fundamental model and various econometric models. It contributes to the scarce literature on short-term fundamental and hybrid EPF models and provides insights about the predictive ability of fundamental-econometric model combinations, the effect of the forecasting horizon on model performance and whether such models should generally be considered in recursive or direct form. In addition, it provides empirical evidence on the forecasting accuracy of popular EPF models over horizons beyond the standard day-ahead horizon.

The remainder of the paper is structured as follows. Section 2 reviews the relevant literature and highlights the contributions of the present study. In Section 3, the different electricity price forecasting models are motivated and presented. In addition, the notion of recursive and direct forecasts is discussed and the considered evaluation framework is outlined. We provide an

overview of the considered dataset and discuss the broad results of the forecasting study in Section 4. Section 5 concludes.

2 Material and methods

2.1 Electricity price modelling and forecasting approaches

The EPF literature generally distinguishes between three forecasting horizons: short, medium and long term. The thresholds between these horizons are not unambiguously defined. Weron and Ziel (2019) formulate a rule of thumb according to which the notion of short term is based upon the availability of reliable (precise) meteorological forecasts for temperature, wind speeds and cloud cover. This covers periods ranging from a few minutes to several days. After this, the medium term covers all horizons beyond reliable meteorological forecasts with horizons ranging from weeks to months to several years. Finally, long-term horizons include everything that follows, starting with a few years up to several decades. Since the data availability and the application of the forecasts vary with the forecasting horizon, different models have been developed and applied. The focus of the present study is on EPF for marketing and operation decisions on spot and reserve capacity markets. We thus consider short-term EPF with horizons of up to a week ahead, as our application study focuses on the German market where reserve capacity has been traded up to one week ahead until recently.

Fundamental approaches are predominantly used in the medium and long term and are applied to the assessment of investment decisions or political measures. The price forecast is often just one of a number of results such as CO_2 emission levels, developments in the power plant fleet or generation volumes. Many of these models are operated and developed in-house and are therefore not publicly available (e.g. Weron (2014)). Ringkjøb et al. (2018) provide an overview of published and partly freely accessible large-scale energy system models.

Short-term fundamental EPF models are extremely rare. Reasons are the very high data requirements and extensive computing times. At the same time, although fundamental electricity market prices can represent average price levels (base prices over certain periods of time) well, they exhibit too low volatility and considerable difficulties in adequately representing extreme prices. Yet, especially these aspects are crucial for potential short-term applications in the energy industry. The few existing short-term fundamental EPF models constitute highly simplified or aggregated models. Beran et al. (2019) develop the simplified fundamental model *ParFuM* building on earlier work by Kallabis et al. (2016) and apply it to explain the price decline in the German day-ahead market in the years 2011 to 2015. For the price assessment, the authors solely use information that is publicly accessible at the time of the respective day-ahead gate closure. A similar approach is taken by Pape et al. (2016), who use the same model architecture to explain

price spreads between the German day-ahead and intraday markets. Marcos et al. (2019a) define a simple optimization model to determine the hourly day-ahead equilibrium prices for the Iberian market area for the year 2017. Although the model consists of a large number of equations and inputs, the computing time is reduced considerably by aggregating power plant classes and is therefore also suitable for usage in short-term markets. We contribute to this relatively scarce literature by presenting and validating a short-term fundamental EPF model for the German day-ahead market.

Econometric or statistical models are predominantly used for short-term EPF, as they are usually better suited to capture volatility and extreme price events. Weron (2014) already documents a substantial literature of statistical approaches, where electricity prices are typically characterized as functions of previous prices and additional exogenous variables. The predictive ability of such models depends mainly on the quality of the considered data, the incorporation of fundamental information and the efficiency of the employed algorithm. In recent years, the econometric EPF literature has grown even further with contributions addressing the comparison of multivariate and univariate model structures (e.g. Ziel and Weron (2018), Gianfreda et al. (2020)), the adoption of regularization techniques (e.g. Uniejewski and Weron (2018)) and the combination of forecasts across different calibration windows (e.g. Serafin et al. (2019)), to name but a few and without referring to the ever-increasing literature on probabilistic EPF. The state-of-the-art econometric model belongs to the class of so-called expert models and represents the electricity price as a function of autoregressive terms, non-linear terms, the price of the last hour of the preceding day and dummy variables that capture calendar information (e.g. Ziel and Weron (2018), Weron and Ziel (2019)). It is common to extend it with additional exogenous information such as predicted production of renewable energy sources (e.g. Gianfreda et al. (2020), Maciejowska et al. (2020)). Yet, not all econometric approaches model the electricity price directly. Ziel and Steinert (2016) propose a time series model extended by endogenous information for the bid volume in predefined price classes that underlie the supply and demand curves of the day-ahead market. Given forecasts for the bid volumes, they construct the resulting supply and demand curves, the intersection of which lends the predicted electricity price. In a follow-up paper, Ziel and Steinert (2018), they present one of the few econometric approaches to mid-term electricity price forecasting based on the preceding methodology. In a first stage, drivers of the physical market situation such as fossil or renewable generation are simulated with stochastic processes and these are translated into day-ahead expectations of fundamentals, supply and demand bid volumes and electricity price forecasts in a second stage.

Given the plethora of approaches to EPF in the literature, investigations into potentially forecast-performance-improving combinations have followed naturally. An approach can generally be

considered hybrid, if it constitutes a combination of two or more EPF techniques (cf. Weron (2014)). In the context of the present study, the terminology is used to describe an approach where the output of a fundamental EPF model constitutes an input to an econometric procedure. Here, the literature is as scarce as the literature on short-term fundamental EPF. Gonzalez et al. (2012) consider a hybrid model to forecast day-ahead baseload prices for Great Britain and report improved forecast accuracy in comparison to a number of purely econometric models. Bello et al. (2017) develop a hybrid forecasting approach for the medium term (several months). Their quantile regression for hourly Spanish electricity prices is based on a fundamental market equilibrium model that incorporates renewable feed-in, cross-border flows and load. They show that the forecasting accuracy for the tails of the electricity price distribution can be significantly increased. More recently, Marcos et al. (2019a) and Marcos et al. (2019b) study hourly day-ahead price forecasts from a hybrid model for the Iberian market. In both studies the fundamental model constitutes the aforementioned cost-production optimization model. Marcos et al. (2019a) include the fundamental market clearing price as an additional feature in a neural network, whereas Marcos et al. (2019b) consider the fundamental market clearing price and technology-specific generation levels as inputs in a neural network, the forecast of which is then combined with the forecast of a neural network that is not based on any output from the fundamental model. These works are thus closely related to our proposed hybrid approach for German day-ahead prices based on a parsimonious approximation of the bid stack.

The majority of contemporaneous studies on short-term EPF has considered the day-ahead forecasting horizon. Thus, very little work has been done on assessing models for forecasting hourly day-ahead prices over longer forecasting horizons, although the notion of longer horizons is not new to the EPF literature. Early studies have focused on predicting daily average electricity prices over horizons of variable length (e.g. Maciejowska and Weron (2013), Maciejowska and Weron (2015)). Muniain and Ziel (2020) have recently revisited the issue by predicting the distribution of daily average peak and off-peak prices over a horizon of seven days. Marcos et al. (2019a) evaluate the proposed hybrid model for Iberian day-ahead prices over both a day-ahead and a week-ahead forecasting horizon. We extend on them by investigating the forecasting performance of our hybrid model at each horizon individually. In addition, we do not only investigate the performance of the hybrid model relative to the benchmark models, but also consider relative performance among the benchmarks, providing empirical evidence on the predictive ability of popular EPF models over increasing forecasting horizons.

The notions of recursive and direct forecasts are directly linked to the forecasting horizon. A multistep-ahead forecast is made recursively, if a one-step ahead model is iterated forward up to the required forecasting horizon, whereas a multistep-ahead forecast is made directly, if a specific

model for each horizon is estimated (e.g. Marcellino et al. (2006)). The relative predictive performance of the two approaches has been studied in other fields of the forecasting literature. In contrast to the theoretical result of increased robustness and reduced bias offered by direct forecasting models (e.g. Marcellino et al. (2006), Taieb and Atiya (2016)), empirical investigations in macroeconometrics have uncovered the opposite for both unconditional and conditional forecasts (e.g. Marcellino et al. (2006), McCracken and McGillicuddy (2019)). To the best of our knowledge, no previous study has compared recursive and direct forecasting models in the context of EPF.

2.2 Testing literature

The Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) constitute the two most popular scores for point forecast evaluation in the EPF literature (e.g. Gürtler and Paulsen (2018), Weron and Ziel (2019)). Whereas the former is a strictly proper scoring rule for median forecasts, the latter is a strictly proper scoring rule for mean forecasts. As both scores are scale dependent and do not easily facilitate forecast comparisons between different data sets, additional measures based on percentage and scaled forecast errors have been considered. Yet, Weron (2014) and Weron and Ziel (2019) conclude that the EPF literature has thus far not established an evaluation standard.

Whereas a comparison of a chosen evaluation measure does allow for a ranking of individual forecasts, it does not allow to establish the statistical significance of deviations in accuracy between them. To this end, statistical tests of equal predictive ability are considered. The Diebold and Mariano (2002) (DM) test is widely applied in the EPF literature (e.g. Uniejewski et al. (2018), Ziel and Weron (2018) and Ugurlu et al. (2018)). It considers whether the loss differential series of two models exhibits an expected value of zero and is usually considered in its multivariate one-sided version (e.g. Ziel and Weron (2018)). Recently, a test outlined in Giacomini and White (2006) has gained popularity (e.g. Marcjasz et al. (2018), Serafin et al. (2019) and Marcjasz et al. (2020)). It is based on an alternative econometric framework to the DM test in the sense that forecasting models are compared at the estimated coefficients rather than the corresponding population values. One can test whether the loss differential series of two models exhibits an expected value of zero both unconditionally or conditionally. To the best of our knowledge, only the conditional version with lagged loss differentials has thus far been considered in the EPF literature (e.g. Serafin et al. (2019)); that is, the conditioning set considered contained a constant and lagged values of the loss differential. We extend on the previous applications of the test and consider both its unconditional form as well as its conditional form with a conditioning set containing a measure of uncertainty of renewable infeed. In addition, to assess relative model

performance over all horizons jointly, we are the first EPF study to consider a test proposed by Quaedvlieg (2021).

3 Forecasting models and evaluation framework

3.1 Electricity price forecasting models

In the following section, we define the different forecasting models from the fundamental, econometric and hybrid model classes. The simplified electricity market model *ParFuM* introduced and extended by Kallabis et al. (2016), Pape et al. (2016) and Beran et al. (2019), among others, represents the class of fundamental models in the present work.[1] It is characterized by accurate forecasting quality with comparatively low data requirements and very short calculation times. These properties are achieved through complexity reduction by considering aggregated technology classes, the absence of coupled time steps and the endogenous modelling of only one market area. The basis of *ParFuM* is a simple supply stack model, consisting of an ascending bid curve, which results from the marginal costs of electricity supply, and a quasi-inelastic demand. Let t denote the day of the forecast up to and at which all hourly spot market prices of the day are completely known and let k denote the considered forecasting horizon in days. We forecast the hourly spot market prices $p_{t+k,h}$ with $h \in \{1, \dots 24\}$ and $k \in \{1, \dots, K\}$. The supply curve corresponds to the aggregated bid stack $B_{t+k,h}\left(D_{t+k,h}\right)$ of the available power plant capacities at a given demand $D_{t+k,h}$. In order to approximate the heterogeneity within the technology classes, different efficiency levels are considered. The demand $D_{t+k,h}$ constitutes a forecast of the residual load, resulting from expected load minus the infeed from wind and solar as well as net cross-border exchange and must-run combined heat and power (CHP) production. The latter is approximated by a temperature-dependent function (cf. Beran et al. (2019)). The fundamental price $p_{t+k,h}^{ParFuM,k}$ of hour h on day $t+k$ results from the intersection of the supply and demand curves and corresponds to the marginal cost of the last power plant needed to satisfy demand. The following therefore applies for the horizon $k = 1$:

$$p_{t+1,h}^{ParFuM,1} = B_{t+1,h}(D_{t+1,h}). \tag{1}$$

We define the pure *ParFuM* price as our first forecasting model, where the spot market price of a particular hour h on day $t+1$ is given by the corresponding *ParFuM* price.

[1] A detailed description of the implemented model can be found in Beran et al. (2019). We adapt the basic structure of the model but provide a broader representation of negative prices. In the case of negative residual load, we obtain negative prices that are linearly interpolated between -100 €/MWh and 0 €/MWh. Our chosen range of values corresponds to the negative prices observed on the market, which result from the guaranteed remuneration payments for renewable feed-in. Beran et al. (2019) set the market price to constant -10 €/MWh in case of negative residual load.

ParFuM Model:

$$p_{t+1,h} = p_{t+1,h}^{ParFuM,1} \tag{2}$$

To account for structural biases in the *ParFuM* price forecast, an additional post-processing step is also considered. The *ParFuM* forecast constitutes the sole explanatory variable in a predictive regression, which allows for the mitigation of the potential biases through the estimated coefficients and lends the *FunR* model.

FunR Model:

$$p_{t+1,h} = \beta_{h,0}^{1}(t+1) + \beta_{h,1}^{1} p_{t+1,h}^{ParFuM,1} + \varepsilon_{t+1,h}, \tag{3}$$

where $\beta_{h,0}^{1}(t+1)$ denotes the following time-varying intercept that is common to all models.

$$\begin{aligned}\beta_{h,0}^{1}(t+1) = \beta_{h,0,0}^{1} &+ \beta_{h,0,1}^{1} \sin\frac{2\pi(t+1)}{365.24} + \beta_{h,0,2}^{1} \cos\frac{2\pi(t+1)}{365.24} \\ &+ \beta_{h,0,3}^{1} \sin\frac{4\pi(t+1)}{365.24} + \beta_{h,0,4}^{1} \cos\frac{4\pi(t+1)}{365.24} + \beta_{h,0,5}^{1} D_{t+1}^{Mo} \\ &+ \beta_{h,0,6}^{1} D_{t+1}^{Fr} + \beta_{h,0,7}^{1} D_{t+1}^{Sa} + \beta_{h,0,8}^{1} D_{t+1}^{Su}\end{aligned} \tag{4}$$

The intercept is thus modelled as the sum of a constant term, a second-order Fourier approximation for seasonal effects and four dummy variables that capture calendar information. The dummies reflect whether the day of the forecast constitutes a Monday, Friday, Saturday or Sunday. In addition, all public holidays are modelled as either Saturday or Sunday, depending on whether they constitute local or nationwide public holidays. The days before and after a public holiday are modelled as Friday and Monday, respectively.

Model *ArR*, the autoregressive model, belongs to the generic class of so-called expert models (e.g. Ziel and Weron (2018)) and represents the state-of-the-art econometric model for short-term EPF. The electricity price of a particular hour h on day $t+1$ is modelled as a function of the day-ahead price of the same hour lagged by one, two and seven days as well as the minimum, maximum and last day-ahead price of the previous day.

ArR Model:

$$\begin{aligned}p_{t+1,h} = \beta_{h,0}^{1}(t+1) &+ \beta_{h,2}^{1} p_{t,h} + \beta_{h,3}^{1} p_{t-1,h} + \beta_{h,4}^{1} p_{t-6,h} + \beta_{h,5}^{1} p_{t,Max} \\ &+ \beta_{h,6}^{1} p_{t,Min} + \beta_{h,7}^{1} p_{t,24} + \varepsilon_{t+1,h}\end{aligned} \tag{5}$$

Similar to previous works in the EPF literature, we consider two extensions of the baseline autoregressive model with exogenous information. First, the fundamental price forecast from the *ParFuM* model is included, lending the *FunArR* model, which constitutes the first hybrid model.

FunArR Model:

$$p_{t+1,h} = \beta^1_{h,0}(t+1) + \beta^1_{h,1} p^{ParFuM,1}_{t+1,h} + \beta^1_{h,2} p_{t,h} + \beta^1_{h,3} p_{t-1,h} + \beta^1_{h,4} p_{t-6,h} + \beta^1_{h,5} p_{t,Max} + \beta^1_{h,6} p_{t,Min} + \beta^1_{h,7} p_{t,24} + \varepsilon_{t+1,h} \quad (6)$$

Second, we obtain model *ArLoR* by including a forecast of the day-ahead residual load $L^1_{t+1,h}$, which is defined as load minus the sum of wind power production, solar power production and net cross-border exchange, i.e. cross-border commercial schedules (CBCS). It thus constitutes a slightly altered definition of residual load to the one used for demand approximation in the *ParFuM* model.

ArLoR Model:

$$p_{t+1,h} = \beta^1_{h,0}(t+1) + \beta^1_{h,2} p_{t,h} + \beta^1_{h,3} p_{t-1,h} + \beta^1_{h,4} p_{t-6,h} + \beta^1_{h,5} p_{t,Max} + \beta^1_{h,6} p_{t,Min} + \beta^1_{h,7} p_{t,24} + \beta^1_{h,8} L^1_{t+1,h} + \varepsilon_{t+1,h} \quad (7)$$

Note that the individual components of the residual load could have been included as separate regressors as in Maciejowska et al. (2020). The present model formulation therefore amounts to an implicit restriction of equality on the individual parameters motivated by the fact that a unit change in either variable should have the same effect on the bid stack and thus the clearing price.

To further assess the informational content of predicted residual load for EPF, we consider two additional models. The *LoR* model results from a predictive regression with the predicted residual load as sole predictor, whereas the *FunLoR* model additionally incorporates the *ParFuM* price forecast.

LoR Model:

$$p_{t+1,h} = \beta^1_{h,0}(t+1) + \beta^1_{h,8} L^1_{t+1,h} + \varepsilon_{t+1,h} \quad (8)$$

FunLoR Model:

$$p_{t+1,h} = \beta^1_{h,0}(t+1) + \beta^1_{h,1} p^{ParFuM,1}_{t+1,h} + \beta^1_{h,8} L^1_{t+1,h} + \varepsilon_{t+1,h} \quad (9)$$

Finally, the combination of all preceding model components lends the full model, labelled *FullR*. It nests all other considered models, which can be derived from it using restrictions on its parameter space.

FullR Model:

$$p_{t+1,h} = \beta^1_{h,0}(t+1) + \beta^1_{h,1} p^{ParFuM,1}_{t+1,h} + \beta^1_{h,2} p_{t,h} + \beta^1_{h,3} p_{t-1,h} + \beta^1_{h,4} p_{t-6,h} + \beta^1_{h,5} p_{t,Max} + \beta^1_{h,6} p_{t,Min} + \beta^1_{h,7} p_{t,24} + \beta^1_{h,8} L^1_{t+1,h} + \varepsilon_{t+1,h} \quad (10)$$

3.2 Forecasting framework

To elucidate the underlying information set, all models have thus far been formulated in their day-ahead form, where the electricity price is modelled based on information available on the previous day t, including day-ahead forecasts of the exogenous variables. It should be noted that

the above formulations therefore constitute the parameter estimation equations of the recursive models, identified by R at the end of the model name. In the recursive framework, the model coefficients are estimated over the day-ahead horizon ($k = 1$), as denoted by the superscripted 1. To forecast over all horizons $k \in \{1, \dots, K\}$, the parameter estimates $\hat{\beta}_h^1$ are fixed and the resulting day-ahead model is applied recursively. Since some recursive models contain autoregressive elements, not all values of the explanatory variables are known at the time of forecasting. Any value of an autoregressive variable that is unknown is replaced by the corresponding forecast from a previous recursion (see Table 5 in Appendix A1). The forecasting equations of the recursive models are presented in Appendix A1.

Additionally, the present paper considers all models in their so-called direct form. In the direct framework, a model is based on exactly the same information available on day t, but rather than estimating the parameters over the day-ahead horizon and applying the resulting day-ahead model recursively, the coefficients are estimated separately for each forecasting horizon k. Consequently, a horizon-specific parameter set γ_h^k is associated with the predictors and forecasts over the different horizons can be directly obtained. It should be noted that the values of all predictors are immediately known at the time of forecast issuance. The estimation and forecasting equations for the direct formulations of all models, identified by D at the end of the model name, are also presented in Appendix A1 and the k-day-ahead estimation horizon is now denoted by the superscripted k on the coefficients.

Figure 1 illustrates the recursive and direct forecasting framework and highlights the basic differences in terms of estimation and forecasting. The estimation of parameters for the individual hours constitutes the first step in both frameworks. It is considered either for the day-ahead horizon only or for each horizon individually. Yet, in both frameworks, parameters are estimated based on a fixed window of size τ rolled forward in time. The available information on the day of forecast issuance t depends on the gate closure of the considered market area. In this paper, we focus on the German day-ahead auction and thus the relevant gate closure is at 12.00 CET/CEST. To ensure that both parameter estimation and forecasting only use information that is known at gate closure, we set the information cut-off time at 11.40 CET/CEST. This definition leaves a small amount of computation time and processing buffer until the actual gate closure at 12.00 CET/CEST (marked in dark grey in Figure 1). This cut-off time implies that at time of parameter estimation the prices of the current day t are completely known from the auction held the day before (known prices are marked in light grey in Figure 1). Based on the estimated parameters, the forecasts for all hours and over all horizons $k \in \{1, \dots, K\}$ are calculated. Note that the first hour to be predicted is the first hour of the following day $t + 1$ (prices to be predicted are marked in white in Figure 1).

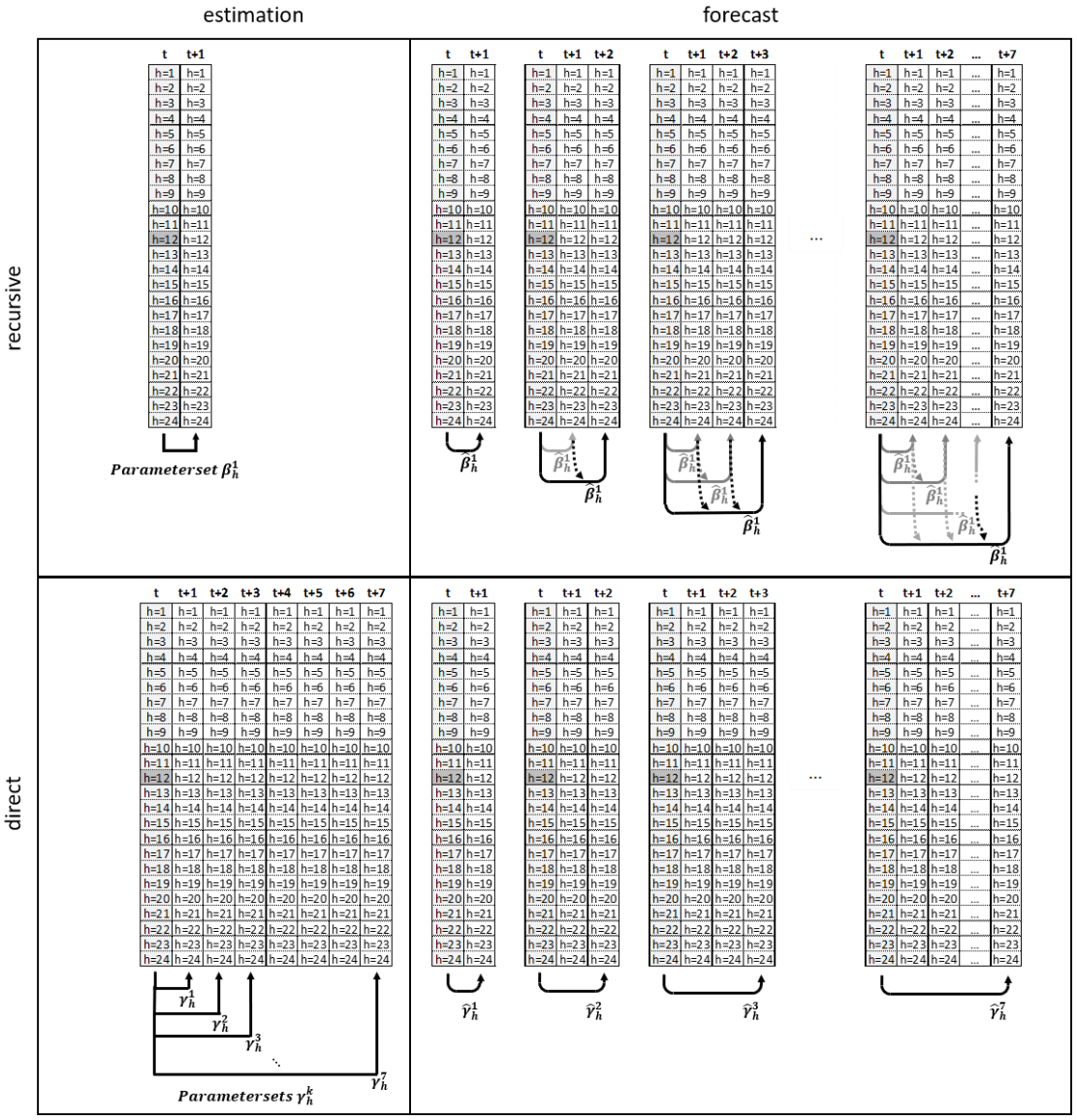

Figure 1: Recursive and direct model estimation and forecasting

3.3 Validation and evaluation

The predictive performance of the considered models is evaluated in an out-of-sample forecasting study. Let T and τ denote the sample size and the estimation window size, respectively. The first k-step-ahead forecast of the electricity price of hour h on day $t+k$ from model i, $\hat{p}_{t+k,h}^{i,k}$, is calculated based on the first τ observations, while consecutive k-step-ahead forecasts are calculated based on τ observations shifted forward in time. Thus, the employed rolling window procedure yields $n = T - \tau - k + 1$ k-step-ahead forecasts indexed $\tau, \dots, T-k$ for evaluation. The out-of-sample sequences of k-step-ahead forecast, realized price and associated k-step-ahead forecast error are given by $\{\hat{p}_{t+k,h}^{i,k}\}_{t=\tau}^{T-k}$, $\{p_{t+k,h}\}_{t=\tau}^{T-k}$ and $\{\hat{\varepsilon}_{t+k,h}^{i,k}\}_{t=\tau}^{T-k}$, respectively.

The first considered metric of predictive performance is the mean absolute error (MAE) across all hours $h \in H$ and all days $t \in \Phi$ defined as

$$MAE_{\Phi,H}^{i,k} = \frac{1}{\#\Phi\#H}\sum_{t\in\Phi}\sum_{h\in H}\left|\hat{\varepsilon}_{t+k,h}^{i,k}\right|, \tag{11}$$

where $\#\Phi$ and $\#H$ denote the cardinality of the sets Φ and H, respectively. This notation is flexible enough to include the MAE over the entire hold-out sample (i.e., $\Phi = \{\tau, \dots, T-k\}$ and $H = \{1, \dots, 24\}$) but also over individual hours $h \in H$ and subperiods (e.g. weeks).

As the level of electricity prices varies strongly over the course of a year, the MAE may provide misleading results when comparing different subperiods within the considered evaluation period. For such comparisons, an error measure adjusted for the respective period's price level is more appropriate. We consider a version of the Weighted Mean Absolute Error (WMAE), for which the MAE is normalized by the mean price of the considered period (cf. Weron (2014)).

$$WMAE_{\Phi,H}^{i,k} = \frac{\frac{1}{\#\Phi\#H}\sum_{t\in\Phi}\sum_{h\in H}\left|\hat{\varepsilon}_{t+k,h}^{i,k}\right|}{\frac{1}{\#\Phi\#H}\sum_{t\in\Phi}\sum_{h\in H} p_{t+k,h}} \tag{12}$$

We use this WMAE statistic particularly for a first graphical analysis of forecasting performance over our evaluation period.

While a comparison of MAE (or also WMAE) values provides a ranking of models, it does not allow to establish conclusions on statistically significant differences in forecasting performance between them. To this end, we define the pairwise loss differential for forecasts over horizon k between models i and j as $d_{t+k}^{ij,k} = \left\|\hat{\varepsilon}_{t+k}^{i,k}\right\|_1 - \left\|\hat{\varepsilon}_{t+k}^{j,k}\right\|_1$ and base hypothesis tests of predictive ability on the out-of-sample loss differential sequence $\left\{d_{t+k}^{ij,k}\right\}_{t=\tau}^{T-k}$.

Giacomini and White (2006) develop a test of equal predictive ability between two models conditional on some information set. The null hypothesis of conditional equal predictive ability (CEPA) is formulated as $\mathrm{E}\left[d_{t+k}^{ij,k}|\mathcal{G}_t\right] = 0$, where $\mathcal{G}_t$ denotes said information set. There are two refinements of this null hypothesis that are of primary interest. First, the conditioning set is equal to the trivial σ-field, i.e. $\mathcal{G}_t = \{\emptyset, \Omega\}$. Second, the conditioning set is equal to the information set available at the time of forecast issuance, i.e. $\mathcal{G}_t = \mathcal{F}_t$. The former is shown to amount to a test of unconditional equal predictive ability (UEPA) in the spirit of Diebold and Mariano (2002).

A test of UEPA considers whether two models exhibit equal forecasting performance on average. If a statistically significant difference is established, the model with the lower average loss is selected. In contrast, a test of CEPA considers whether the relative performance of two models can itself be predicted with information available at the time of forecast issuance and exploits this information for model selection, if the null hypothesis is rejected. Thus, the forecast user may find the models to predict equally well on average (UEPA null hypothesis fails to reject) but may

be able to select a model based on current information rather than just past average performance (CEPA null hypothesis rejects). It should be noted that the null hypothesis of CEPA should be rejected, if the null hypothesis of UEPA is rejected, although Giacomini and White (2006) establish situations where this counterintuitive result may arise.

The present study considers a measure of uncertainty of renewable infeed as conditioning variable. If two of the considered specifications predict equally well on average, we test whether one of the models is more reliable in case of heightened renewable uncertainty, which we consider a case of primary interest in a power system that is increasingly based on renewable energy sources.

The test statistic of the UEPA test with null hypothesis H_0^{UEPA}: $\mathrm{E}\left[d_{t+k}^{ij,k}\right] = 0$ is defined as

$$s_{UEPA}^{ij,k} = \frac{n^{-1}\sum_{t=\tau}^{T-k} d_{t+k}^{ij,k}}{\sqrt{\frac{\widehat{\omega}_n^2}{n}}} \sim N(0,1), \tag{13}$$

where $\widehat{\omega}_n^2$ denotes a suitable estimator of the long-run variance of $\left\{d_{t+k}^{ij,k}\right\}_{t=\tau}^{T-k}$. The UEPA test coincides with the original DM test but the result is derived under assumptions allowing for parameter estimation. Thus, the test allows for the evaluation of forecasting models at the finite-sample estimates of the coefficients rather than their population values.

The null hypothesis of the CEPA test is formulated as H_0^{CEPA}: $\mathrm{E}\left[d_{t+k}^{ij,k}|\mathcal{F}_t\right] = 0$. Yet, to operationalize the test a $q \times 1$ $\mathcal{F}_t$–measurable vector h_t is considered instead of $\mathcal{F}_t$. Thus, h_t contains a constant and the $q-1$ variables believed to account for the difference in forecasting performance of the two models. The test statistic is based on the sequence $\left\{h_t d_{t+k}^{ij,k}\right\}_{t=\tau}^{T-k}$ and defined as

$$S_{CEPA}^{ij,k} = \left(n^{-1}\sum_{t=\tau}^{T-k} h_t d_{t+k}^{ij,k}\right)' \left[\frac{\widehat{\Omega}_n}{n}\right]^{-1} \left(n^{-1}\sum_{t=\tau}^{T-k} h_t d_{t+k}^{ij,k}\right) \sim \chi_q^2, \tag{14}$$

where $\widehat{\Omega}_n$ denotes a suitable estimator of the long-run variance of $\left\{h_t d_{t+k}^{ij,k}\right\}_{t=\tau}^{T-k}$.

It should be noted that the CEPA null hypothesis not only imposes restrictions on the first moment of $h_t d_{t+k}^{ij,k}$ but also on its second moments. Specifically, the lag length to be considered for the heteroskedasticity-and-autocorrelation-consistent (HAC) estimation of $\widehat{\Omega}_n$ is given by $k-1$. No such results can be established under the UEPA null hypothesis, which only imposes restrictions on the first moment of $d_{t+k}^{ij,k}$, and thus the lag length for the HAC estimation of $\widehat{\omega}_n^2$ must be selected by the forecaster. Yet, as the size properties of the Giacomini and White (2006) tests can be

improved using a sample-dependent lag length, we set the lag length to $\left\lfloor 4\left(\frac{n}{100}\right)^{2/9}\right\rfloor$ for the estimation of both $\widehat{\omega}_n^2$ and $\widehat{\Omega}_n$ (cf. McCracken (2019)).

A rejection of a null hypothesis of equal predictive ability subsequently requires a logic to select the better model. Whereas the UEPA test can be considered as a one-sided test, which directly provides such logic, the CEPA test cannot. Note that the rejection of the CEPA null hypothesis effectively means that the relative performance of the two models $d_{t+k}^{ij,k}$ can be predicted using h_t. Thus, one can estimate the regression $d_{t+k}^{ij,k} = \phi^{ij,k\prime} h_t + \eta_{t+k}^{ij,k}$ over the out-of-sample period and use the predicted values of the loss differential, i.e. $\{\hat{d}_{t+k}^{ij,k}\}_{t=\tau}^{T-k} = \{\hat{\phi}^{ij,k\prime} h_t\}_{t=\tau}^{T-k}$, for model selection (e.g. Giacomini and White (2006)). We follow Granz era and Sekhposyan (2019) and construct the statistic

$$I^{ij,k} = \frac{\sum_{t=\tau}^{T-k} \left|\hat{d}_{t+k}^{ij,k}\right| 1\{\hat{d}_{t+k}^{ij,k} \leq 0\}}{\sum_{t=\tau}^{T-k} \left|\hat{d}_{t+k}^{ij,k}\right|}, \tag{15}$$

which constitutes a loss-differential-weighted average of an indicator series showing whether model i is expected to forecast superiorly. Since $I^{ij,k}$ is bounded between 0 and 1, model i is preferred, if $I^{ij,k}$ is greater than 0.5.

The aforementioned tests of equal predictive ability consider the models at each forecasting horizon individually. Consequently, the possibility of inconsistent results of model comparison across different horizons arises. To address this shortcoming, Quaedvlieg (2021) proposes a test of average superior predictive ability (aSPA), that evaluates model performance over multiple horizons jointly. We denote by d_{t+k}^{ij} the weighted average of the loss differential over all considered forecasting horizons $k \in \{1, \dots, K\}$,

$$d_{t+k}^{ij} = [w^1, \dots, w^K] \begin{bmatrix} d_{t+k}^{ij,1} \\ \vdots \\ d_{t+k}^{ij,K} \end{bmatrix}, \tag{16}$$

and test the null hypothesis $H_0^{aSPA}: \mathrm{E}\left[d_{t+k}^{ij}\right] \leq 0$. Note that this allows for outperformance at some horizons to balance out underperformance at other horizons. As in Giacomini and White (2006), the asymptotics of the test are such that models are evaluated at the estimated parameter values. The test statistic is defined as

$$s_{aSPA}^{ij} = \frac{(T-K-\tau+1)^{-1} \sum_{=\tau}^{T-K} d_{t+k}^{ij}}{\sqrt{\frac{\hat{\sigma}^2_{(T-K-\tau+1)}}{(T-K-\tau+1)}}}, \tag{17}$$

where $\hat{\sigma}^2_{(T-K-\tau+1)}$ denotes a suitable estimator of the long-run variance of $\{d^{ij}_{t+k}\}^{T-K}_{t=\tau}$. The aSPA test thus amounts to a DM-type test on the average loss differential across different forecasting horizons. Whereas the test statistic could be compared to a normal distribution, Quaedvlieg (2021) maintains that bootstrapping the critical values is to be preferred.

4 Results and discussion

4.1 Data

The proposed models are evaluated in a forecasting study of hourly German day-ahead prices for the full year 2016.[2] Due to the rolling window approach with an estimation window size of $\tau = 730$ days, the entire sample size is $T = 1096$ and comprises the years 2014, 2015 and 2016. In order to examine the forecasting horizons day ahead until week ahead, we set $K = 7$. We choose to examine this somewhat distant period, because our models require authentic historical forecasts across all horizons. As discussed later in this section, the availability of such forecast data is very limited and data providers are very reluctant to publish their more recent forecasting histories.

Table 1 provides an overview of the data used for estimation and forecasting as well as its availability and sources. Except for the purely autoregressive models *ArR* and *ArD*, the models use fundamental or fundamentally determined information. In accordance with the principle of day-ahead markets, which trade today for tomorrow, forecast values for these fundamental data must be used. Since we forecast prices for the daily horizons $k \in \{1, \dots, 7\}$, predictions of the fundamental regressors are required across all these horizons. Thus, in our setting, the values for *ParFuM* prices, solar infeed, wind infeed and residual load are to be understood as forecasted values at the time of the information cut-off on day t. According to Section 3.1, the residual load is composed of the forecasts for expected load, wind and solar infeed-in, CBCS volumes and, within the *ParFuM* setting, also must-run CHP production. Hourly load data (day-ahead forecasts and total values) are publicly available via the ENTSO-E transparency platform. Unfortunately, hourly load forecasts for horizons $k > 1$ are not publicly available there and, as far as we know, also nowhere else. We therefore generate load forecasts for horizons $k > 1$ by means of a simple SARIMA model. In a first step, we consider annual seasonality through a cosine function and capture weekly structures using dummy variables for the individual hours of the week. Public holidays are treated as either Saturdays or Sundays, in line with standard forecasting practice (see Section 3.1). We subsequently estimate the residuals using a rolling window SARIMA model with a three-year sample size. Since we forecast day-ahead hourly load values in line with the

[2] Before 1st October 2018, the day-ahead spot market auction was held jointly for the German and Austrian market areas. Thus, this joint "EPEX Spot Germany/Austria" price is the focus of our analyses.

definition of the data publications of the ENTSO-E transparency platform, our load forecast data set reflects only about 86% of the actual German load (grid losses, parts of industrial and traction power stations, etc., see Beran et al. (2019) and Hirth and Schumacher (2015)). Thus, we follow Beran et al. (2019) to scale these forecasted load values to the corresponding IEA monthly values.

Table 1: Raw data sources and used data for all horizons

Data	Data description	Resolution[3]	Available fc horizons	Source
Coal price	API#2 (CIF ARA) future (front month)	D	t+1 … t+7	Marex Spectron via Energate-Messenger
Gas price	NCG OTC day-ahead quotation	D	t+1 … t+7	Marex Spectron via Energate-Messenger
Oil price	Europe Brent spot FOB quotation	D	t+1 … t+7	U.S. Energy Information Administration
CO_2 price	EEX 3. Period European carbon futures quotation (front year)	D	t+1 … t+7	EEX via Energate-Messenger
Wind infeed	German wide wind energy infeed	QH	t+1 … t+7	Anonymous professional forecast provider
Solar infeed	German wide solar energy infeed	QH	t+1 … t+7	Anonymous professional forecast provider
Temperature	Average German wide temperature	H	t+1 … t+3[4] t+1 … t+7	Anonymous professional forecast provider
Load	Electricity supplied	M	-	IEA
	Day-ahead hourly load values	H	t+1	ENTSO-E transparency
	Simple fc model for hourly load values	H	t+2 … d+7	SARIMA model
Cross-border trade	Scheduled commercial exchanges	H	t+1	ENTSO-E transparency
	Forecast model for hourly CBCS values	H	t+2 … t+7	Non-linear model (logistic transformation and multiple regression)
Capacity	EEX master data power	D	-	EEX Transparency
	Installed net generation capacity	Y	-	ENTSO-E (2015)ENTSO-E (2016), ENTSO-E (2017), ENTSO-E transparency,
	Power plant list	D	-	BNetzA
Availability	Non-usability generation (ex ante & ex post)	H	t+1 … t+7	EEX Transparency
Must-run CHP	CHP production volumes	Y	-	Öko-Institut (2015)
Electricity price	EPEX spot Germany/Austria Phelix quotation	H	t+1	EPEX SPOT

For wind and solar infeed there exist publicly available forecasts and estimates for the day-ahead horizon, e.g. published by the German transmission system operators and the ENTSO-E transparency platform. However, there is no public data source for hourly historical forecasts over the required horizons $k > 1$. As especially the uncertainty of wind and solar infeed can vary greatly between days, it is crucial to consider authentic data here. We therefore use data of an anonymous professional forecast provider who provides historical forecast data for all years and horizons in our study. The data provider updates its forecasts several times per day. In each case, the forecasts that are as close as possible to the information cut-off deadline are considered in

[3] The resolution column states the temporal resolution of the original data source: Y=Yearly, M=Monthly, D=Daily, H=Hourly, QH=Quarter-hourly. All timeseries are edited to become hourly input data for the fundamental model. Missing data is interpolated.

[4] For the years 2014 and 2015 only temperature forecasts for horizons k=1 to k=3 are available. We set temperature forecasts for horizons k>3 to the k=3 forecast values. For the year 2016, temperature forecasts for all considered horizons are available.

the models. Since Germany's electricity grid is highly interconnected with its European neighbors, cross-border-trade is important for German electricity markets and thus constitutes a substantial part of the residual load considered in all fundamental models described. The most comprehensive public source for these cross-border trade flows is the ENTSO-E transparency platform, which in principle contains data on the day-ahead horizon and total values (i.e. the sum of cross-border flows traded on day-ahead and intraday markets). Unfortunately, these data are extremely incomplete for the years under consideration (cf. Hirth et al. (2018)) and, moreover, they are not published for the required longer horizons $k > 1$. Thus, in a first step, we adjust the published data to provide an hourly aggregated day-ahead cross-border balance value $TB_{t+1,h}$ for Germany for the years 2014 to 2016. Kallabis et al. (2016) and Beran et al. (2019) develop multiple regression models to predict day-ahead CBCS flows. We improve the approximation accuracy even further by developing a multiple regression model as described in equation (18).[5] Equation (19) indicates that the explained variable is an inverse sigmoid transform of the aggregated cross-border balance $TB_{t+1,h}$. It has been chosen as cross-border exchanges are limited by the maximum transmission capacities and the parameter Ref corresponds to the maximum achievable cross-border exchange.[6]

$$z_{t+1,h}^{TB} = \beta_0 + \beta_1 W_{t+1,h} + \beta_2 PV_{t+1,h} + \beta_3 L_{t+1,h} + \beta_4 AvCap_{t+1,h}^{LIG} + \beta_5 AvCap_{t+1,h}^{NUC} + \beta_6 CO_{2\,t+1,h} + \varepsilon_{t+1,h} \tag{18}$$

$$z_{t+1,h}^{TB} = scal \cdot 2 \cdot artanh\left(\frac{TB_{t+1,h}}{Ref}\right) = scal \cdot \ln\left(\frac{1 + \frac{TB_{t+1,h}}{Ref}}{1 - \frac{TB_{t+1,h}}{Ref}}\right) \tag{19}$$

The adopted non-linear approach explains 79% of the variance of the CBCS flow $TB_{t+1,h}$ which is significantly higher than the adj. R^2 values in Kallabis et al. (2016) (~52%) and Beran et al. (2019) (~60%). We estimate the regression parameters with data over the $k = 1$ horizon and use them to forecast the cross-border flows for horizons $k > 1$.[7] These parameter estimates are then used together with predictions of the regressor variables to obtain the CBCS values for the

[5] $AvCap_{t+1,h}^{LIG}$ and $AvCap_{t+1,h}^{NUC}$ in equation (18) correspond to the available capacity of lignite fired and nuclear power plants. These power stations are rather inflexible and produce at low variable cost and thus have a significant effect on hourly CBCS volumes.

[6] The parameter Ref in equation (19) might in principle be estimated using a non-linear regression setting. In order to avoid convergence problems, we instead set $Ref = 25000$, which is an expert estimate of the maximum export/import capability. The scaling parameter $scal$ is introduced to obtain regression coefficients of interpretable size. By choosing a typical cross-border exchange value as reference point $Refpoint$, we compute the scaling factor as $scal = Refpoint / \ln\left(\frac{Ref + Refpoint}{Ref - Refpoint}\right)$. Setting $Refpoint = 5000$ we get $scal = 12332.02$. The estimated coefficients β then reflect the impact strengths of the corresponding factor on the exchange balance at the chosen reference point.

[7] Directly estimating the regression parameters for each horizon resulted in lower adj. R^2 values in contrast to the implemented approach.

forecasting equations (see Appendix A1) of the corresponding horizons. Similarly, the fundamental prices from *ParFuM* are generated as forecasts over all horizons. On the demand side, the residual load described above is additionally adjusted by a temperature-driven share of must-run capacity. The approach is developed in Kallabis et al. (2016) and Pape et al. (2016). The *ParFuM* version implemented here has been extended and described in Beran et al. (2019). For the horizons $k > 1$, temperature forecasts for the different horizons are used. For the construction of the supply curve in *ParFuM*, fuel and CO_2 prices as well as installed capacities and power plant availabilities are additionally required according to Section 3.1. For the calculation of the variable costs of the individual power plant classes, the products with shortest time to maturity quoted on markets for the respective fuels or CO_2 certificates are considered, taking into account the information cut-off date. We assume that the installed capacity considered changes over time, but not over the chosen horizons. The hourly calculation of power plant availability follows the procedure from Beran et al. (2019). However, only those non-availability notifications for the respective horizons that were publicly reported at the information cut-off time are taken into account.

In addition, a measure of uncertainty of renewable infeed is considered as the conditioning variable in the CEPA test introduced by Giacomini and White (2006). It is calculated based on all renewable infeed forecasts for a particular day $t + k$ received prior to the defined cut-off. Specifically, let $\widehat{W}^{k}_{t+k,h}$ denote the predicted wind power infeed for hour h on day $t + k$, which has been received on day t and thus constitutes a forecast over horizon k. We consider the set $\{\widehat{W}^{j}_{t+k,h} : j \geq k\}$ and calculate the standard deviation normalized by the mean of all forecasts in the set. Performing the calculation for all hours and for wind as well as solar predictions, lends a 48×1 vector of normalized standard deviations that is observed for each considered forecasting horizon k over the out-of-sample test set. Following Granz era and Sekhposyan (2019), we consider the first principal component of the normalized standard deviations as conditioning variable. It should be noted that only information that has been available at or prior to the time of forecasting is used and that the transformation via principal component analysis simply reduces the dimensionality of the conditioning set.

4.2 Descriptive and energy economic results

In the following, we assess and compare the forecasting performance of the models for the entire year 2016. Except for the year 2020, which was strongly influenced by the Covid-19 pandemic[8], the considered year 2016 has the lowest base price level of the past 10 years (28.98 €/MWh) and

[8] The Covid-19 pandemic and the associated slowdown in public and economic life resulted in lower electricity demand and thus comparatively low wholesale electricity prices.

shows an average price volatility (coefficient of variation ~43%). The day-ahead prices are characterized by very high prices at the end of January and a very volatile December with initially rather high positive and then strongly negative prices (cf. upper panel of *Figure 2*). The price level was relatively low in the first half of the year due to low coal and gas prices. From August onwards, prices gradually increased and were very high, especially in Q4. In addition to increasing fuel prices, high electricity prices in France influenced the German spot market at the end of the year. French prices were particularly high because a number of nuclear power plants were temporarily shut down due to safety concerns and thus German exports to France were permanently high during that period.[9] The highest price of 104.96 €/MWh was realized on Tuesday, 8th November 2016 and the most negative price of -130.09 €/MWh on Mother's Day, 8th May 2016.

Table 2 provides descriptive statistics of the observed and predicted prices for the day-ahead horizon, i.e. $k = 1$. Note that the recursive and direct specifications are identical over the forecasting horizon of one day. There is not a single model that performs best in terms of all descriptive indicators at the same time. However, it is evident that the autoregressive and hybrid models perform better than the purely fundamental models (*ParFuM* and *FunR/FunD*) over the day-ahead horizon. The state-of-the-art model of the short-term EPF literature, i.e. *ArR/ArD*, is the best in terms of mean as well as minimum and one of the best in terms of the standard deviation, the number of negative prices, prices above 50 €/MWh as well as the average daily price range, where it is narrowly outperformed by the hybrid model ArLoR/ArLoD. Clearly, the purely fundamental models (*ParFuM* and *FunR/FunD*) perform considerably worse. In particular, they fail to adequately predict volatility and negative prices. This is in line with the evaluation literature for fundamental models, which states that fundamental models have systematic difficulties in reproducing and forecasting extreme prices. Interestingly, however, FunR/FunD and *ParFuM* show the best performance in terms of very high prices (75 €/MWh and above) and forecasting of the maximum. Moreover, a simple post-processing of the *ParFuM* price using a regression with seasonal effects brings the fundamental predictions much closer to the observed prices. Thus, post-processing may be very beneficial for a given fundamental model.

[9] A detailed analysis of the impact on the German French cross-border flows and German/Austrian day-ahead prices due to the inspection of the French nuclear power plants is presented by Rinne (2019).

Table 2: Descriptive statistics for horizon k=1. Bold values indicate the best three models regarding each descriptive indicator. Bold and underlined values highlight the best model.

$k = 1$	mean	s.d.	min	max	#neg	#ab50	#ab75	#ab100	mDS
Observed	29.04	12.50	-130.09	104.96	96	391	30	1	23.46
ParFuM	27.78	6.85	6.80	**95.96**	0	54	**8**	0	9.33
FunR/FunD	**28.48**	9.87	-14.88	**98.23**	13	173	**7**	0	22.56
ArR/ArD	**29.32**	**10.93**	**-25.20**	73.62	**40**	**379**	0	0	**22.76**
LoR/LoD	31.10	**10.98**	-6.98	67.10	39	336	0	0	25.25
FunArR/FunArD	28.42	10.56	**-24.74**	**92.01**	36	301	3	0	21.70
ArLoR/ArLoD	29.97	**11.27**	-18.96	74.29	**54**	**383**	0	0	**23.73**
FunLoR/FunLoD	29.86	10.33	-10.99	90.35	37	235	**4**	0	**22.87**
FullR/FullD	**29.40**	10.84	**-19.73**	85.67	**50**	**339**	3	0	22.49

s.d.= standard deviation; #neg=number of negative hours; #ab50/75/100 = number of positive prices above 50/75/100 €/MWh; mDS=mean daily spread defined as $mDS = \frac{1}{n}\sum_{t=\tau}^{T-k} DS_{t+k}$ with $DS_{t+k} = \hat{p}^{k}_{t+k,Max} - \hat{p}^{k}_{t+k,Min}$ and $n = T - \tau - k + 1$.

Table 3 presents the descriptive statistics for all models over the week-ahead horizon, i.e. $k = 7$. It should be noted that the recursive and direct specifications are no longer identical. As the forecasting horizon increases, the models overall reflect the price volatility and range more inaccurately: standard deviations, the number of correctly predicted negative and extreme positive prices decline significantly. Also, the purely autoregressive models *ArR* and *ArD* are not among the best performing models in terms of any of the considered indicators.

Table 3: Descriptive statistics for horizon k=7. Bold values indicate the best three models regarding each descriptive indicator. Bold and underlined values highlight the best models.

$k = 7$	mean	s.d.	min	max	#neg	#ab50	#ab75	#ab100	mDS
Observed	29.04	12.50	-130.09	104.96	96	391	30	1	23.46
ParFuM	25.12	5.49	-97.05	60.51	7	14	0	0	6.63
FunR	**25.82**	**10.33**	-291.57	67.31	21	25	0	0	**23.21**
ArR	30.34	8.76	0.91	64.89	0	95	0	0	22.83
LoR	28.98	**10.45**	-30.99	67.84	**42**	120	0	0	25.08
FunArR	25.36	9.12	**-103.07**	68.63	14	32	0	0	20.84
ArLoR	**27.69**	**10.45**	-31.07	69.86	**56**	106	0	0	**24.02**
FunLoR	27.17	9.37	-30.36	70.18	**42**	49	0	0	22.26
FullR	26.45	9.55	-27.21	**71.99**	**50**	54	0	0	21.43
FunD	30.00	9.42	**-103.39**	71.39	8	158	0	0	22.81
ArD	30.79	8.96	1.21	64.65	0	143	0	0	22.68
LoD	31.49	9.61	-7.29	63.86	6	**208**	0	0	24.18
FunArD	29.85	9.38	**-102.84**	70.72	6	166	0	0	22.40
ArLoD	**31.13**	9.40	-7.91	62.46	5	**196**	0	0	**23.11**
FunLoD	30.91	9.37	-6.86	**72.62**	2	184	0	0	22.81
FullD	30.77	9.29	-0.03	**77.12**	1	**199**	**2**	0	22.33

s.d.= standard deviation; #neg=number of negative hours; #ab50/75/100 = number of positive prices above 50/75/100 €/MWh; mDS=mean daily spread defined as $mDS = \frac{1}{n}\sum_{t=\tau}^{T-k} DS_{t+k}$ with $DS_{t+k} = \hat{p}^{k}_{t+k,Max} - \hat{p}^{k}_{t+k,Min}$ and $n = T - \tau - k + 1$.

The results of the score-based evaluation are shown in Table 4, where we report the MAE values across the out-of-sample period for each individual horizon k and across all horizons. For the day-ahead horizon ($k = 1$), *FullR/FullD* constitutes the overall best model and exhibits a slightly lower MAE than *FunArR/FunArD*. Also, the purely autoregressive model *ArR/ArD* is not among the best models and is outperformed by the combined models. In addition, the *ParFuM*, i.e. the fundamental model without any postprocessing, does not constitute the worst performing model

and no superiority in forecasting performance of *ParFuM*-based models can be established. Beyond the day-ahead horizon, the results reported in Table 4 confirm what the analysis of the descriptive statistics has already suggested and show that the level of the forecast error increases with the forecast horizon for all models, which corresponds to the heightened uncertainty associated with forecasting further into the future. The best performing model per horizon is a model from the direct specifications (*FullD*, *FunD* or *FunArD*). Interestingly, *ParFuM* exhibits the highest MAE value for each horizon, but the best performing model is always a *ParFuM*-based model. Across all forecasting horizons, the FunArD model achieves the best average performance, outperforming the *FunD* model slightly.

Table 4: Mean Absolute Error Values. Bold values indicate the best three models. Bold and underlined values highlight the best models with respect to the forecasting horizon.

	Horizon							
Model	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**All**
ParFuM	5.16	7.40	7.95	8.17	8.28	8.45	8.10	7.64
FunR	4.14	5.63	5.64	5.62	5.73	6.00	6.24	5.57
ArR	4.27	5.39	5.78	6.01	6.09	6.13	6.21	5.70
LoR	5.33	5.37	5.42	5.53	5.76	6.15	6.51	5.72
FunArR	**3.68**	5.00	5.58	5.75	5.95	6.14	6.41	5.50
ArLoR	3.99	4.94	5.36	5.57	5.84	6.09	6.56	5.48
FunLoR	4.17	4.95	4.98	5.08	5.27	**5.64**	6.02	**5.16**
FullR	**3.58**	**4.62**	5.10	5.30	5.54	5.79	6.27	5.17
FunD	4.14	4.91	**4.75**	**4.67**	**4.86**	**5.23**	**5.60**	**5.00**
ArD	4.27	5.48	5.81	6.02	6.08	6.21	6.24	5.98
LoD	5.33	5.59	5.75	5.83	6.03	6.42	6.60	6.04
FunArD	**3.68**	**4.75**	**4.75**	**4.70**	**4.89**	**5.26**	**5.56**	**4.98**
ArLoD	3.99	4.89	5.36	5.60	5.78	6.22	6.35	5.70
FunLoD	4.17	4.91	4.83	**4.88**	5.17	5.61	6.02	5.24
FullD	**3.58**	**4.55**	**4.74**	4.90	**5.16**	5.64	**5.98**	**5.16**

In order to identify drivers of forecast performance, the middle and lower panel of Figure 2 present selected weekly average forecasts and weekly WMAE results, respectively. We compare *ArR* with *ParFuM* and *FullD* (best MAE for $k \in \{1,2,3\}$ and among the best MAE for $k \in \{5,7\}$) as well as *FunArD* (best MAE over all horizons). All considered models perform best in weeks with low observed volatility and most of the hourly prices close to the corresponding weekly average. These are weeks in which less than 15% of the prices are above or below the weekly average +/- weekly standard deviation or, statistically speaking, the best forecasting results are achieved in weeks with platykurtic prices. In 2016, weeks with this characteristic occurred predominantly in the period from mid-April to early September (exclusive of May, in which forecasting quality typically suffers due to three German public holidays). Thus, one of the best forecasted weeks among all models is 25th-31st July 2016, which behaved almost like the preceding week, with very little volatility around the mean price level and no extreme prices (not a single price lay beyond one standard deviation from the weekly mean). All models exhibit the highest errors in weeks with many extreme positive (during the third week of January) and/or negative prices (at

the end of the year). More concretely, the weeks with unexpected high prices, e.g. 18th-24th Jan 2016, and very negative prices, e.g. 2nd-8th May 2016, and the three weeks at year end, e.g. 12th-31st Dec 2016, are the weeks with the highest forecasting errors across all models. In these weeks, more than 48% of the prices lay beyond one standard deviation from the weekly mean. In more than 10% of the hours, the prices lay even beyond three standard deviations from the weekly average price. Except the pre-Christmas week, these mentioned weeks exhibit leptokurtic price patterns.[10]

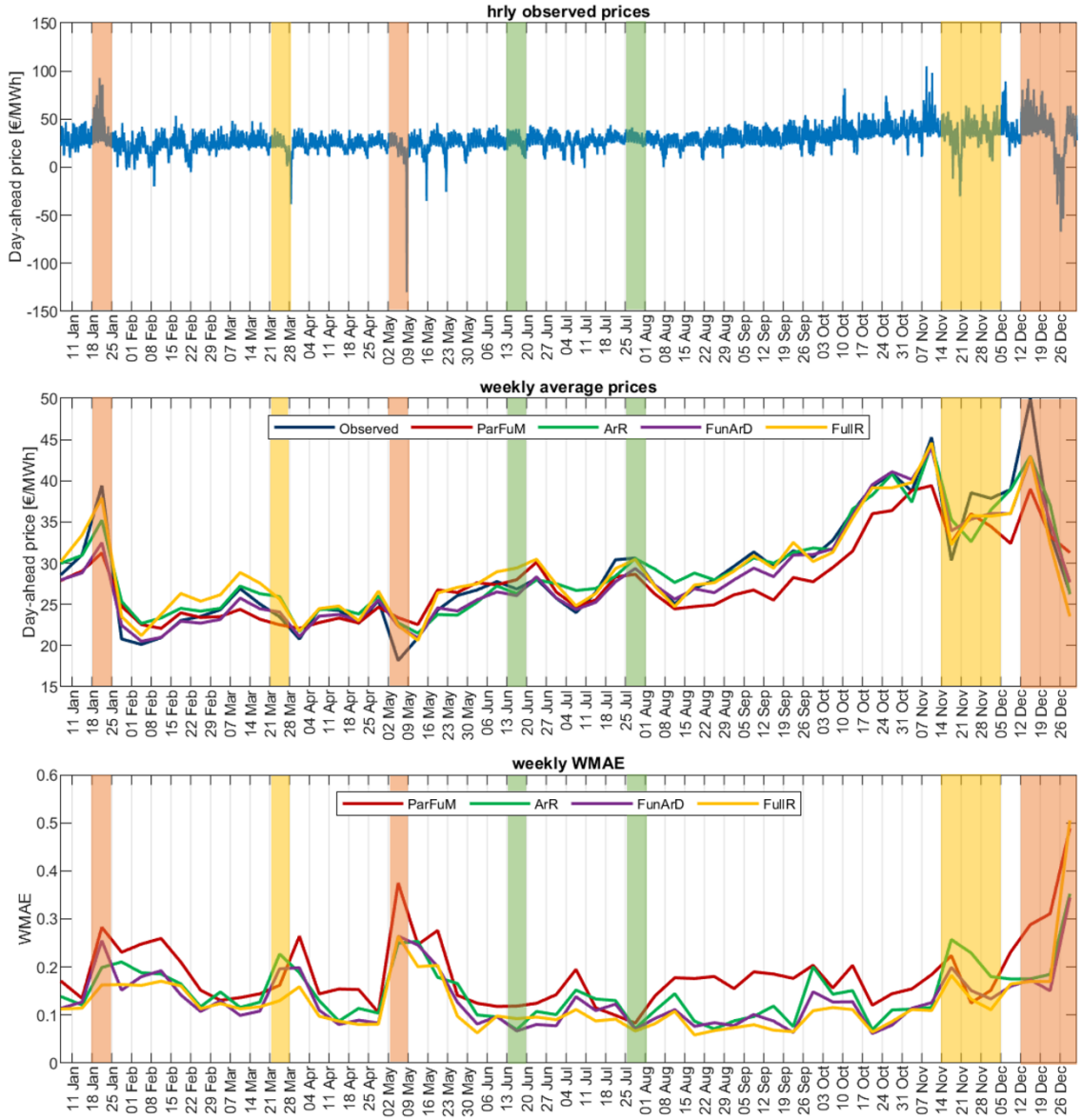

Figure 2: Hourly and weekly price structures and selected model performance for k=1. Green marked weeks indicate periods that can be predicted particularly well by all models and red marked weeks are particularly bad across all models. Yellow highlighted weeks show exemplary periods in which the fundamental model performs better than the ArR model and thus fundamental information significantly improves the forecasting accuracy of the hybrid model FullR.

[10] The highest negative price peak (-130.09 €/MWh) occurred during Mother's Day on Sunday, 08th May 2016. The tested forecasting models also exhibit high error measures in this extreme hour.

The hybrid models *FunArD* and *FullR* show the lowest WMAE values in Figure 2. In most weeks, the *ParFuM* is outperformed by the state-of-the-art model *ArR*. Nevertheless, considering fundamental information via a hybrid model always leads to an improvement in forecasting quality. This effect is particularly large in weeks in which the *ParFuM* performs better than the *ArR* model and can thus make a particularly large contribution to the prediction quality in the hybrid models (yellow areas in Figure 2). We see this improving effect mostly when price patterns change significantly from one week to the next and thus models containing fundamental information perform significantly better than autoregressive-based models.

Figure 3 compares the hourly characteristics of the selected models for horizons $k = 1$ and $k = 7$. The mean forecasted prices of the *ParFuM* are comparatively less volatile over all hours of the day, which is in line with the lows being insufficiently low and the peaks insufficiently high. This effect becomes even stronger with increasing horizon. The models with autoregressive elements are clearly better at forecasting the typical daily shape and capture both peaks and dips significantly better. Over the day-ahead horizon, the *FullR* model outperforms the established *ArR* model by forecasting the typical daily shape more precisely. Thus, the early morning hours (1-6) and the evening peak hours (18-22) in particular are better captured by *FullR*. As the forecasting horizon increases, the *FunArD* model yields better results than the other models, especially for prices from the morning peak until late evening.

Comparing the models in different weeks of the year, it can be seen that all models generally predict the typical daily structure and weekly profiles of the German spot market. However, the forecasts differ in the extent to which they capture these structures. The *ParFuM* identifies fundamental price changes and their directions but fails to capture the magnitude of changes related to diurnal (mostly load and PV) patterns, resulting in too few price spikes and thus too low volatility. The models containing AR components catch the peaks of the spikes much better and also capture negative price episodes more accurately. In a direct comparison of the *ArR* and the *FullD* model, we find that the *ArR* model tends to overrate swings and thus the forecasting quality suffers. In these hours, the *FullD* model performs better by taking fundamental relations into account. It can also be seen that although the *ArR* model achieves a fairly high standard deviation and a high daily price spread, this does not always lead to more accurate forecasting results, but rather simulates the general price behavior. Considering fundamental information can correct this effect and link forecasts with actual market conditions.

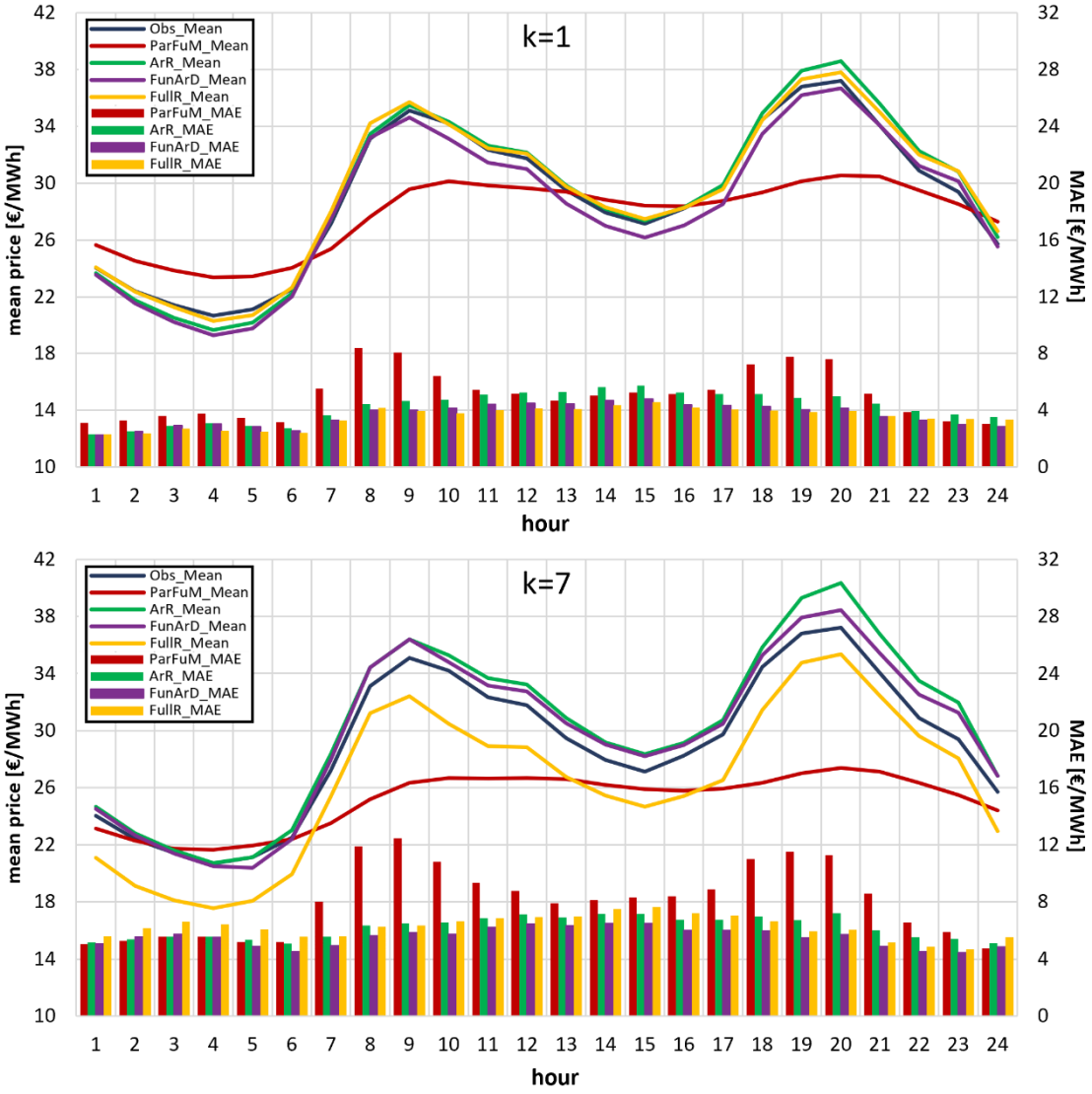

Figure 3: Hourly mean prices and MAEs for k=1 and k=7

Our results indicate on the one hand that the inclusion of autoregressive components enables originally purely fundamental forecasting models to better forecast throughs and peaks as well as positive and negative extreme prices and thus to better capture volatility. On the other hand, the inclusion of fundamental components enables originally purely autoregressive forecasting models to detach themselves from trends and effects of the estimation history and to take short-term systematic effects into account. The hybrid models can thus increase the forecasting quality for all considered horizons. The relative size of these effects on the MAE is above 10% in comparison to the *ArR* and above 30% in comparison to the *ParFuM*. The strongest improvements occur for horizons $k = 4$ and $k = 5$ (see Figure 4).

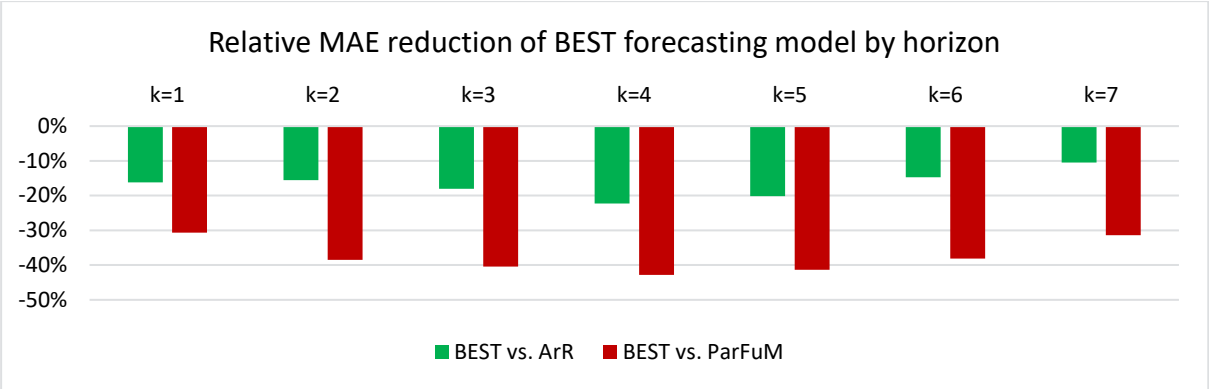

Figure 4: Relative MAE reduction of best forecasting model by horizon. The benchmark "best" models are chosen by MAE comparison. For $k = 1$ to $k = 3$ the best model is FullD, for $k = 4$ to $k = 6$ the best model is FunD and for $k = 7$ the best model is FunArD.

4.3 Test results

To validate the descriptive results, we apply the UEPA and CEPA tests introduced in Section 3.3. Figure 5 summarizes the test results. In each of the seven per-horizon panels, the left plot displays the p-value of the UEPA test against the p-value of the CEPA test. The red dashed lines indicate the five percent level of significance for each test and divide the plot into four quadrants. A point in the upper left quadrant indicates a pairwise model comparison, where in the UEPA test, the null hypothesis is rejected but the CEPA test does not reject the null hypothesis, whereas a point in upper right quadrant indicates that the null hypothesis of neither the UEPA nor the CEPA test is rejected. Recall that the former constitutes a counterintuitive result, although it may arise (e.g. Giacomini and White (2006)). Thus, the lower quadrants constitute the quadrants of primary interest, as they show pairwise comparisons where either both the UEPA and the CEPA null hypotheses are rejected (lower left quadrant) or where we can establish significant differences in forecasting performance conditional on renewable uncertainty (lower right quadrant). It should be noted that the markers for individual model pairs may overlap in the scatter plot.

In each of the seven per-horizon panels, the right chessboard plot displays which model pairs fall into the lower left or lower right quadrant of the left plot, respectively. The squares represent the results of both pairwise tests of equal predictive ability. A white square indicates that no significant improvement in forecasting performance can be uncovered when using the model in the corresponding row instead of the model in the corresponding column. A dark green square indicates that the row model significantly outperforms the column model in the unconditional sense (UEPA test), whereas a light green square indicates that the row model only outperforms the column model in the conditional sense (CEPA test). Thus, in the latter case, renewable uncertainty can explain the relative forecasting performance and the indicator variable $I^{ij,k}$ is larger than 0.5. If, for a given model pair, both associated squares are white, we fail to establish

a difference in predictive ability both in the unconditional and the conditional sense. It should be noted that both tests are considered at the five per cent level of significance and that all p-values have been corrected using the Bonferroni and Holm method.

For the day-ahead horizon ($k = 1$), the results in Figure 5 confirm the preceding discussion based on Table 4. The overall best model *FullR* (and *FullD*) significantly outperforms all but the second-best model *FunArR* (and *FunArD*), whereas the *ParFuM* is significantly outperformed by many but not all other models. Interestingly, one of the most extensively tested models of the short-term EPF literature, i.e. *ArR* (and *ArD*), is significantly outperformed by the two best models, underscoring our preceding finding that the inclusion of a fundamental price signal helps forecasting performance.

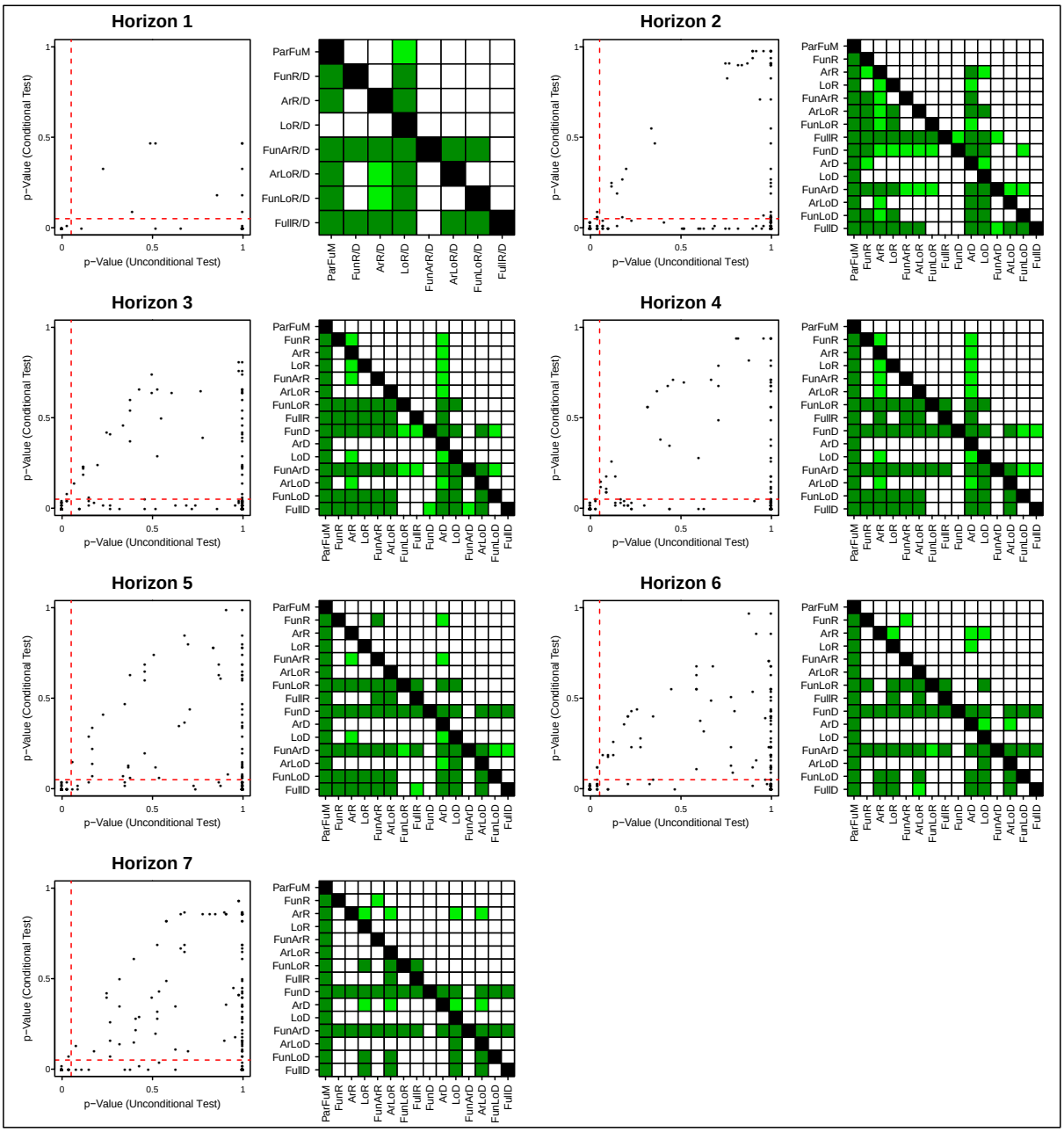

Figure 5: p-Values of Unconditional and Conditional Equal Predictive Ability Test. Note: The recursive and direct

specifications are identical for horizon 1 and thus the recursive and direct models are shown together in the corresponding chessboard plot (e.g., FunR/D represents FunR and FunD).

For the two-day-ahead horizon ($k = 2$), the results in Figure 5 show that *ParFuM* is significantly outperformed by all other models, whereas the overall best model *FullD* significantly outperforms all but four models, all of which nest the *ParFuM*. The *FullR* model significantly outperforms all other recursive models but fails to achieve the same for the direct models. Yet, we do not find the direct specifications to generally predict better than their recursive counterpart based on the same information set. One can establish additional statistically relevant differences in forecasting performance conditional on renewable uncertainty. The *FunD* model fails to significantly outperform most models unconditionally but provides on average better forecasts conditional on renewable uncertainty, suggesting that a fundamental price signal helps forecasting performance in times of heightened uncertainty about renewable infeed. Similar although less pronounced results are established for the *FunArD* model. Interestingly, differences in forecasting performance are also uncovered for both the *ArR* and the *ArD* model, which suggests that such a time series model for electricity prices may be less suited for forecasting beyond the day-ahead horizon. It should be noted that the majority of observed outperformances are conditional for the *ArR* model, whereas they are unconditional for the *ArD* model. Thus, if the time series model is to be used beyond the day-ahead horizon, it should be considered in its recursive form.

As the forecast horizon increases further, the *ParFuM* continues to be significantly outperformed by all other models. In addition, the forecasting performance of both full models (*FullR* and *FullD*) deteriorates significantly, whereas the forecasting performance of the two models with the lowest MAE across all horizons (*FunArD* and *FunD*) continues to improve. One can establish an increasing number of conditional differences in forecasting performance, which subsequently turn into statistically significant unconditional differences in forecasting performance. The number of conditional and unconditional outperformances of the *ArR* and the *ArD* model diminish as the forecasting horizon increases. Thus, the autoregressive models are not improving but they are continuously less bad than the remaining models, which exhibit deteriorations in forecasting performance of their own as the forecasting horizon increases.

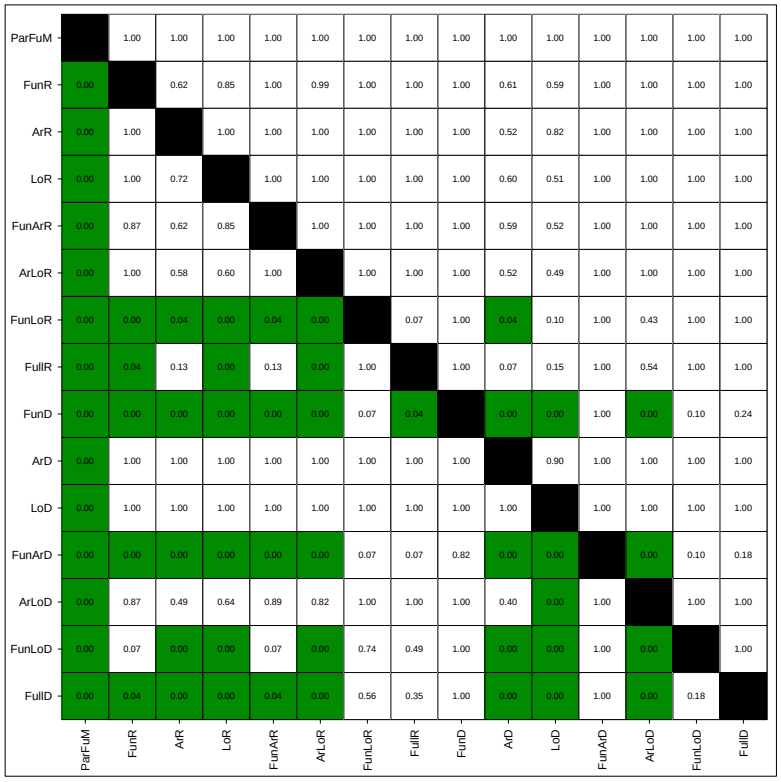

Figure 6: p-Values of average Superior Predictive Ability Test

The results of an evaluation of forecasting performance over multiple horizons jointly are presented in *Figure 6*. A square displays the p-value of a pairwise test of aSPA. White squares indicate that no significant improvement in forecasting performance can be uncovered when using the model in the corresponding row instead of the model in the corresponding column. A dark green square, however, indicates that the model in the row predicts significantly better on average across all horizons than the model in the column at the five per cent significance level. Figure 6 shows that *ParFuM* is again significantly outperformed by all other models, whereas the direct specifications based on *ParFuM* constitute the best models. Yet, only in two out of seven cases, the recursive model forecasts significantly worse than the respective direct model based on the same information set. The two autoregressive models (*ArR* and *ArD*) fail to significantly outperform any other model except *ParFuM*, suggesting that purely autoregressive models can be significantly improved for forecasting horizons greater than the day ahead. Thus, the results based on a joint evaluation across horizons confirm the preceding findings for the individual forecasting horizons.

5 Conclusions

This paper investigates electricity price forecasts for the short term. We do not only consider the widespread day-ahead forecasts, but cover horizons of up to a week ahead, well within the loose notion of short-term forecasting. We use well-established econometric models and a representative of fundamental models that are rather scarce in context of short-term EPF, as well

as their various combinations, so called hybrid models. We consider these models both as recursive and direct variants and examine and compare all model specifications for the individual forecasting horizons $t+1$ to $t+7$ as well as over this entire period using a case study of the German day-ahead market for the year 2016. Careful attention is paid to a realistic data setting and to use only data that was available at the historical hypothetical information cut-off time. At the same time, the models are constructed in a manner that they can also be executed in the short term and can thus be used in real applications in short-term trading. This is especially critical in the operation of fundamental models for the short term, as these often require a considerable database and exhibit long computation times as well as non-transparent price formation. We can overcome these difficulties by using *ParFuM*, which satisfies all requirements and at the same time generates fundamental prices of sufficient accuracy.

The best models across the individual horizons and across all horizons jointly are hybrid model approaches. They incorporate the common autoregressive elements of state-of-the-art EPF models and pair them with the fundamental information of the *ParFuM*. The purely fundamental models are outperformed over all horizons by the autoregressive and hybrid models, thus confirming the well-documented difficulties of fundamental models with regard to the reproduction of extreme prices, price volatility and non-linear relationships. Although the forecasting errors of all models increase with forecasting horizon, our results show that the hybrid model approaches have the lowest errors and significantly outperform the other models. They combine the strengths of autoregressive models in terms of capturing daily - even non-linear - structures with the immediate reactions of fundamental models to short-term events or fundamental changes in the market. Our case study results show that the MAEs improve by at least 10% over the purely autoregressive model and by at least 30% over the purely fundamental model across all forecasting horizons by using the best hybrid model.

Notwithstanding a very conscientious and detailed work, there is still room for improvement in our methodology:

- An approach such as that discussed by Marcos et al. (2019b), among others, could also further improve the hybrid models developed. They suggest not only to consider the fundamental equilibrium price as a regressor, but also other fundamental variables. Although we could not find evidence for simple residual load models to exhibit good forecasting performance, the residual load is part of the best performing model for the horizons $k=1$ to $k=3$ (*FullD*) and thus significantly improves the forecasting accuracy. This effect could be enhanced or supplemented by including other fundamental variables. For the German market area, the availability of elementary base load power plants (e.g. nuclear and lignite) or the must-run CHP quantities would be particularly conceivable.

- In general, the choice of the fundamental model is important and could be further validated. Due to the simplified structure of ParFuM, the fundamental input data used is of particular importance. However, fundamental forecast data beyond the day-ahead horizon is hardly available publicly and therefore corresponding models have to be developed or purchased. Here, other published forecasting methods (most notably for load and rather difficult for cross-border flows) could be applied, developed, or acquired from other third parties.
- Since the model quality of the *ParFuM* decreases rapidly beyond the day-ahead horizon, it would be interesting to investigate how sensitive the fundamental error is with regard to the input forecasts for load, CBCS and wind and PV for the longer horizons. Since an application of *ParFuM* is helpful despite its comparably high error values, the question arises at what point no exact forecast data but simple historical average values are sufficient as fundamental model inputs to integrate fundamental correlations in a hybrid model of longer horizons in a suitable way. This would entail considerable potential savings in terms of data preparation and input generation.
- There is also room for improvement with respect to the incorporation of the fundamental prices, e.g. a time-varying coefficient for the fundamental price and/or a nonlinear transformation to achieve even better “postprocessing” in the sense of a better fit of fundamental prices to the actual prices.
- Another possibility for improvements could be the extension of the considered lags for the models with autoregressive variables. $t-1$, $t-2$ and $t-7$ constitute the state-of-the-art lags but for direct model configurations over longer horizons these reduce to one or two past prices in our setting, which constitutes rather little information. We consider only the aforementioned lags to ensure comparability across all models.
- Since the individual model strengths of the specific configurations come into play at different points, one could consider a dynamic "super" model. This model could use a regime-switching approach to dynamically choose the proven strongest model for the considered period of the year, the hour of the day and the chosen forecasting horizon.

References

Bello, A., Bunn, D.W., Reneses, J., Munoz, A., 2017. Medium-Term Probabilistic Forecasting of Electricity Prices: A Hybrid Approach. IEEE Trans. Power Syst. 32, 334–343. https://doi.org/10.1109/TPWRS.2016.2552983.

Beran, P., Pape, C., Weber, C., 2019. Modelling German electricity wholesale spot prices with a parsimonious fundamental model – Validation & application. Utilities Policy 58, 27–39. https://doi.org/10.1016/j.jup.2019.01.008.

BNetzA, 2020. Kraftwerksliste der Bundesnetzagentur. BNetzA. https://www.bundesnetzagentur.de/EN/Areas/Energy/Companies/SecurityOfSupply/GeneratingCapacity/PowerPlantList/PubliPowerPlantList_node.html (accessed 3 December 2020).

Diebold, F.X., Mariano, R.S., 2002. Comparing Predictive Accuracy. Journal of Business & Economic Statistics 20, 134–144. https://doi.org/10.1198/073500102753410444.

EEX Transparency. Planned and unscheduled non-usability of all reported generating units; Masterdata power. https://www.eex-transparency.com (accessed 3 December 2020).

Energate-Messenger. Marex Spectron Marketdata via Energate-Messenger. www.energate-messenger.de/markt (accessed 3 December 2020).

ENTSO-E, 2015. Statistical Factsheet 2014. Statistical Factsheet. ENTSO-E, Brussels, 8 pp. https://eepublicdownloads.entsoe.eu/clean-documents/Publications/Statistics/Factsheet/entsoe_sfs2014_web.pdf (accessed 3 December 2020).

ENTSO-E, 2016. Statistical Factsheet 2015. Statistical Factsheet. ENTSO-E, Brussels, 8 pp. https://eepublicdownloads.entsoe.eu/clean-documents/Publications/Statistics/Factsheet/entsoe_sfs2015_web.pdf (accessed 3 December 2020).

ENTSO-E, 2017. Statistical Factsheet 2016. Statistical Factsheet. ENTSO-E, Brussels, 9 pp. https://eepublicdownloads.entsoe.eu/clean-documents/Publications/Statistics/Factsheet/entsoe_sfs_2016_web.pdf (accessed 3 December 2020).

ENTSO-E transparency. Transparency Platform. ENTSO-E. https://transparency.entsoe.eu/ (accessed 3 December 2020).

EPEX SPOT. Market Data. Day-Ahead Auction. EPEX SPOT. https://www.epexspot.com/en (accessed 7 December 2020).

Giacomini, R., White, H., 2006. Tests of Conditional Predictive Ability. Econometrica 74, 1545–1578.

Gianfreda, A., Ravazzolo, F., Rossini, L., 2020. Comparing the forecasting performances of linear models for electricity prices with high RES penetration. International Journal of Forecasting 36, 974–986. https://doi.org/10.1016/j.ijforecast.2019.11.002.

Gonzalez, V., Contreras, J., Bunn, D.W., 2012. Forecasting Power Prices Using a Hybrid Fundamental-Econometric Model. IEEE Trans. Power Syst. 27, 363–372. https://doi.org/10.1109/TPWRS.2011.2167689.

Granz era, E., Sekhposyan, T., 2019. Predicting relative forecasting performance: An empirical investigation. International Journal of Forecasting 35, 1636–1657. https://doi.org/10.1016/j.ijforecast.2019.01.010.

Gürtler, M., Paulsen, T., 2018. Forecasting performance of time series models on electricity spot markets: a quasi-meta-analysis. IJESM 12, 103–129. https://doi.org/10.1108/IJESM-06-2017-0004.

Hirth, L., Mühlenpfordt, J., Bulkeley, M., 2018. The ENTSO-E Transparency Platform – A review of Europe's most ambitious electricity data platform. Applied Energy 225, 1054–1067. https://doi.org/10.1016/j.apenergy.2018.04.048.

Hirth, L., Schumacher, M., 2015. How much Electricity do we consume? A Guide to German and European Electricity Consumption and Generation Data. Climate Change and Sustainable Development, 33 pp.

IEA, 2020. Monthly electricity statistics: Revised historical data. IEA. https://www.iea.org/reports/monthly-electricity-statistics (accessed 3 December 2020).

Kallabis, T., Pape, C., Weber, C., 2016. The plunge in German electricity futures prices – Analysis using a parsimonious fundamental model. Energy Policy 95, 280–290. https://doi.org/10.1016/j.enpol.2016.04.025.

Maciejowska, K., Uniejewski, B., Serafin, T., 2020. PCA Forecast Averaging—Predicting Day-Ahead and Intraday Electricity Prices. Energies 13, 3530. https://doi.org/10.3390/en13143530.

Maciejowska, K., Weron, R., 2013. Forecasting of daily electricity spot prices by incorporating intra-day relationships: Evidence form the UK power market, in: 2013 10th International Conference on the European Energy Market (EEM). 2013 10th International Conference on the European Energy Market (EEM 2013), Stockholm, Sweden. 27.05.2013 - 31.05.2013. IEEE.

Maciejowska, K., Weron, R., 2015. Forecasting of daily electricity prices with factor models: utilizing intra-day and inter-zone relationships. Comput Stat 30, 805–819. https://doi.org/10.1007/s00180-014-0531-0.

Marcellino, M., Stock, J.H., Watson, M.W., 2006. A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series. Journal of Econometrics 135, 499–526. https://doi.org/10.1016/j.jeconom.2005.07.020.

Marcjasz, G., Serafin, T., Weron, R., 2018. Selection of Calibration Windows for Day-Ahead Electricity Price Forecasting. Energies 11, 2364. https://doi.org/10.3390/en11092364.

Marcjasz, G., Uniejewski, B., Weron, R., 2020. Beating the Naïve—Combining LASSO with Naïve Intraday Electricity Price Forecasts. Energies 13, 1667. https://doi.org/10.3390/en13071667.

Marcos, R.A. de, Bello, A., Reneses, J., 2019a. Electricity price forecasting in the short term hybridizing fundamental and econometric modelling. Electric Power Systems Research 167, 240–251. https://doi.org/10.1016/j.epsr.2018.10.034.

Marcos, R. de, Bello, A., Reneses, J., 2019b. Short-Term Electricity Price Forecasting with a Composite Fundamental-Econometric Hybrid Methodology. Energies 12, 1067. https://doi.org/10.3390/en12061067.

McCracken, M.W., 2019. Tests of Conditional Predictive Ability: Some Simulation Evidence. Working Paper 2019-011. Federal Reserve Bank of St. Louis, 20 pp.

McCracken, M.W., McGillicuddy, J.T., 2019. An empirical investigation of direct and iterated multistep conditional forecasts. J Appl Econ 34, 181–204. https://doi.org/10.1002/jae.2668.

Muniain, P., Ziel, F., 2020. Probabilistic forecasting in day-ahead electricity markets: Simulating peak and off-peak prices. International Journal of Forecasting 36, 1193–1210. https://doi.org/10.1016/j.ijforecast.2019.11.006.

Nowotarski, J., Weron, R., 2018. Recent advances in electricity price forecasting: A review of probabilistic forecasting. Renewable and Sustainable Energy Reviews 81, 1548–1568. https://doi.org/10.1016/j.rser.2017.05.234.

Öko-Institut, 2015. Aktueller Stand der KWK-Erzeugung (Dezember 2015), Berlin, 58 pp. https://www.oeko.de/oekodoc/2450/2015-607-de.pdf (accessed 1 November 2017).

Pape, C., Hagemann, S., Weber, C., 2016. Are fundamentals enough? Explaining price variations in the German day-ahead and intraday power market. Energy Economics 54, 376–387. https://doi.org/10.1016/j.eneco.2015.12.013.

Quaedvlieg, R., 2021. Multi-Horizon Forecast Comparison. Journal of Business & Economic Statistics 39, 40–53. https://doi.org/10.1080/07350015.2019.1620074.

Ringkjøb, H.-K., Haugan, P.M., Solbrekke, I.M., 2018. A review of modelling tools for energy and electricity systems with large shares of variable renewables. Renewable and Sustainable Energy Reviews 96, 440–459. https://doi.org/10.1016/j.rser.2018.08.002.

Rinne, S., 2019. Radioinactive: Do nuclear power plant outages in France affect the German electricity prices? Energy Economics 84, 104593. https://doi.org/10.1016/j.eneco.2019.104593.

Serafin, T., Uniejewski, B., Weron, R., 2019. Averaging Predictive Distributions Across Calibration Windows for Day-Ahead Electricity Price Forecasting. Energies 12, 2561. https://doi.org/10.3390/en12132561.

Taieb, S.B., Atiya, A.F., 2016. A Bias and Variance Analysis for Multistep-Ahead Time Series Forecasting. IEEE transactions on neural networks and learning systems 27, 62–76. https://doi.org/10.1109/TNNLS.2015.2411629.

U.S. Energy Information Administration. Petroleum & Other Liquids Data. U.S. Energy Information Administration. https://www.eia.gov/dnav/pet/hist/RBRTED.htm (accessed 3 December 2020).

Ugurlu, U., Tas, O., Kaya, A., Oksuz, I., 2018. The Financial Effect of the Electricity Price Forecasts' Inaccuracy on a Hydro-Based Generation Company. Energies 11, 2093. https://doi.org/10.3390/en11082093.

Uniejewski, B., Weron, R., 2018. Efficient Forecasting of Electricity Spot Prices with Expert and LASSO Models. Energies 11, 2039. https://doi.org/10.3390/en11082039.

Uniejewski, B., Weron, R., Ziel, F., 2018. Variance Stabilizing Transformations for Electricity Spot Price Forecasting. IEEE Trans. Power Syst. 33, 2219–2229. https://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. International Journal of Forecasting 30, 1030–1081. https://doi.org/10.1016/j.ijforecast.2014.08.008.

Weron, R., Ziel, F., 2019. Electricity price forecasting, in: Uğur, S., San, R. (Eds.), Routledge Handbook of Energy Economics. Routledge, London, p. 506.

Ziel, F., Steinert, R., 2016. Electricity price forecasting using sale and purchase curves: The X-Model. Energy Economics 59, 435–454. https://doi.org/10.1016/j.eneco.2016.08.008.

Ziel, F., Steinert, R., 2018. Probabilistic mid- and long-term electricity price forecasting. Renewable and Sustainable Energy Reviews 94, 251–266. https://doi.org/10.1016/j.rser.2018.05.038.

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

Appendix

A1 Estimation and forecasting equations of recursive and direct models

Table 5: Estimation and forecasting equations of recursive and direct models

Specification	Name	Formula	Price	Time-varying intercept	ParFuM	AR	Residual Load	Error term
-	*ParFuM*	E	$p_{t+1,h} =$		$p^{ParFuM,k}_{t+1,h}$			
-		F	$\hat{p}^{k}_{t+k,h} =$		$p^{ParFuM,k}_{t+k,h}$			
Recursive	*FunR*	E	$p_{t+1,h} =$	$\beta^{1}_{h,0}(t+1)$	$+\beta^{1}_{h,1} p^{ParFuM,1}_{t+1,h}$			$+\varepsilon_{t+1,h}$
		F	$\hat{p}^{k}_{t+k,h} =$	$\hat{\beta}^{1}_{h,0}(t+k)$	$+\hat{\beta}^{1}_{h,1} p^{ParFuM,k}_{t+k,h}$			
	ArR	E	$p_{t+1,h} =$	$\beta^{1}_{h,0}(t+1)$		$+\beta^{1}_{h,2} p_{t,h} + \beta^{1}_{h,3} p_{t-1,h} + \beta^{1}_{h,4} p_{t-6,h} + \beta^{1}_{h,5} p_{t,Max} + \beta^{1}_{h,6} p_{t,Min} + \beta^{1}_{h,7} p_{t,24}$		$+\varepsilon_{t+1,h}$
		F	$\hat{p}^{k}_{t+k,h} =$	$\hat{\beta}^{1}_{h,0}(t+k)$		$+\hat{\beta}^{1}_{h,2} \hat{p}^{k-1}_{t+k-1,h} + \hat{\beta}^{1}_{h,3} \hat{p}^{k-2}_{t+k-2,h} + \hat{\beta}^{1}_{h,4} \hat{p}^{k-7}_{t+k-7,h} + \hat{\beta}^{1}_{h,5} p^{k-1}_{t+k-1,Max} + \hat{\beta}^{1}_{h,6} p^{k-1}_{t+k-1,Min} + \hat{\beta}^{1}_{h,7} p^{k-1}_{t+k-1,24}$		
	LoR	E	$p_{t+1,h} =$	$\beta^{1}_{h,0}(t+1)$			$+\beta^{1}_{h,8} L^{1}_{t+1,h}$	$+\varepsilon_{t+1,h}$
		F	$\hat{p}^{k}_{t+k,h} =$	$\hat{\beta}^{1}_{h,0}(t+k)$			$+\hat{\beta}^{1}_{h,8} L^{k}_{t+k,h}$	
	FunArR	E	$p_{t+1,h} =$	$\beta^{1}_{h,0}(t+1)$	$+\beta^{1}_{h,1} p^{ParFuM,1}_{t+1,h}$	$+\beta^{1}_{h,2} p_{t,h} + \beta^{1}_{h,3} p_{t-1,h} + \beta^{1}_{h,4} p_{t-6,h} + \beta^{1}_{h,5} p_{t,Max} + \beta^{1}_{h,6} p_{t,Min} + \beta^{1}_{h,7} p_{t,24}$		$+\varepsilon_{t+1,h}$
		F	$\hat{p}^{k}_{t+k,h} =$	$\hat{\beta}^{1}_{h,0}(t+k)$	$+\hat{\beta}^{1}_{h,1} p^{ParFuM,k}_{t+k,h}$	$+\hat{\beta}^{1}_{h,2} \hat{p}^{k-1}_{t+k-1,h} + \hat{\beta}^{1}_{h,3} \hat{p}^{k-2}_{t+k-2,h} + \hat{\beta}^{1}_{h,4} \hat{p}^{k-7}_{t+k-7,h} + \hat{\beta}^{1}_{h,5} p^{k-1}_{t+k-1,Max} + \hat{\beta}^{1}_{h,6} p^{k-1}_{t+k-1,Min} + \hat{\beta}^{1}_{h,7} p^{k-1}_{t+k-1,24}$		
	ArLoR	E	$p_{t+1,h} =$	$\beta^{1}_{h,0}(t+1)$		$+\beta^{1}_{h,2} p_{t,h} + \beta^{1}_{h,3} p_{t-1,h} + \beta^{1}_{h,4} p_{t-6,h} + \beta^{1}_{h,5} p_{t,Max} + \beta^{1}_{h,6} p_{t,Min} + \beta^{1}_{h,7} p_{t,24}$	$+\beta^{1}_{h,8} L^{1}_{t+1,h}$	$+\varepsilon_{t+1,h}$
		F	$\hat{p}^{k}_{t+k,h} =$	$\hat{\beta}^{1}_{h,0}(t+k)$		$+\hat{\beta}^{1}_{h,2} \hat{p}^{k-1}_{t+k-1,h} + \hat{\beta}^{1}_{h,3} \hat{p}^{k-2}_{t+k-2,h} + \hat{\beta}^{1}_{h,4} \hat{p}^{k-7}_{t+k-7,h} + \hat{\beta}^{1}_{h,5} p^{k-1}_{t+k-1,Max} + \hat{\beta}^{1}_{h,6} p^{k-1}_{t+k-1,Min} + \hat{\beta}^{1}_{h,7} p^{k-1}_{t+k-1,24}$	$+\hat{\beta}^{1}_{h,8} L^{k}_{t+k,h}$	
	FunLoR	E	$p_{t+1,h} =$	$\beta^{1}_{h,0}(t+1)$	$+\beta^{1}_{h,1} p^{ParFuM,1}_{t+1,h}$		$+\beta^{1}_{h,8} L^{1}_{t+1,h}$	$+\varepsilon_{t+1,h}$
		F	$\hat{p}^{k}_{t+k,h} =$	$\hat{\beta}^{1}_{h,0}(t+k)$	$+\hat{\beta}^{1}_{h,1} p^{ParFuM,k}_{t+k,h}$		$+\hat{\beta}^{1}_{h,8} L^{k}_{t+k,h}$	
	FullR	E	$p_{t+1,h} =$	$\beta^{1}_{h,0}(t+1)$	$+\beta^{1}_{h,1} p^{ParFuM,1}_{t+1,h}$	$+\beta^{1}_{h,2} p_{t,h} + \beta^{1}_{h,3} p_{t-1,h} + \beta^{1}_{h,4} p_{t-6,h} + \beta^{1}_{h,5} p_{t,Max} + \beta^{1}_{h,6} p_{t,Min} + \beta^{1}_{h,7} p_{t,24}$	$+\beta^{1}_{h,8} L^{1}_{t+1,h}$	$+\varepsilon_{t+1,h}$
		F	$\hat{p}^{k}_{t+k,h} =$	$\hat{\beta}^{1}_{h,0}(t+k)$	$+\hat{\beta}^{1}_{h,1} p^{ParFuM,k}_{t+k,h}$	$+\hat{\beta}^{1}_{h,2} \hat{p}^{k-1}_{t+k-1,h} + \hat{\beta}^{1}_{h,3} \hat{p}^{k-2}_{t+k-2,h} + \hat{\beta}^{1}_{h,4} \hat{p}^{k-7}_{t+k-7,h} + \hat{\beta}^{1}_{h,5} p^{k-1}_{t+k-1,Max} + \hat{\beta}^{1}_{h,6} p^{k-1}_{t+k-1,Min} + \hat{\beta}^{1}_{h,7} p^{k-1}_{t+k-1,24}$	$+\hat{\beta}^{1}_{h,8} L^{k}_{t+k,h}$	
Direct	*FunD*	E	$p_{t+k,h} =$	$\gamma^{k}_{h,0}(t+k)$	$+\gamma^{k}_{h,1} p^{ParFuM,k}_{t+k,h}$			$+\varepsilon_{t+k,h}$
		F	$\hat{p}^{k}_{t+k,h} =$	$\hat{\gamma}^{k}_{h,0}(t+k)$	$+\hat{\gamma}^{k}_{h,1} p^{ParFuM,k}_{t+k,h}$			
	ArD	E	$p_{t+k,h} =$	$\gamma^{k}_{h,0}(t+k)$		$+\gamma^{k}_{h,2} p_{t,h} + \gamma^{k}_{h,3} p_{t-1,h} + \gamma^{k}_{h,4} p_{t+k-7,h} + \gamma^{k}_{h,5} p_{t,Max} + \gamma^{k}_{h,6} p_{t,Min} + \gamma^{k}_{h,7} p_{t,24}$		$+\varepsilon_{t+k,h}$
		F	$\hat{p}^{k}_{t+k,h} =$	$\hat{\gamma}^{k}_{h,0}(t+k)$		$+\hat{\gamma}^{k}_{h,2} p_{t,h} + \hat{\gamma}^{k}_{h,3} p_{t-1,h} + \hat{\gamma}^{k}_{h,4} p_{t+k-7,h} + \hat{\gamma}^{k}_{h,5} p_{t,Max} + \hat{\gamma}^{k}_{h,6} p_{t,Min} + \hat{\gamma}^{k}_{h,7} p_{t,24}$		
	LoD	E	$p_{t+k,h} =$	$\gamma^{k}_{h,0}(t+k)$			$+\gamma^{k}_{h,8} L^{k}_{t+k,h}$	$+\varepsilon_{t+k,h}$
		F	$\hat{p}^{k}_{t+k,h} =$	$\hat{\gamma}^{k}_{h,0}(t+k)$			$+\hat{\gamma}^{k}_{h,8} L^{k}_{t+k,h}$	
	FunArD	E	$p_{t+k,h} =$	$\gamma^{k}_{h,0}(t+k)$	$+\gamma^{k}_{h,1} p^{ParFuM,k}_{t+k,h}$	$+\gamma^{k}_{h,2} p_{t,h} + \gamma^{k}_{h,3} p_{t-1,h} + \gamma^{k}_{h,4} p_{t+k-7,h} + \gamma^{k}_{h,5} p_{t,Max} + \gamma^{k}_{h,6} p_{t,Min} + \gamma^{k}_{h,7} p_{t,24}$		$+\varepsilon_{t+k,h}$
		F	$\hat{p}^{k}_{t+k,h} =$	$\hat{\gamma}^{k}_{h,0}(t+k)$	$+\hat{\gamma}^{k}_{h,1} p^{ParFuM,k}_{t+k,h}$	$+\hat{\gamma}^{k}_{h,2} p_{t,h} + \hat{\gamma}^{k}_{h,3} p_{t-1,h} + \hat{\gamma}^{k}_{h,4} p_{t+k-7,h} + \hat{\gamma}^{k}_{h,5} p_{t,Max} + \hat{\gamma}^{k}_{h,6} p_{t,Min} + \hat{\gamma}^{k}_{h,7} p_{t,24}$		
	ArLoD	E	$p_{t+k,h} =$	$\gamma^{k}_{h,0}(t+k)$		$+\gamma^{k}_{h,2} p_{t,h} + \gamma^{k}_{h,3} p_{t-1,h} + \gamma^{k}_{h,4} p_{t+k-7,h} + \gamma^{k}_{h,5} p_{t,Max} + \gamma^{k}_{h,6} p_{t,Min} + \gamma^{k}_{h,7} p_{t,24}$	$+\gamma^{k}_{h,8} L^{k}_{t+k,h}$	$+\varepsilon_{t+k,h}$
		F	$\hat{p}^{k}_{t+k,h} =$	$\hat{\gamma}^{k}_{h,0}(t+k)$		$+\hat{\gamma}^{k}_{h,2} p_{t,h} + \hat{\gamma}^{k}_{h,3} p_{t-1,h} + \hat{\gamma}^{k}_{h,4} p_{t+k-7,h} + \hat{\gamma}^{k}_{h,5} p_{t,Max} + \hat{\gamma}^{k}_{h,6} p_{t,Min} + \hat{\gamma}^{k}_{h,7} p_{t,24}$	$+\hat{\gamma}^{k}_{h,8} L^{k}_{t+k,h}$	
	FunLoD	E	$p_{t+k,h} =$	$\gamma^{k}_{h,0}(t+k)$	$+\gamma^{k}_{h,1} p^{ParFuM,k}_{t+k,h}$		$+\gamma^{k}_{h,8} L^{k}_{t+k,h}$	$+\varepsilon_{t+k,h}$
		F	$\hat{p}^{k}_{t+k,h} =$	$\hat{\gamma}^{k}_{h,0}(t+k)$	$+\hat{\gamma}^{k}_{h,1} p^{ParFuM,k}_{t+k,h}$		$+\hat{\gamma}^{k}_{h,8} L^{k}_{t+k,h}$	
	FullD	E	$p_{t+k,h} =$	$\gamma^{k}_{h,0}(t+k)$	$+\gamma^{k}_{h,1} p^{ParFuM,k}_{t+k,h}$	$+\gamma^{k}_{h,2} p_{t,h} + \gamma^{k}_{h,3} p_{t-1,h} + \gamma^{k}_{h,4} p_{t+k-7,h} + \gamma^{k}_{h,5} p_{t,Max} + \gamma^{k}_{h,6} p_{t,Min} + \gamma^{k}_{h,7} p_{t,24}$	$+\gamma^{k}_{h,8} L^{k}_{t+k,h}$	$+\varepsilon_{t+k,h}$
		F	$\hat{p}^{k}_{t+k,h} =$	$\hat{\gamma}^{k}_{h,0}(t+k)$	$+\hat{\gamma}^{k}_{h,1} p^{ParFuM,k}_{t+k,h}$	$+\hat{\gamma}^{k}_{h,2} p_{t,h} + \hat{\gamma}^{k}_{h,3} p_{t-1,h} + \hat{\gamma}^{k}_{h,4} p_{t+k-7,h} + \hat{\gamma}^{k}_{h,5} p_{t,Max} + \hat{\gamma}^{k}_{h,6} p_{t,Min} + \hat{\gamma}^{k}_{h,7} p_{t,24}$	$+\hat{\gamma}^{k}_{h,8} L^{k}_{t+k,h}$	

A2 Highlight week model comparison for k=1

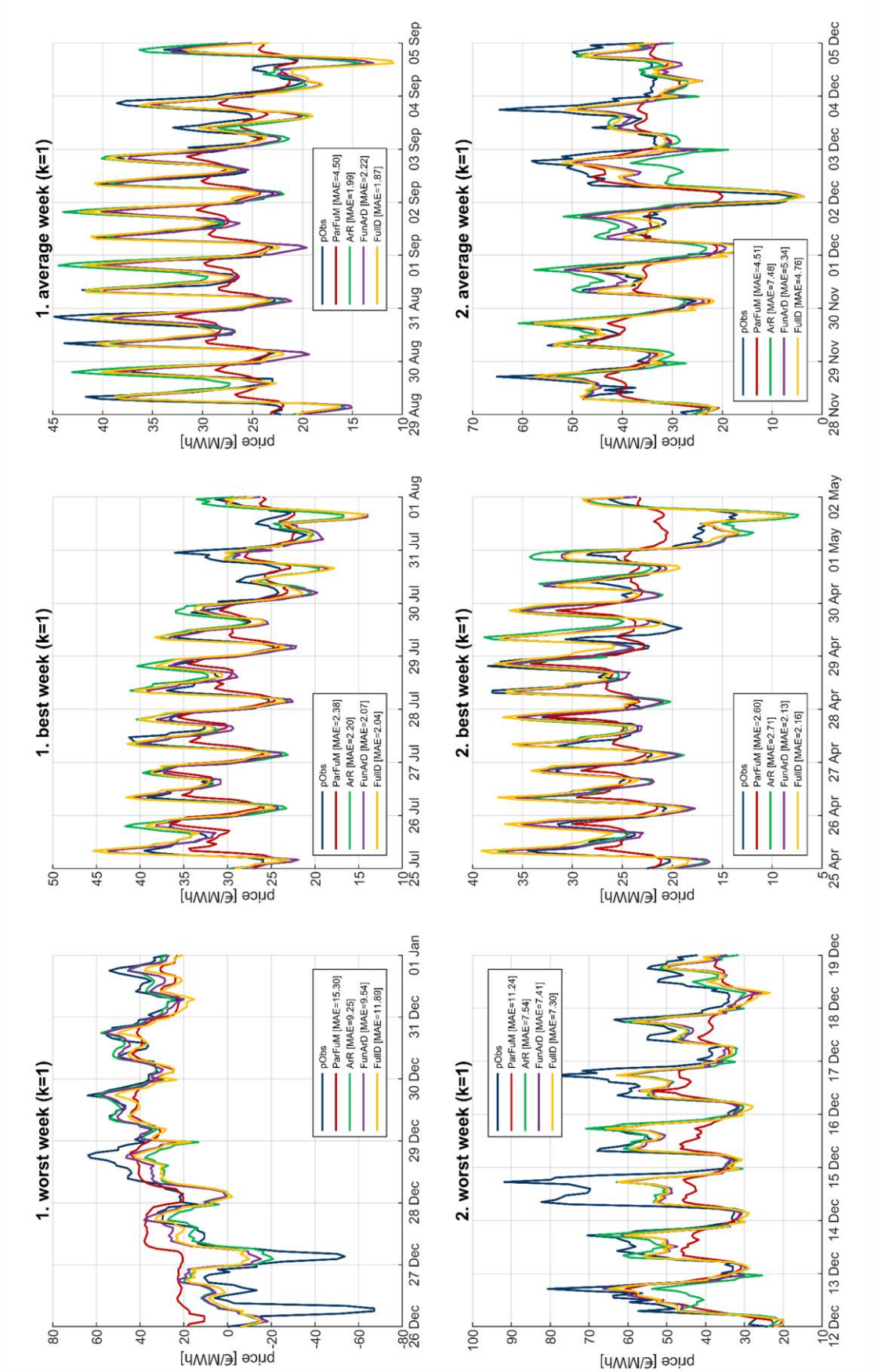

Figure 7: Comparison of full week hourly price forecasts of selected models at horizon k=1. Weeks are chosen by highest (worst), lowest (best)

Correspondence

M.Sc. Philip Beran
E-Mail philip.beran@uni-due.de

M.Sc. Arne Vogler
E-Mail arne.vogler@uni-due.de

Prof. Dr. Christoph Weber
Tel. +49 201 183-2966
E-Mail christoph.weber@uni-due.de

House of Energy Markets and Finance
Chair for Management Science and
Energy Economics
University of Duisburg-Essen,
Campus Essen
Universitätsstr. 12 | 45117 Essen

Tel. +49 201 183-2399
Fax +49 201 183-2703
E-Mail web.hemf@wiwi.uni-due.de
Web www.hemf.net

p