WORKing papers in Management Science

WORMS/20/02

PCA forecast averaging – predicting day-ahead and intraday electricity prices

Katarzyna Maciejowska[1]
Bartosz Uniejewski[1]
Tomasz Serafin[1]

[1] Department of Operations Research and Business Intelligence, Wrocław University of Science and Technology, Poland

WORMS is a joint initiative of the Management Science departments of the Wrocław University of Science and Technology, Wyb. Wyspiańskiego 27, 50-370 Wrocław, Poland

PCA forecast averaging – predicting day-ahead and intraday electricity prices

Katarzyna Maciejowska, Bartosz Uniejewski, Tomasz Serafin

Abstract

Recently, the development in combining point forecasts of electricity prices obtained with different length of calibration windows have provided an extremely efficient and simple tool for improving predictive accuracy. However, the proposed methods are strongly depended on expert knowledge and may not be directly transferred from one to another model or market. Hence, we consider a novel extension and propose to use Principal Component Analysis (PCA) to automate procedure of averaging over a rich pool of predictions. We apply PCA to a panel of over 650 point forecasts obtained for different calibration windows. The robustness of the approach is evaluated with three different forecasting tasks, i.e., forecasting day-ahead prices, forecasting intraday ID3 prices one day in advance and finally very short term forecasting of ID3 prices (i.e., six hours before delivery). The empirical results are compared using the Mean Absolute Error measure and Giacomini and White test for *conditional predictive ability* (CPA). The results indicate that PCA averaging not only yields significantly more accurate forecasts than individual predictions but also outperform other forecast averaging schemes.

Keywords: electricity price forecasting, EPF, day-ahead market, intraday market, forecast averaging, principal component analysis, decision-making

1. Introduction

In recent years, we have observed a dynamic transformation of energy markets, which encompasses changes of the generation structure and a creation of new trading opportunities. Since the establishment of competitive power exchanges, a growing share of electricity has been traded in day-ahead markets, where offers are placed before the noon of the day preceding the delivery. In order to allow for an adoption of trading position to a stochastic generation from intermittent energy sources, spot markets have been complemented by intraday and balancing markets. Operation in such a complex environment becomes challenging for many market participants, as it requires taking various operational decisions, for example generators need to decide, how much electricity to offer on a day-ahead market (see Maciejowska et al. (2019)) or how to structure the intraday trade (Kath & Ziel (2018)). Therefore, an accurate prediction of electricity prices becomes an important issue for utility managers.

Email address: (Katarzyna Maciejowska, Bartosz Uniejewski, Tomasz Serafin)

February 4, 2020

The literature is rich of publications focusing on modelling and forecasting of spot prices (see Weron, 2014; Nowotarski & Weron, 2018, for a comprehensive review). At the same time, there are few articles, which are dedicated to intraday markets Kath & Ziel (2018); Kiesel & Paraschiv (2017); Monteiro et al. (2016). Most of them focus on a very short term – a few hours ahead – forecast, as in Uniejewski et al. (2019a). This type of models could not be directly used by utilities when making operational decisions. Hence, there is a need to develop and evaluate models, which will reflect the timeline of trading decisions, for example day-ahead forecast of intraday prices, as in Maciejowska et al. (2019). On the other hand, electricity markets with their structural changes, seasonal fluctuations and occurrence of both positive and negative spikes, seeks new forecasting methods, which will help to overcome some of these issues. As shown by the literature, forecast averaging may be particularly beneficial, when it is difficult to indicate a single, best preforming model – as in case of electricity prices.

The idea of averaging forecasts has started about 50 years ago. Pioneering papers of Bates & Granger (1969) and Crane & Crotty (1967) inspired many other authors to develop the area of combining forecasts. Since late 60s, subject literature suggested that forecast combinations outperforms individual models, see e.g. Timmermann (2006); Wallis (2011); Nowotarski & Weron (2016). Hibon & Evgeniou (2005) make the very important comment on forecast combination superiority. They claim that not only averaging different forecast perform better in term of accuracy compared to the any individual forecast, but also in practical usage the combinations are always less risky. That is why, in the subject literature a lot of averaging methods have been proposed and empirically compared. Surprisingly, the simple arithmetic mean (i.e. all individual forecasts are weighed equally) is for sure the most popular and increadibly reliable approach Stock & Watson (2004); Genre et al. (2004). The same stands for taking median of forecasts, which in some cases can even outperform a simple average. Using the ordinary least squares (OLS) method to utilize forecast combination is another easy to implement approach in which weights are obtained from the linear regression, where individual forecasts are treated as the explanatory variables. Estimated weights, however, may exhibit unstable behavior (so-called *bouncing betas*) so we need to take into consideration that even slight fluctuations in the data can cause large changes in the final forecast. To prevent this Raviv et al. (2015) suggest to using constrained version of OLS, so-called *CLS averaging*, in which the weights are only positive and sum up to one.

The unwavering popularity of the simplest solutions is a strong evidence of how difficult task is to choose the right tools to average the forecasts. Recently authors suggested that possibly we should turn the question around, and instead of wondering how to average individual forecasts, we should rather develop the tools to select, which forecasts we want to average in the first place. A very important work in this context has been written by Diebold & Shin (2019). They introduce new technique of selection and averaging forecasts based on regularization technique. They propose new variants of the penalty function - egalitarian or partially egalitarian regularization. Even though the results were promising, the study was based only on a few dozen of observations, so it definitely needs further research.

Moreover, recently Hubicka et al. (2019); Marcjasz et al. (2018) shows that also the averaging of results obtained with the same model but calibrated to different portion of data improves the accuracy. In the recent papers Hubicka et al. (2019); Marcjasz et al. (2018); Serafin et al. (2019) regarding averaging forecasts across calibration windows of different lengths, argue that com-

bining forecasts from only a number of carefully selected calibration windows can significantly increase the forecasting performance and outperform the average of all predictions. Authors show that both for point Hubicka et al. (2019); Marcjasz et al. (2018) and probabilistic forecasts Serafin et al. (2019), considering the mix of short and long calibration windows bring statistically significant gains in terms of forecasting accuracy. The intuition behind this result is that electricity prices exhibit both long- and short-term seasonality – therefore combining short and long calibration windows allow us to capture the specific data behaviour. On the other hand, their studies also show that an inappropriate choice of window lengths may noticeably lower the accuracy of our predictions.

Although successful, the above methodology is based on a purely heuristic approach – in this paper we propose an alternative *automated* approach, based on the Principle Component Analysis (PCA) method. The PCA were developed by Stock & Watson (2002); Bai & Ng (2002); Bai (2003) and has been successfully used for modelling big panels of data. There are, however, only a few articles, in which PCA is applied for forecast combination. In Stock & Watson (2004) and Poncela et al. (2011) static factors are employed to extract information from a panel of predictions coming from different models/experts, in order to obtain point forecast of chosen macroeconomic variables. PCA was also adopted by Maciejowska et al. (2016), who used the method for construction of prediction intervals of electricity spot prices. In all the above applications, factors are estimated with relatively small and diversified panels.

In this paper, an alternative setup is explored, in which the panel of predictions is homogeneous and consists of a large number of forecasts based on the same model estimated with different calibration windows, as in Hubicka et al. (2019); Marcjasz et al. (2018); Serafin et al. (2019). We show that in case of slightly misspecified models, the proposed PCA–based procedure significantly outperforms both *ex–post* selected best performing calibration window and Weighted Averaged Windows (WAW) approach Marcjasz et al. (2018). In particular, the results of model for forecasting German day-ahead prices is almost 5% lower in terms of MAE in comparison to the optimal single calibration window and significantly better than any other considered averaging scheme.

The remainder of the paper is structured as follows. In Section 2, we present the datasets illustrating the German electricity market. Section 3 describes the experiment design, introduces Variance Stabilizing Transformation (VST) and defines models used for forecasting of day-ahead and intraday prices. Next, in Section 4, we discuse forecast averaging schemes and introduce PCA forecest combination approach. The performance of presented methods is evaluated in Section 5. Finally, conclusions of the research are presented in Section 6.

2. Datasets

In order to test the proposed methodology, we utilize a number of datasets from the German market – each of them spans from 1.01.2015 to 15.08.2019. We consider two different price time series: the day-ahead hourly electricity prices (*top* panel in Figure 1) and the corresponding time series linked to the intraday market – the ID3 index hourly prices (*bottom* panel in Figure 1). According to the official rules by EPEX SPOT EPEXSpot (2020), the ID3 index is calculated as

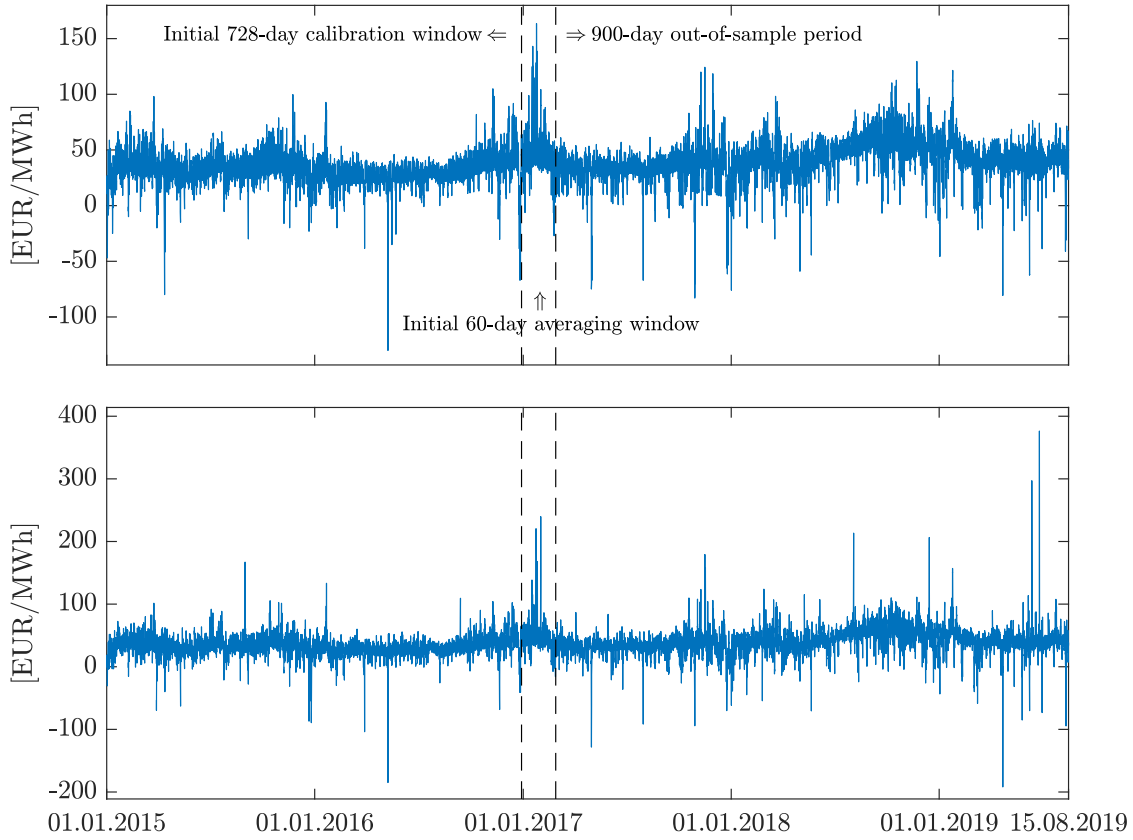

Figure 1: Day-ahead (*top*) and intraday (ID3) (*bottom*) hourly electricity prices for German market, from 1.01.2015 to 15.08.2019. The first vertical dashed line marks the end of the 728–day calibration period and the second marks the end of the initial 60-day calibration window for averaging forecasts.

the volume-weighted average price of all trades within 3 hours before the delivery of the product (up to 30 minutes before delivery).

Additionally we use data for three different types of exogenous variables: day-ahead consumption prognosis (*top* panel in Figure 2) as well as day-ahead wind and solar generation forecasts (respectively: *middle* and *bottom* panels in Figure 2). The wind generation forecast consists of aggregated forecasts of offshore and onshore generation forecasts. The missing or 'doubled' values (most often corresponding to the change in daylight saving time) were replaced by the arithmetic mean of neighbouring observations for the missing ones and the arithmetic mean of both values for 'doubled' hours.

3. Methodology

3.1. Calibration windows

Following the majority of forecasting literature, we consider a so called *rolling window scheme.* Similarly to Hubicka et al. (2019) and Marcjasz et al. (2018), instead of arbitrarily choosing a fixed calibration window length, we consider a set of 673 different window lengths – ranging from 56 (ca. two months) to 728 days (ca. two years) – obtained forecasts are later averaged (see 4). The

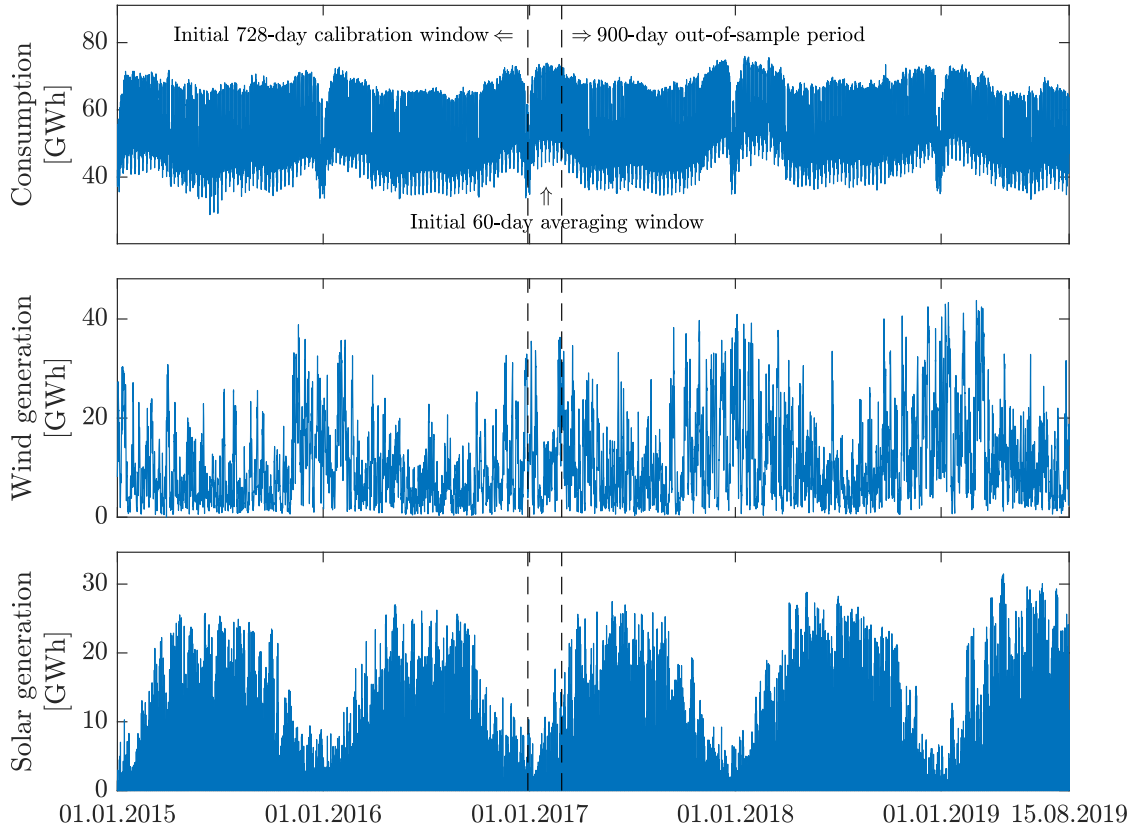

Figure 2: Hourly consumption prognosis (*top*), wind (*middle*) and solar (*bottom*) hourly generation forecasts. Data spans from 1.01.2015 to 15.08.2019. Note that for each series plot, the limits on the y axes are differ. The first vertical dashed line marks the end of the 728–day calibration period and the second marks the end of the initial 60-day calibration window for averaging forecasts.

first 728 (= 2×364) days are used for the initial model calibration. For each day in the period 29 December 2016 to 15 August 2019, we compute 24 point forecasts, one for each hour of the day. In addition, we use a 60–day fixed–length sample to calibrate the averaging methods (see 4) to obtain the average forecast for each hour of the days between 27 February 2017 and 15 August 2019. The forecasts for all 900 days of the out-of-sample period are evaluated in Section 5. The end of the initial 728-day model calibration window (i.e., 1 January 2015 to 28 December 2016) as well as the the end of the initial 60-day calibration window for averaging forecasts (i.e., 29 December 2016 to 26 February 2017) are indicated by dashed, vertical lines in Figures 1 and 2.

3.2. Variance Stabilizing Transformation

Since electricity prices exhibit strong seasonality as well as spiky behaviour, we follow the recommendation of Uniejewski et al. (2018) and apply a so called *variance stabilizing transformation* (VST) to all datasets (to time series of prices as well as to exogenous variables). We apply the N-PIT transformation which is based on the so called *probability integral transform*. The transformed price $X_{d,h}$ for day d and hour h is given by:

$$X_{d,h} = N^{-1}\left(\hat{F}_{P_{d,h}}(P_{d,h})\right), \tag{1}$$

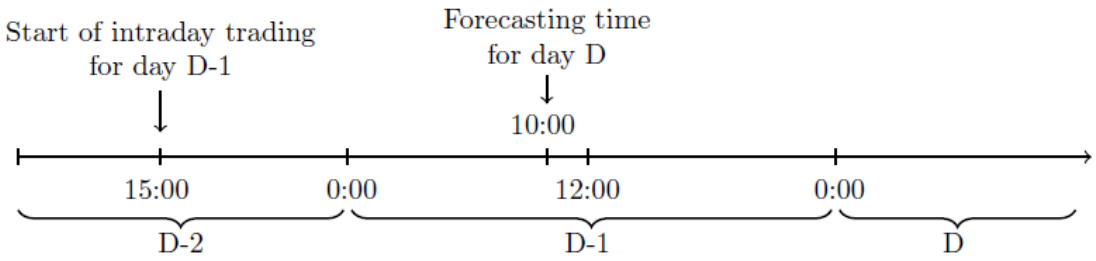

Figure 3: Market trading timeline; the forecasting of DA and IDA for day D is performed at 10:00 of day $D-1$.

where $P_{d,h}$ is the real observation for day d and hour h, $\hat{F}_{P_{d,h}}(\cdot)$ is the empirical cumulative distribution function of $P_{d,h}$ in the calibration sample, and N^{-1} is an inverse of the normal distribution function. We calibrate the models to transformed time series and then apply inverse transformation to the computed forecasts in order to obtain the price predictions:

$$P_{d,h} = \hat{F}_{P_{d,h}}\left(N(X_{d,h})\right). \tag{2}$$

3.3. Models

In this study, we consider both day-ahead and intraday (ID3) prices from the German electricity market. The latter ones can be forecasted day-ahead (before offering the electricity on the day-ahead market) or within a delivery day. The first approach is particularly important, when market participants need to decide, where to sell or buy energy (day ahead vs. intraday market). In such case, they need to predict day-ahead revenues from different trading strategies, as in Maciejowska et al. (2019). Another issue is to perform electricity price forecast of ID3 index six hours before the delivery. In the existing literature, most of researchers Uniejewski et al. (2019b); Narajewski & Ziel (2019); Janke & Steinke (2019) have focused on shorter forecast horizons (from four to three hours before the delivery), which do not leave enough time for adjusting trading strategies to market conditions. Therefore we extend the forecast horizon to 6 hours, as it enables exploiting the future price movements and optimizing the trade Edoli et al. (2016).

Let us focus first on the day-ahead spot prices, $DA_{d,h}$. In order to compute their point forecasts, autoregressive models with exogenous variables (ARX) are applied. This type of models has been extensively used in the EPF literature Ziel & Weron (2018); Uniejewski et al. (2018); Uniejewski & Weron (2018). The classical setup is expanded to include three exogenous variable: TSO forecasts of total load ($L_{d,h}$), wind ($W_{d,h}$) and solar ($S_{d,h}$) generation. The final model, denoted by **DA**, is

described by the following formula:

$$
\begin{aligned}
DA_{d,h} = & \underbrace{\beta_{h,1} DA_{d-1,h} + \beta_{h,2} DA_{d-2,h} + \beta_{h,3} DA_{d-7,h}}_{\text{autoregressive effects}} \\
& + \underbrace{\beta_{h,4} DA_{d-1,min} + \beta_{h,5} DA_{d-1,max}}_{\text{non-linear effects}} + \underbrace{\beta_{h,6} DA_{d-1,24}}_{\text{midnight price}} \\
& + \underbrace{\beta_{h,7} L_{d,h}}_{\text{load forecast}} + \underbrace{\mathbb{1}_{8<h<18} \beta_{h,8} S_{d,h}}_{\text{solar gen. forecast}} + \underbrace{\beta_{h,9} W_{d,h}}_{\text{wind gen. forecast}} + \underbrace{\sum_{i=1}^{7} \beta_{h,9+i} D_i}_{\text{weekday dummies}} + \varepsilon_{d,h},
\end{aligned} \tag{3}
$$

where $DA_{d-1,h}$, $DA_{d-2,h}$, $DA_{d-7,h}$ are the lagged day-ahead prices from previous day, two days before and a week before. $DA_{d-1,min}$ and $DA_{d-1,max}$ refer to the minimum and the maximum price from day $d-1$, $DA_{d-1,24}$ is the last known price from the previous day. Finally, $D_1, \ldots, D_7$ are weekday dummies accounting for the weekly seasonality. Note that the solar generation forecasts, $S_{d,h}$, are included in the model only for hours 9-17, due to the obvious lack of generation during night and early morning hours.

The second task is to predict *day-ahead* the intraday price, namely the value of the ID3 index for day d and hour h. We conduct the forecasting on the day preceding the delivery, as in the DA case. This implies that the intraday and spot prices are modelled in the same manner: all 24 prices for day d are forecasted at the same time, using the same pool of information. The model, denoted by **IDA** (Intraday Day-Ahead), has a structure similar to (3) and extends the model proposed by Maciejowska et al. (2019). It assumes that the data generating process of intraday prices could be described by the following equation:

$$
\begin{aligned}
ID3_{d,h} = & \underbrace{\beta_{h,1} ID3^{*}_{d-1,h} + \beta_{h,2} ID3_{d-2,h} + \beta_{h,3} ID3_{d-7,h}}_{\text{autoregressive effects}} + \underbrace{\beta_{h,4} DA_{d-1,h}}_{\text{Day-ahead market effect}} \\
& + \underbrace{\beta_{h,5} DA_{d-1,min} + \beta_{h,6} DA_{d-1,max}}_{\text{non-linear effects}} + \underbrace{\beta_{h,7} DA_{d-1,24}}_{\text{midnight price}} \\
& + \underbrace{\beta_{h,8} L_{d,h}}_{\text{load forecast}} + \underbrace{\mathbb{1}_{8<h<18} \beta_{h,9} S_{d,h}}_{\text{solar gen. forecast}} + \underbrace{\beta_{h,10} W_{d,h}}_{\text{wind gen. forecast}} + \underbrace{\sum_{i=1}^{7} \beta_{h,10+i} D_i}_{\text{weekday dummies}} + \varepsilon_{d,h},
\end{aligned} \tag{4}
$$

where $IDA^{*}_{d-1,h}$, $IDA_{d-2,h}$, $IDA_{d-7,h}$ are the lagged intraday prices. Due to the transaction timeline (see Figure 3) the prediction are performed at 10:00, when some of the intraday prices $IDA_{d-1,h}$ are yet not known. Therefore, a new variable, $IDA^{*}_{d-1,h}$ is constructed:

$$
ID3^{*}_{d-1,h} = \begin{cases} ID3^{\text{partial}}_{d-1,h} & \text{for } h > 10, \\ ID3_{d-1,h} & \text{for } h \leq 10, \end{cases} \tag{5}
$$

where $ID3^{\text{partial}}_{d-1,h}$ is the volume-weighted average price of all transactions for a certain product, that have been made up to the moment of forecasting. In case there were no transactions, $ID3^{\text{partial}}_{d-1,h}$ is replaced by the corresponding day-ahead price.

Finally, we build a model for a very short-term, 6 hours before the delivery, forecasting of intraday market, which we denote by **ID**. It is based on the results presented in Uniejewski et al. (2019b) and assumes that the ID3 price for day d and hour h is given by:

$$\begin{aligned} ID3_{d,h} = {} & \beta_{h,1} ID3_{d,h-6} + \beta_{h,2} ID3_{d-1,h} + \beta_{h,3} ID3^{\text{partial}}_{d,h} + \beta_{h,4} DA_{d,h} \\ & + \beta_{h,5} DA_{d,h-6} + \underbrace{\beta_{h,6} L_{d,h}}_{\text{load forecast}} + \underbrace{\mathbb{1}_{8<h<18} \beta_{h,7} S_{d,h}}_{\text{solar gen. forecast}} \\ & + \underbrace{\beta_{h,8} W_{d,h}}_{\text{wind gen. forecast}} + \underbrace{\sum_{i=1}^{7} \beta_{h,8+i} D_i}_{\text{weekday dummies}} + \varepsilon_{d,h}, \end{aligned} \tag{6}$$

where $ID3_{d,h-6}$ refer to the ID3 price six hours before delivery and $ID3_{d-1,h}$ is the price for the hour h on the previous day. The $ID3^{\text{partial}}_{d,h}$ is the volume-weighted average price of all transactions for the certain product, that have been made up to the 6 hours before the delivery (the moment of forecasting). Next two variables linked the Intraday and Day-Ahead markets. $DA_{d,h}$ refers to already known day-ahead price for day d and hour h, while $DA_{d,h-6}$ gives the newest information about price level difference between whose two markets. Rest of the predictors are just like in **DA** and **IDA** model.

Note that for better readability, we write $ID3_{d,h+i}$ to mark the product with the delivery i hours after (or before, for $i < 0$) the product (d, h) instead of using the correct notation $\left(d + \left\lfloor \frac{h+i}{24} \right\rfloor, h + i - 24 \left\lfloor \frac{h+i}{24} \right\rfloor\right)$.

4. Forecast averaging

The literature shows that the accuracy of forecasts depends on the length of the calibration window used for estimation of the model parameters. As shown by Hubicka et al. (2019); Marcjasz et al. (2018) this relationship could be non-monotonic and hence the selection of the optimal calibration window length becomes a complex task. On the other hand, diversity of outcomes provide a strong motivation for using *forecast averaging techniques*, which could improve the forecasting performance of models. Moreover, combining predictions could help to solve an issue of the optimal calibration window selection and reduce the model-specification risk.

Somehow interestingly, the concept of averaging forecasts across calibration windows of different lengths is relatively new in the field of electricity price forecasting. The recently published articles of Marcjasz et al. (2018) and Hubicka et al. (2019), were the first papers tackling this overlooked problem in a systematic way.

4.1. Weighed averaged windows

Simple arithmetic average of the selected predictions is one of the most popular forecast combining approaches. This method has been proved successful in a number of different studies across the econometric and forecasting literature. In the presented setup, the *averaged window* (AW) averaging scheme assumes equal weights for all forecasts estimated with the calibration windows of lengths $\tau \in \mathcal{T}$.

$$w = \frac{1}{\#\mathcal{T}} \tag{7}$$

Findings of Marcjasz et al. (2018); Hubicka et al. (2019) demonstrate that reduction of the set of window lengths used in forecast averaging, $\mathcal{T}$, could improve the method performance. They indicate that the average of predictions obtained with three short and three long calibration windows, in most cases, outperforms the single 'optimal' window as well as the average across all window lengths. The solution is also very efficient in terms of the computational cost – it requires calibrating the model to only six different sample lengths.

Marcjasz et al. (2018) extended the idea of simple averaging and proposed an averaging scheme called *weighed averaged windows* (WAW). The weights are computed using the the inverse of *Mean Absolute Error* (MAE) calculated over an averaging window of length D_{ave} (in Marcjasz et al. (2018) $D_{ave} = 1$)

$$w_d^{(\tau)} = \frac{\frac{1}{\mathrm{MAE}^{(\tau)}_{d-D_{ave}:d-1}}}{\sum_{\tau\in\mathcal{T}} \frac{1}{\mathrm{MAE}^{(\tau)}_{d-D_{ave}:d-1}}}, \tag{8}$$

where $w_d^{(\tau)}$ is the weight corresponding to a window of length τ on day d. $\mathrm{MAE}^{(\tau)}_{d_1:d_2}$ is computed as an average of absolute forecast errors obtained from the model calibrated to a training sample of length τ over all hours of days $d \in \{d_1, d_1 + 1, ..., d_2\}$:

$$\mathrm{MAE}^{(\tau)}_{d_1:d_2} = \frac{1}{24(d_2 - d_1 + 1)} \sum_{d=d_1}^{d_2} \sum_{h=1}^{24} |\varepsilon^{(\tau)}_{d,h}|, \tag{9}$$

where $\varepsilon^{(\tau)}_{d,h} = P_{d,h} - \hat{P}^{(\tau)}_{d,h}$. Using this approach, the past performance of each window length is taken into consideration and bigger weights are assigned to forecast obtained from windows that performed well in the past. Despite the computational efficiency and satisfying performance of this method, the choice of calibration window lengths has to be made in an ad-hoc manner and the inappropriate choice may have a significant impact on the forecasting performance.

4.2. PCA averaging

Majority of forecast combination approaches discussed in the literature, either uses a small number of predictions or emphasis the need of prediction selection. Alternatively, one could utilize the information included in a big panel of forecasts by using the Principal Component Analysis (PCA). The idea has been proposed by Stock & Watson (2004), who applied static factors to combine forecasts coming from different models. Similarly, Maciejowska et al. (2016) proposed Factor Quantile Regression Averaging (FQRA) to construct prediction intervals using a panel of point forecasts. In both articles, PCA averaging is applied to relatively small and diversified panels of forecasts based on 27–66 individual models.

In the presented set up, the panel of forecast consists of 673 individual predictions based on different calibration windows. Since the growth of the window size, τ, leads to more stable parameter estimates, the forecasts obtained with large windows, for example $\tau = 721$ and $\tau = 728$ are almost identical. This strong correlation, which is close to collinearity, impedes the classical, regression based methods of forecast averaging. In order to avoid such problems we propose a novel, fully-automated method for averaging forecasts, based on PCA.

In order to utilize all information from the averaging window, the data and predictions are treated as time series, with the time index, $t = 24(d-1)+h$, representing consecutive hours. Similar to WAW approach, we use the information from D_{ave} previous days. Additionally, the data is extended by 24 forecasts of hourly prices from the day d. Therefore the final averaging window consists of $24D_{ave} + 24$ observations. Let us denote by $\hat{P}_{t,\tau}$ the prediction of the variable P_t based on a calibration window of the length τ. The data set $\{\hat{P}_{t,\tau}\}$ could be interpreted as a panel, with the first dimension representing time and the second dimension describing the size of a calibration window. The averaging algorithm consist of the following steps:

1. For each time period t in the averaging window, estimate the mean ($\hat{\mu}_t$) and standard deviation ($\hat{\sigma}_t$) of individual forecasts across different τ
2. Standardize the predictions and the predicted variable

$$\tilde{P}_{t,\tau} = \frac{\hat{P}_{t,\tau} - \hat{\mu}_t}{\hat{\sigma}_t} \tag{10}$$

$$\tilde{P}_t = \frac{P_t - \hat{\mu}_t}{\hat{\sigma}_t} \tag{11}$$

3. Estimate the first $k = 1, ..., K$ principal components, $PC_{t,k}$, of a panel $\{\tilde{P}_{t,\tau}\}$, using the method described by Bai & Ng (2002); Stock & Watson (2004). Notice that the factors have a dimension $(24D_{ave} + 24) \times 1$ as they include the information on the price forecasts on the day d
4. Run a regression using observations from the averaging window, without the last 24 observation

$$\tilde{P}_t = \alpha + \sum_{k=1}^{K} \beta_k PC_{t,k} + \varepsilon_t \tag{12}$$

5. Compute the prediction of the normalized dependent variable on day d at hour h

$$\hat{\tilde{P}}_{24(d-1)+h} = \hat{\alpha} + \sum_{k=1}^{K} \hat{\beta}_k PC_{24(d-1)+h,k} \tag{13}$$

and transform it into its original units

$$\hat{P}_{24(d-1)+h} = \hat{\tilde{P}}_{24(d-1)+h} * \hat{\sigma}_{24(d-1)+h} + \hat{\mu}_{24(d-1)+h} \tag{14}$$

The role of standardization should be emphasised here. The mean, which changes between days, could be interpreted as a first common factor affecting the panel of forecasts. In particular, it represents the forecast based on long calibration windows. The predictions for big τ are, by construction, very similar to each other and have the largest input to the mean. On the other hand, the impact of long windows on the demeaned panel is balanced by larger (in absolute terms) and more variable deviations from mean for short calibration windows.

The standard deviation represents the forecast uncertainty and increases when short and long windows give different predictions. If the original data was used to estimate principal components,

the days with the highest risk would have the largest input to the panel variance and hence would impact strongly factor estimates. Thanks to standardization, all days are equally represented by common factors and the outcomes are stable, even when the outliers are included in sample used for forecast averaging. Finally, the variance estimates could be used for probabilistic forecasting, for example construction of prediction intervals, which is left for further research.

It should be noticed here that the described algorithm is conditioned on a number of factors, K, used in the regression (12). In order to make the choice of K data driven, we use Bayesian Information Criteria (BIC):

$$BIC(K) = \ln(\hat{\sigma}_K^2) + 2\frac{K \ln(24D_{ave})}{24D_{ave}}, \tag{15}$$

where $\hat{\sigma}_K^2$ is an estimated residuals variance from the model (12) with K components. For each day d, the optimal $\hat{K}$ is chosen, which minimizes corresponding BIC.

5. Results

5.1. Forecast Evaluation

We use the *Mean Absolute Error* (MAE) for the full out-of-sample test period of $D = 900$ days (i.e., 27.02.2017 to 15.08.2018, see Figure 1) as the main evaluation criterion. In the paper, two measures are considered

$$MAE_d^{(i)} = \frac{1}{24}\sum_{h=1}^{24} |\varepsilon_{d,h}^{(i)}| \tag{16}$$

$$MAE^{(i)} = \frac{1}{D}\sum_{d=1}^{D} MAE_d^{(i)} \tag{17}$$

where $\varepsilon_{d,h}^{(i)} = P_{d,h} - \widehat{P}_{d,h}^{(i)}$ is the prediction error at day d and hour h based on the averaging method i or $i = \tau$ for models without averaging. The first measure, $MAE_d^{(i)}$ describes the forecast accuracy for a given day, d, and is later use for statistical comparison of individual approaches. Finally $MAE^{(i)}$ describes the overall performance of the method (i). Recall, that the MAE is the most commonly used measure for evaluation forecast accuracy. In case of electricity markets, it reflects the average deviation of the revenue from selling 1 MWh from its expected level. Finally, MAE is more robust to the existence of the outliers than its RMSE counterpart, which makes it suitable for evaluation of electricity price forecasts.

Given couple of results it is hard to properly rank the models accuracy. To solve this issue, following Ziel & Weron (2018); Uniejewski et al. (2018), we introduce the *mean percentage deviation from the best* (m.p.d.f.b.) benchmark, inspired by the m.d.f.b. measure used in Weron & Misiorek (2008); Nowotarski et al. (2014) for comparing models. The m.p.d.f.b. measure for model i compares the model's performance to the best benchmark (it is the best performing calibration window length for each of models $j = DA, IDA, ID$:

$$\text{m.p.d.f.b.}_i = \frac{1}{3}\sum_{j=1}^{3} \frac{MAE_j^{(i)} - MAE_i^{(\tau=\text{Best})}}{MAE_i^{(\tau=\text{Best})}} \times 100\%, \tag{18}$$

The obtained MAE values can be used to provide a ranking of models. Unfortunately, they do not allow to draw statistically significant conclusions on the outperformance of the forecasts of one model by those of another. Therefore, the *conditional predictive ability* (CPA) test of Giacomini & White (2006) is used to compare competitive outcomes. Note that the CPA test could be viewed as a generalization of the popular Diebold & Mariano (1995) test for *unconditional* predictive ability. Here, the test statistic, is computed using the vector of average daily MAE_d:

$$\Delta_{i,j,d} = MAE_d^{(i)} - MAE_d^{(j)}, \tag{19}$$

where $MAE_d^{(i)}$ is the Mean absolute error of forecast obtained with model i on day d. For each pair of window sets and each model we compute the p-value of the CPA test with null $H_0 : \boldsymbol{\phi} = 0$ in the regression Giacomini & White (2006):

$$\Delta_{i,j,d} = \boldsymbol{\phi}' \mathbb{X}_{d-1} + \varepsilon_d, \tag{20}$$

where $\mathbb{X}_{d-1}$ contains elements from the information set on day $d-1$, i.e., a constant and $\Delta_{X,Y,d-1}$.

5.2. Point forecast results

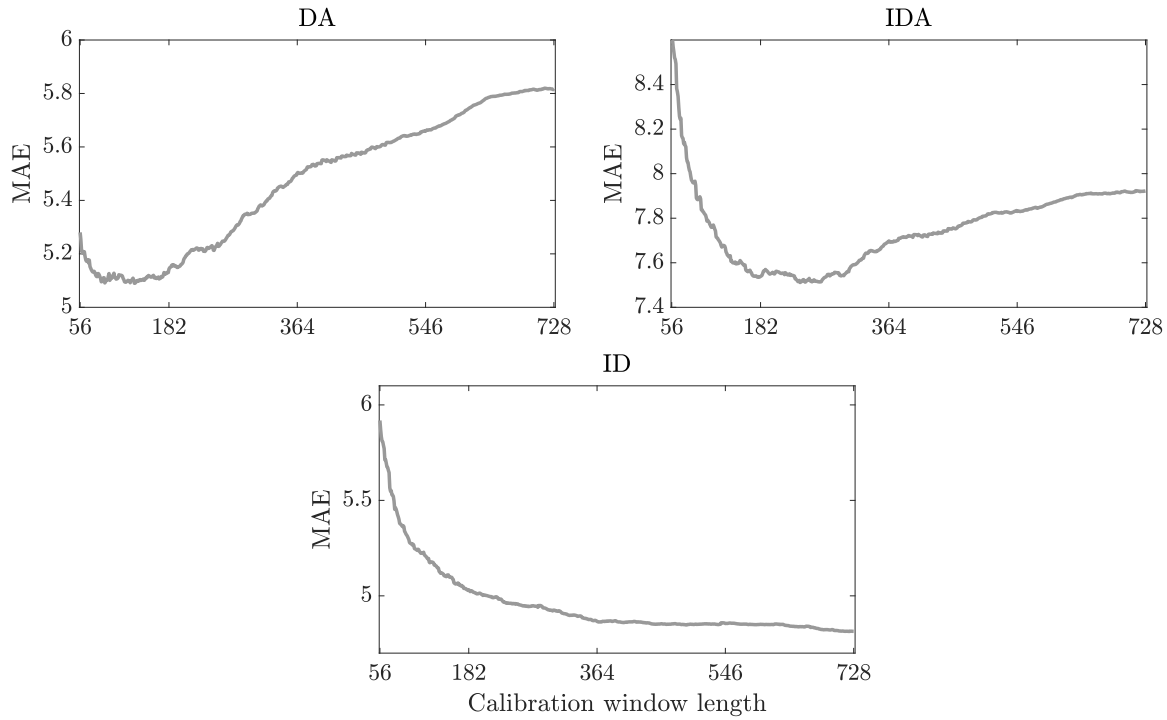

Figure 4: Mean Absolute Error as the function of the calibration sample length. The results for *DA* (*top-left*), *IDA* (*top-right*) and *ID* (*bottom*) models are evaluated over the whole 900-day out-of-sample test period.

As mentioned earlier, in this paper we consider training samples of lengths ranging from 56 to 728 days. Obviously, since the same model calibrated to a sample of different lengths produces differing forecasts, this gives us 673 different 'sub-models' for each of the models. The forecasting

Table 1: Mean absolute errors (MAE) and Mean Percentage Deviation From the Best benchmark (m.p.d.f.b) of the forecast obtained with all three considered model in the 900-days out-of sample period from 27.02.2017 to 15.08.2019. In the first four rows reported are result for model calibrated on single in-sample length, whereas the bottom rows corresponds to models using one of the averaging technique. The percentage change compared to the best *es-post* length of calibration window is reported in columns %chng. Note that the best model in terms of MAE is underlined.

	DA		**IDA**		**ID**		**m.p.d.f.b**
Benchmarks	**MAE**	**%chng**	**MAE**	**%chng**	**MAE**	**%chng**	**%chng**
56	5.280	3.744%	8.658	15.260 %	5.918	22.962%	13.989%
364	5.503	8.110%	7.698	2.484%	4.866	1.098%	3.897%
728	5.818	14.308%	7.922	5.470%	4.814	0.021%	6.600%
Best	5.090 (133)	–	7.511 (238)	–	4.813 (722)	–	–
Averaging							
AW(56:728)	5.199	2.146%	7.458	-0.712%	4.849	0.745%	0.726%
WAW(56:728)	5.160	1.371%	7.446	-0.871%	4.847	0.707%	0.402%
AW(56.84.112.714.721.728)	4.990	-1.964%	7.408	-1.373 %	4.929	2.404%	-0.311%
WAW(56.84.112.714.721.728)	4.930	-3.124%	7.384	-1.700%	4.907	1.953%	-0.957%
PCA(1)	4.890	-3.918%	**7.360**	**-2.012%**	4.860	0.979%	**-1.650%**
PCA(2)	4.883	-4.068%	7.437	-0.996%	4.862	1.022%	-1.347%
PCA(3)	**4.846**	**-4.784%**	7.446	-0.873%	4.869	1.159%	-1.499%
PCA(4)	4.877	-4.186%	7.477	-0.456%	4.878	1.342%	-1.100%
PCA(BIC)	4.879	-4.134%	7.409	-1.356%	4.858	0.930%	-1.520%

performance is evaluated separately for each calibration window, and the results are shown in Figure 4 - each dot represents the MAE of forecasts obtained by calibrating the model to a sample of a certain length. Interestingly, the curves for *DA* and *IDA* models are not monotonic as one may expect, the forecasting error does not strictly fall with the increase of the calibration sample length. This behavior of MAE may suggest that the models are slightly misspecified due to, for example, assumed linearity, time invariant parameters or omitted variables. In such case, the parameter estimates are inconsistent and do not converge to their true values. On the other hand, the curve for the *ID* model is descending. The forecasting accuracy of this model increases with the length of the calibration window, leaving little room for improvement for averaging techniques.

Another conclusion that can be drawn is that none of the calibration window lengths would be the 'optimal' choice for all models – the best-performing calibration window for *DA* model is 91 days, whereas for *IDA* the best forecasting performance would be achieved when calibrating the model on a 255-day sample and for *ID* the longest calibration samples perform the best with the minimum at 722-day. This diversified pattern of behavior shows that there is a need for a more robust way of selecting the length of calibration windows.

5.3. Averaging results

Table 1 presents MAE and m.p.d.f.b results for forecasts obtained with the shortest (56-days), the one-year long (364-days) and the longest (728-days) calibration windows and compare them against a benchmark: the best (optimal) (*ex-post*) calibration window length. Next, outcomes of different averaging technique are reported, starting with AW/WAW for all $\tau \in \{56, 57, ..., 728\}$, AW/WAW for six selected window sizes $\tau \in \{56, 84, 112, 714, 721, 728\}$ as in Hubicka et al.

(2019); Marcjasz et al. (2018); Serafin et al. (2019) and PCA averaging with 1 to 4 factors. Finally, outcomes of PCA(BIC) scheme are presented, in which the number of components is estimated using BIC information criteria. The results are displayed in absolute terms (MAE) and relative to the benchmark (computed as a percentage difference, %chng).

The presented measures are computed with data ranging from 27.02.2017 to 15.08.2019, it is 900-days long out-of-sample period. The three considered models, it is *DA*, *IDA* and *ID*, are evaluated separately and their outcomes are shown in consecutive columns. Finally, the average performance of analyzed forecasting schemes are described by m.p.d.f.b. The results lead to several important conclusions

- In case of *DA* and *IDA* models, the averaged forecasts are more accurate than any of the individual predictions, including those based on the best *ex-post* calibration window length. The gains reach up to 4.784% and 2.012% for *DA* and *IDA*, respectively. At the same time, none of the combined predictions provide results better than the benchmark for *ID* model. This confirms our expectation that averaging across different calibration window lengths may not lead to any improvement of forecast accuracy for well specified models.

- When two similar averaging schemes, AW and WAW, are compared, the results indicate that the extension of WAW, originally proposed by Marcjasz et al. (2018), outperforms the simple arithmetic mean. The superiority of WAW is obtained for all models and both ranges of window lengths, $\mathcal{T}$.

- The outcomes for WAW averaging scheme show that pre-selection of six window lengths improves the forecast accuracy only in case of mis-specified models: *DA* and *IDA*. At the same time, results for *ID* suggest that, for well specified models, the *ad hoc* reduction of the τ dimension increases substantially the MAE measure.

- The PCA forecast averaging approaches lead to more accurate predictions of *DA* and *IDA* than any other combining schemes. They reduce MAE, relative to the benchmark, by 4.784% and 2.012%, respectively. For *ID* model, PCA gives worse results than WAW(56:728), which is the most accurate among averaging schemes. However, the difference is relatively small and equals only 0.23%.

- The optimal number of factors in PCA depends on the model. For *DA* the best accuracy is obtained for three factors, whereas for *IDA* and *ID* for one factor. The outcomes suggest two solutions for this issue. First, usage of PCA(1) leads to results, which are on average the best among all PCA specifications (see m.p.d.f.b. column).

- BIC information criteria is shown to be helpful in selecting the number of components. Although it could not beat the best PCA specification for individual models (*DA*, *IDA* and *ID*), it works very well on average.

Finally, the results are evaluated with the Giacomini-White test Giacomini & White (2006) for the norm of order one. The outcomes are presented on Figure 5, on which a non-black square indicates that the forecasts of a model on the *X*-axis are statistically more accurate than the forecasts

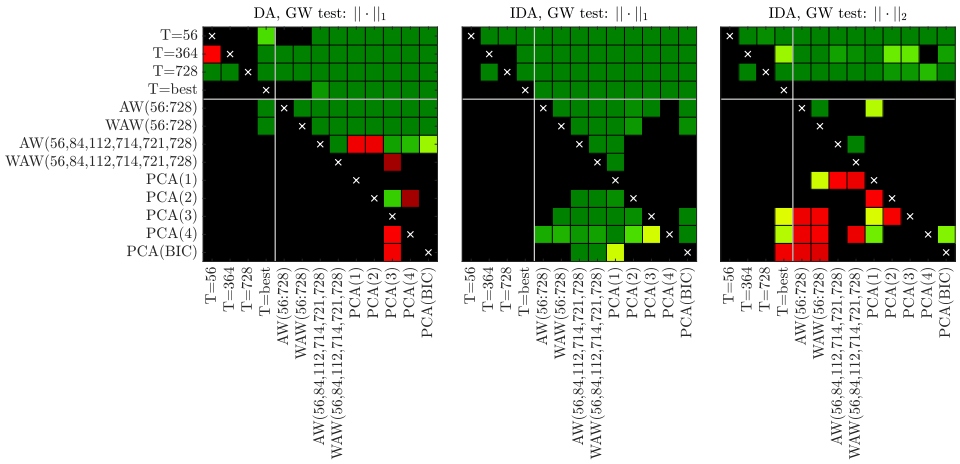

Figure 5: Results of the conditional predictive ability (CPA) test Giacomini & White (2006) for forecasts of all considered models. We use a heat map to indicate the range of the p-values – the closer they are to zero ($\rightarrow$ dark green) the more significant is the difference between the forecasts of a model on the X-axis (better) and the forecasts of a model on the Y-axis (worse).

of a model on the Y-axis. The results confirm previous findings and show that PCA(3) and PCA(1) outperform other forecasting methods for *DA* and *IDA* models. When the *ID* model is considered, the outcomes show that approaches based on the longest and the best calibration window lengths provide forecasts of the same accuracy, which outperform all other prediction methods. Moreover, for this market, forecasts computed with WAW(56:728) are statistically better than those obtained with other averaging schemes. Finally, the results confirm that BIC is useful in determining the optimal number of components used for averaging and hence could be an attractive alternative to *ad hoc* choice of K.

6. Conclusions

In this paper, we model and predict hourly electricity prices on the German market. We consider three forecasting setups: a day-ahead forecast of spot prices, a day-ahead forecast of intraday prices and a short-term, 6 hours ahead prediction of *ID*3 index. The analyzed problems reflect the decision process of market participants and could help in optimizing the selling/buying strategy as in Maciejowska et al. (2019); Kath & Ziel (2018).

We propose a novel approach for calculating the predictions of electricity prices, which utilize forecasts based on models calibrated to windows of different lengths. We extend the previous research of Marcjasz et al. (2018) and Hubicka et al. (2019), which focuses on *ad hoc* selection of the best set of calibration windows. In this study, we propose a Principal Component Analysis (PCA) method for forecast averaging, which enables automatic aggregation of information included in the large panel of predictions. The results indicate that the PCA averaging scheme can, on average, reduce the MAE measure of forecast accuracy, relative to the best *ex-post* calibration window length. It also outperforms other forecast averaging approaches, such as AW and WAW, described by Marcjasz et al. (2018).

Furthermore, we show that *DA*, *IDA* and *ID* models have different characteristics, which correlate with the forecast horizon. The performance of day-ahead forecasts of spot and intraday prices does not improve with the growth of the calibration window length, whereas the short-term predictions of ID3 get more accurate for the longest estimation windows. This difference impacts the potential gains from forecast averaging. For the *ID* model, none of the proposed methods could outperform the forecasts based on the longest calibration window. At the same time, the averaging – and in particular PCA forecast combination – results in a significant decrease of MAE for *DA* and *IDA* models. The forecast accuracy improves relative to the benchmark, by almost 5% for PCA(3) scheme and *DA* model. In case of *IDA* model, the error reduction reaches 2% for PCA(1) approach.

Finally, the results indicate that the simplest PCA forecast combination scheme, with only one component, provides on average the most accurate predictions. It outperforms not only other PCA specifications but also WAW averaging scheme and forecast based on a window of a predefined length. The PCA method with the number of component selected with BIC information criteria provides on average slightly worse predictions than PCA(1), but could be viewed as an interesting alternative for *ad hoc* selection of the number of components. We believe that these results encourage further research on PCA forecast averaging, which could be extended to interval and probabilistic forecasting and be applied to other commodity markets.

References

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, *71*, 135–171.

Bai, J., & Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, *70*, 191–221.

Bates, J. M., & Granger, C. W. J. (1969). The combination of forecasts. *Operational Research Quarterly*, *20*, 451–468.

Crane, D., & Crotty, J. (1967). A two-stage forecasting model: exponential smoothing and multiple regression. *Management Science*, *13*, B501–B507.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, *13*, 253–263.

Diebold, F. X., & Shin, M. (2019). Machine learning for regularized survey forecast combination: Partially-egalitarian lasso and its derivatives. *International Journal of Forecasting*, *35*, 1679–1691.

Edoli, E., Fiorenzani, S., & Vargiolu, T. (2016). Optimal trading strategies in intraday power markets. In *Optimization Methods for Gas and Power Markets: Theory and Cases* (pp. 161–184). Palgrave Macmillan UK.

EPEXSpot (2020). *Description of Epex Spot Market Indices.*. Https://www.epexspot.com/en/indices.

Genre, V., Kenny, G., Meyler, A., & Timmermann, A. (2004). Combining expert forecasts: Can anything beat the simple average? *International Journal of Forecasting*, *29*, 108–121.

Giacomini, R., & White, H. (2006). Tests of conditional predictive ability. *Econometrica*, *74*, 1545–1578.

Hibon, M., & Evgeniou, T. (2005). To combine or not to combine: Selecting among forecasts and their combinations. *International Journal of Forecasting*, *21*, 15–24.

Hubicka, K., Marcjasz, G., & Weron, R. (2019). A note on averaging day-ahead electricity price forecasts across calibration windows. *IEEE Transactions on Sustainable Energy*, *10*, 321–323.

Janke, T., & Steinke, F. (2019). Forecasting the price distribution of continuous intraday electricity trading. *Energies*, *12*, 4262.

Kath, C., & Ziel, F. (2018). The value of forecasts: quantifying the economic gains of accurate quarter-hourly electricity price forecasts. *Energy Economics*, *76*, 411–423.

Kiesel, R., & Paraschiv, F. (2017). Econometric analysis of 15-minute intraday electricity prices. *Energy Economics*, *64*, 77–90.

Maciejowska, K., Nitka, W., & Weron, T. (2019). Day-ahead vs. intraday – forecasting the price spread to maximize economic benefits. *Energies*, *12*, 631.

Maciejowska, K., Nowotarski, J., & Weron, R. (2016). Probabilistic forecasting of electricity spot prices using Factor Quantile Regression Averaging. *International Journal of Forecasting*, *32*, 957–965.

Marcjasz, G., Serafin, T., & Weron, R. (2018). Selection of calibration windows for day-ahead electricity price forecasting. *Energies*, *11*, 2364.

Monteiro, C., Ramirez-Rosado, I., Fernandez-Jimenez, L., & Conde, P. (2016). Short-term price forecasting models based on artificial neural networks for intraday sessions in the Iberian electricity market. *Energies*, *9*, 721.

Narajewski, M., & Ziel, F. (2019). Econometric modelling and forecasting of intraday electricity prices. *Journal of Commodity Markets*, . (DOI: 10.1016/j.jcomm.2019.100107).

Nowotarski, J., Raviv, E., Trück, S., & Weron, R. (2014). An empirical comparison of alternate schemes for combining electricity spot price forecasts. *Energy Economics*, *46*, 395–412.

Nowotarski, J., & Weron, R. (2016). To combine or not to combine? Recent trends in electricity price forecasting. *ARGO*, *9*, 7–14.

Nowotarski, J., & Weron, R. (2018). Recent advances in electricity price forecasting: A review of probabilistic forecasting. *Renewable and Sustainable Energy Reviews*, *81*, 1548–1568.

Poncela, P., Rodriguez, J., Sanchez-Mangas, R., & Senra, E. (2011). Forecast combination through dimension reduction techniques. *International Journal of Forecasting*, *27*, 224–237.

Raviv, E., Bouwman, K. E., & van Dijk, D. (2015). Forecasting day-ahead electricity prices: Utilizing hourly prices.

Serafin, T., Uniejewski, B., & Weron, R. (2019). Averaging predictive distributions across calibration windows for day-ahead electricity price forecasting. *Energies*, *12*, 256.

Stock, J. H., & Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, *97*, 1167–1179.

Stock, J. H., & Watson, M. W. (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting*, *23*, 405–430.

Timmermann, A. G. (2006). Forecast combinations. In G. Elliott, C. W. Granger, & A. Timmermann (Eds.), *Handbook of economic forecasting* (pp. 135–196). Elsevier.

Uniejewski, B., Marcjasz, G., & Weron, R. (2019a). On the importance of the long-term seasonal component in day-ahead electricity price forecasting: Part II – Probabilistic forecasting. *Energy Economics*, *79*, 171–182.

Uniejewski, B., Marcjasz, G., & Weron, R. (2019b). Understanding intraday electricity markets: Variable selection and very short-term price forecasting using lasso. *International Journal of Forecasting*, *35*, 1533–1547.

Uniejewski, B., & Weron, R. (2018). Efficient forecasting of electricity spot prices with expert and LASSO models. *Energies*, *11*, 2039.

Uniejewski, B., Weron, R., & Ziel, F. (2018). Variance stabilizing transformations for electricity spot price forecasting. *IEEE Transactions on Power Systems*, *33*, 2219–2229.

Wallis, K. F. (2011). Combining forecasts – forty years later. *Applied Financial Economics*, *21*, 33–41.

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.

Weron, R., & Misiorek, A. (2008). Forecasting spot electricity prices: A comparison of parametric and semiparametric time series models. *International Journal of Forecasting*, *24*, 744–763.

Ziel, F., & Weron, R. (2018). Day-ahead electricity price forecasting with high-dimensional structures: Univariate vs. multivariate modeling frameworks. *Energy Economics*, *70*, 396–420.